ΚΕΦΑΛΑΙΟ 7

Παλινδρόμηση

Πρόλογος

Αν δύο μεταβλητές συσχετίζονται, η σχέση τους αυτή μπορεί να οδηγήσει σε πρόβλεψη. Για παράδειγμα, αν οι δεξιότητες στους υπολογιστές και μέσοι όροι βαθμολογίας συσχετίζονται, το επίπεδο των δεξιοτήτων στους υπολογιστές μπορεί να χρησιμοποιηθεί για την πρόβλεψη του μέσου όρου βαθμολογίας. Η ακρίβεια της πρόβλεψης αυξάνεται με την ισχύ της αναφερόμενης συσχέτισης.

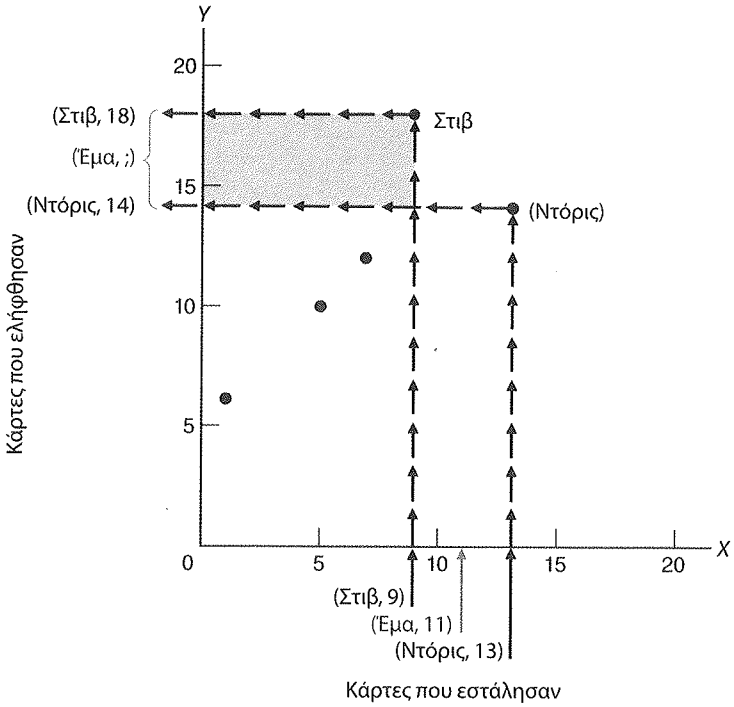
Θα μιλήσουμε επίσης για ένα κυρίαρχο στατιστικό φαινόμενο που αναφέρεται ως «παλινδρόμηση προς τον μέσο». Αυτό το φαινόμενο παρατηρείται συχνά σε υποσύνολα ακραίων παρατηρήσεων, όπως μετά την καλύτερη επίδοση επαγγελματιών αθλητών ή μετά από μια κακή επίδοση παιδιών με προβλήματα μάθησης. Αν ερμηνευτεί λανθασμένα ως πραγματική επίδραση, η παλινδρόμηση προς τον μέσο μπορεί να οδηγήσει σε εσφαλμένα συμπεράσματα.

Μια ανάλυση συσχέτισης της ανταλλαγής ευχετήριων καρτών από πέντε φίλους για την πιο πρόσφατη εορταστική περίοδο αναδεικνύει μια ισχυρή θετική σχέση μεταξύ των καρτών που εστάλησαν και των καρτών που ελήφθησαν. Όταν ενημερώνεται γι' αυτά τα αποτελέσματα, μια άλλη φίλη, η Έμα, η οποία αρέσκεται στο να λαμβάνει ευχετήριες κάρτες, σας ζητά να προβλέψετε πόσες κάρτες θα λάβει κατά την επόμενη εορταστική περίοδο, αν υποθέσουμε ότι σκοπεύει να στείλει 11 κάρτες.

7.1 Δύο γενικές προβλέψεις

Πρόβλεψη ενός «σχετικά μεγάλου αριθμού»

Θα μπορούσατε να προσφέρετε στην Έμα μια πολύ γενική και πρόχειρη πρόβλεψη σκεπτόμενοι ότι οι κάρτες που στέλνονται και οι κάρτες που λαμβάνονται τείνουν να κατέχουν *παρόμοιες* σχετικές θέσεις στις αντίστοιχες κατανομές τους. Επομένως, η Έμα μπορεί να περιμένει ότι θα λάβει έναν *σχετικά μεγάλο* αριθμό καρτών, καθώς σκοπεύει να στείλει έναν *σχετικά μεγάλο* αριθμό καρτών.



ΣΧΗΜΑ 7.1

Μια γενική πρόβλεψη για την Έμα (χρησιμοποιώντας τις κουκκίδες για τους Στιβ και Ντόρις).

Πρόβλεψη «μεταξύ 14 και 18 καρτών»

Για να λάβετε μια ελαφρώς πιο ακριβή πρόβλεψη για την Έμα, ανατρέξτε στο διάγραμμα διασποράς που δημιουργήθηκε για τους πέντε αρχικούς φίλους (βλ. Σχήμα 7.1). Παρατηρήστε ότι το σχέδιο της Έμα να στείλει 11 κάρτες την κατατάσσει στον άξονα *X* μεταξύ των 9 καρτών που έστειλε ο Στιβ και των 13 που έστειλε η Ντόρις. Χρησιμοποιώντας τις κουκκίδες για τον Στιβ και την Ντόρις ως οδηγούς, κατασκευάστε δύο σειρές από βέλη, με τη μία να ξεκινά στο 9 και να τελειώνει στο 18 για τον Στιβ και από το 13 ως το 14 για την Ντόρις. [Η κατεύθυνση των βελών αναπαριστά την προσπάθειά μας να προβλέψουμε τις κάρτες που λαμβάνονται (*Y*) από τις κάρτες που στέλνονται (*X*). Αν και δεν είναι απαραίτητο, συνηθίζουμε να προβλέπουμε από το *X* προς το *Y*.] Εστιάζοντας στο διάστημα κατά μήκος του άξονα *Y* μεταξύ των δύο σειρών βελών, θα μπορούσατε να προβλέψετε ότι η Έμα θα λάβει μεταξύ 14 και 18 καρτών, δηλαδή μεταξύ των αριθμών καρτών που έλαβαν οι Ντόρις και Στιβ.

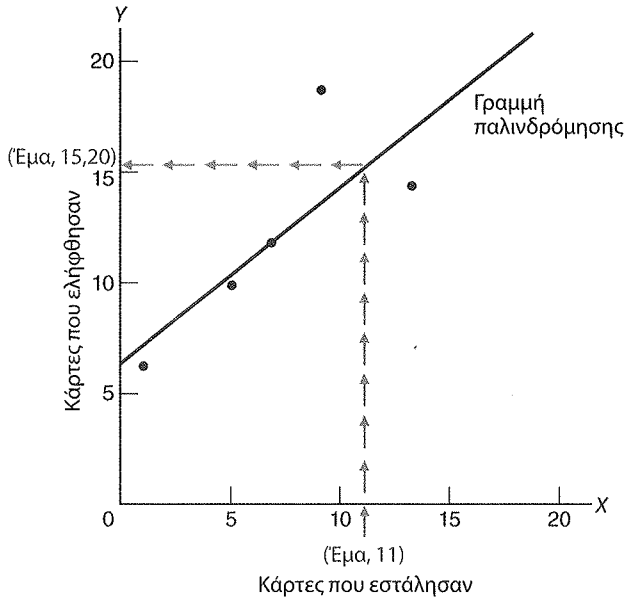
Η τελευταία πρόβλεψη μπορεί να ικανοποιήσει την Έμα, αλλά σίγουρα δεν αξίζει μνείας στα στατιστικά ιστορικά. Αν και οι πέντε κουκκίδες του Σχήματος 7.1 παρέχουν πολύτιμες πληροφορίες για την ανταλλαγή ευχετήριων καρτών, η πρόβλεψή μας για την Έμα βασίζεται μόνο στις δύο κουκκίδες για τους Στιβ και Ντόρις.

7.2 Η γραμμή παλινδρόμησης

Οι πέντε κουκκίδες συνεισφέρουν σε μια πιο ακριβή πρόβλεψη, όπως παρουσιάζεται στο Σχήμα 7.2, σύμφωνα με την οποία η Έμα θα λάβει 15,20 κάρτες. Κοιτάξτε πιο προσεκτικά τη συνεχή γραμμή που ορίζεται ως γραμμή παλινδρόμησης στο Σχήμα 7.2, η οποία καθοδηγεί τη σειρά βελών, ξεκινώντας από το 11 και φτάνοντας στην προβλεπόμενη τιμή 15,20. Η γραμμή παλινδρόμησης είναι ευθεία και όχι καμπύλη γραμμή εξαιτίας της γραμμικής σχέσης μεταξύ των καρτών που εστάλησαν και των καρτών που ελήφθησαν. Όπως θα αποκαλυφθεί αργότερα, μπορεί να χρησιμοποιηθεί κατ' επανάληψη για την πρόβλεψη των καρτών που λαμβάνονται. Ανεξάρτητα από το αν η Έμα αποφασίσει να στείλει 5, 15 ή 25 κάρτες, αυτή η γραμμή θα οδηγεί μια νέα σειρά βελών, ξεκινώντας από το 5 ή το 15 ή το 25, προς μια νέα προβλεπόμενη τιμή κατά μήκος του άξονα *Y*.

Τοποθέτηση της γραμμής

Προς το παρόν, ξεχάστε οποιαδήποτε πρόβλεψη για την Έμα και επικεντρωθείτε στο πώς οι πέντε κουκκίδες υπαγορεύουν την τοποθέτηση της γραμμής παλινδρόμησης. Αν και οι πέντε κουκκίδες όριζαν μία μόνο ευθεία γραμμή, η τοποθέτηση της γραμμής παλινδρόμησης θα ήταν απλή: Απλώς θα την αφήνατε να διέλθει από όλες τις κουκκίδες. Όταν οι κουκκίδες δεν ορίζουν μία μόνο ευθεία γραμμή, όπως στο διάγραμμα διασποράς για τους πέντε φίλους, η τοποθέτηση της γραμμής παλινδρόμησης αναφέρεται σαν να παριστάνει μια «συμβιβαστική» λύση. Περνά μέσα από την κύρια ομάδα, πιθανώς εφαπτόμενη από κάποιες κουκκίδες αλλά όχι απ' όλες.



ΣΧΗΜΑ 7.2

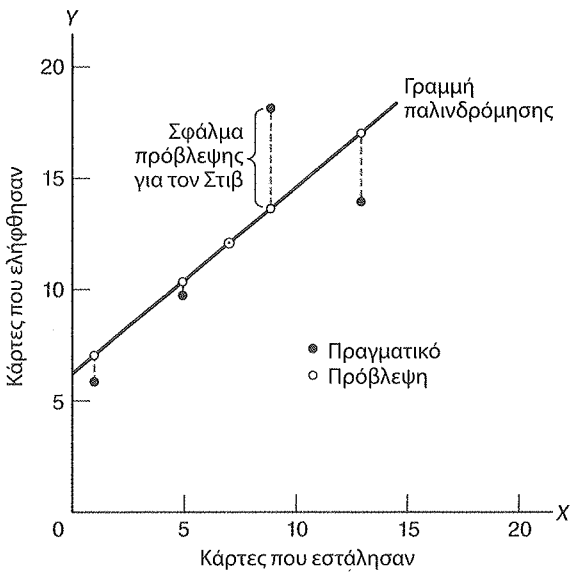
Πρόβλεψη 15,20 για την Έμα (χρησιμοποιώντας τη γραμμή παλινδρόμησης).

Σφάλματα πρόβλεψης

Το Σχήμα 7.3 παρουσιάζει τα σφάλματα πρόβλεψης που θα παρατηρούσατε αν η γραμμή παλινδρόμησης είχε χρησιμοποιηθεί για την πρόβλεψη του αριθμού των καρτών που λαμβάνονται από τους πέντε φίλους. Οι γεμάτες κουκκίδες δείχνουν τον *πραγματικό* αριθμό καρτών που λαμβάνονται και οι χωρίς περιεχόμενο κουκκίδες, οι οποίες βρίσκονται πάντα κατά μήκος της γραμμής παλινδρόμησης, είναι ο *προβλεπόμενος* αριθμός καρτών που στέλνονται. (Για να αποφύγουμε τη σύγχυση στο Σχήμα 7.3, έχουμε παραλείψει τα βέλη. Ωστόσο, θα ήταν χρήσιμο να σκεφτείτε ότι υπάρχουν βέλη που καταλήγουν στον άξονα *Y*, για κάθε κουκκίδα, γεμάτη ή άδεια.) Το μεγαλύτερο σφάλμα πρόβλεψης, το οποίο αναδεικνύεται από μια διακεκομμένη κάθετη γραμμή, παρατηρείται για τον Στιβ, ο οποίος έστειλε 9 κάρτες. Αν και έλαβε 18 κάρτες, θα έπρεπε να είχε λάβει λίγο λιγότερες από 14 σύμφωνα με τη γραμμή παλινδρόμησης. Το μικρότερο σφάλμα πρόβλεψης παρατηρείται για τον Μάικ, ο οποίος έστειλε 7 κάρτες. Ο Μάικ έλαβε 12 κάρτες, ακριβώς όσες θα έπρεπε σύμφωνα με τη γραμμή παλινδρόμησης.

Συνολικό σφάλμα πρόβλεψης

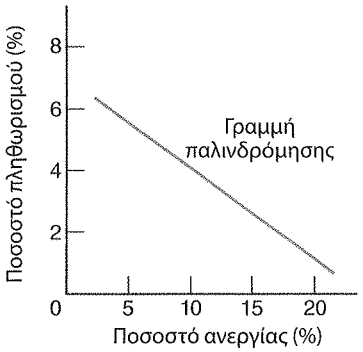
Σε αυτό το σημείο αναφερόμαστε στη «φαινομενικά» ανόητη διαδικασία να προβλέψουμε όσα γνωρίζουμε ήδη ότι είναι πραγματικά για τους πέντε φίλους, έτσι ώστε να βεβαιωθούμε για την ακρίβεια των αποπειρών μας να προβλέψουμε. Όσο μικρότερο είναι το σύνολο όλων των σφαλμάτων πρόβλεψης στο Σχήμα 7.3, τόσο πιο ευνοϊκή θα είναι η πρόγνωση για τις προβλέψεις μας. Είναι σαφές ότι το ζητούμενο για τη γραμμή παλινδρόμησης είναι να βρίσκεται σε μια θέση η οποία *ελαχιστοποιεί* το συνολικό σφάλμα πρόβλεψης, δηλαδή ελαχιστοποιεί το σύνολο των κάθετων ασυμφωνιών μεταξύ των γεμάτων και των άδειων κουκκίδων του Σχήματος 7.3.



ΣΧΗΜΑ 7.3

Σφάλματα πρόβλεψης.

Έλεγχος προόδου *7.1 Για να βεβαιωθείτε ότι έχετε καταλάβει όσα είπαμε στο πρώτο μέρος αυτού του κεφαλαίου, κάντε προβλέψεις χρησιμοποιώντας αυτό το διάγραμμα.



(α) Προβλέψτε το κατά προσέγγιση ποσοστό πληθωρισμού, δεδομένου ποσοστού ανεργίας της τάξης 5%.
(β) Προβλέψτε το κατά προσέγγιση ποσοστό πληθωρισμού, δεδομένου ποσοστού ανεργίας της τάξης 15%.
Απαντήσεις στη σελίδα 538.

7.3 Γραμμή παλινδρόμησης ελαχίστων τετραγώνων

Για να αποφευχθεί η αριθμητική απόκλιση στο μηδέν που παράγεται πάντα όταν προσθέτουμε θετικά και αρνητικά σφάλματα πρόβλεψης (τα οποία σχετίζονται με σφάλματα πάνω και κάτω από τη γραμμή παλινδρόμησης αντίστοιχα), *η τοποθέτηση της γραμμής παλινδρόμησης ελαχιστοποιεί* όχι το συνολικό σφάλμα πρόβλεψης αλλά *το συνολικό τετραγωνικό σφάλμα πρόβλεψης*, δηλαδή το σύνολο για όλα τα τετραγωνικά σφάλματα πρόβλεψης. Όταν βρίσκεται μ' αυτόν τον τρόπο, η *γραμμή παλινδρόμησης* αναφέρεται συχνά ως η *γραμμή παλινδρόμησης ελαχίστων τετραγώνων*. Αν και είναι πιο δύσκολο να το φανταστούμε, αυτή η προσέγγιση συνάδει με τον αρχικό στόχο, δηλαδή την ελαχιστοποίηση του συνολικού σφάλματος πρόβλεψης ή κάποιας εκδοχής του συνολικού σφάλματος πρόβλεψης, παρέχοντας επομένως μια πιο ευνοϊκή πρόγνωση για τις προβλέψεις μας.

Χρειαζόμαστε μια μαθηματική λύση

Χωρίς τη βοήθεια των μαθηματικών, η αναζήτηση μιας γραμμής παλινδρόμησης ελαχίστων τετραγώνων δεν θα μπορούσε παρά μόνο να προκαλέσει απογοήτευση. Τα διαγράμματα διασποράς θα γίνονταν πεδία γεμάτα με αβέβαιες δοκιμαστικές γραμμές παλινδρόμησης, οι οποίες θα απορρίπτονταν εξαιτίας των υπερβολικά μεγάλων συνόλων τους για τις τετραγωνικές ασυμφωνίες. Ακόμα και η πιο χρονοβόρα και συνειδητή προσπάθεια θα κορυφωνόταν σε μια (αρκετά καλή) προσέγγιση στη γραμμή παλινδρόμησης ελαχίστων τετραγώνων.

Εξίσωση παλινδρόμησης ελαχίστων τετραγώνων

Μια εξίσωση απεικονίζει την ακριβή γραμμή παλινδρόμησης ελαχίστων τετραγώνων για κάθε διάγραμμα διασποράς. Πιο γενικά, αυτή η εξίσωση είναι:

Εξίσωση παλινδρόμησης ελαχίστων τετραγώνων

$$Y' = bX + a \qquad (7.1)$$

όπου το Y' αναπαριστά την προβλεπόμενη τιμή (τον προβλεπόμενο αριθμό καρτών που θα ληφθούν από οποιονδήποτε νέο φίλο, όπως την Έμα), το X αναπαριστά τη γνωστή τιμή (τον γνωστό αριθμό καρτών που έστειλε ο νέος φίλος) και τα b και a αναπαριστούν αριθμούς που υπολογίζονται από την αρχική ανάλυση συσχέτισης, όπως περιγράφουμε στη συνέχεια.[18]

18. Ενδεχομένως να αναγνωρίσετε ότι η εξίσωση ελαχίστων τετραγώνων αναπαριστά μια ευθεία γραμμή με κλίση b και την τεταγμένη επί της αρχής των αξόνων να είναι a.

Εύρεση των τιμών b και a

Για να λάβετε μια εξίσωση παλινδρόμησης, θα πρέπει να λύσετε τις παρακάτω εκφράσεις, πρώτα ως προς *b* και έπειτα ως προς *a*, χρησιμοποιώντας δεδομένα από την αρχική ανάλυση συσχέτισης που δίνονται παρακάτω. Η έκφραση για το *b* δίνει:

Λύση ως προς *b*

$$b = r\sqrt{\frac{SS_y}{SS_x}} \qquad (7.2)$$

όπου το *r* αναπαριστά τη συσχέτιση μεταξύ *X* και *Y* (κάρτες που εστάλησαν και ελήφθησαν από τους πέντε φίλους), το SS_y είναι το άθροισμα τετραγώνων για όλα τα αποτελέσματα *Y* (οι κάρτες που ελήφθησαν από τους πέντε φίλους) και το SS_x το άθροισμα τετραγώνων για όλα τα αποτελέσματα *X* (οι κάρτες που εστάλησαν από τους πέντε φίλους).

Η έκφραση για το *a* είναι:

Λύση ως προς *a*

$$a = \overline{Y} - b\overline{X} \qquad (7.3)$$

όπου τα $\overline{Y}$ και $\overline{X}$ αναφέρονται στους δειγματικούς μέσους για όλες τις μεταβλητές *Y* και *X* αντίστοιχα και το *b* ορίζεται από την προηγούμενη έκφραση.

Οι τιμές όλων των όρων στις εκφράσεις για τα *b* και *a* μπορούν να ληφθούν από την αρχική ανάλυση συσχέ-

Πίνακας 7.1
ΠΡΟΣΔΙΟΡΙΣΜΟΣ ΤΗΣ ΕΞΙΣΩΣΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

A. Σειρά πράξεων

Προσδιορισμός των τιμών των SS_x, SS_y, και *r* (1) με αναφορά στην αρχική ανάλυση συσχέτισης στον Πίνακα 6.3.
Αντικατάσταση με αριθμούς στον τύπο (2) και λύση ως προς *b*.
Εκχώρηση τιμών στα $\overline{X}$ και $\overline{Y}$ (3) με αναφορά στην αρχική ανάλυση συσχέτισης στον Πίνακα 6.3.
Αντικατάσταση με αριθμούς στον τύπο (4) και λύση ως προς *a*.
Αντικατάσταση με αριθμούς για τα *b* και *a* στην εξίσωση παλινδρόμησης ελαχίστων τετραγώνων (5).

B. Πράξεις

1 $SS_x = 80^*$

$SS_y = 80^*$

$r = 0{,}80$

2 $b = r\sqrt{\frac{SS_y}{SS_x}} = 0{,}80\sqrt{\frac{80}{80}} = 0{,}80$

3 $\overline{X} = 7^{**}$

$\overline{Y} = 12^{**}$

4 $a = \overline{Y} - (b)(\overline{X}) = 12 - (0{,}80)(7) = 12 - 5{,}60 = 6{,}40$

5 $Y' = (b)(X) + a$

$= (0{,}80)(X) + 6{,}40$

** Οι πράξεις δεν εμφανίζονται. Επαληθεύστε, αν θέλετε, χρησιμοποιώντας τον Τύπο 4.4.*

*** Οι πράξεις δεν εμφανίζονται. Επαληθεύστε, αν θέλετε, χρησιμοποιώντας τον Τύπο 3.1.*

τισης έμμεσα, όπως με την τιμή του *r*, ή άμεσα, όπως με τις τιμές των υπόλοιπων όρων: $SS_{y'}$, $SS_{x'}$, $\overline{Y}$ και $\overline{X}$. Ο Πίνακας 7.1 παρουσιάζει τη σειρά πράξεων που παράγει μια εξίσωση παλινδρόμησης ελαχίστων τετραγώνων για το παράδειγμα των ευχετήριων καρτών, δηλαδή

$$Y' = 0{,}80(X) + 6{,}40$$

όπου τα 0,80 και 6,40 αναπαριστούν τις τιμές που υπολογίζονται για τα *b* και *a* αντίστοιχα.

Εξίσωση παλινδρόμησης ελαχίστων τετραγώνων
Η εξίσωση που ελαχιστοποιεί το σύνολο όλων των τετραγωνικών σφαλμάτων πρόβλεψης για γνωστά αποτελέσματα Y στην αρχική ανάλυση συσχέτισης.

Βασική ιδιότητα

Όταν υπολογιστούν και βρεθούν οι αριθμοί στα *b* και *a* όπως περιγράψαμε, η **εξίσωση παλινδρόμησης ελαχίστων τετραγώνων** εμφανίζεται ως μια εξίσωση με μια ιδιαίτερα επιθυμητή ιδιότητα: *Ελαχιστοποιεί αυτόματα το σύνολο όλων των τετραγωνικών σφαλμάτων πρόβλεψης για γνωστά αποτελέσματα Y στην αρχική ανάλυση συσχέτισης.*

Επίλυση ως προς Y′

Στην παρούσα μορφή της, η εξίσωση παλινδρόμησης μπορεί να χρησιμοποιηθεί για την πρόβλεψη του αριθμού των καρτών που θα λάβει η Έμα, αν υποθέσουμε ότι σκοπεύει να στείλει 11 κάρτες. Απλώς αντικαταστήστε με 11 το *X* και λύστε ως προς *Y′* ως εξής:

$$\begin{aligned} Y' &= 0{,}80(11) + 6{,}40 \\ &= 8{,}80 + 6{,}40 \\ &= 15{,}20 \end{aligned}$$

Πίνακας 7.2
Η ΠΡΟΒΛΕΨΗ ΓΙΑ ΚΑΡΤΕΣ ΠΟΥ ΛΑΜΒΑΝΟΝΤΑΙ (*Y′*) ΓΙΑ ΔΙΑΦΟΡΕΤΙΚΕΣ ΚΑΡΤΕΣ ΠΟΥ ΣΤΕΛΝΟΝΤΑΙ (*X*)

X	*Y′*
0	6,40
4	9,60
8	12,80
10	14,40
12	16,00
20	22,40
30	30,40

Παρατηρήστε ότι ο προβλεπόμενος αριθμός καρτών που θα λάβει η Έμα, 15,20, θεωρείται γνήσια πρόβλεψη, δηλαδή μια πρόβλεψη ενός άγνωστου συμβάντος με βάση πληροφορίες που υπάρχουν από κάποιο γνωστό συμβάν. Αυτή η πρόβλεψη εμφανίστηκε νωρίτερα στο Σχήμα 7.2.

Η εξίσωση παλινδρόμησης παρέχει μια ανεξάντλητη πηγή προβλέψεων για την ανταλλαγή καρτών. Κάθε πρόβλεψη εμφανίζεται εάν απλώς ορίσουμε κάποια τιμή για το *X* και λύσουμε την εξίσωση ως προς *Y′*, όπως περιγράφεται παραπάνω. Ο Πίνακας 7.2 παραθέτει τις προβλεπόμενες λήψεις καρτών για διάφορους αριθμούς καρτών. Επαληθεύστε ότι μπορείτε να λάβετε μερικές από τις τιμές *Y′* που βλέπετε στον Πίνακα 7.2 με βάση την εξίσωση παλινδρόμησης.

Σημειώστε ότι, ακόμα κι όταν δεν γίνεται καμία αποστολή καρτών ($X = 0$), προβλέπουμε ότι θα ληφθούν 6,40 κάρτες εξαιτίας της τιμής του *a*. Επίσης, παρατηρήστε ότι η αποστολή κάθε επιπλέον κάρτας μεταφράζεται σε αύξηση μόλις 0,80 στον αριθμό των προβλεπόμενων καρτών που θα ληφθούν, εξαιτίας της τιμής του *b*. Με άλλα λόγια, όποτε το *b* έχει τιμή μικρότερη από 1,00, οι αυξήσεις στις προβλεπόμενες λήψεις υστερούν –κατά μια ποσότητα που ισούται με την τιμή του *b*, ή 0,80 εν προκειμένω– από τον αριθμό των καρτών που εστάλησαν. Αν η τιμή του *b* ήταν μεγαλύτερη από 1,00, τότε οι αυξήσεις στις προβλεπόμενες λήψεις θα υπερέβαιναν τις αυξήσεις στις αποστολές καρτών. (Αν η τιμή του *b* ήταν αρνητική, εξαιτίας μιας υποκείμενης αρνητικής συσχέτισης, τότε η αποστολή επιπλέον καρτών θα είχε προκαλέσει μειώσεις και όχι αυξήσεις στις προβλεπόμενες λήψεις – και η παράδοση της αποστολής ευχετήριων καρτών στις διακοπές πιθανώς θα εξαφανιζόταν.)

Ένας περιορισμός

Η Έμα μπορεί να μελετήσει αυτές τις προβλέψεις για λήψη καρτών πριν προχωρήσει σε μια συγκεκριμένη επένδυση σε κάρτες. Ωστόσο, αυτή η στρατηγική δεν είναι απαραίτητο ότι θα αποδώσει, επειδή δεν υπάρχει καμία απόδειξη για την ύπαρξη απλής σχέσης *αιτίου-αιτιατού* μεταξύ καρτών που εστάλησαν και καρτών που ελήφθησαν. Η επιθυμητή επίδραση μπορεί να εξαφανιστεί αν, για παράδειγμα, η Έμα αυξήσει τη συνήθη αποστολή καρτών της και συμπεριλάβει απλούς γνωστούς, ή ακόμα και ξένους, πέρα από τους φίλους και συγγενείς.

Έλεγχος προόδου *7.2 Έστω ότι το r 0,30 αποτυπώνει τη σχέση μεταξύ μορφωτικού επιπέδου (αποφοίτηση από την υψηλότερη τάξη) και εκτιμώμενου αριθμού ωρών μελέτης εβδομαδιαίως. Πιο συγκεκριμένα:

Μορφωτικό επίπεδο (X)	Εβδομαδιαίος χρόνος μελέτης (Y)
$\bar{X} = 13$	$\bar{Y} = 8$
$SS_x = 25$	$SS_y = 50$
$r = 0{,}30$	

(α) Υπολογίστε την εξίσωση ελαχίστων τετραγώνων για την πρόβλεψη του εβδομαδιαίου χρόνου μελέτης από το μορφωτικό επίπεδο.
(β) Το μορφωτικό επίπεδο της Φέιθ είναι 15. Ποιος είναι ο προβλεπόμενος χρόνος μελέτης της;
(γ) Το μορφωτικό επίπεδο του Κίγκαν είναι 11. Ποιος είναι ο προβλεπόμενος χρόνος μελέτης του;
Απαντήσεις στη σελίδα 538.

Διαγράμματα ή εξισώσεις;

Αντλώντας έμπνευση από τα Σχήματα 7.2 και 7.3, ενδεχομένως να μπείτε στον πειρασμό να παραγάγετε προβλέψεις από διαγράμματα και όχι από εξισώσεις. Ωστόσο, αν δεν έχουν κατασκευαστεί με εμπειρία και γνώση, τα διαγράμματα δίνουν λιγότερο ακριβείς προβλέψεις από τις εξισώσεις. Μακροπρόθεσμα, είναι πιο ακριβές και εύκολο να παραγάγετε προβλέψεις από εξισώσεις.

7.4 Τυπικό σφάλμα εκτίμησης, $s_{y|x}$

Αν και προβλέψαμε ότι η επένδυση της Έμα σε 11 κάρτες θα της φέρει 15,20 κάρτες, θα αποτελούσε έκπληξη αν πράγματι λάμβανε 15 κάρτες. Είναι πιο πιθανό ότι, εξαιτίας της ατελούς σχέσης μεταξύ καρτών που εστάλησαν και καρτών που ελήφθησαν, η Έμα θα λάβει διαφορετικό αριθμό καρτών από 15. Αν και έχει σχεδιαστεί έτσι ώστε να ελαχιστοποιεί το σφάλμα πρόβλεψης, η εξίσωση ελαχίστων τετραγώνων δεν το εξαλείφει εντελώς. Επομένως, θα πρέπει στη συνέχεια να εκτιμήσουμε την ποσότητα σφάλματος που εμπεριέχουν οι προβλέψεις μας. Όσο μικρότερο είναι το εκτιμώμενο σφάλμα, τόσο καλύτερη θα είναι η πρόγνωση για τις προβλέψεις μας.

Εύρεση του τυπικού σφάλματος εκτίμησης

Η εκτίμηση σφάλματος για νέες προβλέψεις αποτυπώνει την αστοχία μας στην πρόβλεψη του αριθμού των καρτών που λαμβάνονται από τους αρχικούς πέντε φίλους, όπως αποτυπώνεται από τις ασυμφωνίες μεταξύ γεμάτων και άδειων κουκκίδων στο Σχήμα 7.3. Γνωστή ως *τυπικό σφάλμα εκτίμησης* και με το σύμβολο $s_{y|x}$, αυτή η εκτίμηση σφάλματος πρόβλεψης τηρεί τη γενική μορφή για οποιαδήποτε τυπική απόκλιση δείγματος, δηλαδή την τετραγωνική ρίζα ενός όρου αθροίσματος τετραγώνων διά των βαθμών ελευθερίας του. (Βλ. Τύπο 4.10 στη σελίδα 120.) Ο τύπος για το $s_{y|x}$ έχει ως εξής:

Τυπικό σφάλμα εκτίμησης (τύπος ορισμού)

$$s_{y|x} = \sqrt{\frac{SS_{y|x}}{n-2}} = \sqrt{\frac{\sum(Y-Y')^2}{n-2}} \qquad (7.4)$$

όπου ο όρος του αθροίσματος τετραγώνων στον αριθμητή, $SS_{y|x}$, αναπαριστά το άθροισμα των τετραγώνων για σφάλματα πρόβλεψης, $Y - Y'$, και ο όρος για τους βαθμούς ελευθερίας στον παρονομαστή, $n - 2$, αναπαριστά την απώλεια δύο βαθμών ελευθερίας επειδή οποιαδήποτε ευθεία γραμμή, συμπεριλαμβανομένης της γραμμής παλινδρόμησης, μπορεί να κατασκευαστεί έτσι ώστε να συμπίπτει με δύο σημεία δεδομένων. Το σύμβολο $s_{y|x}$ διαβάζεται ως «s υπό y δεδομένου του x».

Αν και μπορούμε να εκτιμήσουμε το συνολικό σφάλμα πρόβλεψης απευθείας από τα σφάλματα πρόβλεψης, $Y - Y'$, είναι πιο αποδοτικό να χρησιμοποιήσουμε τον παρακάτω τύπο υπολογισμού:

Τυπικό σφάλμα εκτίμησης (τύπος υπολογισμού)

$$s_{y|x} = \sqrt{\frac{SS_y(1 - r^2)}{n - 2}} \qquad (7.5)$$

όπου το SS_y είναι το άθροισμα των τετραγώνων για αποτελέσματα Y (κάρτες που ελήφθησαν από τους πέντε φίλους), δηλαδή

$$SS_y = \Sigma(Y - \overline{Y}) = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$$

και r είναι ο συντελεστής συσχέτισης (κάρτες που εστάλησαν και ελήφθησαν).

Τυπικό σφάλμα εκτίμησης ($s_{y|x}$) *Ένα γενικό μέτρο της μέσης ποσότητας του σφάλματος πρόβλεψης.*

Βασική ιδιότητα

Το **τυπικό σφάλμα εκτίμησης** αναπαριστά έναν ειδικό τύπο τυπικής απόκλισης που αντανακλά το μέγεθος του σφάλματος πρόβλεψης.

Θα ήταν χρήσιμο να θεωρούμε το τυπικό σφάλμα εκτίμησης, $s_{y|x}$, ως ένα γενικό μέτρο της μέσης ποσότητας του σφάλματος πρόβλεψης – δηλαδή ένα γενικό μέτρο της μέσης ποσότητας κατά την οποία γνωστές τιμές Y αποκλίνουν από τις προβλεπόμενες τιμές τους Y'.[19]

Η τιμή του 3,10 για το $s_{y|x}$, όπως υπολογίζεται στον Πίνακα 7.3, αναπαριστά την τυπική απόκλιση για τις ασυμφωνίες μεταξύ του γνωστού αριθμού καρτών που λαμβάνονται και της πρόβλεψης γι' αυτόν τον αριθμό, όπως αρχικά παρουσιάστηκε στο Σχήμα 7.3. Στον ρόλο της ως εκτίμησης σφάλματος πρόβλεψης, η τιμή του $s_{y|x}$ μπορεί να χρησιμοποιηθεί σε κάθε νέα πρόβλεψη. Ως εκ τούτου, μια περιεκτική δήλωση πρόβλεψης θα μπορούσε να είναι αυτή: «Η πρόβλεψη για τον αριθμό των καρτών που θα λάβει η Έμα ισούται με 15,20 ± 3,10», όπου ο τελευταίος όρος αποτελεί μια γενική εκτίμηση της μέσης ποσότητας του σφάλματος πρόβλεψης, δηλαδή τη μέση ποσότητα κατά την οποία το 15,20 θα είναι μια εκτίμηση πάνω ή κάτω από τον αριθμό των καρτών που τελικά θα λάβει η Έμα.

Πίνακας 7.3
ΥΠΟΛΟΓΙΣΜΟΣ ΤΟΥ ΤΥΠΙΚΟΥ ΣΦΑΛΜΑΤΟΣ ΕΚΤΙΜΗΣΗΣ, $s_{y|x}$

Α. Ακολουθία πράξεων

Εκχωρείτε τιμές στα SS_y και r (1) με αναφορά σε όσα έγιναν προηγουμένως με την εξίσωση παλινδρόμησης ελαχίστων τετραγώνων στον Πίνακα 7.1.

Αντικαταστήστε με αριθμούς στον τύπο (2) και λύστε ως προς $s_{y|x}$.

Β. Πράξεις

1 $SS_y = 80$

$r = 0{,}80$

2 $$s_{y|x} = \sqrt{\frac{SS_y(1 - r^2)}{n - 2}} = \sqrt{\frac{80(1 - [0{,}80]^2)}{5 - 2}} = \sqrt{\frac{80(0{,}36)}{3}} = \sqrt{\frac{28{,}80}{3}} = \sqrt{9{,}60} = 3{,}10$$

19. Αυστηρά μιλώντας, το τυπικό σφάλμα εκτίμησης είναι μεγαλύτερο από το μέσο σφάλμα πρόβλεψης κατά 10% ως 20%. Ωστόσο, είναι λογικό να αποτυπώνουμε το τυπικό σφάλμα μ' αυτόν τον τρόπο – αρκεί να θυμάστε ότι, όπως με τον αντίστοιχο ορισμό για την τυπική απόκλιση στο Κεφάλαιο 4, περιλαμβάνεται μια κάποια προσέγγιση.

Η σημασία του r

Για να εκτιμήσουμε τη σημασία του συντελεστή συσχέτισης σε οποιαδήποτε προσπάθεια πρόβλεψης, θα αντικαταστήσουμε το *r* με μερικές τιμές στον αριθμητή του Τύπου 7.5 και θα παρατηρήσουμε την τελική επίδραση στο άθροισμα τετραγώνων για τα σφάλματα πρόβλεψης, $SS_{y|x}$. Αντικαθιστώντας με τιμή 1 το *r*, παίρνουμε

$$SS_{y|x} = SS_y (1 - r^2) = SS_y[1 - (1)^2] = SS_y[1 - 1] = SS_y[0] = 0$$

Όπως αναμενόταν, όταν οι προβλέψεις βασίζονται σε τέλειες σχέσεις, το άθροισμα τετραγώνων για σφάλματα πρόβλεψης ισούται με μηδέν και δεν υπάρχει σφάλμα πρόβλεψης. Στο άλλο άκρο, αντικαθιστώντας με τιμή 0 το *r* στον αριθμητή του Τύπου 7.5, παίρνουμε

$$SS_{y|x} = SS_y (1 - r^2) = SS_y[1 - (0)^2] = SS_y[1 - 0] = SS_y[1] = SS_y$$

Και πάλι, όπως αναμενόταν, όταν οι προβλέψεις βασίζονται σε μια μη υπαρκτή σχέση, το άθροισμα τετραγώνων για σφάλματα πρόβλεψης ισούται με SS_y, δηλαδή το άθροισμα τετραγώνων των αποτελεσμάτων *Y* γύρω από το $\overline{Y}$, και δεν υπάρχει καμία μείωση στο σφάλμα πρόβλεψης. Είναι σαφές ότι η πρόγνωση για μια απόπειρα πρόβλεψης είναι πολύ ευνοϊκή όταν οι προβλέψεις βασίζονται σε ισχυρές σχέσεις, όπως δείχνει μια ευμεγέθης θετική ή αρνητική τιμή του *r*. Η πρόγνωση είναι μάλλον ζοφερή –και δεν πρέπει καν να αποπειραθούμε μια προσπάθεια πρόβλεψης– όταν οι προβλέψεις πρέπει να βασίζονται σε ασθενείς ή μη υπαρκτές σχέσεις, όπως συμβαίνει με μια τιμή του *r* κοντά στο 0.

Έλεγχος προόδου *7.3

(α) Υπολογίστε το τυπικό σφάλμα εκτίμησης για τα δεδομένα της Ερώτησης 7.2 στη σελίδα 187, αν υποθέσουμε ότι η συσχέτιση του 0,30 βασίζεται σε $n = 35$ ζεύγη παρατηρήσεων.

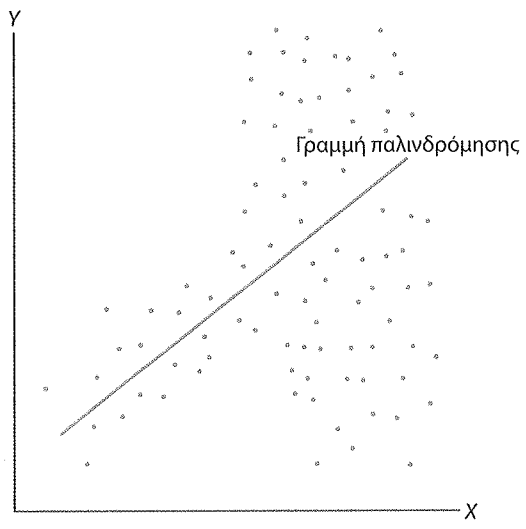
(β) Δώστε μια γενική ερμηνεία του τυπικού σφάλματος εκτίμησης.

Απαντήσεις στη σελίδα 538.

7.5 Υποθέσεις

Γραμμικότητα

Για να χρησιμοποιηθεί η εξίσωση παλινδρόμησης, η υπό εξέταση σχέση πρέπει να είναι γραμμική. Πρέπει να ανησυχείτε για τυχόν παραβιάσεις αυτής της υπόθεσης μόνο όταν το διάγραμμα διασποράς για την αρχική ανάλυση συσχέτισης αποκαλύπτει μια προφανώς έντονα κεκλιμένη ή καμπυλόγραμμη ομάδα κουκκίδων, όπως βλέπετε στο Σχήμα 6.4 στη σελίδα 163. Στην απίθανη περίπτωση μια ομάδα κουκκίδων να αποτυπώνει μια έντονη καμπυλόγραμμη τάση, θα πρέπει να ανατρέξετε σε πιο προχωρημένα βιβλία στατιστικής για να λάβετε κατάλληλες οδηγίες.



ΣΧΗΜΑ 7.4

Παραβίαση της υπόθεσης της ομοσκεδαστικότητας. (Οι κουκκίδες στερούνται ίσης μεταβλητότητας για όλα τα ευθύγραμμα τμήματα.)

Ομοσκεδαστικότητα

Για να χρησιμοποιηθεί το τυπικό σφάλμα εκτίμησης, $s_{y|x}$, εκτός από την τύχη, *οι κουκκίδες στο αρχικό διάγραμμα διασποράς θα διασκορπίζονται εξίσου σε όλα τα τμήματα της γραμμής παλινδρόμησης.* Πρέπει να ανησυχείτε για το αν θα υπάρχουν παραβιάσεις γι' αυτήν την υπόθεση, δηλαδή για την (όχι και τόσο λε-

κτικά φιλική φράση) *ομοσκεδαστικότητα*, μόνο όταν το διάγραμμα διασποράς αποκαλύπτει έναν πολύ διαφορετικό τύπο ομάδας κουκκίδων, όπως την ομάδα του Σχήματος 7.4. Κατ' ελάχιστον, το τυπικό σφάλμα εκτίμησης για τα δεδομένα στο Σχήμα 7.4 θα πρέπει να χρησιμοποιείται προσεκτικά, επειδή η τιμή του υπερεκτιμά τη μεταβλητότητα των κουκκίδων γύρω από το κάτω μισό της γραμμής παλινδρόμησης και υποτιμά τη μεταβλητότητα των κουκκίδων γύρω από το άνω μισό της γραμμής παλινδρόμησης.

7.6 Ερμηνεία του r^2

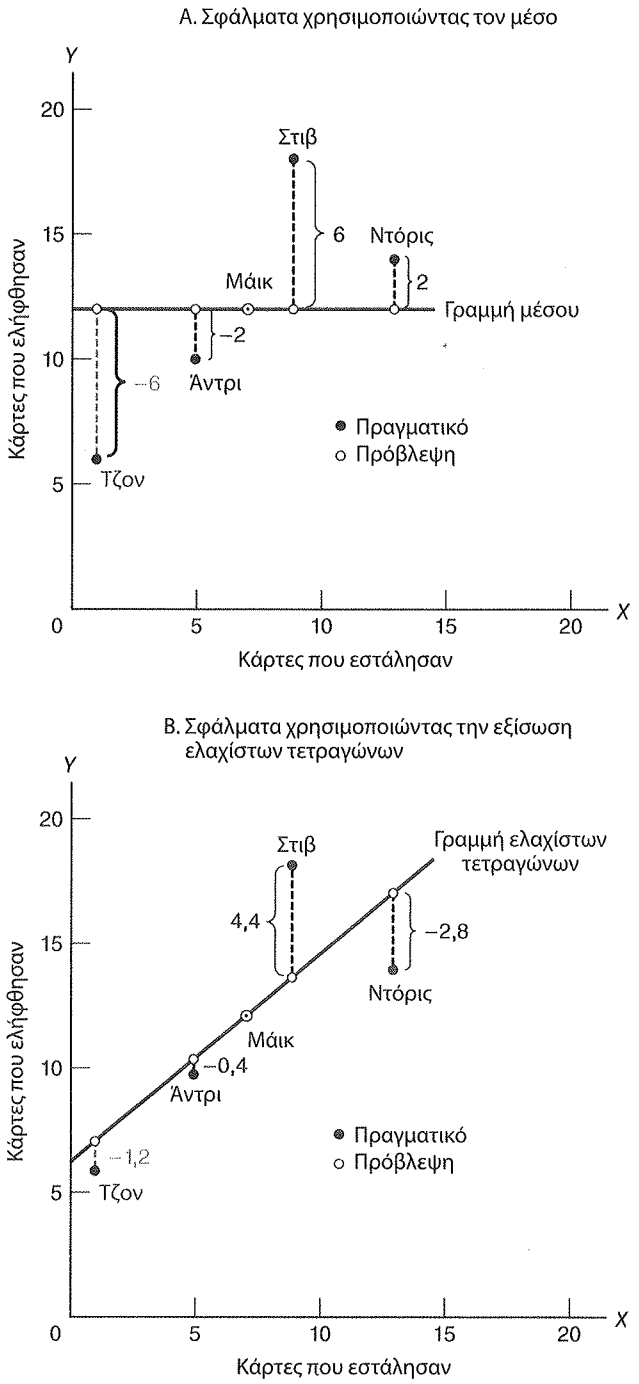
Ο τετραγωνικός συντελεστής συσχέτισης, r^2, μας παρέχει όχι μόνο μια σημαντική ερμηνεία του συντελεστή συσχέτισης, αλλά επίσης ένα μέτρο ακρίβειας της πρόβλεψης το οποίο συμπληρώνει το *τυπικό σφάλμα εκτίμησης*, $s_{y|x}$. Για να κατανοήσουμε τον συντελεστή r^2 πρέπει να επιστρέψουμε στο πρόβλημα της πρόβλεψης του αριθμού ευχετήριων καρτών που έλαβαν οι πέντε φίλοι στο Κεφάλαιο 6. (Μην ξεχνάτε ότι μπαίνουμε στη φαινομενικά ανόητη διαδικασία της πρόβλεψης μιας ποσότητας που ήδη γνωρίζουμε όχι με αυτοσκοπό την εύρεση αυτής της ποσότητας, αλλά ως έναν τρόπο ελέγχου της ακρίβειας της προσπάθειας πρόβλεψης.) Παραδόξως, ακόμα κι αν ο τελικός στόχος μας είναι να δείξουμε τη σχέση μεταξύ r^2 και ακρίβειας πρόβλεψης, θα επικεντρωθούμε αρχικά σε δύο είδη σφαλμάτων πρόβλεψης – σε εκείνα που οφείλονται στην επαναληπτική πρόβλεψη του μέσου και σε εκείνα που οφείλονται στην εξίσωση παλινδρόμησης.

Επαναλαμβανόμενη πρόβλεψη του μέσου

Για να υποστηρίξουμε τα παραπάνω, έστω ότι γνωρίζουμε τα δεδομένα της μεταβλητής *Y* (κάρτες που λαμβάνονται), αλλά όχι τα αντίστοιχα της *X* (κάρτες που στέλνονται), για τους πέντε φίλους. Η έλλειψη πληροφοριών για τη σχέση μεταξύ των δεδομένων των μεταβλητών *X* και *Y* δεν θα μας επέτρεπε να κατασκευάσουμε μια *εξίσωση παλινδρόμησης* και να τη χρησιμοποιήσουμε προκειμένου να παραγάγουμε μια προσαρμοσμένη πρόβλεψη, *Y'*, για κάθε φίλο. Θα μπορούσαμε όμως να υποστηρίξουμε μια αρχική προσπάθεια πρόβλεψης προβλέποντας *πάντα* τον μέσο, $\overline{Y}$, για καθένα από τα αποτελέσματα *Y* των πέντε φίλων. [Δεδομένων των περιορισμών που υπάρχουν τη δεδομένη στιγμή, οι στατιστικολόγοι προτείνουν επαναλαμβανόμενες προβλέψεις του μέσου, $\overline{Y}$, για διάφορους λόγους, συμπεριλαμβανομένου του γεγονότος ότι, αν και το σφάλμα πρόβλεψης για οποιοδήποτε άτομο μπορεί να είναι αρκετά μεγάλο, το *άθροισμα* των πέντε σφαλμάτων πρόβλεψης που προκύπτουν (αποκλίσεις των αποτελεσμάτων *Y* γύρω από το $\overline{Y}$) ισούται πάντα με μηδέν, όπως είπαμε στην Ενότητα 3.3.] Το πιο σημαντικό για τους σκοπούς της παρουσίασής μας είναι ότι χρησιμοποιώντας την επαναλαμβανόμενη πρόβλεψη του $\overline{Y}$ για κάθε αποτέλεσμα *Y* των δεδομένων για τους πέντε φίλους θα λάβουμε ένα *πλαίσιο αναφοράς ως προς το οποίο θα αξιολογούμε τις συνήθεις απόπειρές μας για πρόβλεψη* με βάση τη συσχέτιση μεταξύ των καρτών που εστάλησαν (*X*) και των καρτών που ελήφθησαν (*Y*). Οποιαδήποτε απόπειρα πρόβλεψης που αξιοποιεί μια υπάρχουσα συσχέτιση μεταξύ *X* και *Y* θα πρέπει να είναι σε θέση να παραγάγει μια μικρότερη μεταβλητότητα σφάλματος –και, αντιστρόφως, πιο ακριβείς προβλέψεις του *Y*– από μια αρχική απόπειρα που βασίζεται μόνο στην επαναλαμβανόμενη πρόβλεψη του $\overline{Y}$.

Σφάλματα πρόβλεψης

Το πλαίσιο Α του Σχήματος 7.5 μας παρέχει τα σφάλματα πρόβλεψης για τους πέντε φίλους όταν ο μέσος γι' αυτούς, το $\overline{Y}$, 12 (η μέση γραμμή), χρησιμοποιείται *πάντα* για την πρόβλεψη καθενός από τα πέντε αποτελέσματα *Y*. Το πλαίσιο Β δείχνει τα αντίστοιχα σφάλματα πρόβλεψης για τους πέντε φίλους όταν μια σειρά διαφορετικών τιμών *Y'* που παίρνουμε από την εξίσωση ελαχίστων τετραγώνων (η γραμμή ελαχίστων τετραγώνων) χρησιμοποιείται για την πρόβλεψη καθενός από τα πέντε αποτελέσματα *Y*. Για παράδειγμα, το πλαίσιο Α του Σχήματος 7.5 δείχνει το σφάλμα για τον Τζον όταν ο μέσος για τους πέντε φίλους, $\overline{Y}$, που είναι 12, χρησιμοποιείται για την πρόβλεψη του δικού του αποτελέσματος *Y*, που είναι 6. Το σφάλμα του –6 για τον Τζον, το οποίο αποτυπώνεται ως μια διακεκομμένη κάθετη γραμμή ($Y - \overline{Y} = 6 - 12 = -6$), δείχνει ότι το $\overline{Y}$ υπερεκτιμά το αποτέλεσμα *Y* του Τζον κατά 6 κάρτες. Το πλαίσιο Β δείχνει ένα μικρότερο σφάλμα –1,20 για τον Τζον όταν η τιμή 7,20 για το *Y'* χρησιμοποιείται για την πρόβλεψη του ίδιου αποτελέσματος *Y* (6). Η τιμή 7,20 για το *Y'* προκύπτει από την εξίσωση ελαχίστων τετραγώνων,



ΣΧΗΜΑ 7.5

Σφάλματα πρόβλεψης για πέντε φίλους.

$$\begin{aligned} Y' &= 0{,}80(X) + 6{,}40 \\ &= 0{,}80(1) + 6{,}40 \\ &= 7{,}20 \end{aligned}$$

όπου ο αριθμός των καρτών που έστειλε ο Τζον, 1, έχει αντικαταστήσει το X.

Θετικά και αρνητικά σφάλματα δείχνουν ότι τα αποτελέσματα Y βρίσκονται πάνω ή κάτω από τα αντίστοιχα προβλεπόμενα αποτελέσματα. Συνολικά, όπως αναμενόταν, τα σφάλματα είναι μικρότερα όταν μπορούν να χρησιμοποιηθούν προσαρμοσμένες προβλέψεις του Y' από την εξίσωση ελαχίστων τετραγώνων (επειδή τα αποτελέσματα X είναι γνωστά) παρά όταν μπορεί να χρησιμοποιηθεί μόνο η επαναλαμβανόμενη πρόβλεψη του $\overline{Y}$ (επειδή τα αποτελέσματα X παραλείπονται.) Όπως συμβαίνει με τα περισσότερα φαινόμενα της στατιστικής, υπάρχουν εξαιρέσεις: Το σφάλμα πρόβλεψης για την Ντόρις είναι ελαφρώς μεγαλύτερο όταν χρησιμοποιείται η εξίσωση ελαχίστων τετραγώνων.

Μεταβλητότητα σφάλματος (Άθροισμα τετραγώνων)

Για πιο σωστή αποτίμηση της ακρίβειας των δύο αποπειρών πρόβλεψης που κάνουμε, χρειαζόμαστε κάποιο μέτρο για τα συνολικά σφάλματα που παράγει κάθε μας απόπειρα. Μάλλον δεν θα εκπλαγείτε εάν μάθετε ότι το άθροισμα τετραγώνων είναι κατάλληλο για να παίξει αυτόν τον ρόλο. Το άθροισμα τετραγώνων οποιουδήποτε συνόλου αποκλίσεων, οι οποίες αναφέρονται τώρα ως *σφάλματα*, μπορεί να υπολογιστεί αν πρώτα τετραγωνίσουμε κάθε σφάλμα (για να αφαιρεθούν τα αρνητικά πρόσημα) και έπειτα αθροίσουμε όλα τα τετραγωνικά σφάλματα.

Η μεταβλητότητα σφάλματος για την επαναληπτική πρόβλεψη του μέσου μπορεί να οριστεί ως SS_y, επειδή κάθε αποτέλεσμα Y εκφράζεται ως τετραγωνική απόκλιση από το $\overline{Y}$ και έπειτα αθροίζεται, δηλαδή

$$SS_y = \sum(Y - \overline{Y})^2$$

Χρησιμοποιώντας τα σφάλματα για τους πέντε φίλους που βλέπουμε στο πλαίσιο Α του Σχήματος 7.5, παίρνουμε

$$SS_y = [(-6)^2 + (-2)^2 + 0^2 + 6^2 + 2^2] = 80$$

Η μεταβλητότητα σφάλματος για τις προσαρμοσμένες προβλέψεις από την εξίσωση ελαχίστων τετραγώνων μπορεί να οριστεί ως $SS_{y|x}$, επειδή κάθε αποτέλεσμα Y εκφράζεται ως τετραγωνική απόκλιση από το αντίστοιχό του Y' και έπειτα αθροίζεται, δηλαδή

$$SS_{y|x} = \sum(Y - Y')^2$$

Χρησιμοποιώντας τα σφάλματα από τους πέντε φίλους που βλέπουμε στο πλαίσιο Β του Σχήματος 7.5, παίρνουμε

$$SS_{y|x} = [(-1,2)^2 + (-0,4)^2 + 0^2 + (4,4)^2 + (-2,8)^2] = 28,8$$

Αναλογία προβλεπόμενης μεταβλητότητας

Αν το σκεφτείτε καλά, το SS_y μετρά τη *συνολική* μεταβλητότητα αποτελεσμάτων Y που υπάρχει μόνο αφότου γίνουν αρχικές προβλέψεις που βασίζονται στο $\overline{Y}$ (επειδή τα αποτελέσματα X παραλείπονται), ενώ το $SS_{y|x}$ μετρά την *υπολειμματική* μεταβλητότητα αποτελεσμάτων της Y που παραμένει αφότου γίνουν προσαρμοσμένες προβλέψεις ελαχίστων τετραγώνων (επειδή χρησιμοποιούνται αποτελέσματα X). Η μεταβλητότητα σφάλματος 28,8 για τις προβλέψεις ελαχίστων τετραγώνων είναι πολύ μικρότερη από τη μεταβλητότητα σφάλματος 80 για την επαναλαμβανόμενη πρόβλεψη του $\overline{Y}$, επιβεβαιώνοντας τη μεγαλύτερη ακρίβεια των προβλέψεων ελαχίστων τετραγώνων που βλέπετε στο Σχήμα 7.5.

Για να πάρετε ένα μέτρο SS του πραγματικού *κέρδους στην ακρίβεια* που οφείλεται στις προβλέψεις ελαχίστων τετραγώνων, αφαιρείτε την υπολειμματική μεταβλητότητα από τη συνολική μεταβλητότητα, δηλαδή αφαιρείτε το $SS_{y|x}$ από το SS_y, και παίρνετε

$$SS_y - SS_{y|x} = 80 - 28,8 = 51,2$$

Για να εκφράσετε αυτήν τη διαφορά, 51,2, ως κέρδος στην ακρίβεια *σε σχέση* με την αρχική μεταβλητότητα σφάλματος για την επαναλαμβανόμενη πρόβλεψη του $\overline{Y}$, διαιρείτε την πιο πάνω διαφορά διά SS_y, δηλαδή

$$\frac{SS_y - SS_{y|x}}{SS_y} = \frac{80 - 28,8}{80} = \frac{51,2}{80} = 0,64$$

Αυτό το αποτέλεσμα, 0,64 ή 64%, αναπαριστά το κέρδος αναλογίας ή ποσοστού στην ακρίβεια της πρόβλεψης όταν η επαναλαμβανόμενη πρόβλεψη του $\overline{Y}$ αντικαθίσταται από μια σειρά προσαρμοσμένων προβλέψεων Y' που βασίζονται στην εξίσωση ελαχίστων τετραγώνων. Με άλλα λόγια, το 0,64 ή 64% αναπαριστά την αναλογία ή το ποσοστό της συνολικής μεταβλητότητας του SS_y που προβλέπεται από τη σχέση του με τη μεταβλητή X.

Προς ευαρέσκεια των στατιστικολόγων, όταν υπολογίσουμε το τετράγωνο, η τιμή του συντελεστή συσχέτισης ισούται μ' αυτήν την αναλογία προβλεπόμενης μεταβλητότητας. Έχουμε πει ότι παίρνουμε r 0,80 από τη συσχέτιση μεταξύ των καρτών που εστάλησαν και των καρτών που ελήφθησαν από τους πέντε φίλους και μπορούμε να επαληθεύσουμε ότι $r^2 = (0,80)(0,80) = 0,64$, το οποίο βέβαια είναι επίσης η αναλογία προβλεπόμενης μεταβλητότητας. Δεδομένης αυτής της προοπτικής,

το τετράγωνο του συντελεστή συσχέτισης, r^2, δείχνει *πάντα* την αναλογία της συνολικής μεταβλητότητας σε μία μεταβλητή που μπορεί να προβλεφθεί μέσω της σχέσης με μια άλλη μεταβλητή.

Συντελεστής τετραγωνικής συσχέτισης (r^2)
Η αναλογία της συνολικής μεταβλητότητας μίας μεταβλητής που προβλέπεται μέσω της σχέσης με μια άλλη μεταβλητή.

Εκφράζοντας την εξίσωση για το r^2 με σύμβολα, έχουμε:

Ερμηνεία του r^2

$$r^2 = \frac{SS_{Y'}}{SS_Y} = \frac{SS_Y - SS_{Y|X}}{SS_Y} \qquad (7.6)$$

όπου ο ένας νέος όρος του αθροίσματος τετραγώνων, $SS_{y'}$, είναι απλώς η μεταβλητότητα που εξηγείται ή προβλέπεται από την εξίσωση παλινδρόμησης, δηλαδή

$$SS_{y'} = \sum (Y' - \overline{Y})^2$$

Αναλόγως, ο συντελεστής r^2 μας παρέχει ένα άμεσο μέτρο υπολογισμού της αξίας της προσπάθειας που κάνουμε για να προβλέψουμε τα ελάχιστα τετράγωνα.[20]

Το r^2 δεν ισχύει για μεμονωμένα αποτελέσματα

Μην επιχειρείτε να εφαρμόσετε την ερμηνεία μεταβλητότητας του r^2 σε μεμονωμένα αποτελέσματα. Για παράδειγμα, το γεγονός ότι το 64% της μεταβλητότητας στις κάρτες που ελήφθησαν από τους πέντε φίλους (*Y*) προβλέπεται από τις κάρτες που εστάλησαν (*X*) δεν σηματοδοτεί, επομένως, ότι μπορεί να γίνει τέλεια πρόβλεψη για το 64% των αποτελεσμάτων *Y'* των πέντε φίλων. Όπως μπορείτε να δείτε στο πλαίσιο Β του Σχήματος 7.5, μόνο ένα από τα αποτελέσματα *Y* για τους πέντε φίλους, οι 12 κάρτες που ελήφθησαν από τον Μάικ, προβλέφθηκε τέλεια (επειδή συμπίπτει με τη γραμμή παλινδρόμησης για την εξίσωση ελαχίστων τετραγώνων), και ακόμα κι αυτή η τέλεια πρόβλεψη δεν είναι εγγυημένη απλώς επειδή το r^2 ισούται με 0,64. Αντιθέτως, το 64% πρέπει να ερμηνευτεί όπως ισχύει για τη μεταβλητότητα για *ολόκληρο* το σύνολο των αποτελεσμάτων *Y*. Η συνολική μεταβλητότητα *όλων* των αποτελεσμάτων *Y* –όπως υπολογίζεται από το SS_Y– μπορεί να μειωθεί κατά 64% όταν κάθε αποτέλεσμα *Y* αντικαθίσταται από το αντίστοιχο προβλεπόμενο αποτέλεσμα *Y'* και έπειτα να εκφραστεί ως τετραγωνική απόκλιση από τον μέσο όλων των παρατηρούμενων αποτελεσμάτων. Ως εκ τούτου, το 64% αναπαριστά μια πρόβλεψη στη συνολική μεταβλητότητα για τα πέντε αποτελέσματα *Y* όταν αντικαθίσταται από μια σειρά προβλεπόμενων αποτελεσμάτων, δεδομένης της εξίσωσης ελαχίστων τετραγώνων και διάφορων τιμών του *X*.

Μικρές τιμές του r^2

Όταν γίνεται μεταφορά από το *r* στο r^2, οι οδηγίες του Cohen, τις οποίες αναφέρουμε στη σελίδα 165, δηλώνουν ότι *μια τιμή του συντελεστή r^2 κοντά στο 0,01, στο 0,09 ή στο 0,25 εκφράζει μια αδύναμη, μέτρια ή ισχυρή σχέση αντίστοιχα.* Μην περιμένετε να συναντάτε συχνά μεγάλες τιμές του r^2 σε έρευνες συμπεριφοράς και εκπαίδευσης. Σ' αυτούς τους κλάδους, όπου μέτρα σύνθετων φαινομένων, όπως το νοητικό χάρισμα, η ψυχοπαθητική τάση ή η αυτοεκτίμηση, δεν καταφέρνουν να αποκτήσουν ιδιαίτερη σχέση με οποιαδήποτε μεμονωμένη μεταβλητή, οι τιμές του r^2 που είναι μεγαλύτερες από περίπου 0,25 είναι ιδιαίτερα απίθανες. Ωστόσο, ακόμα και τιμές του r^2 που τείνουν στο μηδέν αξίζουν την προσοχή μας. Για παράδειγμα, αν μόνο το 0,04 (ή το 4%) της μεταβλητότητας αποτελεσμάτων πνευματικής υγείας μαθητών της έκτης δημοτικού θα μπορούσε να προβλεφθεί από μία μόνο μεταβλητή, όπως είναι οι διαφορές τους στην ηλικία απογαλακτισμού, πολλοί ερευνητές πιθανώς θα το θεωρούσαν σημαντικό εύρημα που θα άξιζε περαιτέρω διερεύνηση.

20. Για να υπολογίσετε πραγματικά την τιμή του *r*, μη χρησιμοποιείτε ποτέ τον Τύπο 7.6, ο οποίος έχει σχεδιαστεί αποκλειστικά ως βοήθημα για την κατανόηση της ερμηνείας του r^2. Αντίθετα, χρησιμοποιείτε πάντα τους πολύ πιο αποδοτικούς τύπους των σελίδων 167-168.

Το r^2 δεν εξασφαλίζει σχέση αιτίου-αιτιατού

Το ερώτημα της ύπαρξης της σχέσης αιτίου-αιτιατού, για την οποία μιλήσαμε στην Ενότητα 6.3, δεν μπορεί να λυθεί απλώς λαμβάνοντας το τετράγωνο του συντελεστή συσχέτισης για τη λήψη μιας τιμής για το r^2. Αν η συσχέτιση μεταξύ των αποτελεσμάτων πνευματικής υγείας μαθητών της έκτης δημοτικού και της ηλικίας απογαλακτισμού τους ως βρεφών ισούται με 0,20, δεν μπορούμε να ισχυριστούμε ότι το (0,20)(20) = 0,04 ή το 4% της συνολικής μεταβλητότητας στα αποτελέσματα πνευματικής υγείας *προκαλείται* από τις διαφορές στις ηλικίες απογαλακτισμού. Αντίθετα, μπορούμε να ισχυριστούμε ότι αυτή η συσχέτιση πιθανώς αντανακλά κάποιον πιο βασικό παράγοντα ή παράγοντες, όπως, για παράδειγμα, μια τάση που έχουν οι πιο ασφαλείς οικονομικά και λιγότερο αγχώδεις μητέρες να δημιουργούν ένα οικογενειακό περιβάλλον που διαιωνίζει την καλή πνευματική υγεία και, συμπτωματικά, να θηλάζουν τα βρέφη τους περισσότερο χρόνο. Είναι βέβαιο ότι, όταν δεν υπάρχουν πρόσθετες αποδείξεις, θα ήταν απερίσκεπτο να ενθαρρύνουμε τις μητέρες, ανεξάρτητα από την κατάστασή τους, να αναβάλουν τον απογαλακτισμό εξαιτίας της προβλεπόμενης επίδρασής του σε αποτελέσματα πνευματικής υγείας.

Παρόλο που αναφερόμαστε πολλάκις στον συντελεστή r^2 ως μια ένδειξη της αναλογίας ή του ποσοστού της *προβλέψιμης* μεταβλητότητας, θα συναντήσετε επίσης αναφορές στο r^2 ως ένδειξη της αναλογίας ή του ποσοστού της *εξηγήσιμης* μεταβλητότητας. Σ' αυτό το πλαίσιο, ο όρος «εξηγήσιμη» σημαίνει μόνο προβλεψιμότητα και *όχι* αιτιότητα. Θα μπορούσατε, επομένως, να ισχυριστείτε ότι το 0,04 ή το 4% της μεταβλητότητας σε αποτελέσματα πνευματικής υγείας «εξηγείται» από τις διαφορές στην ηλικία απογαλακτισμού, στον βαθμό που το 0,04 ή το 4% μπορεί να προβλεφθεί από –ή να αποδοθεί στατιστικά σε– διαφορές στην ηλικία απογαλακτισμού.

Έλεγχος προόδου *7.4 Έστω ότι ο συντελεστής r 0,30 περιγράφει τη σχέση μεταξύ μορφωτικού επιπέδου και εκτιμώμενων ωρών μελέτης εβδομαδιαίως.

(α) Σύμφωνα με το r^2, ποιο ποσοστό της μεταβλητότητας στον εβδομαδιαίο χρόνο μελέτης μπορεί να προβλεφθεί από τη σχέση του με το μορφωτικό επίπεδο;

(β) Τι ποσοστό μεταβλητότητας επί του εβδομαδιαίου χρόνου μελέτης δεν μπορεί να προβλεφθεί απ' αυτήν τη σχέση;

(γ) Κάποιος ισχυρίζεται ότι το 9% του εκτιμώμενου χρόνου μελέτης *κάθε* ατόμου μπορεί να προβλεφθεί από τη σχέση. Τι λάθος έχει αυτός ο ισχυρισμός;

Έλεγχος προόδου *7.5 Όπως δείχνει το Σχήμα 6.3 στη σελίδα 162, η συσχέτιση μεταξύ των δεικτών IQ γονέων και παιδιών είναι 0,50 και μεταξύ των δεικτών IQ αναδόχων γονέων και θετών παιδιών είναι 0,27.

(α) Καταλαβαίνουμε από τα παραπάνω ότι η σχέση μεταξύ αναδόχων γονέων και θετών παιδιών έχει περίπου μισή ισχύ σε σύγκριση με τη σχέση μεταξύ γονέων και παιδιών;

(β) Χρησιμοποιήστε το r^2 για να συγκρίνετε την ισχύ αυτών των δύο συσχετίσεων.

Απαντήσεις στη σελίδα 538.

7.7 Εξίσωση πολλαπλής παλινδρόμησης

Οποιαδήποτε σοβαρή προσπάθεια πρόβλεψης συνήθως καταλήγει σε μια πιο περίπλοκη εξίσωση η οποία δεν περιέχει μόνο ένα αλλά πολλά *X* ή *μεταβλητές συνάρτησης πρόβλεψης*. Για παράδειγμα, μια σοβαρή προσπάθεια να προβλέψουμε τον βαθμό πτυχίου μπορεί να καταλήξει σ' αυτήν την εξίσωση:

$$Y' = 0{,}410(X_1) + 0{,}005(X_2) + 0{,}001(X_3) + 1{,}03$$

Εξίσωση πολλαπλής παλινδρόμησης
Μια εξίσωση ελαχίστων τετραγώνων που περιέχει περισσότερες από μία μεταβλητές συνάρτησης πρόβλεψης ή X.

όπου το Y' αναπαριστά τον προβλεπόμενο βαθμό πτυχίου και τα X_1, X_2 και X_3 αναφέρονται στον απολυτήριο βαθμό λυκείου, στον δείκτη IQ και στη βαθμολογία στις εξετάσεις SAT αντίστοιχα. Αξιοποιώντας τη συνδυαστική ισχύ πρόβλεψης πολλών μεταβλητών συνάρτησης πρόβλεψης, αυτή η **εξίσωση πολλαπλής παλινδρόμησης** παρέχει πιο ακριβείς προβλέψεις για το Y' (αναφέρεται συχνά ως η *μεταβλητή κριτηρίου*) απ' όσες θα μπορούσαν να ληφθούν από μια απλή εξίσωση παλινδρόμησης.

Κοινά χαρακτηριστικά

Αν και δεν μπορούμε να το φανταστούμε εύκολα, η εξίσωση πολλαπλής παλινδρόμησης έχει πολλά κοινά χαρακτηριστικά με τις αντίστοιχες απλές. Για παράδειγμα, εξακολουθεί να θεωρείται εξίσωση ελαχίστων τετραγώνων, επειδή ελαχιστοποιεί το άθροισμα των τετραγωνικών σφαλμάτων πρόβλεψης. Κατά τον ίδιο τρόπο, συνοδεύεται από τυπικά σφάλματα εκτίμησης που μετρούν γενικά τις μέσες ποσότητες σφαλμάτων πρόβλεψης. Αυτό το κεφάλαιο μπορεί να αποτελέσει μια καλή αφετηρία αν κάποια στιγμή στο μέλλον χρειαστεί να ασχοληθείτε με εξισώσεις πολλαπλής παλινδρόμησης.

7.8 Παλινδρόμηση προς τον μέσο

Παλινδρόμηση προς τον μέσο *Μια τάση των αποτελεσμάτων (παρατηρήσεων), ιδιαίτερα των ακραίων, να συρρικνώνονται προς τον μέσο.*

Η **παλινδρόμηση προς τον μέσο** *αναφέρεται σε μια τάση που έχουν οι παρατηρήσεις, ιδιαίτερα οι ακραίες, να συρρικνώνονται προς τον μέσο.* Αυτή η τάση εμφανίζεται συχνά μεταξύ υποσυνόλων παρατηρήσεων των οποίων οι τιμές είναι ακραίες και οφείλονται τουλάχιστον εν μέρει στην τύχη. Για παράδειγμα, εξαιτίας της παλινδρόμησης προς τον μέσο, θα περιμέναμε ότι οι φοιτητές που έχουν πετύχει τους πέντε καλύτερους βαθμούς στο πρώτο τεστ στατιστικής δεν θα καταφέρουν το ίδιο στο δεύτερο τεστ στατιστικής. Αν και οι πέντε φοιτητές μπορεί να έχουν βαθμούς πάνω από τον μέσο στο δεύτερο τεστ, κάποιοι από τους βαθμούς τους θα εμφανίσουν κάποια παλινδρόμηση προς τον μέσο. Το πιο πιθανό είναι ότι οι πέντε πρώτοι βαθμοί στο πρώτο τεστ δείχνουν δύο στοιχεία: Το ένα σχετικά μόνιμο στοιχείο είναι ένδειξη του γεγονότος ότι αυτοί οι φοιτητές είναι καλύτεροι χάρη στον σωστό τρόπο μελέτης που εφαρμόζουν, μια ισχυρή ικανότητα για ποσοτική αιτιολόγηση και σε άλλα στοιχεία. Το άλλο σχετικά παροδικό στοιχείο αντανακλά το γεγονός ότι την ημέρα του τεστ κάποιοι τουλάχιστον απ' αυτούς τους φοιτητές ήταν πολύ τυχεροί επειδή διάφοροι μικροί τυχαίοι παράγοντες, όπως η ξεκούραση την προηγούμενη νύχτα ή η ωραία διαδρομή για τη σχολή, λειτούργησαν υπέρ τους. Στο δεύτερο τεστ, ακόμα κι αν οι βαθμοί αυτών των πέντε φοιτητών εξακολουθούν να αναδεικνύουν ότι μόνιμα βρίσκονται πάνω από τον μέσο όρο, κάποιοι από τους βαθμούς δεν θα είναι το ίδιο καλοί εξαιτίας λιγότερο καλής ή ακόμα και κακής τύχης. Η καθαρή επίδραση είναι ότι οι βαθμοί τουλάχιστον μερικών από τους πέντε φοιτητές θα υπολείπονται των πέντε πρώτων βαθμολογιών –δηλαδή θα παρουσιάσουν μια παλινδρόμηση μείωσης προς τον μέσο– στο δεύτερο τεστ. (Όταν παρατηρείται σημαντική παλινδρόμηση προς τον μέσο μετά από μια θεαματική απόδοση ενός π.χ. νέου αθλητή ή πρωτοεμφανιζόμενου συγγραφέα, έχουμε ουσιαστικά το αντίθετο της *τύχης του πρωτάρη*.)

Τα νέα ωστόσο είναι καλά για τους φοιτητές που είχαν τους πέντε χειρότερους βαθμούς στο πρώτο τεστ. Αν και αυτοί οι πέντε φοιτητές μπορεί να έχουν βαθμούς κάτω από τον μέσο στο δεύτερο τεστ, μερικοί από τους βαθμούς τους πιθανώς θα σημειώσουν παλινδρόμηση αύξησης προς τον μέσο. Στο δεύτερο τεστ, μερικοί δεν θα είναι το ίδιο άτυχοι. Η καθαρή επίδραση είναι ότι οι βαθμοί τουλάχιστον μερικών από τους μαθητές που στο πρώτο τεστ ήταν στους τελευταίους πέντε θα μετακινηθούν προς τα πάνω στην κατάταξη στο δεύτερο τεστ.

Εμφανίζεται σε πολλές κατανομές

Η παλινδρόμηση προς τον μέσο εμφανίζεται σε υποσύνολα ακραίων παρατηρήσεων για διάφορες κατανομές. Για παράδειγμα, εμφανίζεται για το υποσύνολο των μετοχών με την καλύτερη (ή τη χειρότερη) απόδοση στο Χρηματιστήριο της Νέας Υόρκης για οποιαδήποτε χρονική περίοδο, όπως μία εβδομάδα, έναν μήνα ή ένα έτος. Εμφανίζεται επίσης στους καλύτερους (ή στους χειρότερους) ρίπτες στο πρωτάθλημα μπέιζμπολ κατά τη διάρκεια διαδοχικών σεζόν. Ο Πίνακας 7.4 παραθέτει τους κορυφαίους 10 ρίπτες των μεγάλων κατηγοριών μπέιζμπολ για το 2014 και δείχνει επίσης πώς συνέχισαν το 2015. Παρατηρήστε ότι οι 7 από τους 10 καλύτερους μέσους όρους ρίψεων παρουσιάζουν καθοδική παλινδρόμηση, προς το 260, τον κατά προσέγγιση μέσο για όλους τους ρίπτες για το 2015. Παρεμπιπτόντως, δεν είναι αλήθεια ότι, αν τους δούμε ως μια ομάδα, όλοι οι ρίπτες των μεγάλων κατηγοριών τείνουν προς τη μετριότητα. Οι 10 καλύτεροι ρίπτες του 2014, οι οποίοι δεν βρίσκονται στους 10 καλύτερους του 2015, αντικαταστάθηκαν από άλλους, κυρίως πάνω από τον μέσο όρο ρίπτες, οι οποίοι ήταν επίσης πολύ τυχεροί το 2015. Η παρατηρούμενη παλινδρόμηση προς τον μέσο συμβαίνει σε άτομα ή υποσύνολα ατόμων αλλά όχι σε ολόκληρες ομάδες.

Πλάνη της παλινδρόμησης
Συμβαίνει όταν η παλινδρόμηση προς τον μέσο ερμηνεύεται ως πραγματική και όχι τυχαία επίδραση.

Η πλάνη της παλινδρόμησης

Η **πλάνη της παλινδρόμησης** *συντελείται κάθε φορά που η παλινδρόμηση προς τον μέσο ερμηνεύεται ως πραγματική και όχι τυχαία επίδραση.* Ένα κλασικό παράδειγμα πλάνης της παλινδρόμησης αναδείχθηκε σε μια μελέτη της εκπαίδευσης πιλότων της Αεροπορίας του Ισραήλ [όπως περιγράφεται στο σύγγραμμα των Tversky, A., & Kahnemann, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131.] Δόθηκαν έπαινοι σε ορισμένους εκπαιδευόμενους πιλότους μετά από μερικές πολύ καλές προσγειώσεις, ενώ υπήρξαν αυστηρές συστάσεις σε άλλους με πολύ κακές προσγειώσεις. Στις επόμενες ασκήσεις, οι πρώτοι είχαν κακά αποτελέσματα και οι δεύτεροι τα πήγαν καλύτερα. Επομένως, μπορούμε να συμπεράνουμε ότι ο έπαινος εμποδίζει, αλλά η επίπληξη συμβάλλει στην καλή απόδοση!

Ένα έγκυρο συμπέρασμα θεωρεί ότι υπάρχει παλινδρόμηση προς τον μέσο. Είναι εύλογο να υποθέσουμε ότι στις προσγειώσεις, εκτός από την ικανότητα, παίζει ρόλο και η τύχη. Μερικοί εκπαιδευόμενοι που έκαναν πολύ καλές προσγειώσεις ήταν τυχεροί, ενώ άλλοι που έκαναν κακές προσγειώσεις ήταν άτυχοι. Επομένως, θα υπήρχε μια τάση, η οποία θα μπορούσε να αποδοθεί στην τύχη, ότι οι καλές προσγειώσεις ακολουθούνται από λιγότερο καλές και οι κακές προσγειώσεις ακολουθούνται από λιγότερο κακές – ακόμα κι αν οι εκπαιδευόμενοι δεν είχαν επαινεθεί μετά από πολύ καλές προσγειώσεις ή επιπληχθεί μετά από πολύ κακές προσγειώσεις.

Αποφυγή της πλάνης της παλινδρόμησης

Η πλάνη της παλινδρόμησης μπορεί να αποφευχθεί αν χωρίσουμε το υποσύνολο των ακραίων παρατηρήσεων σε δύο ομάδες. Στο προηγούμενο παράδειγμα, μία ομάδα εκπαιδευόμενων θα συνέχιζε να δέχεται επαίνους μετά από πολύ καλές προσγειώσεις και επιπλήξεις μετά από πολύ κακές προσγειώσεις. Μια δεύτερη ομάδα εκπαιδευόμενων δεν θα λάμβανε κανένα σχόλιο, ανεξαρτήτως της ποιότητας των προσγειώσεων. Κατά συνέπεια, η δεύτερη ομάδα θα αποτελούσε μέσο ελέγχου για την παλινδρόμηση προς τον μέσο, επειδή οποιαδήποτε μετατόπιση προς τον μέσο στις δεύτερες προσγειώσεις θα οφειλόταν στην τύχη. Πιο σημαντικό είναι ότι οποιαδήποτε παρατηρούμενη διαφορά μεταξύ των δύο ομάδων (που επιβιώνει μιας στατιστικής ανάλυσης όπως περιγράφεται στο Μέρος 2) θα θεωρούνταν πραγματική διαφορά και δεν θα αποδιδόταν στην επίδραση της παλινδρόμησης.

Θα πρέπει να είστε προσεκτικοί στην πλάνη της παλινδρόμησης σε έρευνες για την εκπαίδευση όπου συμμετέχουν ομάδες μαθητών και φοιτητών χωρίς καλούς βαθμούς. Για παράδειγμα, μια ομάδα μαθητών τετάρτης δη-

Πίνακας 7.4
ΠΑΛΙΝΔΡΟΜΗΣΗ ΠΡΟΣ ΤΟΝ ΜΕΣΟ: ΜΕΣΟΙ ΟΡΟΙ ΡΙΨΕΩΝ ΤΩΝ 10 ΚΑΛΥΤΕΡΩΝ ΡΙΠΤΩΝ ΣΤΗΝ ΠΡΩΤΗ ΚΑΤΗΓΟΡΙΑ ΜΠΕΪΖΜΠΟΛ ΓΙΑ ΤΟ 2014 ΚΑΙ Η ΣΥΝΕΧΕΙΑ ΤΟΥΣ ΤΟ 2015

	Μέσοι όροι ρίψεων*		
10 καλύτεροι ρίπτες (2014)	**2014**	**2015**	**Παλινδρόμηση προς τον μέσο;**
1. J. Alture	0,341	0,313	Ναι
2. V. Martinez	0,335	0,282	Ναι
3. M. Brantley	0,327	0,310	Ναι
4. A. Beltre	0,324	0,287	Ναι
5. J. Abreu	0,317	0,290	Ναι
6. R. Cano	0,314	0,287	Ναι
7. A. McCutchen	0,314	0,292	Ναι
8. M. Cabrera	0,313	0,338	Όχι
9. B. Posey	0,311	0,318	Όχι
10. B. Revere	0,306	0,306	Όχι

** Αναλογία επιτυχιών ανά επίσημο αριθμό ρίψεων.*

Πηγή: http://sports.espn.go.com/mlb/stats/batting.

μοτικού, οι οποίοι επιλέγονται για να παρακολουθήσουν ένα ειδικό πρόγραμμα μαθητών με κακές επιδόσεις στην κατανόηση κειμένου, ενδεχομένως να δείξει κάποια βελτίωση. Για να διαπιστωθεί εάν αυτή η βελτίωση θα μπορούσε να αποδοθεί στο ειδικό πρόγραμμα ή σε μια επίδραση της παλινδρόμησης απαιτούνται πληροφορίες από μια ομάδα ελέγχου μαθητών τετάρτης δημοτικού με παρόμοιες κακές επιδόσεις οι οποίοι δεν παρακολουθούν το ειδικό πρόγραμμα. Είναι, επομένως, σημαντικό για την έρευνα να περιλαμβάνει πάντοτε μια ομάδα ελέγχου για παλινδρόμηση προς τον μέσο.

Έλεγχος προόδου *7.6 Αφού μια ομάδα φοιτητών συμμετείχε σε ένα πρόγραμμα μείωσης του άγχους, παρατηρήθηκαν μειώσεις στους δείκτες άγχους εκείνων που πριν από το πρόγραμμα είχαν πολύ υψηλές βαθμολογίες σε μια εξέταση άγχους.

(α) Μπορεί αυτή η μείωση να αποδοθεί στο πρόγραμμα μείωσης άγχους; Εξηγήστε την απάντησή σας.

(β) Τι είδους μελέτη, αν υπάρχει, θα επέτρεπε έγκυρα συμπεράσματα σχετικά με την επίδραση του προγράμματος μείωσης άγχους;

Απαντήσεις στη σελίδα 538.

Περίληψη

Αν υπάρχει μια γραμμική σχέση μεταξύ δύο μεταβλητών, τότε η μία μεταβλητή μπορεί να προβλεφθεί από την άλλη μέσω της εξίσωσης παλινδρόμησης ελαχίστων τετραγώνων, όπως περιγράφεται στους Τύπους 7.1, 7.2 και 7.3.

Η εξίσωση ελαχίστων τετραγώνων ελαχιστοποιεί το άθροισμα όλων των τετραγωνικών σφαλμάτων πρόβλεψης που θα υπήρχαν αν η εξίσωση είχε χρησιμοποιηθεί για την πρόβλεψη γνωστών αποτελεσμάτων *Y* από την αρχική ανάλυση συσχέτισης.

Μια εκτίμηση σφάλματος πρόβλεψης προκύπτει από τον Τύπο 7.5. Γνωστή ως *τυπικό σφάλμα εκτίμησης*, αυτή η εκτίμηση είναι ένα είδος τυπικής απόκλισης που αναπαριστά γενικά τη μέση ποσότητα του σφάλματος πρόβλεψης. Η τιμή του τυπικού σφάλματος της εκτίμησης εξαρτάται κυρίως από το μέγεθος του συντελεστή συσχέτισης. Όσο μεγαλύτερος είναι ο συντελεστής συσχέτισης, προς τη θετική ή προς την αρνητική κατεύθυνση, τόσο μικρότερο είναι το τυπικό σφάλμα της εκτίμησης και τόσο πιο ευνοϊκή είναι η πρόγνωση για προβλέψεις.

Η εξίσωση παλινδρόμησης θεωρεί ότι υπάρχει μια γραμμική σχέση μεταξύ μεταβλητών και το τυπικό σφάλμα εκτίμησης θεωρεί ότι υπάρχει ομοσκεδαστικότητα – περίπου ίσα σημεία διασκόρπισης δεδομένων γύρω από όλα τα τμήματα της γραμμής παλινδρόμησης.

Το τετράγωνο του συντελεστή συσχέτισης, r^2, μας δείχνει την αναλογία της συνολικής μεταβλητότητας σε μία μεταβλητή που μπορεί να προβλεφθεί από τη σχέση της με την άλλη μεταβλητή.

Οι σοβαρές προσπάθειες πρόβλεψης συνήθως περιλαμβάνουν εξισώσεις πολλαπλής παλινδρόμησης που αποτελούνται από περισσότερες από μία μεταβλητές της συνάρτησης πρόβλεψης ή *X*. Η εξίσωση πολλαπλής παλινδρόμησης μοιράζεται πολλά κοινά χαρακτηριστικά με την απλή εξίσωση παλινδρόμησης που περιγράψαμε σ' αυτό το κεφάλαιο.

Η *παλινδρόμηση προς τον μέσο* είναι μια τάση που έχουν οι παρατηρήσεις, ιδιαίτερα οι ακραίες, να συρρικνώνονται προς τον μέσο. Παρατηρείται το φαινόμενο της πλάνης της παλινδρόμησης όταν η παλινδρόμηση προς τον μέσο ερμηνεύεται ως πραγματική και όχι ως τυχαία επίδραση. Για να προφυλασσόμαστε από την πλάνη της παλινδρόμησης, χρησιμοποιούμε ομάδες ελέγχου για την εκτίμηση της επίδρασης της παλινδρόμησης.

Σημαντικοί όροι

Εξίσωση παλινδρόμησης ελαχίστων τετραγώνων
Τυπικό σφάλμα εκτίμησης ($s_{y|x}$)
Τετραγωνική συσχέτιση (r^2)
Εξίσωση πολλαπλής παλινδρόμησης
Παλινδρόμηση προς τον μέσο
Πλάνη της παλινδρόμησης

Κύριες εξισώσεις

ΕΞΙΣΩΣΗ ΠΡΟΒΛΕΨΗΣ

$$Y' = bx + a$$

$$\text{όπου } b = r\sqrt{\frac{SS_Y}{SS_X}}$$

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

Ερωτήσεις επανάληψης

7.7 Έστω ότι ο συντελεστής r –0,80 αποτυπώνει την ισχυρή αρνητική σχέση μεταξύ των ετών καπνίσματος (X) και του προσδόκιμου ζωής (Y). Θεωρούμε επιπλέον ότι οι κατανομές αυτών των δύο παραγόντων έχουν τους μέσους και τα αθροίσματα τετραγώνων που βλέπετε εδώ:

$$\overline{X} = 5 \qquad \overline{Y} = 60$$
$$SS_x = 35 \qquad SS_y = 70$$

(α) Υπολογίστε την εξίσωση παλινδρόμησης ελαχίστων τετραγώνων για την πρόβλεψη του προσδόκιμου ζωής από τα έτη καπνίσματος.
(β) Υπολογίστε το τυπικό σφάλμα εκτίμησης, $s_{y|x}$, αν υποθέσουμε ότι η συσχέτιση –0,80 βασίστηκε σε $n = 50$ ζεύγη παρατηρήσεων.
(γ) Δώστε μια γενική ερμηνεία του $s_{y|x}$.
(δ) Προβλέψτε το προσδόκιμο ζωής για τον Τζον, ο οποίος κάπνιζε για 8 έτη.
(ε) Προβλέψτε το προσδόκιμο ζωής για την Κέιτι, η οποία δεν κάπνισε ποτέ.

7.8 Τα παρακάτω ζεύγη αναπαριστούν τον αριθμό αδειούχων οδηγών (X) και τον αριθμό αυτοκινήτων (Y) για επτά οικογένειες στη γειτονιά μου:

Οδηγοί (X)	Αυτοκίνητα (Y)
5	4
5	3
2	2
2	2
3	2
1	1
2	2

(α) Κατασκευάστε ένα διάγραμμα διασποράς που θα επαληθεύει την έλλειψη έντονης καμπυλότητας.
(β) Υπολογίστε την εξίσωση ελαχίστων τετραγώνων για τα συγκεκριμένα δεδομένα. (Μην ξεχνάτε ότι θα πρέπει να υπολογίσετε πρώτα τα r, SS_y και SS_x.)
(γ) Υπολογίστε το τυπικό σφάλμα εκτίμησης, $s_{y|x}$, εφόσον $n = 7$.
(δ) Προβλέψτε τον αριθμό των αυτοκινήτων για δύο νέες οικογένειες με δύο και πέντε οδηγούς.

7.9 Σε μια μεγάλη τράπεζα, η διάρκεια της υπηρεσίας είναι η καλύτερη μεμονωμένη συνάρτηση πρόβλεψης για τους μισθούς των υπαλλήλων. Μπορούμε, επομένως, να συμπεράνουμε ότι υπάρχει μια σχέση αιτίου-αιτιατού μεταξύ της διάρκειας της υπηρεσίας και του μισθού;

7.10 Έστω ότι ο συντελεστής r^2 ισούται με 0,50 για τη σχέση μεταξύ ύψους και βάρους για ενήλικες. Διευκρινίστε αν οι παρακάτω προτάσεις είναι σωστές ή λάθος.

(α) Το 50% της μεταβλητότητας στα ύψη μπορεί να εξηγηθεί από τη μεταβλητότητα στα βάρη.
(β) Υπάρχει μια σχέση αιτίου-αιτιατού μεταξύ ύψους και βάρους.
(γ) Τα ύψη του 50% των ενηλίκων μπορούν να προβλεφθούν ακριβώς από τα βάρη τους.
(δ) Το 50% της μεταβλητότητας στα βάρη μπορεί να προβλεφθεί από τα ύψη.

***7.11** Γνωρίζουμε καλά από μελέτες που χρονολογούνται από πριν από 100 χρόνια ότι υπάρχει παλινδρόμηση προς τον μέσο μεταξύ του ύψους ενός πατέρα και του ύψους των *ενήλικων* γιων του. Διευκρινίστε αν οι παρακάτω προτάσεις είναι σωστές ή λάθος.
(α) Οι γιοι που έχουν ψηλούς πατέρες θα τείνουν να είναι κοντύτεροι από τους πατέρες τους.
(β) Οι γιοι που έχουν κοντούς πατέρες θα τείνουν να είναι ψηλότεροι από τον μέσο όλων των γιων.
(γ) Κάθε γιος ενός ψηλού πατέρα θα είναι κοντύτερος από τον πατέρα του.
(δ) Αν τους θεωρήσουμε ως μια ομάδα, οι ενήλικες γιοι είναι κοντύτεροι από τους πατέρες τους.
(ε) Οι πατέρες ψηλών γιων θα τείνουν να είναι ψηλότεροι από τους γιους τους.
(στ) Οι πατέρες κοντών γιων θα τείνουν να είναι ψηλότεροι από τους γιους τους, αλλά κοντύτεροι από τον μέσο όλων των πατέρων.
Απαντήσεις στη σελίδα 538.

7.12 Κάποιος ισχυρίζεται ότι μια καλή στρατηγική επενδύσεων θα ήταν να αγοράσει τις πέντε μετοχές με τη χειρότερη πορεία στο Χρηματιστήριο και να κεφαλαιοποιήσει την παλινδρόμηση προς τον μέσο. Τα σχόλιά σας;

7.13 Στην αρχική μελέτη της παλινδρόμησης προς τον μέσο, ο Sir Francis Galton παρατήρησε μια τάση που είχαν απόγονοι ψηλών και κοντών γονέων να μετατοπίζονται προς το μέσο ύψος των απογόνων και ανέφερε αυτήν την τάση ως «παλινδρόμηση προς τη μετριότητα». Τι λάθος έχει το συμπέρασμα ότι τελικά όλα τα ύψη θα είναι κοντά στον μέσο τους;

ΠΑΡΑΡΤΗΜΑ VII

Εκτίμηση του γραμμικού υποδείγματος με την R

Κωνσταντίνος Κουνετάς

Π.VII.1 Εισαγωγή

Στο Παράρτημα του Κεφαλαίου 7 θα παρουσιάσουμε ένα από τα πιο γνωστά εργαλεία της στατιστικής επιστήμης που αφορά το απλό γραμμικό υπόδειγμα. Η βασική ιδέα είναι το πώς θα μπορέσουμε να περιγράψουμε τη σχέση ανάμεσα σε μια εξαρτημένη μεταβλητή (y, ως συνηθίζεται) και σε μια στην παρούσα περίπτωση μεταβλητή (x) η οποία καλείται ανεξάρτητη.

Π.VII.2 Απλό γραμμικό υπόδειγμα

Π.VII.2.1 Αρχικές διαπιστώσεις

Στην ενότητα αυτή θα μιλήσουμε πιο συγκεκριμένα για κάποιες τεχνικές οι οποίες χρησιμοποιούνται στη στατιστική και στην οικονομετρία. Ξεκινάμε με τις παρακάτω εντολές με τις οποίες φορτώνουμε τα δεδομένα μας:

```
> library(readxl)
> FinalData <- read_excel("C:/Users/user/Desktop/Kritiki Book_diorthoseis_Chapter
1_9_Kounetas_K_2018_2019/Data.2017/FinalData.xlsx")
```

Μια πάρα πολύ χρήσιμη εντολή είναι η str(), με την οποία έχουμε σαφή πληροφόρηση για το είδος των μεταβλητών που καλούμαστε να χειριστούμε:

```
> str(FinalData)
Classes 'tbl_df', 'tbl' and 'data.frame':       556 obs. of 126 variables:
$ X__1 : chr "1" "2" "3" "4" ...
$ faculty : chr "ΟΔΕ" "ΑΚΕ" "ΟΔΕ" "ΠΣ" ...
$ deppol : chr NA NA NA "ΜΗΥΠ" ...
$ dephum : chr NA "ΦΛΛ" NA NA ...
$ depmed : chr NA NA NA NA ...
$ depsci : chr NA NA NA NA ...
$ depbus : chr "ΔΕ" NA "ΔΕ" NA ...
$ introyear : num NA 2014 2013 2016 2015 ...
$ semester : num 4 7 13 3 5 3 3 5 3 3 ...
$ entrance : chr "Ειδική κατηγορία" "Πανελλήνιες" "Πανελλήνιες" "Πανελλήνιες" ...
$ attempt : chr "1η" "1η" "1η" "1η" ...
$ sex : chr "M" "M" "M" "M" ...
$ birth : num 1997 1996 1995 1998 1997 ...
$ country : chr "Ελλάδα" "Ελλάδα" "Ελλάδα" "Ελλάδα" ...
$ growthb : chr "Δύο γονείς" "Δύο γονείς" "Δύο γονείς" "Δύο γονείς" ...
$ growtha : chr "Δύο γονείς" "Δύο γονείς" "Δύο γονείς" "Δύο γονείς" ...
$ nkids : num 2 2 2 2 2 3 2 1 2 2 ...
$ econsit : num 4 2 4 4 3 3 4 3 3 3 ...
$ housingneeds : num 5 4 4 5 4 4 4 4 5 5 ...
$ basicneeds : num 5 4 5 5 4 4 5 5 5 4 ...
$ entertainmentneeds : num 4 3 4 4 4 4 4 4 5 4 3 ...
$ culturalneeds : num 5 2 3 3 4 4 4 3 5 4 ...
$ educationneeds : num 5 3 5 5 4 3 4 4 5 4 ...
```

$ urbanity : chr "Αστική" "Ημιαστική" "Ημιαστική" "Αστική" ...
$ typeprimaryschool : chr "Δημόσιο" "Δημόσιο" "Δημόσιο" "Δημόσιο" ...
$ typegymnschool : chr "Δημόσιο" "Δημόσιο" "Δημόσιο" "Δημόσιο" ...
$ typehighschool : chr "Πειραματικό" "Δημόσιο" "Δημόσιο" "Πειραματικό" ...
$ bacgrade : chr "16-17,9" "18-20" "12-13,9" "16-17,9" ...
$ tuition : chr "TRUE" "TRUE" "TRUE" "TRUE" ...
$ tuitionclass : chr "Α Λυκείου" "Γ Λυκείου" "Γυμνάσιο" "Γυμνάσιο" ...
$ typeofcourses : chr "Ιδιαίτερα" "Έως 4 άτομα" "Έως 4 άτομα" "Πάνω από 4 άτομα" ...
$ readinghours : num 3 5 4 3 3 4 2 6 4 1 ...
$ workload : num 3 5 5 5 4 4 3 5 4 5 ...
$ stress : num 4 3 3 3 3 2 5 3 5 5 ...
$ actsports : chr "FALSE" "FALSE" "FALSE" "TRUE" ...
$ actart : chr "FALSE" "FALSE" "FALSE" "FALSE" ...
$ actgroup : chr "FALSE" "TRUE" "TRUE" "TRUE" ...
$ acttravel : chr "FALSE" "FALSE" "TRUE" "FALSE" ...
$ actinternet : chr "TRUE" "TRUE" "TRUE" "FALSE" ...
$ nactivities : num 4 3 2 3 1 3 2 3 3 2 ...
$ sports : chr "TRUE" "FALSE" "FALSE" "TRUE" ...
$ arts : chr "TRUE" "FALSE" "FALSE" "FALSE" ...
$ stopactivity : chr "TRUE" "FALSE" "FALSE" "TRUE" ...
$ parentattend : num 3 1 1 3 3 1 3 2 1 3 ...
$ familyhelp : num 4 5 5 4 5 5 4 5 5 5 ...
$ schoolhelp : num 1 1 1 1 4 4 3 4 5 3 ...
$ tuitionhelp : num 2 4 4 4 5 5 3 5 5 5 ...
$ friendhelp : num 3 4 2 5 4 4 3 4 5 4 ...
$ allhelp : num 10 14 12 14 18 18 13 18 20 17 ...
$ infulence1 : num 5 6 1 5 5 6 5 5 5 6 ...
$ infulence2 : num 2 5 5 1 1 1 1 1 2 1 ...
$ infulence3 : num 6 1 2 2 4 5 2 2 1 2 ...
$ infulence4 : num 3 3 3 4 2 4 3 3 4 5 ...
$ infulence5 : num 4 4 4 6 3 3 4 4 3 4 ...
$ infulence6 : num 1 2 6 3 6 2 6 6 6 3 ...
$ personalinfscore : num 77 72 89 90 90 72 90 90 90 73 ...
$ score : num 14000 17522 12000 14000 13500 ...
$ order : chr ">3η" "1η" ">3η" ">3η" ...
$ hierarchy01 : num 10 10 6 6 9 10 5 10 8 10 ...
$ hierarchy02 : num 8 9 8 3 6 8 1 1 4 4 ...
$ hierarchy03 : num 4 6 7 5 1 9 4 3 5 7 ...
$ hierarchy04 : num 1 7 5 1 7 1 6 4 1 1 ...
$ hierarchy05 : num 5 5 3 4 5 5 7 5 6 9 ...
$ hierarchy06 : num 7 8 1 2 3 3 2 2 9 8 ...
$ hierarchy07 : num 9 3 2 8 4 7 3 7 3 6 ...
$ hierarchy08 : num 3 2 4 7 2 6 8 8 2 5 ...
$ hierarchy09 : num 2 1 9 9 8 4 9 6 7 2 ...
$ hierarchy10 : num 6 4 10 10 10 2 10 9 10 3 ...
$ personalierascore : num 263 272 352 371 338 274 362 307 322 258 ...
$ professionalimp : num 4 4 2 5 5 5 4 5 5 4 ...
$ scientificimp : num 4 3 5 5 5 5 4 3 5 5 ...
$ workstabimp : num 4 5 3 5 5 5 4 5 5 5 ...
$ helpingimp : num 3 4 5 4 5 4 4 3 2 5 ...
$ wealthimp : num 3 2 3 5 4 4 4 4 5 1 ...
$ socprestigeimp : num 5 1 1 5 4 3 3 3 5 4 ...

$ identifiabilityimp : num 4 2 1 5 4 5 3 3 5 3 ...
$ allimportance : num 27 21 20 34 32 31 26 26 32 27 ...
$ coursepercentage : chr "μεγαλύτερο από 75%""μεγαλύτερο από 75%""25% έως 49%""25% έως 49%" ...
$ grade : num 7.4 7.5 8.3 5.5 6 6.5 7.4 7 6.7 6.9 ...
$ courseattfreq : num 4 4 6 4 2 6 4 4 5 6 ...
$ q401 : num 4 2 3 5 1 5 4 3 4 5 ...
$ q402 : num 4 3 4 4 1 5 4 3 4 4 ...
$ q403 : num 4 2 3 3 1 4 3 4 3 2 ...
$ q404 : num 3 3 5 3 3 3 2 2 2 3 ...
$ q405 : num 3 1 5 4 3 2 3 2 2 3 ...
$ q40 : num 18 11 20 19 9 19 16 14 15 17 ...
$ studysatisfaction : num 2 5 3 4 4 4 4 4 5 3 ...
$ q421 : num 3 2 2 2 2 2 3 2 1 2 ...
$ q422 : num 2 3 3 2 3 2 2 3 2 2 ...
$ q423 : num 4 2 1 2 1 2 2 3 1 2 ...
$ q424 : num 2 2 1 4 4 2 2 3 2 2 ...
$ q425 : num 2 5 2 4 4 5 5 4 3 4 ...
$ q426 : num 3 2 2 2 4 3 2 3 2 4 ...
$ q427 : num 2 1 1 2 1 2 2 2 1 2 ...
$ q428 : num 3 3 3 1 1 1 2 2 1 3 ...
$ q42 : num 21 20 15 19 20 19 20 22 13 21 ...
$ workdurstudies : chr "TRUE""FALSE""FALSE""FALSE" ...
$ reasonwork : chr "Εισόδημα" NA NA NA ...
$ relationwstud : chr "FALSE" NA NA NA ...
[list output truncated]

Αρκετές μεταβλητές του αρχείου μας είναι αριθμητικές και κάποιες εξ αυτών είναι χαρακτήρες. Το αρχείο που έχουμε στη διάθεσή μας αποτελείται από 556 παρατηρήσεις, ενώ αναφέρεται σε 126 μεταβλητές. Αντί της εντολής str(), μπορούμε να χρησιμοποιήσουμε τη head(), η οποία μας δίνει ένα πλαίσιο δεδομένων σε μια ορθογώνια και πιο συνοπτική απεικόνιση.

> head(FinalData)
A tibble: 6 x 126
X__1 faculty deppol dephum depmed depsci depbus introyear semester entrance
<*chr*> <*chr*> <*chr*> <*chr*> <*chr*> <*chr*> <*chr*> <*dbl*> <*dbl*> <*chr*>
1 1 ΟΔΕ NA NA NA NA ΔΕ NA 4.00 Ειδική κατηγο~
2 2 ΑΚΕ NA ΦΛΛ NA NA NA 2014 7.00 Πανελλήνιες
3 3 ΟΔΕ NA NA NA NA ΔΕ 2013 13.0 Πανελλήνιες
4 4 ΠΣ ΜΗΥΠ NA NA NA NA 2016 3.00 Πανελλήνιες
5 5 ΟΔΕ NA NA NA NA ΔΕ 2015 5.00 Πανελλήνιες
6 6 ΟΔΕ NA NA NA NA ΔΕ 2016 3.00 Πανελλήνιες
... with 116 more variables: attempt <*chr*>, sex <*chr*>, birth <*dbl*>, country <*chr*>,
growthb <*chr*>, growtha <*chr*>, nkids <*dbl*>, econsit <*dbl*>, housingneeds <*dbl*>,
basicneeds <*dbl*>, entertainmentneeds <*dbl*>, culturalneeds <*dbl*>,
educationneeds <*dbl*>, urbanity <*chr*>, typeprimaryschool <*chr*>,
typegymnschool <*chr*>, typehighschool <*chr*>, bacgrade <*chr*>, tuition <*chr*>,
tuitionclass <*chr*>, typeofcourses <*chr*>, readinghours <*dbl*>, workload <*dbl*>,
stress <*dbl*>, actsports <*chr*>, actart <*chr*>, actgroup <*chr*>, acttravel <*chr*>,
actinternet <*chr*>, nactivities <*dbl*>, sports <*chr*>, arts <*chr*>,
stopactivity <*chr*>, parentattend <*dbl*>, familyhelp <*dbl*>, schoolhelp <*dbl*>,
tuitionhelp <*dbl*>, friendhelp <*dbl*>, allhelp <*dbl*>, infulence1 <*dbl*>,
infulence2 <*dbl*>, infulence3 <*dbl*>, infulence4 <*dbl*>, infulence5 <*dbl*>,
infulence6 <*dbl*>, personalinfscore <*dbl*>, score <*dbl*>, order <*chr*>,

```
# hierarchy01 <dbl>, hierarchy02 <dbl>, hierarchy03 <dbl>, hierarchy04 <dbl>,
# hierarchy05 <dbl>, hierarchy06 <dbl>, hierarchy07 <dbl>, hierarchy08 <dbl>,
# hierarchy09 <dbl>, hierarchy10 <dbl>, personalierascore <dbl>,
# professionalimp <dbl>, scientificimp <dbl>, workstabimp <dbl>, helpingimp <dbl>,
# wealthimp <dbl>, socprestigeimp <dbl>, identifiabilityimp <dbl>,
# allimportance <dbl>, coursepercentage <chr>, grade <dbl>, courseattfreq <dbl>,
# q401 <dbl>, q402 <dbl>, q403 <dbl>, q404 <dbl>, q405 <dbl>, q40 <dbl>,
# studysatisfaction <dbl>, q421 <dbl>, q422 <dbl>, q423 <dbl>, q424 <dbl>,
# q425 <dbl>, q426 <dbl>, q427 <dbl>, q428 <dbl>, q42 <dbl>, workdurstudies <chr>,
# reasonwork <chr>, relationwstud <chr>, furtherstudies <chr>,
# kindofstudies <chr>, educinstit <chr>, sortreasons1 <dbl>, sortreasons2 <dbl>,
# sortreasons3 <dbl>, sortreasons4 <dbl>, sortreasons5 <dbl>, sortreasons6 <dbl>,
# sortreasons7 <dbl>, personalreasonscore <dbl>, ...
```

Τέλος, φαντάζει αρκετά σημαντική η χρήση της εντολής summary().

Στην περίπτωση τώρα που θέλουμε να εξετάσουμε από κοινού δύο μεταβλητές, η χρήση της εντολής xtabs() είναι αναγκαία. Για παράδειγμα, εάν θέλουμε μια απεικόνιση μεταξύ των ωρών διαβάσματος και των ωρών σε φόρτο εργασίας, θα έχουμε το παρακάτω αποτέλεσμα:

```
xtabs(~readinghours+workload, data=FinalData)
            workload
readinghours  1  2  3  4  5
           1  0  0 10  8  3
           2  1  3 12 35 18
           3  0  3 10 62 29
           4  0  2 11 54 50
           5  0  0 12 50 49
           6  0  0  1 23 37
           7  0  0  3  4 12
           8  0  0  0  6 16
           9  0  0  0  2  4
           10 0  1  1  6  4
           11 0  0  0  1  0
           12 0  0  1  0  3
           13 0  0  0  1  0
           14 0  0  0  1  1
           15 0  0  1  2  0
           16 0  0  0  0  1
           17 0  0  0  0  1
           20 0  0  1  0  0
```

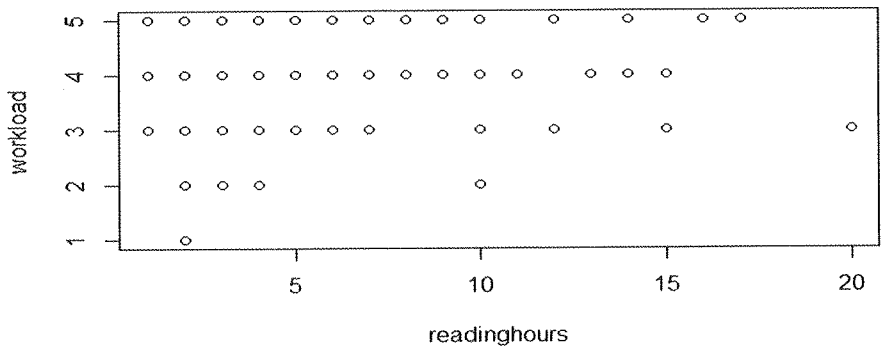
Βέβαια, μπορούμε να αποτυπώσουμε και μεταβλητές οι οποίες δεν είναι ποσοτικές και να λάβουμε τους αντίστοιχους πίνακες συνάφειας.

```
xtabs(~readinghours+sex, data=FinalData)
            sex
readinghours  F  M
           1  2 19
           2 20 49
           3 49 55
           4 58 59
           5 63 48
           6 36 25
           7 11  8
           8 11 11
           9  3  3
           10 9  3
           11 1  0
           12 2  2
           13 1  0
           14 2  0
           15 3  0
           16 1  0
           17 1  0
           20 0  1
```

Προφανώς μπορούμε να επεκτείνουμε λαμβάνοντας σχέσεις με περισσότερες μεταβλητές [π.χ. xtabs(~reading-hours+sex+workload, data=FinalData)].

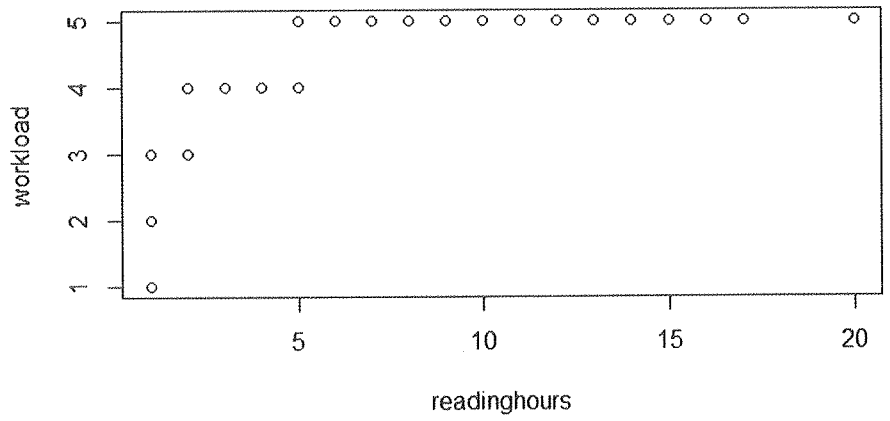
Π.VII.3 Χρήση διαγραμμάτων

Εκτός από τον σημαντικό ρόλο που έχουν τα στατιστικά μέτρα θέσης και μεταβλητότητας, σημαντικό ρόλο έχει και η απεικόνιση μέσω διαφόρων διαγραμμάτων. Για παράδειγμα, στην Εικόνα Π.VII.1 παρουσιάζεται η σχέση ανάμεσα στις ώρες ημερήσιου διαβάσματος και τον φόρτο εργασίας για το δείγμα των 556 ερωτώμενων, ενώ στην Εικόνα Π.VII.2 η απεικόνιση του QQ-plot. Στην Εικόνα Π.VII.3 (σελ. 207) δίνεται η γραφική παράσταση των μεταβλητών σε λογαριθμική κλίμακα.



ΕΙΚΟΝΑ Π.VII.1

Απεικόνιση μεταβλητών ημερήσιων ωρών εργασίας και φόρτου εργασίας.



ΕΙΚΟΝΑ Π.VII.2

QQ-plot απεικόνιση των μεταβλητών ημερήσιων ωρών εργασίας και φόρτου εργασίας.

Π.VII.4 Εκτίμηση γραμμικού υποδείγματος

Στην R η εκτέλεση της γραμμικής παλινδρόμησης γίνεται μέσω της εντολής lm. Πριν ωστόσο προχωρήσουμε στην εκτίμηση, ας εξετάσουμε χωριστά και για τις δύο μεταβλητές ενδιαφέροντος τα βασικά τους στατιστικά μέτρα απομονώνοντας αυτές από το υπόλοιπο σετ δεδομένων. Αυτό γίνεται μέσω των παρακάτω εντολών:

```
banks1<-FinalData [c("readinghours"),("workload")]
summary(banks1)
```

```
> linreg1<-lm(readinghours~workload, data=FinalData)
> summary(linreg1)

Call:
lm(formula = readinghours ~ workload, data = FinalData)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0214 -1.4124 -0.4124  0.9786 16.1965

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9766     0.6061   3.261  0.00118 **
workload      0.6090     0.1403   4.342 1.68e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.424 on 554 degrees of freedom
Multiple R-squared:  0.03291,  Adjusted R-squared:  0.03116
F-statistic: 18.85 on 1 and 554 DF,  p-value: 1.682e-05
```

Ας δώσουμε μεγαλύτερη σημασία στα παραπάνω αποτελέσματα. Στο παραπάνω πλαίσιο αρχικά έχουμε την εντολή για την εκτέλεση της γραμμικής παλινδρόμησης. Ακριβώς κάτω από αυτό έχουμε μια αρχική εικόνα για τα υπόλοιπα του υποδείγματός μας με κάποια στατιστικά μέτρα όπως η διάμεσος, η μικρότερη και η μεγαλύτερη τιμή καθώς και το τρίτο τεταρτημόριο. Το παραπάνω πλαίσιο, και στη γραμμή αναφοράς με την ονομασία coefficients, παρέχει ουσιαστικά τα αποτελέσματα της εκτίμησής μας. Ο σταθερός όρος είναι 1,976, ενώ ο συντελεστής του φόρτου εργασίας είναι 0,6090 (στήλη Estimate – Εκτιμητής). Ωστόσο, μας ενδιαφέρει σε σημαντικό βαθμό να γνωρίζουμε εάν οι εκτιμήσεις μας έχουν και στατιστική σημαντικότητα, δηλαδή εάν έχουν στατιστική ισχύ στο υπόδειγμα που εξετάζουμε. Οι επόμενες στήλες με ονομασίες Std. Error και t value μας απαντούν σε αυτό το ερώτημα συγκρίνοντας την τιμή 3,261 και 4,342 με την αντίστοιχη τιμή (σε ε.σ. 5%), που είναι σε απόλυτη τιμή ίση με 1,96. Παρατηρώντας ότι και οι δύο τιμές είναι μεγαλύτερες από τη συγκεκριμένη τιμή, έχουμε απόρριψη της μηδενικής μας υπόθεσης ότι ο εκτιμητής μας δεν είναι στατιστικά σημαντικός, επομένως οι δύο μεταβλητές που χρησιμοποιούνται έχουν «αξία» για το συγκεκριμένο γραμμικό υπόδειγμα. Στο ίδιο συμπέρασμα θα οδηγηθούμε με βάση την τιμή πιθανότητας (Pr(>|t|), η οποία και στις δύο περιπτώσεις είναι μικρότερη των διαφορετικών επιπέδων σημαντικότητας 1%, 10% και 5%. Να σημειώσουμε ότι οι στατιστικές σημαντικότητες συμβολίζονται με διαφορετικά αστέρια στη συνέχεια της γραμμής Signif. Codes. Τέλος, δίνονται στοιχεία που αφορούν το υπόδειγμά μας γενικά, καθώς αποτυπώνονται τόσο οι συντελεστές προσδιορισμού (διορθωμένος και μη) (Multiple R-squared: 0.03291, Adjusted R-squared: 0.03116) όσο και ο αντίστοιχος έλεγχος (F-statistic: 18.85 on 1 and 554 DF, p-value: 1.682e-05).

Στην περίπτωση που θα θέλαμε να γνωρίσουμε περισσότερα για την εκτίμηση του απλού γραμμικού υποδείγματος, καλούμε την εντολή:

```
->class(linreg1)
->names(linreg1)
> class(linreg1)
[1]"lm"
> names(linreg1)
[1]"coefficients""residuals""effects""rank""fitted.values"
[6]"assign""qr""df.residual""xlevels""call"
[11]"terms""model"
```

όπου μπορούμε να έχουμε αρκετά στοιχεία για την εκτίμηση.

Προφανώς μπορούμε να αναζητήσουμε παραπάνω πληροφόρηση για τα υπόλοιπα (residulas), το υπόδειγμα (model), καθώς και τις προσαρμοσμένες τιμές (fitted.values). Συγκεκριμένα:

```
> linreg1#coefficients

Call:
lm(formula = readinghours ~ workload, data = FinalData)

Coefficients:
(Intercept)     workload
      1.977        0.609

> linreg1$assign
[1] 0 1
> linreg1$residuals[1:5]
          1           2           3           4           5
-0.80346821 -0.02138728 -1.02138728 -2.02138728 -1.41242775
> summary(linreg1$residuals)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-4.0214 -1.4124 -0.4124  0.0000  0.9786 16.1965
> linreg1$fitted.values[1:5]
       1        2        3        4        5
3.803468 5.021387 5.021387 5.021387 4.412428
> summary(linreg1$fitted.values)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.586   4.412   4.412   4.570   5.021   5.021
```

Από την άλλη πλευρά, μπορούμε να προχωρήσουμε και στην ανάλυση ANOVA με τη χρήση της παρακάτω εντολής:

```
> anova(linreg1)
Analysis of Variance Table

Response: readinghours
           Df Sum Sq Mean Sq F value    Pr(>F)
workload    1  110.8 110.769   18.85 1.682e-05 ***
Residuals 554 3255.5   5.876
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  2.586   4.412   4.412   4.570   5.021   5.021
```

Ο πίνακας ANOVA μάς παρέχει πληροφοριακό υλικό σχετικά με το άθροισμα των τετραγώνων γύρω από τον μέσα για την εξαρτημένη μας μεταβλητή θεωρώντας δύο πιθανές διαφορετικές πηγές. Η πρώτη αφορά την παρουσία της ανεξάρτητης μεταβλητής, δηλαδή του φόρτου εργασίας, ενώ η δεύτερη αυτή των σφαλμάτων. Στην παρούσα φάση δεν θα θέλαμε να επεκταθούμε περαιτέρω σ' αυτήν την ενότητα.

Προχωρούμε τώρα με την εκτίμηση της σχέσης ανάμεσα στις ώρες ημερήσιου διαβάσματος και τον τελικό βαθμό εισαγωγής.

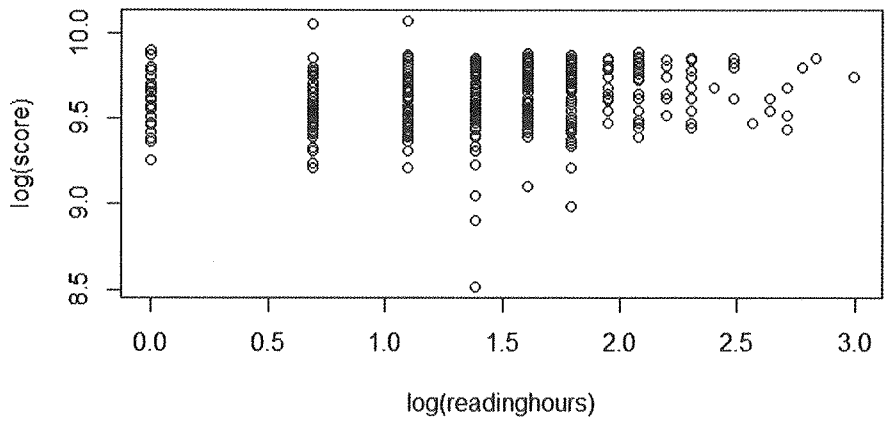
```
> linreg2<-lm(log(score)~log(readinghours),data=FinalData)
> summary(linreg2)

Call:
lm(formula = log(score) ~ log(readinghours), data = FinalData)

Residuals:
Min 1Q Median 3Q Max
-1.09564 -0.11551 0.00297 0.13912 0.47732
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.53039 0.02006 475.137 < 2e-16 ***
log(readinghours) 0.05947 0.01352 4.399 1.31e-05 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1658 on 553 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared: 0.03381,    Adjusted R-squared: 0.03206
F-statistic: 19.35 on 1 and 553 DF, p-value: 1.306e-05
```

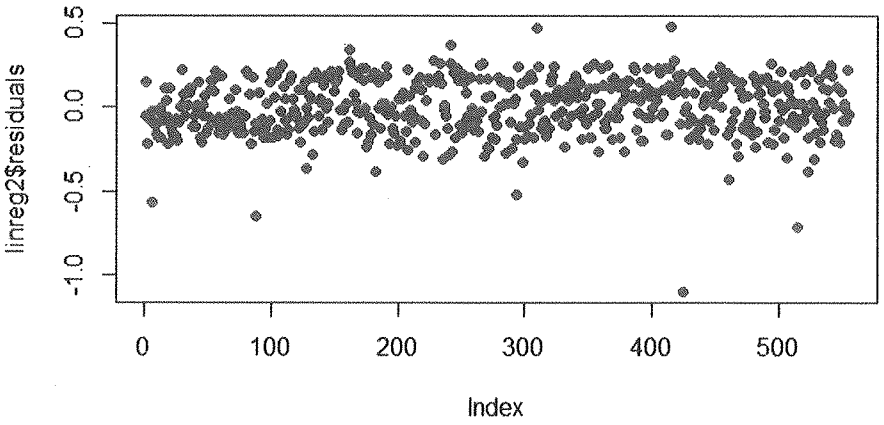


ΕΙΚΟΝΑ Π.VII.3

Απεικόνιση των μεταβλητών ημερήσιων ωρών εργασίας και βαθμού εισαγωγής (σε λογαριθμική κλίμακα).

Τέλος, παρατίθεται η γραφική παράσταση των υπολοίπων της εκτιμώμενης παλινδρόμησης (Εικόνα Π.VII.4).

```
> abline(linreg2)
> class(linreg2)
[1] "lm"
> plot(linreg2$residuals, pch = 16, col = "red")
```



ΕΙΚΟΝΑ Π.VII.4

Διαγνωστικό διάγραμμα υπολοίπων από την εκτίμηση της σχέσης μεταξύ των μεταβλητών ημερήσιων ωρών εργασίας και βαθμού εισαγωγής (σε λογαριθμική κλίμακα).

Στόχος από το παρόν διάγραμμα είναι σε πολύ γενικές γραμμές η μη ύπαρξη κάποιου προτύπου, αλλά η τυχόν κατανομή τους στον δισδιάστατο χώρο (όπως παρατηρούμε).

Τέλος, η παρακάτω εντολή μάς παρέχει και τις προβλέψεις:

predict(linreg2, FinalData)

Βιβλιογραφία

Bruce, P., and Bruce, A. (2017). *Practical Statistics for Data Scientists*. O'Reilly Media.

Chang, W. (2012). *R Graphics Cookbook: Practical recipes for visualizing data.* «O'Reilly Media, Inc.».

Crawley, M. J. (2005) *Statistics: An Introduction Using R* (John Wiley & Sons, Chichester).

Croissant, Y., & Millo, G. (2018). *Panel Data Econometrics with R*. John Wiley & Sons.

Field, A. P., Miles, J., and Field, Z. (2013). *Discovering Statistics Using R* (Sage, London).

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With applications in R*. Springer Publishing Company, Incorporated.

Keen, K. J. (2010). *Graphics for Statistics and Data Analysis with R*. CRC Press.

Kleiber, C., & Zeileis, A. (2008). *Applied Econometrics with R*. Springer Science & Business Media.

Raykov, T., and Marcoulides, G. A. (2013). *Basic Statistics: An introduction with R* (Rowman and Littlefield, Plymouth).

Καρλής, Δ., και Ντζούφρας, Ι. (2015). *Εισαγωγή στον Προγραμματισμό και τη Στατιστική Ανάλυση με R*. (https://repository.kallipos.gr/handle/11419/2601)

Φωκιανός, Κ., και Χαραλάμπους, Χ. (2010). *Εισαγωγή στην R* – Πρόχειρες Σημειώσεις, 2η έκδοση. Τμήμα Μαθηματικών και Στατιστικής, Πανεπιστήμιο Κύπρου. (https://cran.r-project.org/doc/contrib/mainfokianoscharalambous.pdf)

ΜΕΡΟΣ 2

Επαγωγική στατιστική

Γενίκευση πέρα από τα δεδομένα

Πρόλογος

Τα επόμενα κεφάλαια πραγματεύονται το πρόβλημα της γενίκευσης πέρα από σύνολα πραγματικών παρατηρήσεων. Τα δύο επόμενα κεφάλαια αναπτύσσουν κάποιες βασικές θεωρίες και εργαλεία για την επαγωγική στατιστική, ενώ τα επόμενα παρουσιάζουν μια σειρά από στατιστικούς ελέγχους ή διαδικασίες που μας επιτρέπουν να γενικεύουμε πέρα από ένα παρατηρούμενο αποτέλεσμα, από μια έρευνα ή ένα πείραμα, εξετάζοντας τις επιδράσεις της τύχης.