VÉRIFICATION SÉLECTIVE DES DONNÉES ET QUALITÉ DU PRÉDICTEUR UTILISÉ

Ph.Brion
consultant

9ème colloque francophone sur les sondages
Gatineau, Canada - octobre 2016

La vérification sélective

- Une technique appliquée pour le contrôle des données, qui consiste à:
  - Séparer l'ensemble des données en deux lots
  - L'un contrôlé de façon manuelle, l'autre traité de façon automatique
  - Pour ce faire, on utilise souvent une fonction de score

La vérification sélective (suite)

- La fonction DIFF

$$w_i\left(Y_i - Y_i^p\right)$$

- Où Y_i est la donnée observée, Y_i^p un « prédicteur », et w_i le poids de sondage

La vérification sélective(suite)

- La pratique montre que ce type de méthode fonctionne bien pour certaines variables, et moins bien pour d'autres
- Objectif de la présentation:
  - Modéliser l'impact de la qualité du prédicteur
  - En se limitant à une situation simplifiée: recensement, et étude limitée à une seule variable

La modélisation utilisée

- Trois valeurs, pour une variable et une unité *i*:
- Y_i^v est la valeur vraie (mais inconnue)
- Y_i est la valeur observée
- Y_i^p est la valeur du prédicteur

La modélisation utilisée (suite)

- On introduit une variable égale à l'erreur de mesure :

$$e_i = Y_i - Y_i^v$$

- On postule pour cette erreur de mesure une loi contaminée :

$$f(e_i) = (1-\pi_i)\delta(0) + \pi_i N(e_i, e_{i0}, \sigma_{ei}^2)$$

La modélisation utilisée (suite)

- On introduit une autre variable relative à la qualité du prédicteur:

$$u_i = Y_i^p - Y_i^v$$

- On postule une loi normale pour cette variable:

$$f(u_i) = N(u_i, 0, \sigma_{ui}^2)$$

Détermination de la loi de l'erreur a posteriori

- Ce qu'on connait : Y_i et u_i-e_i
- On applique la formule de Bayes pour déterminer la loi de e_i, sachant u_i-e_i
- On trouve (en oubliant l'indice *i*):

$$f(e/u-e) = (1-\tilde{\pi})\delta(0) + \tilde{\pi}N(e, \frac{e_0\sigma_u^2 - (u-e)\sigma_e^2}{\sigma_u^2+\sigma_e^2}, \frac{\sigma_u^2\sigma_e^2}{\sigma_u^2+\sigma_e^2})$$

Détermination de la loi de l'erreur a posteriori (suite)

- La probabilité de l'erreur est elle-même modifiée
- Cas où $e_0=0$

$$\tilde{\pi} = \frac{\pi \frac{\sigma_u}{\sqrt{\sigma_u^2+\sigma_e^2}} \exp\left(\frac{\sigma_e^2 (u-e)^2}{2\sigma_u^2(\sigma_e^2+\sigma_u^2)}\right)}{1-\pi+\pi \frac{\sigma_u}{\sqrt{\sigma_u^2+\sigma_e^2}} \exp\left(\frac{\sigma_e^2 (u-e)^2}{2\sigma_u^2(\sigma_e^2+\sigma_u^2)}\right)}$$

Utilisation de cette loi pour la vérification sélective

- Si on dispose de valeurs « proxy » pour σ_{ui} σ_{ei} et e_{0i} , on peut déterminer les unités à contrôler manuellement, ceci afin de limiter l'EQM due aux erreurs d'observation potentielles, par exemple en triant les unités selon la valeur de

$$E(e_i^2 / u_i - e_i)$$

Utilisation de cette loi pour la vérification sélective (suite)

- Dans le cas où $e_0=0$

$$E(e_i^2 / u_i - e_i) = \tilde{\pi}_i^2 \frac{((u_i - e_i)\sigma_{ei}^2)^2}{(\sigma_{ei}^2 + \sigma_{ui}^2)^2} + \tilde{\pi}_i^2 \frac{\sigma_{ei}^2 \sigma_{ui}^2}{\sigma_{ei}^2 + \sigma_{ui}^2}$$

- Le critère de sélection est donc plus complexe que la seule fonction DIFF

Quel est l'impact de la qualité du prédicteur ?

- Quand σ_{ui} tend vers 0, $\widetilde{\pi}_i$ tend vers 1 : l'usage de la fonction DIFF est justifié
- Quand σ_{ui} tend vers l'infini, $\widetilde{\pi}_i$ tend vers sa valeur initiale (aucune valeur ajoutée du prédicteur), et il est donc préférable d'utiliser une autre méthode

Conclusion

- Pour plus de détails se reporter au papier
- Développements possibles:
  - Quantifier sur des enquêtes passées, réaliser des simulations pour comparer le classement obtenu avec DIFF et la méthode proposée ici
  - Se placer dans le cadre d'un sondage
  - Se placer dans le cadre de l'utilisation d'un « redresseur » pour les unités détectées en erreur mais pas vérifiées de façon manuelle

Merci pour votre attention !

Philippe.brion55@gmail.com