Le modèle logit

Christophe Benavent
christophe.benavent@free.fr

Michel Calciu
mihai.calciu@iae.univ-lille1.fr

Limites d’une formalisation linéaire d’un modèle causal

Le problème de la spécification

Dans de nombreux cas la spécification d'un modèle linéaire est incorrecte. Un de ces cas est celui où la variable prédite est dichotomique ou polytomique.

Exemple

Examinons le cas suivant : on suppose que le revenu explique le fait d'avoir acheté un téléphone cellulaire. La variable dépendante prend ici pour valeur 1, si l'achat a eu lieu, 0 dans le cas contraire. La variable indépendante est une variable continue, bornée à gauche (on suppose qu'il n'y a pas de revenu négatif). A priori rien n'empêche d'estimer un modèle de régression simple.

Illustration graphique

Examinons graphiquement, le type de résultats que l'on risque d'obtenir. Dans cet exemple, caricatural, on se rend compte très clairement que les individus dont les revenus sont les plus élevés possèdent tous un téléphone mobile, alors que les autres n'en possèdent pas. La droite de régression, qui passe par les centres de gravités des deux groupes de consommateurs, permet de prédire l'achat ou le non-achatt. Elle donne un score qui peut être assimilé sans difficulté à une probabilité d'achat. Ainsi pour l'individu A, la valeur d'achat est de l'ordre de 0,1. De même l'individu médian a une probabilité de 0,5.

Interprétation

Le problème posé par l'utilisation du modèle de régression se manifeste avec l'individu C pour lequel la valeur prédite est négative. Pour d'autres individus, la valeur prédite est supérieure à 1. Dans les deux cas, l'assimilation de la valeur prédite à une probabilité est remise en cause. Le modèle est théoriquement inconsistant. Il est donc inutile de se pencher sur d'autres défauts plus formels de l'utilisation d'un modèle de régression (distribution des erreurs, etc.).

Il est donc nécessaire de définir un nouveau modèle, dont la caractéristique principale sera de prédire une valeur comprise entre 0 et 1.

Spécification d'un modèle de probabilité

La notion de logit

Au lieu d'estimer Yi, l'objectif est d'estimer P(Yi=1). avec Pi compris ente 0 et 1. Le problème posé est qu'une mesure de probabilité est bornée à droite et à gauche. Il convient donc de trouver un moyen de supprimer ces bornes.

En divisant p par (1-p), la borne à gauche est annulée car lorsque p tend vers 1, p/(1-p) tend vers l'infini (+). Si p tend vers 0 alors p/(1-p) tend aussi vers zéro. On applique alors une seconde transformation, de type logarithmique, de telle manière à ce que lorsque p tend vers 0, la transformation tend vers moins l'infini. Cette double transformation est appelée le Logit de p.

Logit = ln(p/(1-p))

Illustration graphique

Cette transformation est indiquée dans le graphe suivant :

Formalisation

Le modèle que l'on va donc chercher à estimer, prend donc la forme suivante :

ln (p/(1-p)) = bX = Z

En manipulant de manière adéquate le modèle précédent on obtient un modèle

qui exprime p en fonction de z :

p = eZ/(1+eZ)

La première forme est la linéarisation de cette seconde expression.

L'estimation des paramètres

La régression linéaire

Le problème posé est d'estimer les paramètres de notre modèle, ceci ne poserait pas de problème si on avait une première estimation de p. Or tout ce que l'on connaît est le fait que les individus ont acheté ou n'ont pas acheté.

Reconsidérons le modèle. Le premier point à préciser est que p(A=1), représente en fait une moyenne : c'est la proportion des individus qui ont acheté et qui possèdent un vecteur commun de caractéristiques X. Un moyen simple d'estimer le modèle, serait alors de regrouper les individus qui ont un même profil, de calculer cette proportion, et d'utiliser celle-ci comme valeur de la probabilité. On calcule ensuite le logit, et il n'y a plus qu'à appliquer une méthode de MCO.

Cette méthode peut être raffinée, mais ce n'est pas celle qui est effectivement employée, même si d'un point de vue pédagogique, elle a un intérêt certain.

La méthode du maximum de vraisemblance

Plutôt que de raisonner en deux étapes, on emploie, une technique d'estimation appelée maximum de vraisemblance. Dans le cadre du modèle logit cette fonction s'écrit :

La signification de cette quantité est claire. Si un individu a pour valeur y=1, que la probabilité calculée est de 0.8, la vraisemblance pour cet individu est de 0,81 x 0,2 0 =0,8. Dans le cas contraire (y=0 et même probabilité) on a li=0,2. On se rend compte ainsi que si les estimations des probabilités sont en accord avec l'observation, la vraisemblance est maximisée. Puisque p dépend du vecteur de paramètre # et du vecteur de variable X, on va chercher à maximiser la vraisemblance en les manipulant. Naturellement on ne pourra pas toucher au vecteur X, car ce sont les données. Par contre on cherchera quelles sont les valeurs de # qui maximisent cette quantité l.

D'un point de vue pratique, il est plus commode d'utiliser la log-vraisemblance, notée L, celle-ci transformant les produits en somme. Maximiser cette quantité, revient à maximiser la vraisemblance.

A partir de ce moment le problème devient simple, puisque pour trouver le maximum de cette fonction, il suffit d'égaler sa dérivée à 0. En pratique, on utilise des méthodes numériques telles que l'algorithme de Newton-Raphson, pour trouver les valeurs recherchées.

L'interprétation des paramètres.

Analogies avec la régression

L'interprétation des valeurs des paramètres ne peut être conduite comme en régression. En effet dans un modèle linéaire, le paramètre a une interprétation simple : c'est la variation de Y qui suit une variation d'une unité de X. Par construction, il y a constance de l'effet. Le modèle logit, n'est pas linéaire, l'interprétation précédente est valable mais uniquement pour la forme linéarisée.

Particularités des modèles logit

Par contre l'interprétation du coefficient b quant à son influence sur p est plus délicate : elle varie selon la variable X (variable centrée) comme cela est illustré dans le diagramme suivant.

La courbe en forme de cloche indique la variation de la probabilité pour 1/4 d'unité de X. On se rend compte que l'impact maximal est obtenu autour de X=0, et que plus la valeur absolue de X est grande moins l'impact est important. Il en résulte qu'il est difficile d'interpréter la valeur b directement. C'est pourquoi, on peut être tenté d'utiliser un autre indicateur dont l'effet soit constant.

Si l'on réécrit le modèle sous la forme suivante :

p/(1-p)= eZ =

On s'aperçoit que la quantité exp(b) représente le facteur par lequel le rapport p/(1-p) augmente lorsque X varie d'une unité. Cette quantité est indépendante de X. Ainsi, avec cette expression on obtient une valeur de paramètre qui est plus interprétable que le b dans la forme linéarisée (il est difficile de se représenter un logit) et qui n'a pas l'inconvénient souligné précédemment dans la forme non-linéaire.

Dans l'exemple précédent le coefficient b était égal à 0,5. Exp(0,5)=1,64. Ceci signifie que pour une augmentation de une unité de X, le rapport p/(1-p) est augmenté de 1,64 fois. Si b=0, exp(b)=1, le rapport reste inchangé, ce qui signifie simplement que les probabilités ne dépendent pas de X. Si b est égal à 0, exp(b) sera inférieur à 1, ce qui signifie simplement que l'effet est négatif.

Evaluation du modèle

Introduction

Pour évaluer un modèle de régression logistique plusieurs techniques peuvent être employées de manière concurrente. Ces techniques peuvent être groupées en deux catégories, d'abord des tests de vraisemblance, ensuite des analyses de discriminance.

Vraisemblance

Un bon modèle est un modèle dont la vraisemblance est grande, c'est à dire qui tend vers 1. En pratique on utilise la "moins double log-vraisemblance": -2LL, de sorte que lorsque la vraisemblance tend vers 1 alors -2LL tend vers 0.

Pour tester l'adéquation d'un modèle, on utilise l'hypothèse suivante :

H0: -2LL = 0

La -2LL se distribuant comme le CHI2, avec N-p degré de liberté (N: taille de l'échantillon et p nombre de paramètres). Si la probabilité est inférieure au seuil de risque on est amené à rejeter l'hypothèse nulle que le modèle convient. (Si cette probabilité est supérieure on ne peut rejeter l'hypothèse nulle, on conviendra alors que le modèle est correct).

Le niveau d'adéquation

Une autre statistique est celle du niveau d'adéquation (goodness of fit). Elle est

donné par

cette statistique est, elle aussi, distribuée selon un Chi2 à N-p degré de liberté.

En fait des tests directs sont peu employés. On préfère généralement adopter une approche légèrement différente, qui consiste à comparer un modèle sans paramètre autre que la constante, et le modèle que l'on veut tester. On calcule pour chacun des deux modèles la -2LL. Le test consiste simplement à calculer la différence entre les deux valeurs, celle-ci se distribuant comme le chi2 à k degré de liberté correspondant au nombre de paramètres que comprend le modèle. L'hypothèse correspondante au test étant que la différence est nulle. Autrement dit que l'incorporation des variables n'apporte rien à la vraisemblance.

Le reclassement

Si on attribue un individu à un groupe si sa probabilité d'appartenance est supérieure à un certain niveau (généralement 0,5), il suffira d'examiner le tableau de répartition des valeurs observées entre les valeurs prédites. Une inadéquation absolue est obtenue lorsque l'on est capable de ne reclasser que 50% des individus. (possibilité d'un test Z de Fisher).

Une technique proche consiste à représenter l'histogramme des probabilités estimées. L'idéal est que les deux groupes soient bien différenciés.

Dans les deux cas la qualité des prédictions obtenues dépend étroitement du seuil choisit. Ceci mérite une discussion plus approfondie.

Remarques par rapport aux critères de reclassement

Le choix d'un seuil à 50%, a une signification probabiliste précise : c'est la probabilité de type classique, elle coïncide avec une définition fréquentielle lorsqu'il y a equi-probabilité, ce qui n'est pas toujours le cas. Dans ces cas là la probabilité seuil à retenir sera la probabilité marginale.

Une deuxième remarque est relative à l'équivalence des deux événements. Faire une mauvaise prédiction n'a pas forcément la même valeur pour les deux événements. Un exemple classique est dans le cas de la construction d'un modèle de réponse dans le cas de mailing. Si envoyer un mailing à une personne qui ne répondra pas coûte 10f, ne pas envoyer un mailing à une personne qui répondra peut produire un manque à gagner de 100f. Il vaut mieux donc dans ce cas choisir un seuil moins exigeant dans le second cas que dans le premier, et de favoriser des critères d'espérances.

Modèles stepwise

Comme pour la plupart des méthodes prédictives, des procédures d'introduction progressive des variables ont été mises au point.

Test des paramètres

Statistiques de Wald

Les paramètres obtenus dans le modèle logit font l'objet d'un test analogue à celui employé dans le cas de la régression ordinaire, mais au lieu d'employer un test t, on emploie la statistique de Wald qui se distribue selon la loi du Chi2 à un degré de liberté:

Wald = (b/sb)2

Un problème est posé lorsque le paramètre b a une valeur absolue importante, dans ce cas l'erreur type est trop grande, conduisant à ne pas rejeter l'hypothèse nulle à tors. C'est pourquoi, il est généralement recommandé d'utiliser un autre test : celui des différences du Chi2, dont on détaille le résultat dans la section suivante.

Corrélations partielles

De même qu'en régression multiple, la contribution des variables est difficile à déterminer lorsque les variables explicatives sont corrélées. Un moyen détourné est en régression ordinaire le calcul de corrélation partielle. Il en est de même pour la régression logistique. Cette corrélation étant calculée de la manière suivante :

Si la statistique de Wald est inférieure à deux k, R est simplifié à 0.

Tests des différences du CHI2

Le principe est simple, il suffit de tester un modèle avec la variable considérée, puis sans la variable, de mesurer la différence entre les -2LL (le nombre de degré de liberté est la différence entre les deux). Il s'agit donc de reprendre la méthode de test global, mais en comparant un modèle à k paramètre par rapport à un modèle à k-1 paramètres. Une procédure adéquate pour tester systématiquement les paramètres consiste dans un premier temps à établir un modèle général dont l'ajustement global soit le meilleur. Puis systématiquement, il s'agit d'ôter la variable à tester et d'effectuer la comparaison.