

TP: INTÉGRATION D'UN OFFSET DANS LE CAS DU MODÈLE BINOMIAL (RÉGRESSION LOGISTIQUE)

On se propose dans ce TP d'essayer de comprendre comment intégrer une exposition au risque dans le cadre d'un modèle de fréquence binomial. Nous avons déjà vu en cours comment l'effectuer dans le cas d'un modèle de fréquence qui suit la loi de Poisson.

Préparation des données :

- Simuler 1000 réalisations d'une loi Uniforme sur $[0, 1]$, que vous stockerez dans un vecteur de covariable X . Quelle est la moyenne empirique de ces simulations ?
- Simuler 1000 réalisations d'une loi Exponentielle de moyenne théorique 5, que vous stockerez dans un vecteur N . Quelle en est la moyenne empirique ?
- Simuler 1000 réalisations d'une loi Uniforme dont les bornes inférieures dépendent des réalisations des 1000 premières simulations de la première question (la borne supérieure vaut toujours 1). Stockez ces valeurs dans un vecteur d'exposition E . Quelle est l'exposition minimale ? Maximale ? Moyenne ?
- Construisez un vecteur Y dont la valeur vaut 1 si $N > 10$, 0 sinon. Quelle est la moyenne de Y ? Ce vecteur représente l'indicatrice de l'événement de sinistre en une année d'exposition au risque.
- Construisez un data frame à 3 colonnes, avec les 1000 observations de Y , X et E .

On se propose dans la suite d'effectuer les étapes suivantes pour illustrer le phénomène de prise en compte de l'offset.

- (1) Effectuer une première régression de Y sur X avec offset E en log (comme dans le modèle poissonnien) dans le modèle binomial (avec lien log). Le phénomène à expliquer sera l'événement $Y = 1$. Interpréter les résultats (s'il y en a).
- (2) Estimer maintenant un modèle GLM log-Poisson avec les mêmes informations à expliquer et explicatives. Qu'observez-vous ? A quoi cela pourrait vous servir ? Interpréter les résultats.
- (3) Tenter de re-estimer le modèle de la question 1) en spécifiant des valeurs initiales dans l'algorithme d'optimisation (à travers un argument de la fonction `glm`), par exemple en prenant les coefficients obtenus à la question 2). Qu'observez-vous ?
- (4) Estimons maintenant les paramètres d'une régression logistique avec un lien `complementary log-log`. Qu'obtenez-vous ? Interprétez les résultats.
- (5) Faites des prévisions à partir de ce modèle : pour cela, prévoyez les réponses moyennes pour des covariables dont la valeur est comprise en 0 et 1 par pas de 0,01 ; avec une exposition identique pour tous les individus, égale à la moyenne de

votre exposition.

Remarque : ce vecteur entre 0 et 1 a la même moyenne que le vecteur X initial, donc les prévisions moyennes de proportion de réponses égales à 1 devraient se ressembler si l'exposition est semblable.

Vous semble-t-il que l'offset a bien été pris en compte ?

- (6) Effectuez maintenant des prévisions avec une exposition au risque égale à 1 pour l'ensemble des individus. Quelle est la proportion moyenne de 1 prédite par le modèle ? Cela vous semble-t-il logique ? Interprétez.
- (7) Estimer un modèle GLM binomial avec lien logit, et offset en log. Comparer les paramètres estimés avec le modèle précédent.
- (8) Estimer un modèle GLM binomial avec lien logit, mais sans offset. L'information sur l'exposition au risque des individus sera intégrée grâce au paramètre `weight` de la fonction `glm`. Comparer les paramètres estimés avec le modèle précédent.
- (9) Effectuez les prévisions relatives à ces 2 dernières modélisations, pour des individus exposés toute l'année. Quelle est la proportion moyenne d'événement dans chacun de ces 2 modèles ? A-t-on tendance à sous-estimer ou à surestimer le taux de sinistralité ? Quel est le pire modèle ?
- (10) Faire un graphique où vous superposerez les prévisions des 3 modèles construits en fonction de la valeur de X , pour X variant de 0 à 1 par pas de 0,01.
- (11) Conclure : quelle est le bon modèle GLM binomial (quelle fonction de lien) pour la prise en compte de l'exposition quand l'offset est considéré en log ?

Nous allons maintenant retrouver ce résultat en posant le modèle "théoriquement". Pour cela, nous allons redéfinir la vraisemblance du modèle, que nous optimiserons "à la main".

- (1) Comment s'exprime la probabilité d'observer l'événement $Y = 1$ dans le modèle GLM binomial avec lien logit ? Sachant que l'on part du principe que les événements suivent un processus de Poisson, intégrez dans cette expression l'idée de l'exposition au risque.
- (2) En revenant à la définition même de la vraisemblance, programmez une fonction R qui donne l'expression de la log-vraisemblance pour l'ensemble de l'échantillon.
- (3) Optimisez cette fonction afin de trouver les paramètres du modèle, grâce à la fonction `optim`. Ces paramètres sont-ils différents du modèle de référence estimé dans la partie précédente (cf conclusion) ?
- (4) Effectuer les prévisions des probabilités d'occurrence de l'événement pour des valeurs de covariable identiques à la première partie du TP (par défaut l'exposition sera de 1 dans ces prévisions).
- (5) Quelle est la probabilité moyenne de ces prévisions ? Comparer avec le modèle retenu dans la section précédente.

Remarque : nous aurions aussi pu adopter une stratégie plus intuitive pour reconstruire ce résultat. Une autre solution consisterait à prendre le plus petit intervalle de temps d'exposition (par ex. 1 mois plutôt qu'un an), et dupliquer les assurés pour ensuite lancer une régression logistique standard qui prédit la probabilité mensuelle d'observer l'événement.