Enquête sur le milieu de travail et les employés - Estimation

Le poids déterminé par le plan d'échantillonnage de chaque unité dépend de la sélection de l'échantillon initial. Durant tout le processus d'enquête, les pondérations de plan initiales peuvent subir plusieurs ajustements, qui visent à maintenir la représentativité de l'échantillon. Pour l'EMTE, on effectue deux ajustements : un afin de compenser les non-réponses complètes et l'autre afin de réduire l'influence sur les estimations des sauteuses de strate (de grandes unités qu'on croit petites et vice-versa). Pour la correction de la non-réponse, on multiplie les poids initiaux déterminés par le plan d'échantillonnage des unités déclarantes par un ratio de l'ensemble des unités échantillonnées sur l'ensemble des unités déclarantes dans chaque strate. Cette méthode est fondée sur l'hypothèse selon laquelle les répondants et les non répondants se comportent de la même façon. Cette hypothèse n'est pas déraisonnable, puisque les non-réponses sont principalement le fait des petites unités.

La correction pour les « sauteuses » de strate est plus complexe car il y a au moins trois méthodes pour traiter ce problème en général. On peut réduire la pondération de plan de la sauteuse de strate et répartir la différence entre les unités restantes à l'intérieur de la strate, ou on peut réduire ses valeurs, ou encore supprimer entièrement l'unité et la traiter comme un cas de non-réponse. Nous avons choisi la première méthode et avons ciblé environ 30 employeurs en vue d'un ajustement des pondérations de plan.

L'utilisation des poids déterminés par le plan d'échantillonnage, qu'il s'agisse des poids initiaux ou des poids corrigés, donne des estimations non déformées mais parfois inefficientes. Pour accroître l'efficacité du processus d'estimation, on peut étalonner, ou calibrer, l'échantillon suivant un ensemble de chiffres (totaux) pour une population connue ou estimée efficacement. Dans le cadre de l'EMTE, cela se fait à l'aide de l'emploi total estimé grâce à l'EERH au niveau de l'industrie par région, où les estimations découlant de l'Enquête sur le milieu de travail et les employés doivent correspondre aux estimations découlant de l'EERH. Les facteurs de correction ainsi obtenus sont appliqués aux poids corrigés déterminés par le plan d'échantillonnage. L'établissement de repères est le plus avantageux dans les situations où la variable de calibration (l'emploi dans l'EMTE) est en étroite corrélation avec les variables d'intérêt.

Le produit du poids corrigé déterminé par le plan d'échantillonnage et du facteur de calibration est le poids final de l'emplacement. On obtient le poids relié final en corrigeant le poids de l'emplacement pour tenir compte des employeurs actifs sans employés répondants avant d'appliquer le facteur de calibration. Le poids final des employés reflète la sélection des employés et la non-réponse supplémentaire des employés. Ces poids finals servent à calculer des statistiques comme les totaux, les moyennes, les coefficients de régression, etc. Pour estimer la variance de ces statistiques, il faut utiliser des progiciels qui

permettent à l'utilisateur de préciser le plan d'échantillonnage. Si on utilise des produits comme SAS sans transformer de façon appropriée les pondérations de l'enquête, la sous-estimation en découlant de la variance peut être assez marquée.

Estimation de la variance

Les analystes qui désirent produire de bonnes estimations de la variance disposent de plusieurs moyens. L'une d'elles est le recours au Système généralisé d'estimation (SGE) de Statistique Canada qui se chargera de l'estimation des totaux, des moyennes et des ratios pour différents plans. L'utilisation du SGE par des chercheurs externes pourra s'avérer dispendieuse en raison des coûts élevés liés à l'obtention d'une licence.

La deuxième possibilité est de loin la plus générale et la plus facile à mettre à exécution. Elle englobe l'utilisation de pondérations bootstrap. La méthode bootstrap est une technique statistique au moyen de laquelle on suit une procédure de ré échantillonnage pour produire un certain nombre d'ensembles de pondérations qui, si on les utilise correctement, saisissent la variabilité de bien des statistiques. L'idée consiste à calculer un grand nombre d'estimations bootstrap, puis leur variance.

Une fois les pondérations bootstrap calculées, on peut les spécifier dans l'énoncé des pondérations à l'intérieur de toute procédure du SAS qui en comporte un. Pour estimer la variance d'une statistique, il faut produire une estimation fondée sur chaque ensemble de poids bootstrap. Ensuite, on utilise la variabilité entre ces estimations bootstrap pour produire une bonne estimation de la variance de la statistique désirée.

L'utilisation de pondérations *bootstrap* pour le calcul de variances conformes au plan

Lorsqu'on calcule les variances d'estimations fondées sur des échantillons provenant de populations finies, il faut tenir compte du plan d'échantillonnage, ce qui ne se fait pas facilement à l'intérieur de la plupart des progiciels d'analyse statistique. S'ils permettent l'utilisation de poids, la plupart ne les utilisent pas de la bonne manière, ce qui amène souvent une sous-estimation de la variance. Cela pourrait avoir des conséquences désastreuses sur les vérifications d'hypothèse et sur la construction d'intervalles de confiance.

Au fil des ans, les bureaux de la statistique ont mis au point des systèmes pour traiter les populations finies; la plupart de ces systèmes n'ont cependant pas la flexibilité nécessaire pour effectuer l'analyse des données. C'est là où intervient la technique *BOOTSTRAP*, qui est fondée sur le ré échantillonnage. Cette technique consiste à utiliser l'échantillon initial, à partir duquel on sélectionne un échantillon aléatoire simple, et on remplace autant d'unités qu'on en avait au

départ. On répète cette procédure bien des fois pour garantir la convergence. Cela donne plusieurs ensembles de poids bootstrap. Dans l'EMTE, on utilise la méthodologie des poids bootstrap moyens, où chaque ensemble de poids bootstrap est, de fait, obtenu comme moyenne de nombreux ensembles de poids bootstrap (50 dans l'EMTE).

Une fois les pondérations *bootstrap* calculées, on peut les préciser à l'intérieur de l'énoncé des pondérations dans le cadre de toute procédure SAS qui en comporte un. Pour estimer la variance d'une statistique désirée, il faut produire une estimation fondée sur chaque ensemble de poids bootstrap. Ensuite, on calcule la variabilité entre ces estimations bootstrap pour produire une estimation appropriée de la variance de la statistique désirée. Ci-dessous figurent deux exemples de la façon dont on peut procéder pour des totaux et pour des coefficients de corrélation.

Selon votre analyse, vous utiliseriez wkp_bsw1-wkp_bsw100 (poids bootstrap de l'emplacement), emp_bsw1-emp_bsw100 (poids bootstrap de l'employé) ou lnk_bsw1-lnk_bsw100 (poids bootstrap liés). Les utilisateurs de SPSS utiliseront wkp_b1-wkp_b100, emp_b1-emp_b100, lnk_b1_lnk_b100. L'exemple suivant vise l'information sur le milieu de travail.

```
PROC SUMMARY DATA = WES NWAY;
CLASS DOM_IND;
VAR WKP_FINAL_WT WKP_BSW1-WKP_BSW100;
WEIGHT TTL_EMP;
OUTPUT OUT = ESTIM (DROP = _FREQ_ _TYPE_)
SUM = EMPL WKP_BSW1-WKP_BSW100;
RUN;
PROC TRANSPOSE DATA = ESTIM
OUT = T_ESTIM (DROP = _NAME_ RENAME = (COL1 = ESTIM));
VAR WKP_BSW1-WKP_BSW100;
BY DOM_IND;
RUN;
PROC SUMMARY DATA = T_ESTIM NWAY;
CLASS DOM_IND;

VAR ESTIM;
OUTPUT OUT = VAR (DROP = _FREQ_ _TYPE_)
CSS = VAR;
RUN;
DATA ESTIM;
MERGE ESTIM (KEEP = DOM_IND EMPL)
VAR;
BY DOM_IND;
CV = ROUND (SQRT(50 / 100 * VAR) / EMPL, 0.01);
RUN;
```

La première procédure SOMMAIRE fait appel à un truc qui permet de calculer toutes les estimations nécessaires en une seule étape simple, ce qui peut se faire lorsqu'on produit des estimations pour une variable unique. Le truc en question consiste à préciser que les pondérations *bootstrap* sont les variables

d’analyse et à utiliser la variable d’analyse comme s’il s’agissait de la pondération. On calcule les estimations au niveau de l’industrie du domaine précisé par l’énoncé des classes.

Une fois les estimations calculées, transposées et renommées, on utilise une autre procédure SOMMAIRE pour calculer leur variance. En fait, leur somme corrigée des carrés, CSS dans le SAS. Finalement, la multiplication de CSS par 50/100 produit la bonne variance. Le dénominateur (100) est l’ajustement normal *n* qui donne la variance classique. Le numérateur (50) reflète le fait que la moyenne de chaque ensemble de pondération *bootstrap* a été calculée à partir de 50 itérations, ce qui fournit une pondération *bootstrap* moyenne. La correction permet donc de réinjecter la variabilité que l’on avait perdue en utilisant la moyenne.

L’exemple qui suit illustre l’utilisation de pondérations *bootstrap* pour le calcul de coefficients de corrélation. On doit ici utiliser une macro pour calculer chacun des coefficients, étant donné qu’il est impossible d’employer facilement le truc fourni ci-dessus.

```
%MACRO COR_COEF;
            %DO I = 1 %TO 100;
                  PROC CORR DATA = BOOT OUTP = CORRS NOPRINT;
VAR TTL_EMP CBA_EMP;
BY DOM_IND;
WEIGHT WKP_BSW&I;
                  RUN;
                  DATA CORRS (KEEP = DOM_IND CBA_EMP RENAME = (CBA_EMP
            = CORR));
SET CORRS (WHERE = (_TYPE_ = 'CORR' & _NAME_ = 'TTL_EMP'));
                  RUN;
                  PROC DATASETS FORCE NOLIST;
APPEND BASE = ESTIM DATA = CORRS;
QUIT;
                  RUN;
            %END;
      %MEND;
      %COR_COEF;
      PROC SUMMARY DATA = ESTIM NWAY;
CLASS DOM_IND;
AR CORR;
VOUTPUT OUT = VAR (DROP = _FREQ_ _TYPE_)
CSS = VAR;
      RUN;
      PROC CORR DATA = BOOT OUTP = CORRS NOPRINT;
VAR TTL_EMP CBA_EMP;
BY DOM_IND;
WEIGHT WKP_FINAL_WT;
      RUN;
      DATA CORRS (KEEP = DOM_IND CBA_EMP RENAME = (CBA_EMP =
EST_CORR));

SET CORRS (WHERE = (_TYPE_ = 'CORR' & _NAME_ = 'TTL_EMP'));
      RUN;
```

```
      DATA ESTIM;
MERGE VAR CORRS;
BY DOM_IND;
CV = ROUND(SQRT(50 / 100 * VAR) / EST_CORR * 100, 0.01);
      RUN;
```

La macro COR_COEF calcule des coefficients de corrélation fondés sur chaque ensemble de pondérations *bootstrap*. L'exemple donné ici englobe deux variables continues, mais peut être facilement étendu à des variables multiples à la fois continues et catégoriques. Une fois les estimations calculées, on produit la somme corrigée des carrés, ainsi qu'un coefficient de corrélation qui est fondé sur les pondérations finales.

On fusionne ensuite les deux fichiers, on ajuste la somme corrigée des carrés et on calcule un coefficient de variation. On devrait suivre des étapes similaires pour calculer des variances d'estimations de régression, des composantes principales et d'autres statistiques. Les totaux d'une variable simple mis à part, on ne peut effectuer les calculs en une seule étape. On recommande, pour écourter le temps de calcul par itération, de réduire l'ensemble de données de départ aux variables d'analyse.

Sont inclus dans \CODE d'autres codes élaborés en STATA et en SAS qui indiquent comment utiliser la pondération bootstrap EMTE pour exécuter une grande diversité d'analyses statistiques. Cet ensemble de codes repose sur des travaux antérieurs de François Brisebois (macros SPSS et SAS pour l'ENSP), de Pierre Felx (macros SAS pour l'EMTE), de Tony Fang (macros STATA pour l'EMTE) et de Dominic Grenier (macros STATA et SAS pour l'ELIC). Ces macros ne sont pas là pour l'estimation de moyennes, de totaux ni de rapports; ces programmes visent principalement à illustrer l'application de la pondération bootstrap EMTE à la modélisation statistique. Les codes permettent d'effectuer les types suivants d'analyses :

- régression linéaire;
- test T;
- analyse de variance;
- analyse de covariance;
- régression logistique;
- modélisation par probits;
- régression logistique multinomiale;
- modélisation logistique (par logits) ordinale;
- modélisation ordinale par probits;
- équation d'estimation généralisée (EEG);
- modélisation linéaire généralisée (toute la famille de modèles);
- tests d'ajustement, d'homogénéité et d'association avec les corrections de Rao-Scott de premier et de second ordre

Les programmes visés sont souples, aisément reproductibles, faciles à utiliser et généralisables à toute enquête pour laquelle il existe une pondération bootstrap.

Souplesse :
Les programmes ne se présentent pas sous forme de fichiers ADO en STATA ni de macros en SAS à sauvegarder dans une bibliothèque de macros. L'utilisateur expérimenté et les gens qui ont moins l'expérience de STATA et de SAS peuvent, en y travaillant un peu, adapter ces codes aux problèmes particuliers qu'ils ont à résoudre. Ils peuvent facilement les étendre ou les réduire. L'utilisateur moins expérimenté pourrait désirer les employer tels quels, c'est-à-dire dans leur formulation actuelle.

Facilité de reproduction :
Une même structure de programmation revient dans tous les programmes. Cette structure est souvent extensible ou reproductible dans d'autres modèles statistiques pour lesquels il n'y a pas de codes bootstrap explicites.

Facilité d'utilisation :
D'abord, l'utilisateur prépare un ensemble de données avec les variables d'intérêt pour les modèles à mettre en ajustement. L'ensemble peut être augmenté de la pondération bootstrap et, selon la nature de l'analyse, la pondération finale d'enquête (auprès des employés, liés ou des employeurs) sera aussi prise en compte.

Ensuite, à la ligne de commande du modèle STATA, l'utilisateur spécifie le nom de ses propres variables et la pondération finale à utiliser, comme dans les exemples cités. Dans de tels programmes, il emploie la désignation de base de la variable de pondération bootstrap, qui peut être emp_bsw pour la partie « employés » ou wkp_bsw pour la partie « milieux de travail » aux fins de l'analyse.

Dans les macros on SAS, l'utilisateur doit spécifier en début de programme la pondération finale d'enquête, le nombre de valeurs de pondération bootstrap et d'itérations, l'ensemble de données à utiliser et la désignation de base des poids bootstrap.

<u>Voici ce qui est spécifié au départ :</u>

```
%let bsw = emp_bsw;/* dans le fichier « employés », utiliser emp_bsw et, dans le fichier « employeurs », remplacer cette variable par wkp_bsw*/.

%let fwgt = emp_final_wt; /* utiliser le nom de la variable en pondération finale, c'est-à-dire emp_final_wt pour le fichier « employés » et wkp_final_wt pour le fichier « employeurs »*/

%let dsn=boot_data;/* cet ensemble de données comporte un sous-ensemble de variables d'intérêt pour l'analyse et la pondération bootstrap*/
%let b=100;/* le nombre de valeurs de pondération bootstrap disponibles dans le fichier*/
        %let iter=50;
```

Voici ce qui est spécifié à la fin :

```
%linregress(boot_data,hr_waget,age) /* le nombre d'éléments
à cette ligne dépend des modèles et des autres paramètres
nécessaires pour les macros. À cette ligne s'exécute une
analyse de régression avec le salaire horaire comme
fonction de l'âge des employés selon l'ensemble de données
boot_data.
```

Enfin, on sauvegarde les résultats dans un répertoire fourni par l'utilisateur en remplaçant le chemin « c:\Documents and Settings\decaywe\bootstrap_ yves\res.dta » par un chemin propre.

Généralisabilité :
Ces fichiers de programme peuvent servir dans toute enquête qui fournit une pondération bootstrap. Ce qui distingue l'EMTE est que, dans le calcul de la variance, on tient compte de ce que chaque valeur de pondération bootstrap représente une moyenne de 50 itérations. Dans le programme STATA, on applique à cet égard l'instruction « local iter = 50 ». Dans d'autres enquêtes, l'utilisateur a seulement à remplacer 50 par 1 à cette ligne. Ajoutons que, si une enquête fournit 1 000 valeurs de pondération bootstrap, on remplace seulement 100 par 1 000 à la ligne de commande « local bs = 100 ». C'est tout ce dont on a besoin.

On peut faire de même avec les macros en SAS en remplaçant

« %let iter= 50" par "%let iter = 1 ».

S'il y a 1 000 valeurs de pondération bootstrap, on remplace

« %let b =100" par "%let b =1000 ».

On peut recourir à des progiciels sur le marché comme SUDAAN et WesVar pour effectuer une estimation de variance bootstrap si le mode d'estimation est spécifié comme BRR et que les variables de pondération bootstrap le sont comme poids BRR (D. Binder et G. Roberts, 2004, dans « Statistical inference in survey data analysis: Where does the sample design fit in? »). Avec les valeurs de pondération EMTE, les résultats produits par ces progiciels devraient être mis en ajustement, car on aura pris la moyenne de chaque ensemble de valeurs sur 50 itérations, ce qui aura donné une pondération moyenne bootstrap. Les codes que nous présentons tiennent compte de ces itérations et les rendent donc spécifiques à l'EMTE. Il reste que, en fixant le nombre d'itérations à l'unité, il y a

généralisation à l’ensemble des enquêtes qui fournissent une pondération bootstrap à l’utilisateur.