Travail sur les données

> *Datasets are the single most valuable (and most expensive) outputs of our research. If an evaluation cost one million dollars, shouldn't we value the underlying dataset at one million dollars as well?*
> *–(Bhavani) Prathap Kasina, former Associate Director Research, J-PAL South Asia, current Associate Programs Director, IPA.*

Il est essentiel de veiller à la sauvegarde et à la sécurité des données. Les données doivent être sauvegardées et protégées en permanence, par des procédures prises en conformité avec le protocole de recherche et les recommandations du comité d'éthique. Une fois la sécurité des données garantie, des protocoles adaptés en matière de nettoyage et d'analyse des données sont nécessaires pour que les procédés d'analyse des données soient transparents et reproductibles. Les trois principaux types de gestion des données qui sont abordés dans cette partie sont :

1. Sécurité, sauvegarde et stockage des données
2. Nettoyage des données
3. Analyse des données

Sécurité, sauvegarde et stockage des données

Le stockage et la sécurité des données soulèvent les questions suivantes :

- Où est-il logique de stocker les données ? Est-il préférable de les stocker sur le cloud ? Si non, pourquoi ?
- L'ensemble des données doivent-elles être protégées par un mot de passe ? Qu'est-ce qu'un bon mot de passe ? Quels sont les bons protocoles de transmission des mots de passe entre membres de l'équipe ?
- Quelles sont les informations d'identification personnelle et les autres données sensibles qui doivent être encryptées ?
- Pour avoir une classification des informations considérées comme sensibles, se référer au tableau de classification des données de Harvard.
- Bien que TrueCrypt ne soit plus en développement ni mis à jour, ce logiciel reste téléchargeable. Il est utilisé dans la plupart des projets du réseau J-PAL/IPA.
- BoxCryptor est une autre solution largement utilisée pour l'encryptage de données stockées sur le cloud (à l'aide de services tels que Box, Dropbox, Google Drive, etc.)
- Les systèmes d'exploitation Windows Vista et ultérieurs proposent BitLocker, un logiciel qui encrypte toutes les données stockées sur l'ordinateur.
- Si les données sont stockées dans un emplacement non sécurisé, quelle est la marche à suivre pour supprimer les identifiants ? Les données anonymisées peuvent-elles être partagées avec des parties non approuvées par le comité d'éthique pour traiter des données personnelles ?
- À quelle fréquence faut-il effectuer une copie de sauvegarde des données ? et dans combien d'endroits la sauvegarde doit-elle être effectuée?
- Comment transférer des données en interne et vers l'extérieur ? Qu'est-ce qui constitue un canal sécurisé pour le transfert de données ?

Nettoyage de données

Même les données tirées des enquêtes les mieux conçues doivent généralement faire l'objet d'une préparation et d'un nettoyage avant de pouvoir être analysées. Les questions qui se posent concernant les données entrantes sont notamment :

- Comment nommer, recoder et étiqueter les données de la manière la plus intuitive possible ?
- Existe-t-il des bonnes pratiques pour la préparation des données avant leur analyse ?
- Quelles vérifications standard doit-on mener sur les identifiants uniques et les valeurs variables ?
- Comment faire face aux conflits dans les données ?
- Comment gérer les données manquantes?
- Quel est le meilleur moyen de vérifier la cohérence logique des réponses et de vérifier les réponses au sein des données par rapport aux options d'enquête?

Analyse de données

Lors de l'analyse des données relatives à un projet, on s'intéresse principalement au lien entre deux variables. Par exemple, l'analyse peut être utilisée pour : vérifier des théories, comprendre les liens qui unissent différentes variables, prévoir des résultats et exécuter des simulations. L'analyse d'impact peut-être basique, par exemple pour déterminer s'il existe une différence de résultats statistiquement significative entre le groupe test et le groupe témoin, ou plus complexe, comme lorsqu'il s'agit d'utiliser les données pour étudier des impacts hétérogènes, des effets d'externalité, etc.

- Pour avoir un vaste aperçu de la théorie qui sous-tend

l'inférence causale dans les évaluations aléatoires, consulter le Guide d'utilisation de la randomisation.

- Mastering 'Metrics (Maîtriser l'économétrie) de Joshua Angrist et Jörn-Steffen Pischke est un bon guide d'économétrie sur laquelle s'appuie l'inférence causale dans différents modèles d'étude, notamment l'évaluation aléatoire.
- Colin Cameron et Praveen Trivedi ont rédigé un guide complet ayant trait à la réalisation d'analyses micro-économétriques à l'aide du logiciel Stata.
- Chuck Huber et David Drukker ont signé une série d'articles sur le blog de Stata concernant l'utilisation de la fonctionnalité des effets du programme de ce logiciel.
- Les diapositives de Christopher Baum (2013) sur l'utilisation de Stata en matière de gestion de données et de recherches reproductibles contiennent de nombreuses informations et astuces utiles.
- L'ouvrage Analysis of Household Surveys (Analyse des études sur les ménages, 1998) d'Angus Deaton aborde les questions de l'élaboration, de l'échantillonnage et de l'analyse dans le cadre des travaux de recherche, en particulier concernant les mesures de la consommation.

Pour plus de ressources techniques sur l'utilisation de logiciels dans l'optique de gérer et d'analyser des données, se reporter à la partie Outils logiciels.

Published on *The Abdul Latif Jameel Poverty Action Lab* (https://www.povertyactionlab.org)