Estimation

Fabian Bastin

Avril 2011

Approche de vraisemblance

Nous supposons que nous disposons d'une population de taille I, de K alternatives, des choix individuels, ainsi que des valeurs des attributs pour toutes les alternatives dans toutes les situations de choix.

Les variables explicatives sont supposées exogènes à la situation de choix.

L' approche classique consiste à chercher les paramètres β qui maximise la probabilité des choix observés, autrement dit,

$$\max_{\beta} L(\beta) = \prod_{i=1}^{I} \prod_{k=1}^{K} \left(P_{i,k}(\beta)\right)^{Y_{i,k}},$$

où

$$Y_{i,k} = \begin{cases} 1 & \text{si } i \text{ a choisi } k; \\ 0 & \text{sinon.} \end{cases}$$

Fonction de vraisemblance

De manière équivalente,

$$\max_{\beta} L(\beta) = \prod_{i=1}^{I} P_i(\beta),$$

où nous avons omis l'indice de l'alternative choisie dans $P_i(\beta)$.

Note : la formulation précédente suppose que les observations sont indépendantes.

$\forall\, i,\ 0 < P_i(\beta) < 1$. Numériquement : instable.

Solution : transformer le produit en somme :

$$\max_{\beta} LL(\beta) = \sum_{i=1}^{I} \ln P_i(\beta).$$

Procédures d'optimisation

Nous sommes donc confrontés à un problème de maximisation sans contraintes :

$$\max_{\beta} LL(\beta) = \frac{1}{I} \sum_{i=1}^{I} \ln P_i(\beta),$$

où nous avons introduit le facteur $\frac{1}{I}$ pour éviter que la fonction n'explose quand I croît à l'infini.

Dans nombre de cas, la fonction sera même concave, autrement dit, nous aurons un problème d'optimisation convexe.

La division par la taille de la population n'est pas toujours reprise, mais cela permet d'obtenir une valeur non sensible à la taille de l'échantillonnage.

Algorithmes d'optimisation

Nombre de logiciel d'estimation de modèles de choix discrets (LIMDEP, ALOGIT, codes de Kenneth Train,...) sont basés sur Newton-Raphson.
L'approche se base sur l'approximation issue du développement de Taylor au second ordre

$$LL(\beta_{t+1}) \approx LL(\beta_t)+(\beta_{t+1}-\beta_t)^T \nabla_\beta LL(\beta_{t+1})+\frac{1}{2}(\beta_{t+1}-\beta_t)^T H_t(\beta_{t+1}-\beta_t),$$

où H_t est le hessien de $LL(\beta)$ en β_t (ou une approximation de ce hessien).

La maximisation de cette approximation donne

$$\beta_{t+1} = \beta_t - H_t^{-1} \nabla_\beta LL(\beta_t).$$

Recherche linéaire

Une extension immédiate est l'incorporation d'une longueur de pas α_t :

$$\beta_{t+1} = \beta + t - \alpha_t H_t^{-1} \nabla_\beta LL(\beta_t).$$

α_t sera choisi de manière à garantir une croissance suffisante de la fonction de log-vraisemblance.

On peut généraliser, et prouver la convergence vers un minimum local (global si la fonction est concave). Voir par exemple Nocedal et Wright, "Numerical Optimization", Springer, 1999.

Approximation du hessien

Comme d'ordinaire pour les méthode de recherche linéaire, on utilisera souvent une approximation du hessien plutôt que le hessien exact. Les méthodes classiques, telles que DFP ou BFGS fonctionnent.

Une autre mise-à-jour du hessien a été proposé, permettant de mieux exploiter la structure de la fonction de log-vraisemblance. Il s'agit de l'approximation BHHH, du nom de ses inventeurs : Berndt, Hall, Hall, et Hausman.

Approximation BHHH

Le *score* d'une observation est le gradient du logarithme de la probabilité de cette observations, par rapport au vecteur de paramètres β :

$$s_i(\beta_t) = \nabla_\beta \ln P_i(\beta_t),$$

Dès lors, le gradient de la fonction de log-vraisemblance est la moyenne empirique des scores :

$$\nabla_\beta LL(\beta_t) = \frac{1}{I} \sum_{i=1}^{I} s_i(\beta_t).$$

In approximation du hessien est obtenue en prenant la moyenne empirique des produits externes des scores :

$$B_t = \frac{1}{I} \sum_{i=1}^{I} s_i(\beta_t) s_i(\beta_t)^T.$$

BHHH : forces et faiblesses

Au point annulant la moyenne des scores, B_t donne aussi une approximation la matrice de variance-covariance des paramètres (voir la partie consacrée à l'analyse asymptotique).

Similaire à Gauss-Newton.

- Exploitation directe de la forme du problème pour construire dès les premières itérations une approximation sensée du hessien.
- Repose sur des hypothèses fortent, conduisant à la propriété dite "identité de l'information". Grosso-modo, il faut un modèle correctement formulé. Ce n'est jamais le cas en pratique, et pour des modèles complexes, la méthodes peut ne pas converger.

Critères de convergence

Vu que nous sommes en optimisation sans contraintes, nous pouvons chercher à vérifier la condition du premier ordre d'annulation du gradient (ou le gradient relatif).

Nous pouvons aussi utiliser la statistique

$$m_t = \nabla_\beta LL(\beta_t)^T(-H_t^{-1})\nabla_\beta LL(\beta_t).$$

m_t est la statistique de test pour l'hypothèse que tous les éléments du gradients sont nuls. Elle suit une χ^2 à n degrés de liberté (où n est la dimension de β).