1 Η αναζήτηση της αιτιότητας

Πώς ξέρουμε αυτά που ξέρουμε; Ή, τουλάχιστον, γιατί σκεφτόμαστε ό,τι σκεφτόμαστε; Η σύγχρονη απάντηση είναι τα αποδεικτικά στοιχεία. Για να πείσουμε τους άλλους –και για να πείσουμε *τον εαυτό μας*– θα πρέπει να παρέχουμε πληροφορίες που μπορούν να επαληθευτούν. Κάτι που γνωρίζουμε διαισθητικά ή κάτι που απλώς «ξέρουμε» μπορεί να είναι σημαντικό αλλά δεν αποτελεί αποδεικτικό στοιχείο, όπως αυτά που καθοδηγούν τη σύγχρονη επιστημονική έρευνα.

Ποια είναι η βάση των αποδείξεών μας; Σε ορισμένες περιπτώσεις, μπορούμε να δούμε την αιτία και το αποτέλεσμα. Βλέπουμε ένα αναμμένο κερί να αναποδογυρίζει και να ανάβει φωτιά. Τώρα ξέρουμε τι προκάλεσε τη φωτιά. Αυτή είναι μια πολύ καλή γνώση. Μερικές φορές στις πολιτικές επιστήμες και στην πολιτική εντοπίζουμε μια αλυσίδα αιτιότητας με παρόμοιο τρόπο. Ωστόσο, αυτή η διαδικασία μπορεί να γίνει περίπλοκη. Γιατί ορισμένες οικονομίες μένουν στάσιμες ενώ άλλες ευδοκιμούν; Ποιες είναι οι οικονομικές και κοινωνικές επιδράσεις του διεθνούς εμπορίου; Γιατί ο Donald Trump κέρδισε τις προεδρικές εκλογές το 2016; Γιατί έχει μειωθεί η εγκληματικότητα στις Ηνωμένες Πολιτείες; Για τέτοιου είδους ερωτήσεις, δεν μπορούμε να λαμβάνουμε υπόψη μόνο ένα αναμμένο κερί· υπάρχουν και οι κεραυνοί, τα ελαττωματικά καλώδια, οι εμπρηστές και ποιος ξέρει τι άλλο. Σαφώς, σε αυτές τις σύνθετες περιπτώσεις είναι πολύ πιο δύσκολο να εντοπιστούν η αιτία και το αποτέλεσμα.

Όταν δεν είναι δυνατή η άμεση παρατήρηση της αιτίας και του αποτελέσματος, στρεφόμαστε φυσικά στα δεδομένα. Και τα δεδομένα υπόσχονται πολλά. Κάποιο κτί-

συσχέτιση ≠ αιτιότητα

ΣΧΗΜΑ 1.1: Κανόνας #1

ριο καταρρέει κατά τη διάρκεια ενός σεισμού. Τι ήταν αυτό που προκάλεσε την κατάρρευση του κτιρίου – και όχι άλλων στην ίδια πόλη; Ευθύνονται τα οικοδομικά υλικά; Το ύψος του; Ο σχεδιασμός του; Η ηλικία του; Η εσφαλμένη τοποθεσία; Παρότι ενδέχεται να μην είναι δυνατό να δούμε άμεσα την αιτία, μπορούμε να συλλέξουμε πληροφορίες για τα κτίρια που κατέρρευσαν και για εκείνα που έμειναν όρθια. Αν τα παλαιότερα κτίρια είναι πιο πιθανό να καταρρεύσουν, θα μπορούσαμε εύλογα να υποψιαστούμε ότι η ηλικία του κτιρίου έπαιξε ρόλο στην κατάρρευσή του. Αν τα κτίρια που κατασκευάζονταν χωρίς ενίσχυση από χάλυβα είναι πιο πιθανό να καταρρεύσουν ανεξάρτητα από την ηλικία τους, θα μπορούσαμε εύλογα να υποψιαστούμε ότι η κατασκευή του κτιρίου σχετίζεται με την κατάρρευσή του.

Κι όμως, δεν πρέπει να νιώθουμε απόλυτη σιγουριά. Ακόμη κι αν τα παλαιά κτίρια είναι πιο πιθανό να καταρρεύσουν, δεν γνωρίζουμε με βεβαιότητα ότι η ηλικία του κτιρίου είναι η κύρια εξήγηση για την κατάρρευση. Θα μπορούσε να ευθύνεται το γεγονός ότι τα περισσότερα κτίρια σχεδιάζονταν με έναν συγκεκριμένο τρόπο μια ορισμένη χρονική περίοδο. Θα μπορούσε να οφείλεται στην ύπαρξη περισσότερων παλαιών κτιρίων σε μια συνοικία που πλήττεται με μεγαλύτερη σφοδρότητα από τη σεισμική δραστηριότητα. Ή η κατάρρευση πολλών κτιρίων τα οποία έτυχε να είναι παλαιά θα μπορούσε να αποτελεί μια τεράστια σύμπτωση. Με άλλα λόγια, η συσχέτιση δεν είναι το ίδιο με την αιτιότητα. Επισημαίνουμε αυτό το γεγονός με μεγάλα μπλε γράμματα στο Σχήμα 1.1 επειδή συνιστά θεμελιώδες σημείο εκκίνησης σε κάθε σοβαρή ανάλυση δεδομένων.

Η οικονομετρία που μαθαίνουμε σε αυτό το βιβλίο θα μας βοηθήσει να εντοπίζουμε τις αιτίες και να διατυπώνουμε επιχειρήματα σχετικά με τις πραγματικές αιτίες ενός γεγονότος. Αν η συσχέτιση δεν είναι αιτιότητα, τότε *τι* σημαίνει αιτιότητα; Θα χρειαστεί ολόκληρο το βιβλίο για να δοθεί μια πλήρης απάντηση, αλλά ιδού μια σύντομη εκδοχή: *αν μπορούμε να εντοπίσουμε εξωγενή μεταβλητότητα*, τότε η συσχέτιση είναι πιθανώς αιτιότητα. Επομένως, θα πρέπει να καταλάβουμε τι σημαίνει εξωγενής μεταβλητότητα και πώς να διακρίνουμε την τυχαιότητα από την αιτιότητα όσο καλύτερα μπορούμε.

Σε αυτό το κεφάλαιο, εισάγουμε τρεις έννοιες που βρίσκονται στον πυρήνα του βιβλίου. Η Ενότητα 1.1 εξηγεί το βασικό υπόδειγμα που χρησιμοποιούμε σε όλο το σύγγραμμα. Η Ενότητα 1.2 παρουσιάζει δύο μεγάλες προκλήσεις που μπορεί να δυσχεράνουν τη χρήση δεδομένων για την κατανόηση του κόσμου. Δεν είναι τα μαθηματικά. (Αλήθεια!) Η πρώτη πρόκληση είναι η τυχαιότητα: μερικές φορές το τυχαίο της δειγματοληψίας μάς ωθεί να παρατηρούμε σχέσεις που δεν είναι πραγματικές. Άλλες φορές η τυχαία πιθανότητα μας κάνει να αγνοούμε σχέσεις που είναι πραγματικές. Η δεύτερη πρόκληση είναι η ενδογένεια, ένα φαινόμενο που μπορεί να μας κάνει να πιστεύουμε λανθασμένα ότι μια μεταβλητή προκαλεί κάποιο αποτέλεσμα ενώ αυτό δεν ισχύει. Στην Ενότητα 1.3 παρουσιάζονται τα τυχαιοποιημένα πειράματα ως ο ιδανικός τρόπος

για να ξεπεραστεί η ενδογένεια. Συνήθως, αυτά τα πειράματα δεν είναι δυνατό να διεξαχθούν, αλλά ακόμη και όταν υπάρχει αυτή η δυνατότητα, τα πράγματα μπορεί να πάνε στραβά. Ως εκ τούτου, το υπόλοιπο βιβλίο αφορά την ανάπτυξη μιας εργαλειοθήκης που μας βοηθά να ανταποκριθούμε (ή να προσεγγίσουμε) στο εξιδανικευμένο πρότυπο των τυχαιοποιημένων πειραμάτων.

1.1 Το βασικό υπόδειγμα

Όταν μιλάμε για αιτία και αποτέλεσμα, θα αναφερόμαστε στο αποτέλεσμα ενδιαφέροντος ως την **εξαρτημένη μεταβλητή** (dependent variable). Θα αναφερόμαστε σε μια πιθανή αιτία ως την **ανεξάρτητη μεταβλητή** (independent variable). Η εξαρτημένη μεταβλητή, που συνήθως συμβολίζεται με *Y*, ονομάζεται έτσι επειδή η τιμή της *εξαρτάται* από την ανεξάρτητη μεταβλητή. Η ανεξάρτητη μεταβλητή, που συνήθως συμβολίζεται με *X*, ονομάζεται έτσι γιατί κάνει ό,τι θέλει. Είναι δυνητικά η αιτία κάποιας μεταβολής στην εξαρτημένη μεταβλητή.

▶ **εξαρτημένη μεταβλητή**
Το αποτέλεσμα ενδιαφέροντος, που συνήθως συμβολίζεται με *Y*.

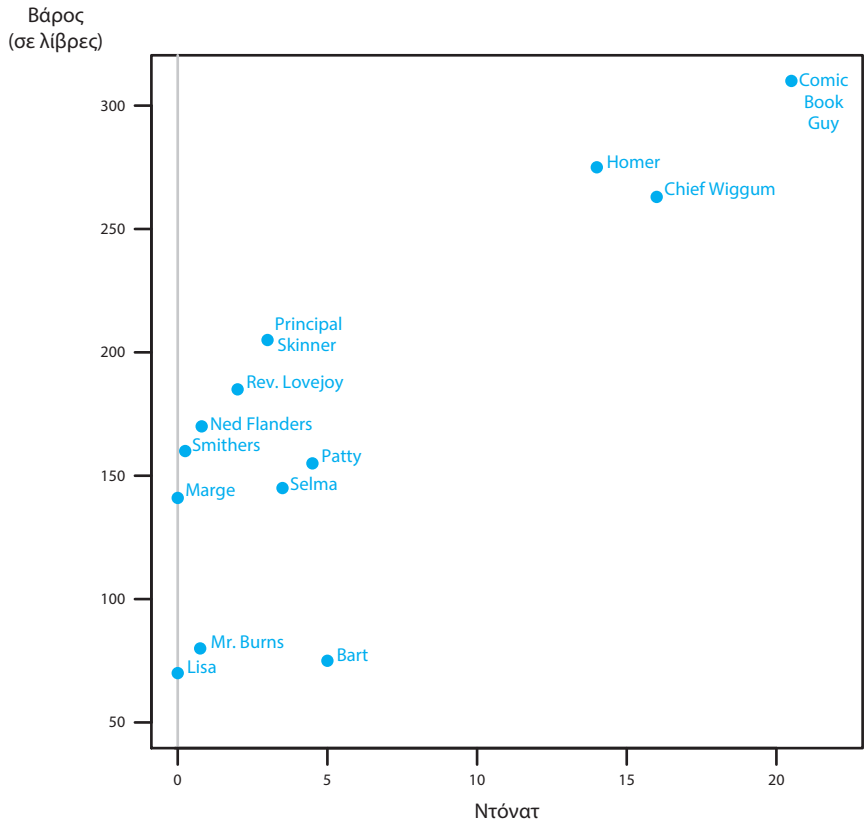
▶ **ανεξάρτητη μεταβλητή**
Μια μεταβλητή που πιθανώς επηρεάζει την τιμή της εξαρτημένης μεταβλητής.

Στον πυρήνα τους, οι κοινωνικές επιστημονικές θεωρίες υποστηρίζουν ότι μια μεταβολή σε κάτι (στην ανεξάρτητη μεταβλητή) οδηγεί σε μεταβολή σε κάτι άλλο (στην εξαρτημένη μεταβλητή). Θα εκφράσουμε με πιο τυπικό τρόπο αυτή τη σχέση στη συνέχεια. Τώρα, ας ξεκινήσουμε με ένα παράδειγμα. Ας υποθέσουμε ότι μας ενδιαφέρει η μεγάλη αύξηση της παχυσαρκίας στις ΗΠΑ και θέλουμε να αναλύσουμε την επίδραση της διατροφής στην υγεία. Μπορεί να αναρωτηθούμε, για παράδειγμα, αν τα ντόνατ προκαλούν προβλήματα υγείας. Στο υπόδειγμά μας η κατανάλωση ντόνατ (*Donuts*, μεταβλητή *X*, η ανεξάρτητη μεταβλητή μας) προκαλεί κάποια μεταβολή στο βάρος (*Weight*, μεταβλητή *Y*, η εξαρτημένη μεταβλητή μας). Αν καταφέρουμε να βρούμε δεδομένα για την ποσότητα των ντόνατ που κατανάλωσαν τα άτομα και το βάρος τους, μπορεί να βρεθούμε στα πρόθυρα μιας επιστημονικής ανακάλυψης.

Ας φανταστούμε μια μικρή πόλη στα μεσοδυτικά των ΗΠΑ και ας κάνουμε μια μικρή έρευνα. Το Σχήμα 1.2[1] απεικονίζει την κατανάλωση ντόνατ και το βάρος 13 ατόμων από μια τυχαία επιλεγμένη πόλη: το Σπρίνγκφιλντ των ΗΠΑ. Τα ακατέργαστα δεδομένα μας εμφανίζονται στον Πίνακα 1.1. Κάθε άτομο έχει μια γραμμή στον πίνακα. Ο Homer είναι η παρατήρηση 1. Εφόσον έφαγε 14 ντόνατ την εβδομάδα, $Donuts_1 = 14$. Θα αναφερόμαστε συχνά στη X_i ή στην Y_i, που είναι οι τιμές των *X* και *Y* για το άτομο *i* στο σύνολο δεδομένων. Το βάρος του έβδομου ατόμου στο σύνολο δεδομένων, του Smithers, είναι 160 λίβρες,[2] που σημαίνει $Weight_7 = 160$, κ.ο.κ.

1. Κατά σύμβαση, στα σχήματα και στους πίνακες του συγγράμματος οι ονομασίες των μεταβλητών παρουσιάζονται στα αγγλικά, όπως εμφανίζονται στα υπολογιστικά προγράμματα, ενώ τα υπόλοιπα στοιχεία που σχετίζονται με τη θεωρία παρουσιάζονται στα ελληνικά προς διευκόλυνση των φοιτητών. (Σ.τ.Ε.)

2. Η 1 λίβρα είναι ίση με περίπου 0.454 κιλά. (Σ.τ.Ε.)



ΣΧΗΜΑ 1.2: Βάρος και ντόνατ στο Σπρίνγκφιλντ

ΠΙΝΑΚΑΣ 1.1: Κατανάλωση ντόνατ και βάρος

Αριθμός παρατήρησης	Όνομα	Ντόνατ ανά εβδομάδα	Βάρος (λίβρες)
1	Homer	14	275
2	Marge	0	141
3	Lisa	0	70
4	Bart	5	75
5	Comic Book Guy	20	310
6	Mr. Burns	0.75	80
7	Smithers	0.25	160
8	Chief Wiggum	16	263
9	Principal Skinner	3	205
10	Rev. Lovejoy	2	185
11	Ned Flanders	0.8	170
12	Patty	5	155
13	Selma	4	145

Το Σχήμα 1.2 είναι ένα **διάγραμμα διασποράς** (scatterplot) δεδομένων, με κάθε παρατήρηση να βρίσκεται στις συντεταγμένες που ορίζονται από τις ανεξάρτητες και τις εξαρτημένες μεταβλητές. Η τιμή των ντόνατ ανά εβδομάδα βρίσκεται στον άξονα *X* και το βάρος στον άξονα *Y*. Κοιτάζοντας αυτό το διάγραμμα, αντιλαμβανόμαστε ότι υπάρχει μια θετική σχέση μεταξύ των ντόνατ και του βάρους, επειδή όσο περισσότερα ντόνατ καταναλώνονται από κάποιο άτομο, τόσο υψηλότερο τείνει να είναι το βάρος του.

▶ **διάγραμμα διασποράς** Ένα διάγραμμα δεδομένων στο οποίο κάθε παρατήρηση βρίσκεται στις συντεταγμένες που ορίζονται από τις ανεξάρτητες και τις εξαρτημένες μεταβλητές.

Χρησιμοποιούμε μια απλή εξίσωση για να χαρακτηρίσουμε τη σχέση μεταξύ των δύο μεταβλητών:

$$Weight_i = \beta_0 + \beta_1 Donuts_i + \epsilon_i \quad (1.1)$$

- Η εξαρτημένη μεταβλητή, $Weight_i$, είναι το βάρος του ατόμου *i*.
- Η ανεξάρτητη μεταβλητή, $Donuts_i$, είναι ο αριθμός των ντόνατ που καταναλώνει αυτό το άτομο σε μία εβδομάδα.
- Το β_1 είναι ο **συντελεστής κλίσης** (slope coefficient) στα ντόνατ, που δείχνει πόσο περισσότερο[3] ζυγίζει ένα άτομο για κάθε ντόνατ που τρώει.
- Το β_0 είναι ο **σταθερός όρος** (constant) ή **σημείο τομής** (intercept), που δείχνει το αναμενόμενο βάρος των ατόμων που τρώνε μηδέν ντόνατ.
- Το ϵ_i είναι ο **όρος σφάλματος** (error term), που καταγράφει οτιδήποτε άλλο επηρεάζει το βάρος.

▶ **συντελεστής κλίσης** Ο συντελεστής μιας ανεξάρτητης μεταβλητής. Δείχνει κατά πόσο αυξάνεται η εξαρτημένη μεταβλητή όταν η ανεξάρτητη μεταβλητή αυξάνεται κατά μία μονάδα.

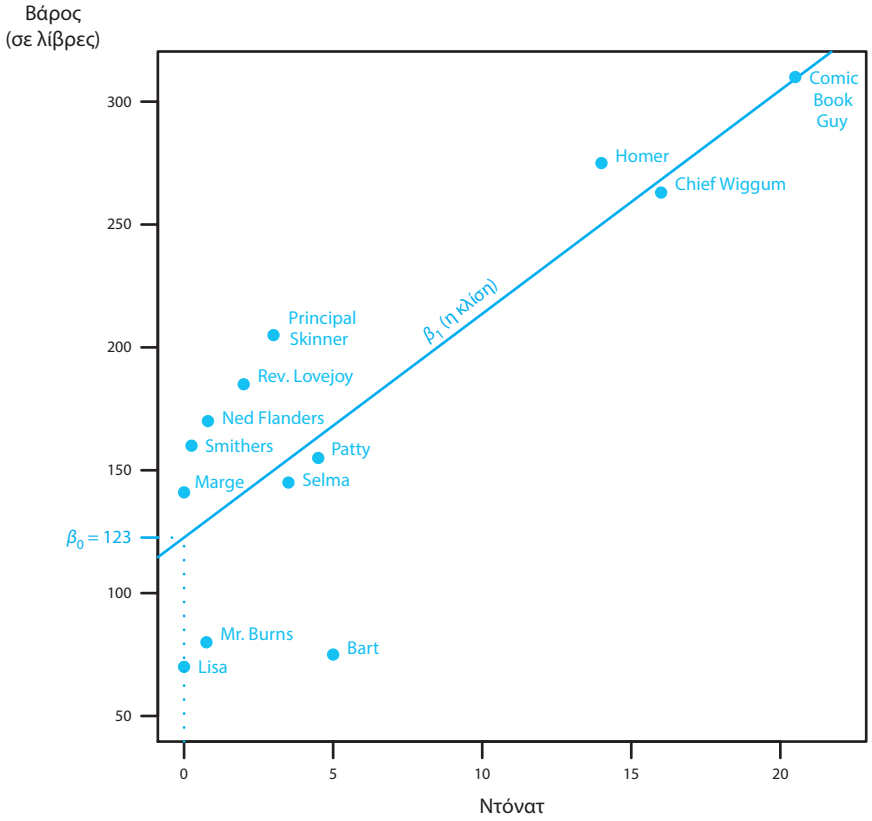
▶ **σταθερός όρος** Η παράμετρος β_0 σε ένα υπόδειγμα παλινδρόμησης. Είναι το σημείο στο οποίο μια γραμμή παλινδρόμησης τέμνει τον άξονα *Y*. Αναφέρεται επίσης ως *σημείο τομής*.

▶ **όρος σφάλματος** Ο όρος που σχετίζεται με μη μετρούμενους παράγοντες σε ένα υπόδειγμα παλινδρόμησης· τυπικά συμβολίζεται με ϵ.

Αυτή η εξίσωση θα μας βοηθήσει να εκτιμήσουμε τις δύο παραμέτρους που είναι απαραίτητες για να χαρακτηριστεί μια γραμμή. Θυμάστε τη σχέση $Y = mX + b$ από τα προηγούμενα χρόνια στο σχολείο; Αυτή είναι η εξίσωση για μια γραμμή όπου *Y* είναι η τιμή της γραμμής στον κάθετο άξονα, *X* είναι η τιμή στον οριζόντιο άξονα, *m* είναι η κλίση και *b* ο σταθερός όρος ή η τιμή της *Y* όταν η *X* είναι μηδέν. Η Εξίσωση 1.1 είναι ουσιαστικά η ίδια, μόνο που αναφερόμαστε στον όρο «*b*» ως β_0 και ονομάζουμε τον όρο «*m*» β_1.

Το Σχήμα 1.3 παρουσιάζει ένα παράδειγμα μιας πιθανής γραμμής από αυτό το υπόδειγμα για τα δεδομένα μας στο Σπρίνγκφιλντ. Το σημείο τομής (β_0) είναι η τιμή του βάρους όταν η κατανάλωση ντόνατ είναι μηδέν ($X = 0$). Η κλίση (β_1) είναι η ποσότητα κατά την οποία αυξάνεται το βάρος για κάθε ντόνατ που καταναλώνεται. Σε αυτή την περίπτωση, ο σταθερός όρος είναι περίπου 123, το οποίο σημαίνει ότι το αναμενόμενο βά-

3. Ή λιγότερο – ας είμαστε αισιόδοξοι!



ΣΧΗΜΑ 1.3: Γραμμή παλινδρόμησης για το βάρος και τα ντόνατ στο Σπρίνγκφιλντ

ρος για όσους δεν τρώνε κανένα ντόνατ είναι περίπου 123 λίβρες. Η κλίση είναι περίπου 9.1,[4] το οποίο σημαίνει ότι για κάθε ντόνατ που καταναλώνεται την εβδομάδα το βάρος αυξάνεται κατά περίπου 9.1 λίβρες.

Γενικότερα, το βασικό μας υπόδειγμα μπορεί να γραφεί ως

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1.2}$$

όπου β_0 είναι ο σταθερός όρος που δείχνει την τιμή της Y όταν $X = 0$ και β_1 είναι η κλίση που δείχνει πόση μεταβολή στην Y αναμένεται αν η X αυξηθεί κατά μία μονάδα. Σχεδόν πάντα ενδιαφερόμαστε για το β_1, που χαρακτηρίζει τη σχέση μεταξύ X και Y. Συνήθως δεν μας ενδιαφέρει πολύ ο όρος β_0. Μας βοηθάει, βέβαια, να φέρουμε τη γραμμή στη σωστή θέση, αλλά ο προσδιορισμός της τιμής της Y όταν η X είναι μηδέν σπάνια αποτελεί το βασικό ερευνητικό μας ενδιαφέρον.

4. Στο σύγγραμμα ακολουθείται το αγγλοσαξονικό σύστημα στους δεκαδικούς και στις χιλιάδες ώστε να συνάδουν με τον τρόπο εμφάνισής τους στα υπολογιστικά προγράμματα. (Σ.τ.Ε.)