Comparaison de méthodes
pour la correction de la non-réponse totale par repondération





Insee Pays de la Loire

Service Etudes Diffusion

PISE - Pôle Ingénierie Statistique Entreprises

Nicolas Sigler
Méthodologue

INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Plan de l'intervention

Contexte, objectif poursuivi

Description du protocole

- Étapes de la simulation
- Correction par repondération :
  - en recourant à la **méthode des scores**
  - en recourant à la **segmentation par arbres**

Comparaison des résultats obtenus

- Indicateurs utilisés
- Résultats

Conclusion…

Contexte, objectif poursuivi

Contexte :

- Mémoire de FCDA (Formation continue diplômante des attachés) d'Émilie Dequidt
- Réalisé au PISE (Pôle Ingénierie Statistique Entreprises)
- Support : l'enquête thématique TIC 2011

  sur les technologies de l'information et de la communication et le commerce électronique

Objectif poursuivi :

Comparer 2 méthodes de constitution des GRH

(classes de repondération utilisées pour corriger la non-réponse totale)

- la méthode des scores (après régression logistique)
- la segmentation par arbres

Objectif **pratique**

Contexte, objectif poursuivi

L'enquête TIC 2011

- enquête européenne annuelle auprès des entreprises
- base de sondage : répertoire Sirène ;
- plan de sondage : stratifié secteur x taille ;
- des variables cibles qualitatives :
  - accès à internet, existence d'un site web, facturation électronique…

  et quantitatives :

  nombre d'utilisateurs d'internet, montant du CA généré par le web…

Méthode de correction de la NRT (# 20 %) utilisée au Pise :

- Repondération à l'intérieur de « groupes de réponse homogène »

  définis en modélisant la probabilité de réponse
- Méthode : PROC LOGISTIC de SAS
- Variables utilisées :

  comportement de réponse en N–1, localisation, secteur, taille…

Description du protocole : Étapes

Reconstitution d'une population fictive

- en partant des répondantes du fichier Tic 2011 redressé
- en dupliquant les répondantes selon leur poids de calage

Échantillonnage

On tire 1000 échantillons (PROC SURVEYSELECT de SAS)

Génération de la non-réponse selon 21 scenarii :

- 7 mécanismes de réponse
- 3 taux de réponse : 70 %, 80 % et 90 %
  - utilisation de la PROC SURVEYSELECT

Description du protocole : les mécanismes testés

7 mécanismes de réponse :

- aléatoire simple sans remise
- variable *cachée* : ZAU (zonage en aires urbaines)
- variable *cachée* : taux d'endettement de l'entreprise
- GRH déterminés dans le redressement réel de TIC 2011
- GRH x taux d'endettement
- secteur d'activité x taille
- secteur d'activité x taille x taux d'endettement

Description du protocole

1e méthode testée : méthode des scores

Modélisation de la probabilité de réponse
avec la PROC LOGISTIC de SAS

Variables auxiliaires utilisées :

- comportement de réponse à l'enquête précédente
- localisation géographique (DOM/Paris/Province)
- secteur d'activité (23 postes)
- taille de l'entreprise (5 tranches)
- chiffre d'affaires (5 tranches)
- appartenance à un groupe ou non

Génération de 25 GRH (au maximum)

par la méthode des quantiles égaux (probas classées par ordre croissant)

-> avantage : facilement automatisable

Description du protocole

2e méthode testée : segmentation par arbres

Segmentation pour expliquer l'indicatrice de réponse :

- méthode : CHAID
- utilisation de la macro TREEDISC implémentée sous SAS

Variables auxiliaires utilisées :

- les mêmes que pour la méthode des scores

Paramétrage :

- seuil de significativité du Chi-2 : 10 %

- critères d'arrêt :

- nb mini d'obs d'un nœud pour le subdivisé : 160
- nb mini d'obs pour constituer une feuille : 80
- nb maxi de niveaux de l'arbre : 20

Comparaison des résultats :
Génération d'estimateurs aux différentes étapes

Estimateurs cibles,
- après génération de la population fictive

Estimateurs « 100 % de réponse »
- sur les 1000 échantillons

Estimateurs sur les répondants,
- après génération de la non-réponse

Estimateurs « score »
- pour les 1000 échantillons et les 21 scenarii de NRT

Estimateurs « segmentation »
- pour les 1000 échantillons et les 21 scenarii de NRT

Comparaison des résultats :
Indicateurs utilisés

Biais relatif de Monte-Carlo :

$$RB_{MC}\left(\hat{Y}\right)=\frac{1}{R}\sum_{j=1}^{R}\frac{\hat{Y}_j-Y}{Y}\times 100 \text{ (en \%)}$$

Erreur relative de Monte-Carlo :

$$RMSE_{MC}\left(\hat{Y}\right)=\frac{MSE_{MC}\left(\hat{Y}\right)}{MSE_{MC}\left(\hat{Y}_{Réf}\right)}=\frac{\sum_{j=1}^{R}\left(\hat{Y}_j-Y\right)^2}{\sum_{j=1}^{R}\left(\hat{Y}_{j,Réf}-Y\right)^2}$$

Analyse des résultats :
1 – Méthode des Scores

Principaux modèles trouvés par régression logistique selon le scénario (en % sur les 1 000 échantillons) :

Modélisation		Modèles trouvés : variables significatives													
Secteur				x	x	x	x	x	x	x	x	x	x	Autres modèles	Ensemble des modèles
Taille								x	x	x	x	x			
Localisation géographique			x			x	x	x			x	x	x		
Comportement de réponse en 2010						x	x	x			x	x	x		
Appartenance à un groupe								x					x		
Chiffre d'affaires					x		x			x		x			
Scénario															
Aléatoire simple	70 %	75												25	100
	80 %	75												25	100
	90 %	76												24	100
ZAU	70 %	56	16											28	100
	80 %	37	29											34	100
	90 %		53											47	100
Taux d'endettement	70 %	65												35	100
	80 %	60												40	100
	90 %	41		19										40	100
GRH	70 %					41	20	22			12			5	100
	80 %					15	28	25			26			6	100
	90 %						27				25	18	11	19	100
GRH x taux d'endettement	70 %					41	22	20			11			7	100
	80 %					16	35	28			17			5	100
	90 %						44	11			14	17		15	100
Secteur x taille	70 %				48				36					16	100
	80 %								74					26	100
	90 %								58	26				16	100
Secteur x taille x taux d'endettement	70 %				64				22					14	100
	80 %				44				41					15	100
	90 %				41				37					22	100

Note : les cases blanches représentent moins de 10 % des échantillons.
Lecture : lorsque le mécanisme de réponse est fondé sur les variables secteur x taille avec un taux de réponse de 70 %, pour 48 % des échantillons, le modèle trouvé comprend le secteur et le chiffre d'affaires.

Analyse des résultats : 1 – Méthode des Scores

Analyse des modèles trouvés :

Logiquement, peu de modèles trouvés :

- dans les scenarii ‘aléatoire simple’ ou avec variable cachée seule
- sauf quand taux de réponse = 90 % -> trouve variable corrélée (‘localisation’ pour ‘ZAU’, ‘secteur’ pour ‘taux d’endettement’)

Modèles en général bien reconstitués ou approchés :

- dans les autres scenarii (secteur x taille en particulier),
- même s’ils sont combinés à une variable cachée ;
- le taux de réponse intervient peu

Analyse des résultats : 1 – Méthode des Scores

Analyse des estimateurs ‘score’

Pour les variables quantitatives :

- biais relatif toujours faible (< 1,3 %)
- erreur relative faible (< 4 %),
  les scenarii avec GRH sont moins bien corrigés
- le taux de réponse est déterminant pour l’erreur relative

Pour les variables qualitatives :

- résultats plus nuancés
  (car estimateurs sur les seuls répondants sont déjà bons)
- biais relatif toujours faible (< 1 %),
  sauf dans le scénario secteur x taille x taux d’endettement
- erreur relative faible (< 1,7 %),
  -> qui diminue quand le nombre de répondants augmente
  -> et lorsqu’il y a une variable cachée ;
  - erreur maxi avec 90 % de réponse : 1,4 % pour ‘secteur x taille’

Analyse des résultats : 2 – Segmentation par arbres

Principaux « modèles » issus de la segmentation selon le scénario (en % sur les 1 000 échantillons) :

Variables			"Modèles" trouvés															
		Secteur				x	x	x	x	x	x	x	x	x	x	x	x	Autres	Ensemble
		Taille								x	x	x	x	x	x	x	x	
		Localisation géographique			x			x	x				x	x	x	x	x	
		Comportement de réponse en 2010						x	x			x		x	x	x	x	
		Appartenance à un groupe					x		x		x	x	x		x	x	x	
		Chiffre d'affaires		x		x	x	x	x	x	x	x	x	x			x	
Scénario	Aléatoire simple	70 %	63														37	100
		80 %	65														35	100
		90 %	65														35	100
	ZAU	70 %	44														56	100
		80 %	27		14												59	100
		90 %			19												81	100
	Taux d'endettement	70 %	52														48	100
		80 %	48														52	100
		90 %	29	10													61	100
	GRH	70 %							20					13	15	22	30	100
		80 %							12					12	24	40	12	100
		90 %							11						21	66	2	100
	GRH x taux d'endettement	70 %						12	17					11	13	21	26	100
		80 %							16					15	17	40	12	100
		90 %													13	68	19	100
	Secteur x taille	70 %				13				12							75	100
		80 %									17	16	13			11	45	100
		90 %									19	16	12			13	40	100
	Secteur x taille x taux d'endettement	70 %				19	13										68	100
		80 %								16	18	12					54	100
		90 %									19	19	16			14	32	100

Analyse des résultats : 2 – Segmentation par arbres

Analyse des modèles trouvés :

Souvent, davantage de variables utilisées :

- dans les scenarii avec GRH,
  -> toutes les variables sont souvent conservées
  -> (surtout quand le taux de réponse est élevé)
- grande variété de « modèles » trouvés dans le scenario secteur x taille

Le nombre de GRH généré varie entre 3 et 35

Analyse des résultats : 2 – Segmentation par arbres

Analyse des estimateurs ‘segmentation’

Pour les variables quantitatives :

- biais relatif toujours faible (< 1 %)
- erreur relative faible (< 4,5 %),
  du même ordre que ceux observés pour la méthode des scores
- le taux de réponse améliore nettement les résultats sur l’erreur relative

Pour les variables qualitatives :

- biais relatif toujours faible (< 1%),
  sauf dans le scénario secteur x taille x taux d’endettement
- erreur relative faible (< 1,8 %),
  -> qui diminue quand le nombre de répondants augmente
  -> résultats très proches de ceux obtenus
  par la méthode des scores

Comparaison : synthèse

« Modélisation » :

- la segmentation permet une interprétation plus simple des différents GRH

- la segmentation utilise davantage l'information auxiliaire
  (recours à davantage de variables)

- mais la segmentation nécessite d'ajuster le paramétrage

- et sa robustesse dans le cas de petits échantillons reste à vérifier

Comparaison : synthèse

Indicateurs de Monte-Carlo

- des résultats souvent proches,
- -> mais plus souvent en faveur de la segmentation
  (aussi bien pour les variables quantitatives que qualitatives)
- la méthode des scores s'en sort mieux en termes de biais quand le taux de répondants est plus faible

Proportion de cas où la segmentation est meilleure selon l'indicateur de Monte-Carlo (en % sur les 294 cas) :

		Biais relatif				Erreur relative			
		70 %	80 %	90 %	Total	70 %	80 %	90 %	Total
Type de variable	Variables quantitatives	43	51	63	52	71	51	83	69
	Variables qualitatives	44	48	62	51	62	51	63	59
Mécanisme de réponse	Aléatoire simple	14	43	36	31	93	93	43	76
	ZAU	79	71	79	76	57	57	64	60
	Taux d'endettement	57	64	57	60	43	57	79	60
	GRH	57	50	64	57	57	43	64	55
	GRH x taux d'endettement	43	50	71	55	50	43	57	50
	Secteur x taille	29	29	79	45	50	43	86	60
	Secteur x taille x taux d'endettement	29	36	50	38	50	36	79	55
Ensemble		**44**	**49**	**62**	**52**	**57**	**53**	**67**	**59**

Note : Les cases sont surlignées en jaune lorsque la segmentation est meilleure dans la majorité des cas.

Conclusion…

Léger avantage à la segmentation

- pour faciliter le regroupement de modalités
- pour faciliter la description des GRH constitués

  (vertu pédagogique : facile à présenter à un public non averti)
- par rapport aux différents indicateurs (proches toutefois)

Pistes :

- On peut poursuivre les tests en faisant varier le paramétrage (seuil de significativité…)
- La stabilité des « modèles » de segmentation pourrait être testée (validation croisée, ou utilisation d'un échantillon d'apprentissage et d'un échantillon-test)
- Dans la pratique, la régression logistique pourrait être conservée en phase préliminaire, pour sélectionner les variables auxiliaires, avant d'utiliser la segmentation par arbres

Merci de votre attention !

Contacts
M. Nicolas Sigler
Tél. : 02 40 41 78 23
Courriel : nicolas.sigler@insee.fr

M. Benoît Buisson
Courriel :
benoit.buisson@dgfip.finances.gouv.fr

Mme Emilie Dequidt
Tél. : 02 40 41 76 06
Courriel : emilie.dequidt@insee.fr

D.R. de l'Insee des Pays de la Loire
Service Études Diffusion
Pôle Ingénierie Statistique Entreprises
105 rue des Français Libres
BP 67401
44274 NANTES CEDEX 2

Intranet :
http://www.agora.insee.fr/jahia/Jahia/site/dr-pays-de-la-loire/PISE/pid/57410

INSEE