2 Ακατέργαστα στατιστικά στοιχεία: Καλές πρακτικές διαχείρισης δεδομένων

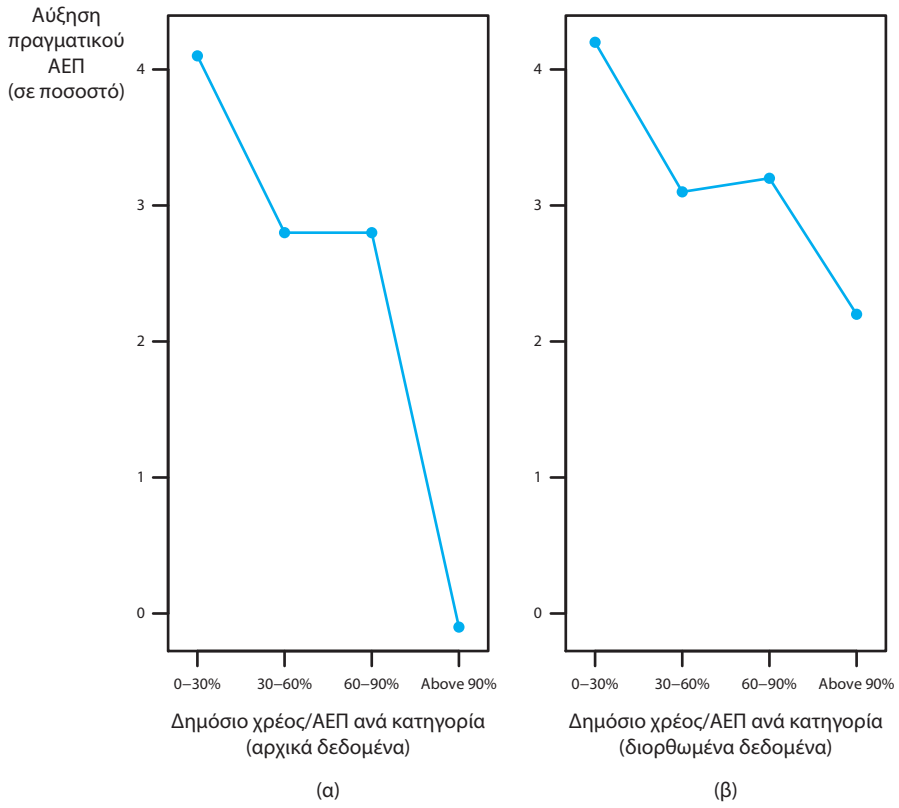
Στόχος μας είναι να χρησιμοποιούμε δεδομένα για να κατανοούμε καλύτερα τον κόσμο. Είδαμε στο προηγούμενο κεφάλαιο ότι η τυχαιότητα και η ενδογένεια δυσχεραίνουν την επίτευξη αυτού του στόχου. Μεγάλο μέρος του βιβλίου αφορά τον τρόπο αντιμετώπισης αυτών και άλλων προκλήσεων.

Ας πάρουμε όμως τα πράγματα από την αρχή: Η οικονομετρία απαιτεί δεδομένα. Και αν καταστρέψουμε τα δεδομένα μας, κανένα από τα εργαλεία που θα χρησιμοποιήσουμε αργότερα δεν θα μπορέσει να μας σώσει.

Είναι εύκολο να υποτιμάμε τη συλλογή και την οργάνωση δεδομένων ως μια ανιαρή εργασία. Είναι, ωστόσο, εξαιρετικά σημαντικές. Ας αναλογιστούμε τι συνέβη όταν οι οικονομολόγοι Carmen Reinhart και Ken Rogoff (2010) επιχείρησαν να διαπιστώσουν εάν το δημόσιο χρέος επηρέασε την οικονομική μεγέθυνση. Πρόκειται για ένα εξαιρετικά σημαντικό ερώτημα, καθώς όσο καλύτερα κατανοούμε τη μεγέθυνση τόσο αποτελεσματικότερα μπορούμε να καταπολεμήσουμε την ανεργία, η οποία καταστρέφει ζωές, απειλεί την υγεία και γενικά δημιουργεί μεγάλα προβλήματα.

Οι Reinhart και Rogoff συγκέντρωσαν περισσότερες από 3.700 ετήσιες παρατηρήσεις για την οικονομική μεγέθυνση από ένα μεγάλο δείγμα χωρών. Το πλαίσιο (α) του Σχήματος 2.1 παρουσιάζει ένα από τα βασικά αποτελέσματά τους, ομαδοποιώντας τη μέση εθνική αύξηση του ακαθάριστου εγχώριου προϊόντος (ΑΕΠ) σε τέσσερις κατηγορίες ανάλογα με τον λόγο του δημόσιου χρέους προς το ΑΕΠ κάθε χώρας. Το συγκλονιστικό εύρημα ήταν ότι η μέση οικονομική μεγέθυνση κατακρημνίστηκε στις χώρες όπου το δημόσιο χρέος ξεπέρασε το 90% του ΑΕΠ. Η επίπτωση ήταν προφανής: οι κυβερνήσεις θα πρέπει να είναι πολύ προσεκτικές όταν χρησιμοποιούν ελλειμματικές δαπάνες για την καταπολέμηση της ανεργίας.

Ωστόσο, υπήρχε ένα πρόβλημα στα ευρήματα των οικονομολόγων. Τα δεδομένα δεν συμφωνούσαν απόλυτα με τα συμπεράσματά τους. Οι Herndon, Ash και Pollin (2014) έκαναν κάποια έρευνα και διαπίστωσαν ότι ορισμένες παρατηρήσεις είχαν πα-



ΣΧΗΜΑ 2.1: Δύο εκδοχές δεδομένων για το χρέος και τη μεγέθυνση

ραλειφθεί, άλλες ήταν τυπογραφικά λάθη και, το πιο θλιβερό, ορισμένοι υπολογισμοί στο αρχικό υπολογιστικό φύλλο Excel των Reinhart και Rogoff ήταν λανθασμένοι. Μετά τη διόρθωση των δεδομένων, το γράφημα πήρε τη μορφή που φαίνεται στο πλαίσιο (β) του Σχήματος 2.1, δείχνοντας κάτι αρκετά διαφορετικό.

Η οικονομική μεγέθυνση δεν σημείωσε κατακόρυφη πτώση όταν το δημόσιο χρέος ξεπέρασε το 90% του ΑΕΠ. Παρότι θα μπορούσαμε να συζητήσουμε κατά πόσο η κλίση της καμπύλης στο πλαίσιο (β) είναι λίγο η περισσότερο απότομη, σαφώς δεν μοιάζει με την κατακρήμνιση που έδειχναν τα δεδομένα αρχικά.[1]

Θα μπορούσαμε να επωφεληθούμε από την ατυχή έρευνα των Reinhart και Rogoff αν συνειδητοποιήσουμε ότι ακόμη και κορυφαίοι μελετητές μπορεί να κάνουν λάθη

1. Ένα βαθύτερο ερώτημα είναι αν θα πρέπει να θεωρούμε ότι αυτά τα παρατηρούμενα δεδομένα έχουν οποιαδήποτε αιτιώδη ισχύ. Τα επίπεδα του δημόσιου χρέους είναι πιθανόν να σχετίζονται με άλλους παράγοντες οι οποίοι επηρεάζουν την οικονομική μεγέθυνση, όπως είναι οι πόλεμοι και η ποιότητα των θεσμών μιας χώρας. Με άλλα λόγια, υπάρχει η πιθανότητα το δημόσιο χρέος να είναι ενδογενές, γεγονός που σημαίνει ότι πιθανότατα δεν μπορούμε να καταλήξουμε σε συμπεράσματα σχετικά με τις επιπτώσεις του χρέους στη μεγέθυνση χωρίς να εφαρμόσουμε τεχνικές που θα παρουσιάσουμε αργότερα σε αυτό το βιβλίο.

στα δεδομένα. Ως εκ τούτου, θα πρέπει να αναπτύξουμε μεθόδους ώστε να ελαχιστοποιήσουμε τα λάθη και, εφόσον δεν καταφέρουμε να τα αποφύγουμε, να μεγιστοποιήσουμε την πιθανότητα να τα εντοπίσουν οι άλλοι.

Αυτό το κεφάλαιο εστιάζει στα κρίσιμα πρώτα βήματα κάθε οικονομετρικής ανάλυσης. Πρώτα, πρέπει να κατανοήσουμε τα δεδομένα μας. Η Ενότητα 2.1 παρουσιάζει εργαλεία για την περιγραφή των δεδομένων και τον εντοπισμό πιθανών σφαλμάτων ή ανωμαλιών. Δεύτερον, πρέπει να είμαστε προετοιμασμένοι να πείσουμε τους άλλους. Αν οι άλλοι δεν μπορούν να αναπαράγουν τα αποτελέσματά μας, τότε αυτά δεν θα πρέπει να λαμβάνονται υπόψη. Επομένως, η Ενότητα 2.2 μας βοηθά να αναπτύξουμε κατάλληλες μεθόδους, ώστε ο κώδικάς μας να είναι κατανοητός τόσο από εμάς όσο και από τους άλλους. Τέλος, θα πρέπει να διασφαλίσουμε ότι δεν θα κάνουμε όλη αυτή τη δουλειά με το χέρι. Έτσι, η Ενότητα 2.3 παρουσιάζει δύο σημαντικά στατιστικά προγράμματα λογισμικού, το Stata και το R. Αυτό το κεφάλαιο είναι σύντομο γιατί θα αφιερώσουμε επίσης χρόνο για να εξοικειωθούμε με τη χρήση του λογισμικού μας.

2.1 Γνωρίστε τα δεδομένα μας

Στην ιδανική περίπτωση, τα δεδομένα μας παράγονται σε καθαρά δωμάτια που στελεχώνονται από ρομπότ τελευταίας τεχνολογίας. Ωστόσο, δεν λειτουργεί έτσι ο κόσμος. Τα πειράματα κοινωνικών επιστημών, εφόσον μπορούν να διεξαχθούν τέτοιου είδους πειράματα, μπορεί να παράγουν αρκετά «ακατάστατα» δεδομένα. Τα δεδομένα που προέρχονται από παρατηρούμενη συμπεριφορά και όχι από πείραμα είναι ακόμη πιο «ακατάστατα».[2]

Επομένως, η πρώτη δουλειά στην ανάλυση δεδομένων είναι να γνωρίσουμε τα δεδομένα μας. Αυτός ο κανόνας φαίνεται προφανής και απλός, αλλά δεν εφαρμόζεται πάντα, οδηγώντας μερικές φορές σε ατυχή συμπεράσματα. Για κάθε μεταβλητή, θα πρέπει να γνωρίζουμε τον αριθμό των παρατηρήσεων, τη μέση και τυπική απόκλιση και τις ελάχιστες και μέγιστες τιμές. Η γνώση αυτών των πληροφοριών μάς δίνει μια καλή αίσθηση για τα δεδομένα, βοηθώντας μας να κατανοήσουμε αν λείπουν δεδομένα και ποιες είναι οι κλίμακες και τα εύρη των μεταβλητών. Ο Πίνακας 2.1 παρουσιάζει ένα παράδειγμα για τα δεδομένα που αφορούν τη σχέση της κατανάλωσης ντόνατ (Donuts) με το βάρος (Weight) τα οποία συζητήσαμε στη σελίδα 3. Ο αριθμός των παρατηρήσεων, που συχνά αναφέρεται ως «*N*» (από το αγγλικό number), είναι ο ίδιος για όλες τις μεταβλητές σε αυτό το παράδειγμα, αλλά ποικίλλει μεταξύ των μεταβλητών αν λείπουν κάποια στοιχεία. Όλοι γνωρίζουμε τον μέσο (γνωστό και ως αριθμητικός μέσος). Η **τυπική απόκλιση** μετρά πόσο ευρέως διασκορπισμένες εί-

▶ **τυπική απόκλιση** Η τυπική απόκλιση περιγράφει το εύρος των δεδομένων.

2. Ο Chris Achen (1982, 53) σημειώνει εύστοχα: «Αν οι πληροφορίες έχουν κωδικοποιηθεί από μη επαγγελματίες και δεν έχουν υποστεί καμία επεξεργασία, όπως συμβαίνει συχνά σε εργασίες πολιτικής ανάλυσης, είναι πιθανώς "βρόμικες"».

ΠΙΝΑΚΑΣ 2.1: Περιγραφική στατιστική για τα δεδομένα ντόνατ και βάρους

Μεταβλητή	Παρατηρήσεις (N)	Μέσος	Τυπική απόκλιση	Ελάχιστο	Μέγιστο
Weight	13	171.85	76.16	70	310
Donuts	13	5.41	6.85	0	20.5

ΠΙΝΑΚΑΣ 2.2: Πίνακας συχνοτήτων για τη μεταβλητή των ανδρών στο σύνολο δεδομένων των ντόνατ

Τιμή	Παρατηρήσεις
0	4
1	9

ΠΙΝΑΚΑΣ 2.3: Πίνακας συχνοτήτων για τη μεταβλητή των ανδρών στο δεύτερο σύνολο δεδομένων των ντόνατ

Τιμή	Παρατηρήσεις
0	4
1	8
100	1

ναι οι τιμές της παρατήρησης.[3] Το *ελάχιστο* και το *μέγιστο*, που μας δείχνουν το εύρος των δεδομένων, μπορούν να υποδεικνύουν παράλογες τιμές μιας μεταβλητής όταν το ελάχιστο ή το μέγιστο δεν έχει νόημα.

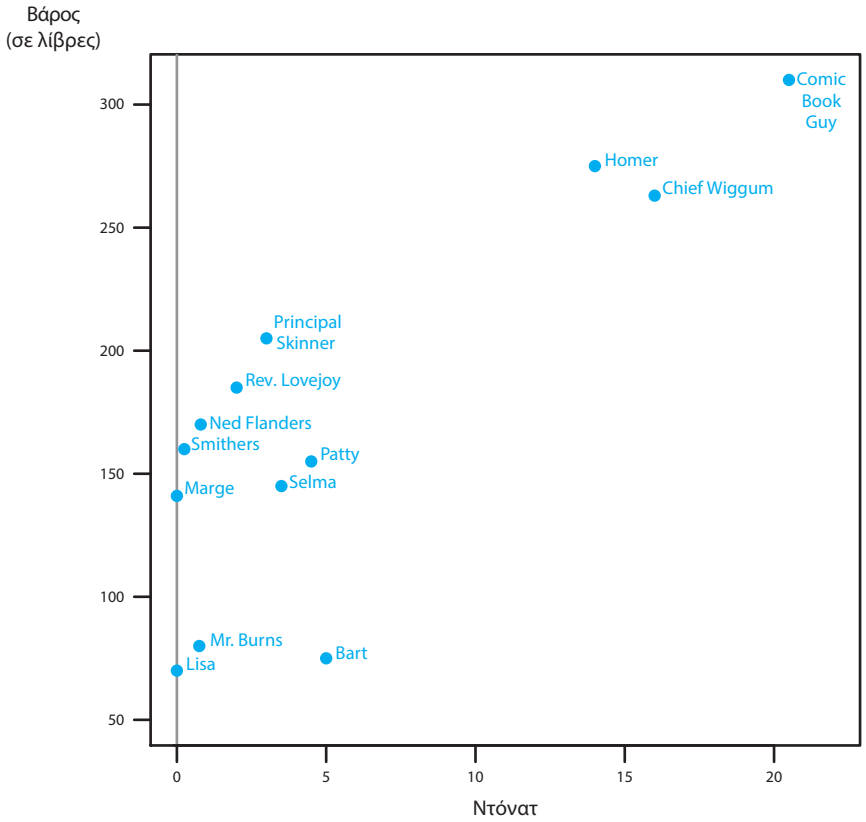
Αν μια μεταβλητή λαμβάνει λίγες μόνο τιμές, είναι επίσης χρήσιμο να εξετάσουμε την κατανομή των παρατηρούμενων τιμών. Ο Πίνακας 2.2 είναι ένας πίνακας συχνοτήτων για τη μεταβλητή των ανδρών, που ισούται με 1 για τους άνδρες και 0 για τις γυναίκες. Ο πίνακας δείχνει ότι το σύνολο δεδομένων των ντόνατ αποτελείται από εννέα άνδρες και τέσσερις γυναίκες. Αρκετά λογικό. Ας υποθέσουμε, όμως, ότι ο πίνακας συ-

3. Το Παράρτημα Γ περιλαμβάνει περισσότερες λεπτομέρειες. Εδώ κάνουμε μια σύντομη ανασκόπηση. Η τυπική απόκλιση του X είναι ένα μέτρο της διασποράς του X. Όσο μεγαλύτερη η τυπική απόκλιση τόσο πιο διασκορπισμένες είναι οι τιμές. Η τυπική απόκλιση υπολογίζεται ως $\sqrt{\frac{1}{N}\sum (X_i - \overline{X})^2}$, όπου $\overline{X}$ είναι ο μέσος του X. Καταγράφουμε πόσο απέχει κάθε παρατήρηση από τον μέσο όρο. Στη συνέχεια, τετραγωνίζουμε κάθε τιμή, καθώς προκειμένου να υπολογίσουμε τη διασπορά, δεν διακρίνουμε αν μια τιμή είναι κάτω από τον μέσο όρο ή πάνω από αυτόν. Όταν υψωθούν στο τετράγωνο, όλες αυτές οι τιμές γίνονται θετικοί αριθμοί. Καταγράφουμε τον μέσο όρο αυτών των τετραγωνικών τιμών. Τέλος, δεδομένου ότι είναι τετραγωνισμένες τιμές, λαμβάνοντας την τετραγωνική ρίζα του μέσου όρου προκύπτει η τελική τιμή πίσω στην κλίμακα της αρχικής μεταβλητής.

χνοτήτων μας έμοιαζε με τον Πίνακα 2.3. Είτε ο άνδρας που έχουμε στο δείγμα κάνει για πολλούς, είτε (το πιθανότερο) έχουμε κάνει κάποιο λάθος στα δεδομένα μας. Τα οικονομετρικά εργαλεία που χρησιμοποιούμε αργότερα σε αυτό το βιβλίο δεν θα επισημαίνουν απαραίτητα τέτοια ζητήματα, επομένως πρέπει να βρισκόμαστε σε εγρήγορση.

Η δημιουργία γραφικών δεδομένων είναι χρήσιμη επειδή μας επιτρέπει να βλέπουμε σχέσεις και να εντοπίζουμε ασυνήθιστες παρατηρήσεις. Τα εργαλεία που θα αναπτύξουμε αργότερα ποσοτικοποιούν αυτές τις σχέσεις, αλλά το να τις εντοπίζουμε μόνοι μας είναι ένα εξαιρετικό και απαραίτητο πρώτο βήμα. Για παράδειγμα, το Σχήμα 2.2 παρουσιάζει το διάγραμμα διασποράς των δεδομένων για τη σχέση βάρους και κατανάλωσης ντόνατ που είδαμε νωρίτερα. Μπορούμε να διαπιστώσουμε ότι φαίνεται να υπάρχει μια σχέση μεταξύ των δύο μεταβλητών.

Βλέπουμε, επίσης, κάποιες σχέσεις που μπορεί να μην εντοπίζαμε χωρίς γραφήματα. Η Lisa και ο Bart, για παράδειγμα, είναι παιδιά, επομένως το βάρος τους είναι πολύ χαμηλότερο. Αυτό θα θέλαμε πιθανώς να το λάβουμε υπόψη στην ανάλυσή μας. Οι γυναίκες φαίνεται επίσης να ζυγίζουν λιγότερο.



ΣΧΗΜΑ 2.2: Βάρος και ντόνατ στο Σπρίνγκφιλντ