ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Μοντέλα για τη διερεύνηση και μελέτη της σχέσης μεταξύ μιας εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών

ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

α) Η εξαρτημένη και οι ανεξάρτητες μεταβλητές είναι ποσοτικές.

β) Αναζητείται, κατασκευάζεται και ελέγχεται η ορθότητα της γραμμικής σχέσης που συνδέει την εξαρτημένη με τις ανεξάρτητες μεταβλητές. Η γραμμική σχέση που κατασκευάζεται χρησιμοποιείται για την πρόβλεψη της εξαρτημένης μεταβλητής για δεδομένες τιμές των ανεξάρτητων μεταβλητών.

ΜΟΝΤΕΛΑ ΑΝΑΛΥΣΗΣ ΔΙΑΚΥΜΑΝΣΗΣ

α) Η εξαρτημένη μεταβλητή είναι ποσοτική και οι ανεξάρτητες μεταβλητές είναι, κατά κανόνα, ποιοτικές.

β) Δεν κατασκευάζεται κάποια γραμμική σχέση μεταξύ της εξαρτημένης και των ανεξάρτητων μεταβλητών αλλά το ενδιαφέρον εστιάζεται στην ανάλυση των επιδράσεων μιας ή περισσότερων ανεξάρτητων μεταβλητών στην εξαρτημένη μεταβλητή. Αναζητούνται οι τιμές των ανεξάρτητων μεταβλητών που ασκούν τη σημαντικότερη επίδραση στην εξαρτημένη μεταβλητή.

Μοντέλα ανάλυσης διακύμανσης

↓

Γραμμικά μοντέλα

↙ ↘

Μοντέλα παλινδρόμησης

1) Ποσοτικές μεταβλητές.
   Y $X_1, \ldots, X_p$
2) Προσπαθούμε να κατασκευάσουμε τη γραμμική σχέση που συνδέει την Y με τις $X_1 \ldots X_p$, να ελέγξουμε την εγκυρότητα της σχέσης και να τη χρησιμοποιήσουμε προκειμένου να κάνουμε προβλέψεις.

Μοντέλα ανάλυσης διακύμανσης

1) Θεωρούμε μια εξαρτημένη ποσοτική μεταβλητή Y και μια ή περισσότερες κατά κανόνα ποιοτικές μεταβλητές
2) Στα μοντέλα ανάλυσης διακύμανσης δεν ενδιαφερόμαστε για την κατασκευή κάποιας συγκεκριμένης σχέσης, αλλά ενδιαφερόμαστε στο να εντοπίσουμε, ποια ή ποιες από τις κατηγορίες των ποιοτικών μεταβλητών ασκούν τη σημαντικότερη επίδραση στην Y.

Κάποιες φορές ποσοτικά δεδομένα τα μετατρέπουμε σε ποιοτικά π.χ ύψος
- 150 και κάτω → κοντός
- 150-165 → μέτριος
- 165-180 → ψηλός
- 180-2 → πολύ ψηλός

Ορολογία-Συμβολισμός

$Y \equiv$ ελίδωση (επίπεδα)

↑ (παράγοντας)

$X \equiv$ δείκτης νοημοσύνης ← χαμηλός, μέτριος, ψηλός

δεν έχει νόημα να συνδέσω κάτι που μετριέται με κάτι που χαρακτηρίζεται

Παράγοντας → ταυτόσημη με την ανεξάρτητη μεταβλητή.
Κάθε τιμή του παράγοντα (που είναι μια κατηγορία, αφού ο παράγοντας είναι ποιοτική μεταβλητή) ονομάζεται **επίπεδο** του παράγοντα.
Προβλήματα στα οποία υπεισέρχεται **ένας** παράγοντας, ονομάζονται **προβλήματα ανάλυσης διακύμανσης κατά ένα παράγοντα.**
Προβλήμα στα οποία υπεισέρχονται **δύο** παράγοντες, ονομάζονται **προβλήματα ανάλυσης διακύμανσης κατά δύο παράγοντες**
Προβλήματα με περισσότερους από δύο παράγοντες, ονομάζονται προβλήματα **πειραματικών σχεδιασμών**
Δοκιμασία: Κάθε συνδυασμός επιπέδων των παραγόντων. (δύο ή περισσότερων)

ΠΑΡΑΔΕΙΓΜΑ: Ενδιαφέρει η επίδοση Υ μαθητών (**ποσοτική**). Ενδιαφέρεται να διερευνήσει πως το Υ επηρεάζεται από το επίπεδο μόρφωσης Πατέρα (Ε.Μ.Π)
Το επίπεδο μόρφωσης πατέρα είναι ποιοτική μεταβλητή η οποία κατηγοριοποιείται ως εξής: 1) Απόφοιτος Υποχρεωτικής (ΑΥ) 2) Λυκείου (Λ) 3) Πανεπιστημίου (Π) 4) Μεταπτυχιακό (Μ) 5) Διδακτορικό (Δ)

Πρόβλημα ανάλυσης διακύμανσης κατά ένα παράγοντα.
Παράγοντας ≡ ΕΜΠ, Επίπεδα παράγοντα

Αν εισάγουμε το επίπεδο μόρφωσης μητέρας → κατά δύο παράγοντες

1ος παράγοντας: ΕΜΠ.
2ος παράγοντας: ΕΜΜ. (μητέρα)

Δοκιμασία: (ΑΥ, ΑΥ), (ΑΥ, Λ) κ.ο.κ.

(Περιγράφεται από μια μέση τιμή, μια διακύμανση, μια κατανομή)

(Ι) Ανάλυση διακύμανσης κατά ένα παράγοντα

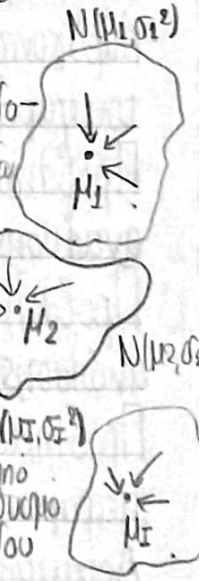
Θεωρώ μια ποσοτική Υ και ένα παράγοντα ο οποίος απαντάται σε Ι επίπεδα.

Μορφή δεδομένων

δεδομένα για τη Υ (τ.δ μεγέθους J_1)
δεδομένα από τον πληθυσμό του 1ου επιπέδου
δεδομένα από τον πληθυσμό του 2ου επιπέδου

Παράγοντας	Δεδομένα για την Υ	Σύνολα	Μέσος όρος
1 επίπεδο	$Y_{11}, Y_{12}, \dots, Y_{1J_1}$ (1ος παιδί, 2ος παιδί)	$Y_{1\cdot}$	$\bar{Y}_{1\cdot}$
2 επίπεδο	$Y_{21}, Y_{22}, \dots, Y_{2J_2}$	$Y_{2\cdot}$	$\bar{Y}_{2\cdot}$
⋮	⋮	⋮	⋮
Ι-επίπεδο	$Y_{I1}, Y_{I2}, \dots, Y_{IJ_I}$	$Y_{I\cdot}$	$\bar{Y}_{I\cdot}$

Y_{ij} η j-παρατήρηση του i επίπεδο του παράγοντα



όπου $Y_{i\cdot} = \sum_{j=1}^{J_i} Y_{ij}$ | $\bar{Y}_{i\cdot} = \frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij} = \frac{1}{J_i} Y_{i\cdot}$ δειγματικό μέσο | $\bar{Y}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J_i} Y_{ij}$ | $N = \sum_{i=1}^{I} J_i$, $N = J_1 + J_2 + \dots + J_I$

Μοντέλο ανάλυσης διακύμανσης κατά ένα παράγοντα

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad , i = 1, \dots, I \quad , \quad J = 1, \dots, J_i$$

Εκτιμητές ελαχίστων τετραγώνων

$$S = \sum_{i=1}^{I} \sum_{j=1}^{J_i} \varepsilon_{ij}^2 = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \mu_i)^2$$

$$\frac{\partial S}{\partial \mu_i} = 0 \Rightarrow \dots \Rightarrow \text{ΕΕΤ} : \hat{\mu}_i = \overline{Y}_{i\cdot} \quad , i = 1, \dots, I$$

(Υποθέσεις για σφάλματα):

1) $E(\varepsilon_{ij}) = 0$, $i = 1, \dots, I = k$
2) $Var(\varepsilon_{ij}) = \sigma^2$, $j = 1, \dots, J_i = \ell$
3) $Cov(\varepsilon_{ij}, \varepsilon_{k\ell}) = 0$, $k \neq \ell$.
4) $\varepsilon_{ij} \sim N(0, \sigma^2)$

$\longleftrightarrow$ αντιστοίχιση

① $E(Y_{ij}) = \mu_i$, $i = 1, \dots, I$
② $Var(Y_{ij}) = \sigma^2 / J_i$, $i = 1, \dots, I$
③ $Cov(Y_{ij}, Y_{k\ell}) = 0$
④ $Y_{ij} \sim N(\mu_i, \sigma^2)$

ΙΔΙΟΤΗΤΕΣ

1) $E(\hat{\mu}_i) = \mu_i$, $i = 1, \dots, I$
2) $Var(\hat{\mu}_i) = \frac{\sigma^2}{J_i}$, $i = 1, \dots, I$

ΑΠΟΔΕΙΞΗ

($E(\varepsilon_{ij}) = 0$ από υπόθεση)

1) $E(\hat{\mu}_i) = E(\overline{Y}_{i\cdot}) = E\left(\frac{1}{J_i} \sum_{j=1}^{J_i} Y_{ij}\right) = \frac{1}{J_i} \sum_{j=1}^{J_i} E(Y_{ij}) = \frac{1}{J_i} \cdot \sum_{j=1}^{J_i} \mu_i = \frac{1}{J_i} \cdot J_i \, \mu_i = \mu_i$

Ισοδύναμο μοντέλο (μαθηματικά $\mu_i = \mu + a_i$)

$$Y_{ij} = \mu + a_i + \varepsilon_{ij} \quad , \quad i = 1, \dots, I \quad , \quad j = 1, \dots, J_i$$

Εκτιμητές ελαχίστων τετραγώνων

$$S = \sum_{i=1}^{I} \sum_{j=1}^{J_i} \varepsilon_{ij}^2 = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \mu - a_i)^2$$

$$\left.\begin{array}{l} \frac{\partial \Sigma}{\partial \mu} = 0 \\ \frac{\partial \Sigma}{\partial a_i} = 0 \end{array}\right\} \Rightarrow \left.\begin{array}{l} \sum_{i=1}^{I} a_i J_i + \mu \sum_{i=1}^{I} J_i = \sum_{i=1}^{I}\sum_{j=1}^{J_i} Y_{ij} \quad \leftarrow \text{πρώτη εξίσωση} \\ a_{ij} J_i + \mu J_i = \sum_{j=1}^{J_i} Y_{ij}, \quad i = 1, \dots, I \end{array}\right\} \text{κανονικές εξισώσεις}$$

Το σύστημα κανονικών εξισώσεων δεν οδηγεί σε κανονική λύση, αφού αν αθροίσω την 2η εξίσωση προκύπτει η 1η ως προς a_i (γραμ. εξαρτημένα)

Για να πετύχω μοναδική λύση θεωρώ πλευρική συνθήκη. Το ερώτημα είναι ποια θα είναι η πλευρική συνθήκη;
Έχουν προταθεί διάφορες. Από τις διάφορες καλύτερη φυσική ερμηνεία έχει εκείνη που οδηγεί σε αποδεκτούς διαισθητικά εκτιμητές.
Π.Χ. διαισθητικά αποδεκτός εκτιμητής για το μ είναι $\hat{\mu} = \bar{Y}_{..}$ → γενικός δειγματικός μέσος
(αφού το μ σχετίζεται με όλες τις παρατηρήσεις δηλαδή διαισθητικά του δεχόμαι)

Για να πετύχω ως εκτιμητή του γενικού δειγματικού μέσου αρκεί να υποθέσω ότι $\sum_{i=1}^{I} a_i J_i = 0$ (από την πρώτη εξίσωση)

οπότε θεωρώ πλευρική συνθήκη: $\sum_{i=1}^{I} a_i J_i = 0$.

Υπό αυτή τη πλευρική συνθήκη οι ΕΕΤ είναι $\hat{\mu} = \bar{Y}_{..}$
$\hat{a}_i = \bar{Y}_{i.} - \bar{Y}_{..}$

Πίνακας Ανάδια του μοντέλου ανάλυσης διακύμανσης κατά ένα παράγοντα

- ολική μεταβλητότητα = δειγματική διακύμανση χωρίς τα μεγέθη δείγματος
- δειγματική διακύμανση $= \frac{1}{N-1} \sum_{i=1}^{I}\sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{..})^2$
  ↳ χωρίς μεγέθη δείγματος

ΘΕΩΡΗΜΑ: $SStot = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{I} J_i(\bar{Y}_{i.} - \bar{Y}_{..}) + \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i.})^2$

$SStr$ $SSres$

ΑΠΟΔΕΙΞΗ: $\sum\sum(Y_{ij} - \bar{Y}_{..} + \bar{Y}_{i.} - \bar{Y}_{i.})^2 =$ ... πράξεις

ΠΙΝΑΚΑΣ ΑΝΑΛΥΣΗΣ

πηγή μεταβλητότητας	SS	B.E	MS	F-πηλίκο
μοντέλο	$SStr = \sum_{i=1}^{I} J_i(\bar{Y}_{i.} - \bar{Y}_{..})$	$I-1$	$MStr = \frac{SStr}{I-1}$	$F = \frac{MStr}{MSres}$
υπόλοιπα	$SSres = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i.})^2$	$N-I$	$MSres = \frac{SSres}{N-I}$	
ολική	$SStot = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{..})^2$	$N-1$		