Linearni regrese

Mate k dispozici fragment z vystupu linearni regrese:

Dependent
Regression Equation Section

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)
Intercept	2,6461	0,6080
x1	0,0921	0,0109
x2	-0,0604	0,0145
x3	-0,0581	0,0105

Analysis of Variance Section

Source	DF	Sum Of Squares	Mean Square
Model	3	134,7	44,9
Error	30	15,3	0,51

a. Napiste tvar regresniho modelu

$$\hat{y} = 2,6461 + 0.0921 \cdot x_1 - 0.0604 \cdot x_2 - 0.0581 \cdot x_3$$

b. Jaky byl rozsah vyberu (kolik radku v datove matici)?

$$n - k - 1 = Error$$

$$n = Error + k + 1 = Error + pocetpromennych + 1$$

$$n = 30 + 3 + 1 = 34$$

c. Lze zamitnout hypotezy, ze regresni koeficienty jsou nulove?

t-statistika:
$t_0 = |\frac{b_0}{Sb_0}| = |\frac{2.6461}{0.6080}| = 3.72$
$t_1 = |\frac{b_1}{Sb_1}| = |\frac{0.0921}{0.0109}| = 8.45$
$t_2 = |\frac{b_2}{Sb_2}| = |\frac{-0.0604}{0.0145}| = 4.17$
$t_3 = |\frac{b_3}{Sb_3}| = |\frac{-0.0581}{0.0105}| = 5.53$
kriticka hodnota z tabulek pro $df = 30$ a $\alpha = 0.05$ je 2.042
hypotezu o nulovosti koeficentu lze zamitnout

d. Spocitejte koeficent determinace

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{RSS}{MSS + RSS} = 1 - \frac{15.3}{134.7 + 15.3} = 0.88$$

e. Odhadnete rozptyl nahodne slozky modelu

$$s^2 = \frac{RSS}{n - k - 1} = \frac{15.3}{30} = 0.51$$

d. Je model uspokojivy pro pro vysvetleni zavislosti Y na regresorech?

F-statistika:
$f_{krit} = 2.92$

$$F = \frac{MMS}{RMS} = \frac{44.9}{0.51} = 88$$

$F \gg f \implies$ model je statisticky vyznamny

Logisticka regrese

V logistickem regresnim modelu vyjadrujicim zavislost poskozeni sdravi (velicina POSKOZENI ma hodnotu 0, 1) na pobytu v rizikovem prostredi (velicina RIZIKO ma hodnoty 0, 1) a na intenzite rehabilitace (velicina REHAB, spojita s kladnymi hodnotami) ma nasledujici vysledky:

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	p
Intercept	4.175	1.883	0.02662
rehab	-0.365	0.125	0.00357
riziko	1.713	0.764	0.02494

a. Napiste tvar regresniho modelu

pro $\alpha = 0.05$:

$$ln(\frac{p}{1-p}) = 4.175 - 0.365 \cdot rehab + 1.713 \cdot riziko$$

b. Meni se pravdepodobnost poskozeni zdravi intenzitou rehabilitace? Pokud ano, tak jak?

pavdepodobnost klesa, protoze koeficient je zaporny

c. Urcete 95%ni interval spolehlivosti pro pomer sanci(odds ratio) poskozeni zdravi vystavenim riziku. Interpretujte tento vysledek.

Bodovy odhad pro koureni ($\beta = 1.713$):

$$OR = e^{\beta} = e^{1.713} = 5.546$$

Interval spolehlivosti:

$$meze = \beta \pm z_{\alpha/2} \times S_i = 1.713 \pm (1.96 \times 0.764)$$

$$meze = [0.2156, 3.2104]$$

transformace na OR:

$$OR_{dolni} = e^{0.2156} = 1.241$$

$$OR_{horni} = e^{3.2104} = 24.789$$

Vysledek:

$$95\% IS = (1.241; 24.789)$$

Interpretace:

- interval neobsahuje 1, takze muzeme povazovat faktor ma vyznamny pozitivni efekt na sanci vyskytu udalosti
- skutecna hodnota OR lezi s 95% jistotou v tomto intervalu, faktor zvysuje sanci minimalne 1,24x a maximalne 24,79x
- interval je siroky, mozna malo dat?

d. Napiste klasifikacni pravidlo umoznujici podle hodnot roky a koureni klasifikovat osoby do kategorie nemocen/zdrav

$$p = \frac{1}{1 + e^{-x^T\beta}} = \frac{1}{1 + e^{-1(4.175 - 0.365 \cdot rehab + 1.731 \cdot riziko)}}$$

pokud $p \geq c$ klasifikujeme jako NEMOCEN

PCA

Vyberova korelacni matice ma nasledujici vlastni cisla

No.	Eigenvalue
1	6.097632
2	1.225727
3	0.246754

No.	Eigenvalue
4	0.161526
5	0.102103
6	0.097033
7	0.056346
8	0.012879

Co z toho muzete ususzovat? Jaky rozmer ma korelacni matice a jaky je spravny rozmer matice dat?

korelacni matice ma rozmer 8×8
matice dat ma rozmer $n \times 8$

kde je $\lambda \approx 0$ tam je temer linearni zavislost
pokud vicekrat $\implies$ multikolinearita

efektivni dimenze = 2, prvni dve komponenty

$$K = \frac{\lambda_{max}}{\lambda_{min}} = \frac{6.097632}{0.012879} = 473$$

$K > 100 \implies$ multikolinearity

Jakou cast celkoveho rozptylu vysvetluji prvni dve hlavni komponenty?

$$\frac{6.097632 + 1.225727}{8} = \frac{7.323359}{8} = 0.915 = 91.5\%$$

Diskriminacni analyza

Z nasledujiciho vystupu zformulujte klasifikacni pravidlo a vypoctete procento spravne klasifikace. Napiste predpoklady potrebne pro vyuziti linearni diskriminacni funkce.

Linear Discriminant
Functions

	skup	
Variable Constant	1	2
Constant	-22.840	-47.031
x1	2.575	3.638
x3	1.151	1.711

Classification Count Table for skup

	Predicted		
Actual	0	1	Total
0	23	1	24
1	1	25	26
Total	24	26	50

Klasifikacni pravidlo

$$LDF_1 = -22.840 + 2.575 \times x_1 + 1.151 \times x_3$$

$$LDF_2 = -47.031 + 3.638 \times x_1 + 1.711 \times x_3$$

$LDF_1 > LDF_2 \implies skupina1$
$LDF_2 > LDF_1 \implies skupina2$
$LDF_1 = LDF_2 \implies nahodne$

Nebo jednoduseji pro dve skupiny:

$$(-22.840 - (-47.031)) + x_1 \times (2.575 - 3.638) + x_3 \times (1.151 - 1.711) > 0 \implies skupina1$$

Procento spravne klasifikace

$$\frac{spravne}{celkem} = \frac{23 + 25}{50} = 0.96 = 96\%$$

Podminky pro vyuziti LDF:

nahodny vektor ma normalni rozdeleni a skupiny se lisi jen vektorem strednich hodnot

Matice

$$a = [1, 1, 1]^T$$

$$x = [x_1, x_2, x_3]^T$$

$$B = \begin{bmatrix} 1 & -2 & 3 \\ -2 & 3 & 3 \\ 3 & 3 & 2 \end{bmatrix}$$

a. $a^T x$

$$a^T x = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1 + x_2 + x_3$$

b. ax^T

$$ax^T = \begin{bmatrix}1\\1\\1\end{bmatrix}\begin{bmatrix}x_1 & x_2 & x_3\end{bmatrix} = \begin{bmatrix}x_1 & x_2 & x_3\\x_1 & x_2 & x_3\\x_1 & x_2 & x_3\end{bmatrix}$$

c. Ba

$$Ba = \begin{bmatrix}1 & -2 & 3\\-2 & 3 & 3\\3 & 3 & 2\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix} = \begin{bmatrix}2\\4\\8\end{bmatrix}$$

Matice

$$a = [1,1,1]^T$$

$$B = \begin{bmatrix}1 & 2 & 3\\0 & 1 & 1\\1 & 3 & 2\end{bmatrix}$$

c. Ba

$$Ba = \begin{bmatrix}1 & 2 & 3\\0 & 1 & 1\\1 & 3 & 2\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix} = \begin{bmatrix}6\\2\\6\end{bmatrix}$$

Predpoklady klasickeho linearniho modelu:

- stredni hodnota nahodne slozky nulova
- nahodne slozky jsou nekorelovane
- matice dat X je nenahodna matice
- sloupce matice jsou linearne nezavisle, tedy $h(X) = k + 1 \leq n$

Zapiste maticove linearni regresni model

$$y = X\beta + \epsilon$$

y: $n \times 1$
ϵ: $n \times 1$
X: $n \times (k+1)$
β: $(k+1) \times 1$

Projekcni matice:

$$H = X(X^TX)^{-1}X^T$$

$$\hat{y} = Hy$$

promita y do roviny dane regresory
H: $n \times n$

co znamenaji hodnoty diagonalnich prvku:
Jak velky vliv ma dane pozorovani na regresni model

Rozsah: $0 \geq h_{ii} \geq 1$

Soucet diagonalnich prvku $\sum h_{ii} = k + 1$