Model regresji wielokrotnej—Wykład 14 (2.06.2008)

Przykład — ceny domów w Chicago

Poniżej są przedstawione dane dotyczące cen domów w Chicago (źródło: Sen, A., Srivastava, M., Regression Analysis, Springer, 1990, str. 32).

	Price	BDR	FLR	FP	RMS	ST	LOT	TAX	BTH	CON	GAR	CDN	L1	L2
1	53	2	967	0	5	0	39	652	1.5	1	0.0	0	1	0
2	55	2	815	1	5	0	33	1000	1.0	1	2.0	1	1	0
3	56	3	900	0	5	1	35	897	1.5	1	1.0	0	1	0
							...							
25	50	2	691	0	6	0	30	549	1.0	0	2.0	1	0	0
26	65	3	1023	0	7	1	30	900	2.0	1	1.0	0	1	0

Oznaczenia zmiennych

Price: cena (w tys. dolarów);

BDR: liczba pokoi sypialnych (*bedrooms*);

FLR: powierzchnia w stopach kw.;

FP: liczba kominków (*fireplaces*);

RMS: liczba pokoi;

ST: okna sztormowe (1- jeśli są; 0- jeśli nie ma);

LOT: szerokość działki (w stopach);

TAX: podatki (roczne);

BTH: liczba łazienek;

CON: konstrukcja (1- jeśli z cegieł; 0- w przeciwnym przypadku);

GAR: (0- żadnego garazu; 1- garaż na 1 sam.; 2- garaże na 2 sam.)

CDN: stan (1— potrzeba remontu; 0 — nie potrzeba);

L1: położenie (L1=1 jeśli dom położony w str. A; 0 w przec. przyp.);

L2: położenie (L2=1 jeśli dom położony w str. B; 0 w przec. przyp.).

Zależność ceny od powierzchni

Można próbować znaleźć zależność pomiędzy FLR (zmienną niezależną) i Price (zmienną zależną).

```
> data(E2.2)  # zbiór E2.2 w pakiecie SenSrivastava
> l1<-lm(Price~FLR)
> summary(l1)

Call:
lm(formula = Price ~ FLR)

Residuals:
     Min       1Q   Median       3Q      Max
-21.8986  -4.6090   0.8269   5.4978  13.3056

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.916831   4.510747   7.519 9.29e-08 ***
FLR          0.020213   0.003796   5.325 1.83e-05 ***
---
Sig. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.694 on 24 degrees of freedom
Multiple R-Squared: 0.5416,     Adjusted R-squared: 0.5225
F-statistic: 28.35 on 1 and 24 DF,  p-value: 1.830e-05
```

— wsółczynnik determinacji (*Multiple R-Squared*) jest równy $0{,}5416$ — czyli prosta MNK o równaniu $y = 0{,}020213x + 33{,}916831$ wyjaśnia 54 procent zmienności w naszych danych.

Problem

czy można znależc zależność "regresyjną" zmiennej `Price` od większej liczby zmiennych niezależnych (`BDR, FLR, FP,` itd.) — np. zależność postaci:

$$Price = \beta_0 + \beta_1 BDR + \beta_2 FLR + \beta_3 FP + \epsilon$$

Normalny model regresji wielokrotnej

$$Y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{im}\beta_m + \epsilon_i, \tag{1}$$

gdzie: ϵ_i ma rozkład normalny $N(0, \sigma)$, gdzie $\sigma > 0$ jest nieznanym parametrem;
zmienne $\epsilon_1, \ldots, \epsilon_n$ sa niezależne;
$\beta_0, \beta_1, \ldots, \beta_m$ są nieznanymi, stałymi liczbami (parametrami);
współczynniki

$$x_{ij}, \quad i = 1, \ldots, n, \quad i = 1, \ldots, m.$$

są znane. Zakładamy, że $n > m + 1$.
Oczywiście, Y_i są zmiennymi losowymi o rozkładzie normalnym $N(\mu, \sigma)$, gdzie

$$\mu = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{im}\beta_m.$$

Zapis macierzowy

Niech $X_0 = (x_{ij})_{i=1,\ldots,n;j=1,\ldots,m}$ oraz

$$X = [1_n | X_0],$$

gdzie 1_n oznacza n-wymiarowy wektor kolumnowy, którego wszystkie współrzędne sa równe 1,

$$\beta = (\beta_0, \beta_1, \ldots, \beta_m)'$$

oraz

$$\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'.$$

Model (1) może być zapisany w postaci:

$$Y = X\beta + \epsilon.$$

Uwaga. Y, X, β można też zapisać w postaci:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \vdots & x_{nm} \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{2}$$

Estymacja wektora β

Załóżmy, że rząd macierzy X jest równy $m+1$, tj. że X jest pełnego rzędu kolumnowego. Można udowodnić, że w tym przypadku macierz $X^{'}X$ jest nieosobliwa. Wektor $b = (b_0, b_1, \ldots, b_m)^{'}$ określony przez:

$$b = (X^{'}X)^{-1}X^{'}Y$$

jest estymatorem nieobciązonym wektora β posiadającym szereg atrakcyjnych własności teoretycznych.

Testowanie hipotez

Analogicznie jak w przypadku modelu regresji z jedną zmienną niezależną, w naszym modelu można testować hipotezy (korzystając z faktu, że odpowiednie statystyki mają rozkład t-Studenta):

$$H_0 : \beta_0 = 0 \quad \text{przeciw} H_1 : \beta_0 \neq 0 \tag{3}$$

$$H_0^{'} : \beta_1 = 0 \quad \text{przeciw} H_1^{'} : \beta_1 \neq 0 \tag{4}$$

$$H_0^{''} : \beta_2 = 0 \quad \text{przeciw} H_1^{''} : \beta_2 \neq 0 \quad \text{itd.} \tag{5}$$

oraz hipotezę

$$\bar{H}_0 : \beta_1 = 0 \wedge \beta_2 = 0 \wedge \ldots \wedge \beta_m = 0 \quad \text{przeciw: } \bar{H}_1 : \text{istnieje } i > 0 \text{ takie, że } \beta_i \neq 0 \tag{6}$$

(w tym przypadku statystyka testowa ma rozkład F z odpowiednią liczbą stopni swobody).

Ceny domów w Chicago — przykład obliczeń

Poniżej znajdują się wyniki obliczeń w środowisku R dla modelu, w którym zmiennymi niezależnymi są: FLR,RMS,BDR,BTH,GAR,LOT,FP i ST.

```
> l2<-lm(Price~FLR+RMS+BDR+BTH+GAR+LOT+FP+ST)
> summary(l2)

Call:
lm(formula = Price ~ FLR + RMS + BDR + BTH + GAR + LOT + FP +
    ST)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3058  -2.8417  -0.1511   3.2882   7.9518

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 18.637664   5.240957   3.556 0.002429 **
FLR          0.017570   0.003235   5.431 4.49e-05 ***
RMS          3.904374   1.615617   2.417 0.027194 *
BDR         -7.697444   1.829426  -4.208 0.000592 ***
BTH          2.374591   2.557865   0.928 0.366221
GAR          1.770861   1.404310   1.261 0.224334
LOT          0.263522   0.135109   1.950 0.067808 .
FP           6.909765   3.083583   2.241 0.038680 *
ST          10.818663   2.300203   4.703 0.000205 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.717 on 17 degrees of freedom
Multiple R-Squared: 0.9044,     Adjusted R-squared: 0.8595
F-statistic: 20.11 on 8 and 17 DF,  p-value: 3.147e-07
```

Z obliczeń wynika, że włączenie do modelu zmiennych BTH, GAR i LOT

jest problematyczne (p-wartości odpowiadające tym zmiennym są duże). Można oczekiwać, że w modelu ze zmiennymi niezależnymi: FLR, RMS, BDR, FP i ST żadnej z hipotez odpowiadających tym zmiennym nie będzie można odrzucić na poziomie istotności $\alpha = 0{,}05$.

P-wartość odpowiadająca hipotezie (6) jest równa $3.147e-07$ — wskazuje ona na to, że przynajmniej niektóre spośród współczynników $\beta_1, \ldots, \beta_m$ są różne od zera.

Wsółczynnik determinacji (*Multiple R-Squared*) jest równy $0{,}9044$ — a więc nasza zależność regresyjna w dużym stopniu objaśnia zmienność danych.

Ceny domów w Chicago— c.d.

Korzystając z powyższych obliczeń: dla domu dla którego wartości zmiennych:
FLR,RMS,BDR,BTH,GAR,LOT,FP i ST
są równe $x_{FLR}, x_{RMS}, x_{BDR}, x_{BTH}, x_{GAR}, x_{LOT}, x_{FP}$ i x_{ST}, wartość oczekiwana (średnia) ceny Price jest równa:

$$\begin{aligned} E(Price) = 18{,}637664 &+ 0{,}017570 x_{FLR} + 3{,}904374 x_{RMS} \\ &- 7{,}697444 x_{BDR} + 2{,}374591 x_{BTH} + 2{,}374591 x_{GAR} \\ &+ 0{,}263522 x_{LOT} + 6{,}909765 x_{FP} + 10{,}818663 x_{ST} \end{aligned}$$

Uwaga 1. Z powyższego równania wynika np. że zwiększenie powierzchni domu o 1 stopę kw. powoduje zwiększenie ceny domu o
$0{,}017570 \times 1000 \sim 18$ USD.

Uwaga 2. Z równania wynika, że zwiększenie liczby pokoi sypialnych o jeden — bez zmiany pozostałych „parametrów” — spowoduje spadek wartości domu o $7{,}697444 \times 1000 \sim 7697$ USD. Interpretacja tego zaskakującego faktu jest następująca: przy mniejszej liczbie pokoi sypialnych powierzchnia domu jest zajęta przez bardziej „kosztowne” pomieszczenia, takie jak łazienki.

Uwaga 3. Do naszego „równania regresyjnego” nie włączyliśmy zmiennych CON, CDN, L1 i L2. Zmienne te mogą przyjmować tylko dwie wartości — są to tzw. zmienne indykatorowe. Użycie takich zmiennych w budowaniu modeli regresyjnych zostało omówione np. w rozdz. 4 książki Sena i Srivastavy.

Polecana literatura

Koronacki, J., Mielniczuk, J. Statystyka dla studentów kierunków technicznych i przyrodniczych, WNT 2001, rozdz. 4.3.

Kala, R., Statystyka dla przyrodników, Wyd. Akademii Rolniczej w Poznaniu, Poznań 2002.