La régression logistique, quelques notes

Par Claire Durand,
professeur,
Département de sociologie,
Université de Montréal

La régression logistique, quelques notes...

Introduction

Les usages les plus fréquents de la régression logistique semblent se trouver en épidémiologie. Lorsque l'on nous fait part de résultats du type "Vos chances d'avoir un cancer de tel type sont trois fois plus grandes si vous avez telle caractéristique, si vous fumez ou si vous mangez de telle manière", les informations proviennent d'une régression logistique. La régression logistique donne des résultats sous forme de probabilités, de rapport de chance. La variable dépendante est une fonction logarithmique (d'où le nom logistique) du rapport entre la probabilité qu'un événement survienne et la probabilité qu'il ne survienne pas. Par exemple, on peut prédire la probabilité qu'une personne meure du sida après un temps déterminé étant donné ses caractéristiques physiologiques et le traitement qu'elle a reçu. La présence de certaines caractéristiques de chaque variable indépendante pourra faire augmenter ou diminuer le logarithme de la probabilité que cette mort survienne.

Les sciences sociales peuvent utiliser également cette même procédure pour prédire des événements tels la probabilité qu'une personne change d'emploi, la probabilité qu'un étudiant s'inscrive dans une université plutôt qu'une autre, la probabilité qu'un étudiant termine ses études dans un temps déterminé, etc. Il faut noter que lorsque des données sont datées, il est préférable d'utiliser la régression de survie qui permet de prédire la probabilité qu'un événement survienne à chaque unité de temps qui passe.

Tout comme pour les autres types de régression, la régression logistique peut être standard (toutes les variables sont entrées en même temps), hiérarchique ou séquentielle (entrée des variables par bloc) ou "statistique" (entrée en fonction de critères statistiques avec diverses méthodes dont la plus courante est le pas-à-pas *(stepwise*)). Tout comme pour les autres types de régression, on peut également évaluer la solution -- l'équation de régression -- selon deux critères: la quantité de prédiction et la qualité de la prédiction. Pour ce qui est des variables indépendantes, on se demandera si leur contribution est significative et comment on peut définir cette contribution.

Informations

Pour le modèle:
- Le χ^2 de maximum de vraisemblance (différentes versions dont celle de C.C. Brown et celle de Hosmer et Lemeshow) indique si les données sont compatibles avec le modèle, ***auquel cas la probabilité du χ^2 est plus grande que .05***. On inverse le processus normal de test statistique. Habituellement, l'hypothèse nulle postule l'indépendance entre les deux variables ou entre le groupe de variables et la variable dépendante; lorsque l'on rejette l'hypothèse nulle ($p<.05$), on conclut qu'il est peu probable que la relation observée dans l'échantillon soit due au hasard et par conséquent, il est plausible que la relation existe réellement dans la population. *Dans le cas du χ^2 de maximum de vraisemblance*, l'hypothèse nulle postule que les relations observées entre les variables dans l'échantillon sont conformes à l'existence des mêmes relations dans la population. Lorsque l'on

rejette l'hypothèse nulle, on doit conclure que les données de l'échantillon s'écartent de façon significative du modèle de relation postulé. Par conséquent, on **cherche à ne pas rejeter l'hypothèse nulle (donc p>,05),** auquel cas on peut conclure que les relations observées sont compatibles avec le "modèle" postulé.

Par contre, le χ^2 du modèle est lui un test classique de χ^2 et le fait que la probabilité du test soit inférieure à ,05 indique que les variables qui sont dans le modèle apportent une contribution significative à l'explication.

- La table de classification est un autre indice de justesse du modèle. Il faut toutefois l'utiliser en faisant attention à sa pertinence et au *point de césure*. La table de classification indique la proportion de cas dans l'échantillon qui seraient bien classés si on décidait de classer dans la catégorie 1 tous les cas dont la probabilité prédite d'appartenir à la catégorie 1 est supérieure au point de césure, compte tenu de leurs valeurs sur les variables indépendantes qui sont dans le modèle. Il est possible de modifier le point de césure vers le haut ou vers le bas (en utilisant la proportion de cas pour lesquelles la variable dépendant vaut 1) de façon à trouver le point de césure optimal qui permettra de prédire le mieux possible autant la probabilité que l'événement survienne (code 1) que la probabilité qu'il ne survienne pas (code 0).

Enfin, tout comme pour la régression linéaire, l'analyse des résidus indique jusqu'à quel point l'ensemble des cas sont bien prédits par le modèle.

Pour les prédicteurs:

Les indices qui permettront de comprendre l'apport de chaque variable sont les suivants:

- la contribution d'une variable indépendante ou d'un bloc de variables indépendantes à l'explication est significative si la probabilité du χ^2 de la variable ou du bloc de variables est inférieure à ,05, comme pour un test de χ^2 habituel.

- Pour une variable indépendante nominale (en catégories), chaque catégorie est liée significativement à la variable dépendante si la probabilité du χ^2 de Wald est inférieure à ,05. Pour une variable continue, la variable elle-même est liée à la V.D. selon les mêmes critères.

- Le signe du coefficient de régression informe sur le sens de la relation (positif ou négatif) soit, si la présence de la caractéristique chez un répondant fait augmenter ou diminuer la probabilité de connaître l'événement.

- Pour chaque catégorie d'une variable nominale et pour chaque variable continue, le rapport de cote (*odds ratio)*, qui est la fonction exponentielle du coefficient, indique combien de fois plus ou moins de "chances" on a de connaître l'événement représenté par la valeur 1 de la variable dépendante selon que l'on possède telle caractéristique indiquée par la variable indépendante (plutôt que la caractéristique de référence).

Note1:

-Il est possible de sauvegarder la valeur de la probabilité prédite ainsi que le groupe prédit comme variables. On peut ensuite faire des graphiques des relations entre les V.I. et la V.D. (sous forme de probabilité ou d'appartenance à un groupe), ce qui donne une bonne vision de la qualité du modèle.

Note2:

- On peut sélectionner des cas sur lesquels on base la prédiction et voir si celle-ci s'applique bien à d'autres cas. Exemple: Je pourrais prendre des cas au hasard mais je peux aussi voir si le modèle de prédiction basé sur les hommes tient pour les femmes ou si le modèle élaboré à partir des employés réguliers tient pour les employés occasionnels.

Comparaison entre régression ordinaire et logistique

Information	**Régression linéaire**	**Régression logistique**
Variable dépendante	métrique ou "pseudo-métrique" (0,1)	nominale ou ordinale, transformée ln ((P(y=1|x)/(P(y=0)|x))
Variables indépendantes	métrique ou "pseudo-métrique" (0,1) interactions	Toutes formes (nominale, ordinale, métrique) -interactions (en général plus faciles à entrer et à interpréter qu'en régression linéaire)
Justesse du modèle (Model-fit)	- Test F (significatif?) - Variance expliquée (quantité) - Analyse des résidus (qualité)	- χ^2 de maximum de vraisemblance; On cherche à de pas rejeter l'hypothèse nulle - χ^2 du modèle - Analyse de la classification obtenue - Analyse des résidus
Test par variable ou groupes de variables	T=coeff/s.e. F(change)	χ^2 - Test par catégorie: Wald=coeff^2/s.e.^{2} - Improvement χ^2
Types	standard hiérarchique statistique	standard hiérarchique statistique
coefficient de régression	b = augmentation de valeur de la V.D. pour un point d'augmentation de la V.I.	$E^b=\Psi$ = rapport de cote (*odds ratio*) : Jusqu'à quel point on a plus de chance de vivre l'événement Y=1 si on est dans la catégorie i de la V.I. plutôt que dans la catégorie de référence.