

14-1-2026

Γραμμικά Μοντέλα Με κατηγορικές μεταβλητές

Κατηγορίες , Μεταβλητών

1) X : κατηγορική (qualitative)

πάρει τιμές σε ένα διακριτό σύνολο

πχ. $\{R, G, B\}$ (χρώμα)
 $\{No, Yes\}$

2) X : διάταξη (ordinal)

υπάρχει διάταξη μεταξύ τους

πχ. Likert scale (καθόλου, λίγο, αρκετά, ...)

3) X : κλίμακας/ποσοτικές (scale/quantitative)

έχει μονάδα μέτρησης

Σε μοντέλα παλινδρόμησης $\begin{cases} Y \\ X_1, \dots, X_k \end{cases}$ scale variables

Αν σε πραγματικά δεδομένα υπάρχουν παράγοντες
που δεν είναι ποσοτικές για να ενταχθούν

σε ένα μοντέλο παλινδρόμησης χρειάζεται

κατάλληλη κωδικοποίηση

πχ. t-test για σύγκριση ΕΥ μεταξύ Μ & F

Πώς γενικεύεται αυτή η μεθοδολογία
αν έχουμε > 2 κατηγορίες

Παράδειγμα Apartments dataset.

Y = τιμή πώλησης διαμερίσματος

σε 3 πόλεις $\begin{cases} \text{CHI} \\ \text{BOS} \\ \text{NY} \end{cases}$

Μέθοδος ANOVA (Ανάλυση Διασποράς)

$$H_0 : \mu_{\text{BOS}} = \mu_{\text{CHI}} = \mu_{\text{NY}}$$

H_1 : τουλάχιστον
δύο διαφέρουν
μεταξύ τους

3 άγνωστοι παράμετροι ($\mu_{\text{BOS}}, \mu_{\text{CHI}}, \mu_{\text{NY}}$)

Μοντέλο

$$Y | \text{Chi} \sim \mathcal{N}(\mu_{\text{CHI}}, \sigma^2) \Leftrightarrow Y | \text{Chi} = \mu_{\text{CHI}} + \varepsilon$$

$$Y | \text{Bos} \sim \mathcal{N}(\mu_{\text{BOS}}, \sigma^2) \quad Y | \text{BOS} = \mu_{\text{BOS}} + \varepsilon$$

$$Y | \text{NY} \sim \mathcal{N}(\mu_{\text{NY}}, \sigma^2) \quad Y | \text{NY} = \mu_{\text{NY}} + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$\left[\begin{array}{l} \text{ομο-} \\ \text{σκε-} \\ \text{δαση} \end{array} \right]$

$X = \text{City}$

$$Y = f(X) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$f(x) = \begin{cases} \mu_{\text{CHI}} & , X = \text{Chi} \\ \mu_{\text{BOS}} & , X = \text{Bos} \\ \mu_{\text{NY}} & , X = \text{NY} \end{cases}$$

"συνάρτηση
παραπρόσμεσης"

Αντιστοιχεί σε γραμμικό μοντέλο?

$$f(x) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 \dots ?$$

$$\bar{y}_{i \cdot k}$$

$$\bar{y}_{\cdot j k}$$

$$\bar{y}_{i \cdot \cdot}$$

Μοντέλο ANOVA : Ειδική περίπτωση γραμμικού μοντέλου
Ποια είναι η κατάλληλη κωδικοποίηση?

Στο dataset apart1.Rdata

$$\hat{\mu}_{CHI} = 191, \quad \hat{\mu}_{BOS} = 201, \quad \hat{\mu}_{NY} = 259$$

? $H_0: \mu_{CHI} = \mu_{BOS} = \mu_{NY}$ H_1 τουλάχιστον 2 διαφορετικά

3 άγνωστες παράμετροι ($\mu_{BOS}, \mu_{CHI}, \mu_{NY}$)
 $df_{er} = n - 3$ Αφαιρούμε 3 df για εκτίμηση σ^2

Πως κωδικοποιείται η $X = city$

Π.χ.

$$CHI = 1$$

$$BOS = 2$$

$$NY = 3$$

$$X' = \begin{cases} 1 & \text{αν } X = CHI \\ 2 & \text{αν } X = BOS \\ 3 & \text{αν } X = NY \end{cases}$$

$$\text{Erm } Y = b_0 + b_1 X' + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$df_{\text{er}} = n - 2$$

$Y \text{ nicht } \varepsilon$!!

$$E(Y_{\text{CHI}}) = b_0 + b_1 \cdot 1 = \mu_{\text{CHI}}$$

$$E(Y_{\text{Bos}}) = b_0 + 2b_1 = \mu_{\text{Bos}}$$

$$E(Y_{\text{NY}}) = b_0 + 3b_1 = \mu_{\text{NY}}$$

$$\mu_{\text{Bos}} - \mu_{\text{CHI}} = b_1$$

$$\mu_{\text{NY}} - \mu_{\text{Bos}} = b_1$$

$$\mu_{\text{NY}} - \mu_{\text{CHI}} = 2b_1$$

Π.χ. 2

$$X' = \begin{cases} 1 & \text{NY} \\ 2 & \text{CHI} \\ 3 & \text{Bos} \end{cases}$$

$$Y = b_0 + b_1 X'$$

$$\mu_{\text{CHI}} - \mu_{\text{NY}} = b_1$$

$$\mu_{\text{Bos}} - \mu_{\text{CHI}} = b_1$$

$$\mu_{\text{Bos}} - \mu_{\text{NY}} = 2b_1$$

Π.χ.

$$X'' = \begin{cases} 1, & \text{Chi} \\ 5, & \text{Bos} \\ -2, & \text{NY} \end{cases}$$

Κωδικοποίηση μέσω ψευδομεταβλητών

City $\Rightarrow$
$$\begin{cases} X_{CHI} = \begin{cases} 1, & \text{αν } City = Chi \\ 0, & \text{αν } City \neq Chi \end{cases} \\ X_{BOS} = \begin{cases} 1, & \text{αν } City = Bos \\ 0, & \text{αν } City \neq Bos \end{cases} \\ X_{NY} = \begin{cases} 1, & \text{αν } City = NY \\ 0, & \text{αν } City \neq NY \end{cases} \end{cases}$$

(δείχνει
Chicago)

$$Y = b_0 + b_1 X_{CHI} + b_2 X_{BOS} + b_3 X_{NY} \quad \left(\begin{array}{l} 4 \text{ άγνωστες} \\ \text{παραμέτρους} \\ ?? \end{array} \right)$$

Δείγμα

X	X_{CHI}	X_{BOS}	X_{NY}
Chi	1	0	0
Bos	0	1	0
Chi	1	0	0
NY	0	0	1
⋮	⋮	⋮	⋮

$$X_{CHI} + X_{BOS} + X_{NY} = 1$$

παρατήρηση

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

(3 άγνωστες
παραμέτρους ✓)

↓ LSE

① $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 \iff \hat{\mu}_{CHI}, \hat{\mu}_{BOS}, \hat{\mu}_{NY}$

② $H_0: \mu_{CHI} = \mu_{BOS} = \mu_{NY}$ $H_1: \text{max. 2 διαφέρει}$
 Two given;

City	EY	$EY = b_0 + b_1 X_{CHI} + b_2 X_{BOS}$	X_{CHI}	X_{BOS}
CHI	μ_{CHI}	$EY = b_0 + b_1$	1	0
BOS	μ_{BOS}	$EY = b_0 + b_2$	0	1
NY	μ_{NY}	$EY = b_0$	0	0

$\Rightarrow$

$$b_0 = \mu_{NY}$$

$$b_1 = \mu_{CHI} - \mu_{NY} \leftarrow \text{effect CHI} \right\} \text{ over NY}$$

$$b_2 = \mu_{BOS} - \mu_{NY} \leftarrow \text{effect BOS} \right\}$$

② Επεξοδος

Ans $H_0: \mu_{CHI} = \mu_{NY} \Rightarrow b_1 = 0$
 $\mu_{BOS} = \mu_{NY} \Rightarrow b_2 = 0$

$$H_0: \mu_{CHI} = \mu_{NY} = \mu_{BOS} \Leftrightarrow H_0: \beta_1 = \beta_2 = 0$$

Επεξήγηση: $H_0: \beta_1 = \beta_2 = 0$. H_1 τουλάχιστον ένα $b_1, b_2 \neq 0$

F-test ουσιαστικής μινιμάλωσης

Εκτίμηση b_0, b_1, b_2

$$\hat{\beta}_0 = 258.50$$

$$\hat{\beta}_1 = -67.50$$

$$\hat{\beta}_2 = -56.86$$

Γενική Μεθοδολογία για δίτιμες μεταβλητές

Έστω X : κατηγορική μεταβλητή

με τιμές : $a_1, a_2, \dots, a_k$
levels (ενότητες)

[Έστω R μια κατηγορική μεταβλ. $\Rightarrow$ factor]

1) Επιλέγουμε (αυθαίρετα) ένα επίπεδο
επίπεδο αναφοράς (reference level)
πχ a_k .

2) Για κάθε ένα από τα υπολοίπα επίπεδα
ορίζουμε μια δίτιμη μεταβλητή 0-1

$$X_1 = \begin{cases} 1, & \text{αν } X = a_1 \\ 0, & \text{διαφορετικά} \end{cases}$$

$$X_2 = \begin{cases} 1, & \text{αν } X = a_2 \\ 0, & \text{διαφ.} \end{cases}$$

$\vdots$

$$X_{k-1} = \begin{cases} 1, & \text{αν } X = a_{k-1} \\ 0, & \text{διαφ.} \end{cases}$$

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_{k-1} X_{k-1} + \varepsilon$$

$$\hat{b}_0 = \hat{\mu}_{a_k} = E(Y|X=a_k)$$

$$\hat{b}_1 = \hat{\mu}_{a_1} - \hat{\mu}_{a_k}$$

$$\hat{b}_2 = \hat{\mu}_{a_2} - \hat{\mu}_{a_k}$$

$$\vdots$$

$$\hat{b}_{k-1} = \hat{\mu}_{a_{k-1}} - \hat{\mu}_{a_k}$$

$$df_{\text{er}} = n - k$$

$$H_0: \mu_{a_1} = \mu_{a_2} = \dots = \mu_{a_k}$$



H_1 τουλάχιστον 2
διαφορετικές
μέσες τιμές των

$$H_0: \beta_1 = \beta_2 = \dots = \beta_{k-1} = 0$$

H_1 : τουλάχιστον ένα $\neq 0$

Γερίκευον σε αλφισοόζες, ταχυορικής μετὰβηζής

Παρά Y : εξαρτημένη

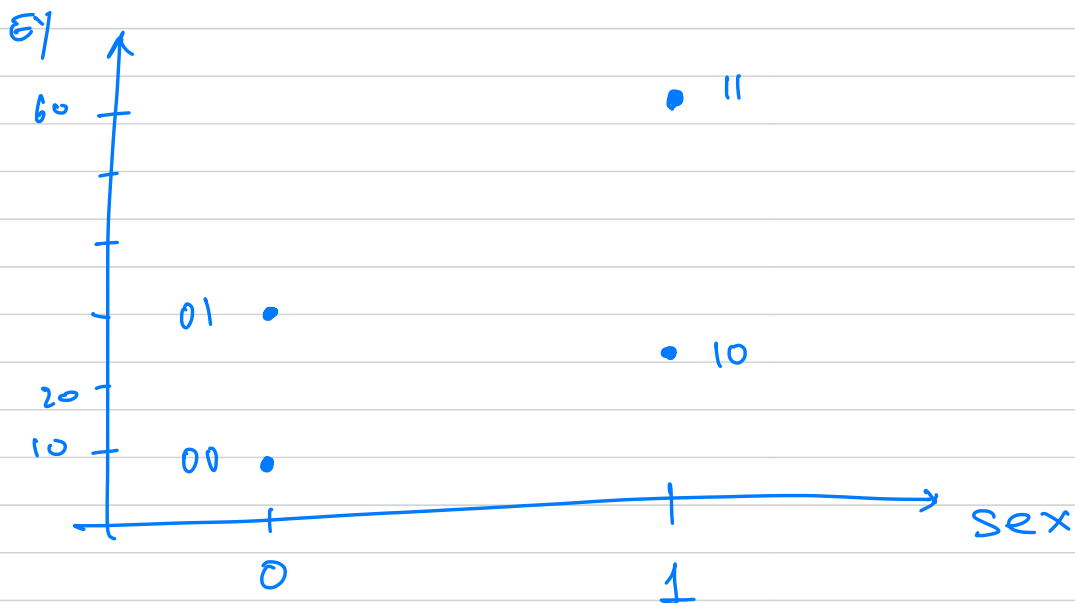
Sex : $\begin{cases} M \\ F \end{cases}$

SMOKING (SM) : $\begin{cases} NO \\ YES \end{cases}$

		SM	
		(0) No	(1) YES
(0) M	sex	μ_{MN}	μ_{MY}
(1) F	sex	μ_{FN}	μ_{FY}

4 'εξείδερες' παράμετροι

(dataset = data1)



Main effects model

$$Y = b_0 + b_1 \overset{\text{effect sex}}{\underset{\text{effect SM}}{X_F}} + b_2 \overset{\text{effect sex}}{\underset{\text{effect SM}}{X_Y}}$$

3 παράμετροι
 b_0, b_1, b_2

Sex	Smok.	X_F	X_Y	$\hat{Y} = b_0 + b_1 X_F + b_2 X_Y$
M	N	0	0	$\hat{\mu}_{MN} = b_0$
M	Y	0	1	$\mu_{MY} = b_0 + b_2$
F	N	1	0	$\mu_{FN} = b_0 + b_1$
F	Y	1	1	$\mu_{FY} = b_0 + b_1 + b_2$

$$b_0 = \hat{\mu}_{MN}$$

$$b_1 = \mu_{FN} - \mu_{MN} = \text{ενιαζ. sex σε No}$$

$$b_2 = \mu_{MY} - \mu_{MN} = \text{ενιαζ. smoking σε M}$$

$$b_1 = \mu_{FY} - \mu_{MY} = \text{ενιαζ. sex σε Y}$$

$$b_2 = \mu_{FY} - \mu_{FN} = \text{ενιαζ. smoking σε F}$$

