

TP: RÉPONSES CATÉGORIELLES DÉSÉQUILIBRÉES

On se propose dans ce TP d'illustrer le biais introduit par un jeu de données dont la réponse à expliquer ne comporte que peu d'événements d'intérêt. Nous mettons ensuite en application les techniques statistiques nécessaires pour corriger les estimateurs du biais induit par ce phénomène, afin de quantifier l'impact sur un jeu de données réel.

Cas paramétrique

On va construire un échantillon de 1000 observations avec une réponse de type Bernoulli(p) ($Y = 0$ ou $Y = 1$).

On introduit une dépendance de Y avec un prédicteur X ($X \sim \mathcal{N}(0, 1)$) suivant le modèle suivant :

$$\text{logit}(p) = \ln \left(\frac{E[Y|X=x]}{1 - E[Y|X=x]} \right) = -3.35 + 2x.$$

Ce modèle doit aboutir à environ 10% d'observations Y égales à 1.

A partir de ce jeu de données, on crée ensuite un nouvel échantillon basé sur toutes les observations égales à 1 du premier échantillon, et (1/9) des observations égales à 0 (toujours du premier échantillon). On va calculer les distributions de probabilité selon les approches suivantes vues en cours :

- sans aucune correction du biais,
- avec une correction via la methode weighting method,
- avec une correction via la prior correction.

Questions :

- (1) Générer le jeu de données dans R : pour cela, 1) simuler les X_i , 2) en déduire les p_i correspondants avec le modèle théorique spécifié, 3) simuler des réalisations de Y_i avec $Y_i \sim \mathcal{B}(p_i)$.
- (2) Créer un jeu de données dont les modalités de la réponse ont été re-équilibrées suivant la description ci-dessus.
- (3) Estimer un modèle logistique sur le jeu de données déséquilibré. Regarder les résultats, interpréter les coefficients de régression et la pertinence du modèle par rapport au modèle théorique ayant servi à générer les données.
- (4) Sur l'échantillon re-équilibré, estimer un modèle logistique sans ajustement des coefficients de régression (pas de correction du biais). Regarder les résultats, interpréter les coefficients de régression et la pertinence du modèle par rapport au modèle théorique ayant servi à générer les données.
- (5) Faire de même, mais en corrigeant le biais du coefficient de régression adéquat par la méthode de repondération (weighting method). Qu'en déduisez-vous ?

- (6) Faire de nouveau de même, mais en corrigeant le biais du coefficient de régression adéquat par la méthode *prior correction*. Qu'en déduisez-vous ? Cette méthode est-elle la plus adaptée ? Pourquoi ?
- (7) Tracer sur un même graphique les probabilités estimées (prévisions) de chaque modélisation en fonction des valeurs de x et déduisez-en le meilleur des différents modèles construits. Le biais se visualise-t-il bien ?

Cas non paramétrique

On travaille ici avec le jeu de données `Caravan` de la librairie `R ISLR`.