Robustní odhad vícerozměrného modelu lineární regrese

Petr Jonáš

Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Katedra pravěpodobnosti a matematické statisitky
email: `jonas@karlin.mff.cuni.cz`

Abstrakt: V klasické statistice se k odhadování vícerozměrného modelu lineární regrese používá metoda nejmenších čtverců, která je velmi citlivá na odlehlá pozorování. Proto se v tomto příspěvku budeme zabývat robustními alternativami k této metodě a představíme metodu MLWS (Multivariate Least Weighted Squares), která je přímým zobecněním metody LWS (Least Weighted Squares).

VÍCEROZMĚRNÝ MODEL LINEÁRNÍ REGRESE.

Používá se v situacích, kdy máme vícerozměrná pozorování vysvětlujících proměnných ($X_t \in \mathbb{R}^q$) a vysvětlovaných proměnných ($Y_t \in \mathbb{R}^p$) a předpokládáme následující model

$$Y_t = B^{'} X_t + U_t. \tag{1}$$

Pro $t = 1, \dots, n$ jsou $U_t \in \mathbb{R}^p$ nezávislé, stejně rozdělené (iid) náhodné vektory s hustotou

$$f_{U_t}(x) = \frac{g(x^{'}\Sigma^{-1}x)}{(\det(\Sigma))^{1/2}},$$

kde Σ je positivně definitní matice (*scatter matrix*) a g je kladná funkce. Pokud exitují druhé momenty U_t, je Σ rovna kovarianční matici (až na vynásobení konstantou).

Použijeme-li značení $X = (X_1, \dots, X_T)^{'}$, $Y = (Y_1, \dots, Y_T)^{'}$ a $U = (U_1, \dots, U_T)^{'}$ můžeme model (1) zapsat maticově

$$Y = XB + U.$$

K odhadnutí tohoto modelu se nejčastěji používá metoda nejměnších čtverců (LS). LS odhad má tvar (viz [**1**])

$$\hat{B}_{LS} = (X^{'}X)^{-1}X^{'}Y,$$

a nestranný odhad matice Σ

$$\hat{\Sigma}_{LS} = \frac{1}{T-K}(Y - X\hat{B}_{OLS})^{'}(Y - X\hat{B}_{OLS}).$$

LS je velice citlivá na přítomnost odlehlých pozorování, které mohou náš odhad velmi zkreslit. Proto má smysl hledat její robustní alternativy.

NEJMENŠÍ VÁŽENÉ ČTVERCE (LWS).

Používá se pro (jednorozměrný) model lineární regrese. LWS odhad je definován jako řešení následující úlohy

$$\hat{\beta}^{n,w}_{LWS} := \underset{\beta \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) r^2_{(i)}(\beta). \tag{2}$$

Všimněme si, že váhy nejsou (na rozdíl od klasického WLS) reziduím přiřazeny přímo, ale jsou jim přiřazeny implicitně podle velikosti reziduí. Vhodnou volbou váhové funkce dostaneme například nejmenší useknuté čtverce (LTS). Metoda LWS dává konzistentní odhad parametru β při splnění poměrně obecných podmínek. Tento odhad může být velmi robustní. Podrobnosti o tomto odhadu lze nalézt například v [**2**].

ROBUSTNÍ ODHAD VÍCEROZMĚRNÉHO MODELU LIN. REGRESE.

Abychom mohli rozšířit LWS na vícerozměrný model, je potřeba nahradit vzdálenost reziduí ($r^2(\beta)$) vícerozměrnou vzdálenosti. Jako vhodná se ukazuje Mahalanobisova vzdálenost

$$d_t(B, \Sigma) = ((Y_t - B^{'}X_t)^{'}\Sigma^{-1}(Y_t - B^{'}X_t))^{1/2}. \tag{3}$$

Jako vhodné rozšíření LTS pro vícerozměrný případ je v článku [**1**] navrhnut odhad

$$\hat{B}_{MLTS} = \underset{B,\Sigma;|\Sigma|=1}{\arg\min} \sum_{s=1}^{h} d^2_{(s)}(B, \Sigma), \tag{4}$$

kde $d_{(1)}(B, \Sigma) \leq d_{(2)}(B, \Sigma) \leq \cdots \leq d_{(n)}(B, \Sigma)$ jsou uspořádané Mahalanobisovy vzdálenosti a $|\Sigma|$ značí determinant matice Σ. Stejným způsobem jako v jednorozměrném případě lze přidáním vah tento odhad zobecnit na vícerozměrné nejmenší vážené čtverce (MLTS)

$$\hat{B}^{n,w}_{MLWS} = \underset{B,\Sigma;|\Sigma|=1}{\arg\min} \sum_{s=1}^{n} w\left(\frac{s-1}{n}\right) d^2_{(s)}(B, \Sigma). \tag{5}$$

PLÁN DO BUDOUCNA.

O odhadu MLWS zatím umíme ukázat pouze existenci řešení úlohy (5). Do budoucna je ještě potřeba dokázat zda a za jakých podmínek bude náš odhad konzistentní. Neméně důležité bude navrhnout a implementovat rychlý a spolehlivý algoritmus.

Poděkování. Tato práce vznikla za podpory grantu GAČR 402/09/0557 a příspěvku od MFF UK.

Reference.

[**1**] Agullo, J., Croux, C.,Van Aelst,S. (2008). *The multivariate least-trimmed squares estimator.* Journal of multivariate analysis 99, 311-338.

[**2**] Mašíček, L. (2004). *Diagnostika a senzitivita robustních modelů.* Disertace MFF UK.