Κεφάλαιο 10

Η Ανάλυση Παλινδρόμησης

Η Ανάλυση Παλινδρόμησης

- Το στατιστικό κριτήριο που μας επιτρέπει να **προβλέψουμε** τις τιμές μιας μεταβλητής από τις τιμές μιας ή πολλών άλλων γνωστών μεταβλητών
- Η σχέση ανάμεσα στις μεταβλητές που μελετώνται χαρακτηρίζονται **αιτιώδης**, διότι οι τιμές της μιας μεταβλητής ερμηνεύουν τις τιμές της άλλης
- Μπορούμε να προσδιορίσουμε το βαθμό στον οποίο μια ή περισσότερες ανεξάρτητες μεταβλητές **επηρεάζουν** μια εξαρτημένη

2

Βασικές Έννοιες

- **Ερμηνευτική ή προβλεπτική μεταβλητή:**
  Η μεταβλητή της οποίας γνωρίζουμε τις τιμές
- **Μεταβλητή Κριτήριο:**
  Η μεταβλητή της οποίας θέλουμε να προβλέψουμε τις τιμές
- **Απλή Παλινδρόμηση:**
  Όταν έχουμε μια προβλεπτική μεταβλητή
- **Πολλαπλή Παλινδρόμηση:**
  Όταν έχουμε πολλές προβλεπτικές μεταβλητές

3

Κριτήρια για την εφαρμογή του κριτηρίου

- Οι δύο μεταβλητές να **συσχετίζονται** μεταξύ τους σε ικανοποιητικό βαθμό
- Να έχουν **ευθύγραμμη** σχέση μεταξύ τους

4

Ένα Παράδειγμα

- Ένας ερευνητής θέλει να μελετήσει τη σχέση που υπάρχει ανάμεσα στην **Ευσυνειδησία** (διάσταση της προσωπικότητας) και την **Επαγγελματική Ικανοποίηση** (ικανοποίηση από την εργασία)
- Πιο συγκεκριμένα, θέλει να διαπιστώσει κατά πόσο μπορεί κάποιος να **προβλέψει** την επαγγελματική ικανοποίηση του ατόμου, βασιζόμενος στην επίδοση που πέτυχε στην κλίμακα της Ευσυνειδησίας

5

Γιατί να μας ενδιαφέρουν τέτοιου είδους δεδομένα;

- Γιατί ίσως να θέλουμε να κάνουμε μια **πρόβλεψη**
- Επιπρόσθετα, ίσως να θέλουμε να καταλάβουμε τη **σχέση** που έχουν αυτές οι μεταβλητές μεταξύ τους
  - π.χ. Πόσο αυξάνεται η επαγγελματική ικανοποίηση του ατόμου όταν αυξηθεί κατά μία μονάδα η επίδοσή του στην κλίμακα της Ευσυνειδησίας;

6

Τα Δεδομένα της Έρευνας

Άτομα	Ευσυνειδησία	Επαγγελματική Ικανοποίηση
1	74	56
2	57	58
3	45	43
4	81	75
5	56	61
6	62	60
7	33	42
8	83	93
9	72	59
10	53	64

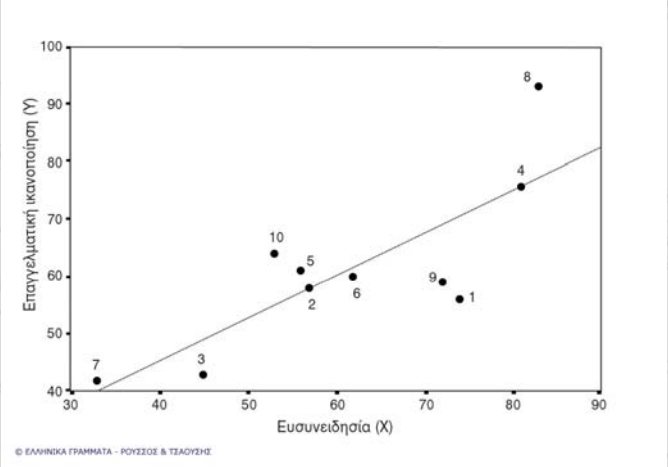
7

Ικανοποιούνται οι προϋποθέσεις για τη χρήση της παλινδρόμησης;

- Ο δείκτης συσχέτισης μεταξύ των δύο μεταβλητών είναι r = **0.80**
- Η σχέση μεταξύ τους είναι **ευθύγραμμη** όπως φαίνεται και από το διάγραμμα σκεδασμού (επόμενη διαφάνεια)

8

Διάγραμμα Σκεδασμού και η Γραμμή Παλινδρόμησης για της μεταβλητές της Έρευνας



Η γραμμή Παλινδρόμησης

- Η γραμμή που περνά όσο το δυνατόν **πλησιέστερα** από τα περισσότερα σημεία του διαγράμματος σκεδασμού
  - **Ο Τύπος…**

$$\hat{Y} = a + bX$$

- $\hat{Y}$ = η τιμή του *Y* που θέλουμε να προβλέψουμε (επαγγελματική ικανοποίηση)
- ***X*** = η επίδοση των ατόμων στην κλίμακα Ευσυνειδησίας

10

Συντελεστές Παλινδρόμησης

- Οι "Συντελεστές" είναι το **α** και το ***b***
- **α = σταθερός όρος** (intercept)
  - Η τιμή του $\hat{Y}$ όταν το *X* = 0
- ***b* = κλίση** (slope)
  - Η αλλαγή που συντελείται στην προβλεπόμενη τιμή του Y, για κάθε μία μονάδα του *X*

11

Πώς υπολογίζονται

- **Κλίση**

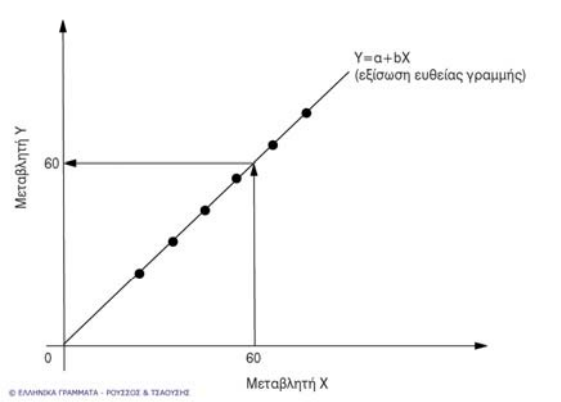
$$b = \frac{N\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{N\left(\sum X^2\right) - \left(\sum X\right)^2}$$

- **Σταθερός Όρος**

$$a = \overline{Y} - b\overline{X}$$

12

Παράδειγμα



α = 0, **b** = 45° = 1,

Y = 0 + 1X ή Y = X

π.χ.

X = 60, Y=;

Παράδειγμα

- **Χρησιμοποιώντας τα δεδομένα του παραδείγματος:**

N = 10, **ΣXY** = 39341, **ΣX** = 616, **ΣY** = 611, **ΣX²** = 40262, & **(ΣX)²** = 379456

$$b = \frac{10(39341) - (616)(611)}{10(40262) - (379456)} = \frac{393410 - 376376}{402620 - 379456} = \frac{17034}{23164} = 0.735$$

$$a = 61.1 - (0.735 \times 61.6) = 61.1 - 45.3 = 15.82$$

Πρόβλεψη…

- Εάν κάποιος θέλει να προβλέψει ποια θα είναι η επίδοση ενός ατόμου στο τεστ επαγγελματικής ικανοποίησης όταν έχει **επίδοση 50** στην κλίμακα ευσυνειδησίας:

$$\hat{Y} = a + bX$$

$$\hat{Y} = 15.82 + (0.735 \times 50)$$

= 15.82 + (36.75) = **52.57**

15

Παράδειγμα

- Θέλουμε να μελετήσουμε τη σχέση που υπάρχει ανάμεσα στον αριθμό των **τσιγάρων** και τον **καρκίνο του πνεύμονα**

- Θέλουμε να **προβλέψουμε** το ποσοστό των θανάτων από καρκίνο του πνεύμονα μελετώντας τον αριθμό των τσιγάρων που κατά μέσο όρο καπνίζουν την ημέρα οι κάτοικοι 21 χωρών

16

Τα δεδομένα της Έρευνας

Χώρες	Νο Τσιγάρων	ΚΠΝ
1	11	26
2	9	21
3	9	24
4	9	21
5	8	19
6	8	13
7	8	19
8	6	11
9	6	23
10	5	15
11	5	13
12	5	4
13	5	18
14	5	12
15	5	3
16	4	11
17	4	15
18	4	6
19	3	13
20	3	4
21	3	14

Νο Τσιγάρων = Αριθμός Τσιγάρων ανά άτομο την ημέρα

ΚΠΝ = Θάνατοι από καρκίνο του πνεύμονα ανά 10.000 κατοίκους

17

Τα αποτελέσματα από το SPSS

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,367	2,941		,805	,431
	CIGAR	2,042	,461	,713	4,426	,000

a. Dependent Variable: CHD

Σημειώσεις:

- Ο σταθερός όρος ονομάζεται και **constant**
- Η κλίση συχνά αντικαθίσταται από την **προβλεπτική μεταβλητή**

18

Πώς μπορούμε να προβλέψουμε τον αριθμό των θανάτων από καρκίνο

- Ας υποθέσουμε ότι θέλουμε να βρούμε τον αριθμό των ατόμων που θα πεθάνουν από καρκίνο του πνεύμονα όταν ο αριθμός των τσιγάρων που καπνίζουν σε μια χώρα κατά μέσο όρο θα είναι 6

$$\hat{Y} = a + bX = 2.37 + 2.04X$$

$$\hat{Y} = 2.37 + 2.04 * 6 = 14.61$$

- Προβλέπουμε ότι 14.61 ανά 10.000 άτομα θα πεθάνουν από καρκίνο του πνεύμονα

19

Σφάλματα στην Πρόβλεψη

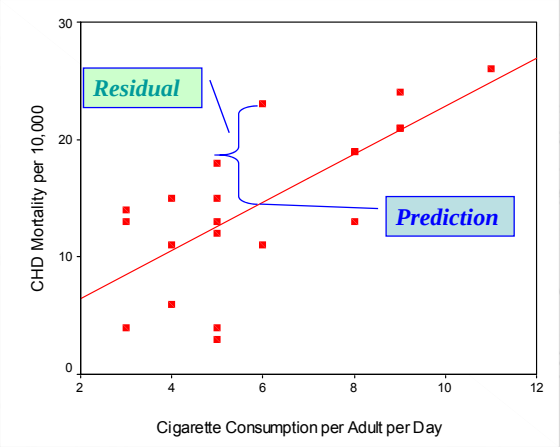
- Κάθε προσπάθεια να προβλέψουμε το Y για μια συγκεκριμένη τιμή του X εμπεριέχει κάποιο **σφάλμα**
- Γι' αυτό υπολογίζουμε το $\hat{Y}$, το οποίο είναι η **καλύτερη δυνατή εκτίμηση** που μπορούμε να κάνουμε
- Η διαφορά ανάμεσα στο Y και το $\hat{Y}$, μας δείχνει το **μέγεθος του σφάλματος**

20

Σφάλματα στην Πρόβλεψη

- Οι Φιλανδοί καπνίζουν κατά μέσο όρο **6 τσιγάρα** την ημέρα. Προβλέψαμε ότι θα έχουν **14.61** θανάτους ανά 10.000 πληθυσμό
- Στην πραγματικότητα όμως, έχουν **23 θανάτους** ανά 10.000 πληθυσμό
- Το **σφάλμα** μας (residual) είναι 23 - 14.61 = **8.39**
  - Αυτό είναι ένα σημαντικό σφάλμα

21



Μπορεί να νομίζουμε ότι μπορούμε να προβλέψουμε μια τιμή αλλά μεσολαβεί και το σφάλμα της μέτρησης που επηρεάζει αυτή την πρόβλεψη

22

Σφάλματα στην Πρόβλεψη

- **Διακύμανση των υπολοίπων** (residual variance)
  - Η διακύμανση των προβλεπόμενων τιμών

$$s^2_{Y-\hat{Y}} = \frac{\Sigma(Y-\hat{Y})^2}{N-2}$$

- **Τυπικό σφάλμα εκτίμησης**
  - Η τυπική απόκλιση των προβλεπόμενων τιμών

23

Τυπικό Σφάλμα Εκτίμησης (Standard Error of Estimate)

- Η **τυπική απόκλιση** των σημείων γύρω από τη γραμμή παλινδρόμησης

$$s_{X.Y} = \sqrt{\frac{\Sigma(Y-\hat{Y})^2}{N-2}}$$

- Ένας δείκτης που μας πληροφορεί για την **ακρίβεια** της πρόβλεψης που κάναμε
  - Θέλουμε να είναι όσο **μικρότερος** γίνεται

24

Ο δείκτης Προσδιορισμού r^2

- Είναι ο δείκτης που μας πληροφορεί για το **κοινό** ποσοστό της διακύμανσης που ερμηνεύουν οι δύο μεταβλητές X και Y

$$r^2 = \frac{\sum\left(\hat{Y} - \bar{Y}\right)^2}{\sum\left(Y - \bar{Y}\right)^2}$$

- Στην ουσία πρόκειται για το **τετράγωνο** του δείκτη συσχέτισης

25

Ένα Παράδειγμα

- r = .713
- r^2 = .713² =.508

Περίπου το 50% της διακύμανσης του αριθμού των θανάτων από καρκίνο του πνεύμονα σχετίζεται τη διακύμανση του αριθμού των τσιγάρων

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,713[a]	,508	,482	4,81640

a. Predictors: (Constant), CIGAR

26

Έλεγχος Υποθέσεων

- **Μηδενική Υπόθεση**
  - $b^* = 0$
  - $\alpha^* = 0$
    - Συνήθως επικεντρωνόμαστε μόνο στο b
  - **Δείκτης συσχέτισης πληθυσμού** (ρ) = 0
    - Ο κλασικός έλεγχος υποθέσεων για το δείκτη συσχέτισης

27

Στατιστικός Έλεγχος για την Κλίση

Αυτός γίνεται με τη χρήση του στατιστικού **κριτηρίου *t***. Στη συγκεκριμένη περίπτωση δεχόμαστε την εναλλακτική υπόθεση που θέλει την κλίση να είναι στατιστικά σημαντική ($p < 0.05$). Αυτό σημαίνει ότι η επίδραση του αριθμού των τσιγάρων στον καρκίνο του πνεύμονα **δεν είναι μηδενική**. Άρα, υπάρχει επίδραση της προβλεπτικής μεταβλητής (αριθμός τσιγάρων την ημέρα) στο ποσοστό των θανάτων από καρκίνο του πνεύμονα)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,367	2,941		,805	,431
	CIGAR	2,042	,461	,713	4,426	,000

a. Dependent Variable: CHD

Ανάλυση Πολλαπλής Παλινδρόμησης

- Όταν έχουμε **περισσότερες** από μια προβλεπτικές μεταβλητές

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_n X_n$$

- Στόχος είναι να υπολογίσουμε τη συνολική **προβλεπτική ικανότητα** των μεταβλητών που χρησιμοποιούμε

29

Ο δείκτης πολλαπλής Συσχέτισης R

- Είναι ο δείκτης που μας δείχνει το δείκτη συσχέτισης της μεταβλητής κριτηρίου με όλες τις προβλεπτικές μεταβλητές **ταυτόχρονα**

- Υψώνοντας στο τετράγωνο το δείκτη R ($\mathbf{R^2}$), μπορούμε να εκτιμήσουμε το ποσοστό της **συνολικής διακύμανσης** που ερμηνεύουν οι προβλεπτικές μεταβλητές που μελετάμε

30

Τυποποιημένοι Συντελεστές παλινδρόμησης β (beta)

- Είναι οι συντελεστές (εκφρασμένοι σε z-τιμές) που μας πληροφορούν για το **βαθμό σπουδαιότητας** της κάθε προβλεπτικής μεταβλητής

- Για να μετατρέψουμε τους συντελεστές παλινδρόμησης σε τυποποιημένους συντελεστές παλινδρόμησης χρησιμοποιούμε τον παρακάτω τύπο

$$\hat{\beta} = b\left(\frac{s_X}{s_Y}\right)$$

31

Ένα Παράδειγμα

- Η επίδραση της **προσωπικότητας** στη **διοικητική αποτελεσματικότητα** ενός μάνατζερ

- **Μεταβλητή Κριτήριο**
  - **Διοικητική Αποτελεσματικότητα**
- **Προβλεπτικές Μεταβλητές**
  - **Εξωστρέφεια**
  - **Νευρωτισμός**
  - **Δεκτικότητα στην Εμπειρία**
  - **Προσήνεια**
  - **Ευσυνειδησία**

32

Τα αποτελέσματα από το SPSS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,409[a]	,167	,157	14,67999

a. Predictors: (Constant), Ευσυνειδησία, Δεκτικότητα στην Εμπειρία, Προσήνεια, Νευρωτισμός, Εξωστρέφεια

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	133,000	10,944		12,153	,000
	Εξωστρέφεια	,165	,115	,076	1,440	,151
	Νευρωτισμός	,111	,088	,064	1,263	,207
	Δεκτικότητα στην Εμπειρία	,220	,099	,105	2,231	,026
	Προσήνεια	,708	,118	,293	6,003	,000
	Ευσυνειδησία	,291	,108	,141	2,698	,007

a. Dependent Variable: Δ.Α (Δ.Α. Συνολική Βαθμολογία - Στάσεις)

Συμπεράσματα από την Ανάλυση

- Το συνολικό ποσοστό της διακύμανσης που ερμηνεύει η προσωπικότητα είναι **17%**
- Από τους πέντε παράγοντες, οι μόνοι που είναι στατιστικά σημαντικοί (και συνεισφέρουν στην πρόβλεψη της διοικητικής αποτελεσματικότητας) ήταν τρεις: **Δεκτικότητα, Προσήνεια, Ευσυνειδησία**
- Από τις τρεις μεταβλητές που επηρεάζουν τη διοικητική αποτελεσματικότητα, η πιο σημαντική είναι η **Προσήνεια** (beta = .29), η επόμενη ήταν η **Ευσυνειδησία** (beta = .14), και τέλος η **Δεκτικότητα** (beta = .11)

34