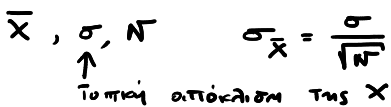


ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

13-03-2023

$s_N \xrightarrow{N\to\infty} \sigma$

t - κατανομή $\xrightarrow{N\to\infty}$ Κανονική κατανομή

Στη συνέχεια θα περιγράψουμε το διάστημα εμπιστοσύνης για την μέση τιμή του πληθυσμού στις ακόλουθες περιπτώσεις:

1. Η μεταβλητή X ακολουθεί κανονική κατανομή
2. Η μεταβλητή X δεν ακολουθεί κανονική κατανομή
   - Σε αυτή τη περίπτωση υποθέτουμε ότι το δείγμα είναι αρκετά μεγάλο ($\mathrm{n} \geq 30$)

- Επίσης θα εξετάσουμε χωρίστα αν γνωρίζουμε την τυπική απόκλιση σ ή όχι.
- Όταν το σ είναι άγνωστο χρειαζόμαστε τη t-κατανομή.

- όταν το σ είναι γνωστό θα έχουμε

Τυπική απόκλιση της $\bar{X}$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

Διάστημα εμπιστοσύνης της μ

Το $(1-a)*100\%$ διάστημα εμπιστοσύνης για την μ είναι:

$$[\bar{X} - z\sigma_{\bar{X}}, \bar{X} + z\sigma_{\bar{X}}]$$

όπου το z (z-score) λαμβάνεται έτσι ώστε

$$P(Z < z) = 1 - a/2$$

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \sqrt{n}\,\frac{\bar{x} - \mu}{\sigma}$$

$$Z \sim N(0,1)$$

- Περιθώριο σφάλματος: $E = z\sigma_{\bar{X}}$

$T \sim t_{df}$

- Είναι γνωστή και ως Student's t distribution και σχετίζεται με την τυπική κανονική κατανομή.
- Όπως και η τυπική κανονική κατανομή η t-κατανομή είναι συμμετρική γύρω από το μηδέν, έχει καμπανοειδη μορφή και η συνάρτηση πυκνότητας πιθανότητας είναι παντού θετική.
- Παρουσιάζει μεγαλύτερη διασπορά τιμών σε σχέση τη τυπική κανονική κατανομή.
- Η μορφή της εξαρτάται από το μέγεθος του δείγματος N. Μάλιστα η μοναδική παράμετρος της συμβολίζεται με df και είναι άμεσα συνδεδεμένη με το N.

$$\mathrm{df} = \mathrm{N} - 1 \quad (\text{βαθμοί ελευθερίας})$$

$df \to \infty \Rightarrow t_{df \to \infty} \to N(0,1)$

- Όσο το df αυξάνει η t-κατανομή προσεγγίζει όλο και περισσότερο την τυπική κανονική κατανομή.

- Την t-κατανομή με df βαθμούς ελευθερίας θα την συμβολίζουμε ως $\mathrm{t_{df}}$
- Συνάρτηση πυκνότητας πιθανότητας

$$\mathrm{p(t)} = \frac{\Gamma(\frac{\mathrm{df}+1}{2})}{\sqrt{\mathrm{df}\pi}\Gamma(\frac{\mathrm{df}}{2})}\left(1+\frac{\mathrm{t}^2}{\mathrm{df}}\right)^{-\frac{\mathrm{df}+1}{2}}$$

- Μέση τιμή

$$\mathbb{E}(\mathrm{T}) = 0$$

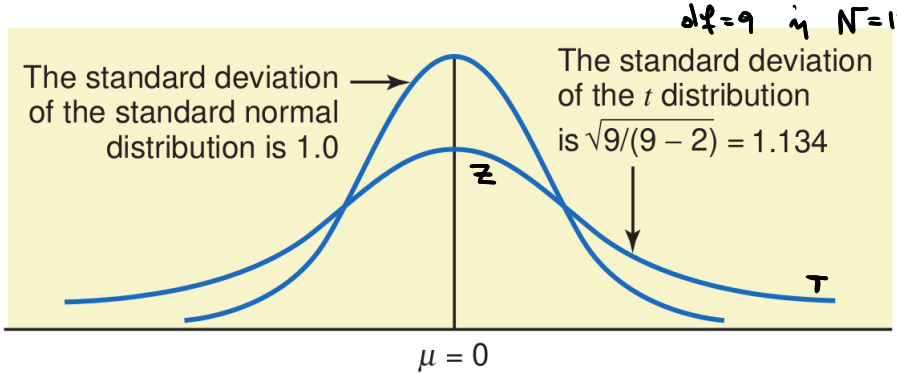
$$\mathbb{E}[T] = \int_{-\infty}^{+\infty} t\, p(t)\, dt = 0$$

- Διασπορά

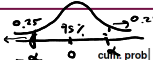
$$\mathbb{V}(\mathrm{T}) = \mathrm{df}/(\mathrm{df}-2) \quad , \mathrm{df} > 2$$

$$V[T] = \mathbb{E}[T^2] - (\mathbb{E}[T])^2$$

$$\sigma_T^2 = V[T] = \frac{N-1}{N-3} \quad , N > 3$$


df=9 ή N=10
The standard deviation of the standard normal distribution is 1.0
The standard deviation of the t distribution is $\sqrt{9/(9-2)} = 1.134$
Z
T
$\mu = 0$

t-Κατανομή (t-distribution)



Τιθασσώτητες

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
→ two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
→ 20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
→ 24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
→ ∞ z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

t-scores

df = 20 → ∞ → z

$\mathbb{P}\left[T \in [-2.086, 2.086]\right] =$

$= 0.95$

$\mathbb{P}\left[T \in [-1.325, 1.325]\right) =$

$= 0.8$

$z = 1.96$ 95%

$[\bar{x} - 1.96\,\sigma_{\bar{x}}, \bar{x} + 1.96\,\sigma_{\bar{x}}]$

Υπενθύμιση του πίνακα των z-scores

$N(0,1)$

$Z \sim N(0,1)$

Standard Normal Probabilities

$P[Z \leq 0.94] = 0.8264$

$P[Z \leq -0.4] = 1 - P[Z \leq 0.4]$

$= 1 - 0.6554$

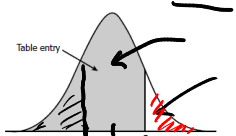


Table entry for z is the area under the standard normal curve to the left of z.

z-score

πιθανότητες

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

- όταν το σ δεν είναι γνωστό δεν μπορούμε να υπολογίσουμε το $\sigma_{\bar{X}}$

Εκτιμήτρια της τυπικής απόκλισης της $\bar{X}$

$$s_{\bar{X}} = \frac{s}{\sqrt{N}}$$

Διάστημα εμπιστοσύνης της μ

Το $(1-a)*100\%$ διάστημα εμπιστοσύνης για την μ είναι:

$$[\bar{X} - t s_{\bar{X}}, \bar{X} + t s_{\bar{X}}]$$

$\left([\bar{X} - z\sigma_{\bar{x}}, \bar{X} + z\sigma_{\bar{x}}] \right)$

όπου το t λαμβάνεται από την $t_{df},\ df = N - 1$ έτσι ώστε

$$P(T < t) = 1 - a/2$$

- Περιθώριο σφάλματος: $E = t s_{\bar{X}}$

Παράδειγμα

Έστω ότι η μεταβλητή X ακολουθεί κανονική κατανομή. Έστω επίσης ότι για ένα δείγμα με 25 στοιχεία λάβαμε:

$$\bar{\mathrm{X}} = 186, \quad \mathrm{s} = 12$$

1. Κατασκευάστε το 95 % διάστημα εμπιστοσύνης για την μέση τιμή μ.
2. Εάν για τη μελέτη μας το περιθώριο του σφάλματος θεωρείται μεγάλο τι θα μπορούσαμε να κάνουμε για να το μειώσουμε;
3. τι θα άλλαζε αν γνωρίζαμε ότι $\sigma = 12$.

(1) σ άγνωστο, $S_{\bar{x}} = \frac{S}{\sqrt{25}} = \frac{12}{5} = 2.4$ $t_{24} = 2.064$ $z = 1.96$

$\mu \in [186 - 2.4\, t_{24}, 186 + 2.4\, t_{24}] = [181.0464, 190.9536]$

(3) $\mu \in [186 - 2.4 \cdot z, 186 + 2.4 \cdot z] = [181.296, 190.704]$

Διάστημα Εμπιστοσύνης του μ

Δύο μεταβλητές που αναφέρονται στα ίδια στοιχεία λέμε ότι σχετίζονται αν κάποιες τιμές της μια μεταβλητής τείνουν να εμφανίζουν πιο συχνά όταν η δεύτερη μεταβλητή λαμβάνει συγκεκριμένες τιμές.

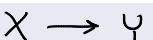
Εξαρτημένη μεταβλητή

Ονομάζεται η μεταβλητή για την οποία θέλουμε να περιγράψουμε και να εξηγήσουμε την συμπεριφορά της. Συνήθως συμβολίζεται με Y.

Ανεξάρτητη μεταβλητή

Ονομάζεται η μεταβλητή η οποία χρησιμοποιείται για να δικαιολογήσει τις αλλαγές των τιμών της εξαρτημένης μεταβλητής. Συνήθως συμβολίζεται με X.

Παράδειγμα



Το αλκοόλ προκαλεί πολλές παρενέργειες στον οργανισμό όπως είναι η πτώση της θερμοκρασίας. Για τη μελέτη του φαινομένου, οι ερευνητές δίνουν διαφορετικές ποσότητες αλκοόλης σε ποντίκια και έπειτα μετρούν την αλλαγή της θερμοκρασίας τους 15 λεπτά μετά τη λήψη. Η **ποσότητα της αλκοόλης** είναι η **ανεξάρτητη μεταβλητή** ενώ **η μεταβολή της θερμοκρασίας** είναι η **εξαρτημένη μεταβλητή**.

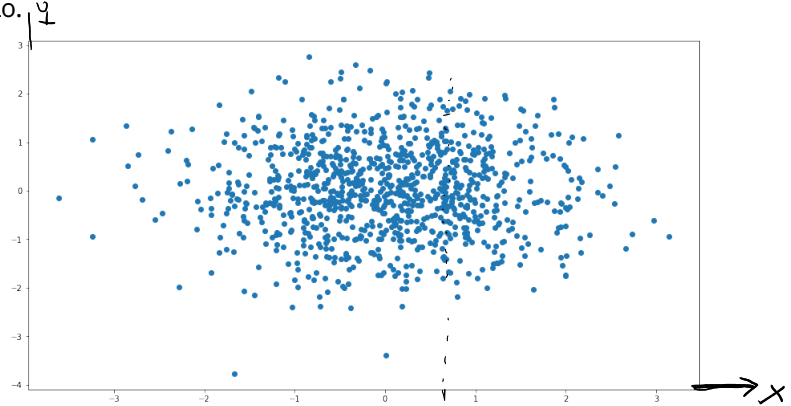
$X \quad \{x_1, x_2, \ldots, x_N\} \qquad \{(x_j, y_j)\}_{j=1}^{N}$

$Y \quad \{y_1, y_2, \ldots, y_N\}$

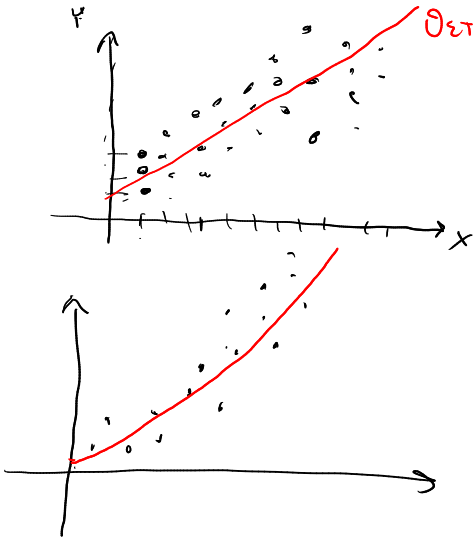
Για τη μελέτη του κατά πόσο δύο μεταβλητές συσχετίζονται, ακολουθούμε τα ακόλουθα βήματα:

- Γραφική αναπαράσταση και υπολογισμός των περιγραφικών μέτρων
- Αναγνώριση προτύπων και μελέτη των αποκλίσεων των τιμών.
- Όταν τα πρότυπα είναι αρκετά ευδιάκριτα, επιλογή κατάλληλου μαθηματικού μοντέλου για τη περιγραφή τους.

Το **διάγραμμα διασποράς** παρουσιάζει τη σχέση μεταξύ των τιμών δύο ποσοτικών μεταβλητών που αναφέρονται στα ίδια στοιχεία. Ο οριζόντιος άξονας εκφράζει τις τιμές της μιας μεταβλητής (συνήθως της ανεξάρτητης μεταβλητής) ενώ ο κάθετος τις τιμές της άλλης μεταβλητής (συνήθως της εξαρτημένης μεταβλητής).
Κάθε ζεύγος τιμών (x, y) για τα στοιχεία του πληθυσμού ή του δείγματος απεικονίζοντε με ένα συμβολο.



Προσθήκη Ποιοτικής μεταβλητής στο διάγραμμα διασποράς



Θετική συσχέτιση

αρνητική συσχέτιση

Θετικά συσχετισμένες μεταβλητές

Όσο μεγαλύτερες τιμές μιας μεταβλητής τείνουν να συνοδεύονται με όλο και μεγαλύτερες τιμές της άλλης μεταβλητής.

Αρνητικά συσχετισμένες μεταβλητές

Όσο μεγαλύτερες τιμές μιας μεταβλητής τείνουν να συνοδεύονται με όλο και μικρότερες τιμές της άλλης μεταβλητής.

- Αν X είναι η ανεξάρτητη μεταβλητή και Y είναι η εξαρτημένη μεταβλητή η συναρτησιακή σχέση των δύο μεταβλητών περιγράφεται μέσω μιας συνάρτησης f στη μορφή $\mathrm{Y} = \mathrm{f}(\mathrm{X})$.
- Για δεδομένη τιμή x της ανεξάρτητης μεταβλητής, η συνάρτηση f δίνει την αντιστοιχη τιμή y της εξαρτημένης μεταβλητής Y.
- Η f δύναται να είναι στοχαστική συνάρτηση. Σε αυτή την περίπτωση ακόμη και για ίδιες τιμές της μεταβλητής X μπορούν να προκύψουν διαφορετικές τιμές για την Y.

$f(x) = x + (3.5 - \eta)$, η - σιποτέλεσμα ρίψης ενός ζαριού

$f(1) = 1 + 3.5 - 4 = 0.5$

$f(1) = 1 + 3.5 - 1 = 3.5$

Παλινδρόμηση

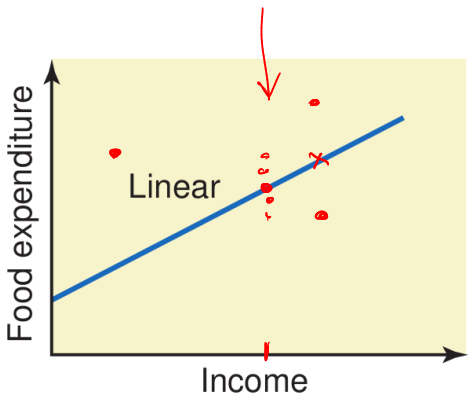
Ένα μοντέλο παλινδρόμησης είναι μια μαθηματική εξίσωση που περιγράφει την σχέση μεταξύ δύο ή περισσότερων μεταβλητών. Το μοντέλο παλινδρόμησης με δύο μεταβλητές, μια ανεξάρτητη και μια εξαρτημένη ονομάζεται **μοντέλο απλής παλινδρόμησης**.

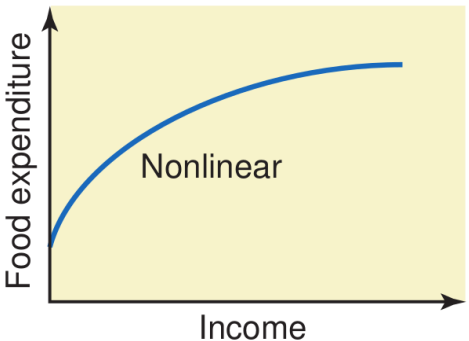
μια ελεύθερη μεταβλητή

Απλή Γραμμική Παλινδρόμηση (Simple Linear Regression)

Ένα μοντέλο παλινδρόμησης το οποίο συνδέει με γραμμικό τρόπο την ανεξάρτητη με την εξαρτημένη μεταβλητή ονομάζεται **μοντέλο απλής γραμμικής παλινδρόμησης**.

ευθεία


Food expenditure
Linear
Income


Food expenditure
Nonlinear
Income

Αιτιοκρατικό μοντέλο

$$y = A + Bx$$

Πιθανοθεωρητικό μοντέλο - Μοντέλο απλής γραμμικής παλινδρόμησης

$$y = A + Bx + \epsilon, \quad \epsilon : \text{όρος τυχαίου σφάλματος}$$

A : σταθερός όρος (constant term), B : κλίση (slope)

$\varepsilon \sim N(0, \sigma_\varepsilon^2)$ $\quad \sigma_\varepsilon^2 = \mathbb{E}[x^2] - (\mathbb{E}[x])^2$

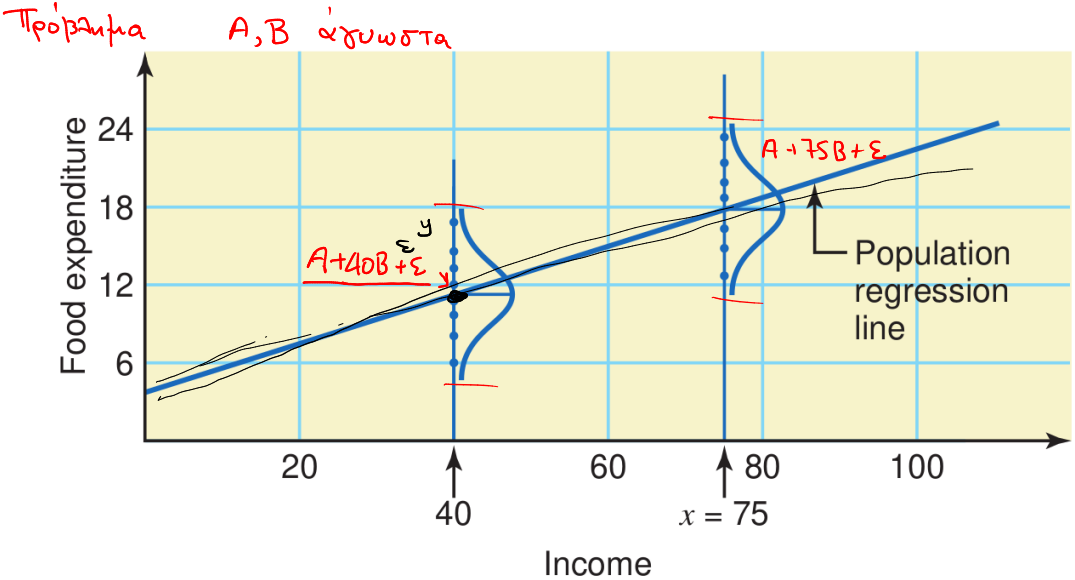
$\frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2}{6}$

Παραδοχές

- Για δοσμένο x το ϵ ακολουθεί ~~τυπική~~ κανονική κατανομή. με μέση τιμή 0
- Τα τυχαία σφάλματα διαφορετικών παρατηρήσεων είναι ανεξάρτητα.
- Για κάθε x οι κατανομές των τυχαίων σφαλμάτων παρουσιάζουν την ίδια τυπική απόκλιση.

Ευθεία παλινδρόμησης για τον πληθυσμό

$$\mu_{y|x} = A + Bx$$


Πρόβλημα
A, B άγνωστα
Food expenditure
24
18
12
6
20
40
60
x = 75
80
100
Income
A+40B+ε
A+75B+ε
Population
regression
line

Δειγματικό μοντέλο απλής γραμμικής παλινδρόμησης

$$\hat{y} = a + bx$$

$\mu_{y|x} = A + Bx$

$\hat{y} = \alpha + bx$

$\alpha \xrightarrow{N \to \infty} A$

$b \to B$

- a είναι δειγματική προσέγγιση του A
- b είναι δειγματική προσέγγιση του B
- $\hat{y}$ είναι η εκτιμώμενη τιμή του y για δοσμένο x

Τυχαίο σφάλμα του δειγματικού μοντέλου απλής γραμμικής παλινδρόμησης

$y = A + Bx + \varepsilon$

$$e = y - \hat{y}$$

$\hat{y} = \alpha + bx$

Έστω το τυχαίο δείγμα

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$$

Για το τυχαίο σφάλμα του δειγματικού μοντέλου απλής γραμμικής παλινδρόμησης έχουμε:

$$e_n = y_n - \hat{y}_n, \quad n = 1, \ldots, N$$

όπου η προσέγγιση του κάθε y_n δίνεται ώς

$$\hat{y}_n = a + bx_n$$

Άθροισμα τετραγωνικών σφαλμάτων

$$\text{SSE} = \sum_{n=1}^{N} e_n^2$$

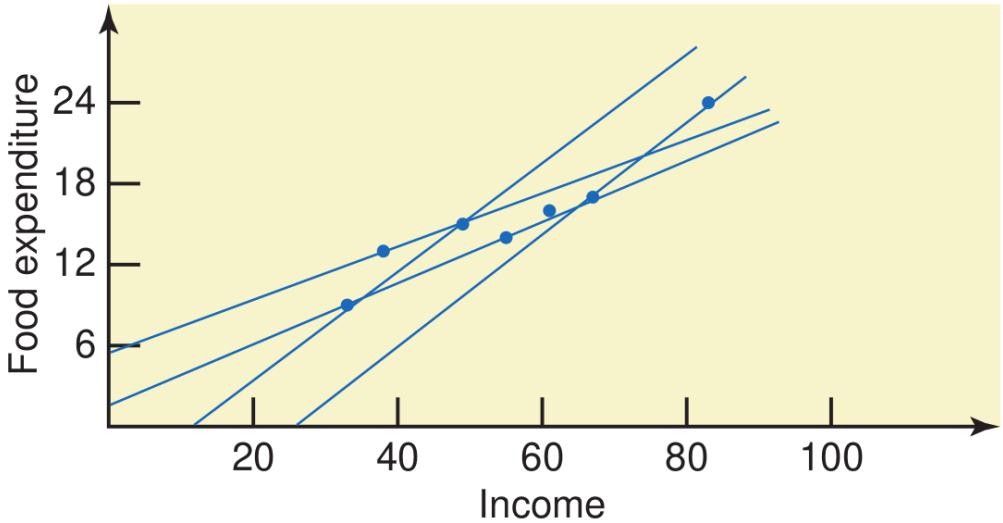
Άθροισμα τετραγωνικών σφαλμάτων συναρτήσει των παραμέτρων του δειγματικού μοντέλου

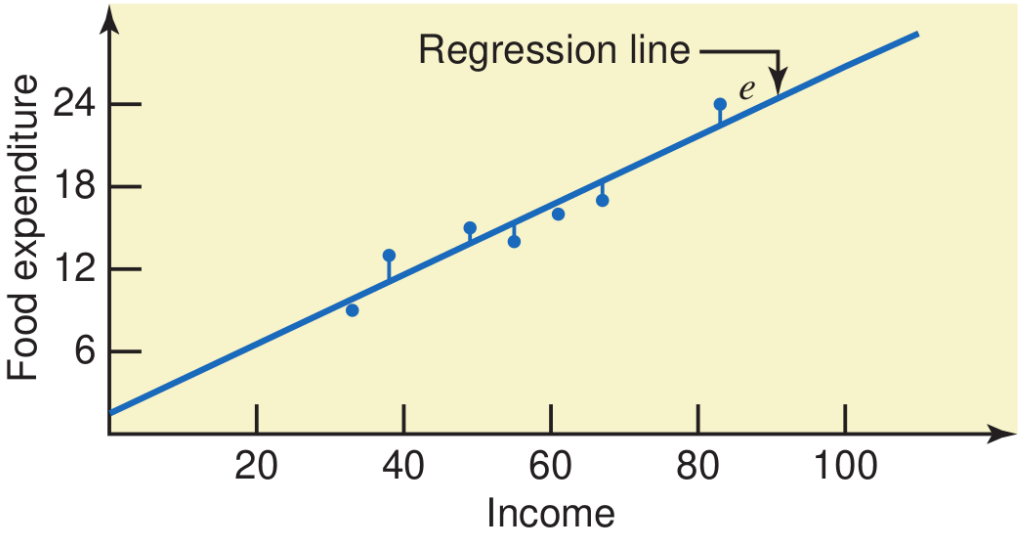
$$\mathcal{Q}(\mathrm{a}, \mathrm{b}) = \mathsf{SSE} = \sum_{\mathrm{n}=1}^{\mathrm{N}} (\mathrm{y_n} - \mathrm{a} - \mathrm{bx_n})^2$$

Εκτίμηση ελαχίστων τετραγώνων

Ως εκτίμησεις των a, b λαμβάνουμε τις τιμές $\mathrm{a}^*, \mathrm{b}^*$ που ελαχιστοποιούν το άθροισμα των τετραγωνικών σφαλμάτων.

$$\mathrm{a}, \mathrm{b} = \arg\min_{\mathrm{a}', \mathrm{b}'} \mathcal{Q}(\mathrm{a}', \mathrm{b}')$$


Food expenditure
24
18
12
6
20
40
60
80
100
Income


Regression line
e
Food expenditure
24
18
12
6
20
40
60
80
100
Income