Maximálně věrohodné odhady a lineární regrese ve výběrových šetřeních

Šedová M. [1,2] **a Kulich M.** [1]

1. Katedra pravděpodobnosti a matematické statistiky, MFF UK, Sokolovská 83, Praha, ČR
2. EuroMISE centrum, Oddělení medicínské informatiky, ÚI AV ČR, v.v.i., Pod Vodárenskou věží 2, Praha, ČR

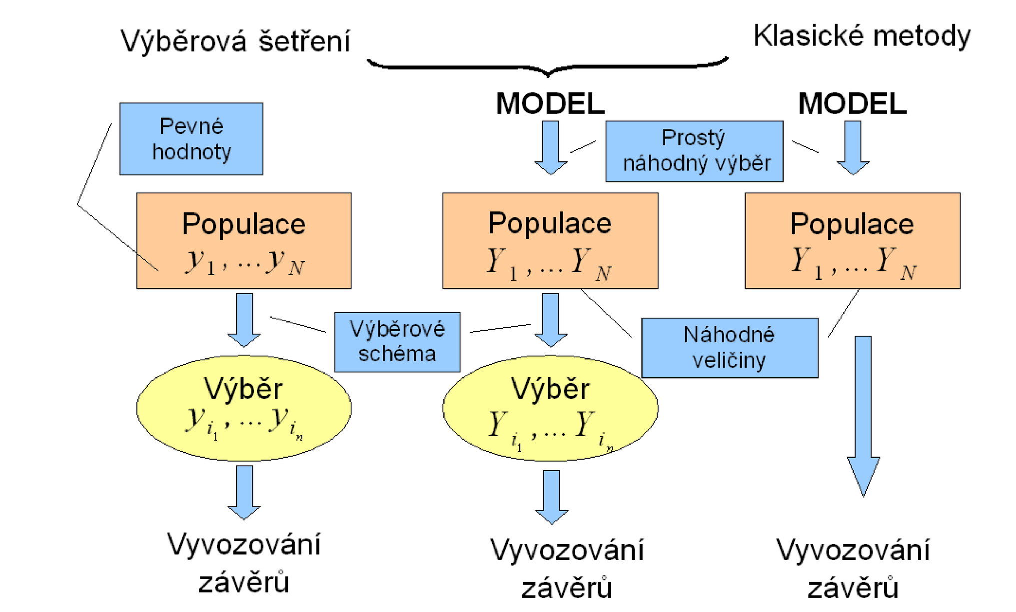
sedova@karlin.mff.cuni.cz

Abstrakt

V klasické teorii výběrových šetření jsou předmětem studia parametry charakterizující konečnou populaci, jako např. úhrn nebo průměr N pevných hodnot. Někdy je však vhodnější považovat pozorování za náhodné veličiny a zároveň brát v úvahu, že není k dispozici prostý náhodný výběr. Popisujeme alternativu maximálně věrohodných odhadů parametrů, která zohledňuje výběrové schéma, a výsledek ilustrujeme na lineárním modelu.

1. Pevná hodnota nebo náhodná veličina?

V kontextu výběrových šetření se zpravidla zabýváme parametry, které charakterizují konečnou populaci (např. úhrn nebo průměr N pevných hodnot). Někdy však může nastat situace, kdy bychom rádi výsledky zobecnili na jiné populace, nebo i tutéž populaci v jiném čase. Navíc, připustíme-li, že sesbíraná data nemusí být zcela spolehlivá, vidíme, že je vhodné chápat naše pozorování jako realizace náhodných veličin. Takto přistupují k datům klasické statistické metody. Ty však předkládají, že je k dispozici prostý náhodný výběr, což v kontextu výběrových šetření často není možné. Např. můžeme mít dvoustupňové výběrové schéma, kde jsou nejprve (se stejnou pravděpodobností) vybrány domácnosti a poté je ze všech členů dané domácnosti náhodně určen jeden a zařazen do studie. Tak vznikne výběr, který nadhodnocuje počet členů malých domácností a naopak podhodnocuje zastoupení domácností velkých.
Proto je někdy potřebné zvolit postup analýzy dat, který kombinuje oba tyto přístupy. Znamená to modifikovat metody tak, aby zohledňovaly dané výběrové schéma. Rozdíl v přístupu teorie výběrových šetření, klasických metodách a našem postupu (kombinace obojího) je schématicky popsán na obrázku 1.
Zde se zaměříme na analogii maximálně věrohodných odhadů a lineární regrese.



Obrázek 1: *Přístup teorie výběrových šetření, klasických metod a kombinace obojího.*

2. Odhad střední hodnoty

$W_i \dots$ diskrétní náhodná veličina, odpovídá stratu v populaci, $W_i \in \{1, 2, \dots, K\}$, $p_k = \mathrm{P}(W_i = k)$
$Y_i \dots$ spojitá nebo diskrétní náhodná veličina, odpovídá sledované veličině

$$\mathrm{E}\, Y_i = \theta = \sum_{k=1}^{K} p_k \theta_k, \text{ kde } \theta_k = \mathrm{E}\,(Y_i | W_i = k)$$

$\xi_i \dots$ náhodná veličina,

$$\xi_i = \begin{cases} 1 & \text{jedinec } i \text{ byl zahrnut do výběru} \\ 0 & \text{jedinec } i \text{ nebyl zahrnut do výběru} \end{cases}, \text{ nezávislé}$$

$\pi_i \dots$ pravděpodobnost výběru jedince i
$N \dots$ velikost populace

$$\mathrm{E}\,(\xi_i | W_i = k) = \mathrm{P}(\xi_i = 1 | W_i = k) = \pi_k, \text{ pro } i = 1, 2, \dots, N$$

$W_i \dots$ pozorované pro všech N členů populace
$Y_i \dots$ pozorované pouze pro jedince z výběru, tj. pro $\xi_i = 1$
Pro libovolnou náhodnou veličinu Z_i značíme

$$\mathrm{var}_k(Z_i) = \mathrm{var}\,(Z_i | W_i = k).$$

Odhad parametru θ definujeme

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \Big(\frac{1}{\hat{\pi}_k} \xi_i Y_i \Big) \mathrm{I}(W_i = k),$$

kde

$$\hat{\pi}_k = \frac{1}{n_k} \sum_{i=1}^{N} \xi_i \mathrm{I}(W_i = k),$$

$n_k \dots$ počet jedinců ve stratu k.

Tvrzení 1 *Předpokládejme, že vektory (Y_i, W_i, ξ_i) jsou nezávislé, stejně rozdělené (iid) a ξ_i je nezávislé s Y_i za podmínky W_i, pro $i = 1, 2, \dots, N$. Předpokládejme, že $\mathrm{var}\, Y_1 < \infty$. Pak*

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma),$$

kde

$$\Sigma = \mathrm{var}\, Y_1 + \sum_{k=1}^{K} p_k \frac{1 - \pi_k}{\pi_k} \mathrm{var}_k Y_1$$

Důkaz viz [1].

3. Maximálně věrohodné odhady

Nechť Y_i, $i = 1, 2, \dots, N$, jsou iid náhodné veličiny s hustotou $f(y, \boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p) \in \Theta$. Klasický maximálně věrohodný odhad parametru $\boldsymbol{\theta}$ se získá maximalizací logaritmu věrohodnostní funkce $L_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} L_i(\boldsymbol{\theta} | y_i)$. To většinou vede na soustavu rovnic

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{N} \boldsymbol{U}_i(\boldsymbol{\theta}) = 0, \quad \text{kde} \quad \boldsymbol{U}_i(\boldsymbol{\theta}) = \Big(\frac{\partial L_i(\boldsymbol{\theta} | y_i)}{\partial \theta_j} \Big)_{j=1}^{p}. \tag{1}$$

Pro výběrové schéma soustavu rovnic (1) modifikujeme na

$$\boldsymbol{V}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\xi_i}{\hat{\pi}_k} \boldsymbol{U}_i(\boldsymbol{\theta}) \mathrm{I}(W_i = k) = 0. \tag{2}$$

Řešením (2) získáme odhad parametru θ, zohledňující výběrové schéma (ZVS).

Tvrzení 2 *Nechť platí předpoklady Tvrzení 1. Označme $\hat{\theta}$ řešení rovnice (2). Potom*

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(0, J(\boldsymbol{\theta})^{-1} \Sigma J(\boldsymbol{\theta})^{-1}),$$

kde

$$\Sigma = J(\boldsymbol{\theta}) + \sum_{k=1}^{K} p_k \frac{1 - \pi_k}{\pi_k} J_k(\boldsymbol{\theta}).$$

$J(\boldsymbol{\theta}) \dots$ *Fisherova informace,* $J_k(\boldsymbol{\theta}) = \mathrm{var}(\boldsymbol{U}_i(\boldsymbol{\theta}) | W_i = k)$.

4. Lineární model

$(Y_i, \boldsymbol{x}_i)$, $i = 1, 2, \dots, N$ jsou iid náhodné vektory

$$Y_i | \boldsymbol{x}_i \sim (\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2), \quad i = 1, 2, \dots, N. \tag{3}$$

Soustava rovnic pro odhad parametru $\boldsymbol{\beta}$ (neuvažujeme-li výběrové schéma) je potom

$$\sum\nolimits_{i=1}^{N} \boldsymbol{x}_i (Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) = 0. \tag{4}$$

Předpokládejme nyní opět výběrové schéma popsané výše. Může se stát, že v každém stratu platí model (3), tj.

$$(Y_i | \boldsymbol{x}_i, W_i) \sim (\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2), \quad i = 1, 2, \dots, N. \tag{5}$$

Potom výběrové schéma není třeba zohledňovat a lze použít klasickou teorii lineárních modelů. Obecně však v každém stratu platí jiný vztah než (5), přičemž nás zajímá marginální model (3).
Příklad Uvažujme, že vektor $\boldsymbol{x}_i$ zahrnuje všechny prediktory Y_i (kromě W_i) a že ve stratu k platí

$$(Y_i | \boldsymbol{x}_i, W_i = k) \sim (\boldsymbol{x}_i^T \boldsymbol{\beta}_k, \sigma_k^2), \quad i = 1, 2, \dots, N.$$

Potom

$$\begin{aligned} \mathrm{E}\,(Y_i | \boldsymbol{x}_i) &= \sum_{k=1}^{K} p_k \boldsymbol{x}_i^T \boldsymbol{\beta}_k = \boldsymbol{x}_i^T \boldsymbol{\beta} \qquad (6) \\ \mathrm{var}\,(Y_i | \boldsymbol{x}_i) &= \mathrm{E}\, \sigma_{W_i}^2 + \mathrm{var}\, \boldsymbol{x}_i^T \boldsymbol{\beta}_{W_i} \\ &= \sum_{k=1}^{K} p_k \sigma_k^2 + \boldsymbol{x}_i^T \sum_{k=1}^{K} p_k (\boldsymbol{\beta}_k - \boldsymbol{\beta})^2 \boldsymbol{x}_i, \end{aligned}$$

kde $\sigma_{W_i} = \sigma_k$ a $\beta_{W_i} = \beta_k$ pro $W_i = k$.

Zde se již nejedná o klasický homoskedastický lineární model, neboť rozptyl Y_i závisí na prediktorech $\boldsymbol{x}_i$. V klasickém přístupu je $\hat{\boldsymbol{\beta}}$ stále řešením soustavy rovnic (4), avšak bereme-li v úvahu výběrové schéma, je ji nutné modifikovat podle (2) a dostáváme

$$\sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\xi_i}{\hat{\pi}_i} \boldsymbol{x}_i (Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) \mathrm{I}(W_i = k) = 0. \tag{7}$$

$i_1, \dots, i_n$ jsou všecha i taková, že $\xi_{i_l} = 1$ pro $l = 1, \dots, n$

$$D = \text{Diag}(d_{i_1}, \dots, d_{i_n}), \quad d_{i_l} = \sum_{k=1}^{K} \frac{1}{\hat{\pi}_k} \mathrm{I}(W_{i_l} = k),$$

$$\boldsymbol{X} = (x_{i_l j})_{n \times p} \quad \text{a} \quad \boldsymbol{Y} = (Y_{i_l})_{n \times 1} \quad \text{pro} \quad l = 1 \dots, n.$$

Řešením rovnice (7) je

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{D} \boldsymbol{X})^{-1} (\boldsymbol{X}^T \boldsymbol{D} \boldsymbol{Y}).$$

Pro $\hat{\boldsymbol{\beta}}$ platí

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, A^{-1} \Sigma A^{-1}),$$

kde

$$A = \mathrm{E}\, \boldsymbol{x}_i \boldsymbol{x}_i^T \quad \text{a} \quad \Sigma = \mathrm{var}\, \boldsymbol{U}_i + \sum_{k=1}^{K} p_k \frac{1 - \pi_k}{\pi_k} \mathrm{var}_k(\boldsymbol{U}_i)$$

pro $\boldsymbol{U}_i = \boldsymbol{x}_i (Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})$.
Konzistentní odhad rozptylu $A^{-1} \Sigma A^{-1}$ získáme nahrazením neznámých hodnot jejich konzistentními odhady. Pro $\mathrm{var}\, \boldsymbol{U}_i$ resp. $\mathrm{var}_k(\boldsymbol{U}_i)$ se použije tzv. robustní odhad, opět ZVS

$$\begin{aligned} \mathrm{var}_E U_i &= \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\xi_i}{\hat{\pi}_i} U_i U_i^T \mathrm{I}(W_i = k) \\ \mathrm{var}_E(U_i | W_i = k) &= \frac{\sum_{i=1}^{N} \xi_i U_i U_i^T \mathrm{I}(W_i = k)}{\sum_{i=1}^{N} \xi_i \mathrm{I}(W_i = k)}. \end{aligned}$$

5. Ilustrace

Výsledky ilustrujeme na malé simulační studii. Zajímá nás výše měsíčního platu v závislosti na vzdělání a délce praxe. Nechť výše platu závisí ale také na pohlaví.
Předpokládejme, že platí lineární model

$$\begin{aligned} plat = {} & \beta_0 + \beta_z * z + (\beta_p + \beta_{pz} * z) * p \\ & + \beta_s * S + \beta_m * M + (\beta_v + \beta_{vz} * z) * V + e, \end{aligned} \tag{8}$$

$z \dots$ dichotomická proměnná značící pohlaví
$p \dots$ délka praxe v letech
$S, M, V \dots$ faktor značící pořadě (alespoň) SŠ vzdělání bez maturity, s maturitou nebo VŠ vzdělání
$e \dots$ náhodná veličina, $e \sim N(0, 5000^2)$.
V simulované populaci jsou hodnoty parametrů následující

β_0 = 15 000 β_z = -5 000 β_p = 370 β_{pz} = -70
β_s = 3 000 β_m = 9 000 β_v = 27 000 β_{vz} = -9 000

Nechť je podíl žen mezi výdělečně činnými osobami 0.4. Marginální model závislosti platů na délce praxe a vzdělání je podle (6)

$$plat = 13000 + 342p + 3000S + 9000M + 23400V + e.$$

Postup Nejprve byla vygenerována populace o velikosti 10 000, dle modelu (8). Potom byl proveden náhodný výběr, pravděpodobnost zahrnutí žen 0.1, mužů 0.3. Parametry marginálního modelu byly odhadnuty klasickým způsobem i postupem ZVS. Výsledky jsou uvedené v Tabulce 1.

Tabulka 1. Odhady parametru pro jednu nasimulovanou populaci

	Skutečná hodnota	Odhad klasický	Odhad ZVS	Směr. chyba
Konstanta	13 000	14 360	13 197	597
Praxe	342	338	324	26
Vzdělání bez mat.	3 000	2 941	2 811	611
Vzdělání s mat.	9 000	8 759	8 609	612
Vzdělání VŠ	23 400	25 882	24 600	800

Postup ZVS odhadl lépe ty složky parametru β, které jsou pro každé pohlaví výrazně jiné (konstanta, ohodnocení VŠ). Poté byl tento proces zopakován 1 000 × (Tabulka 2).

Tabulka 2. Průměrné výsledky simulace (1 000 opakování)

	Skut. hodnota	Klas. odhad*	Odhad ZVS*	Odhad SE*	SE** odhadu
Konstanta	13 000	14 101	13 018	536	542
Praxe	342	358	343	26	26
Vzděl. bez mat.	3 000	2 980	2 977	551	571
Vzděl. s mat.	9 000	8 985	8 987	553	556
Vzděl. VŠ	23 400	25 333	23 374	789	798

*průměrný; **empirická
Průměrný odhad ZVS parametru je blízko skutečným hodnotám, průměrný odhad směrodatné chyby odhadu ZVS je blízký empirické směrodatné chybě odhadů.

Reference

[1] Šedová M., Kulich M. (2007): Statistical Methods for Analysis of Survey Data. in *WDS'07 Proceedings of Contributed Papers*, Prague, Matfyzpress, pp. 181–186.

Poděkování

Práce byla částečně podpořena výzkumným záměrem MSM 0021620839. Poděkování také patří KPMS a ČSOB za umožnění účasti M. Šedové na konferenci Robust 2008.