Προσαρμογή καμπύλης με τη μέθοδο των ελαχίστων τετραγώνων

Διδάσκοντες: Χ. Κυρανούδης, Γ. Μαυρωτάς

Εισαγωγή

- Με βάση κάποιο δείγμα (Χ,Υ) ζητούμε να εξάγουμε συμπεράσματα για τη σχέση μεταξύ εξαρτημένης μεταβλητής (Υ) και της ανεξάρτητης (Χ)
- Εξίσωση παλινδρόμησης
- Π.χ. Γραμμική σχέση: $Y= \alpha X + \beta$
  - Υ: εξαρτημένη μεταβλητή
  - Χ: ανεξάρτητη μεταβλητή
  - α, β: Παράμετροι της εξίσωσης παλινδρόμησης

- Τα δεδομένα δεν ταιριάζουν ποτέ ακριβώς επάνω στα δεδομένα λόγω παραγόντων λάθους (πειραματικά και άλλου τύπου σφάλματα)
- Μέθοδος ελαχίστων τετραγώνων (least squares)
  - Η μέθοδος ελαχίστων τετραγώνων είναι επιδιώκει την ελαχιστοποίηση του σφάλματος μεταξύ του μοντέλου και των δεδομένων.
  - Ορίζουμε σαν συνάρτηση σφάλματος το άθροισμα των τετραγώνων των σφαλμάτων (SSE).

$$SSE = \sum_{i=1}^{n} (Y_i - Y_i^{est})^2$$

Y_i : πραγματική τιμή i-παρατήρησης
Y_i^{est}: εκτιμούμενη τιμή i-παρατήρησης

Μέθοδος ελαχίστων τετραγώνων

- Για γραμμικό μοντέλο ισχύει:

$$SSE = \sum_{i=1}^{n} (Y_i - Y_i^{est})^2 \quad \begin{matrix} \text{γραμμικό μοντέλο} \\ = \end{matrix} \quad \sum_{i=1}^{n} (Y_i - aX_i - \beta)^2$$

- Για να ελαχιστοποιήσουμε μια παράσταση θέτουμε τις μερικές παραγώγους ίσες με μηδέν
- Οι τιμές των παραμέτρων α, β προκύπτουν από τη λύση του συστήματος:

$$\partial SSE/\partial \alpha = 0$$

$$\partial SSE/\partial \beta = 0$$

- Προκύπτουν δύο εξισώσεις (όσες κι οι παράμετροι)

Γραμμικά μοντέλα

Ειδικά για γραμμικά μοντέλα ισχύει:

$$\partial SSE/\partial\alpha = 0 \Rightarrow \quad \partial[\sum_{i=1}^{n}(Y_i - \alpha X_i - \beta)^2]/\partial\alpha = 0 \Rightarrow \quad \sum_{i=1}^{n}(2\alpha X_i^{\,2} - 2X_iY_i + 2\beta X_iY_i) = 0$$

$$\partial SSE/\partial\beta = 0 \Rightarrow \quad \partial[\sum_{i=1}^{n}(Y_i - \alpha X_i - \beta)^2]/\partial\alpha = 0 \Rightarrow \quad \sum_{i=1}^{n}(2\beta - 2Y_i + 2\alpha X_iY_i) = 0$$

- Υπολογισμός των παραμέτρων α, β από το γραμμικό σύστημα που προκύπτει:

$$\sum_{i=1}^{n}(2X_i^{\,2})\alpha + \quad \sum_{i=1}^{n}(2X_iY_i)\beta = \quad \sum_{i=1}^{n}(2X_iY_i)$$

$$\sum_{i=1}^{n}(2X_iY_i)\alpha + \qquad 2n\beta = \qquad \sum_{i=1}^{n}(2Y_i)$$

$$\alpha = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1}^{n}X_iY_i - n\cdot\overline{X}\cdot\overline{Y}}{\sum_{i=1}^{n}X_i^2 - n\cdot\overline{X}^2}$$

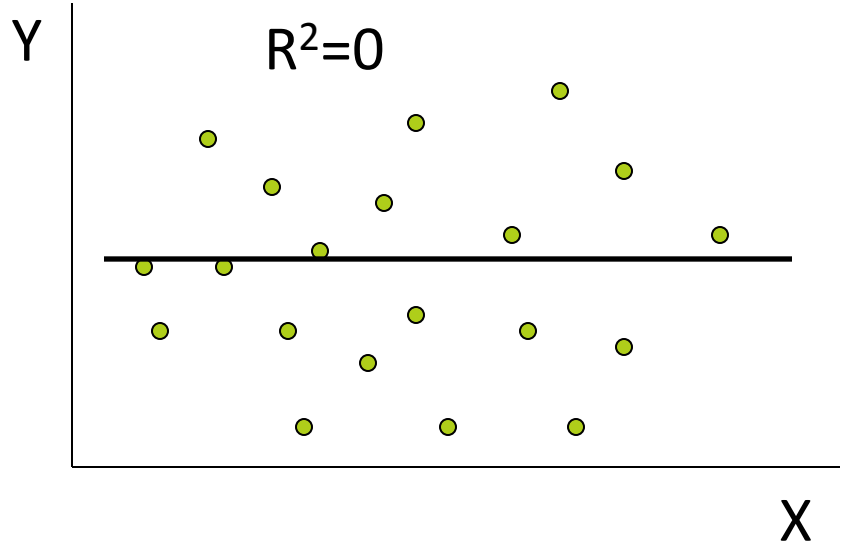
$$\beta = \overline{Y} - \alpha\cdot\overline{X}$$

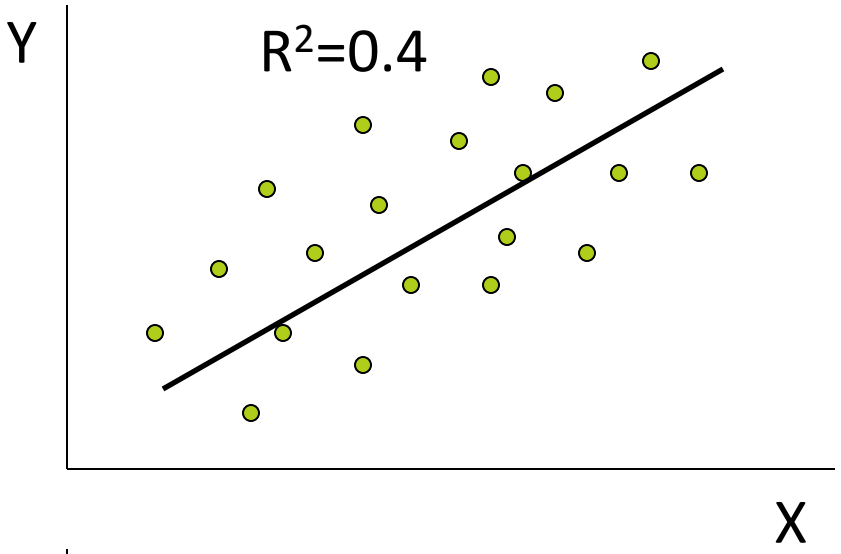
Συντελεστής προσδιορισμού

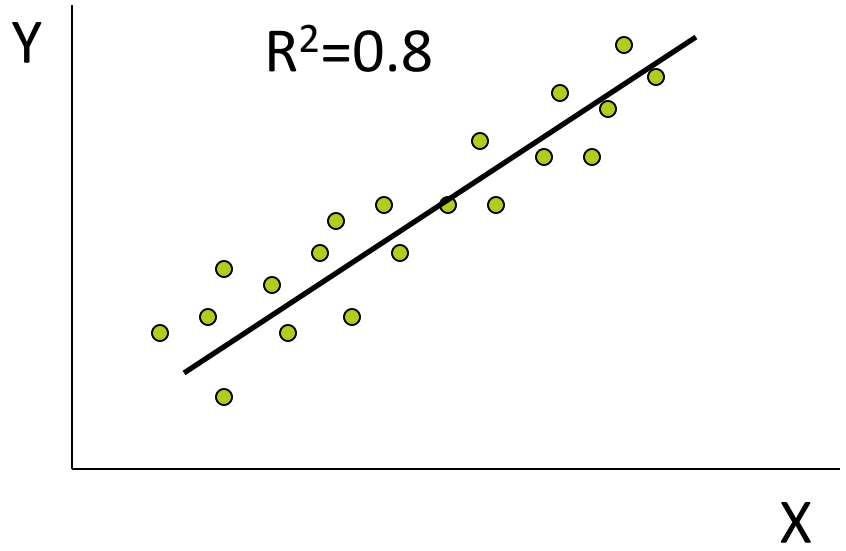
- Ενδεικτικό της καλής προσαρμογής του μοντέλου παλινδρόμησης στα δεδομένα είναι ο συντελεστής προσδιορισμού R^2.

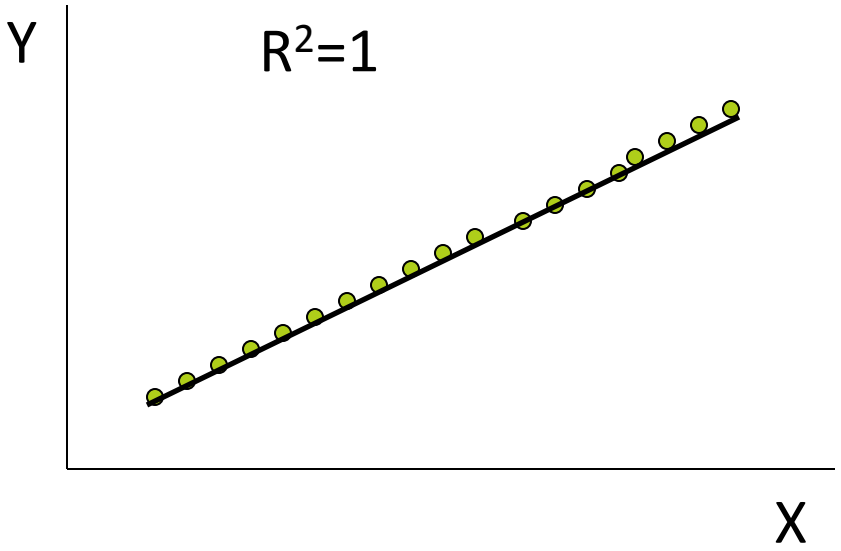
$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n} (Y_i - Y_i^{est})^2}{\sum_{i=1}^{n} (Y_i - Y^{avg})^2}$$

Συντελεστής προσδιορισμού (2)

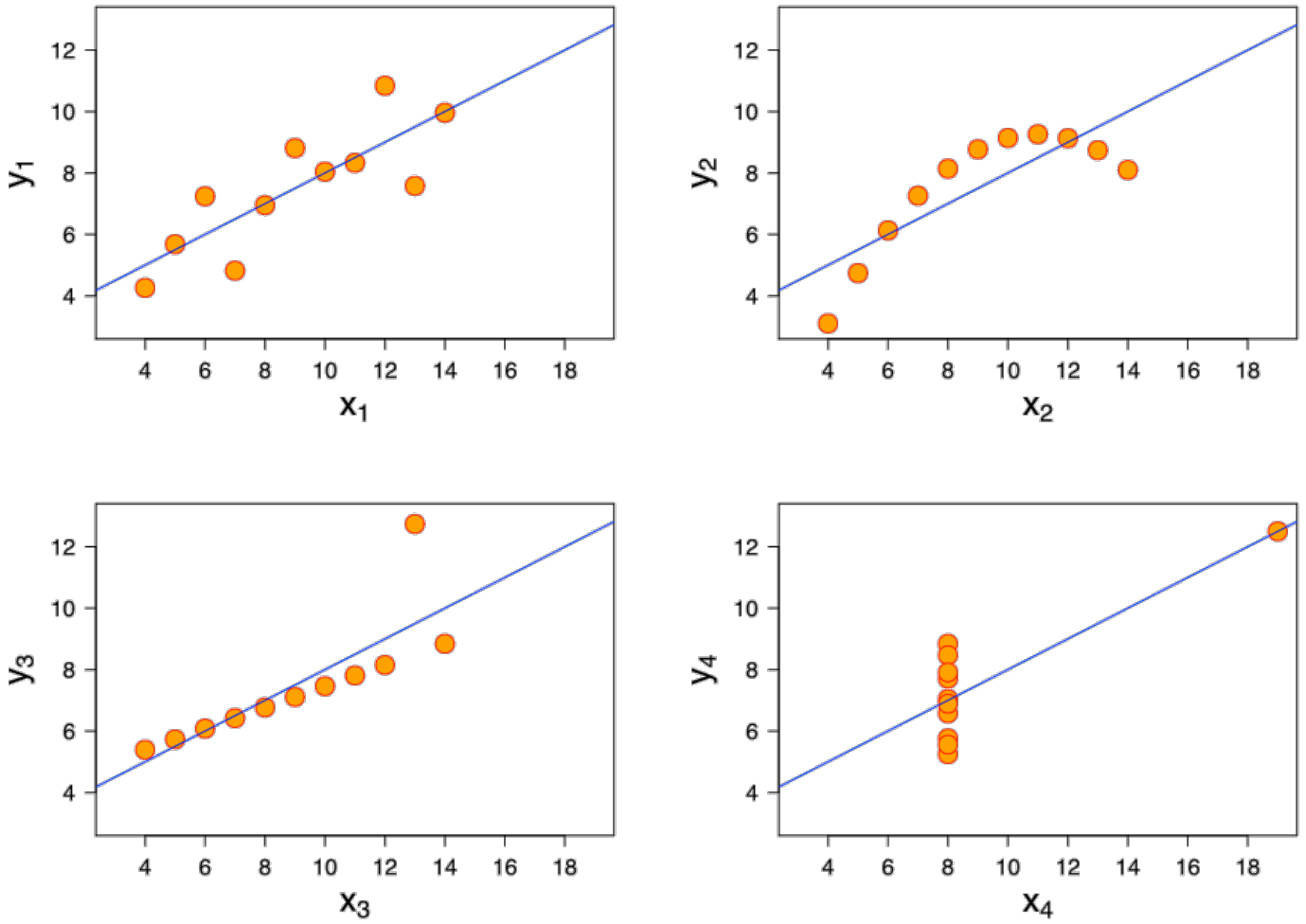






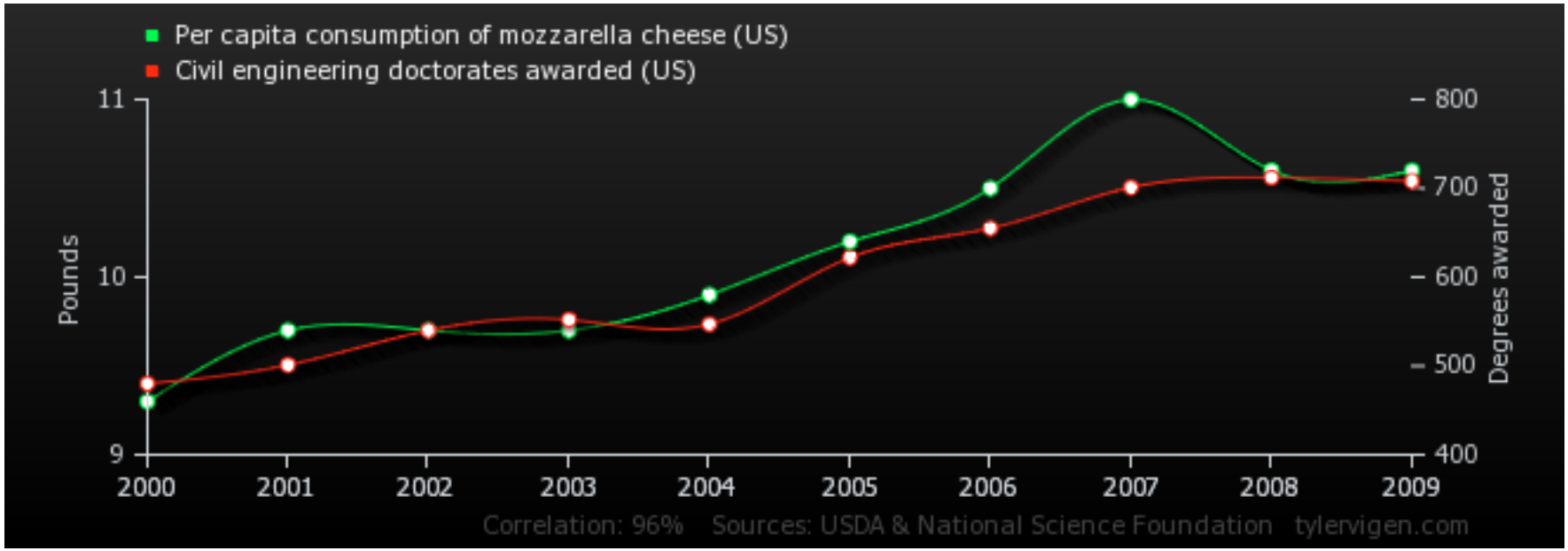


Προσοχή-Κοινή εξίσωση προσέγγισης δεν σημαίνει ίδια δεδομένα



Τα δεδομένα αυτά προτάθηκαν από τον Francis Anscombe για να δείξουν την αξία της βαθύτερης μελέτης των δεδομένων
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Προσοχή - Μαθηματική συσχέτιση δεν σημαίνει πάντα πραγματική σχέση



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Κατά κεφαλή κατανάλωση mozzarella (USA) - Pounds (USDA)	9.3	9.7	9.7	9.7	9.9	10.2	10.5	11	10.6	10.6
Πλήθος νέων διδακτόρων Πολιτικών Μηχανικών που αναγορεύτηκαν (USA) (National Science Foundation)	480	501	540	552	547	622	655	701	712	708

R^2: 0.958648

http://www.tylervigen.com/

Παράδειγμα στο Excel

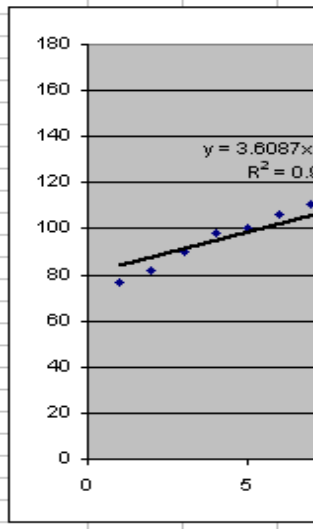
Microsoft Excel - example_simple_regression

File Edit View Insert Format Tools Data Stats Window Help Adobe PDF

Arial 10 B I U 75%

C28

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1															
2		X	Y		X*Y	X^2	Y^2		Yest	e=Y-Yest	(Y-Yest)^2				
3	1986	1	77		77	1	5929		84.427	-7.427	55.159				
4	1987	2	82		164	4	6724		88.036	-6.036	36.428				
5	1988	3	90		270	9	8100		91.644	-1.644	2.704				
6	1989	4	98		392	16	9604		95.253	2.747	7.546				
7	1990	5	100		500	25	10000		98.862	1.138	1.296				
8	1991	6	106		636	36	11236		102.470	3.530	12.458				
9	1992	7	111		777	49	12321		106.079	4.921	24.216				
10	1993	8	108		864	64	11664		109.688	-1.688	2.848				
11	1994	9	110		990	81	12100		113.296	-3.296	10.867				
12	1995	10	120		1200	100	14400		116.905	3.095	9.578				
13	1996	11	130		1430	121	16900		120.514	9.486	89.987				
14	1997	12	130		1560	144	16900		124.123	5.877	34.545				
15	1998	13	117		1521	169	13689		127.731	-10.731	115.159				
16	1999	14	130		1820	196	16900		131.340	-1.340	1.795				
17	2000	15	138		2070	225	19044		134.949	3.051	9.311				
18	2001	16	146		2336	256	21316		138.557	7.443	55.394				
19	2002	17	152		2584	289	23104		142.166	9.834	96.707				
20	2003	18	147		2646	324	21609		145.775	1.225	1.501				
21	2004	19	150		2850	361	22500		149.383	0.617	0.380				
22	2005	20	138		2760	400	19044		152.992	-14.992	224.763				
23	2006	21	147		3087	441	21609		156.601	-9.601	92.175				
24	2007	22	164		3608	484	26896		160.209	3.791	14.368				
25															
26	Σ	253	2691		34142	3795	341589				899.186				
27															
28										Rsq=1-SSE/SST					
29										Rsq=	0.92766				
30		SSxx	885.5												
31		SSyy	12430.8												
32		SSxy	3195.5												
33		Xavg	11.5												
34		Yavg	122.318												
35															
36		b1	3.6087												
37		b0	80.8182												
38															
39															
40															



Βήμα 1

Βήμα 2

Βήμα 3

Μη γραμμικά μοντέλα

- Π.χ. $Y = \alpha + \beta X + \gamma X^2$ ή $$Y = \frac{a}{1 + 3e^{-\beta X}}$$
- Μετασχηματισμό σε αντίστοιχα γραμμικά
  - Π.χ. $Y = \alpha \cdot e^{-\beta X}$ (με λογαριθμοποίηση)
- Αν η γραμμικοποίηση δεν είναι δυνατή τότε αναζητούμε τις τιμές των παραμέτρων α, β που ελαχιστοποιούν το SSE με μια μέθοδο βελτιστοποίησης

Γραμμική παρεμβολή

Για να βρίσκουμε τιμές μεταξύ δύο τιμών πίνακα

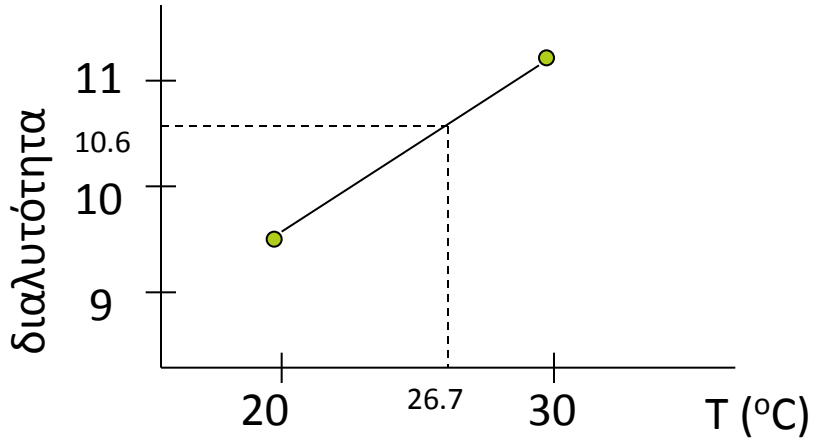
T (°C)	60	50	40	30	20	10
C (mol/lt)	16.4	14.45	12.7	11.1	9.6	8.15

Έστω ότι ψάχνω την διαλυτότητα για T= 26.7°C

Η ευθεία που ορίζεται από τα σημεία Α(20°C, 9.6) και Β(30°C, 11.1) έχει εξίσωση:

$$(x-20)/(30-20) = (y-9.6)/(11.1-9.6)$$

Αντικαθιστώντας όπου x το 26.7 => (26.7-20)/(30-20) = (y-9.6)/(11.1-9.6) => y= 10.6



Επίλυση με Solver Excel

- Δίπλα στη στήλη με τα πραγματικά δεδομένα βάζουμε τα estimated από το μοντέλο (προσοχή: απόλυτη αναφορά με $ στα κελιά των παραμέτρων του μοντέλου)

- Υπολογίζουμε σε διπλανή στήλη το τετράγωνο των διαφορών για κάθε σημείο
- Κάτω από τη στήλη αυτή βάζουμε την παράσταση με το:
  SSE (=sum([actual-estimated]^2) σε ένα κελί

- Τρέχουμε τον Solver με κελί προς ελαχιστοποίηση αυτό που περιέχει το SSE και μεταβαλλόμενα κελιά αυτά που περιέχουν τις παραμέτρους του μοντέλου

- Υπολογίζονται αυτόματα όλες οι τιμές του μοντέλου