Multiple Imputation and Modeling : Some Experiences from the KOF/ETHZ' s Innovation Survey 1999

Laurent Donzé
Swiss Federal Institute of Technology Zurich
ETH-Zentrum
CH – 8092 Zurich, Switzerland
E-mail : laurent.donze@kof.gess.ethz.ch
URL : http://www.dplanet.ch/users/ldonze

1. Introduction

Tous les trois ans, le KOF/ETHZ mène auprès des entreprises suisses une vaste enquête sur leurs activités d' innovation. En 1999, cette enquête a été adressée par courrier à un échantillon de quelque 6436 entreprises des secteurs de l' industrie, de la construction et des services. Un taux global de réponse de l' ordre de 34% a été obtenu. Evidemment, la complexité de l' enquête – questions quantitatives et qualitatives – conduit également à un fort taux de non-réponse partielle. Depuis quelques années, nous utilisons la technique de l' imputation pour la correction de cette non-réponse. Nous devons distinguer à ce sujet deux problématiques. La première consiste à imputer de manière la « meilleure possible » les valeurs manquantes tandis que la seconde est l' utilisation de ces valeurs imputées. Tels sont les éléments traités dans cette contribution.

2. L' imputation des données

La technique de l' imputation multiple, développée essentiellement par **Rubin**, consiste à générer un ensemble de *K* bases de données complètes, c' est-à-dire dont les valeurs manquantes ont été imputées. Plus le nombre *K* d' imputations est grand, plus les estimateurs seront précis. Cependant, en pratique, on constate qu' on a de bons résultats déjà à partir d' un petit nombre d' imputations, par exemple en fixant *K*=5.

La première exigence à remplir lorsqu' on impute des données manquantes est de choisir une *procédure d' imputation des données « appropriées »* (« proper imputation »). Une telle procédure incorpore la variabilité adéquate parmi les *K* ensembles d' imputation dans le cadre d' un modèle (cf. définition exacte in **Rubin** (1987), chap. 4). Toutes les procédures ne sont pas « appropriées ». **Rubin** propose une procédure d' imputation, simple et générale, qui a ces propriétés : *la procédure ABB* (« Approximate Bayesian Bootstrap »). C' est une procédure de type « hot-deck » qui incorpore les idées des méthodes bootstrap. Elle suppose que l' on ait au préalable classé les répondants et non-répondants en classes homogènes, appelées cellules d' ajustement ou d' imputation. Ces cellules peuvent être relativement facilement construites, par exemple en appliquant la méthode dite des *« propensity scores »* (cf. **Rosenbaum** et **Rubin** (1985)), ou en admettant que les strates de notre échantillon sont suffisamment homogènes.

L' imputation d' une variable ou d' un groupe de variables (cf. ci-dessous) suit les deux étapes suivantes. D' abord, la construction de cellules d' imputation. Ensuite, l' imputation des données manquantes par la méthode ABB. On répète la seconde étape 5 fois de manière à constituer pour une variable *X* à valeurs manquantes, 5 variables *XI1*, …, *XI5* à valeurs complètes. Ces variables sont ajoutées à la base de données initiales. Cette façon de faire permet d' utiliser aisément les données dans un cadre de travail uniforme et unique.

Dans le cadre de notre enquête, nous avons affaire à un grand nombre de variables qui peuvent être de nature très différente les unes des autres. En outre, la structure du questionnaire est telle que nous sommes souvent confrontés à des groupes de variables, par exemple lorsque la réponse à une question est détaillée selon différents aspects. Si nous voulons être en mesure d' imputer rapidement et de la manière la « meilleure possible » nos données manquantes pour toutes les variables d' intérêt du questionnaire, nous devons bénéficier de l' apport d' une procédure générale et efficace.

C'est pourquoi, nous avons opté pour la méthode ABB d'imputation qui s'effectue pour chaque variable, ou chaque groupe de variables, à imputer, et par cellule d'imputation. Le cas des groupes de variables est traité comme suit. On assigne aux variables du groupe désigné comme manquant, les valeurs des variables d'un groupe reconnu non manquant. Les cellules d'imputation sont obtenues par la méthode des « propensity scores » qui consiste essentiellement à modéliser et estimer la probabilité de réponse à une question, à regrouper ces probabilité en classes, par exemple en quintiles, puis finalement à identifier les observations correspondantes à ces probabilités. Dans notre cas, l'expérience nous a montré qu'on peut en général de manière satisfaisante estimer par un modèle de type logit la probabilité de réponse à une question dont les variables explicatives sont les variables de structure de l'enquête : branches économiques, tailles des entreprises, variables régionales.

3. L'utilisation des données imputées

Dans l'analyse de données imputées, par exemple dans le cas d'un calcul d'une moyenne ou l'estimation d'un modèle de régression, on applique la même procédure à chacune des *K* (vecteurs de) variables imputées. Les résultats sont ensuite combinés —en en prenant la moyenne —de manière à refléter la variabilité supplémentaire due à l'imputation des données manquantes, ce que ne permet pas une imputation unique, du moins simplement. On peut aussi facilement estimer la variance des estimateurs qui sont asymptotiquement normaux.

Toutes les analyses descriptives de l'enquête sont effectuées avec les données imputées. En général, on observe une relative stabilité des différentes imputations ainsi que du résultat final par rapport à la variable initiale. L'estimation économétrique de différents modèles d'innovation, par exemple l'estimation d'un modèle de type logit de l'innovation-produit, profite largement de données imputées, notamment en évitant la perte substantielle de données due à la combinaison de différentes variables incomplètes. Dans notre travail d'estimation, nous avons procédé ainsi. D'abord, nous avons trouvé un modèle satisfaisant sur la base des variables initiales, non imputées. Ensuite, nous avons testé ce modèle avec les données imputées. Cela permet d'évaluer l'effet de la perte d'informations due aux données manquantes et à l'élimination de données, et de confirmer la significativité et le signe des paramètres estimés.

REFERENCE

Donzé, L. (2001) : "L'imputation des données manquantes, la technique de l'imputation multiple, les conséquences sur l'analyse des données : l'enquête 1999 KOF/ETHZ sur l'innovation, exposé présenté au Congrès annuel de la Société suisse d'économie et de statistique, les 15 et 16 mars 2001, à Genève.

Rosenbaum, P. R., Rubin, D. B. (1985) : "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score", The American Statistician, February 1985, Vol. 39, No. 1, pp. 33-38.

Rubin, D. B. (1987) : Multiple Imputation for Nonresponse in Surveys, John Wiley & Sons.

ABSTRACT

To analyze the KOF/ETHZ's innovation survey of 1999, we proceeded, in order to correct the item nonresponse, to the imputation of missing data. We have chosen the multiple imputation, a general method. First, we have constructed adjustment cells with the "propensity scores" method. Then, we have imputed the missing data with the "Approximate Bayesian Bootstrap" method. We have repeated this second step for 5 times and therefore obtained 5 other variables with imputed data for each one. The new database is successfully used to estimate the descriptive statistics and econometric models of our innovation survey.