
Régression sur données censurées avec indicateurs manquants

Feriel Bouhadjera*¹

¹Bouhadjera – Cedric-Cnam – France

Résumé

Les données censurées représentent une réelle difficulté dans la modélisation d'un problème de régression. Parmi les méthodes classiques permettant d'estimer la fonction de régression, nous pouvons notamment citer les méthodes à noyaux, dites *non-paramétriques*. Toutefois, dans le cas où l'indicateur de censure est manquant, il devient encore plus complexe d'estimer la fonction de régression. Ce travail porte sur un estimateur à noyaux de la fonction de régression basé sur la méthode locale linéaire lorsque la fonction de perte est l'erreur relative quadratique moyenne dans un contexte de données censurées avec indicateurs de censure manquants. Il s'agit d'une extension de Bouhadjera and Ould Saïd (2022) au contexte de données manquantes aléatoirement. L'approche proposée présente certains avantages notamment la stabilité en présence de données censurées et aberrantes avec indicateurs de censure manquants. Elle présente également l'avantage d'atténuer les effets de bords et de réduire le biais. Ainsi, sous des hypothèses appropriées, nous présentons un résultat de convergence asymptotique. Outre cette étude théorique, ce travail compare numériquement l'approche proposée avec des compétiteurs de la littérature, adaptés au contexte des données manquantes au hasard, dites MAR (Rubin, 1976). Parmi ces compétiteurs, nous retrouvons: la régression classique (Guessoum and Ould Said, 2008), la régression locale linéaire (Bouhadjera et al., 2022), la régression relative (Khardani and Slaoui, 2019) et la régression relative locale linéaire (Bouhadjera and Ould Said, 2022). Les simulations sont réalisées sur un jeu de données générées sous différents scénarios dépendants du modèle considéré (linéaire ou non-linéaire), de la taille d'échantillon, du taux de valeurs aberrantes, du taux de censure et du taux de valeurs manquantes. L'étude est complétée par une simulation basée sur un jeu de données réels illustrant les bonnes propriétés de l'approche.

Mots-Clés: Données censurées, données aberrantes, indicateurs manquants, lissage à noyaux, mécanisme MAR, régression relative, régression locale linéaire.

*Intervenant