Nº 12-002-X au catalogue — Nº 2014001
ISSN 1710-2200

Le Bulletin technique et d'information des centres de données de recherche

Hiver 2014, vol. 6 no 1

Statistics Canada Statistique Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Comment accéder à ce produit

Le produit nº 12-002-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Publication autorisée par le ministre responsable de Statistique Canada

This publication is also available in English.

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, ses entreprises, ses administrations et les autres établissements. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

À propos du Bulletin technique et d'information

Le Bulletin technique et d'information des centres de données de recherche est un forum permettant aux utilisateurs actuels et prospectifs des centres de partager de l'information et les techniques d'analyse des données disponibles dans les centres. Le bulletin paraît au printemps et à l'automne, et l'on publiera à l'occasion des numéros spéciaux sur des questions d'actualité.

Objectifs

Les objectifs principaux de ce bulletin sont les suivants :

- l'accroissement et la diffusion de la connaissance concernant les données de Statistique Canada;
- les échanges d'idées parmi les utilisateurs membres des centres de données de recherche (CDR);
- l'aide aux nouveaux utilisateurs du programme des CDR; et
- offrir des occasions supplémentaires permettant aux chercheurs dans les centres de communiquer avec les spécialistes et divisions spécialisées au sein de Statistique Canada.

Contenu

Nous souhaitons publier des articles qui contribueront à accroître la qualité des travaux de recherche menés dans les centres de données de recherche de Statistique Canada et qui fourniront des conseils méthodologiques aux chercheurs travaillant dans les CDR.

Les articles portent principalement sur :

- l'analyse et la modélisation des données;
- la gestion des données;
- les pratiques statistiques, informatiques ou scientifiques éprouvées, ou au contraire, inefficaces;
- le contenu en données;
- les effets associés au libellé des questionnaires;
- la comparaison d'ensembles de données;
- l'examen des méthodes et de leur application;
- les problèmes associés aux données et leurs solutions;
- les outils innovateurs faisant appel aux enquêtes et aux logiciels pertinents des CDR.

Ceux et celles qui souhaitent soumettre un article au Bulletin technique et d'information sont priés de suivre les directives pour les auteurs.

Les rédacteurs et les auteurs tiennent à remercier les réviseurs de leurs commentaires précieux.

Articles

Estimation pondérée et estimation de la variance bootstrap pour analyser des données d'enquête : Comment les effectuer dans certains logiciels choisis?

Christian Gagné, Georgia Roberts & Leslie-Anne Keown
Centre de ressources en analyse de données, Direction de la méthodologie

Table des matières

Introduction

Ce document est destiné aux analystes/chercheurs qui envisagent d'effectuer de la recherche avec des données issues d'une enquête pour lesquelles des poids d'enquête et des poids bootstrap sont fournis dans les fichiers de données.

Ce document donne, pour certains progiciels choisis, des instructions sur la façon d'utiliser des poids d'enquête et des poids bootstrap pour effectuer une analyse de données d'enquête.

Les logiciels choisis aux fins de ce document sont:

1. SUDAAN 10
2. Les fonctions *svy* dans STATA 12
3. WesVar 5.1
4. Procédures *Survey* dans SAS 9.2
5. Bootvar 3.2 pour SAS

Les détails concernant d'autres progiciels qui peuvent effectuer une estimation pondérée appropriée et qui peuvent aussi faciliter le calcul des poids bootstrap d'enquête pour effectuer l'estimation de la variance seront ajoutés dans ce document dans le futur.

Nous fournissons d'abord une liste de contrôle d'un nombre d'éléments dont un analyste devrait tenir compte une fois qu'il a déterminé ses principales questions de recherche et qu'il a effectué une recherche préliminaire de sa ou ses sources de données potentielles. Certaines des questions contenues dans la liste de contrôle sont spécifiques au problème analytique particulier qui a été abordé, tandis que les autres questions sont reliées à un logiciel particulier qui est pris en compte. Afin de que la liste de contrôle soit exhaustive, elle peut contenir des questions qui ne concernent pas toutes les analyses.

Ensuite, nous présentons un exemple d'analyse effectuée à partir d'une enquête précise. Cet exemple sera utilisé dans le contexte de chacun des logiciels choisis pour ce document. D'autres exemples pourraient être ajoutés dans une édition ultérieure de ce document.

Enfin, nous donnons de brèves instructions sur la façon d'obtenir des estimations fondées sur des enquêtes pondérées, des estimations de la variance bootstrap (ainsi que d'autres erreurs de quantités souhaitées) et quelques tests statistiques classiques pour chaque progiciel. Même si ces directives sont seulement fournies pour les exemples choisis, nous donnons des renseignements sur l'étendue des analyses pondérées utilisant les poids bootstrap qui peuvent être effectuées par chaque logiciel.

La section de la bibliographie contient des publications qui sont mentionnées dans ce document, ainsi que certains livres qui contiennent des explications sur la raison et la manière selon laquelle le plan de sondage devrait être pris en compte au moment de l'analyse. Dans les annexes, nous donnons des renseignements sur d'autres sujets qui ont été identifiés par les chercheurs et les analystes qui analysent des données d'enquête, notamment une discussion sur la normalisation des poids. D'autres sujets seront ajoutés au fil du temps.

Afin que les chercheurs puissent tirer le maximum d'information de ce document, nous leur suggérons d'abord de commencer par réviser les listes de contrôle ci-dessous, et ensuite de faire la lecture de l'exemple et de l'achèvement de la liste de contrôle des problèmes pour cet exemple. Ensuite, ils peuvent consulter les sections du document qui ont trait au logiciel qu'ils envisagent d'utiliser pour effectuer leurs analyses. Ainsi, ils peuvent faire la lecture de la liste de contrôle de ce logiciel pour cet exemple avant d'utiliser ce logiciel pour pouvoir effectuer leurs analyses.

Liste de contrôle pour l'analyse faite au moyen de la pondération et de l'estimation de la variance bootstrap

Liste de contrôle des problèmes

1. Avez-vous exposé clairement en détail les questions auxquelles vous voulez répondre dans votre analyse?
2. Est-ce qu'un échantillon analytique convenable avec des variables appropriées a été déterminé?
3. Avez-vous déterminé clairement quels sont les types d'analyses qui devront être effectuées? (p. ex. types de statistiques descriptives, types de modèles)
4. Avez-vous conclu que les estimations pondérées et les estimations de la variance bootstrap (erreurs) sont propices pour les analyses? (Même si, pour la plupart des analyses, l'estimation pondérée et l'estimation de la variance fondée sur le plan de sondage au moyen du bootstrap peuvent être recommandées, il existe des situations où cela ne constitue peut-être pas la meilleure méthode pour effectuer une analyse.)
5. Avez-vous déterminé quels seront les tests et statistiques dont vous aurez besoin?
6. Avez-vous consulté le guide de l'utilisateur de l'enquête (et d'autres documents) et déterminé les éléments suivants :
   - **a.** règles de non-divulgation;
   - **b.** toutes les mises en garde sur les méthodes analytiques appropriées;
   - **c.** variable de poids d'enquête appropriée (p. ex. poids de la personne ou du ménage; poids longitudinal ou transversal) et les variables de poids bootstrap correspondantes;
   - **d.** si les variables du bootstrap moyen sont des poids bootstrap et, le cas échéant, le nombre de d'échantillons bootstrap qui ont été utilisées pour produire chaque poids de bootstrap moyen (qui sont nécessaires pour effectuer un ajustement de poids bootstrap)? [L'annexe C du guide de l'utilisateur de BootVar fournit ces renseignements pour plusieurs d'enquêtes.]

Liste de contrôle du logiciel

1. Avez-vous déterminé, si les estimations pondérées et les estimations de la variance bootstrap (erreurs) requises peuvent être calculées avec le logiciel que vous utilisez? S'il est possible de produire les statistiques/effectuer les tests nécessaires avec le logiciel que vous utilisez?
2. Si les tests ne peuvent être effectués et que les statistiques spécifiques souhaitées ne peuvent être calculées, est-il possible de développer un programme post-estimation pour calculer ces statistiques et effectuer les tests dans le logiciel que vous utilisez?
3. Est-il possible, dans le logiciel, de restreindre l'échantillon ou d'éliminer les observations qui ne concernent pas l'échantillon du fichier de données complet de l'enquête?
4. Si le poids de l'enquête, les poids bootstrap et les variables d'analyse ne sont pas dans le même fichier, savez-vous comment fusionner les différentes sources dans le logiciel que vous utilisez? (En supposant que le logiciel utilisé requiert que tous ces renseignements soient dans le même fichier.)
5. Tout en effectuant votre analyse, avez-vous vérifié les résultats de sortie de votre logiciel pour déterminer :
   - **a.** que la bonne taille de l'échantillon a été utilisée;
   - **b.** que la bonne variable de poids a été utilisée;
   - **c.** que l'ensemble complet des poids bootstrap a été utilisé;
   - **d.** que l'ajustement du bootstrap moyen a été effectué de la façon appropriée (si nécessaire);
   - **e.** s'il existait des échantillons bootstrap pour lesquels il n'était pas possible de faire des estimations?

Exemple tiré de l'Enquête sociale générale (ESG) : Utiliser les données du cycle 22 de l'ESG

Le programme de l'ESG, qui existe depuis 1985, permet de mener des enquêtes téléphoniques dans les dix provinces du Canada. L'ESG recueille des données transversales sur une base régulière qui permettent de faire une analyse des tendances. Il est reconnu pour sa capacité à mettre à l'essai et développer de nouveaux concepts qui se penchent sur des problèmes émergents. Le programme de l'ESG permet de recueillir des données sur des sujets sociaux dans le but de surveiller les changements dans les conditions de vie et le bien-être des Canadiens au fil du temps, et pour fournir des renseignements immédiats sur des questions précises liées aux politiques sociales d'intérêt qui concernent l'actualité ou qui sont émergentes.

Le cycle 22 est le deuxième cycle de l'ESG à recueillir des données sur l'engagement social et les réseaux sociaux. Le premier cycle à le faire, en 2003, était le cycle 17 – Engagement social.

La population cible du cycle 22 de l'ESG incluait toutes les personnes de 15 ans et plus au Canada en 2008, sauf :

1. les résidents du Yukon, des Territoires du Nord-Ouest et du Nunavut;
2. les résidents à plein temps des établissements institutionnels.

Afin de recueillir les données pour le cycle 22 de l'ESG, la méthode de l'interview téléphonique assistée par ordinateur (ITAO) a été utilisée. En ce qui a trait à l'échantillonnage, la population cible a été divisée en strates géographiques. Les ménages ont été sélectionnés au moyen de la méthode de l'échantillonnage par composition aléatoire, qui génère au hasard une liste de numéros de téléphone servant à prendre contact avec les ménages. Une fois qu'une prise de contact avec un ménage choisi a été effectuée, un répondant à l'enquête de 15 ans et plus est choisi pour participer à l'enquête. Les répondants ont été interviewés dans la langue officielle de leur choix et les interviews par personne interposée n'étaient pas permises. Les données du cycle 22 de l'ESG ont été recueillies en cinq vagues, de février à novembre 2008.

Liste de contrôle des problèmes pour l'exemple tiré de l'ESG

Un aspect de l'engagement social, qui est le centre d'intérêt du cycle 22 de l'ESG, est la participation communautaire. Un sujet couvert par l'enquête, qui fait partie de la participation communautaire, était l'engagement politique. Le chercheur souhaite savoir si les gens votent ou non, et si certains facteurs précis semblent liés à leur décision de voter ou non.

1. **Avez-vous exposé clairement en détail les questions auxquelles vous voulez répondre dans votre analyse?**
   Il y a deux questions d'intérêt :
   - Quelle a été la réponse des gens lorsqu'on leur a demandé s'ils avaient voté lors de la dernière élection fédérale? Quelle est la proportion de la population qui est estimée avoir déclaré chaque différent type de réponse?
   - Étant donné qu'une personne est âgée de 20 à 64 ans et que cette personne accepte de révéler si elle a voté ou non lors de la dernière élection fédérale, est-ce que la probabilité de voter est liée au sexe, groupe d'âge (20 à 29 ans, 30 à 44 ans, 45 à 64 ans) et du fait qu'elle vive en milieu rural ou urbain? Une fois que le sexe et le milieu rural ou urbain ont été pris en compte, est-ce que le groupe d'âge est toujours significativement lié à la probabilité de voter?

2. Est-ce qu'un échantillon analytique convenable avec des variables appropriées a été identifié?
Le cycle 22 de l'ESG semble contenir les variables nécessaires pour les analyses et paraît approprié pour cette population d'intérêt spéciale pour le chercheur. La population cible du cycle 22 de l'ESG est un peu plus vaste que la population d'intérêt en ce qui concerne l'intervalle d'âges (parce qu'elle couvre les personnes de 15 ans et plus et que nous nous intéressons seulement à celles de 20 à 64 ans), mais pourrait s'avérer un peu plus limitée que souhaité en ce qui a trait au type de personne faisant l'objet de la couverture (car elle ne couvre pas les résidents du Yukon, des Territoires du Nord-Ouest et du Nunavut ou les résidents à plein temps des établissements institutionnels). Le chercheur devra déterminer si ces contraintes ont de l'importance dans le cadre de son analyse.

Le fichier de microdonnées à grande diffusion (FMGD) ou le fichier de données confidentielles du cycle 22 de l'ESG pourraient être utilisé pour effectuer l'analyse. Nous illustrerons cet exemple en utilisant le FMGD, car ce fichier est plus accessible et permet d'effectuer la mise à l'essai de différents logiciels dans des endroits autres qu'un centre de données de recherche. Les variables analytiques (et leurs valeurs) contenues dans le FMGD qui ont été identifiées pour l'analyse sont les suivantes :

Tableau 1a Variables du fichier FMGD du cycle 22 de l'ESG, PER_Q110 variable

Beaucoup de personnes trouvent qu'il est difficile de sortir pour aller voter. Avez-vous voté lors des dernières élections fédérales?	
1	Oui
2	Non
7	Non demandé
8	Non déclaré
9	Ne sais pas

Tableau 1b Variables du fichier FMGD du cycle 22 de l'ESG, AGEGR5 variable

Groupe d'âge du répondant	
1	15 à 17
2	18 à 19
3	20 à 24
4	25 à 29
5	30 à 34
6	35 à 39
7	40 à 44
8	45 à 49
9	50 à 54
10	55 à 59
11	60 à 64
12	65 à 69
13	70 à 74
14	75 à 79
15	80 ans et plus

Tableau 1c Variables du fichier FMGD du cycle 22 de l'ESG, SEX variable

Sexe du répondant	
1	Homme
2	Femme

Tableau 1d Variables du fichier FMGD du cycle 22 de l'ESG, LUC_RST variable

Indicateur urbain/rural	
1	Grands centres urbains (RMR/AR)
2	Rural et petite municipalité (Autre qu'une RMR/AR)
3	Île-du-Prince-Édouard

On peut remarquer que la variable du fichier de données liée au fait qu'une personne ait voté ou non lors de la dernière élection fédérale, **PER_Q110**, a d'autres valeurs possibles que seulement « Oui » et « Non ». Le chercheur devra décider comment traiter cette variable dans son analyse. La variable **AGEGR5** fournit l'âge des répondants selon les catégories de groupe d'âge; les catégories 03 à 11 couvrent au complet l'intervalle d'âge d'intérêt et permettent la création des groupes d'âge d'intérêt. La variable de l'indicateur urbain/rural, **LUC_RST**, possède une catégorie pour « Urbain » et une autre pour « Rural », ainsi qu'une troisième pour la catégorie « Île-du-Prince-Édouard ». Encore une fois, le chercheur devra décider comment traiter cette variable dans son analyse. Pour le présent exemple, nous avons décidé d'inclure les observations de l'« Île-du-Prince-Édouard » dans la catégorie « Rural et petite municipalité ». Cela a fréquemment du sens de recoder les variables qui posent problème, en se fondant sur les décisions prises, au lieu de travailler avec les variables originales.

3. Avez-vous déterminé clairement quels sont les types d'analyses qui devront être effectuées?

Pour la question 1a) ci-dessus, une procédure qui permet d'estimer les proportions d'une population avec des caractéristiques précises sera nécessaire. Il sera également nécessaire d'obtenir des estimations de la variabilité de ces proportions estimées, comme les estimations des variances, les erreurs types ou les coefficients de variation.

Pour la question 1b) ci-dessus, une procédure qui permet d'exécuter une régression logistique pour la sous-population cible (c'est-à-dire les personnes de 20 à 64 ans qui ont révélé s'ils ont voté ou non) semblerait convenir. La variable dépendante pourrait être le logit de la probabilité de voter, tandis que les variables indépendantes pourraient être le sexe (avec « Femme » comme catégorie de référence), l'âge (deux variables avec « 45-64 » comme catégorie de référence) et urbain/rural (avec « Urbain » comme catégorie de référence et l'Île-du-Prince-Édouard serait considérée comme faisant partie de la catégorie « Rural »). Nous devrons également examiner la variabilité des coefficients du modèle estimés, de même qu'effectuer des tests précis sur les coefficients du modèle.

4. Avez-vous conclu que les estimations pondérées et les estimations de la variance bootstrap (erreurs) sont propices aux analyses?

Oui. Les estimations pondérées et les estimations de la variance bootstrap sont propices aux estimations des proportions ainsi qu'à la mise au point et la mise à l'essai d'un modèle de régression logistique.

5. **Avez-vous déterminé quels seront les tests et statistiques supplémentaires dont vous aurez besoin?**
Habituellement, pour un modèle logistique, un chercheur souhaite tester la signification de chaque coefficient dans le modèle; des tests de ce genre font normalement partie du résultat par défaut issu d'une régression logistique. Toutefois, le chercheur a précisé qu'il souhaite aussi vérifier si l'âge contribue de façon significative au modèle, après en avoir fait le contrôle pour le sexe et l'emplacement urbain/rural. En d'autres mots, un test est requis pour savoir si les coefficients des deux variables de la catégorie de référence « âge » dans le modèle à l'étude sont significativement différents de 0, étant donné que le sexe et la l'emplacement rural/urbain sont dans le modèle.

6. **Avez-vous consulté le guide de l'utilisateur de l'enquête (et d'autres documents) et déterminé les éléments suivants :**

   a. **Règles de non-divulgation?**
   Le guide de l'utilisateur du cycle 22 de l'ESG mentionne que : « L'utilisateur doit déterminer le nombre d'enregistrements dans le fichier de microdonnées qui ont fourni les données entrant dans le calcul d'une estimation. Ce nombre devrait être au moins 15 dans le cas des personnes. Si le nombre d'enregistrements contribuant à l'établissement de l'estimation pondérée est de moins de 15, celle-ci ne doit généralement pas être diffusée, quelle que soit la valeur de son coefficient de variation approximatif. Si l'estimation est malgré tout diffusée, elle doit l'être avec beaucoup de prudence et le nombre insuffisant d'enregistrements sur lesquels elle est fondée doit être indiqué clairement. »
   Les lignes directrices concernant les coefficients de variation doivent également être prises en compte. Voir le guide de l'utilisateur du cycle 22 de l'ESG.

   b. **Toutes les mises en garde sur les méthodes analytiques appropriées?**
   Dans le guide de l'utilisateur du cycle 22 de l'ESG, les utilisateurs sont encouragés à utiliser des estimations pondérées et les variances de l'estimation au moyen des poids bootstrap.

   c. **Variable de poids d'enquête appropriée et les variables de poids bootstrap correspondantes?**
   Dans le FMGD du cycle 22 de l'ESG, la variable de poids de la personne et les poids bootstrap correspondants sont respectivement nommés «wght_per » et « wtbs_001 » - « wtbs_500 ». (Il existe aussi une variable de poids du ménage, mais aucuns poids bootstrap correspondants ne sont fournis. Par conséquent, davantage de directives provenant du fournisseur de données seraient nécessaires avant de faire des analyses avec le poids du ménage.)

   d. **Si les variables du bootstrap sont des poids bootstrap et, le cas échéant, le nombre d'échantillons bootstrap qui ont été utilisées pour produire chaque poids bootstrap moyen (qui sont nécessaires pour effectuer un ajustement de poids bootstrap)?**
   Les poids bootstrap sont générés par la méthode du « bootstrap moyen ». Chaque poids du bootstrap moyen est généré à partir de 25 échantillons bootstrap réguliers. (Voir Phillips (2004) pour une brève description des poids de bootstrap moyen.)

SUDAAN 10

Aperçu

SUDAAN est un progiciel statistique qui sert à analyser des données corrélées, y compris des données corrélées provenant d'enquêtes par sondage complexes. SUDAAN 10 a été lancé en août 2008 et est composé de neuf procédures analytiques et de deux procédures pré-analytiques. Il permet de faire des estimations qui tiennent compte d'éléments d'un plan de sondage complexe d'une enquête, tels que la pondération inégale, la stratification, les plans à plusieurs degrés et en grappes, les mesures répétées, la corrélation de grappe générale et l'analyse des variables imputées à plusieurs reprises.

SUDAAN 10 offre trois méthodes d'estimation de la variance : séries de Taylor, jackknife et répliques répétées équilibrées. L'option des répliques répétées équilibrées peut être utilisée avec des poids bootstrap spécifiques à l'utilisateur, afin d'obtenir des estimations de la variance bootstrap (voir Phillips, 2004).

La version exécutable en SAS de SUDAAN 10 est celle qui est utilisée dans ce document. Dans cette version, les procédures de SUDAAN sont intégrées dans un programme SAS, ce qui signifie que les avantages reliées à la gestion de données qu'offre SAS peuvent être combinés aux forces d'analyse de données d'enquête de SUDAAN.

Le tableau suivant souligne les principaux types d'analyses qui peuvent être effectuées par SUDAAN 10, en utilisant l'estimation pondérée et l'estimation de la variance bootstrap. Les procédures particulières de SUDAAN pour obtenir ces analyses sont également fournies.

Tableau 2 Principaux types d'analyses dans SUDAAN 10

Type d'analyse	Procédure SUDAAN
Moyennes (y compris les moyennes géométriques)	proc descript
Totaux	proc descript
Quantiles/percentiles	proc descript
Ratios	proc ratio
Proportions/pourcentages	proc descript, proc crosstab
Tests d'indépendance dans les tableaux croisés	proc crosstab
Régression linéaire	proc regress
Régression logistique	proc rlogist
Logit multinominal	proc multilog
Cotes proportionnelles	proc multilog
Régression loglinéaire et de Poisson	proc loglink
Risques proportionnels (Cox)	proc survival
Courbes de survie de Kaplan-Meier	proc kapmeier
Analyse des variables imputées à plusieurs reprises	Toutes les procédures

Les détails à propos de chaque procédure peuvent être consultés dans le manuel du langage de SUDAAN et dans le manuel des exemples de SUDAAN, qui sont disponibles en ligne lorsque le logiciel est installé.

Liste de contrôle du logiciel pour l'exemple tiré de l'ESG

1. Avez-vous déterminé :

a. Si les estimations pondérées et les estimations de la variance bootstrap (erreurs) requises peuvent être calculées avec le logiciel que vous utilisez?

SUDAAN peut calculer les estimations pondérées et les estimations de la variance bootstrap nécessaires pour les types d'analyses propices à l'exemple, et pour plusieurs autres types d'analyses, comme on peut le constater dans le tableau ci-dessus. Toutefois, SUDANN ne permet pas de calculer des estimations de coefficients de variation qui peuvent être nécessaires pour certaines quantités, afin d'en vérifier la conformité avec les lignes directrices.

Comme décrit dans Phillips (2004), le fait de choisir l'option des répliques répétées équilibrées dans SUDDAN pour effectuer l'estimation de la variance et fournir les variables de poids bootstrap se soldera en estimations de la variance bootstrap. Phillips (2004) mentionne également comment le fait de choisir l'option des répliques répétées équilibrées avec un ajustement de Fay permettra de calculer correctement les estimations de la variance du bootstrap moyen.

Par conséquent, pour obtenir l'estimation pondérée et l'estimation de la variance bootstrap pour toute procédure de SUDAAN, il faut inclure les éléments suivants :

- Dans l'énoncé **PROC**, inclure l'option **DESIGN=BRR**.
- Inclure un énoncé **WEIGHT** pour identifier la variable de poids d'enquête à être utilisée pour calculer l'estimation pondérée.
- Inclure un énoncé **REPWEIGHT** pour indiquer les noms des variables de poids bootstrap dans le fichier de données. Si les variables de poids bootstrap sont des variables du bootstrap moyen, ajouter l'option **ADJFAY** à l'énoncé REPWEIGHT. La valeur ADJFAY consiste simplement en le nombre d'échantillons bootstrap utilisés pour produire chaque variable du bootstrap moyen. (L'option ADJFAY est omise si vous avez des poids bootstrap « réguliers ».)

Pour l'exemple tiré de l'ESG, où la variable de poids est wght_per, les 500 variables de poids bootstrap moyens sont wtbs_001 à wtbs_500 et chacune est formée de 25 échantillons bootstrap; toutes les procédures de SUDAAN utilisées contiendront les éléments suivants :

```
PROC procedurename data=SAS_datafile_name design=BRR;
WEIGHT wght_per;
REPWEIGHT wtbs_001-wtbs_500 / ADJFAY=25;
+Other statements required by the procedure
```

Nota : La documentation de SUDAAN mentionne qu'un énoncé WEIGHT est optionnel. Toutefois, si l'énoncé WEIGHT est omis, la moyenne des poids bootstrap est utilisée en tant que variable de poids. Cela change les valeurs des estimations pondérées et des estimations de la variance bootstrap de ce qu'on obtiendrait avec SUDAAN si un énoncé WEIGHT était inclus.

b. S'il est possible de produire les statistiques/effectuer les tests nécessaires avec le logiciel que vous utilisez?

Pour l'exemple tiré de l'ESG, nous voulons vérifier que chaque coefficient du modèle logistique est 0, et aussi que les coefficients des variables nominales des deux groupes d'âge sont simultanément 0. Les résultats de tests de la sorte font partie du résultat par défaut obtenu avec PROC RLOGIST, comme nous le démontrerons plus loin dans le document. Toutefois, il est également possible de recourir à d'autres statistiques pour vérifier la même chose ou pour tester des relations plus complexes parmi les coefficients du modèle en utilisant PROC RLOGIST.

2. **Si les tests ne peuvent être effectués et que les statistiques spécifiques souhaitées ne peuvent être calculées, est-il possible de développer un programme post-estimation pour calculer ces statistiques et effectuer les tests dans le logiciel que vous utilisez?**
   Pour cet exemple précis, tout test souhaité peut être exécuté avec SUDAAN. Toutefois, afin de fournir un aperçu des analyses pour lesquelles ce n'est pas le cas, les renseignements suivants sont fournis.

   Il est impossible de développer un programme post-estimation dans SUDAAN pour effectuer les tests qui ne sont pas inclus dans SUDAAN. Toutefois, puisque les procédures de SUDAAN sont intégrées dans un programme SAS, il est habituellement possible d'utiliser un large éventail des résultats de sortie de SUDAAN dans un programme SAS pour calculer ce qui est requis. Au moyen d'un énoncé OUTPUT, une gamme d'estimations peuvent être obtenues dans un format clé en mains. Par exemple, la matrice variance-covariance entière d'un ensemble d'estimations peut être obtenue; cette matrice (au lieu de seulement les estimations de variance) est nécessaire dans plusieurs tests statistiques. Une situation particulière où il serait utile de produire des quantités estimées et ensuite d'utiliser ces quantités dans un programme SAS serait le calcul des coefficients de variation, ce qui est impossible de faire avec SUDAAN. (Notons que le coefficient de variation d'une estimation est simplement le ratio de l'erreur type de l'estimation par rapport à l'estimation elle-même.)

3. **Est-il possible, dans le logiciel, de restreindre l'échantillon ou d'éliminer les observations qui ne concernent pas l'échantillon du fichier de données complet de l'enquête?**
   Une analyse avec SUDAAN peut être restreinte à l'échantillon avec des caractéristiques particulières (souvent nommé l'échantillon dans une sous-population particulière) en utilisant un énoncé SUBPOPN dans une procédure de SUDAAN. L'énoncé SUBPOPN doit être inclus dans toutes les procédures de SUDAAN pour lesquelles un échantillon restreint est requis.

   D'un autre côté, il est également simple de restreindre l'échantillon ou d'éliminer les observations qui ne concernent pas l'échantillon du fichier de données complet de l'enquête en écrivant le code pour une étape DATA d'un programme SAS avant de recourir à une procédure de SUDAAN. Cela produit un fichier de données plus petit qui peut ensuite être utilisé comme un fichier de données d'entrée pour toutes les procédures de SUDAAN pour lesquelles l'échantillon restreint est requis.

4. **Si le poids de l'enquête, les poids bootstrap et les variables d'analyse ne sont pas dans le même fichier, savez-vous comment fusionner les différentes sources dans le logiciel que vous utilisez? (En supposant que le logiciel utilisé requiert que tous ces renseignements soient dans le même fichier.)**
   La fusion des différentes sources devra être effectuée dans SAS avant d'utiliser les procédures de SUDAAN. Toutefois, dans SUDAAN, il n'est pas nécessaire que les poids bootstrap soient dans le même fichier de données que le poids de l'enquête et les variables d'analyse. Le manuel du langage de SUDAAN donne des directives sur la manière de spécifier un fichier différent pour les poids bootstrap.

5. **Tout en effectuant votre analyse, avez-vous vérifié les résultats de sortie de votre logiciel pour déterminer :**
   a. **que la bonne taille d'échantillon a été utilisée?**
      Le résultat par défaut obtenu d'une procédure de SUDAAN donne le nombre d'observations lues et le nombre d'observations utilisées dans l'analyse effectuée par la procédure. Il donne également les valeurs pondérées de ces deux quantités.
   b. **que la bonne variable de poids a été utilisée?**
      Le résultat par défaut obtenu d'une procédure de SUDAAN donne le nom de la variable de poids d'enquête utilisée.

c. **que l'ensemble complet des poids bootstrap a été utilisé?**
Le résultat par défaut obtenu par une procédure de SUDAAN produit une liste des noms de toutes les variables de poids bootstrap qui ont été utilisées.

d. **que l'ajustement pour le bootstrap moyen a été effectué de façon appropriée (si nécessaire)?**
Il est possible de déterminer si le bon ajustement pour le bootstrap moyen a été effectué en vérifiant que l'énoncé de sortie « Multiplicateur associé aux poids de ré-échantillonnage » contient le nombre d'échantillons bootstrap standards qui ont été combinés pour produire chaque bootstrap moyen. Dans le cas du cycle 22 de l'ESG, cette valeur devrait être 25.

e. **s'il existait des échantillons bootstrap pour lesquels il n'était pas possible de faire des estimations?**
S'il y a des échantillons bootstrap pour lesquels il n'a pas été possible de produire des estimations, ils sont identifiés dans les données de sortie de SUDAAN. Cela ne s'est pas produit dans l'exemple tiré de l'ESG.

Programme SAS/SUDAAN et résultats pour l'exemple tiré de l'ESG

Programme

```
/* PARTIE 1 */
options linesize=80;
libname pumfl '\\SASD6\Sasd-Dssea-
Public\DATA\GSS\DLI\CYCLE22\C22MDFSasAndCode-EngFr';

data c22pumf;
 set pumfl.c22pumf;
run;

/* PARTIE 2 */
/*analyse descriptive préliminaire*/
proc crosstab data= c22pumf design=brr;
 weight wght_per;
 repwgt wtbs_001- wtbs_500 / ADJFAY=25;
 class Per_Q110 /nofreq;
 tables Per_Q110;
 setenv colwidth=20 decwidth=4;
 print nsum="sampsize" wsum="popsize" rowper serow lowrow uprow;
run;

/* PARTIE 3 */
/*Recoder les variables et choisir l'échantillon de la  sous-population*/
data c22pumfn;
 set pumfl.c22pumf;

 /*Sous-population des électeurs de 20 à 64 ans*/
 if agegr5 ge 03 and agegr5 le 11;
 if Per_Q110 =1 or Per_Q110 =2;

 /*Recodage de la variable URBAIN*/
 if LUC_RST=1 then Urban=1;
 else Urban=0;
```

```
 /*Recodage de la variable ÂGE en trois catégories*/
 if agegr5 le 04 then age = 1;
 else if agegr5 le 07 then age =2;
 else if agegr5 le 11 then age =3;

 /*Recodage de la variable VOTE*/
 if Per_Q110 =1 then Vote =1;
 else Vote =0;
run;

/* PARTIE 4 */
proc crosstab data= c22pumfn design=brr;
 weight wght_per;
 repwgt wtbs_001- wtbs_500 / ADJFAY=25;
 class Vote /nofreq;
 tables Vote;
 setenv colwidth=20 decwidth=4;
 print nsum="sampsize" wsum="popsize" rowper serow lowrow uprow;
run;

/* PARTIE 5 */
proc rlogist data=c22pumfn design=BRR;
 weight wght_per;
 repwgt wtbs_001- wtbs_500 / ADJFAY=25;
 class sex age Urban;
 model Vote = sex age Urban ;
run;
```

Commentaires au sujet du programme et des résultats :

PARTIE 1

Cette partie du programme est de la programmation SAS où l'ensemble de données SAS du FMGD initial est spécifié. À la suite de l'exécution de l'étape DATA, le log SAS (pas présenté) indique que l'ensemble de données c22pumf renferme 20 401 enregistrements. Cette partie du programme ne produit aucun résultat.

PARTIE 2

Cette partie du programme fait appel à SUDAAN PROC CROSSTAB pour obtenir des estimations des proportions de la population totale qui a donné les différents types de réponses à la question concernant l'action de voter à la dernière élection fédérale. Cela permet de faire une inspection préliminaire de la variable Per_Q110.

Comme c'est le cas à chaque fois où l'on utilise une procédure de SUDAAN, le résultat fournit les noms de la variable de poids et des variables de poids bootstrap, la méthode d'estimation de la variance utilisée, la taille de l'échantillon et la taille de la population estimée.

Afin de limiter le résultat de CROSSTAB aux quantités d'intérêt du chercheur, un énoncé PRINT a été utilisé. En plus des proportions estimées et leurs erreurs types, la taille de l'échantillon et la taille de la population estimée et des limites de confiance de 95 % étaient requises. Il faut noter que SUDAAN ne permet pas de calculer d'estimations des coefficients de variation.

Comme on peut le constater avec le résultat, 94 % (c'est-à-dire, 68,98+25,09=94,07) de la population ciblée par le cycle 22 de l'ESG est estimée avoir répondu soit oui (Per_Q110=1) ou non (Per_Q110=2) à la question de l'action de voter lors de la dernière élection fédérale. Le reste de la population n'a pas répondu par oui ou non pour une variété de raisons (valeurs de 7,8 ou 9 pour Per_Q110).

PROC CROSSTAB calcule, par défaut, les intervalles de confiance asymétriques pour les proportions, en utilisant une transformation logistique. Si, au lieu, on utilise PROC DESCRIPT pour calculer la proportion d'un cas spécial d'une moyenne, des intervalles de confiance symétriques sont produits, fondés sur l'hypothèse d'une distribution normale approximative pour le ratio de la proportion sur son erreur type.

Figure 1 Résultats de SUDAAN pour PER_Q110

```
                          S U D A A N
            Software for the Statistical Analysis of Correlated Data
           Copyright      Research Triangle Institute     August 2008
                                Release 10.0

DESIGN SUMMARY: Variances will be computed using the
Balanced Repeated Replication (BRR) Method
  Sample Weight: WGHT_PER
  Replicate Sample Weights:
        WTBS_001 WTBS_002 WTBS_003 WTBS_004 WTBS_005 WTBS_006 WTBS_007
(names of other weights are omitted in this handbook output)....
        WTBS_498 WTBS_499 WTBS_500
    Multiplier Associated with Replicate Weights: 25

Number of observations read    :  20401    Weighted count : 27261810
Denominator degrees of freedom :    500
```

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

		PER_Q110
		Total
sampsize		20401.0000
popsize		27261809.6964
Row Percent		100.0000
SE Row Percent		0.0000
Lower 95% Limit		
ROWPER		.
Upper 95% Limit		
ROWPER		.

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

		PER_Q110
		1
sampsize		14941.0000
popsize		18805257.7595
Row Percent		68.9802
SE Row Percent		0.3674
Lower 95% Limit		
ROWPER		68.2538
Upper 95% Limit		
ROWPER		69.6974

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

		PER_Q110
		2
	sampsize	4660.0000
	popsize	6841080.2797
	Row Percent	25.0940
	SE Row Percent	0.3804
	Lower 95% Limit	
	ROWPER	24.3541
	Upper 95% Limit	
	ROWPER	25.8487

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

		PER_Q110
		7
	sampsize	666.0000
	popsize	1437778.6657
	Row Percent	5.2740
	SE Row Percent	0.1370
	Lower 95% Limit	
	ROWPER	5.0111
	Upper 95% Limit	
	ROWPER	5.5498

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

		PER_Q110
		8
	sampsize	45.0000
	popsize	54861.2709
	Row Percent	0.2012
	SE Row Percent	0.0375
	Lower 95% Limit	
	ROWPER	0.1395
	Upper 95% Limit	
	ROWPER	0.2902

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

		PER_Q110
		9
	sampsize	89.0000
	popsize	122831.7206
	Row Percent	0.4506
	SE Row Percent	0.0647
	Lower 95% Limit	
	ROWPER	0.3397
	Upper 95% Limit	
	ROWPER	0.5974

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

PARTIE 3

Cette partie du programme est une étape DATA dans SAS. Dans cette partie, des observations de l'échantillon dans la sous-population d'intérêt sont choisies (c'est-à-dire, des observations de l'échantillon pour les personnes de 20 à 64 ans et répondre « oui » ou « non » à la question sur l'action de voter à la dernière élection fédérale). Puis, certaines des variables sont recodées afin de jumeler les catégories requises pour effectuer la régression logistique. Voici les détails des recodages effectués :

Tableau 3 Recodage des variables

PER_Q110	Vote	AGEGR5	Âge	LUC_RST ->	Urbain
1 Yes	1	03,04	1	1	1
2 No	0	05,06,07	2	2,3	0
		08,09,10,11	3		

Après avoir exécuté l'étape DATA, le log SAS (pas présenté) indique que le nouvel ensemble de données c22pumf contient 14 813 enregistrements. Cette étape DATA ne produit aucun résultat.

PARTIE 4

Cette partie du programme fait appel à PROC CROSSTAB dans SUDAAN, afin d'obtenir des estimations du pourcentage de personnes de 20 à 64 ans qui ont répondu « oui » (VOTE=1) ou « non » (VOTE=0) à la question concernant l'action de voter lors de la dernière élection fédérale, en supposant qu'ils ont répondu par une de ces deux réponses à la question. Il faut noter que la nouvelle variable VOTE est utilisée, avec l'échantillon restreint. Il aurait été possible d'avoir utilisé, au lieu, l'ensemble complet de données, mais de le restreindre à la sous-population d'intérêt en incluant un énoncé SUBPOPN dans PROC CROSSTAB.

Figure 2 SUDAAN sommaire du plan et de la variable VOTE

```
DESIGN SUMMARY: Variances will be computed using the Balanced Repeated
Replication (BRR) Method
  Sample Weight: WGHT_PER
  Replicate Sample Weights:
        WTBS_001 WTBS_002 WTBS_003 WTBS_004 WTBS_005 WTBS_006 WTBS_007
...
        WTBS_498 WTBS_499 WTBS_500
    Multiplier Associated with Replicate Weights: 25

Number of observations read    :  14813    Weighted count : 20625842
Denominator degrees of freedom :    500
```

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

		VOTE
		Total
	sampsize	14813.0000
	popsize	20625841.5165
	Row Percent	100.0000
	SE Row Percent	0.0000
	Lower 95% Limit ROWPER	.
	Upper 95% Limit ROWPER	.

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

		VOTE
		0
	sampsize	3852.0000
	popsize	5742406.3343
	Row Percent	27.8408
	SE Row Percent	0.4592
	Lower 95% Limit ROWPER	26.9476
	Upper 95% Limit ROWPER	28.7520

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

		VOTE
		1
	sampsize	10961.0000
	popsize	14883435.1822
	Row Percent	72.1592
	SE Row Percent	0.4592
	Lower 95% Limit ROWPER	71.2480
	Upper 95% Limit ROWPER	73.0524

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

PARTIE 5

Cette partie du programme est l'ajustement du modèle logistique à l'échantillon restreint, au moyen de PROC RLOGIST dans SUDAAN. La procédure modélise le logit de la probabilité que VOTE=1 en utilisant l'échantillon restreint. Notons que toutes les informations sur le poids, les poids bootstrap, etc. doivent être incluses lorsqu'on utilise la procédure.

Il faut noter que le résultat par défaut fournit des renseignements sur les variables SEXE, ÂGE et URBAIN qui sont identifiées comme étant catégorielles dans un énoncé CLASS. Par défaut, la valeur la plus élevée de chaque variable sera la catégorie de référence lorsque le programme crée des variables nominales. Il est aussi

possible de déclarer quelles sont les valeurs que vous voulez choisir comme catégories de référence en utilisant une option REFLEVEL, comme décrit dans le guide de l'utilisateur.

Figure 3 SUDAAN sommaire du plan et des variables SEX, AGE, et URBAN

```
DESIGN SUMMARY: Variances will be computed using the Balanced Repeated
Replication (BRR) Method
  Sample Weight: WGHT_PER
  Replicate Sample Weights:
        WTBS_001 WTBS_002 WTBS_003 WTBS_004 WTBS_005 WTBS_006 WTBS_007
...
        WTBS_498 WTBS_499 WTBS_500
    Multiplier Associated with Replicate Weights: 25


Number of zero responses     :  3852
Number of non-zero responses : 10961

Independence parameters have converged in 6 iterations

Number of observations read       :  14813    Weighted count: 20625842
Observations used in the analysis :  14813    Weighted count: 20625842
Denominator degrees of freedom    :    500


Maximum number of estimable parameters for the model is  5

Sample and Population Counts for Response Variable VOTE
Based on observations used in the analysis
0:  Sample Count     3852    Population Count   5742406
1:  Sample Count    10961    Population Count  14883435
```

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

```
R-Square for dependent variable VOTE (Cox & Snell, 1989): 0.049179

-2 * Normalized Log-Likelihood with Intercepts Only : 17522.11

-2 * Normalized Log-Likelihood Full Model          : 16775.10
Approximate Chi-Square (-2 * Log-L Ratio)          :   747.01
Degrees of Freedom                                 :        4

Note: The approximate Chi-Square is not adjusted for clustering.
      Refer to hypothesis test table for adjusted test.
```

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

```
Frequencies and Values for CLASS Variables
by: SEX.
---------------------------------
SEX             Frequency    Value
---------------------------------
Ordered
  Position:
  1                  6576        1
Ordered
  Position:
  2                  8237        2
---------------------------------
```

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

```
Frequencies and Values for CLASS Variables
by: AGE.
---------------------------------
AGE             Frequency    Value
---------------------------------
Ordered
  Position:
  1                  2109        1
Ordered
  Position:
  2                  5031        2
Ordered
  Position:
  3                  7673        3
---------------------------------
```

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

```
Frequencies and Values for CLASS Variables
by: URBAN.
---------------------------------
URBAN           Frequency    Value
---------------------------------
Ordered
  Position:
  1                  3534        0
Ordered
  Position:
  2                 11279        1
```

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Le modèle ajusté est présenté de deux façons dans deux tableaux différents : le premier présente les coefficients ajustés; le deuxième présente les coefficients sous forme de rapports de cotes. Un troisième tableau présente les résultats des test t sur chaque coefficient individuel.

Figure 4 SUDAAN modèle logit sur la variable dépendante VOTE

```
Link Function: Logit
Response variable VOTE: VOTE
by: Independent Variables and Effects.
------------------------------------------------------------------------
Independent
  Variables and        Beta                      Lower 95%    Upper 95%
  Effects              Coeff.          SE Beta   Limit Beta   Limit Beta
------------------------------------------------------------------------
Intercept                   1.55         0.04         1.46         1.64
SEX
  1                        -0.00         0.05        -0.10         0.09
  2                         0.00         0.00         0.00         0.00
AGE
  1                        -1.25         0.06        -1.38        -1.13
  2                        -0.79         0.05        -0.89        -0.69
  3                         0.00         0.00         0.00         0.00
URBAN
  0                         0.03         0.06        -0.08         0.14
  1                         0.00         0.00         0.00         0.00
------------------------------------------------------------------------
```

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Independent Variables and Effects	Odds Ratio	Lower 95% Limit OR	Upper 95% Limit OR
Intercept	4.71	4.32	5.13
SEX			
1	1.00	0.91	1.10
2	1.00	1.00	1.00
AGE			
1	0.29	0.25	0.32
2	0.45	0.41	0.50
3	1.00	1.00	1.00
URBAN			
0	1.03	0.92	1.15
1	1.00	1.00	1.00

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Independent Variables and Effects	T-Test B=0	P-value T-Test B=0
Intercept	35.02	0.0000
SEX		
1	-0.01	0.9955
2	.	.
AGE		
1	-19.31	0.0000
2	-15.35	0.0000
3	.	.
URBAN		
0	0.48	0.6298
1	.	.

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Le tableau ci-dessous, qui fait partie du résultat par défaut, présente les résultats des tests servant à déterminer quelle variable contribue de façon significative au modèle, à la suite de l'inclusion des autres variables. Ainsi, ce tableau nous permet de vérifier si la variable AGE (avec deux degrés de liberté) contribue de façon significative au modèle, sachant que la variable SEX et RURAL/URBAN sont déjà présente. C'était une des questions d'intérêt du chercheur. Une approche de rechange pour évaluer la même hypothèse serait d'utiliser un énoncé CONTRAST ou EFFECT lorsqu'on utilise PROC RLOGIST.

Figure 5 SUDAAN régression

Contrast	Degrees of Freedom	Wald F	P-value Wald F
OVERALL MODEL	5	471.41	0.0000
MODEL MINUS INTERCEPT	4	114.69	0.0000
INTERCEPT	.	.	.
SEX	1	0.00	0.9955
AGE	2	228.33	0.0000
URBAN	1	0.23	0.6298

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

STATA 12

Aperçu

Stata 12 a été lancé en 2010. Cette version inclut plusieurs caractéristiques conçues spécialement pour travailler avec les poids bootstrap à Statistique Canada. Stata utilise l'ajout d'un préfixe (*svy*) avec plusieurs fonctions, ce qui les adapte pour fonctionner avec le plan de sondage, lequel est spécifié dans un énoncé de programme (*svyset*). On peut utiliser Stata soit en utilisant la syntaxe (énoncés de programme) sous la forme d'un fichier « do » au moyen de menus déroulants. Tous les exemples illustrés dans ce document utilisent des énoncés de syntaxe.

Stata 12 offre une ensemble complet d'options d'estimation de la variance fondée sur le plan avec des fonctions *svy* : Séries de Taylor, jackknife, répliques répétées équilibrées et bootstrap. L'option bootstrap peut être utilisée avec des poids bootstrap d'enquête spécifiques à l'utilisateur, tels que ceux fournis par plusieurs enquêtes de Statistique Canada, afin d'obtenir des estimations de la variance bootstrap. La méthode pour utiliser les versions antérieures de Stata afin d'obtenir des estimations de la variance bootstrap est décrite à l'annexe 2.

Le tableau suivant illustre certains types d'analyses communes qui peuvent être effectués au moyen des fonctions *svy* dans Stata 12, où des estimations pondérées et des estimations de la variance bootstrap sont produites. Il existe plusieurs autres techniques d'analyse en utilisant les fonctions *svy*. On trouve une liste complète dans l'annexe A1 et aussi dans le manuel des données de l'enquête, la procédure d'aide et la documentation en format PDF dans Stata, ou en utilisant la FAQ ou les archives statalist sur le site Web de Stata, ou encore, en communiquant avec le service de soutien technique de Stata.

Tableau 4 Certaines analyses communes au moyen des fonctions svy dans STATA 12

Type d'analyse	Procédure
Moyennes	svy:mean
Totaux	svy:total
Ratios	svy:ratio
Proportions/pourcentages	svy:proportion; svy:tabulate
Tests d'indépendance dans les tableaux croisés	svy:tabulate
Régression linéaire	svy:regress
Régression logistique	svy:logistic ou svy:logit
Logit multinominal	svy:mlogit
Cotes proportionnelles	svy:ologit
Régression loglinéaire et de Poisson	proc svy:poisson
Modèles de risques proportionnels (Cox)	svy:stcox
Courbe de survie de Kaplan-Meyer	svy: stcox avec une covariable simple, suivie par stcurve

En plus des fonctions *svy*, plusieurs fonctions de Stata 12 acceptent des poids d'enquête dans un énoncé pweight. Étant donné que ces fonctions n'utilisent pas les poids bootstrap, l'estimation de la variance bootstrap fondée sur un modèle n'est pas effectuée. On trouve davantage de renseignements sur ce sujet dans l'annexe sur les poids normalisés (Annexe 4).

Liste de contrôle du logiciel pour l'exemple tiré de l'ESG

1. Avez-vous déterminé :

a. Si les estimations pondérées et les estimations de la variance bootstrap (erreurs) requises peuvent être calculées avec le logiciel que vous utilisez?

Stata peut calculer les estimations pondérées et les estimations de la variance bootstrap requises pour les types d'analyses pour cet exemple, et pour plusieurs autres types d'analyses, comme on peut le constater dans le tableau ci-dessus et dans les annexes. Fondamentalement, pour produire des résultats analytiques qui incluent le plan de sondage au moyen de poids d'enquête et des poids bootstrap, il faut procéder de la façon suivante :

1. Utiliser une fonction *svyset* avec les caractéristiques suivantes :
   a. Assigner *pweight* à la bonne variable de poids d'enquête.
   b. Poids bootstrap identifiés avec l'option *bsrweight*.
   c. Type d'estimation de la variance assigné à *bootstrap* avec l'option *vce*.
   d. Degrés de liberté assignés à leur valeur par défaut (nombre de poids bootstrap) ou, dans le cas de quelques enquêtes, à une plus petite valeur appropriée. (Une bonne approximation des degrés de liberté se calcule par le nombre d'unités primaires d'échantillonnage contenues dans l'échantillon de la population qui a fait l'objet d'une analyse moins le nombre de strates contenues dans l'échantillon de cette population. Toutefois, ces quantités sont fréquemment inconnues pour le chercheur. Dans la plupart des cas, le fait d'utiliser le nombre de poids bootstrap au lieu de cette approximation aura peu d'effet sur les résultats.)
   e. L'option de l'erreur quadratique moyenne est assignée grâce à l'option *mse*. (Cette option permet de calculer la variance bootstrap en utilisant les écarts quadratiques entre les estimations bootstrap et l'estimation de l'échantillon au complet. Si cette option n'est pas choisie, la variance bootstrap est calculée en utilisant les écarts quadratiques entre les estimations bootstrap et la moyenne des estimations bootstrap.)
   f. L'ajustement du bootstrap moyen est assigné, si nécessaire, à l'option *bsn*, qui devrait avoir la valeur du nombre d'échantillons bootstrap utilisés pour produire chaque poids bootstrap moyen.
2. Pour l'exemple tiré de l'ESG, la variable de poids d'enquête est *wght_per*, les 500 variables de poids bootstrap moyen sont *wtbs_001* à *wtbs_500* et chacune est formée à partir de 25 échantillons bootstrap. Votre fonction *svyset* apparaîtra comme ceci :

```
svyset [pweight=wght_per], bsrweight(wtbs_001- wtbs_500) bsn(25)
vce(bootstrap) dof(500) mse
```

3. Après avoir utilisé la fonction *svyset*, Stata produira des estimations pondérées et des estimations bootstrap à chaque fois que vous utilisez le préfixe *svy* au début de l'écriture de la fonction qui accepte ce préfixe (voir annexe A1) – p. ex. svy:mean.

b. S'il est possible de produire les statistiques/effectuer les tests nécessaires avec le logiciel que vous utilisez?

Pour l'exemple tiré de l'ESG, évaluer si chaque coefficient du modèle logistique sous-jacent qui est significativement différent de 0 est obtenu dans le résultat par défaut du modèle. Déterminer si l'âge reste significatif à la suite du contrôle pour le sexe et l'emplacement urbain/rural se fait au moyen d'un test conjoint des coefficients des deux groupes d'âge en utilisant la syntaxe test. Ça peut être obtenu par une fonction post-estimation.

D'autres tests et procédures post-estimation sont disponibles dans Stata. Afin de s'assurer que ces tests sont appropriés pour les résultats obtenus au moyen d'une fonction *svy* particulière, veuillez consulter le guide de l'utilisateur ou le service de soutien technique de Stata.

2. **Si les tests ne peuvent être effectués et que les statistiques spécifiques souhaitées ne peuvent être calculées, est-il possible de développer un programme post-estimation pour calculer ces statistiques et effectuer les tests dans le logiciel que vous utilisez?**
   Il est possible d'utiliser des programmes développés par l'utilisateur dans Stata. De plus, Stata possède la capacité de télécharger des programmes développés par d'autres personnes pour accomplir certaines tâches. Pour obtenir des précisions sur ces possibilités, voir le guide de programmation et le guide MATA disponibles dans Stata ou en ligne. Stata permet aussi de calculer presque toutes les estimations, matrices, etc. qui peuvent être utilisées pour effectuer des tests post-estimation dans des programmes développés par l'utilisateur dans Stata. Toutefois, c'est la responsabilité du chercheur de déterminer quels programmes à télécharger ou développés par l'utilisateur conviennent à son étude et qu'ils utilisent les poids appropriés et estimations bootstrap dans leurs calculs.

3. **Est-il possible, dans le logiciel, de restreindre l'échantillon ou d'éliminer les observations qui ne concernent pas l'échantillon du fichier de données complet de l'enquête?**
   Sample Les restrictions de l'échantillon peuvent être effectuées à l'aide d'énoncés *if* dans les fonctions, en maintenant ou en éliminant des enregistrements au moyen d'une fonction *keep* ou *drop*, ou en utilisant des options de sous-population dans les fonctions *svy*. Dans l'exemple tiré de l'ESG, deux fonctions *keep* ont été utilisées pour restreindre l'échantillon à des observations appropriées.

4. **Si le poids de l'enquête, les poids bootstrap et les variables d'analyse ne sont pas dans le même fichier, savez-vous comment fusionner les différentes sources dans le logiciel que vous utilisez? (En supposant que le logiciel utilisé requiert que tous ces renseignements soient dans le même fichier.)**
   Dans Stata, le poids d'enquête, les poids bootstrap ainsi que les variables d'analyse doivent apparaître dans le même fichier. Ceci peut être effectué en utilisant une fusion une à une dans Stata.

5. **Tout en effectuant votre analyse, avez-vous vérifié les résultats de sortie de votre logiciel pour déterminer :**
   a. **Que la bonne taille d'échantillon a été utilisée?**
      Le résultat par défaut obtenu à l'aide d'une procédure Stata donne le nombre d'observations lues et le nombre d'observations utilisées dans l'analyse effectuée. Il est aussi possible d'obtenir des valeurs pondérées de ces deux quantités.
   b. **Que la bonne variable de poids a été utilisée?**
      Le résultat par défaut obtenu à l'aide d'une fonction *svyset* donne le nom de la variable de poids d'enquête utilisée.
   c. **Que l'ensemble complet des poids bootstrap a été utilisé?**
      Le résultat par défaut obtenu à l'aide de la fonction *svyset* produit une liste des noms de toutes les variables de poids bootstrap utilisées. Vérifiez attentivement cet énoncé pour vous assurer que tous les poids bootstrap ont été utilisés. Dans la syntaxe montrée ici, nous présumons que tous les poids bootstrap sont ordonnés de façon contigüe dans le fichier. Une autre façon de spécifier les poids est d'utiliser un caractère de remplacement (c'est-à-dire wtbs_*).

d. Que l'ajustement pour le bootstrap moyen a été effectué de façon appropriée (si nécessaire)?
Le résultat obtenu à l'aide de la fonction *svyset* montre si l'ajustement du bootstrap moyen a été effectué.

e. S'il existait des échantillons bootstrap pour lesquels il n'était pas possible de faire des estimations?
S'il y a des échantillons bootstrap pour lesquels des estimations n'auraient pu être effectuées, ils sont identifiés dans les données de sortie de Stata. Cela ne s'est pas produit dans l'exemple tiré de l'ESG.

Programme Stata et résultats de l'exemple tiré le l'ESG

Syntaxe du fichier « do »

```
*Ouvrir le fichier log
log using «F:\GSS22\pumf\example.log», replace

*Ouvrir le fichier données
use "F:\GSS22\pumf\c22pumf_eng.dta", clear

*Restreindre l'échantillon par l'âge
keep if agegr5 >2
keep if agegr5 <12

*Recoder per_q110 en une nouvelle variable VOTE où 1=oui et 0=non
 7,8,9=données manquantes (.)
recode per_q110 (1=1) (2=0) (7/9=.), gen (vote)
label define vote 1 «yes» 0 «no», modify
label values vote vote

*Recoder la variable agegr5 en ÂGE avec 3 catégories
recode agegr5 (1/2=.) (3/4=1) (5/7=2) (8/11=3), gen (age)
label define age 1»20 to 29 years»  2»30 to 44 years» 3 «45 to 64 years», modify
label values age age

*Recoder luc_rst en URBAIN
recode luc_rst (1=1) (2/3=0), gen (urban)
label define urban 0 «rural « 1 «urban», modify
label values urban urban

*Configurer l'information du plan de sondage pour l'utiliser avec le préfixe
svy
svyset [pweight=wght_per], bsrweight(wtbs_001- wtbs_500) bsn(25) vce(bootstrap)
dof(500) mse

*Tableau de fréquence avec des écarts-types et des coefficients de variation bootstrap
pour chaque variable pour vérifier les descriptifs et les recodages
svy:tab per_q110, obs count se  cv format(%14.4g)
svy:tab vote, obs count se  cv format(%14.4g)
svy:tab vote, obs se cv format(%14.4g)
svy:tab age, obs count se cv format(%14.4g)
svy:tab age, obs se cv format(%14.4g)
svy:tab urban, obs count se cv format(%14.4g)
```

```
svy:tab urban, obs se cv format(%14.4g)
svy:tab sex, obs count se cv format(%14.4g)
svy:tab sex, obs se cv format(%14.4g)

*Tableaux croisés de l'ÂGE et URBAIN avec VOTE et les tests
svy:tab vote age, col obs se ci cv format(%14.4g)
svy:tab vote urban, col obs se ci cv format(%14.4g)

*Régression logistique
svy:logit vote ib2.sex ib3.age ib1.urban
svy:logistic vote ib3.age ib1.urban ib2.sex
test 1.age 2.age
```

Résultats de Stata

Ces premières fonctions permettent d'ouvrir le fichier, de restreindre l'échantillon à l'intervalle d'âge souhaité ainsi que de recoder certaines variables.

```
. *Ouvrir le fichier de données
. use «F:\GSS22\pumf\c22pumf_eng.dta», clear

. *Restreindre l'échantillon par l'ÂGE
. keep if agegr5 >2
(1057 observations deleted)

. keep if agegr5 <12
(4421 observations deleted)

. *Recoder per_q110 en une nouvelle variable VOTE où 1=oui and 0=non 7,8,9=données
manquantes (.)
. recode per_q110 (1=1) (2=0) (7/9=.), gen (vote)
(3962 differences between per_q110 and vote)
. label define vote 1 «yes» 0 «no», modify
. label values vote vote
. *recoder la variable agegr5 en ÂGE avec 3 catégories
. recode agegr5 (1/2=.) (3/4=1) (5/7=2) (8/11=3), gen (age)
(14923 differences between agegr5 and age)
. label define age 1»20 to 29 years»  2»30 to 44 years» 3 «45 to 64 years», modify
. label values age age
. *recodeR luc_rst en URBAIN
. recode luc_rst (1=1) (2/3=0), gen (urban)
(3559 differences between luc_rst and urban)
. label define urban 0 «rural « 1 «urban», modify
. label values urban urban
```

La fonction *svyset* ci-dessus informe Stata qu'il y a un poids de probabilité (wght_per) à utiliser dans le calcul des estimations et un ensemble de poids bootstrap à utiliser pour l'estimation de la variance fondée sur le plan. Elle informe également Stata qu'il est requis de corriger pour le bootstrap moyen (bsn=25). De plus, le préfixe *mse* dit à Stata d'utiliser la différence entre les estimations bootstrap et les estimations de l'échantillon au complet pour effectuer les calculs d'estimation de variance. Les résultats démontrent que tout ceci a été effectué.

```
. *Configurer l'information du plan de sondage pour l'utiliser avec le préfixe svy
. svyset [pweight=wght_per], bsrweight(wtbs_001- wtbs_500) bsn(25) vce(bootstrap)
dof(500) mse

      pweight: wght_per
          VCE: bootstrap
          MSE: on
    bsrweight: wtbs_001 wtbs_002 wtbs_003 wtbs_004 wtbs_005 wtbs_006 wtbs_007 wtbs_008
wtbs_009
               wtbs_010 wtbs_011 wtbs_012 wtbs_013 wtbs_014 wtbs_015 wtbs_016 wtbs_017
wtbs_018
…(output omitted)
               wtbs_496 wtbs_497 wtbs_498 wtbs_499 wtbs_500
          bsn: 25
    Design df: 500
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
```

Cette prochaine section nous permet de voir notre nombre d'observations, les valeurs pondérées et les pourcentages pondérés avec les coefficients de variation pour chaque variable de notre analyse. Seul le résultat pour les variables Per_q110 et VOTE est présenté. Veuillez noter que le nombre d'observations baisse de 14 923 (pour Per_q110) à 14 813 (pour VOTE), car les observations pour lesquelles VOTE a une valeur manquante sont exclues de l'estimation.

Figure 6 STATA analyse de fréquences bootstrap pour PER_Q110 et VOTE

```
. *frequency table with bootstrapped for each variable to check basic descriptives and recodes

. svy:tab per_q110, obs count se  cv format(%14.4g)
(running tabulate on estimation sample)

Bstrap *: for cell counts

                                                Number of obs      =     14923
                                                Population size    =  20780386
                                                Replications       =       500
                                                Design df          =       500

------------------------------------------------------------
lots of   |
people    |
find it   |
difficult |
to get    |
out and   |
vote. did |
you       |      count          se          cv         obs
----------+-------------------------------------------------
      yes |   14883435       95794       .6436       10961
       no |    5742406       94867       1.652        3852
 not aske |       8137        4254       52.29           4
 not stat |      33752        7655       22.68          29
 don't kn |     112656       17305       15.36          77
          |
    Total |   20780386                                14923
------------------------------------------------------------
  Key:  count     =  weighted counts
        se        =  bootstrap standard errors of weighted counts
        cv        =  coefficients of variation of weighted counts
        obs       =  number of observations
```

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

```
. svy:tab vote, obs count se  cv format(%14.4g)
(running tabulate on estimation sample)

Bstrap *: for cell counts

                                                Number of obs      =     14813
                                                Population size    =  20625842
                                                Replications       =       500
                                                Design df          =       500

------------------------------------------------------------
Vote      |      count          se          cv         obs
----------+-------------------------------------------------
       no |    5742406       94867       1.652        3852
      yes |   14883435       95794       .6436       10961
          |
    Total |   20625842                                14813
------------------------------------------------------------
  Key:  count     =  weighted counts
        se        =  bootstrap standard errors of weighted counts
        cv        =  coefficients of variation of weighted counts
        obs       =  number of observations
```

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

```
. svy:tab vote, obs se cv format(%14.4g)
(running tabulate on estimation sample)

                                            Number of obs      =     14813
                                            Population size    =  20625842
                                            Replications       =       500
                                            Design df          =       500

------------------------------------------------------------
Vote      | proportions           se           cv          obs
----------+-------------------------------------------------
       no |       .2784      .004592        1.649         3852
      yes |       .7216      .004592        .6364        10961
          |
    Total |           1                                 14813
------------------------------------------------------------
  Key:  proportions  =  cell proportions
        se           =  bootstrap standard errors of cell proportions
        cv           =  coefficients of variation of cell proportions
        obs          =  number of observations
```

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Ensuite, nous nous intéressons aux tableaux croisés entre notre variable d'intérêt (VOTE) et chaque covariable. Seul le tableau croisé entre VOTE et AGE est présenté. Pour chaque cellule au sein de chaque tableau nous pouvons voir une estimation de la proportion et de l'écart-type pour la proportion, le coefficient de variation, l'intervalle de confiance, et les comptes non pondérés. Il y a aussi un test d'indépendance au bas de chaque tableau. Veuillez noter que les intervalles de confiance sont asymétriques car ils sont calculés en utilisant une transformation logit.

Figure 7 STATA tableau croisé pour VOTE et AGE

```
*crosstabs of age and urban with vote with tests
. svy:tab vote age, col obs se ci cv format(%14.4g)
(running tabulate on estimation sample)

Bstrap *: for columns

                                             Number of obs      =     14813
                                             Population size    =  20625842
                                             Replications       =       500
                                             Design df          =       500

------------------------------------------------------------------------------
          |      RECODE of agegr5 (age group of the respondent.)
  Vote    |      20 to 29       30 to 44       45 to 64          Total
----------+-------------------------------------------------------------------
       no |          .426          .3185          .1744          .2784
          |      (.01354)      (.007999)      (.005157)      (.004592)
          |         3.178          2.511          2.957          1.649
          | [.3996,.4528]   [.303,.3344]  [.1645,.1848]  [.2695,.2875]
          |           913           1569           1370           3852
          |
      yes |          .574          .6815          .8256          .7216
          |      (.01354)      (.007999)      (.005157)      (.004592)
          |         2.359          1.174          .6247          .6364
          | [.5472,.6004]   [.6656,.697]  [.8152,.8355]  [.7125,.7305]
          |           1196           3462           6303          10961
          |
    Total |             1              1              1              1
          |
          |
          |
          |          2109           5031           7673          14813
------------------------------------------------------------------------------
  Key:  column proportions
        (bootstrap standard errors of column proportions)
        coefficients of variation of column proportions
        [95% confidence intervals for column proportions]
        number of observations

  Pearson:
    Uncorrected   chi2(2)         =  741.0659
    Design-based  F(1.90, 947.99) =  213.2548     P = 0.0000
```

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Deux résultats du modèle sont présentés ci-dessous. Leur seule différence a trait à ce que le premier (svy: logit) montre les coefficients, et le deuxième (svy: logistic) montre les rapports de cotes. Il faut noter l'utilisation du préfixe d'enquête *svy* au début de l'écriture de la fonction logit ou logistique. Les tests t pour calculer les coefficients ou les rapports de cotes se trouvent juste à côté des estimations. La fonction test qui suit l'ajustement des tests du modèle logistique permet de voir si les deux variables nominales pour l'âge ont une incidence significative sur le modèle, une fois que la variable sexe et emplacement urbain/rural sont contrôlées.

Figure 8 STATA modèle logit et tests t

```
*logistic regression
. svy:logit vote ib2.sex ib3.age ib1.urban
Survey: Logistic regression                     Number of obs      =     14813
                                                Population size    =  20625842
                                                Replications       =       500
                                                Design df          =       500
                                                F(   4,    497)    =    114.00
                                                Prob > F           =    0.0000
------------------------------------------------------------------------------
             |   Observed   Bstrap *
        vote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.sex |  -.0002705   .0484234    -0.01   0.996     -.095409     .094868
             |
         age |
          1  |  -1.254564   .0649763   -19.31   0.000    -1.382225   -1.126904
          2  |  -.7928308    .051653   -15.35   0.000    -.8943144   -.6913472
             |
     0.urban |    .027307   .0566123     0.48   0.630    -.0839204    .1385344
       _cons |   1.549015   .0442291    35.02   0.000     1.462117    1.635912
------------------------------------------------------------------------------

. svy:logistic vote ib3.age ib1.urban ib2.sex
Survey: Logistic regression                     Number of obs      =     14813
                                                Population size    =  20625842
                                                Replications       =       500
                                                Design df          =       500
                                                F(   4,    497)    =    114.00
                                                Prob > F           =    0.0000

------------------------------------------------------------------------------
             |   Observed   Bstrap *
        vote | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |
          1  |      .2852   .0185312   -19.31   0.000     .2510195    .3240348
          2  |   .4525619   .0233762   -15.35   0.000     .4088878    .5009008
             |
     0.urban |   1.027683   .0581795     0.48   0.630     .9195045    1.148589
       1.sex |   .9997295   .0484104    -0.01   0.996     .9090011    1.099514
------------------------------------------------------------------------------

. test 1.age 2.age

Adjusted Wald test
 ( 1)  [vote]1.age = 0
 ( 2)  [vote]2.age = 0

       F(  2,   499) =  227.88
            Prob > F =    0.0000
```

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

WesVar 5.1

Aperçu

WesVar est un progiciel produit par l'organisme Westat. Une version récente du progiciel peut être téléchargée gratuitement à http://www.westat.com/statistical_software/WesVar/index.cfm.

WesVar exécute diverses analyses de données d'enquête en utilisant exclusivement des méthodes de répliques pour calculer l'estimation de la variance. Une des méthodes est celle des répliques répétées équilibrées avec un ajustement de Fay, qui, comme cela est détaillé dans Phillips (2004), peut être utilisée pour obtenir des estimations de la variance bootstrap si les variables de poids bootstrap sont fournies par le chercheur. Dans WesVar, la méthode d'estimation de la variance est précisée lors de la création d'un nouveau fichier de données WesVar. Le fichier qui en découle est ensuite utilisé pour définir des classeurs dans lesquels des tableaux et des régressions peuvent être effectués.

Le tableau suivant illustre les principaux types d'analyses qui peuvent être effectuées avec WesVar 5.1, au moyen de l'estimation pondérée et de l'estimation de la variance bootstrap. Les emplacements où trouver ces analyses dans le logiciel sont aussi mentionnés.

Tableau 5 Principaux types d'analyses avec WesVar 5.1

Type d'analyse	Emplacement
Moyennes	Dans l'onglet tableau de la boîte de nouvelle requête du classeur.
Totaux	
Proportions	
Ratios	
Tests d'indépendance	En choisissant soit RS2 ou RS3 dans la partie de l'ensemble du tableau du classeur.
Quantiles	Dans l'onglet des statistiques calculées du nouveau tableau de requête. Choisir la fonction Quantile.
Régression linéaire	Dans l'onglet de régression dans la boîte de nouvelle requête du classeur.
Régression logistique	
Logit multinominal	

Des instructions claires pour expliquer comment utiliser WesVar sont fournies dans le guide de l'utilisateur, qui peut également être téléchargé gratuitement à http://www.westat.com/statistical_software/WesVar/index.cfm.

WesVar est un programme autonome. Puisqu'il est capable d'importer une grande variété de formats de fichier, il peut être prêt à utiliser par les chercheurs qui possèdent des fichiers de données dans des formats tels que les ensembles de données SAS ou SPSS. L'utilisateur peut également produire les résultats à partir de l'ensemble des classeurs ou de seulement une section dans un ou plusieurs fichiers de textes délimités par des tabulations.

WesVar est doté d'une interface visuelle. Ainsi, les chercheurs qui préfèrent les menus déroulants pour effectuer des analyses devraient être à l'aise avec WesVar.

Liste de vérification du logiciel pour l'exemple tiré de l'ESG

1. **Avez-vous déterminé :**

   a. **Si les estimations pondérées et les estimations de la variance bootstrap (erreurs) requises peuvent être calculées avec le logiciel que vous utilisez?**
   WesVar peut calculer les estimations pondérées et les estimations de la variance bootstrap nécessaires pour les types d'analyses dans l'exemple, et pour plusieurs autres types d'analyses, comme on peut le voir dans le tableau ci-dessus. La sélection des variables de poids bootstrap et la méthode des répliques répétées équilibrées au moment de créer le fichier de données produira de bonnes estimations de la variance bootstrap, si les poids bootstrap sont des poids bootstrap « réguliers ». Si les variables de poids bootstrap sont des bootstrap moyens, la méthode de Fay doit être choisie, ainsi que la spécification d'une valeur de l'ajustement de Fay, afin de calculer correctement les estimations de la variance du bootstrap moyen.

   b. **S'il est possible de produire les statistiques/effectuer les tests nécessaires avec le logiciel que vous utilisez?**
   Pour l'exemple tiré de l'ESG, nous voulons vérifier que chaque coefficient du modèle dans la régression logistique est 0, et aussi que les coefficients des variables nominales des deux groupes d'âge sont simultanément 0. Les résultats de tels tests font partie du résultat par défaut du tableau des coefficients estimés dans l'onglet des données de sortie de la régression, comme nous le démontrerons plus loin. Toutefois, il est également possible d'avoir besoin d'autres statistiques pour vérifier la même chose ou pour tester des relations plus complexes parmi les coefficients du modèle.

2. **Si les tests ne peuvent être effectués et que les statistiques spécifiques souhaitées ne peuvent être calculées, est-il possible de développer un programme post-estimation pour calculer ces statistiques et effectuer les tests dans le logiciel que vous utilisez?**
   Vous ne pouvez développer un programme post-estimation dans WesVar. Toutefois, vous serez peut-être en mesure d'exporter les quantités requises des résultats de sortie de WesVar, et de créer ensuite un programme dans un autre logiciel, en fonction des tests que vous voulez effectuer.

3. **Est-il possible, dans le logiciel, de restreindre l'échantillon ou d'éliminer les observations qui ne concernent pas l'échantillon du fichier de données complet de l'enquête?**
   Lors de la création de l'ensemble de données WesVar, l'onglet du sous-ensemble de la population peut être utilisé pour restreindre l'échantillon à une sous-population donnée.

4. **Si le poids de l'enquête, les poids bootstrap et les variables d'analyse ne sont pas dans le même fichier, savez-vous comment fusionner les différentes sources dans le logiciel que vous utilisez? (En supposant que le logiciel utilisé requiert que tous ces renseignements soient dans le même fichier.)**
   La fusion des différentes sources devra être effectuée dans un autre logiciel avant d'utiliser WesVar. WesVar exige que les poids bootstrap soient dans le même fichier de données que le poids d'enquête et les variables d'analyse.

5. **Tout en effectuant votre analyse, avez-vous vérifié les résultats de sortie de votre logiciel pour déterminer :**
   a. **que la bonne taille d'échantillon a été utilisée?**
      Les données de sortie de la requête indiquent le nombre d'observations lues et fournissent une estimation pondérée de la taille de la population représentée par les observations.
   b. **que la bonne variable de poids a été utilisée?**
      Les données de sortie de la requête donnent le nom de la variable de poids d'enquête utilisée.
   c. **que l'ensemble complet des poids bootstrap a été utilisé?**
      Les données de sortie de la requête produisent une liste des variables des poids bootstrap qui ont été utilisées.
   d. **que l'ajustement pour le bootstrap moyen a été effectué de façon appropriée (si nécessaire)?**
      Il est possible de déterminer si le bon ajustement du bootstrap moyen a été effectué en constatant que l'énoncé de sortie « Fay's Factor » contient l'ajustement de poids bootstrap moyen. Dans le cas du cycle 22 de l'ESG, cette valeur devrait être 0,8 parce qu'elle devrait être égale à $1-C^{-1/2}$, où C est le nombre, 25, d'échantillons bootstrap utilisés pour produire chaque variable de poids bootstrap moyen.
   e. **s'il existait des échantillons bootstrap pour lesquels il n'était pas possible de faire des estimations?**
      Dans l'option de contrôle de l'onglet de demande de régression, dans la partie des fichiers de résultats auxiliaires, la boîte du coefficient de répliques peut être sélectionnée. Cela produit un fichier qui contient tous les coefficients de régression pour chaque réplique bootstrap. S'il est impossible de faire une estimation pour une réplique donnée, ce fichier permettra de faire son identification.

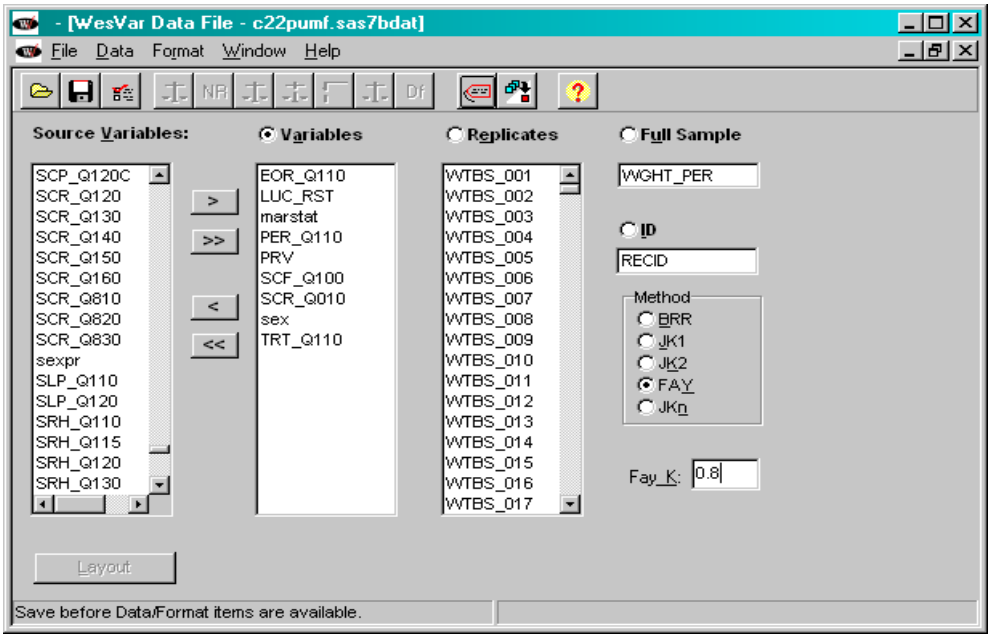
Démonstration de WesVar pour l'exemple tiré de l'ESG

Créer un fichier de données WesVar

WesVar peut accepter plusieurs types de fichier au départ, tels que des fichiers de données Stata, SAS et SPSS. Dans cet exemple, un fichier SAS a été utilisé pour créer le fichier de données WesVar. Pour créer un fichier de données WesVar avec des poids du bootstrap moyen, il faut :

a. transférer les variables de poids de réplique (c'est-à-dire, wtbs-001 à wtbs_500) dans la boîte *Replicates*;
b. transférer la variable de poids d'enquête (c'est-à-dire, wght_per) dans la boîte *Full sample*;
c. pour le bootstrap moyen, préciser la *Méthod* comme étant *Fay* et préciser que *Fay*_K=,8;
d. transférer toutes les variables d'analyse dans la boîte *Variables* et, facultativement, une variable d'identificateur unique dans la boîte ID. Pour cet exemple, RECID est un identificateur d'enregistrement unique;
e. sauvegarder le fichier.

Figure 9 WESVAR création de fichier



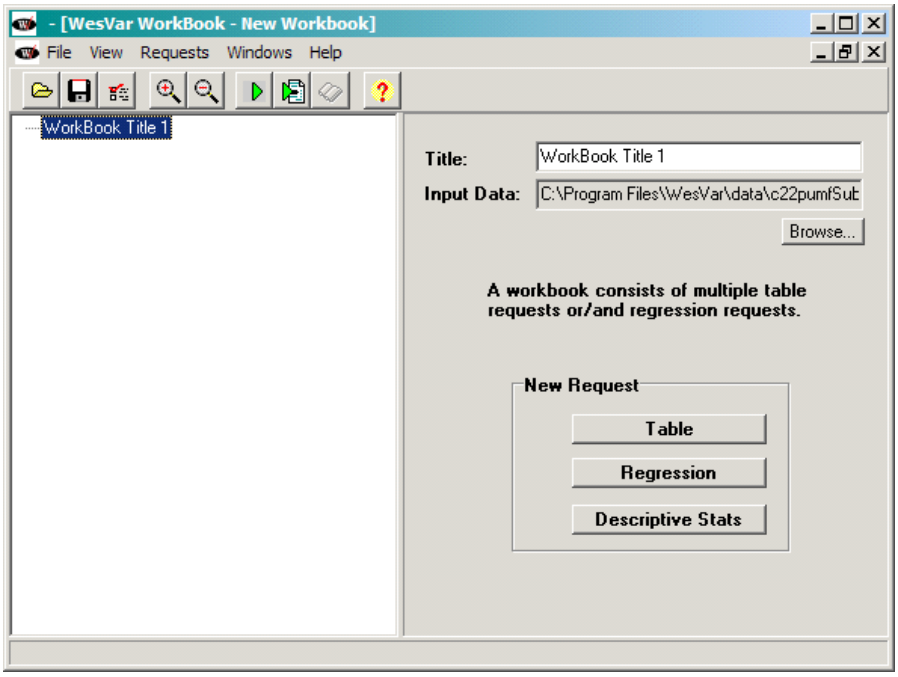
Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Créer un classeur

Comme mentionné dans le guide de l'utilisateur, « En plus d'un fichier de données WesVar, vous devez créer un classeur. Ce classeur est une façon d'organiser les analyses pour un projet ou un ensemble de données. À l'intérieur du classeur, il est possible de définir et d'exécuter des requêtes de tableau et de régression, et aussi d'ouvrir, de voir et d'imprimer leurs résultats de sortie. Un classeur peut contenir plusieurs requêtes de tableau ou de régression ».

Comme on peut le voir ci-dessous, l'écran du classeur de WesVar est divisé en deux parties; le côté droit vous permet de définir et de changer les requêtes d'analyse et le côté gauche, de voir l'organisation du classeur.

Figure 10 WESVAR écran de classeur



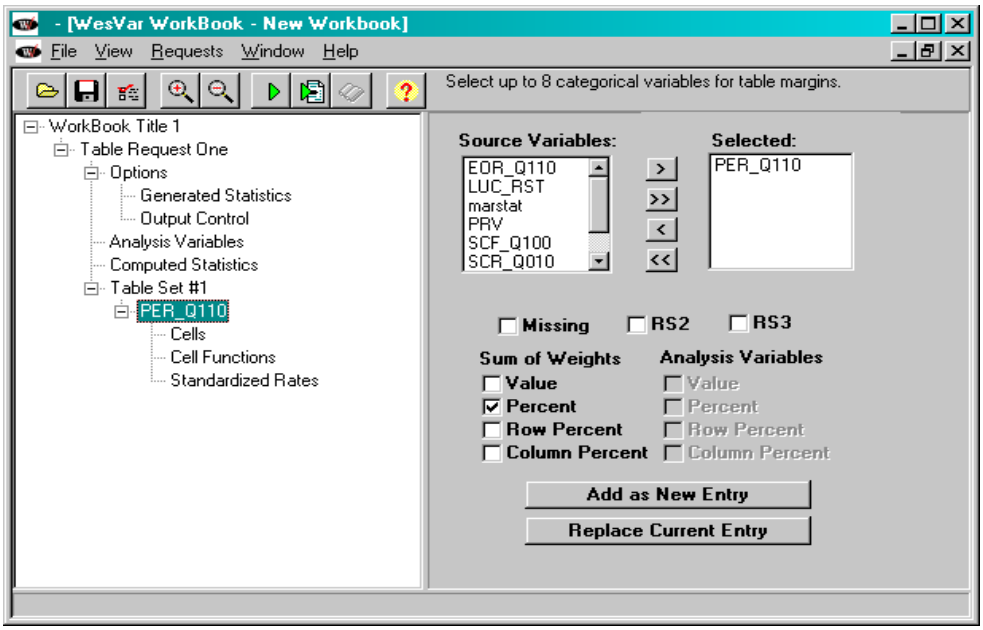
Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Estimation des proportions de la population ayant donné chaque réponse possible à la question de l'action de voter lors de la dernière élection fédérale

Pour préparer cette analyse, le bouton « Table », comme on peut le voir dans le tableau ci-dessus, est sélectionné dans la boîte « New Request ».

Ensuite, il y a un écran sur lequel les détails de la requête d'un tableau sont choisis. Aux fins de cette analyse, la variable PER_Q110 est choisie, car elle constitue la variable qui indique la réponse à la question de l'action de voter lors de la dernière élection fédérale. Puisque le chercheur souhaite obtenir le pourcentage de personnes avec chaque valeur de PER_Q110, la boîte de pourcentage est sélectionnée, comme on peut le voir ci-dessous.

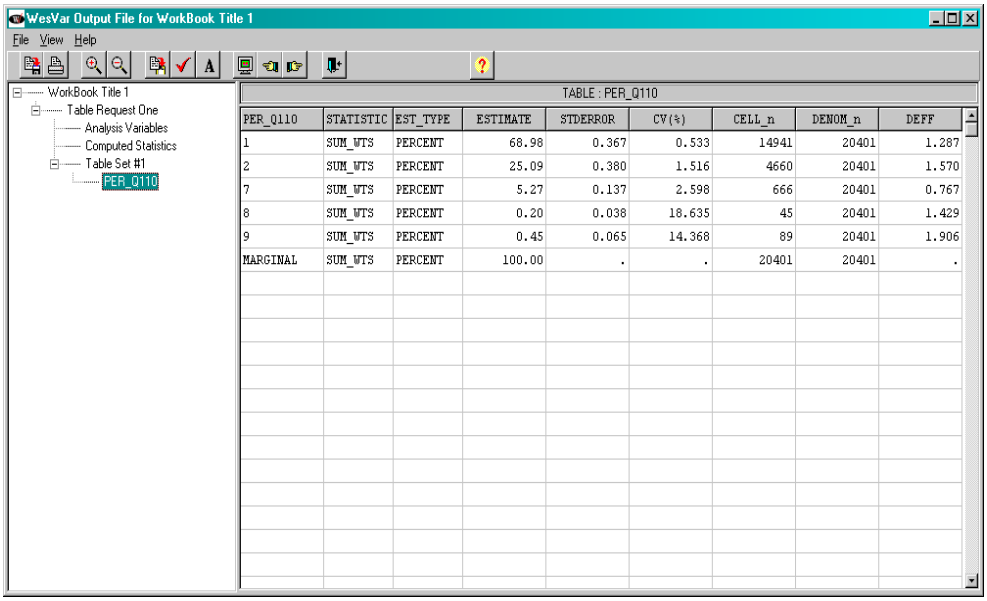
Figure 11 WESVAR requête d'un tableau pour PER_Q110



Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Lorsque toutes les sélections sont effectuées, choisissez «Add as New Entry», et utilisez ensuite le bouton «Run Selected Request». Le résultat, dévoilé ci-dessus, peut être consulté à l'aide du bouton «View Output».

Figure 12 WESVAR sortie



PER_Q110	STATISTIC	EST_TYPE	ESTIMATE	STDERROR	CV(%)	CELL_n	DENOM_n	DEFF
1	SUM_WTS	PERCENT	68.98	0.367	0.533	14941	20401	1.287
2	SUM_WTS	PERCENT	25.09	0.380	1.516	4660	20401	1.570
7	SUM_WTS	PERCENT	5.27	0.137	2.598	666	20401	0.767
8	SUM_WTS	PERCENT	0.20	0.038	18.635	45	20401	1.429
9	SUM_WTS	PERCENT	0.45	0.065	14.368	89	20401	1.906
MARGINAL	SUM_WTS	PERCENT	100.00	.	.	20401	20401	.

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

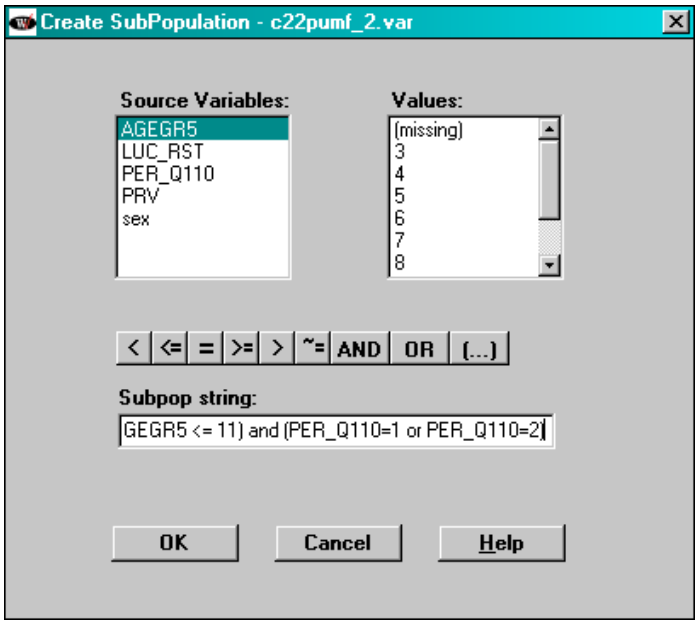
Choisir une sous-population

La sous-population à l'étude est composée des personnes de 20 à 64 ans qui acceptent de révéler s'ils ont voté ou non lors de la dernière élection fédérale (et l'échantillon dans cette sous-population est composé des personnes de 20 à 64 ans qui avaient une valeur de 1 ou 2 pour la variable PER_Q110).

Le bouton du sous-ensemble de la population se trouve dans la fenêtre du fichier de données. L'instruction de sous-population à entrer dans cette fenêtre est la suivante :

```
(AGEGR5  >=  03 and AGEGR5 <= 11) and (PER_Q110=1 or PER_Q110=2)
```

Figure 13 WESVAR sous-population



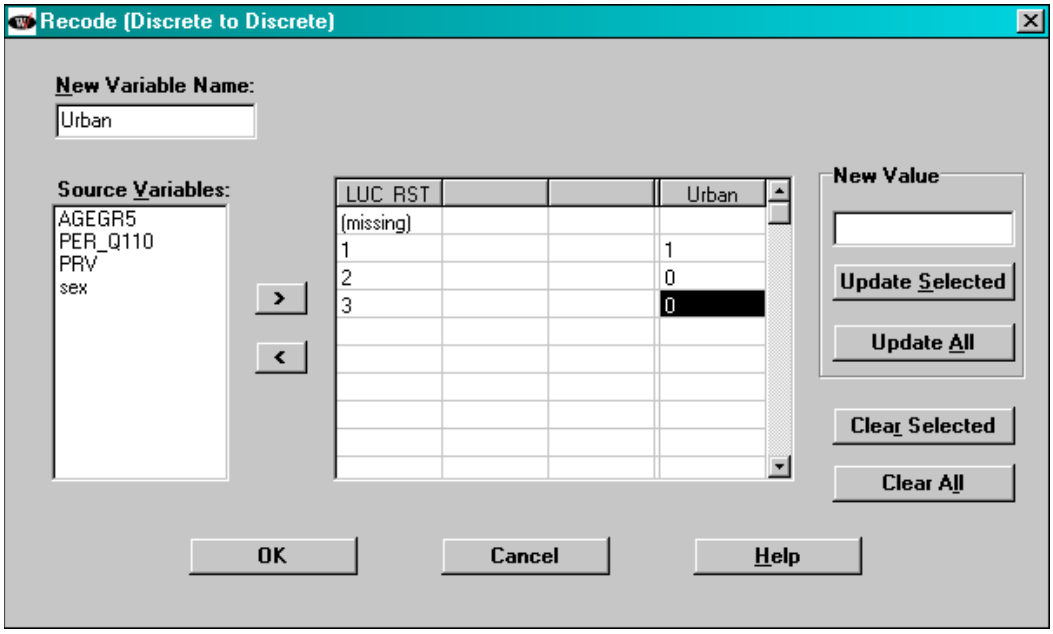
Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Recodage des variables

Le recodage peut être effectué pour créer de nouvelles variables ou pour changer les valeurs des variables. Le bouton de recodage est l'avant-dernier à la droite de la fenêtre «Data File».

Dans la fenêtre ci-dessous, une variable 0/1 nommée « URBAN » est créée à partir de la variable LUC_RST. Dans « URBAN », nous donnons une valeur de 0 aux observations rurales et aux observations de l'Île-du-Prince-Édouard.

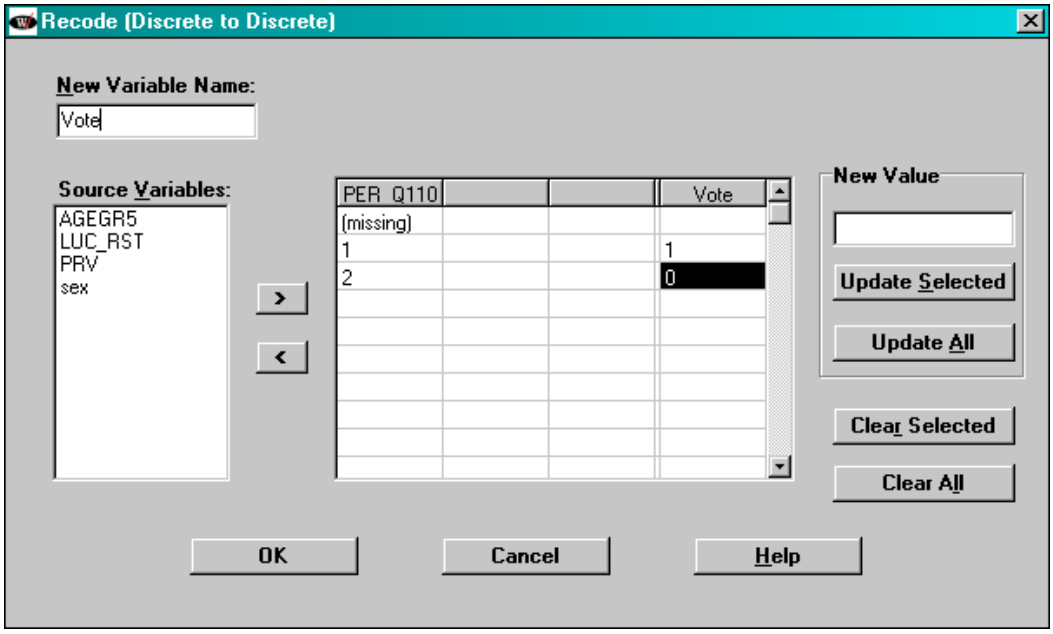
Figure 14 WESVAR recodage de URBAN



Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Dans la fenêtre ci-dessous, une variable 0/1 nommée « VOTE » est créée à partir de PER_Q110. Dans l'échantillon de la sous-population, PER_Q110 accepte seulement les valeurs 1 et 2.

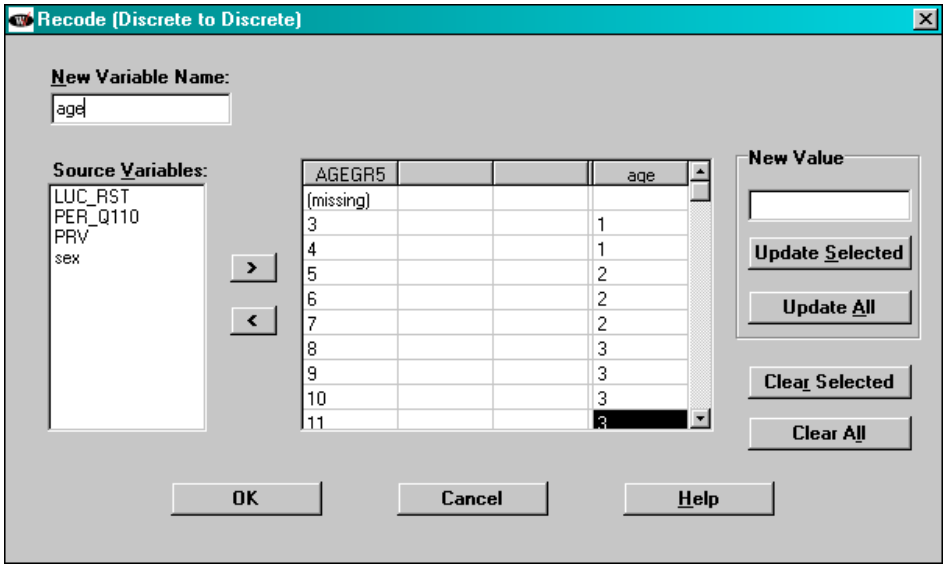
Figure 15 WESVAR recodage de VOTE



Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

La variable « AGE », créée à partir de la variable AGEGR5, a trois catégories : 1, 2, 3 pour les personnes de 20 à 29 ans, 30 à 44 ans et 45 à 64 ans, respectivement. Voir ci-dessous.

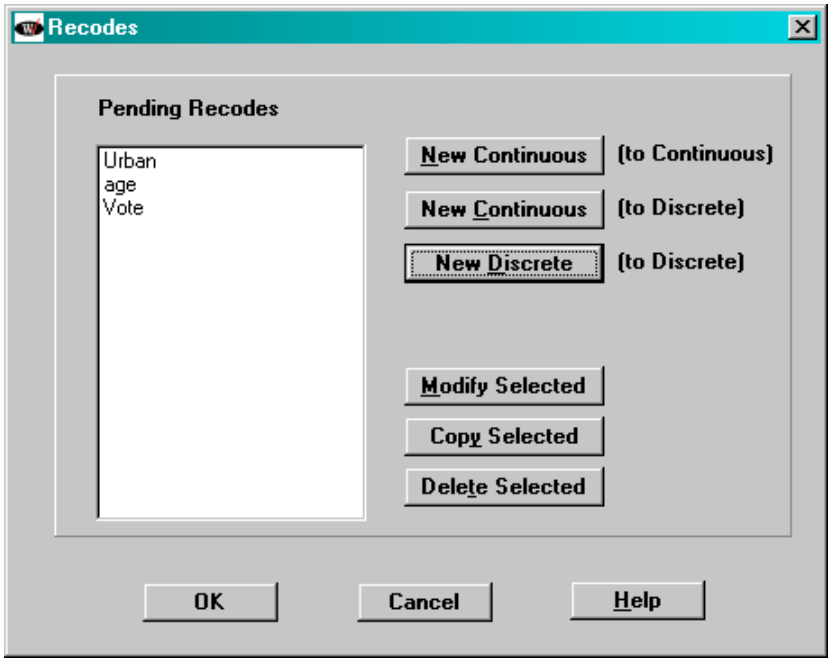
Figure 16 WESVAR recodage de AGE



Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

En cliquant sur le bouton « OK » de l'écran ci-dessous, les nouvelles variables seront ajoutées au fichier de données.

Figure 17 WESVAR recodage

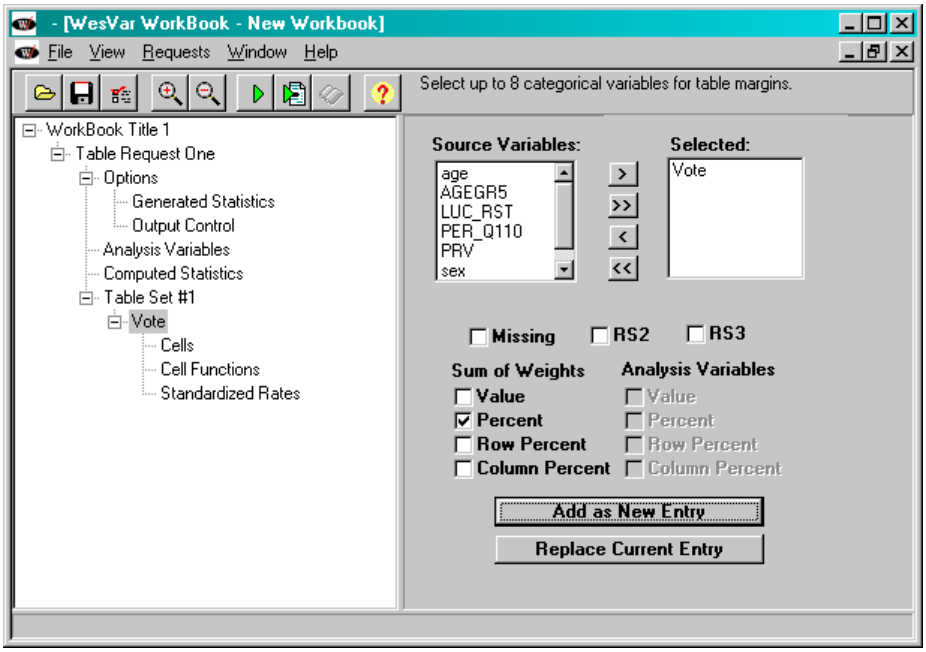


Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Analyse de la variable VOTE

Il faut maintenant estimer la proportion des personnes qui ont répondu « oui » à la question de l'action de voter lors de la dernière élection fédérale, ainsi que la proportion des personnes qui ont répondu « non » à cette même question, étant donné qu'ils font partie de la sous-population d'intérêt. Pour ce faire, une nouvelle requête est créée dans le classeur au moyen du bouton « Table ». Ensuite, la variable VOTE est sélectionnée dans les variables sources. Il faut cliquer sur le bouton «Add as New Entry» pour créer cette requête spécifique.

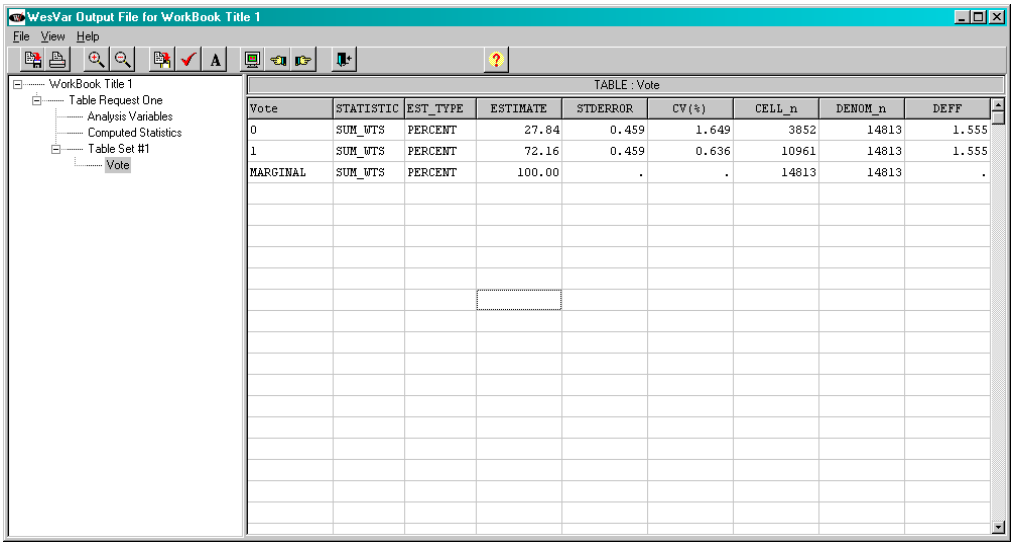
Figure 18 WESVAR requête de tableau pour VOTE



Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Les résultats de cette requête apparaissent ci-dessous.

Figure 19 WESVAR sortie



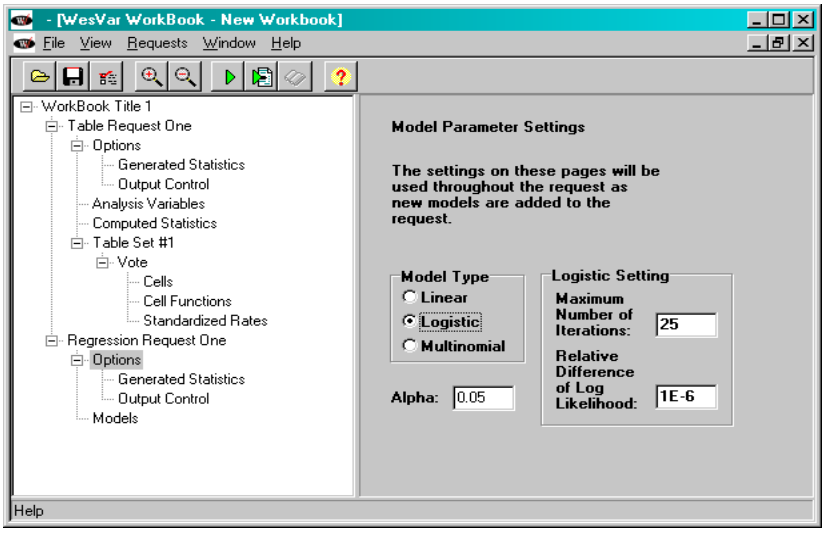
Vote	STATISTIC	EST_TYPE	ESTIMATE	STDERROR	CV(%)	CELL_n	DENOM_n	DEFF
0	SUM_WTS	PERCENT	27.84	0.459	1.649	3852	14813	1.555
1	SUM_WTS	PERCENT	72.16	0.459	0.636	10961	14813	1.555
MARGINAL	SUM_WTS	PERCENT	100.00	.	.	14813	14813	.

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Régression logistique

La régression logistique est utilisée pour modéliser la probabilité de voter lors de la dernière élection fédérale en fonction du sexe, de l'âge et du fait que la personne habitait dans une ville ou non. Le modèle doit être ajusté à l'échantillon dans la sous-population d'intérêt. Après avoir choisi le bouton « Régression » dans la page du classeur, il faut sélectionner « Logistique » dans le type de modèle de la page d'option, comme on peut le voir ci-dessous.

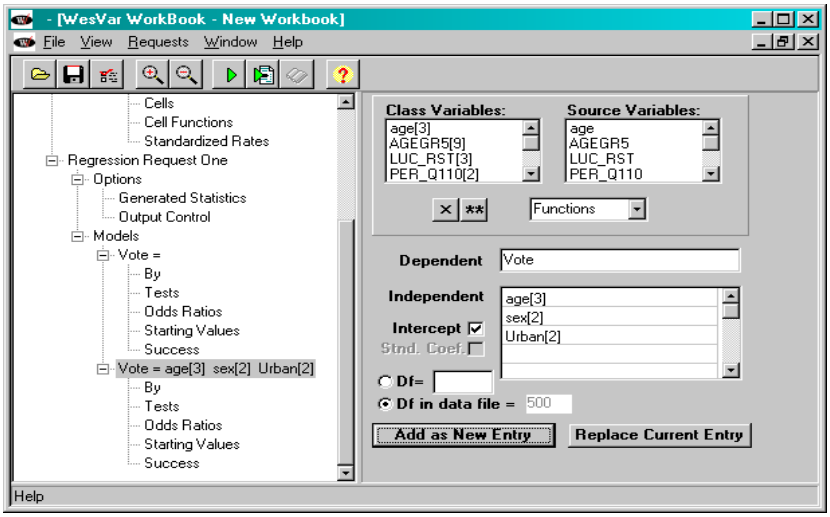
Figure 20 WESVAR régression logistique



Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Ensuite, dans le panneau du modèle de régression, il faut choisir des variables dépendantes et indépendantes en les éliminant de la liste des variables de classification et de la liste des variables sources. (Toute variable avec moins de 256 catégories de réponse apparaîtra dans la liste des variables de classification, avec le nombre de niveaux, et aussi dans la liste des variables sources.) Ici, les variables de classification AGE[3], SEX[2] et URBAN[2] sont sélectionnées pour ensuite être transférées dans la boîte de variable indépendante, ce qui signifie que WesVar créera ensuite le nombre approprié de variables 0/1 dans le modèle pour ces variables. La variable dépendante VOTE doit être choisie à partir de la liste des variables sources. Voir ci-dessous.

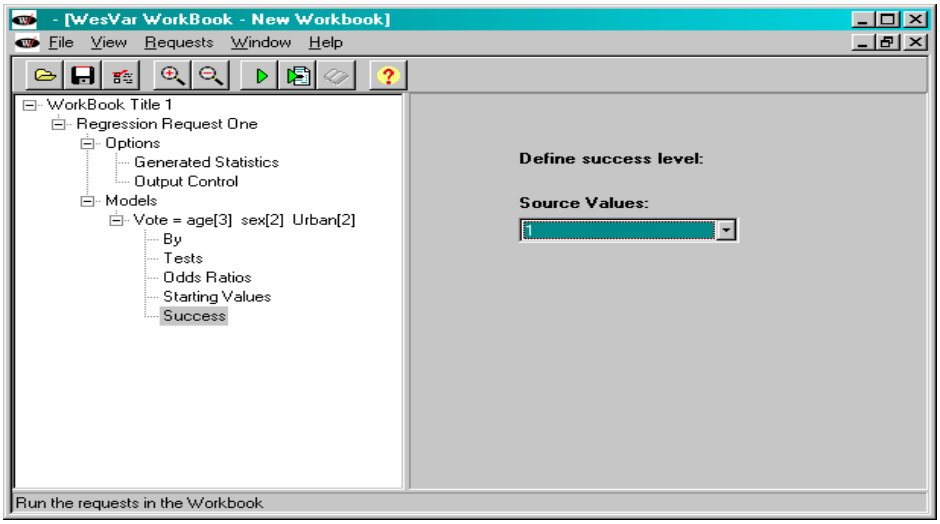
Figure 21 WESVAR regression



Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

L'événement modélisé dans une régression logistique est souvent reconnu comme un « succès ». Pour notre exemple précis, « succès » fait référence à la réponse « oui » d'une personne à la question de voter lors de la dernière élection fédérale. Afin de s'assurer que nous modélisons la probabilité de voter au lieu de la probabilité de ne pas voter, nous choisissons « 1 » comme valeur source dans l'onglet «Success » de l'arbre du classeur. (Rappel : VOTE=1 si une personne a répondu qu'elle a voté et VOTE=0 si la personne a répondu qu'elle n'a pas voté.) Si aucune valeur « succès » n'est précisée, la valeur par défaut est la plus petite des deux valeurs de la variable dépendante. Ce qui signifie que, pour une variable dépendante 0/1, la valeur par défaut « 0 » est définit comme un « succès ».

Figure 22 WESVAR régression logistique

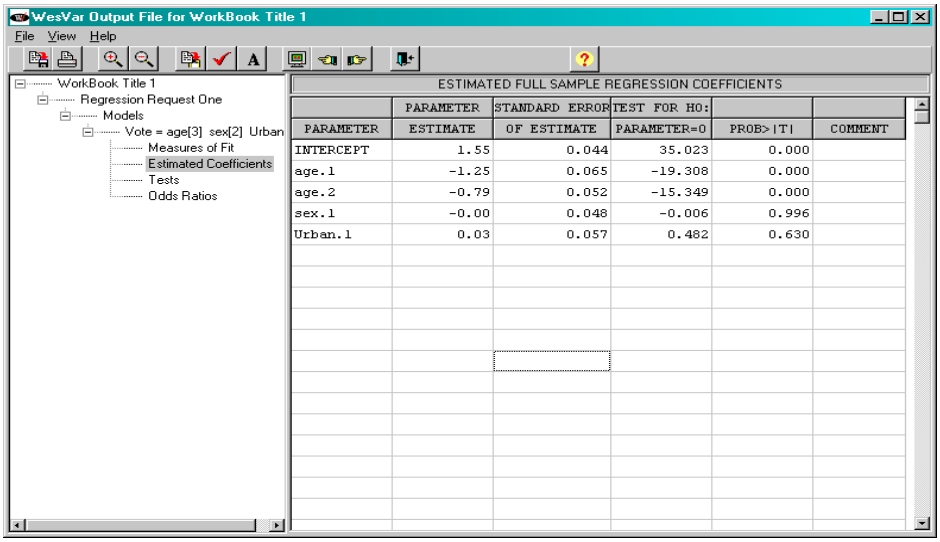


Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

À la suite de l'envoi de la requête d'un modèle logistique, une variété de résultats peuvent être consultés en cliquant sur le bon nœud dans le panneau à la gauche de l'écran. On peut voir ci-dessous les coefficients du

modèle estimés, ainsi que les erreurs types estimées et les résultats des tests standards pour savoir si chaque coefficient sous-jacent est 0.

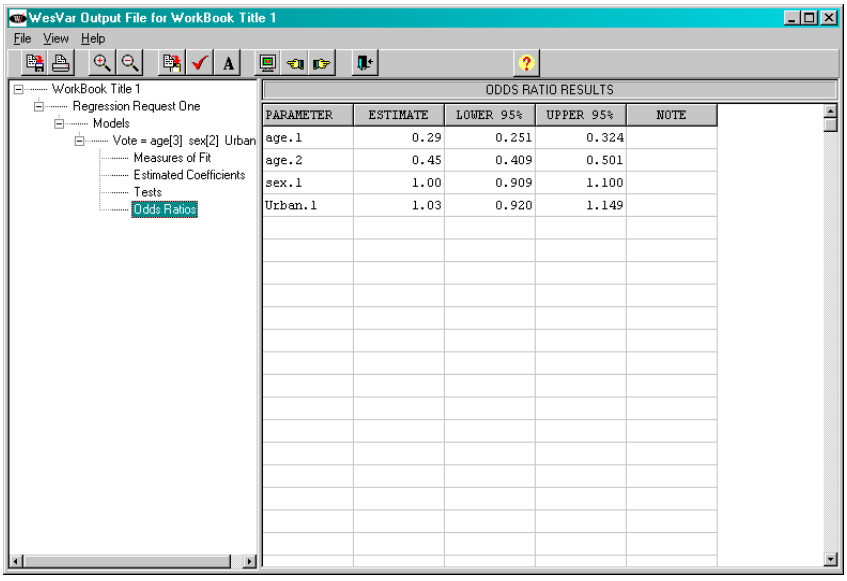
Figure 23 WESVAR sortie



Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

L'image ci-dessous montre le modèle ajusté sous la forme de rapports de cotes, ainsi qu'un intervalle de confiance pour chaque cote.

Figure 24 WESVAR rapports de cotes

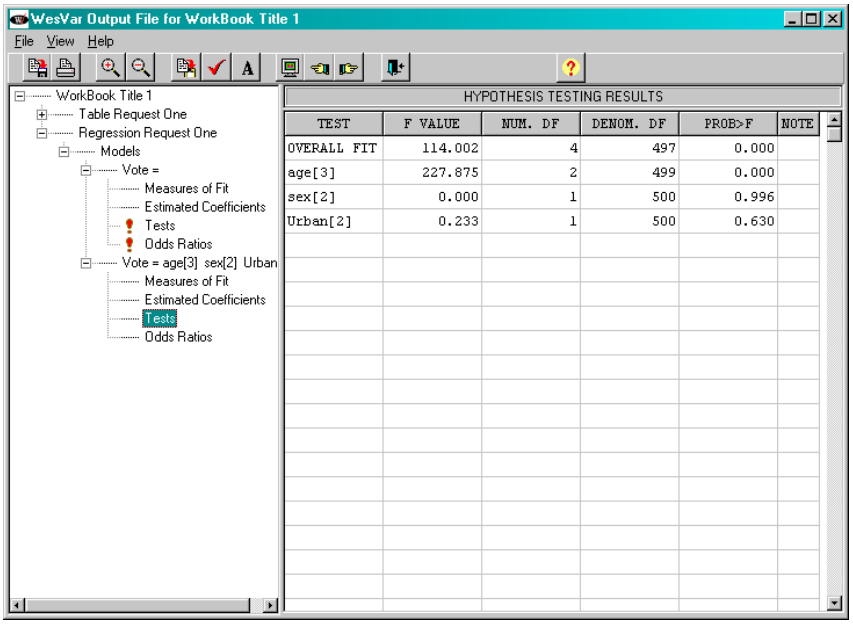


Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Les données de sortie ci-dessus fournissent les résultats d'un test d'ajustement global, ainsi qu'un test de la contribution de chaque variable indépendante au modèle, étant donné que les autres variables sont déjà dans

le modèle. Le test pour l'âge vérifie, pour sa part, que les coefficients sous-jacents des variables nominales des deux groupes d'âge sont simultanément égaux à 0. C'est le test qui intéresse le chercheur.

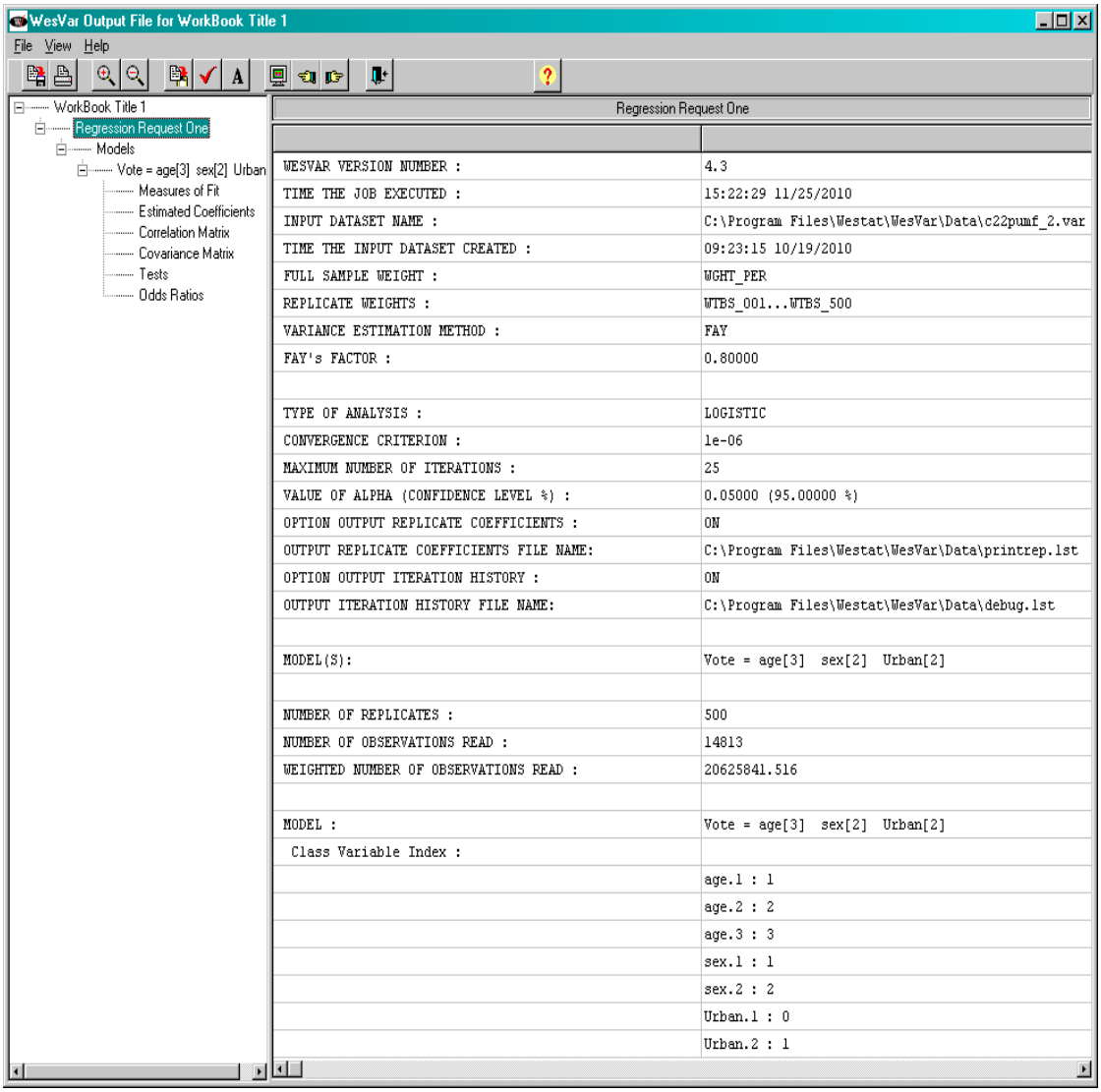
Figure 25 WESVAR tests d'hypothèse



Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Les résultats de sortie incluent aussi un log qui résume la requête de régression logistique. Une portion du log est montrée ci-dessous. Ce log aide le chercheur à vérifier si l'analyse a été effectuée comme voulue.

Figure 26 WESVAR journal de la régression logistique



WESVAR VERSION NUMBER :	4.3
TIME THE JOB EXECUTED :	15:22:29 11/25/2010
INPUT DATASET NAME :	C:\Program Files\Westat\WesVar\Data\c22pumf_2.var
TIME THE INPUT DATASET CREATED :	09:23:15 10/19/2010
FULL SAMPLE WEIGHT :	WGHT_PER
REPLICATE WEIGHTS :	WTBS_001...WTBS_500
VARIANCE ESTIMATION METHOD :	FAY
FAY's FACTOR :	0.80000
TYPE OF ANALYSIS :	LOGISTIC
CONVERGENCE CRITERION :	1e-06
MAXIMUM NUMBER OF ITERATIONS :	25
VALUE OF ALPHA (CONFIDENCE LEVEL %) :	0.05000 (95.00000 %)
OPTION OUTPUT REPLICATE COEFFICIENTS :	ON
OUTPUT REPLICATE COEFFICIENTS FILE NAME:	C:\Program Files\Westat\WesVar\Data\printrep.lst
OPTION OUTPUT ITERATION HISTORY :	ON
OUTPUT ITERATION HISTORY FILE NAME:	C:\Program Files\Westat\WesVar\Data\debug.lst
MODEL(S):	Vote = age[3] sex[2] Urban[2]
NUMBER OF REPLICATES :	500
NUMBER OF OBSERVATIONS READ :	14813
WEIGHTED NUMBER OF OBSERVATIONS READ :	20625841.516
MODEL :	Vote = age[3] sex[2] Urban[2]
Class Variable Index :	
	age.1 : 1
	age.2 : 2
	age.3 : 3
	sex.1 : 1
	sex.2 : 2
	Urban.1 : 0
	Urban.2 : 1

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

SAS 9.2

Aperçu

SAS 9.2 est la première version de SAS qui offre quelques méthodes de répliques pour calculer l'estimation de la variance dans ses quatre procédures d'analyses d'enquête. L'option des répliques répétées équilibrées peut être utilisée avec les poids bootstrap fournis par l'utilisateur, afin d'obtenir des estimations de la variance bootstrap (tel qu'expliqué dans Phillips (2004)).

Le tableau suivant illustre les principaux types d'analyses qui peuvent être effectuées avec SAS 9.2, au moyen de l'estimation pondérée et de l'estimation de la variance bootstrap. Le nom des procédures SAS pour obtenir ces analyses est également fourni.

Tableau 6 Principaux types d'analyses avec SAS 9.2

Type d'analyse	Procédure
Moyennes (y compris les moyennes géométriques)	Proc surveymeans
Totaux	Proc surveymeans
Proportions/pourcentages	Proc surveyfreq
Tests d'indépendance dans les tableaux croisés	Proc surveyfreq
Régression linéaire	Proc surveyreg
Régression logistique	Proc surveylogistic
Logit multinominal	Proc surveylogistic, Link=Clogit, Ref=
Cotes proportionnelles	Proc surveylogistic, Link=Glogit
Régression probit	Proc surveylogistic, Link=Probit
Log-log complémentaire	Proc surveylogistic, Link=Cloglog
Risques proportionnels (Cox)	Proc Surveyphreg

On peut noter que, dans le tableau et dans le guide de l'utilisateur de SAS 9.2, plusieurs techniques analytiques sont incluses dans chaque procédure. Toutefois, il faut noter que les estimations de la variance pour les centiles ne peuvent être calculées lorsque l'option des répliques répétées équilibrées est utilisée, même si le guide de l'utilisateur ne l'indique pas. Les estimations de la variance pour les centiles, en utilisant l'option des répliques répétées équilibrées, peuvent être calculées dans la version SAS 9.3.

Liste de vérification du logiciel pour l'exemple tiré de l'ESG

1. **Avez-vous déterminé :**

   a. **Si les estimations pondérées et les estimations de la variance bootstrap (erreurs) requises peuvent être calculées avec le logiciel que vous utilisez?**
   SAS peut calculer les estimations pondérées et les estimations de la variance bootstrap nécessaires pour ces types d'analyses aux fins de cet exemple, et pour plusieurs autres types d'analyses, comme on peut le voir dans le tableau ci-dessus. Le fait de choisir l'option des répliques répétées équilibrées en SAS pour calculer l'estimation de la variance et fournir les variables de poids bootstrap produira des estimations de la variance bootstrap. Le fait de choisir l'option des répliques répétées équilibrées avec un ajustement de Fay permettra de calculer correctement les estimations de la variance du bootstrap moyen.
   Ainsi, pour toutes les procédures SAS, il faut :
   1. Dans l'énoncé **PROC**, inclure l'option **varmethod=BRR(FAY=*f*)**.
      Fay=f est l'ajustement du bootstrap moyen et f devrait être établi à $1-C^{-1/2}$, où C est le nombre d'échantillons bootstrap utilisés pour produire chaque variable de poids du bootstrap moyen. (Nota : Si vous avez des poids bootstrap « réguliers » au lieu de poids bootstrap moyens, vous avez seulement à inclure varmethod=BRR dans l'énoncé PROC).ment to identify the weight variable to be used for weighted estimation.
   2. Inclure un énoncé **WEIGHT** pour identifier la variable de poids à être utilisée pour obtenir l'estimation pondérée.
   3. Inclure un énoncé **REPWEIGHT** pour indiquer le nom des variables de poids bootstrap dans le fichier de données.

   Pour l'exemple tiré de l'ESG, pour lequel la variable de poids est wght_per, les 500 variables de poids du bootstrap moyen sont wtbs_001 à wtbs_500 et chaque variable est formée de 25 échantillons bootstrap, chaque procédure bootstrap utilisée devrait contenir les éléments suivants :

```
PROC procedurename  data=SAS_datafile_name  varmethod=BRR(FAY=0.8);
WEIGHT wght_per;
REPWEIGHT wtbs_001-wtbs_;
+Other statements required by the procedure
```

   b. **S'il est possible de produire les statistiques/effectuer les tests nécessaires avec le logiciel que vous utilisez?**
   Pour l'exemple tiré de l'ESG, nous voulons vérifier que chaque coefficient du modèle dans la régression logistique est 0, et aussi que les coefficients des variables nominales des deux groupes d'âge sont simultanément 0. Les résultats de ces tests font partie du résultat par défaut de PROC SURVEYLOGISTIC, comme nous le démontrerons plus loin. Toutefois, il est également possible de faire d'autres tests statistiques pour vérifier la même chose ou pour tester des relations plus complexes parmi les coefficients du modèle.

2. **Si les tests ne peuvent être effectués et que les statistiques spécifiques souhaitées ne peuvent être calculées, est-il possible de développer un programme post-estimation pour calculer ces statistiques et effectuer les tests dans le logiciel que vous utilisez?**
   Il est habituellement possible d'entrer les résultats SAS dans un fichier de données SAS et ensuite de développer un programme SAS pour calculer ce que l'on souhaite calculer. Cela n'était pas nécessaire aux fins de notre exemple.

3. **Est-il possible, dans le logiciel, de restreindre l'échantillon ou d'éliminer les observations qui ne concernent pas l'échantillon du fichier de données complet de l'enquête?**
   Une analyse avec SAS peut être restreinte à un échantillon avec des caractéristiques particulières (souvent nommé échantillon d'une sous-population particulière) en utilisant un énoncé WHERE dans une procédure SAS. L'énoncé WHERE doit être inclus dans chaque procédure SAS pour laquelle un échantillon restreint est utilisé.

   Par contre, une solution plus simple consiste à restreindre l'échantillon ou à éliminer les observations qui ne concernent pas l'échantillon du fichier de donnée d'enquête complet en procédant au codage d'une étape DATA avant d'utiliser une procédure d'enquête SAS. Cela produit un plus petit fichier de données qui peut ensuite être utilisé comme fichier de données d'entré pour toutes les procédures SAS pour lesquelles un échantillon restreint est requis.

4. **Si le poids de l'enquête, les poids bootstrap et les variables d'analyse ne sont pas dans le même fichier, savez-vous comment fusionner les différentes sources dans le logiciel que vous utilisez? (En supposant que le logiciel utilisé requiert que tous ces renseignements soient dans le même fichier.)**
   La fusion des différentes sources devra être effectuée dans une étape DATA ou en SQL avant d'utiliser les procédures d'enquête SAS. Toutefois, SAS n'exige pas que les poids bootstrap soient dans le même fichier de données que le poids d'enquête et les variables d'analyse. Le guide d'utilisateur SAS donne des directives sur la façon de préciser un fichier différent pour les poids bootstrap.

5. **Tout en effectuant votre analyse, avez-vous vérifié les résultats de sortie de votre logiciel pour déterminer :**
   a. **Que la bonne taille d'échantillon a été utilisée?**
      Les résultats par défaut d'une procédure SAS donnent le nombre d'observations lues et le nombre d'observations utilisées dans l'analyse effectuée par la procédure. Ils fournissent également les valeurs pondérés de ces quantités.
   b. **Que la bonne variable de poids a été utilisée?**
      Les résultats par défaut d'une procédure SAS donnent le nom de la variable de poids d'enquête utilisée.
   c. **Que l'ensemble complet des poids bootstrap a été utilisé?**
      Les résultats par défaut d'une procédure SAS produisent une liste de nombre de poids bootstrap qui ont été utilisés.
   d. **Que l'ajustement pour le bootstrap moyen a été effectué de façon appropriée (si nécessaire)?**
      Il est possible de déterminer si le bon ajustement pour le bootstrap moyen a été effectué en observant si l'énoncé de sortie «Fay Coefficient» contient l'ajustement du bootstrap moyen. Dans le cas du cycle 22 de l'ESG, cette valeur devrait être 0,8.
   e. **S'il existait des échantillons bootstrap pour lesquels il n'était pas possible de faire des estimations?**
      S'il y a des échantillons bootstrap pour lesquels il était impossible de faire des estimations, ils sont identifiés dans les résultats de sortie du SAS. Cela ne s'est pas produit pour l'exemple tiré de l'ESG.

Programme SAS et résultats pour l'exemple tiré de l'ESG

Programme

```
/* PARTIE 1 */
options linesize=80;
libname pumfl '\\SASD6\Sasd-Dssea-Public\DATA\GSS\DLI\CYCLE22\C22MDFSasAndCode-EngFr';

data c22pumf;
 set pumfl.c22pumf;
run;

/* PARTIE 2 */
/*Analyse descriptive préliminaire*/
proc surveyfreq data=c22pumf varmethod=BRR(FAY=0.8);
 tables PER_Q110 ;
 repweights  wtbs_001 - wtbs_500 ;
 weight wght_per  ;
run;

/* PARTIE 3 */
/*Recoder les variables et sélectionner les observations dans la sous-population*/
data c22pumf_sub;
 set pumfl.c22pumf;
  /*Subpopulation of voters aged 20 to 64*/
 if agegr5 ge 03 and agegr5 le 11;
 if Per_Q110 =1 or Per_Q110 =2;
 /*Recode of the Urban variable*/
 if LUC_RST=1 then Urban=1;
 else Urban=0;
 /*Recode of age variable into 3 categories*/
 if agegr5 le 04 then age = 1;
 else if agegr5 le 07 then age =2;
 else if agegr5 le 11 then age =3;
 /*Recode of the voted variable*/
 if Per_Q110 =1 then Vote =1;
 else Vote =0;
run;

/* PARTIE 4 */
proc surveyfreq data=c22pumf_sub varmethod=BRR(FAY=0.8);
 tables Vote ;
 repweights  wtbs_001 - wtbs_500 ;
 weight wght_per ;
run;

/* PARTIE 5 */
/*Si RURAL et SEXE sont inclus dans l'énoncé de classe sans l'option PARAM=REF, SAS
utilisera ses propres paramètres de codage pour les variables nominales (p. ex. -1, 1
au lieu de 0; 1, toutefois, aura un effet sur les valeurs du paramètre)*/
proc surveylogistic data=c22pumf_sub varmethod=BRR(FAY=0.8);
 class Vote sex(REF=LAST) age(REF=LAST)  Urban(REF=LAST) / PARAM=REF;
 model Vote(event='1')= sex age Urban;
 repweights wtbs_001 - wtbs_500;
 weight wght_per;
run;
```

Commentaires au sujet du programme et des résultats

PARTIE 1

Cette partie du programme est de la programmation SAS, où l'ensemble de données SAS du FMGD initial est précisé. À la suite de l'exécution de l'étape DATA, le log SAS (pas présenté) indique que l'ensemble de données c22pumf contient 20 401 enregistrements. Cette partie du programme ne produit aucun résultat.

PARTIE 2

Cette partie du programme fait appel à PROC SURVEYFREQ pour obtenir des estimations des proportions de la totalité de la population qui a donné différents types de réponses à la question de l'action de voter lors de la dernière élection fédérale. Cela permet de faire une inspection préliminaire de la variable Per_Q110.

Les résultats fournissent la méthode d'estimation de la variance utilisée, le nombre de répliques utilisées, la taille de l'échantillon et la taille estimée de la population.

À la lumière des résultats, 94 % (c'est-à-dire, 68,98+25,09=94,07) de la population ciblée par le cycle 22 de l'ESG est estimée avoir répondu oui ou non à la question de l'action de voter lors de la dernière élection fédérale. Le reste de la population n'a donné aucune des ces deux réponses pour diverses raisons.

Figure 27 SAS SURVEYFREQ PER_Q110

The SAS System

The SURVEYFREQ Procedure

Data Summary

Number of Observations	20401
Sum of Weights	27261809.7

Variance Estimation

Method	BRR
Replicate Weights	C22PUMF
Number of Replicates	500
Fay Coefficient	0.800

Table of PER_Q11

PER_Q110	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
1	14941	18805258	100161	68.9802	0.3674
2	4660	6841080	103698	25.0940	0.3804
7	666	1437779	37358	5.2740	0.1370
8	45	54861	10224	0.2012	0.0375
9	89	122832	17648	0.4506	0.0647
Total	20401	27261810	7.29343E-7	100.000	

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

PARTIE 3

Même si ce ne fut pas le cas pour l'exemple, il est possible dans l'énoncé TABLES de SURVEYFREQ d'obtenir les coefficients de variation et les intervalles de confiance pour l'estimation des pourcentages. Les limites de confiance sont par défaut fondées sur l'hypothèse que le ratio d'une proportion et de son erreur type a une distribution t avec un nombre de degrés de liberté égal au nombre de poids bootstrap. Toutefois, plusieurs autres méthodes pour calculer un intervalle de confiance peuvent être choisies, comme il est décrit dans le guide de l'utilisateur de SAS.

PARTIE 4

Cette partie du programme fait appel à PROC SURVEYFREQ dans SAS pour obtenir des estimations du pourcentage des personnes de 20 à 64 ans qui ont répondu « oui » et le pourcentage des personnes qui ont répondu « non » à la question de l'action de voter lors de la dernière élection fédérale, en supposant qu'ils ont donné une de ces deux réponses à la question. Il faut noter que la nouvelle variable VOTE est utilisée avec l'échantillon restreint. Il aurait été possible, au lieu d'avoir utilisé l'ensemble de données complet, de le restreindre à la sous-population d'intérêt en incluant un énoncé WHERE dans PROC SURVEYFREQ.

Figure 28 SAS SURVEYFREQ VOTE

```
                                The SAS System

.

                           The SURVEYFREQ Procedure

                                 Data Summary

                      Number of Observations         14813
                      Sum of Weights            20625841.5


                             Variance Estimation

                       Method                         BRR
                       Replicate Weights          C22PUMF
                       Number of Replicates           500
                       Fay Coefficient              0.800


                                Table of Vote

                              Weighted    Std Dev of               Std Err of
 Vote     Frequency     Frequency      Wgt Freq    Percent       Percent
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
    0          3852       5742406         94867    27.8408        0.4592
    1         10961      14883435         95794    72.1592        0.4592

Total         14813      20625842         19623    100.000
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
```

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

PARTIE 5

Cette partie du programme est l'ajustement du modèle logistique à l'échantillon restreint au moyen de PROC SURVEYLOGISTIC dans SAS. La procédure modélise le logit de la probabilité que VOTE=1 dans l'échantillon restreint. Il faut noter que toutes les informations à propos des poids, des poids bootstrap, etc. doivent être incluses dans la procédure utilisée.

Notons que le résultat par défaut fournit des informations sur les variables catégorielles SEX, AGE et URBAN qui sont identifiées dans l'énoncé CLASS.

Figure 29 SAS SURVEYLOGISTIC information du modèle

The SAS
The SURVEYLOGISTIC Procedure
Model Information

Data Set	WORK.C22PUMF
Response Variable	Vote
Number of Response Levels	2
Weight Variable	WGHT_PER
Model	Binary Logit
Optimization Technique	Fisher's Scoring

Number of Observations Read	14813
Number of Observations Used	14813
Sum of Weights Read	20625842
Sum of Weights Used	20625842

Response Profile

Ordered Value	Vote	Total Frequency	Total Weight
1	0	3852	5742406
2	1	10961	14883435

Probability modeled is Vote=1.

Class Level Information

Class	Value	Design Variables		
sex	1	1		
	2	0		
age	1	1	0	
	2	0	1	
	3	0	0	
Urban	0	1		
	1	0		

Variance Estimation

Method	BRR
Number of Replicates	500
Fay Coefficient	0.8
Replicate Weights Data Set	C22PUMF

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Figure 30 SAS SURVEYLOGISTIC convergence

The SURVEYLOGISTIC Procedure
Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	24398053	23357911
SC	24398060	23357949
-2 Log L	24398051	23357901

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1040149.87	4	<.0001
Score	1032256.59	4	<.0001
Wald	458.7692	4	<.0001

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Le tableau ci-dessous fournit le résultat des tests pour vérifier si chaque variable indépendante contribue de façon significative au modèle. Le test statistique pour l'âge vérifiera si les coefficients sous-jacents des variables des deux groupes d'âge sont simultanément 0, étant donné que les autres variables sont dans le modèle.

Figure 31 SAS analyse des effets

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	0.0000	0.9955
age	2	456.6735	<.0001
Urban	1	0.2327	0.6296

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Le modèle ajusté est présenté en deux formes et dans deux tableaux différents; le premier, où les coefficients ajustés sont montrés, et le deuxième, où les coefficients sont exprimés comme des rapports de cotes. Le test du khi carré de Wald est utilisé pour les paramètres du modèle dans le premier tableau, tandis que dans le deuxième tableau, on peut voir les intervalles de confiance pour le rapport de cotes.

Figure 32 SAS estimation du maximum de vraisemblances

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.5490	0.0442	1226.6123	<.0001
sex	1	1	-0.00027	0.0484	0.0000	0.9955
age	1	1	-1.2546	0.0650	372.8043	<.0001
age	2	1	-0.7928	0.0517	235.6012	<.0001
Urban	0	1	0.0273	0.0566	0.2327	0.6296

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Figure 33 SAS rapports de cotes

The SAS System

The SURVEYLOGISTIC Procedure

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
sex	1 vs 2	1.000	0.909	1.099
age	1 vs 3	0.285	0.251	0.324
age	2 vs 3	0.453	0.409	0.501
Urban	0 vs 1	1.028	0.920	1.148

Association of Predicted Probabilities and Observed Responses

Percent Concordant	51.3	Somers' D	0.255
Percent Discordant	25.8	Gamma	0.330
Percent Tied	22.9	Tau-a	0.098
Pairs	42221772	c	0.627

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

BootVar 3.2 pour SAS

Aperçu

La version SAS de BootVar a été développée par des méthodologistes de Statistique Canada pour estimer les variances au moyen de la méthode du bootstrap. BootVar ne produit pas de poids bootstrap, mais utilise ceux fournis dans les fichiers de données d'enquête.

Les macros de BootVar calculent une variance en utilisant les écarts quadratiques des estimations bootstrap à partir de la moyenne des estimations bootstrap. C'est la même chose que la méthode par défaut dans Stata, tandis que dans le cas de SUDAAN, SAS et WesVar, ils calculent une variance en utilisant les écarts quadratiques des estimations bootstrap à partir de l'estimation de l'échantillon au complet. Les deux méthodes peuvent produire des estimations de la variance légèrement différentes.

Le tableau suivant illustre les principaux types d'analyses qui peuvent être effectuées par BootVar 3.2, au moyen de l'estimation pondérée et de l'estimation de la variance bootstrap. Le nom de la macro BootVar pour obtenir chaque analyse est également donné.

Tableau 7 Principaux types d'analyses avec BootVar 3.2

Type d'analyse	Macro de BootVar
Moyennes	%ratio
Totaux	%total
Proportions	%ratio
Tests d'indépendance	%chi2
Quantiles	%prcntle
Régression linéaire	%regress
Régression logistique	%logreg
Difference entre les deux rapports	%diffrat

Les détails concernant chaque macro peuvent être consultés dans le guide de l'utilisateur de BootVar, qui est accessible lorsque le logiciel est installé. Il est important de noter que les statistiques de sortie produites avec BootVar sont plus limitées que celles produites par les autres logiciels dont il est question dans ce document.

Le guide de l'utilisateur de BootVar mentionne que « la version 3.2 de BootVar pour SAS a été mise à l'essai et fonctionne avec la version 9.1 de SAS. Des résultats apropriés ne sont pas garantis lorsqu'on utilise le programme avec des versions plus vieilles ou plus récentes de SAS ».

Liste de vérification du logiciel pour l'exemple tiré de l'ESG

1. **Avez-vous déterminé :**

   a. **Si les estimations pondérées et les estimations de la variance bootstrap (erreurs) requises peuvent être calculées avec le logiciel que vous utilisez?**
   BootVar peut calculer les estimations pondérées et les estimations de la variance bootstrap requises pour ces types d'analyses aux fins de l'exemple, ainsi que pour plusieurs autres types d'analyses, tel qu'illustré dans le tableau ci-dessus. Le fait de choisir la variable macro « R » permet de calculer correctement les estimations de la variance du bootstrap moyen.

   Pour l'exemple tiré de l'ESG, la variable de poids est wght_per, les 500 variables de poids du bootstrap moyen sont wtbs_001 à wtbs_500 et chaque variable est composée de 25 échantillons bootstrap standards. De plus, la variable RECID contenue dans le fichier de données est un identificateur d'enregistrement. Avant d'exécuter toute macro de BootVar avec ce fichier de données, il faut s'assurer d'avoir les éléments suivants dans un programme SAS :

   1. %let ident = RECID;
      indique les noms des variables d'identification unique pour chaque observation;
   2. %let fwgt = wght_per;
      identifie la variable de poids à utiliser pour l'estimation pondérée;
   3. %let bsw = wtbs_;
      indique le préfixe du nom des variables de poids bootstrap dans le fichier de données;
   4. %let R = 25;
      indique que les variables de poids bootstrap sont des variables du bootstrap moyen, chaque variable créée à partir de 25 échantillons bootstrap. Si vous avez des poids bootstrap « réguliers », établissez « R » à 1;
   5. %let B = 500 ;
      le nombre de poids bootstrap.

   b. **S'il est possible de produire les statistiques/effectuer les tests nécessaires avec le logiciel que vous utilisez?**
   Pour l'exemple tiré de l'ESG, nous voulons vérifier que chaque coefficient du modèle dans la régression logistique soit 0, et aussi que les coefficients des variables nominales des deux âges soient 0. Tandis que BootVar permet d'effectuer des tests pour les coefficients individuels en tant que résultat par défaut de **%logreg**, il n'a pas la capacité d'effectuer un examen des cas plus complexes pour voir si les coefficients des variables des deux âges sont simultanément 0.

2. **Si les tests ne peuvent être effectués et que les statistiques spécifiques souhaitées ne peuvent être calculées, est-il possible de développer un programme post-estimation pour calculer ces statistiques et effectuer les tests dans le logiciel que vous utilisez?**
   Citation du guide de l'utilisateur : « Puisque BootVar est distribué en tant que code source ouvert, les utilisateurs possédant de l'expérience dans la programmation SAS peuvent modifier le code du programme, afin de satisfaire les besoins qui ne sont pas abordés par BootVar ».

3. **Est-il possible, dans le logiciel, de restreindre l'échantillon ou d'éliminer les observations qui ne concernent pas l'échantillon du fichier de données complet de l'enquête?**
   Une analyse avec BootVar peut être restreinte à un échantillon avec des caractéristiques particulières (qu'on nomme souvent l'échantillon dans une sous-population particulière). Restreindre l'échantillon ou éliminer les observations qui ne font pas partie de l'échantillon du fichier de données d'enquête complet est effectué en écrivant le code pour une étape DATA d'un programme SAS avant d'utiliser une macro BootVar. Cela produit un plus petit fichier de données qui peut ensuite être utilisé par BootVar lorsque un échantillon restreint est requis.
4. **Si le poids de l'enquête, les poids bootstrap et les variables d'analyse ne sont pas dans le même fichier, savez-vous comment fusionner les différentes sources dans le logiciel que vous utilisez? (En supposant que le logiciel utilisé requiert que tous ces renseignements soient dans le même fichier.)**
   BootVar n'exige pas que les poids d'enquête et les poids bootstrap se trouvent dans un fichier de données différent que celui qui contient les variables d'analyse. S'ils sont dans des fichiers différents, les deux fichiers doivent contenir les mêmes variables d'identificateur unique. Le chercheur n'a pas besoin de faire de fusion par lui-même. Le nom du fichier d'analyse et du fichier contenant les poids doit être fourni en énoncés « %let… » avant d'exécuter toute macro BootVar. Les deux énoncés « %let… » pointeront vers le même fichier si les variable d'analyse et les variables de poids sont dans le même fichier.
5. **Tout en effectuant votre analyse, avez-vous vérifié les résultats de sortie de votre logiciel pour déterminer :**
   a. **Que la bonne taille d'échantillon a été utilisée?**
      Dans le log SAS, le nombre d'observations utilisées est affiché.
   b. **Que la bonne variable de poids a été utilisée?**
      Le nom de la variable de poids utilisé n'est pas affiché.
   c. **Que l'ensemble complet des poids bootstrap a été utilisé?**
      La variable rep_mod indique combien de poids bootstrap ont été utilisés.
   d. **Que l'ajustement pour le bootstrap moyen a été effectué de façon appropriée (si nécessaire)?**
      L'ajustement du bootstrap moyen n'est pas affiché.
   e. **S'il existait des échantillons bootstrap pour lesquels il n'était pas possible de faire des estimations?**
      S'il y a des échantillons bootstrap pour lesquels il était impossible d'effectuer des estimations, cela est indiqué dans le log SAS.

Un examen minutieux des valeurs d'entrée dans les variables macro est très important dans le cas de BootVar. Il n'y a pas beaucoup d'information dans les résultats de sortie pour vérifier que les bons paramètres ont été utilisés. Il faut également vérifier le log SAS pour s'assurer que toutes les observations ont été utilisées de la bonne façon.

Programme SAS/BootVar et résultats de l'exemple tiré de l'ESG

Programme

```
/* PARTIE 1 */
libname pumfl '\\SASD6\Sasd-Dssea-Public\DATA\GSS\DLI\CYCLE22\C22MDFSasAndCode-EngFr';
options linesize=60;

/*À partir du fichier de données original, produire un nouveau fichier de données qui
contient les variables requises pour la première partie de l'analyse. Certaines des
nouvelles variables sont créées. Il faut aussi noter que ce fichier contient également
la variable de poids et les variables de poids bootstrap*/

data c22pumf (keep= recid  one PER_q110 dp1 dp2 dp7 dp8 dp9 wght_per  wtbs_001 -
wtbs_500);
 format per_q110 1. ;
 set pumfl.c22pumf;
 one=1; /* Variable avec une valeur de 1 pour toutes les observations */
/* Créer un variable nominale pour chaque valeur différente de PER_Q110 */
 if per_q110 =1 then dp1=1; else dp1=0;
 if per_q110 =2 then dp2=1; else dp2=0;
 if per_q110 =7 then dp7=1; else dp7=0;
 if per_q110 =8 then dp8=1; else dp8=0;
 if per_q110 =9 then dp9=1; else dp9=0;
run;

/* PARTIE 2 */
/* Estimation des proportions de la population totale qui ont des valeurs différentes
de la variable Per_Q110, en utilisant BootVar */

%let Mfile = c22pumf;
%let bsamp = c22pumf;
%let classes =.;
%let ident = RECID;
%let fwgt = wght_per;
%let bsw = wtbs_;
%let R = 25;
%let B = 500 ;
%include «F:\DARC\BootVar\MACROE_V32.SAS»;
%ratio(dp1,one);
%ratio(dp2,one);
%ratio(dp7,one);
%ratio(dp8,one);
%ratio(dp9,one);
%output;
```

```
/* PARTIE 3 */
/* À partir du fichier de données original, créer un fichier de données qui contient
l'échantillon de la sous-population pour la régression logistique. Recoder les variables
nécessaires pour faire l'analyse*/

data c22pumf_sub (keep= recid age2034 age3554 sexM Urban Vote one PER_q110 wght_per
wtbs_001 - wtbs_500);
 set pumfl.c22pumf;
 /*Sous-population des électeurs de 20 à 64 ans*/
 if agegr5 ge 03 and agegr5 le 11;
 if Per_Q110 =1 or Per_Q110 =2;
 /*Recoder la variable URBAIN */
 if LUC_RST=1 then Urban=1;
 else Urban=0;
 /*Recoder la variable ÂGE en trois catégories*/
 if agegr5 le 04 then age = 1;
 else if agegr5 le 07 then age =2;
 else if agegr5 le 11 then age =3;
 /*Recoder la variable du VOTE*/
 if Per_Q110 =1 then Vote =1;
 else Vote =0;
 one=1;
 /*Créer des variables nominales pour l'âge, nécessaires pour la régression logistique*/
 if age=1 then age2034 = 1;
      else age2034=0;
 if age=2 then age3554 = 1;
      else age3554=0;
 if sex=1 then sexM=1;
      else sexM=0;
run;

/* PARTIE 4*/
/* Estimer les proportions avec des valeurs différentes de VOTE et effectuer une
estimation de la régression logistique pour la variable VOTE au moyen de BootVar*/
%let Mfile = c22pumf_sub;
%let bsamp = c22pumf_sub;
%let classes =.;
%let ident = RECID;
%let fwgt = wght_per;
%let bsw = wtbs_;
%let R = 25;
%let B = 500 ;
%include «F:\DARC\BootVar\MACROE_V32.SAS»;
%ratio(Vote,one);
%logreg(Vote,sexM age2034 age3554 Urban);
%output;
```

Commentaires au sujet du programme et des résultats

PARTIE 1

Cette partie du programme est du code SAS, qui consiste en une étape DATA où un ensemble de données nommé c22pumf est créé à partir d'un ensemble de données SAS du FMGD. Le fichier c22pumf a un nombre réduit de variables, qui sont nécessaires pour faire l'estimation des proportions de la population qui a donné des réponses différentes à la question de l'action de voter lors de la dernière élection fédérale. Certaines variables requises pour la macro BootVar du ratio sont créées lors de cette étape DATA : une variable « one » qui prend la valeur de 1 pour toutes les observations, et des variables 0/1 DP1, DP2, etc.; une pour chaque valeur possible de la variable PER_Q110. Afin de créer ces variables 0/1, le chercheur doit connaître les différentes valeurs possibles que PER_Q110 peut prendre, ce qui pourrait être déterminé au moyen d'un dictionnaire de données ou à l'aide de PROC FREQ. Aussi, ce qui est retenu dans le C22pumf est le RECID, qui est un identificateur d'enregistrement unique, la variable de poids de l'enquête wght_per, et toutes les variables de poids bootstrap.

Cette partie du programme ne produit pas de résultats de sortie. Toutefois, après avoir exécuté l'étape DATA, le log SAS (pas présenté) indique que l'ensemble de données c22pumf bootstrap_wght contient 20 401 enregistrements.

PARTIE 2

Cette partie du programme permet d'obtenir les estimations des proportions de la population complète qui a donné des différents types de réponses à la question de voter lors de la dernière élection fédérale. Cela permet de faire une inspection préliminaire de la variable Per_Q110.

Chaque proportion estimée exige un appel distinct à la macro BootVar pour le ratio. Toutefois, si l'énoncé %output est utilisé après les cinq appels à %ratio, les résultats sont affichés dans un seul tableau, comme on peut le voir ci-dessous.

Afin d'estimer la proportion pour une réponse spécifique, une variable 0/1 doit être créée pour prendre la valeur 1 pour les enregistrements ayant la réponse souhaitée dans l'échantillon et la valeur 0 pour tous les autres enregistrements. Cette variable 0/1 est ensuite utilisée au numérateur lorsqu'on décide d'utiliser %ratio. En ce qui a trait à notre exemple, les variables 0/1 pour les réponses différentes ont les noms DP1, DP2, DP7, DP8 et DP9, afin de correspondre aux valeurs de réponses possibles de 1,2,7,8 et 9 pour Per_Q110.

Puisque les poids bootstrap sont dans le même fichier que les variables d'analyses, Mfile et bsample sont assignés aux mêmes valeurs. Si les poids bootstrap étaient dans un fichier différent, les valeurs Mfile et bsample auraient indiqué le nom des différents fichiers. (Dans le cas de plusieurs enquêtes de Statistique Canada, les poids bootstrap sont fournis dans différents fichiers que ceux contenant les variables d'analyse. Toutefois, pour le cycle 22 de l'ESG, le fichier FMGD contient les variables d'analyse et les poids bootstrap.)

Figure 34 SAS BOOTVAR estimation de variance d'un ratio

Variance Estimation for a RATIO
using 500 bootstrap replicates

Numerator	Denominator	Numerator size	Ratio	Standard error	Coeff. of variation	Lower limit confidence interval 95%	Upper limit confidence interval 95%
dp1	one	14941	0.6898	0.0037	0.53	0.6826	0.6970
dp2	one	4660	0.2509	0.0038	1.51	0.2435	0.2584
dp7	one	666	0.0527	0.0014	2.60	0.0501	0.0554
dp8	one	45	0.0020	0.0004	18.63	0.0013	0.0027
dp9	one	89	0.0045	0.0006	14.35	0.0032	0.0058

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

PARTIE 3

Cette partie du programme est une étape DATA en SAS. Dans cette étape, des observations de l'échantillon dans la sous-population d'intérêt sont choisies (c'est-à-dire, observations de l'échantillon pour les personnes de 20 à 64 ans et qui ont répondu « oui » ou « non » à la question sur l'action d'avoir voté lors de la dernière élection fédérale). Ensuite, certaines des variables sont recodées pour produire des variables nominales 0/1 que le chercheur souhaite utiliser comme variable dépendante et indépendante pour la régression logistique, étant donné que BootVar n'est pas en mesure de le faire.

À la suite de l'exécution de l'étape DATA, le log SAS (pas présenté) indique que le nouvel ensemble de données c22pumf_sub contient 14 813 enregistrements. Cette étape DATA ne produit aucun résultat.

PARTIE 4

On retrouve, ci-dessous, les résultats de sortie de la macro de régression logistique. On peut voir dans le programme que le chercheur a utilisé le nouvel ensemble de données c22pumf_sub comme source pour toutes ses variables d'analyse, ses poids et ses poids bootstrap. Les résultats de l'ajustement du modèle sont présentés sous la forme de coefficients estimés (dans la colonne de coefficient de régression partielle réduit du tableau) et de rapports de cotes (dans la colonne de rapport de cotes du tableau).

L'erreur-type de chaque coefficient estimé est fournie, ainsi qu'une statistique de Wald pour vérifier si chaque coefficient du modèle est zéro, étant donné que toutes les autres variables sont dans le modèle. La valeur p pour chaque test de Wald est également présenté. Des intervalles de confiance de 95 % pour chaque rapport de cotes sont également fournis.

Comme mentionné dans la liste de vérification du logiciel, BootVar ne peut effectuer un test pour déterminer si les deux coefficients d'âge sous-jacents dans le modèle logistique sont simultanément zéro, à la suite de l'inclusion de la variable du sexe et de l'emplacement (rural/urbain) dans le modèle.

Figure 35 SAS BOOTVAR estimation de variance d'une régression logistique

10:32 Friday, November 26, 2010 16

Variance Estimation for a LOGISTIC REGRESSION
using 500 bootstrap replicates

---------- Model=1: Dependent Variable = Vote ----------

Independent variables	Beta	Odds ratio	Standard error	Wald	p value	Odds ratio lower limit conf. int. 95%	Odds ratio upper limit conf. int. 95%
Intercept	1.5763	4.84	0.0630	626.32	0.0000	4.2754	5.4727
sexM	-0.0003	1.00	0.0484	0.00	0.9955	0.9092	1.0992
age2034	-1.2546	0.29	0.0649	373.66	0.0000	0.2511	0.3239
age3554	-0.7928	0.45	0.0516	235.68	0.0000	0.4090	0.5008
Urban	-0.0273	0.97	0.0566	0.23	0.6295	0.8709	1.0872

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Même si cela ne fait pas partie de l'analyse exigée, un autre appel à la macro ratio a été effectué dans cette partie du programme. Cela a permis d'estimer la proportion de la sous-population qui a répondu soit « oui », soit « non » à la question sur l'action d'avoir voté lors de la dernière élection fédérale. Dans le numérateur de %ratio, on retrouvait la variable VOTE qui avait une valeur de 1 pour les personnes qui avaient répondu « oui » et de 0 pour les autres personnes. L'ensemble de données c22pumf_sub contenait seulement les enregistrements pour les personnes qui avaient répondu « oui » ou « non ». On peut voir le résultat ci-dessous.

Figure 36 SAS BOOTVAR estimation de variance pour VOTE

Variance Estimation for a RATIO
using 500 bootstrap replicates

Numerator	Denominator	Numerator size	Ratio	Standard error	Coeff. of variation	Lower limit confidence interval 95%	Upper limit confidence interval 95%
Vote	one	10961	0.7216	0.0046	0.64	0.7126	0.7306

Veuillez noter que les saisies d'images sont offertes en anglais seulement, parce que le logiciel dans lequel elles sont prises n'est pas disponible en français.

Bibliographie

La liste ci-dessous inclut le matériel qui a été utilisé pour rédiger ce document.

Chowhan, J. et Buckley, N. 2005. « Utilisation de poids « bootstrap » moyens dans Stata : une révision de BSWREG ». *Le Bulletin technique et d'information des Centres de données de recherche*, vol. 2, nº 1, p. 23 à 37, nº 12-002-X au catalogue de Statistique Canada, http://www5.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-002-X20050018031&lang=fra. (consulté le 4 avril 2012).

Phillips, O.. 2004. « Comment utiliser les poids « bootstrap » avec WesVar et SUDAAN ». *Le Bulletin technique et d'information des Centres de données de recherche*, vol. 1 nº. 2, p. 6 à 15. Voir : http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-002-X20040027032&lang=fra. (consulté le 17 juillet 2013)

La liste ci-dessous comprend des livres qui traitent des méthodes d'analyse de données issues d'une enquête avec un plan de sondage complexe.

Chambers, R.L. et Skinner, C.J. (Eds.). 2003. *Analysis of Survey Data*, John Wiley and Sons.

Cochran, W.G. 1977. *Sampling Techniques*, 3e édition, John Wiley and Sons.

Heeringa, S.G., West, B.T., et Berglund, P.A. 2010. *Applied Survey Data Analysis*. Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences.

Korn, E.L. et Graubard, B.I. 1999. *Analysis of Health Surveys*, John Wiley and Sons.

Lumley, Thomas S. 2010. *Complex Surveys: A Guide to Analysis using R*, John Wiley and Sons.

Lohr, Sharon L. 1999. *Sampling: Design and Analysis*, Duxbury Press.

Pfeffermann, Daniel. et C.R. Rao. 2010. *Handbook of Statistics - Sample Surveys: Inference and Analysis*, vol. 29, Elsevier.

Skinner, C.J., D. Holt et T.M.F. Smith. 1989. *Analysis of Complex Surveys*, John Wiley and Sons.

Annexes

Annexe 1 Fonctions svy dans STATA 12, liste exhaustive

Le préfixe *svy* peut être utilisé avec plusieurs fonctions d'estimation dans Stata. Voici la liste des fonctions d'estimation qui fonctionnent avec le préfixe *svy*.

Statistiques descriptives

mean	[R] mean — Estimation des moyennes
proportion	[R] proportion — Estimation des proportions
ratio	[R] ratio — Estimation des ratios
total	[R] total — Estimation des totaux
tabulate oneway	[svy] tabulate oneway – Tableau de fréquence à une dimension
tabulate twoway	[svy] tabulate twoway – Tableau de fréquence à deux dimensions

Modèles de régression linéaire

cnsreg	[R] cnsreg — Régression linéaire contrainte
glm	[R] glm — Modèles linéaires généralisés
intreg	[R] intreg — Régression d'intervalle
nl	[R] nl — Estimation non-linéaire des moindres carrés
regress	[R] regress — Régression linéaire
tobit	[R] tobit — Régression tobit
treatreg	[R] treatreg — Modèle d'effets de traitement
truncreg	[R] truncreg — Régression tronquée

Modèles de régression de données de survie

stcox	[ST] stcox — Modèle de régression à risque proportionnels de Cox
streg	[ST] streg — Modèles de survie paramétrique

Modèles de régression de réponse binaire

biprobit	[R] biprobit — Régression probit bivariée
cloglog	[R] cloglog — Régression log-log complémentaire
hetprob	[R] hetprob — Modèle probit hétéroscédastique
logistic	[R] logistic — Régression logistique - rapports de cotes
logit	[R] logit — Régression logistique - coefficients
probit	[R] probit — Régression probit
scobit	[R] scobit — Régression logistique asymétrique

Modèles de régression de réponse discrète

clogit	[R] clogit — Régression logistique conditionnelle (effets-fixes)
mlogit	[R] mlogit — Régression logistique multinominale (polytomique)
mprobit	[R] mprobit — Régression probit multinominale
ologit	[R] ologit — Régression logistique ordonnée
oprobit	[R] oprobit — Régression probit ordonnée
slogit	[R] slogit — Régression logistique de stéréotype

Modèles de régression de Poisson

gnbreg	Régression binominale négative généralisée dans [R] nbreg
nbreg	[R] nbreg — Régression binominale négative
poisson	[R] poisson — Régression de Poisson
zinb	[R] zinb — Régression binominale négative gonflée à zéros
zip	[R] zip — Régression de Poisson gonflée à zéros
ztnb	[R] ztnb — Régression binominale négative tronquée à zéros
ztp	[R] ztp — Régression de Poisson tronquée à zéros

Modèles de régression des variables instrumentales

ivprobit	[R] ivprobit — Modèle probit avec régresseurs endogènes continus
ivregress	[R] ivregress — Régression des variables instrumentales d'une seule équation
ivtobit	[R] ivtobit — Modèle tobit avec régresseurs endogènes continus

Modèles de régression avec sélection

heckman	[R] heckman — Modèle de sélection de Heckman
heckprob	[R] heckprob — Modèle probit avec sélection d'échantillon

Annexe 2 STATA avant la version 12, comment utiliser les poids et les poids bootstrap

Version 10

Dans la version 10 de Stata, il est nécessaire d'effectuer une brève révision de la façon dont les procédures d'enquête sont configurées. À la suite de cette brève révision, toutes les analyses devraient être effectuées de la même manière que celle décrite dans Stata 12 de la partie principale de ce document.

La fonction de configuration de l'enquête (*svyset*) doit comporter les précisions suivantes :

1. *pweight* assigné à la bonne variable de poids d'enquête.
2. Poids bootstrap identifiés au moyen de l'option *brrweight.*
3. Type d'estimation de la variance établi à *brr* avec l'option *vce*.
4. Établissement de l'option de l'erreur quadratique moyenne en utilisant l'option *mse*. (Cette option permet de calculer la variance bootstrap comme la moyenne des écarts quadratiques entre les estimations bootstrap et l'estimation de l'échantillon complet. Si cette option n'est pas choisie, la variance bootstrap est calculée comme la moyenne des écarts quadratiques entre les estimations bootstrap et la moyenne des estimations bootstrap.)
5. Établir l'ajustement du bootstrap moyen, si nécessaire, avec l'option *Fay*. La valeur à entrer comme un ajustement de Fay devrait être fondée sur un calcul en utilisant la formule suivante :

   $1 - C^{-\frac{1}{2}}$, où C est le nombre d'échantillons bootstrap utilisés pour produire chaque variable de poids bootstrap moyen.

Version 9

Dans la version 9, le préfixe d'enquête n'est pas disponible, et cela change les procédures pour obtenir des estimations de la variance bootstrap dans Stata de façon radicale. Ces procédures sont illustrées dans *Le Bulletin technique et d'information des Centres de données de recherche*, printemps 2005, vol. 2, nº 1, dans un article intitulé « Utilisation de poids « bootstrap » moyens dans Stata : une révision de BSWREG » rédigé par James Chowhan et Neil Buckley. Cet article peut être consulté sur le site Web de Statistique Canada à http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-002-X20050018031&lang=fra.

Annexe 3 SPSS et utilisation des poids bootstrap

Même si SPSS possède un module additionnel pour traiter les échantillons complexes qui offre plusieurs outils d'analyse de données d'enquête, il ne peut fournir aucune méthodes de répliques pour l'estimation de la variance fondée sur le plan. Par conséquent, il ne peut calculer l'estimation de la variance bootstrap en utilisant les poids bootstrap fournis par plusieurs enquêtes de Statistique Canada.

Dans les versions précédentes de SPSS, il y avait une version SPSS de BootVar développée par des méthodologistes de Statistique Canada qui permettait de calculer les estimations de la variance bootstrap pour une gamme de procédures analytiques. Ce programme n'est plus utilisé et n'est plus mis à jour.

Les personnes qui utilisent SPSS pour effectuer d'autres types d'analyses doivent, par conséquent, utiliser un autre progiciel afin d'utiliser les poids bootstrap. Elles peuvent choisir ce progiciel en se fondant sur leur style d'analyse favori et de leurs problèmes analytiques propres à eux. Par exemple, si un chercheur préfère utiliser des menus déroulants, il devrait envisager d'utiliser WesVar ou Stata. Plusieurs autres progiciels accepteront un fichier de données SPSS en tant que données d'entrée.

Annexe 4 Normaliser les poids

Un chercheur peut avoir entendu parler de « normaliser » ou « mettre à l'échelle » une variable de poids d'enquête. Cela signifie que le poids d'enquête pour chaque membre de l'échantillon dans la sous-population analysée est divisé par la moyenne des poids d'enquête des membres de l'échantillon dans la sous-population, c'est-à-dire que le poids normalisé a la valeur :

$$w_{norm,i} = \frac{w_i}{\{\Sigma_{i \epsilon s_D} w_i\}/n_D},$$

où w_i est le poids d'enquête pour le i^e membre de l'échantillon dans la sous-population d'intérêt, s_D représente l'échantillon dans cette sous-population, et n_D est la taille de cet échantillon. Les poids normalisés qui en découlent auront un valeur moyenne de 1,0 et les poids normalisés pour tous les cas d'échantillon dans la sous-population seront ajoutés à n_D, la taille de l'échantillon dans la sous-population.

Normaliser les poids n'est pas nécessaire lorsque les analystes utilisent un logiciel capable de tenir compte correctement du plan de sondage dans une analyse et lorsque l'information contenue dans un plan de sondage convenable est disponible pour l'analyste. Ainsi, si un analyste a un poids d'enquête et des poids bootstrap correspondants, et s'il possède un logiciel analytique approprié pour des données d'enquête, il n'y a aucune raison de procéder à une normalisation des poids.

Alors, pourquoi un analyste n'utiliserait pas un logiciel adapté pour tenir compte correctement du plan de sondage? Deux situations nous viennent à l'esprit :

1. L'analyste possède des poids d'enquête, mais ne possède pas de poids bootstrap ou d'autres renseignements appropriés sur le plan de sondage pour estimer les variances dans un logiciel analytique pour des données d'enquête. Par conséquent, le logiciel analytique pour des données d'enquête ne peut être utilisé de la bonne façon. Une telle situation pourrait se produire, par exemple, avec une version à usage public d'un fichier de données où, pour des raisons de confidentialité, seuls des poids d'enquête peuvent être fournis.
2. L'analyste a en main des poids d'enquête et les poids bootstrap correspondants, mais souhaiterait effectuer certaines analyses préliminaires en utilisant un autre logiciel qu'un logiciel analytique pour des données d'enquête et souhaiterait aussi utiliser les poids d'enquête. Cela permettrait d'économiser du temps ou d'utiliser certaines caractéristiques spéciales de l'autre logiciel, telles que certaines fonctions permettant de produire des graphiques diagnostiques.

Toutefois, normaliser les poids et utiliser ceux-ci dans un logiciel qui n'est pas conçu précisément pour effectuer des analyses de données d'enquête constitue une méthode provisoire. Même si les estimations ponctuelles peuvent être correctes, cette méthode pourrait vous amener à émettre des conclusions incorrectes, étant donné que nous ne savons pas si les mesures de variabilité estimée seront plus grandes ou plus faibles que s'ils avaient été obtenus par une méthode fondée sur le plan de sondage complet. La taille des estimations de la variabilité a une incidence, par exemple, sur l'ampleur du test statistique et des valeurs p, et aussi sur l'étendue des intervalles de confiance. Même si cela ne s'avère pas vrai dans tous les cas, la tendance générale est de sous-estimer les mesures de la variabilité lorsqu'on choisit de ne pas utiliser une méthode fondée sur un plan de sondage complet, ce qui signifie que les valeurs p tendent à être trop faibles et les intervalles de confiance trop étroits.

Si l'analyste n'a d'autres choix que d'utiliser un logiciel de rechange, il doit examiner la façon dont ce logiciel utilise une variable de poids et si les résultats seraient « plus concluants » si la variable de poids avait été normalisée. Voici deux façons communes qu'un logiciel non conçu pour analyser des données d'enquête devrait traiter un poids :

1. en étant un compteur de fréquence qui indique combien d'observations ont exactement les mêmes valeurs pour toutes les variables;
2. en étant l'inverse de la variance de l'observation.

Le fait d'utiliser la variable de poids telle que fournie, ou à la suite de sa normalisation, produira généralement les mêmes estimations ponctuelles, sauf pour l'estimation d'un total. Ces estimations ponctuelles seront les mêmes que les estimations ponctuelles qui seraient obtenues à l'aide d'un logiciel analytique conçu pour analyser des données d'enquête. Toutefois, là ou il pourrait y avoir une nette différence entre le fait d'utiliser un poids normalisé et un poids non normalisé dans un logiciel de rechange serait au niveau des estimations de la variabilité et des tests statistiques fondés sur des estimations de la variabilité. Souvent, un poids normalisé produira une variabilité et des résultats de test dans un logiciel de rechange qui ont des valeurs plus semblables à la variabilité et aux résultats du test obtenus au moyen d'un logiciel analytique pour des données d'enquête que si un poids non normalisé est utilisé dans le logiciel de rechange. Par contre, pour que cela se produise, il est important que la normalisation soit effectuée pour les unités d'échantillonnage dans la sous-population étudiée dans une analyse précise, et non dans l'échantillon complet d'une enquête.

Dans Stata, plusieurs fonctions permettent d'utiliser un poids et elles ne sont pas des fonctions *svy*. Si un *pweight* peut être précisé dans une de ces fonctions, alors la normalisation de la variable de poids n'aura aucune incidence sur les résultats.

Dans SAS, plusieurs procédures autres que les procédures d'enquête permettent d'utiliser un énoncé WEIGHT. La façon dont la variable de poids est utilisée dans les calculs varie d'une procédure à l'autre et certaines de ces procédures offrent une option NORM. Il est conseillé de normaliser la variable de poids dans l'échantillon (sous-échantillon) utilisé aux fins de l'analyse, peu importe la méthode requise pour une procédure particulière utilisée.

Dans SPSS, plusieurs procédures permettent l'attribution d'une variable de poids. Dans la plupart des cas, les poids sont considérés comme étant des poids de fréquence. Par contre, il est conseillé de normaliser la variable de poids dans l'échantillon (sous-échantillon) utilisé aux fins de l'analyse.

Directives pour les auteurs

Les articles portant sur les questions méthodologiques et les sujets techniques reliés aux ensembles de données qui se trouvent dans les CDR sont appropriés pour le Bulletin technique et d'information.

Langage du matériel soumis

Les manuscrits peuvent être soumis en français ou en anglais. Une fois acceptés, les manuscrits seront traduits dans la deuxième langue officielle avant de les publier.

Longueur d'une soumission

Les articles ne doivent pas dépasser 20 pages à double interligne, en excluant les programmes et les annexes. En plus des explications en profondeur et des questions techniques, le Bulletin accepte également les notes et les commentaires brefs (idéalement, trois pages ou moins) traitant de solutions rapide aux problèmes analytiques et commentaires soulevées antérieurement dans le Bulletin ou part les chercheurs collègues.

Le format électronique et la mise en page des manuscrits

Les manuscrits doivent être en format « Microsoft Word (.doc) ». Les auteurs peuvent les soumettre par courrier ordinaire sur disquette ou disque compact. Ils peuvent également les envoyer comme attachement à un courriel.

Les noms des auteurs, le nom de l'établissement principal, et les coordonnées (numéro de téléphone, adresse postale et adresse électronique) du chercheur principal doivent paraître à la page couverture du manuscrit.

Les auteurs doivent se servir de la police Times New Roman de 12 points, interligne double, et des marges de 1 pouce (2,5 cm) en rédigeant leurs manuscrits.

Nous mettons la majuscule qu'au premier mot du titre (p. ex. Pour une utilisation plus conviviale de la méthode bootstrap…).

Nous nous servons des caractères gras que pour les entêtes. Il ne faut pas souligner les mots ou les phrases ni se servir des caractères en italiques pour les entêtes.

Les notes en bas de page et la bibliographie doivent être à simple interligne. Les auteurs sont invités de consulter *Le Guide du rédacteur*, 2e édition.

Le format et mise en page des graphiques et tableaux

Les tableaux et graphiques doivent être soumis en format « Microsoft Excel (.xls) » ou en format séparé par virgule (.csv). Le nom des fichiers doit indiquer le contenu (p. ex. tableau1, graphique6, etc.).

Les auteurs peuvent les soumettre par courrier ordinaire sur disquette ou disque compact. Ils peuvent également les envoyer comme attachement à un courriel. Indiquez dans le texte l'emplacement des tableaux et graphiques plutôt que de les placer pas dans le texte. Servez vous du titre suivi par le nom du fichier entre parenthèses, p. ex :

Graphique 6. La consommation du chocolat par les enfants au Canada, 2000 (graphique6)

Les expressions mathématiques

Toutes les expressions mathématiques doivent être dissociées du texte. Les équations doivent être numérotées, le numéro devant figurer à la droite de l'équation, aligné à la marge.

Guide de rédaction à l'intention des auteurs

Les auteurs sont priés de se servir de *Le Guide du rédacteur*, 2e édition. Vous pouvez en acheter une copie des Publications du gouvernement du Canada, Travaux publics et Services gouvernementaux Canada.

Où soumettre les manuscrits

Envoyez les manuscrits et toutes communications reliées au Bulletin au Comité de révision.

- Adresse électronique — DAM-SSOBC@statcan.gc.ca

Révision des soumissions

Le processus de révision initiale des articles relève du Comité de rédaction. Les rédacteurs peuvent inviter des auteurs ayant déjà publié des articles dans le Bulletin ou des spécialistes à participer au processus. Les articles soumis au Bulletin font l'objet d'une révision permettant d'en assurer l'exactitude, la cohérence et la qualité.

Au terme du processus de révision initiale par le Comité de rédaction, les articles sont soumis à un examen par les pairs et à un examen interne. L'examen par les pairs sera effectué conformément à la *Politique concernant l'évaluation des produits d'information* de Statistique Canada. En outre, des cadres supérieurs de Statistique Canada procéderont à des examens internes pour s'assurer que le matériel respecte les directives et les normes de l'organisme et qu'il ne porte pas atteinte à la réputation d'impartialité politique, d'objectivité et de neutralité de Statistique Canada.

Veuillez communiquer avec le Comité de révision à l'adresse ci-haut pour des obtenir de plus amples renseignements.