Université Grenoble Alpes
Statistique en grande dimension
emilie.devijver@univ-grenoble-alpes.fr

Régression linéaire de Poisson

Les modèles de régression de Poisson sont couramment utilisés pour modéliser des données de comptage, où la variable réponse représente le nombre d'événements se produisant dans un intervalle donné. Ce TP vous permettra d'explorer ces modèles en utilisant un ensemble de données réelles, en appliquant une régularisation Lasso pour traiter la grande dimension des variables. Vous utiliserez le jeu de données *bike sharing* disponible sur UCI, qui contient des informations sur les locations de vélo. On se concentrera sur les données quotidiennes (le fichier day.csv).

Préparation des données

- Le jeu de données *bike sharing* est disponible sur UCI.
- Définition des variables :
  - Variable réponse y : le nombre total de vélos empruntés chaque jour (variable `cnt`).
  - Variables explicatives X : toutes les autres variables, telles que la température, l'humidité, la vitesse du vent, etc.
- Vérifiez les dimensions des données et assurez-vous qu'elles sont cohérentes avec le modèle.
- Nettoyez les données en vérifiant l'absence de valeurs manquantes ou de valeurs aberrantes.

Estimation et stabilité de l'estimateur

- Ajustez un modèle de régression de Poisson avec régularisation Lasso à l'aide du package `glmnet`.
- Utilisez la validation croisée pour sélectionner la meilleure valeur du paramètre de régularisation (par exemple, `lambda.min` ou `lambda.1se`).
- Stabilité de la sélection de variables
  - Effectuez un bootstrap sur les données (par exemple, 10 itérations) pour ajuster le modèle. Analysez si les variables sélectionnées ($\beta_j \neq 0$) restent stables entre les itérations lorsque le paramètre de régularisation est `lambda.1se`.
  - Que peut-on conclure sur la stabilité des variables sélectionnées ?
  - Que signifie la stabilité dans la sélection de variables ?

Sélection post-inférence

- Construction d'un intervalle de confiance
  - On note $\hat{S}$ l'ensemble des variables actives (coefficients non nuls) estimé par le `glmnet`. Calculez $\hat{\beta}$ l'estimateur du maximum de vraisemblance restreint à $\hat{S}$.

– Construisez l'intervalle de confiance suivant (basé sur la normalité asymptotique de l'estimateur du maximum de vraisemblance) :

$$\hat{\mathrm{IC}}_\alpha^j = \left[\hat{\beta}_j - \frac{q_{1-\alpha/2}\hat{\sigma}_j}{\sqrt{n}}; \hat{T}_j + \frac{q_{1-\alpha/2}\hat{\sigma}_j}{\sqrt{n}}\right]$$

où $q_{1-\alpha/2}$ est le quantile de la loi normale standard et $\hat{\sigma}_j$ est l'écart-type estimé de l'estimateur $\hat{\beta}_j$ basé sur la vraisemblance restreinte.

- Évaluation du niveau empirique :
  - Répétez le processus suivant 100 fois :
    - Générez de nouvelles données simulées à partir du modèle (à partir de la première estimation des paramètres, générer un échantillon de la même taille).
    - Ajustez un modèle régularisé et sélectionnez $\hat{S}$.
    - Calculez l'intervalle de confiance pour chaque coefficient actif.
    - Évaluez si le coefficient vrai (celui utilisé dans al génération, celui estimé au début du TP) est inclus dans l'intervalle ($\beta_j \in \hat{\mathrm{IC}}_\alpha^j$).
  - Calculez le niveau empirique moyen, et comparez ce niveau à la vraie valeur $\alpha = 0.05$.
- Que concluez-vous sur l'inférence après sélection de variables dans ce cadre ? Quels sont les défis spécifiques à cette approche ?