Projet de Techniques d'Enquêtes

Souleymane Faye Max Michel Koffi Sébastien Culié

Magistère Ingénieur Economiste 1
AMSE - Ecole d'Economie d'Aix Marseille

20 Avril 2022

Introduction

Dans ce projet, nous étudions le traitement post-enquête de la non-réponse sur une variable unique
La population est un groupe d'élèves de Licence 1 à Aix-Marseille Université
La question posée est la suivante : **Pensez-vous être suffisamment informé sur les mesures anti-Covid-19 prises à AMU ?**
Notre objet d'études est la variable 24 de notre jeu de données : Mesures stop Covid.
Il y a quatre réponses possibles :

- *Oui, tout à fait*
- *Plutôt oui*
- *Plutôt non*
- *Non, pas du tout*

Pour simplifier nous codons sur SAS cette variable en deux modalités :

- *Oui*
- *Non*

Sommaire

1. Le problème de la non-réponse dans les enquêtes
2. Traitement de la non-réponse : concepts théoriques
3. Imputation par la régression : programmation SAS
4. Analyse empirique et résultats obtenus
5. Conclusion : Takeaway

Le problème de non-réponse dans les enquêtes

- Il y a non-réponse vis-à-vis de la variable Y pour l'individu échantillonné i dès lors que l'on ne dispose pas de la valeur Y_i relative à cet individu, quelle qu'en soit la cause
- La non-réponse est partielle pour les individus ici
- La non-réponse peut-être expliquée par l'indiscrétion de la question ou un état d'incertitude trop important. La deuxième semble plus probable dans le cas de la variable 24
- Introduction d'un biais de non-réponse : la non-réponse de l'individu à la question de la variable 24 nous donne une information sur celui-ci
- De plus, une diminution de la précision est causée par la non-réponse
- Si nous voulons corriger la non-réponse nous devons faire des hypothèses fortes sur le comportement des répondants
- En effet, il faut que la relation entre la variable dépendante et les variables explicatives qu'on modélise sur les répondants soit encore valable sur les non-répondants

Traitement de la non-réponse : concepts théoriques

Il est possible de traiter la non réponse par trois méthodes :

1. Répondération par calage sur marge
2. Imputation par la régression
3. Imputation par Hot Deck

- La répondération a un avantage dans le sens ou il n'y a aucun problème de spécification du modèle
- La contrainte dans l'imputation par régression est que les non-répondants doivent bien respecter le modèle que l'on va spécifier
- Nous avons tout de même préféré l'utiliser en faisant cette hypothèse
- La variable d'intérêt Y_i qui mesure la probabilité de réponse de l'individu est estimée par un modèle *Logit*

Imputation par la régression : programmation SAS

Pour traiter la non-réponse nous avons procédé de la manière suivante :

1. *Importation des données* et code en Oui versus Non de nos modalités
2. Code des variables administratives BN à CI en 0 et 1 : on ne conserve que les variables qui pourraient avoir un impact sur notre variable 24
3. *Double estimation* : avec toutes nos variables puis seulement avec les variables significatives
4. On fait la *prédiction de réponse des non répondants* à la question de notre variable 24
5. *Estimation de la proportion des individus ayant répondu oui* à la question dans la grande population

Analyse empirique et résultats obtenus

- Résultats obtenus pour les coefficients : interprétation des coefficients de la matrice β
  - Un coefficient positif dans la matrice β associé à la variable explicative x_{ik} veut dire que les individus ayant cette caractéristique ont une probabilité plus grande de répondre oui à la question d'intérêt; un coefficient négatif de répondre non
- Test de significativité statistique des variables qui nous amène à un second modèle
  - Le test de Wald et le Score de Fischer nous font revenir à un modèle de deux variables au lieu des sept que nous avions retenu

Conclusion - Takeaway

- L'estimation nous a donné presque exclusivement des réponses oui à la question d'intérêt pour les individus non-répondants
- Les résultats de l'imputation ont des limites :
  - Notre interprétation dans les variables administratives que l'on retient
  - L'erreur qu'on utilise pour l'estimation du modèle Logit : plus on prend une erreur importante, moins les coefficients retenus sont réellement significatifs
  - En conséquence, nous avons pris une erreur de 0,01 pour une plus grande robustesse
- L'estimateur de la proportion dans la grande population donne 0,53 pour une réponse positive à la question 24 lorsque les réponses manquantes ont été imputées par régression :
  - Il est intéressant de le comparer à l'estimation de la proportion dans la grande population des individus ayant répondu à la question 24