8 Výběr regresorů v mnohorozměrné regresi

Průvodce studiem

Kapitola se zabývá důležitou a často aplikovanou úlohou hledání vhodného regresního modelu. Na tuto kapitolu počítejte nejméně se třemi až čtyřmi hodinami studia. Důkladně prostudujte a promyslete i řešený příklad na konci této kapitoly.

Poměrně často se v analýze dat setkáváme s úlohami, které formálně mohou být zapsány jako klasický lineární regresní model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

ale v matici $\mathbf{X}$ typu $n \times (p+1)$ je počet regresorů p velký. Velký počet regresorů má často za následek, že pak pro mnoho z parametrů β_i, $i = 1, 2, \ldots, p$ nemůžeme zamítnout $H_0 : \beta_i = 0$, tzn. i-tý regresor nevysvětluje změny hodnot veličiny y.

Naším cílem je najít jednodušší model s k regresory ($k < p$), obsahující jen takové regresory, které významnou měrou vysvětlují variabilitu hodnot y.

Řešit takou úlohu prozkoumáním všech lineárních modelů k regresory, $k = 1, 2, \ldots, p$, je pro větší hodnoty p časově neúnosné. Znamenalo by to prozkoumat

$$\sum_{k=0}^{p} \binom{p}{k} = 2^p$$

modelů, tedy např. jen pro $p = 10$ výpočet a interpretaci výsledků odhadů 1024 modelů, což by představovalo práci na několik týdnů. Navíc pro velké hodnoty p bývá často matice $\mathbf{X}$ *špatně podmíněná*, tzn. determinant

$$\left|\mathbf{X}^T\mathbf{X}\right| \doteq 0$$

a odhady parametrů jsou pak numericky nestabilní a mají velké rozptyly, takže výsledky nejsou prakticky využitelné.

8.1 Kroková regrese

Jedním ze způsobů, jak nalézt podmnožinu k regresorů do vhodného lineárního regresního modelu, je *kroková (stepwise)* regrese. Ještě dříve, než vysvětlíme tento postup, připomeneme základní pojmy a myšlenky, na kterých je založen. Víme, že celkový součet čtverců lze rozložit:

$$\mathrm{TSS} = \mathrm{MSS} + \mathrm{RSS}$$

Dále, i modelovou sumu MSS můžeme rozložit. Představme si model s k regresory. Potom část MSS připadající i-tému regresoru,

$$\mathrm{MSS}(i \cdot 1, 2, \ldots, i-1, i+1, \ldots, k), \quad \text{označme ji zkratkou} \quad \mathrm{MSS}_{k(-i)}$$

$\mathrm{MSS}_{k(-i)}$ je tedy rozdíl modelové sumy čtverců $\mathrm{MSS}(k)$ při zařazení všech k regresorů a modelové sumy čtverců $\mathrm{MSS}(k-1)$ s $(k-1)$ regresory ($i-$tý regresor vynechán):

$$\mathrm{MSS}_{k(-i)} = \mathrm{MSS}(k) - \mathrm{MSS}(k-1).$$

Přidáním $i-$tého regresoru se současně změní odpovídajícím způsobem i residuální součet čtverců

$$\Delta\mathrm{RSS}_i = \mathrm{RSS}(k-1) - \mathrm{RSS}(k) = \mathrm{MSS}(k) - \mathrm{MSS}(k-1) = \mathrm{MSS}_{k(-i)},$$

neboť platí, že

$$\mathrm{TSS} = \mathrm{MSS}(k) + \mathrm{RSS}(k) = \mathrm{MSS}(k-1) + \mathrm{RSS}(k-1).$$

Současně víme, že $\Delta\mathrm{RSS}_i \geq 0$, tzn. přidáním regresoru se residuální součet čtverců nezvýší.

Myšlenka krokové regrese spočívá v tom, že v každém kroku (předpokládejme, že $k-1$ regresorů je už zařazeno v modelu) budeme z dosud nezařazených vybírat takový regresor, který nejvíce snižuje residuální sumu čtverců, tj. ten, jehož $\Delta\mathrm{RSS}_i$ je největší. Přitom zařadíme jen takový regresor, který residuální sumu čtverců snižuje významně. Kritériem (statistikou) pro posouzení významnosti je tzv. *parciální F*, což je

$$\frac{\Delta\mathrm{MSS}_i}{s^2} \sim F_{1,\nu},$$

kde $\Delta\mathrm{MSS}_i$ označuje zvýšení modelové sumy čtverců odpovídající zařazení $i-$tého regresoru z dosud v modelu nezařazených, s^2 je nestranný odhad parametru σ^2 a ν je jeho počet stupňů volnosti.

Implementace procedury krokové regrese se mohou lišit v tom, jakým způsobem je počítán s^2. Jedna z možností je počítat s^2 z residuální sumy čtverců v aktuálním kroku, tj.

$$s^2 = \frac{\mathrm{RSS}(k-1)}{n-k}.$$

Pak parciální F_i $i-$tého nezařazeného regresoru lze určit z parciálního a celkového korelačního koeficientu

$$F_i = \frac{[r^2_{iY\cdot(k-1)}]/(n-k)}{1 - r^2_{Y\cdot(k-1)}} = \frac{\Delta\mathrm{RSS}_i}{\mathrm{RSS}(k-1)/(n-k)},$$

kde $r_{iY\cdot(k-1)}$ je parciální korelační koeficient Y a x_i po „odečtení vlivu“ $(k-1)$ už zařazených regresorů a $r_{Y\cdot(k-1)}$ je mnohonásobný (celkový) koeficient korelace Y s $(k-1)$ už zařazenými regresory.

V $k-$tém kroku tedy zařadíme ten regresor, který má největší parciální F_i, a to jen tehdy, je-li F_i větší, než zadaná hodnota *F-to-entry*, kterou obvykle volíme jako takový kvantil F rozdělení, aby parciální F_i a tudíž i změna v residuálním součtu čtverců byly významné, tedy *F-to-entry* $= F_{1,(n-k-1)}(1-\alpha_1)$, kde α_1 je zvolená hladina významnosti pro zařazení regresoru do modelu.

Po zařazení $i-$tého regresoru může kvůli korelaci mezi zařazenými regresory nastat situace, že parciální F některého ze zařazených regresorů přestane být významné. Jinými slovy, vypuštění tohoto regresoru z modelu pak nezvýší významně residuální sumu čtverců, tzn. regresor je v modelu nadbytečný. Proto se po zařazení regresoru spočítají parciální F_i všech dosud zařazených regresorů

$$F_i == \frac{\Delta\mathrm{RSS}(i)}{s^2},$$

kde $\Delta\mathrm{RSS}(i)$ znamená změnu (zvýšení) residuální sumy čtverců při vypuštění $i-$tého regresoru z modelu. Najde se nejmenší z těchto parciálních F_i a posuzuje se, zda bychom vypuštěním tohoto regresoru zvýšili RSS jen nepodstatně. Kriterium pro toto rozhodování je to, zda minimální F_i je menší než zadaná hodnota F-to-remove. Většinou volíme F-to-remove $= F_{1,(n-k-1)}(1-\alpha_2)$, kde α_2 je zvolená hladina významnosti pro vypuštění regresoru z modelu. Abychom předešli možnosti nekonečného cyklu zařazování a vyřazování regresorů, obvykle se volí F-to-remove < F-to-entry, tj. $\alpha_2 > \alpha_1$.

Po tomto vysvětlení tedy můžeme algoritmus krokové regrese zapsat takto:

krok 0: zvol model se žádným regresorem , tj. $\hat{y} = \bar{y}$, a z p nezařazených regresorů zvol ten, který má největší absolutní hodnotu korelačního koeficientu s vysvětlovanou veličinou y (při jednom zařazeném regresoru je $R^2 = r_{xy}^2$, tedy nejvíce korelovaný regresor nejvíce snižuje residuální sumu čtverců)
if $F_i <$ F-to-entry **then** konec
else $k = 1$

krok k: mezi nezařazenými regresory vyber ten s největším F_i.
if $F_i <$ F-to-entry **then** konec
else zařaď $i-$tý regresor, $k = k+1$
mezi zařazenými regresory najdi ten s nejmenším F_i,
if min $F_i <$ F-to-remove **then** vyřaď $i-$tý regresor, $k = k-1$
go to krok k

Analogický krokový (stepwise) postup se, jak uvidíme dále, užívá nejen v lineární regresi, ale i v dalších metodách analýzy mnohorozměrných dat. Existují i některé další varianty, tzv. postupných procedur výběru veličin do modelu, např. zpětná *backward* procedura, která vychází z modelu, ve kterém je zařazeno všech p veličin a postupně vyřazuje nevýznamné, nebo dopředná *forward* procedura, která je podobná výše popsané krokové proceduře, avšak neumožňuje vypouštění zařazených veličin, které se stanou nevýznamné.

Obecně lze říci, že krokové procedury jsou užitečným nástrojem pro hledání vhodných modelů v mnohorozměrných datech, ale negarantují nalezení nejvhodnějšího modelu, neboť ho mohou „minout“. Pro hledání vhodného lineárního regresního modelu je spolehlivější procedura popsaná v další kapitole, ale ta je výpočetně podstatně náročnější, takže pro velmi rozsáhlá data může být její využití problematické.

8.2 Hledání nejlepší množiny regresorů

Systematičtěji než kroková regrese pracují procedury označované jako „all possible regressions“ nebo „best subset of regressors“. Pro každé $k = 1, \ldots, p$ hledají takovou k-tici regresorů, aby $R^2 =$ bylo pro daný počet regresorů maximální. Jak bylo dříve uvedeno, počet modelů roste exponenciálně s počtem potenciálních regresorů p, procedury využívající jen hrubou sílu, tj. opravdu zkoumají všechny modely, mohou být užity jen pro poměrně malý počet potenciálních regresorů p, např. v NCSS 2000 je to $p \leq 15$. V některých statistických programech je implementována heuristika, která sice nezajišťuje vyčerpávající prohledání všech modelů, ale zato dovoluje větší počet regresorů. V NCSS je to procedura Multivariate Selection v Regression – Variable selection routines.

Jelikož s rostoucím k index determinace R^2 neklesá (obvykle roste), není vhodným kritériem pro optimalizaci modelu. Vhodnějším kritériem je adjustovaný index determinace

$$R^2_{adj} = 1 - \frac{n-1}{n-k-1}\left(1-R^2\right) = R^2 - \left(1-R^2\right)\frac{k}{n-k-1}$$

nebo nejčastěji užívaná *Mallowsova statistika* C_p

$$C_p = [n-(k+1)]\,\frac{s_k^2}{s^2} - [n-2(k+1)] = n\left(\frac{s_k^2}{s^2}-1\right) - (k+1)\left(\frac{s_k^2}{s^2}-2\right),$$

kde s_k^2 je residuální rozptyl při k zařazených regresorech a s^2 je residuální rozptyl při všech zařazených regresorech. Střední hodnota této statistiky je

$$E(C_p) = k+1.$$

Když $(s_k^2/s^2) \approx 1$, tzn. model už nelze podstatně vylepšit, pak $C_p \approx k+1$ a zařazením zbytečného dalšího regresoru se C_p zvětší o 1. Tedy vzhledem k C_p je nejlepší ten model, který má C_p nejmenší, přibližně rovné počtu zařazených regresorů zvětšených o jedničku. Obvykle je však C_p jen jedním z kritérií při hledání nejvhodnějšího modelu, musíme vzít do úvahy i residuální rozptyl a další, většinou nestatistická kriteria, jako počet regresorů (čím méně, tím obvykle lépe), cena jejich měření (levnější má přednost), interpretaci vlivu regresoru na vysvětlovanou proměnnou atd. v závislosti na konkrétní úloze.

Příklad 8.1 Užití krokové regrese a prohledání všech podmnožin regresorů při hledání vhodného regresního modelu si ukážeme na datech, která jsou v souboru STEPWISE.XLS. V datech máme 30 pozorování vysvětlované veličiny y a deseti potenciálních regresorů, $x1, x2, \ldots, x10$. Úkolem je najít vhodný lineární regresní model. Pro výpočty byly užity programy stepwise regression a all subset z [14]. Výstupy jsou opět uvedeny v surovém stavu, jen s drobným zkrácením.

```
Stepwise Regression Report

Dependent Y

Iteration Detail Section
```

Iter. No.	Action	Variab	R-Squared	Sqrt(MSE)	Max R-Sqrd Other X's
0	Unchanged		0.000000	3.010984	0.000000
1	Added	x1	0.765361	1.484322	0.000000
2	Unchanged		0.765361	1.484322	0.000000
3	Added	x5	0.826587	1.29947	0.300558
4	Unchanged		0.826587	1.29947	0.300558
5	Added	x6	0.985724	0.3799527	0.822143
6	Unchanged		0.985724	0.3799527	0.822143
7	Added	x8	0.988579	0.3465689	0.918500
8	Unchanged		0.988579	0.3465689	0.918500
9	Added	x2	0.990822	0.3170781	0.978202
10	Unchanged		0.990822	0.3170781	0.978202
11	Added	x3	0.993080	0.2812623	0.978713
12	Unchanged		0.993080	0.2812623	0.978713

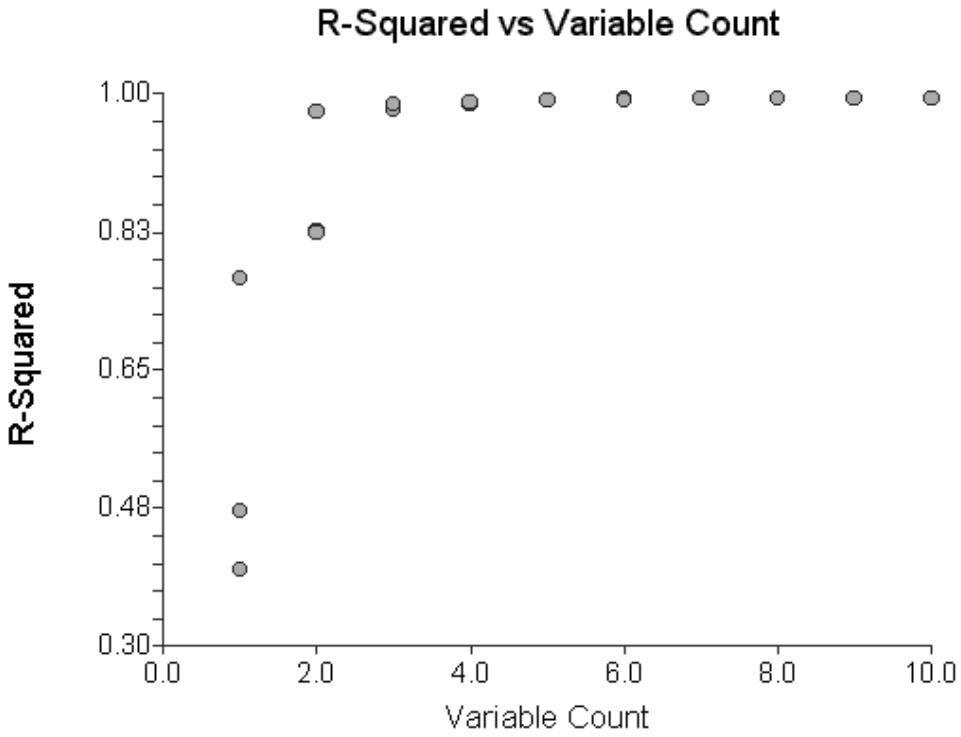
Při implicitním nastavení kritérií pro zařazování a vyřazování regresorů ($\alpha_1 = 0.05$, $\alpha_2 = 0.10$) byly postupně zařazovány regresory $x1$, $x5$, $x6$, $x8$, $x2$ a $x3$, žádný nebyl vyřazen. Při pohledu na výsledky vidíme, podstatná změna v R^2 a residuální směrodatné odchylky nastala po zařazení regresoru $x6$, tedy pro model se třemi regresory $x1$, $x5$, $x6$. Přidávání dalších regresorů už index determinace R^2 nijak významně nezvýšilo a ani zmenšení residuální směrodatné odchylky není nikterak dramatické. Model se třemi regresory $x1$, $x5$, $x6$ je tedy nejnadějněším kandidátem na model, který vhodně vysvětluje variabilitu veličiny y. Zda je to opravdu vhodný model je nutno zkoumat podrobněji s využitím postupů uvedených v příkladu o odhadu regresních parametrů a pak i posoudit věcné souvislosti s řešeným problémem.

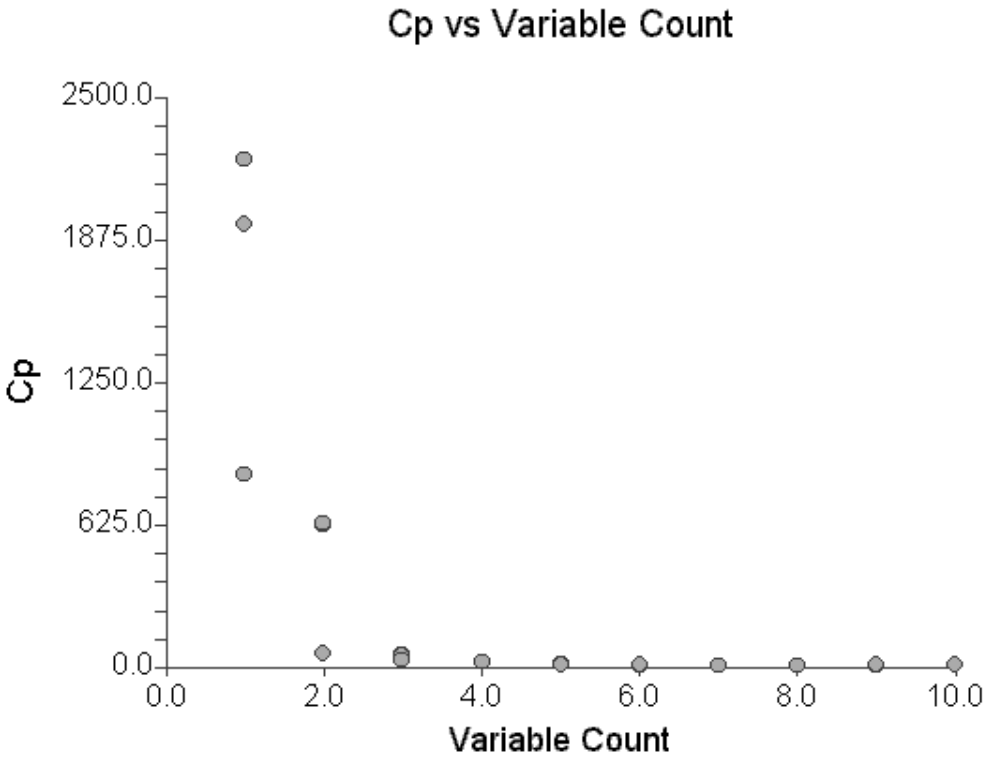
All Possible Results Section

Model Size	R-Squared	Root MSE	Cp	Model
1	0.765361	1.484322	848.389398	A (x1)
1	0.471363	2.227958	1943.984931	E (x5)
1	0.395810	2.381853	2225.534947	C (x3)
1	0.377170	2.418317	2295.000669	G (x7)
1	0.356763	2.457615	2371.046612	H (x8)
1	0.350940	2.468714	2392.745524	I (x9)
1	0.347381	2.475473	2406.008366	F (x6)
1	0.235375	2.679494	2823.404786	J (x10)
1	0.126059	2.864636	3230.773865	D (x4)
1	0.065507	2.962214	3456.422815	B (x2)
2	0.976958	0.4736817	61.867264	BE
2	0.826587	1.29947	622.230164	AE
2	0.823812	1.309826	632.570989	AC
3	0.985724	0.3799527	31.201415	AEF
3	0.980175	0.4477463	51.880208	BCE
3	0.978947	0.4614055	56.456614	BDE
4	0.988579	0.3465689	22.560822	AEFH
4	0.988347	0.3500752	23.426346	ABEF
4	0.987601	0.3611051	26.205962	ACEF
5	0.991966	0.2966695	11.939729	ACDEF
5	0.990822	0.3170781	16.200645	ABEFH
5	0.990702	0.3191541	16.649960	ACEFH
6	0.993525	0.2720509	8.127863	ACDEFH
6	0.993080	0.2812623	9.789421	ABCEFH
6	0.992527	0.29228	11.849453	ABDEFH
7	0.994075	0.2660938	8.079170	ABCDEFH
...				
8	0.994333	0.266362	9.118089	ABCDEFHJ
...				
10	0.994901	0.2656163	11.000000	ABCDEFGHIJ

Ve výstupu z programu All possible si povšimněme nejlepšího modelu se dvěma regresory $x2$ a $x5$, který má výrazně vyšší R^2 než ostatní modely se dvěma regresory a přibližuje se hodnotám R^2 s větším počtem regresorů. Přitom kroková procedura tento model „minula“. To je dosti názorná ilustrace nevýhod stepwise procedur, které jsou sice výpočetně méně náročné než úplné prohledávání, ale za cenu rizika takového minutí vhodného modelu.

Mallowsovo C_p má nejmenší hodnotu pro model se sedmi regresory, jen o málo je C_p větší pro nejlepší model se šesti regresory. Všimněme si, že tento nejlepší model se šesti regresory není shodný s tím, který byl nalezen krokovou procedurou, na místo regresoru $x2$ je zařazen regresor $x4$. Vidíme, že procedura All possible nám nabízí více kandidátů na vhodný model než procedura Stepwise. Mezi těmito kandidáty je nutno pečlivě vybírat, rozhodně není jediný vhodný model s minimálním C_p. Pro výběr vhodných modelů jsou užitečná i grafická zobrazení statistik pro nalezené modely proti počtu regesorů. Jako příklad uvádíme grafy pro index determinace a Mallowsovo C_p, na kterých je jasně vidět výrazný skok v hodnotách statistik pro nejlepší model se dvěma regresory. Podobně je užitečný i graf závislosti residuální směrodatné odchylky.





Opravdu vhodný model je však možno doporučit až po podrobnější analýze a porovnání jednotlivých kandidátů. Jak kroková procedura, tak All possible regressions nám jen generují návrhy, které je nutno podrobněji analyzovat.

Shrnutí

- *výběr regresorů do modelu*
- *kroková (stepwise) regrese*
- *hledání nejlepší množiny regresorů*
- *kriteria pro posouzení vhodnosti modelu*

Kontrolní otázky

1. *Vysvětlete principy a algoritmus krokové regrese.*
2. *Proč maximalizace R^2 není dobrou strategií při hledání vhodného modelu?*
3. *Proč minimalize Mallowsovy statistiky je přijatelnou strategií při hledání vhodného modelu?*
4. *Podle čeho se posuzuje, který model je vhodný pro danou úlohu?*
5. *Porovnejte výhody a nevýhody krokové regrese a prohledávání všech modelů.*

Korespondenční úloha

Korespondenční úlohy budou zadávány ke každému kursu samostatně.