2023-12-22

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \varepsilon \qquad X_2 = X_1^2$$

① Η στατιστική σημαντικότητα των συντελεστών (μέσω t-test) υπολογίζεται κ' δίνεται υποθέτοντας ότι η μεταβλητή αυτή μπαίνει τελευταία στο μοντέλο που ήδη περιέχει όλες τις υπόλοιπες

② Η σημαντικότητα μιας μεταβλητής εξαρτάται από τις άλλες μεταβλητές που ήδη βρίσκονται στο μοντέλο.

③ $$df_{error} = n - \underbrace{(k+1)}_{\text{αρ. παρ. (b) στο μοντέλο}}$$

$$Y = b_0 + b_1 X_1 + \cdots + b_k X_k + \varepsilon$$

Εμπειρικός κανόνας $\begin{cases} df_{er} > 30 \\ \text{πιο αυστηρός } df_{er} > 10k \end{cases}$

Ελεγχοι F στο μοντέλο πολλαπλής παλινδρόμησης

Ελεγχος F γενικά λόγος $\dfrac{MS_1}{MS_2}$

$$MS = \frac{SS}{df} \qquad \left(\begin{array}{l} MSR = \dfrac{SSR}{df_r} \\[2ex] MSE = \dfrac{SSE}{df_{er}} \end{array} \right)$$

$$df_{er} = n-(k+1) = n - \#\beta$$

$df_{reg} = df_{mod} = k = \#\beta - 1 = \#$ ανεξ. μεταβλητών.

$$MSR = \frac{SSR}{df_r} = \frac{\text{Εξηγημένη μεταβλητότητα της } Y}{k}$$

① Ελεγχος F για όλο το μοντέλο

$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0 \qquad H_1:$ τουλάχιστον ένα $\neq 0$

$$F = \frac{MSR}{MSE} = \frac{\dfrac{SSR}{df_r = k}}{\dfrac{SSE}{df_{er}}}$$

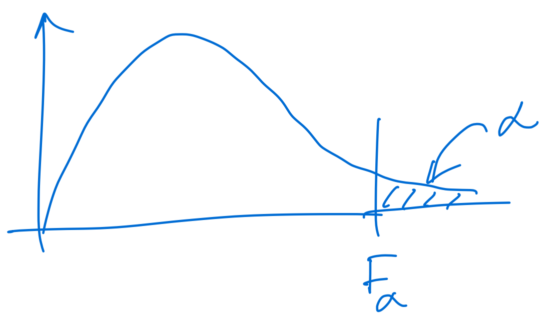
$\nearrow F \gg \;\rightarrow\; H_1$

$\searrow F \ll \;\rightarrow\; H_0$

κάτω από H_0: $F \sim F_{df_r, df_{er}}$

accept H_0: $F \leq F_\alpha$ $(p \geq \alpha)$

reject H_0: $F > F_\alpha$ $(p < \alpha)$



② Έλεγχος F για μέρος του μοντέλου

Μοντ. 1: $Y = b_0 + b_1 X_1 + b_2 X_2 + \varepsilon$

Μοντ. 2: $Y = b_0 + b_1 X_1 + b_2 X_2 + \underline{b_3 X_3 + b_4 X_4} + \varepsilon$

Μοντ. 1 "φωλιασμένο" (nested) στο Μοντ. 2
↓ part ↓ full

Στο Μοντ. 2: έστω ο έλεγχος

$H_0: b_3 = b_4 = 0$ $\qquad H_1: b_3 \neq 0$ ή $b_4 \neq 0$ ή κ' τα δύο

Θέλουμε να ελέγξουμε αν το μοντέλο 2 είναι στατιστικά σημαντικά διαφορετικό (καλύτερο) από το Μοντ. 1

Έστω $SSR_{part} = SSR(X_1, X_2)$

$SSR_{full} = SSR(X_1, X_2, X_3, X_4) \geq SSR_{part}$

$$\underbrace{SSR_{full} - SSR_{part}}_{\text{αύξηση της εξηγ. μετ. της } Y} = SSR(X_3, X_4 \mid X_1, X_2)$$

$$F = \frac{\dfrac{SSR_{full} - SSR_{part}}{df_{r,full} - df_{r,part}}}{MSR_{full}},$$

Αν Ho

$F \sim F_{df_{r,full} - df_{r,part},\ df_{r,full}}$

3 Ελεγχοι F typeI (ή sequential F-tests)

Στο μοντέλο $Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \varepsilon$

κ' στο $Y = b_0 + b_1 X_3 + b_2 X_1 + b_3 X_2 + \varepsilon$

το t-test της X_1

θα είναι στο Μοντ 1: $H_0: b_1 = 0$ $H_1: b_1 \neq 0$

" Μοντ 2: $H_0: b_2 = 0$, $H_1: b_2 \neq 0$

θα έχουν ακριβώς το ίδιο αποτέλεσμα

Όμως κάποιες φορές ενδιαφέρει να βρούμε την στατ. σημαντικότητα μιας μεταβλητής ανάλογα με τη σειρά με την οποία έχει μπει (!)

M1 : $Y_0 : b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$

μόνη της μονοπαραγοντικό

sequential F-tests

$H_0: b_1 = 0 \quad : H_1: b_1 \neq 0$ μονοπαραγ (X_1) (X_1)

μετά της X_1 !!

$H_0: b_2 = 0 \quad : H_1: b_2 \neq 0$ X_1, X_2 $(X_2 | X_1)$

$H_0: b_3 = 0 \quad : H_1: b_3 \neq 0$ X_1, X_2 X_3 $(X_3 | X_1, X_2)$

$M2: \underbrace{Y_0 = b_0 + b_1 X_2 + b_2 X_3 + b_3 X_1}$

$H_0: b_1 = 0 \quad H_1: b_1 \neq 0 \quad (X_2)$

$b_2 \quad (X_3 | X_2)$

$b_3 \quad (X_1 | X_2, X_3)$

π.χ. X_1 = age

X_2 = BP

X_3 = αρ. τσιγ/ημέρα

Y εξαρτ.

: $Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$

seq. F-test:

1) Age : stat significant

2) BP | Age : " "

3) smoking | Age, BP " "

---

4 Partial F-tests : (Type III)

(~ t-test)

$Y_0 : b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$

$H_0: b_1 = 0, \quad H_1: b_1 \neq 0 \quad : \quad X_1 | X_2, X_3$

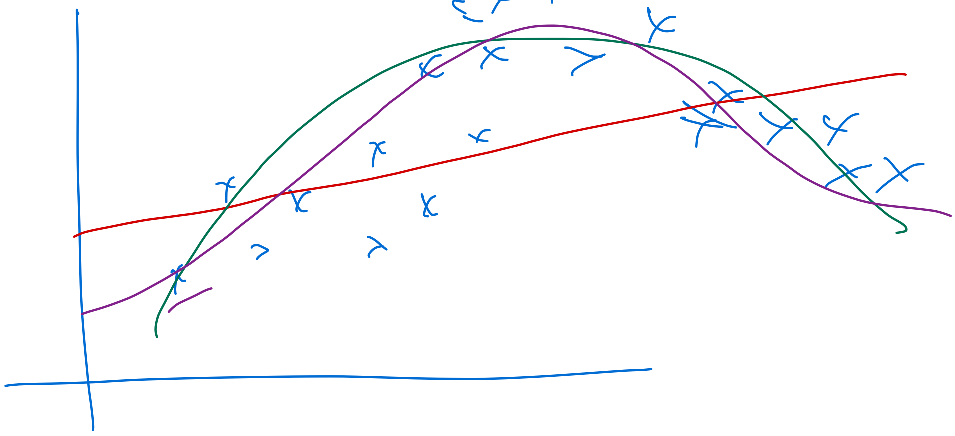
$H_0 \quad b_2 \quad \ldots \quad X_2 | X_1, X_3$

$b_3 \quad X_3 | X_1, X_2$

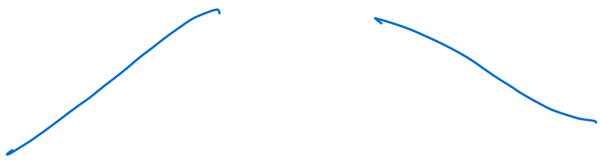
Πολυωνυμική Παλινδρόμηση

$$Y = b_0 + b_1 X_1 + b_2 X_2 \qquad X_2 = X_1^2$$

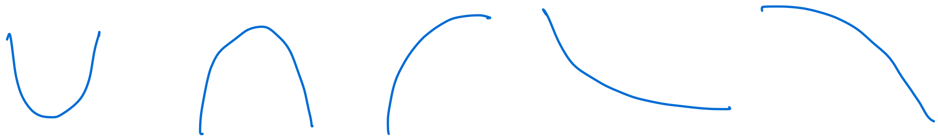
$$Y = b_0 + b_1 X_1 + b_2 X_1^2$$ ← πολυωνυμική παλινδρόμηση βαθμού 2



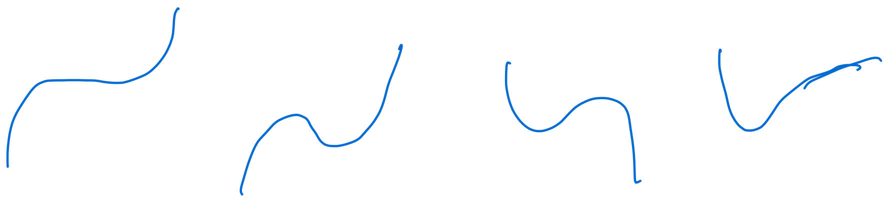
$b_0 + b_1 X$



$b_0 + b_1 X + b_2 X^2$



$b_0 + b_1 x + b_2 x^2 + b_3 x^3$



ΚΑΤΗΓΟΡΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ (ΑΝΕΞΑΡΤΗΤΕΣ)

Στο μοντέλο παλινδρόμησης $Y = b_0 + b_1 X_1 + b_2 X_2 + \varepsilon$
υποθέσαμε ότι οι X_1, X_2 είναι ποσοτικές

Όμως σε εφαρμογές κάποιες ανεξ. μετ. είναι κατηγορικές

Παράδειγμα: Τιμές διαμερισμάτων σε 3 πόλεις

X: πόλη : categorical

Τύποι Μεταβλητών

1) X: κατηγορική (nominal) (categorical) : οι δυνατές τιμές της είναι διακριτές κ' μη συγκρίσιμες (π.χ. χρώμα)

2) X: διάταξης (ordinal) : οι τιμές της μπορούν να μπουν σε αύξουσα σειρά αλλά όχι αναλογικά
(Κλίμακα Likert σε ερωτηματολόγια
1: καθόλου
2: λίγο
3: πολύ
4: εξαιρετικά)

3) X: κλίμακας (scale ποσοτικές) : οι τιμές μετρούνται με βάση μονάδα μέτρησης
π.χ. ύψος (σε cm) $X_1 = 200$, $X_2 = 100$, $X_1 = 2X_2$

Μέθοδος LSE

Παράδειγμα: (τιμές διαμερισμάτων)

Y = τιμή πώλησης

δείγμα $n=32$

$\bar{Y} = 219,97$

$\bar{Y}_{NY} = 258.5$

$\bar{Y}_{CHI} = 191$

$\bar{Y}_{BOS} = 201$

φαίνεται ότι η πόλη εξηγεί μέρος της μεταβλητότητας της τιμής

ίδια διασπορά

Έστω Y_{NY} : τιμή στη NY $\sim N(\mu_{NY}, \sigma^2)$

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

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

Θα θέλαμε να κάνουμε τον έλεγχο:

3 άγνωστες παράμετροι ($\mu_{NY}, \mu_{BOS}, \mu_{CHI}$.)

$$\boxed{H_0 : \mu_{NY} = \mu_{BOS} = \mu_{CHI} \qquad H_1 : \text{τουλάχιστον δύο διαφορετικές}}$$

Στη Βιοστατιστική έχουμε δει t-test για σύγκριση μέσων δύο πληθυσμών

$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$

① Μέθοδος ANOVA (Analysis of Variance)

(διαφορετική έννοια από πίνακα ANOVA)

(Δεν χρησιμοποιείται)

② Παλινδρόμηση μέσω κατάλληλης κωδικοποίησης της $X = $ πόλη → ποσοτική

$$\text{City} = \begin{cases} \text{Chi} \\ \text{Bos} \\ \text{NY} \end{cases} \longrightarrow \text{ποσοτική}$$

City : κατηγορική σε 3 επίπεδα (levels) {Ch, Bos, NY} →

$$\left.\begin{aligned} \text{Chi} &= 1 \\ \text{Bos} &= 2 \\ \text{NY} &= 3 \end{aligned}\right\} = X$$

$$Y = b_0 + b_1 X + \varepsilon$$

Αν City = Chi ⇒ $X = 1$ ⇒ $EY = b_0 + b_1 = \mu_{CHI}$

" Bos ⇒ $X = 2$ ⇒ $EY = b_0 + 2b_1 = \mu_{Bos}$

NY ⇒ $X = 3$ ⇒ $EY = b_0 + 3b_1 = \mu_{NY}$

$$\underbrace{\mu_{Bos} - \mu_{CHI}}_{10} = \underbrace{\mu_{NY} - \mu_{Bos}}_{\sim 60} = b_1 \quad \times \; ??$$

Κωδικοποίηση μέσω δυαδικών (0–1) μεταβλητών (binary, dummy variables)

← νέες μεταβλητές

Levels	X_{CHI}	X_{BOS}	X_{NY}
Chi	1	0	0
Bos	0	1	0
NY	0	0	1

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

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

$$X_{NY} = \begin{cases} 1, & \text{city} = \text{NY} \\ 0, & \text{διαφ.} \end{cases}$$

Δείγμα

Price	~~City~~	X_{CH}	X_{BOS}	~~X_{NY}~~
'	Bos	0	1	0
'	NY	0	0	1
'	Chi	1	0	0
'	NY	0	0	1
'	'			
'	'			

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

→ → City = ?

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

4 μεταβλητές

στο αρχικό μόνο 3: $\mu_{CHI}, \mu_{BOS}, \mu_{NY}$

X_{CHI}	X_{BOS}	City
0	0	$\longrightarrow$ NY
0	1	
1	0	
0	0	
1	1	

Παραλείπεται η NY

$\Downarrow$

NY: επίπεδο αναφοράς

$$Y = \beta_0 + \beta_1 X_{CH} + \beta_2 X_{BOS}$$

(3 παράμετροι b_0, b_1, b_2)

Αν σε αυτό το μοντέλο εκτιμήσουμε $\hat{b}_0, \hat{b}_1, \hat{b}_2$

τότε είναι οι εκτιμήτριες $\hat{\mu}_{CH}, \hat{\mu}_{BOS}, \hat{\mu}_{NY}$?

$$EY = b_0 + b_1 X_{CH} + b_2 X_{BOS}$$

A: City = NY $\quad X_{CHI}=0,\ X_{BOS}=0 \Rightarrow EY = b_0 + b_1 \cdot 0 + b_2 \cdot 0 = \mu_{NY}$

$$\Rightarrow \boxed{\mu_{NY} = b_0}$$

City = Chi. $\quad X_{CHI}=1,\ X_{BOS}=0 \Rightarrow EY = b_0 + b_1 \cdot 1 + b_2 \cdot 0 = \mu_{CHI}$

$$\Rightarrow \boxed{\mu_{CHI} = b_0 + b_1}$$

City = Bos $\quad X_{BOS}=1,\ X_{CHI}=0 \Rightarrow EY = b_0 + b_1 \cdot 0 + b_2 \cdot 1 = \mu_{BOS}$

$$\Rightarrow \boxed{\mu_{BOS} = b_0 + b_2}$$

$$\left.\begin{array}{l} \hat{b}_0 = 258.5 \\ \hat{b}_1 = -67.5 \\ \hat{b}_2 = -56.9 \end{array}\right\} \Rightarrow \boxed{\begin{array}{l} \hat{\mu}_{NY} = 258.5 \\ \hat{\mu}_{CHI} = \hat{b}_0 + \hat{b}_1 = 258.5 - 67.5 = \underline{191} \\ \hat{\mu}_{BOS} = \hat{b}_0 + \hat{b}_2 = 258.5 - 56.9 \simeq \underline{202} \end{array}}$$

Equivalently

$$\left.\begin{array}{l} \mu_{NY} = b_0 \\ \mu_{CHI} = b_0 + b_1 \\ \mu_{BOS} = b_0 + b_2 \end{array}\right\} \Rightarrow \boxed{\begin{array}{l} b_0 = \mu_{NY} \\ b_1 = \mu_{CHI} - \mu_{NY} \\ b_2 = \mu_{BOS} - \mu_{NY} \end{array}}$$

Γενικά: $b_0 = E(Y \mid \text{επίπεδο αναφοράς})$

$b_1 = E(Y \mid \text{επίπεδο } 1) - E(Y \mid \text{αναφ})$

$b_2 = E(Y \mid \text{επιπ } 2) - E(Y \mid \text{αναφ})$

Ο αρχικός έλεγχος

$H_0: \mu_{NY} = \mu_{CHI} = \mu_{BOS}$ vs H_1: τουλαχ. 2 διαφορετικά

$\Downarrow$

$\mu_{NY} = \mu_{CHI} \Rightarrow \beta_0 = \beta_0 + \beta_1 \Rightarrow \beta_1 = 0$

$\mu_{NY} = \mu_{BOS} \Rightarrow \beta_0 = \beta_0 + \beta_2 \Rightarrow \beta_2 = 0$

ισοδύναμος έλεγχος: $H_0: b_1 = b_2 = 0$, H_1: τουλάχιστον ένα $\neq 0$

Έλεγχος F για πλήρες μοντέλο

$p = 3 \cdot 10^{-8} \simeq 0$

$\Rightarrow$ απορρίπτεται σχεδόν για κάθε α