14-3-2023

Μοντέλα Παλινδρόμησης με Κατηγορικές Ανεξάρτητες Μεταβλητές.

ANOVA

Y ποσοτική

$X_1, \dots, X_k$: κατηγορικές μεταβλητές

ANOVA Method

REGRESSION

παλινδρόμηση

Y ποσοτική

$X_1, \dots, X_k$ ποσοτικές

ANACOVA

Y ποσοτική

$\underbrace{X_1, \dots, X_k}_{\text{ποσοτικές}}$ $\underbrace{X_{k+1}, \dots, X_p}_{\text{κατηγορικές}}$

Ενιαίο Μοντέλο

$$Y = b_0 + b_1 X_1 + \dots$$

Μεταβλητές

1) Κατηγορικές (Nominal) — Επίπεδα

π.χ. X = χρώμα
{ άσπρο, μαύρο, γκρι, καφέ }

2) Ordinal (Διατεταγμένες)

π.χ. Likert scale
{ λίγο, πολύ, πάρα πολύ }

3) Ποσοτικές / Κλίμακας (Scale)
έχουν μονάδα μέτρησης } παλινδρόμηση.

Παράδειγμα

Y = Τιμή πώλησης διαμερίσματος

Πόλη : { Bos, Chi, NY }

City	Price
Chi	150
Bos	200
Chi	120
NY	300
Chi	.
NY	
:	:

Έστω

$$Y_{NY} \sim \mathcal{N}(\mu_{NY}, \sigma^2)$$

$$Y_{BOS} \sim \mathcal{N}(\mu_{BOS}, \sigma^2)$$

$$Y_{CHI} \sim \mathcal{N}(\mu_{CHI}, \sigma^2)$$

πολλαπλοί έλεγχοι

$H_0 : \mu_{NY} = \mu_{CHI} \qquad H_1 : \mu_{NY} \neq \mu_{CHI}$

$H_0 \quad \mu_{NY} = \mu_{BOS} \qquad H_1 : \mu_{NY} \neq \mu_{BOS}$

H_0 : CHI − BOS

1ος P("σωστό") : 0.95

2ος P " : 0.95

P (...) : 0.95

P(όλα σωστά) $= 0.95 \times 0.95 \times 0.95 < 0.95$.

Θα θέλαμε έναν έλεγχο

$H_0: \mu_{NY} = \mu_{Bos} = \mu_{CHI}$: H_1: τουλάχιστον δύο διαφορετικά μεταξύ τους

Κωδικοποίηση μιας κατηγορικής μεταβλητής.

(1)

City	X
NY	= 1
Chi	= 2
Bos	= 3

$\Rightarrow$ $\boxed{Y = \beta_0 + \beta_1 X + \varepsilon}$

$E(Y/NY) =$

$E(Y/X=1) = \beta_0 + \beta_1$

dataset

X	Y
1	150
2	300
3	200
1	:
3	:
2	:

$$E(Y|NY) = E(Y|X=1) = \beta_0 + \beta_1 = \mu_{NY}$$
$$E(Y|Chi) = E(Y|X=2) = \beta_0 + 2\beta_1 = \mu_{CHI}$$
$$E(Y|Bos) = E(Y|X=3) = \beta_0 + 3\beta_1 = \mu_{Bos} \Rightarrow$$

$$\mu_{CHI} - \mu_{NY} = \beta_1$$
$$\mu_{Bos} - \mu_{CHI} = \beta_1$$
$$\mu_{Bos} - \mu_{NY} = 2\beta_1$$

$\Rightarrow \boxed{\mu_{CHI} - \mu_{NY} = \mu_{Bos} - \mu_{CHI} = \frac{1}{2}(\mu_{Bos} - \mu_{NY})}$

(2)

Bos	1
NY	2
Chi	3

$Y = b_0 + b_1 X + \varepsilon$

$\Rightarrow$ - - -

$$\boxed{\mu_{NY} = \mu_{BOS} = \mu_{CHI} = \mu_{NY} = \frac{1}{2}(\mu_{CHI} - \mu_{BOS})}$$

---

(3)

Bos	-5
NY	4
Chi	7

$\Rightarrow$ - - - -

"Έυχη" Κωδικοποίηση (Δίτιμες (binary) μεταβλητές)

City | 3 επίπεδα: Bos, Chi, NY → 3 νέες μεταβλητές

$$X_{NY} = \begin{cases} 1 & \text{αν City} = NY \\ 0 & \text{αν City} \neq NY \end{cases}$$

$$X_{Bos} = \begin{cases} 1, & \text{City} = Bos \\ 0, & \text{City} \neq Bos \end{cases}$$

$$X_{CHI} = \begin{cases} 1, & \text{City} = Chi \\ 0, & \text{city} \neq Chi \end{cases}$$

$$\boxed{X_{NY} + X_{Bos} + X_{CHI} = 1}$$

$$Y = b_0 + b_1 X_{NY} + b_2 X_{CHI} + b_3 X_{Bos}$$

4 παράμετροι > 3 !!

Y	City	X_{NY}	X_{CHI}	X_{Bos}
	Chi	0	1	0
	Chi	0	1	0
	Bos	0	0	1
	NY	1	0	0
	NY	1	0	0
	Bos	0	0	1

⋮

City	X_{NY}	X_{CHI}	X_{BOS}
Chi	0	1	
Chi	0	1	
Bos	0	0	
NY	1	0	
NY	1	0	
Bos.	0	0	
⋮			

X_{BOS} → παραλ. Bos επίπεδο αναφοράς

$$Y = b_0 + b_1 X_{NY} + b_2 X_{CHI}$$ ← 3 παραμ.

Κωδικοποίηση

Γενικά: Για κάθε επίπεδο εκτός του επιπέδου αναφοράς ορίζουμε μια δίτιμη μεταβλητή

Έστω το ~~μοντέλο~~

$$Y = b_0 + b_1 X_{NY} + b_2 X_{CHI}$$

Ερμηνείες των β.

Για κάθε επίπεδο υπολογίζουμε τη $E(Y)$ με βάση το μοντέλο

City : Bos $\quad X_{NY}=0 \quad X_{CH_1}=0$

$$Y = b_0 \Rightarrow E(Y|Bos) = \boxed{b_0 = \mu_{Bos}}$$

City	Y
Bos	$\mu_{Bos} = \beta_0$
NY	$\mu_{NY} = \beta_0 + \beta_1$
Chi	$\mu_{CH_1} = \beta_0 + \beta_2$

$Y = \beta_0 + \beta_1 NY + \beta_2 Chi$

$\beta_0 = \mu_{Bos}$

$\beta_1 = \mu_{NY} - \beta_0 = \mu_{NY} - \mu_{Bos}$

$\beta_2 = \mu_{CH_1} - \mu_{Bos}$

Γενικά

$\beta_0 = E(Y | X : \text{επίπεδο αναφοράς})$

$\beta_1 = E(Y | X = \text{επίπεδο } 1) - E(Y | \text{αναφ.})$

$\beta_2 = E(Y | X = \text{επιπ. } 2) - E(Y | \text{αναφ.})$

$\vdots$

Ftest για συνολικό μοντέλο

$H_0 : \beta_1 = \beta_2 = 0 \qquad H_1 : \beta_1 \neq 0$ ή $\beta_2 \neq 0$ ή κ' τα δύο

$$H_0 : \begin{array}{l} \beta_1 = 0 \Rightarrow \mu_{NY} = \mu_{Bos} \\ \beta_2 = 0 \Rightarrow \mu_{CH_1} = \mu_{Bos} \end{array} \Bigg\} \Rightarrow \boxed{\mu_{NY} = \mu_{CH_1} = \mu_{Bos}}$$

$\Longleftrightarrow \quad H_0 : \mu_{NY} = \mu_{CHI} = \mu_{BOS}$

$$Y = \beta_0 + b_1 X_{NY} + b_2 X_{CHI}$$

$\beta_0 = 201.636$

$\beta_1 = 56.864$

$b_2 = -10.636$

$\mu_{BOS} = 201.636$

$\mu_{NY} = 201.636 + 56.864 = 258.5$

$\mu_{CHI} = 201.636 - 10.636 = 191$

συμφωνούν με τις εκτιμήσεις των $\mu_{NY}, \mu_{BOS}, \mu_{CHI}$.

F-test : $p = 7 \cdot 10^{-8}$

υπάρχει συσχέτιση ανάμεσα σε πόλη κ' τιμή διαμερίσματος

Model2 : $Y = b_0 + b_1 X_{BOS} + b_2 X_{CHI}$ (NY : επίπ. αναφ.)

$\hat{b}_0 = \hat{\mu}_{NY} = 258.5$

$\hat{b}_1 = -56.87 \Rightarrow \hat{\mu}_{BOS} = 258.5 - 56.87 = 201.63$

$\hat{b}_2 = -67.5 \Rightarrow \hat{\mu}_{CHI} = 258.5 - 67.5 = 191$

$H_0 : b_1 = b_2 = 0 \qquad H_1 \; \ldots \qquad F = 30.54$

$p = 7 \cdot 10^{-8}$

Μοντέλα με ≥ 2 κατηγορικές μεταβλητές

X_1 k_1 επίπεδα
X_2 k_2 "
$\vdots$
X_p k_p επίπεδα

Παράδειγμα phys.Rdata

sex	smoking	Y (score)
M	Y	167
F	N	150
F	N	.
F	Y	.
M	N	.

4 παραμ.

$\mu_{MN} = E(Y \mid \text{male, no smoking})$
μ_{MY}
μ_{FN}
μ_{FY}

Sex (2 levels) Έστω F: reference level

$$X_M = \begin{cases} 1 & \text{if sex} = M \\ 0 & \text{if sex} \neq M \end{cases}$$

Smoking (2 levels) N: reference level

$$X_Y = \begin{cases} 1 & \text{if smoking} = \text{Yes} \\ 0 & \neq \text{Yes} \end{cases}$$

$$Y = b_0 + b_1 X_M + b_2 X_Y$$ [3 παραμ.]

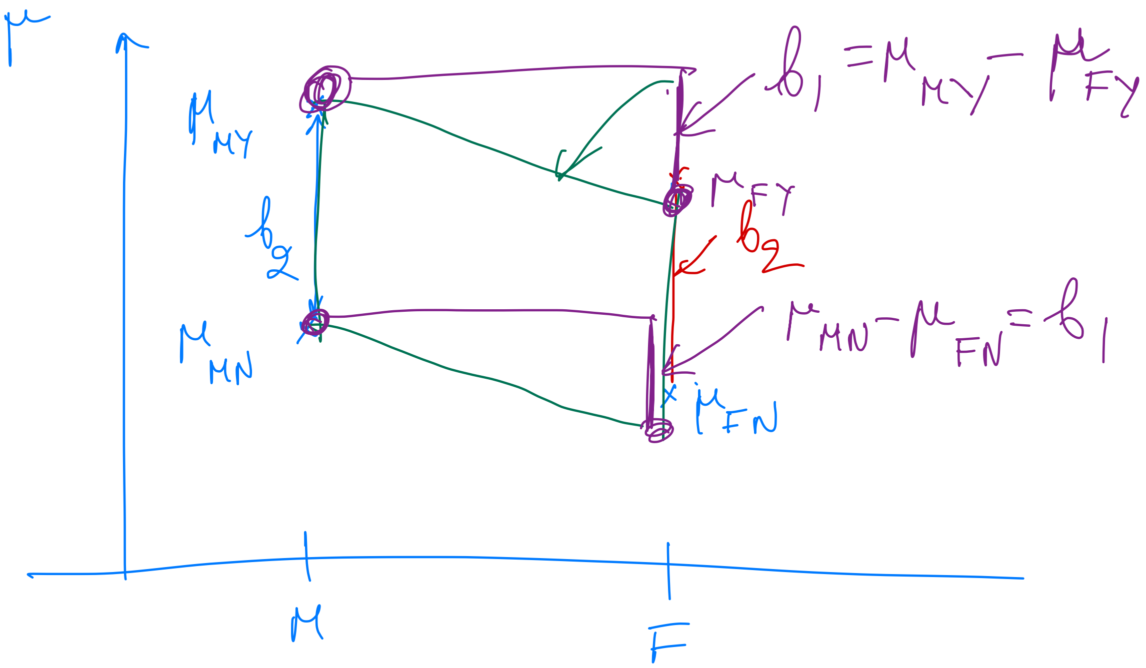
Sex	Smoking	X_M	X_Y	Y
Fem	No	0	0	$\mu_{FN} = b_0$
Fem	Yes	0	1	$\mu_{FY} = b_0 + b_2$
Male	No	1	0	$\mu_{MN} = b_0 + b_1$
Male	Yes	1	1	$\mu_{MY} = b_0 + b_1 + b_2$

$$\hat{b}_0 = \hat{\mu}_{FN}$$

$$\hat{b}_1 = \mu_{MN} - \mu_{FN} = \mu_{MY} - \mu_{FY}$$

$$\hat{b}_2 = \mu_{FY} - \mu_{FN} = \boxed{\mu_{MY} - \mu_{MN}}$$

ισοδύναμες



Effect φύλου για N : $\mu_{MN} - \mu_{FN}$

'' '' για Y : $\mu_{MY} - \mu_{FY}$

Το μοντέλο υποθέτει ίσα

Αντίστοιχα για το effect του καπνίσματος

$\hat{b}_1$: effect sex (ίδιο για smoking Y or N)

$\hat{b}_2$: effect smoking (ίδιο για sex M, F)

main effects

κύριες επιδράσεις

$$Y = b_0 + b_1 X_1 + b_2 X_2$$: main effects model.