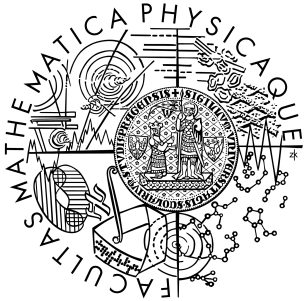
Katedra pravděpodobnosti a matematické statistiky



MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

Tomáš Krupa

Ekonometrický seminář 1

Referát

6. marca 2024

Zdroj

Sampson, M. (1990).
A Markov chain model for unskilled workers and the highly mobile.
Journal of the American Statistical Association,**85**,177-180.
https://www.jstor.org/stable/2289541

Motivácia

Chceli by sme štatistickým modelom popísať zmeny zamestnania mladých (resp. nekvalifikovaných) mužov v určitých odvetviach. Konkrétne:

- poľnohospodárstvo, lesníctvo, rybolov a banský priemysel
- stavebníctvo
- výrobný priemysel
- doprava, správa ciest a inžinierskych sietí
- veľkoobchod a maloobchod
- finančný sektor, zábavný a rekreačný priemysel
- odborné služby a verejná správa

Teda nás zaujíma, koľko ľudí prejde z jedného obdobia do druhého za jeden rok.

Model

Nazrime na tento problém ako na náhodný proces s konečnou množinou stavov (budú to zmienené odvetvia). Predstavíme a porovnáme tieto 3 modely:

- Markovov reťazec
- nový model navrhnutý v zmienenom článku
- Mover-Stayer model

Mover-Stayer model sa používa (resp. používal v dobe publikácie, t.j. 1990) na popis viacerých spoločenských javov, napr. migrácia, vernosť zákazníka nejakej značke, zmena miezd.

Majme diskrétnu, konečnú stavovú množinu $S = \{1, \ldots, 7\}$. Markovov reťazec na nej je charakterizovaný maticou pravdepodobnostných prechodov $\mathcal{Q}$, typu $7x7$.

$$\mathcal{Q} = \{p_{ij}\}_{i,j=1}^{7}$$

Prvky p_{ij} sú pravdepodobnosti prechodu (počas jedného kroku) zo stavu i do stavu j, teda platí:

$$\sum_{j=1}^{7} p_{ij} = 1$$

Odhad parametrov odvodíme metódou maximálnej vierohodnosti.

$$P(\mathbf{X}_1^{(n)} = \mathbf{x}_1^{(n)}) = P(X_1 = x_1) * \prod_{t=2}^{n} P(X_t = x_t | \mathbf{X}_1^{(t-1)} = \mathbf{x}_1^{(t-1)})$$

$$P(\mathbf{X}_1^{(n)} = \mathbf{x}_1^{(n)}) = P(X_1 = x_1) * \prod_{t=2}^{n} P(X_t = x_t | X_{t-1} = x_{t-1})$$

$$L(p) = \Pr(X_1 = x_1) \prod_{t=2}^{n} p_{x_{t-1} x_t}$$

$$L(p) = \Pr(X_1 = x_1) \prod_{i} \prod_{j} p_{ij}^{n_{ij}}$$

Kde n_{ij} je počet prechodov zo stavu i do stavu j.

$$l(p) = \log L(p) = \log \Pr(X_1 = x_1) + \sum_{i,j} n_{ij} \log p_{ij}$$

Tento výraz chceme maximalizovať a pritom dodržať podmienku:

$$\sum_j p_{ij} = 1$$

Použijeme metódu Lagrangeových multiplikátorov, teda maximalizujeme výraz:

$$l(p) - \sum_i \lambda_i \left(\sum_j p_{ij} - 1 \right)$$

Toto postupne zderivujeme podľa p_{ij} a postavíme rovné 0.

$$0 = \frac{n_{ij}}{p_{ij}} - \lambda_i, \quad \forall i, j$$

$$p_{ij} = \frac{n_{ij}}{\lambda_i}, \quad \forall i, j$$

Podľa podmienky platí:

$$\sum_j \frac{n_{ij}}{\lambda_i} = 1, \quad \forall i$$

Teda:

$$\sum_j n_{ij} = \lambda_i, \quad \forall i$$

Ako MLE dostávame:

$$\hat{p}_{ij} = \frac{n_{ij}}{\sum_j n_{ij}}$$

”nový model”

Tento model získame použitím pomerne silného predpokladu, že odvetvie, ktoré pracovníci opúšťajú, nemá vplyv na ich ďalšie uplatnenie (resp. sú nekvalifikovaní). Pripúšťame, že relatívne množstvo opúšťajúcich pracovníkov sa môže pre rôzne odvetvia líšiť. Matica prechodu bude mať tvar:

$$\mathcal{Q} = \theta + (\mathcal{I} - \theta)\mathbf{1}\mathbf{p}^\top$$

$$q_{ij} = \theta_i * \mathbb{1}\{i = j\} + (1 - \theta_i)p_j$$

Kde θ je diagonálna matica s prvkami $\theta_i \in (0, 1)$, ktoré reprezentujú pravdepodobnosti, s akou pracovníci v jednotlivých odvetviach zostanú vo svojom momentálnom zamestnaní. $\mathbf{p}$ je vektor s pravdepodobnosťami získania nového zamestnania v danom odvetví (pre všetkých pracovníkov nezávisle na ich predchádzajúcom zamestnaní). Odvodenie MLE je zložitejšie ako pre Markovov reťazec, dokonca nemá explicitný výsledok. Odhad získame Newtonovou iteračnou metódou.

Mover-Stayer model

Predpokladajme, že v každom odvetví sú od začiatku ľudia, ktorí meniť zamestnanie nebudú (stayers). V takom prípade by sme chceli použiť model, ktorý bude ”hýbať” iba časťou pracovníkov. Pohyblivú časť pracovníkov popíšeme nejakou maticou prechodu $\mathcal{M}$. Matica k-tého prechodu bude vyzerať takto:

$$\mathcal{Q}^{(k)} = \mathcal{S} + (\mathcal{I} - \mathcal{S})\mathcal{M}^k$$

$$q_{ij}(k) = s_i * \mathbb{1}\{i = j\} + (1 - s_i)m_{ij}(k)$$

Kde $\mathcal{S}$ je diagonálna matica s prvkami $s_i \in (0, 1)$, ktoré predstavujú množstvo ”stayers” v danom odvetví. Maticu prechodu $\mathcal{M}$ môžeme uvažovať ako maticu Markovovho reťazca, alebo ju môžeme skonštruovať podľa ”nového modelu” (podľa výsledkov prezentovaných v článku táto voľba neovplyvní MLE pre relatívne množstvo ”stayers” v jednotlivých odvetviach). Na MLE opäť potrebujeme Newtonovu metódu (ak sme použili maticu prechodu ”nového modelu”).

Testovanie na menšom množstve dát

V rámci testovania na dátach z rokov 1966 a 1967 porovnáme Markovov reťazec a nový model, nebudeme uvažovať Mover-Stayer. Naše dáta vyzerajú takto:

$$[n_{ij}] = \begin{bmatrix} 110 & 10 & 12 & 3 & 13 & 5 & 3 \\ 2 & 69 & 17 & 6 & 7 & 2 & 6 \\ 11 & 14 & 369 & 9 & 29 & 13 & 11 \\ 0 & 5 & 7 & 50 & 5 & 1 & 1 \\ 2 & 8 & 33 & 10 & 198 & 23 & 17 \\ 1 & 6 & 14 & 4 & 21 & 59 & 12 \\ 1 & 4 & 9 & 2 & 6 & 5 & 119 \end{bmatrix},$$

Testovanie na menšom množstve dát

MLE matice prechodov Markovovho reťazca:

$$\hat{Q} = \begin{bmatrix} .705 & .064 & .077 & .019 & .083 & .032 & .019 \\ .018 & .633 & .156 & .055 & .064 & .018 & .055 \\ .024 & .031 & .809 & .020 & .064 & .029 & .024 \\ .000 & .073 & .101 & .725 & .073 & .015 & .015 \\ .007 & .028 & .113 & .034 & .680 & .079 & .058 \\ .009 & .051 & .120 & .034 & .180 & .504 & .103 \\ .007 & .027 & .062 & .014 & .041 & .034 & .815 \end{bmatrix}$$

MLE matice θ a vektoru p.

$$\hat{p} = [.042 \quad .115 \quad .275 \quad .078 \quad .245 \quad .126 \quad .118]$$

$$\hat{\theta} = \text{diag}[.692 \quad .585 \quad .737 \quad .701 \quad .577 \quad .433 \quad .790],$$

Testovanie na menšom množstve dát

Samotný test bude Likelihood-ratio typu s takouto nulovou hypotézou:

$$H_0 : \mathcal{Q} = \theta + (\mathcal{I} - \theta)\mathbf{1}\mathbf{p}^\top$$

$$H_1 : (\mathcal{Q} \in \mathcal{R}^{7x7})\&(\neg H_0)$$

Rozdiel v počte parametrov je $7(7-1) - (7+6) = 29$. Teda kritická hodnota je $c = 42.56$. Hodnota testovej štatistiky je $lr = 42.54$. Nulovú hypotézu síce zamietnuť nemôžeme, ale nie je to veľmi presvedčivý výsledok.

Testovanie na väčšom množste dát

V rámci testovania na dátach z rokov 1966 až 1971 porovnáme Markovov reťazec, a obe spomenuté varianty modelu Mover-Stayer. Naše dáta vyzerajú takto:

$$[n_{ij}] = \begin{bmatrix} 435 & 33 & 53 & 10 & 33 & 12 & 10 \\ 26 & 456 & 59 & 19 & 24 & 14 & 22 \\ 41 & 73 & 1,989 & 37 & 127 & 50 & 45 \\ 9 & 23 & 28 & 343 & 17 & 7 & 9 \\ 16 & 33 & 139 & 40 & 979 & 79 & 59 \\ 8 & 15 & 52 & 12 & 79 & 344 & 37 \\ 7 & 15 & 45 & 11 & 30 & 31 & 685 \end{bmatrix},$$

$$[r] = [50 \quad 44 \quad 250 \quad 34 \quad 91 \quad 26 \quad 77],$$

$$[n(0)] = [156 \quad 109 \quad 456 \quad 69 \quad 291 \quad 117 \quad 146],$$

Kde r je počet pracovníkov v daných odvetviach, ktorý nezmenili zamestnanie ani raz a $n(0)$ je počiatočné množstvo pracovníkov v daných odvetviach.

Testovanie na väčšom množste dát

MLE odhady budú vyzerať takto:

$$\hat{Q} = \begin{bmatrix} .606 & .086 & .138 & .026 & .086 & .031 & .026 \\ .060 & .621 & .136 & .044 & .055 & .032 & .051 \\ .030 & .053 & .730 & .027 & .092 & .036 & .033 \\ .030 & .077 & .094 & .688 & .057 & .024 & .030 \\ .016 & .033 & .138 & .040 & .636 & .079 & .059 \\ .018 & .034 & .119 & .027 & .180 & .537 & .084 \\ .013 & .028 & .083 & .020 & .056 & .058 & .742 \end{bmatrix}$$

$$\hat{\theta} = \text{diag}[.579 \quad .571 \quad .621 \quad .663 \quad .529 \quad .473 \quad .711],$$

$$[\hat{s}] = [.260 \quad .343 \quad .430 \quad .401 \quad .233 \quad .186 \quad .390],$$

$$\hat{p} = [.064 \quad .117 \quad .287 \quad .075 \quad .228 \quad .121 \quad .109]$$

Kde $[\hat{s}]$ je odhad počtu pracovníkov so stabilným zamestnaním (stayers).

Testovanie na väčšom množste dát

Boli vykonané 2 testy Likelihood-ratio typu, prvý testoval či má zmysel použiť Mover-Stayer model, druhý testoval 2 varianty Mover-Stayer modelov proti sebe. Nulové hypotézy boli v danom poradí:

$$H_0 : s_i = 0, \quad \forall i$$

$$H_1 : (\mathcal{Q}^{(k)} = \mathcal{S} + (\mathcal{I} - \mathcal{S})\mathcal{M}^k) \& (\neg H_0)$$

$$H_0 : \mathcal{Q} = \theta + (\mathcal{I} - \theta)\mathbf{1}\mathbf{p}^\top$$

$$H_1 : (\mathcal{Q}^{(k)} = \mathcal{S} + (\mathcal{I} - \mathcal{S})\mathcal{M}^k) \& (\neg H_0)$$

Rozdiel v počte parametrov v prvom teste je 7, v druhom teste to je 29. Hodnoty testových štatistík boli $lr = 267.39; 137.97$ s príslušnými kritickými hodnotami $c = 14.07; 42.56$. Oba testy zamietali, na základe čoho súdime, že najvhodneší model v tejto situácii je Mover-Stayer (pôvodný, nie nový).

Ďakujem za pozornosť!