5. ΤΟ ΓΕΝΙΚΟ ΓΡΑΜΜΙΚΟ ΜΟΝΤΕΛΟ (GENERAL LINEAR MODEL)

5.1 Εναλλακτικά μοντέλα του απλού γραμμικού μοντέλου: Το εκθετικό μοντέλο

Ένα εναλλακτικό μοντέλο της απλής γραμμικής παλινδρόμησης (που χρησιμοποιήθηκε μέχρι τώρα στις αναλύσεις μας) είναι το **εκθετικό μοντέλο**. Το μοντέλο αυτό έχει την μορφή

$$Y = \alpha(1 + r)^X$$

Μια πιο συνήθες μορφή του εκθετικού μοντέλου είναι η

$$Y = e^{\alpha + \beta X}$$

Το μοντέλο αυτό θεωρείται ότι ανήκει στην κατηγορία των γενικών γραμμικών μοντέλων γιατί, αν πάρουμε τους λογαρίθμους και των δύο μελών του μοντέλου θα έχουμε ότι

$$\ln Y = \alpha + \beta X$$

Έχουμε δηλαδή ότι ο λογάριθμος της μεταβλητής Y συνδέεται γραμμικά με την ανεξάρτητη μεταβλητή X.
Για την εκτίμηση των παραμέτρων α και β χρησιμοποιείται η μέθοδος των ελαχίστων τετραγώνων για την σχέση που αναφέρεται στον λογάριθμο της Y.

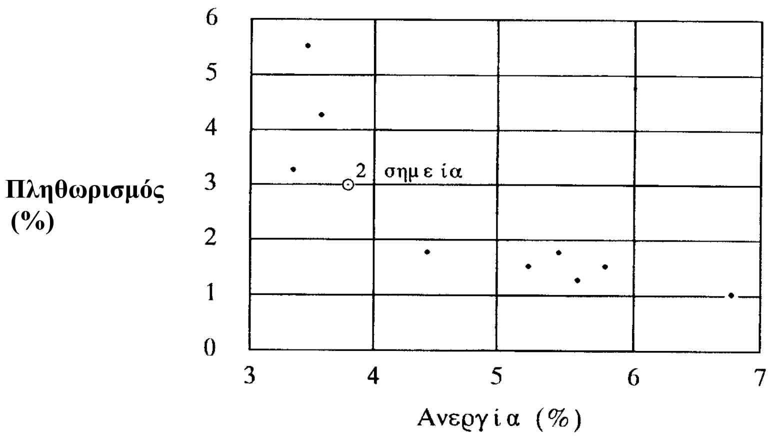
Πρόβλημα: Το 1958 ο Βρετανός οικονομολόγος A.W.H. Phillips δημοσίευσε μια εργασία με την οποία προσπάθησε να προσδιορίσει την σχέση μεταξύ πληθωρισμού και ανεργίας. Την εποχή εκείνη ο προσδιορισμός της σχέσης αυτής εθεωρείτο σαν ένα από τα μεγαλύτερα άλυτα προβλήματα στην μακροοικονομική θεωρία. Η σχέση την οποία προσδιόρισε ο Phillips δεν ήταν βασισμένη στην θεωρία αλλά στην προσαρμογή μιας γραμμής σε παρατηρηθέντα δεδομένα. Ο Phillips προσάρμοσε την καμπύλη γραμμή που χρησιμοποίησε σε βρετανικά δεδομένα για την περίοδο 1861 έως 1913 και στη συνέχεια "επιβεβαίωσε" την μέθοδό του χρησιμοποιώντας την καμπύλη αυτή για τα τότε δεδομένα της

περιόδου 1948 έως 1957 όπου και έδειξε ότι η προηγουμένως προσδιορισθείσα καμπύλη προσαρμόζεται εξίσου ικανοποιητικά και στα δεδομένα αυτά.
Τα αντίστοιχα δεδομένα για την αμερικανική οικονομία της δεκαετίαςτου ’60 δίνονται στον πίνακα που ακολουθεί και στο αντίστοιχο διάγραμμα.

Οπως φαίνεται από τη γραφική παράσταση των σημείων η ευθεία παλινδρόμησης δεν είναι το καταλληλότερο μοντέλο για να εκφράσει τα συγκεκριμένα δεδομένα.

Ρυθμοί πληθωρισμού και ανεργίας για την περίοδο 1960 έως 1969 της Αμερικανικής Οικονομίας

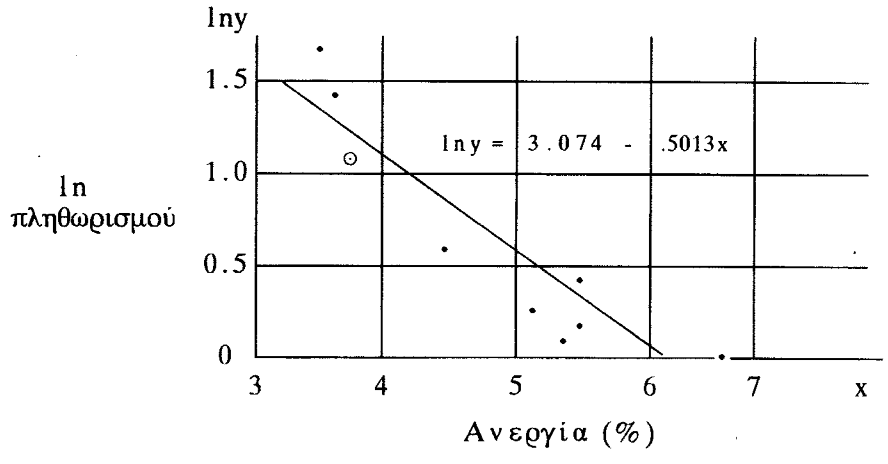
Ετος	Ποσοστό ρυθμού Πληθωρισμού (Y)	Ποσοστό ρυθμού ανεργίας (X)
1960	1.6	5.5
61	1.0	6.7
62	1.1	5.5
63	1.2	5.7
64	1.3	5.2
65	1.7	4.5
66	2.9	3.8
67	2.9	3.8
68	4.2	3.6
69	5.4	3.5



Αν, αντιθέτως, χρησιμοποιήσουμε την μέθοδο των ελαχίστων τετραγώνων γαι να προσαρμόσουμε το εκθετικό μοντέλο (εκφρασμένο μέσω των λογαρίθμων) θα καταλήξουμε στο συμπέρασμα ότι η εκτιμήτρια της εξίσωσης είναι η

$$\ln\hat{y} = 3.0754 - 0.5013x.$$

Η εξίσωση αυτή φαίνεται στο γράφημα που ακολουθεί.



Γραφική παράσταση των ζευγών (x, $\ln\hat{y}$**)**
με βάση τα στοιχεία του πίνακα και
της εξίσωσης $\ln\hat{y} = 3.0754 - 0.5013x.$

Οι υπολογισμοί που απαιτούνται για τον προσδιορισμό της εξίσωσης αυτής δίνονται στον πίνακα που ακολουθεί.

Ετος	ln Y	X	$(\ln Y)^2$	X^2	X(ln Y)
1960	0.4700	5.5	0.2209	30.25	2.5850
1961	0.0000	6.7	0.0000	44.89	0.0000
1962	0.0953	5.5	0.0091	30.25	0.5442
1963	0.1823	5.7	0.0332	32.49	1.0392
1964	0.2624	5.2	0.0688	27.04	1.3643
1965	0.5306	4.5	0.2816	20.25	2.3878
1966	1.0647	3.8	1.1336	14.44	4.0459
1967	1.0647	3.8	1.1336	14.44	4.0459
1968	1.4351	3.6	2.0595	12.96	5.1663
1969	1.6864	3.5	2.8439	12.25	5.9024
Σύνολο	6.7915	47.8	7.7842	239.26	27.0610

$$b = \frac{27.0610 \ - \ 6.7915\,(47.8) \ / \ 10}{239.26 \ - \ (47.8)^2 \ / \ 10} = 0.5013$$

$$a = [6.7915 - (0.5013)(47.8)] /10 = 3.0754$$

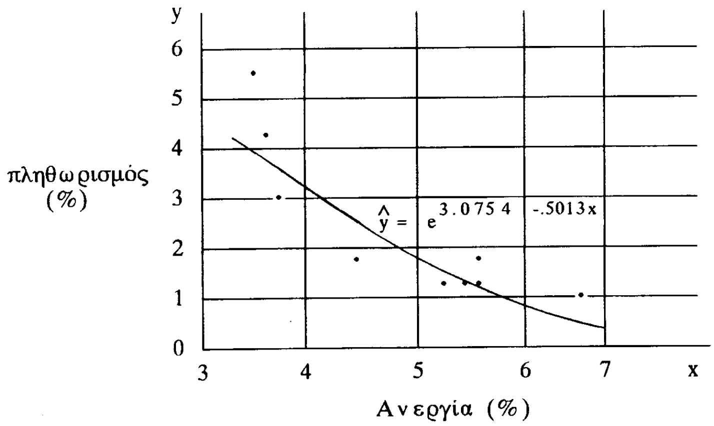
Αν θεωρήσουμε και τα δύο μέλη της εκτιμηθείσας ευθείας παλινδρόμησης ως εκθέτες του e θα έχουμε

$$e^{\ln y} = e^{3.0754 - 0.5013x}$$

ή

$$y = e^{3.0754 - 0.5013x}$$

Η γραφική παράσταση της εξίσωσης αυτής δίνεται στο σχήμα που ακολουθεί.



Γραφική παράσταση των ζευγών (x, y)
με βάση τα στοιχεία του πίνακα
και της εξίσωσης $\hat{y} = e^{3.0754-0.013x}$

Σημείωση: Αν παρατηρήσουμε τα συγκεκριμένα δεδομένα θα δούμε ότι υπάρχει μια μονότονη σχέση αφού ο ρυθμός πληθωρισμού φαίνεται να ελλατώνεται σχεδόν μονότονα, καθώς ο αντίστοιχος ρυθμός ανεργίας αυξάνει. Σύμφωνα επομένως με όσα έχουμε ήδη πει είναι δυνατόν να εφαρμόσουμε την μέθοδο της παλινδρόμησης τάξης μεγέθους (*rank regression*).
Οι σχετικοί υπολογισμοί για την εφαρμογή της μεθόδου αυτής δίνονται στον πίνακα που ακολουθεί.

Έτος	R_Y	R_X	R_Y^2	R_X^2	$R_X R_Y$
1960	5	7.5	25.00	56.25	37.50
1961	1	10	1.00	100.00	10.00
1962	2	7.5	4.00	56.25	15.00
1963	3	9	9.00	81.00	27.00
1964	4	6	16.00	36.00	24.00
1965	6	5.	36.00	25.00	30.00
1966	7.5	3.5	56.25	12.25	26.25
1967	7.5	3.5	56.25	12.25	26.25
1968	9	2	81.00	4.00	18.00
1969	10	1	100.00	1.00	10.00
Σύνολο	55	55	384.50	384.00	224.00

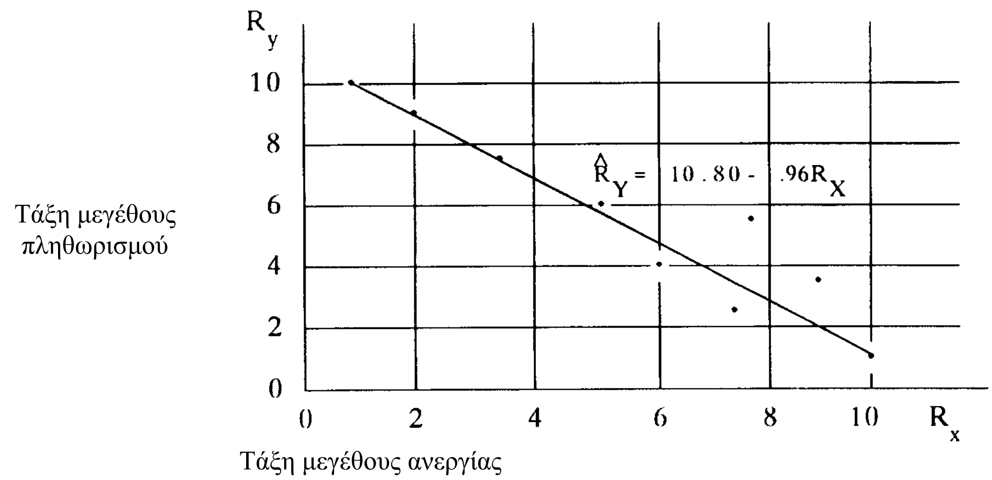
$$b = \frac{224 - 55(55) / 10}{384 - (55)^2 / 10} = -0.9632$$

$$a = [55 - (-0.9632)55]/10 = 10.7976$$

Επομένως, η παλινδρόμηση των τάξεων μεγέθους δίνεται από τη σχέση

$$\hat{R}_Y = 10.7976 - 0.9662R_X$$

Η ευθεία παλινδρόμησης για τις τάξεις μεγέθους φαίνεται στο σχήμα που ακολουθεί:



Γραφική παράσταση της ευθείας παλινδρόμησης για τις τάξεις μεγέθους

Στον πίνακα που ακολουθεί δίνονται οι προβλεπόμενες τιμές $\hat{Y}$ για τις τιμές του x όπως επίσης και τα κατάλοιπα $Y - \hat{Y}$ για την ευθεία παλινδρόμησης, το εκθετικό μοντέλο παλινδρόμησης και την παλινδρόμηση τάξης μεγέθους.

Παρατηρηθείσες τιμές του Y και προβλεπόμενες τιμές $\hat{Y}$ του Y με τη χρησιμοποίηση διαφόρων μοντέλων και το άθροισμα τετραγώνων των λαθών για κάθε μοντέλο

		Ευθεία		Εκθετική		Παλινδρόμηση τάξης μεγέθους	
Έτος	Y(%)	$\hat{Y}$	$Y-\hat{Y}$	$\hat{Y}$	$Y-\hat{Y}$	$\hat{Y}$	$Y-\hat{Y}$
1960	1.6	1.494	0.106	1.375	0.225	1.257	0.343
1961	1.0	0.100	0.900	0.753	0.247	1.017	-0.017
1962	1.1	1.494	-0.394	1.375	-0.275	1.257	-0.157
1963	1.2	1.262	-0.062	1.244	-0.044	1.113	0.087
1964	1.3	1.842	-0.542	1.598	-0.298	1.602	-0.302
1965	1.7	2.655	-0.955	2.269	-0.569	1.698	0.002
1966	2.9	3.468	-0.568	3.223	-0.323	2.841	0.059
1967	2.9	3.468	-0.568	3.223	-0.323	2.841	0.059
1968	4.2	3.700	0.500	3.563	0.637	4.088	0.112
1969	5.4	3.816	1.584	3.747	1.653	5.201	0.199
SSE =	$\Sigma(Y-\hat{Y})^2$		5.590		3.949		0.300

Συγκρίσεις

Από τη σύγκριση των SSE του εκθετικού μοντέλου παλινδρόμησης και του μοντέλου γραμμικής παλινδρόμησης προκύπτει ότι το εκθετικό μοντέλο παλινδρόμησης αποδίδει καλύτερα τα συγκεκριμένα δεδομένα.

Από τα δεδομένα επίσης προκύπτει ότι το συνολικό άθροισμα τετραγώνων για τα Y (SST) είναι,

$$SST = \sum_i \left(Y_i - \overline{Y}\right)^2 = \sum Y_i^2 - \left(\sum Y_i\right)^2 \Big/ n$$

$$= 74.41 - (23.3)^2/10 = 20.121$$

Επομένως, το ποσοστό r^2 της διακύμανσης στα Y που μπορεί να αποδοθεί μέσω του εκθετικού μοντέλου παλινδρόμησης είναι

$$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{3.949}{20.121} = 0.804$$

ενώ αντίστοιχα, για την ευθεία παλινδρόμησης είναι:

$$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{5.590}{20.121} = 0.722$$

Επομένως, από τη χρήση του συντελεστή προσδιορισμού προκύπτει ότι το εκθετικό μοντέλο παλινδρόμησης είναι καλύτερο από ότι η ευθεία παλινδρόμησης.

Σημείωση: Σύγκριση και των τριών μοντέλων με βάση το άθροισμα τετραγώνων των λαθών (SSE) οδηγεί στο συμπέρασμ α ότι η παλινδρόμηση τάξης μεγέθους δίνει την καλύτερη προσαρμογή στα δεδομένα αυτά.

5.2 Άλλα Μοντέλα για Διμεταβλητά Δεδομένα

Τα μοντέλα που έχουμε χρησιμοποιήσει μέχρι τώρα στις αναλύσεις μας, έχουν δύο παραμέτρους α και β. Υπάρχουν όμως και μοντέλα με περισσότερες από δύο παραμέτρους. Τέτοια μοντέλα χρησιμοποιούν, για παράδειγμα, την πολυωνυμική εξίσωση

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots$$

Εκτός τούτου όροι της μορφής 1/x, $\sqrt{x}$, $x^{3/2}$, κ.λ.π. είναι δυνατόν να χρησιμοποιηθούν στο γενικό γραμμικό μοντέλο.

Θα πρέπει να σημειωθεί ότι τα μοντέλα με περισσότερες παραμέτρους δίνουν, εν γένει, καλύτερη προσαρμογή στα δεδομένα εξαιτίας της φύσεως της μεθόδου των ελαχίστων τετραγώνων στην εκτίμηση των παραμέτρων. Παρ’ όλα αυτά θα πρέπει να τονισθεί ότι τα μοντέλα με περισσότερες παραμέτρους που δίνουν καλύτερη προσαρμογή δεν είναι πάντοτε τα καλύτερα που μπορούν να χρησιμοποιηθούν για προβλέψεις. Θα δούμε στη συνέχεια παραδείγματα που αποδεικνύουν την άποψη αυτή.

5.2.1 Προσαρμογή μοντέλων με τρεις ή περισσότερες παραμέτρους

Η αρχή που χρησιμοποιείται στη μέθοδο των ελεχίστων τετραγώνων για δύο παραμέτρους χρησιμοποιείται επίσης στην προσαρμογή μοντέλων με τρεις, τέσσερις ή περισσότερες παραμέτρους. Αυτό σημαίνει ότι οι παράμετροι εκτιμώνται με τρόπο ώστε να ελαχιστοποιείται το άθροισμα των τετραγώνων των λαθών. Είναι φυσικό, ότι όσο περισσότερες είναι οι παράμετροι τόσο πιο πολύπλοκοι γίνονται οι υπολογισμοί. Για το λόγο αυτό για προβλήματα με περισσότερες από δύο παραμέτρους χρησιμοποιούνται στατιστικά πακέτα.

Στον πίνκα που ακολουθεί εμφανίζεται η εκτύπωση από το στατιστικό πρόγραμμα SAS της προσαρμογής με τη μέθοδο των ελαχίστων τετραγώνων του μοντέλου

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2$$

στα δεδομένα πληθωρισμού - ανεργίας που χρησιμοποιήθηκαν προηγουμένως.

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F
MODEL	2	17.5109	8.7555	23.48	.0008
ERROR	7	2.6101	.3729		
TOTAL	9	20.1210			
				r SQUARE .8703	

PARAMETER	ESTIMATE	T FOR H_0 PARAMETER=0	PR>|T|	STD ERROR OF ESTIMATE
INTERCEPT	21.2667	4.41	.0031	4.8226
X	-6.8602	-3.39	.0116	2.0249
X * X	0.5791	2.83	.0255	0.2049

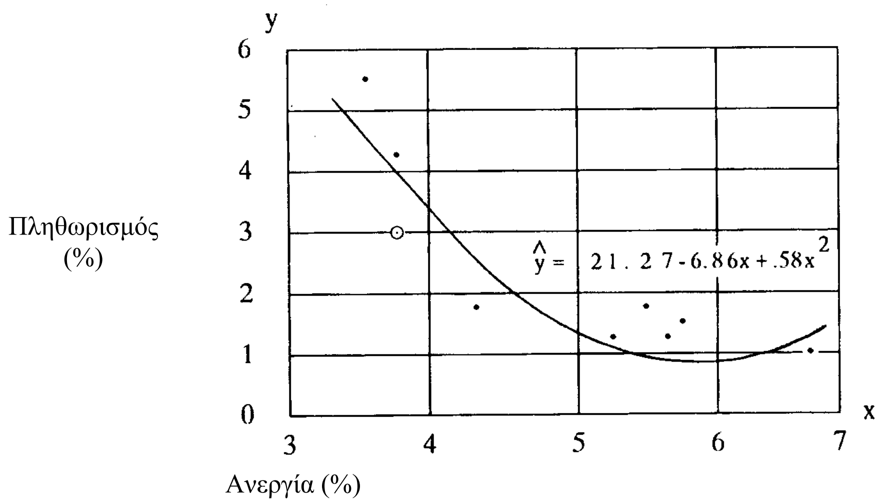
Ο πίνακας αυτός προήλθε από τη χρησιμοποίηση του στατιστικού πακέτου SAS. Οι τιμές των εκτιμητριών $\hat{\beta}_0$, $\hat{\beta}_1$ και $\hat{\beta}_2$ εμφανίζονται στον πίνακα κάτω από την επικεφαλίδα ESTIMATE. Οι τιμές αυτές είναι

$$\hat{\beta}_0 = 21.2667 \qquad \hat{\beta}_1 = -6.8602 \qquad \hat{\beta}_2 = 0.5791$$

Επομένως η τετραγωνική εξίσωση που δίνει την καλύετρη προσαρμογή με τη χρήση της μεθόδου των ελαχίστων τετραγώνων στα δεδομένα πληθωρισμού - ανεργίας είναι η

$$\hat{\mu}_{Y|x} = 21.27 - 6.86x + 0.579x^2$$

Η γραφική παράσταση της εξίσωσης αυτής εμφανίζεται στο σχήμα που ακολουθεί.



Γραφική παράσταση των ζευγών (x y) με βάση τα στοιχεία του πίνακα και της εξίσωσης $\hat{\mu}_{Y|x}$ = 21.27 - 6.86x + 0.579x^2

5.2.2 Ένα άλλο εναλλακτικό μοντέλο

Μέχρι τώρα έχουμε χρησιμοποιήσει αρκετά εναλλακτικά μοντέλα προκειμένου να προσαρμόσουμε τα δεδομένα πληθωρισμού - ανεργίας με τη χρήση περιορισμένου αριθμού παραμέτρων. Συμπληρώνουμε την ενότητα αυτή με την παρουσίαση ενός ακόμα μοντέλου. Η χρησιμοποίηση του μοντέλου αυτού οφείλεται στο γεγονός ότι η γραφική παράσταση της συνάρτησης $Y=\beta_2/X$ είναι μια καμπύλη που ελαττώνεται ως προς Y όσο αυξάνει το X και φαίνεται να ανταποκρίνεται στα συγκεκριμένα δεδομένα. Το μοντέλο αυτό είναι

$$Y = \beta_0 + \beta_1 X + \beta_2/X$$

Ένας πρόσθετος λόγος που παρουσιάζουμε και το μοντέλο αυτό είναι για να δώσουμε έμφαση στο επιχείρημα ότι υπάρχουν πολλά εναλλακτικά μοντέλα που ενδεχομένως δίνουν καλή προσαρμογή σε ένα συγκεκριμένο σύνολο δεδομένων.

Η επιλογή της καμπύλης προσαρμογής είναι μια διαδικασία που συνδυάζει καλή γνώση των διαφόρων συνερτήσεων και των καμπύλων που αντιστοιχούν σε αυτές και της ικανότητας του ερευνητή να εκτιμά ποιά από τις μαθηματικές συναρτήσεις εφαρμόζει καλύτερα στα συγκεκριμένα δεδομένα που πρόκειται να αναλυθούν.

Ένα ερώτημα που τίθεται στις περιπτώσεις εφαρμογών μοντέλων με περισσότερες από δύο παραμέτρους είναι ακτά πόσο κάποια από τις παραμέτρους χρειάζεται. Το ερώτημα αυτό απαντιέται από την εκτιμώμενη τιμή της παραμέτρου και το σχετικό έλεγχο υποθέσεως για το κατά πόσο η παράμετρος αυτή διαφέρει στατιστικά σημαντικά από το μηδέν.

Η εκτύπωση της εφαρμογής του μοντέλου αυτού στα δεδομένα πληθωρισμού - ανεργίας δίνεται στον πίνακα που ακολουθεί:

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR>F
MODEL	2	18.2600	9.1299	34.34	.002
ERROR	7	1.8610	.2659		
TOTAL	9	20.1210			

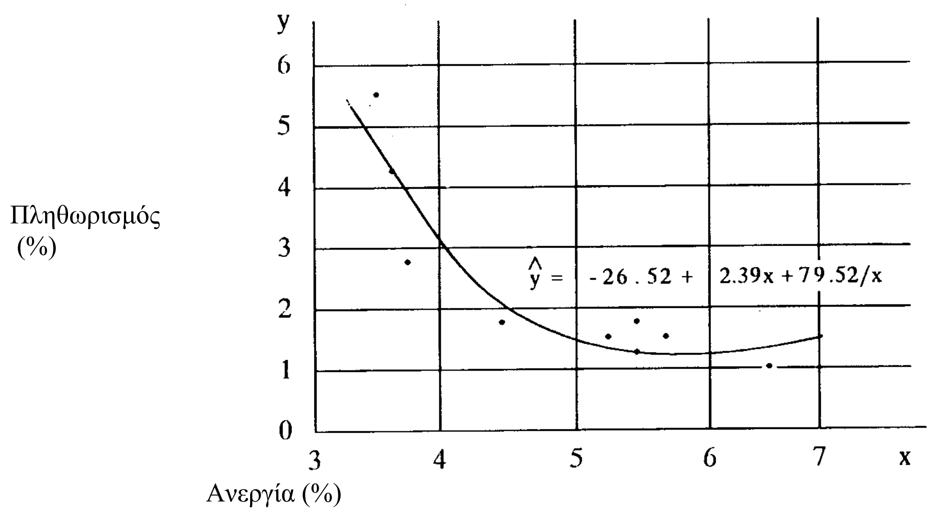
r SQUARE
.9075

PARAMETER	ESTIMATE	T FOR H_0 PARAMETER=0	PR>|T|	STD ERROR OF ESTIMATE
INTERCEPT	-26.5182	-2.88	.0238	9.2187
X	2.38762	2.499	.0419	0.9607
1 / X	79.5166	3.74	.0072	21.2357

Επομένως η εκτιμώμενη καμπύλη είναι βάσει των δεδομένων η

$$\hat{\mu}_{Y|x} = -26.52 + 2.39x + 79.52/x$$

Η γραφική παράσταση της συνάρτησης αυτής δίνεται στο σχήμα που ακολουθεί.



Γραφική παράσταση των ζευγών (x y) με βάση τα στοιχεία του πίνακα και της εξίσωσης $\hat{\mu}_{Y|x}$ = -26.52 + 2.39x + 79.52/x

5.2.3 Παρατηρήσεις και συγκρίσεις

Οι προβλέψεις $\hat{Y}$ για κάθε τιμή του x για τα δύο τελευταία μοντέλα που χρησιμοποιήθηκαν, όπως επίσης τα κατά λοιπα Y - $\hat{Y}$ και τα αθροίσματα τετραγώνων των λαθών (SSE) δίνονται στον πίνακα που ακολουθεί:

Παρατηρηθείσες τιμές του Y, προβλέψεις $\hat{Y}$ του Y, κατάλοιπα και άθροισμα τετραγώνων λαθών του "τετραγωνικού" και "αντίστροφου" μοντέλου παλινδρόμησης για τα δεδομένα πληθωρισμού - ανεργίας

		"Τετραγωνική"		"Αντίστροφη"	
Ετος	(Y%)	$\hat{Y}$	$Y-\hat{Y}$	$\hat{Y}$	$Y-\hat{Y}$
1960	1.6	1.053	0.547	1.071	0.529
1961	1.0	1.299	-0.299	1.347	-0.347
1962	1.1	1.053	0.047	1.071	0.029
1963	1.2	0.979	0.221	1.041	0.159
1964	1.3	1.253	0.047	1.189	0.111
1965	1.7	2.123	-0.423	2.896	-0.196
1966	2.9	3.560	-0.660	3.480	-0.580
1967	2.9	3.560	-0.660	3.480	-0.580
1968	4.2	4.075	0.125	4.165	0.035
1969	5.4	4.350	1.050	4.557	0.843
SSE =	$\Sigma\,(Y-\hat{Y})^2$		2.610		1.861

Όπως παρατηρούμε, τόσο από τον πίνακα αυτό όσο και από τις δύο εκτυπώσεις του προγράμματος SAS, το άθροισμα τετραγώνων των λαθών για τα δύο μοντέλα είναι,

$$SSE = \sum\left(Y_i - \hat{Y}\right)^2 = 2.610 \quad \text{και} \quad SSE = \sum\left(Y_i - \hat{Y}\right)^2 = 1.861$$

αντίστοιχα.

Το συνολικό άθροισμα των τετραγώνων των αποκλίσεων είναι και στις δύο περιπτώσεις

$$SST = \sum\left(Y_i - \overline{Y}\right) = 20.1210$$

Οι τιμές του συντελεστή προσδιορισμού για τα δύο μοντέλα (που προκύπτουν και από τις εκτυπώσεις του SAS) είναι

$$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{2.610}{20.121} = 0.8700$$

$$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{1.861}{20.121} = 0.9075$$

αντίστοιχα.

Σύγκριση του "τετραγωνικού" μοντέλου με αυτό της απλής γραμμικής παλινδρόμησης δείχνει μεταβολές των δύο "κοινών" παραμέτρων οι οποίες οδηγούν σε ελάττωση του αθροίσματος των τετραγώνων των λαθών από 5.590 στο απλό γραμμικό μοντέλο σε 2.610 στο "τετραγωνικό"μοντέλο. Αυτό έχει ως συνέπεια αύξηση της τιμής του συντελεστή προσδιορισμού r^2 από 0.722 στο απλό γραμμικό μοντέλο σε 0.870 στο "τετραγωνικό". Το μοντέλο τετραγωνικής παλινδρόμησης επομένως αυξάνει κατά 14.8% τη διακύμανση των τιμών του Y που εξηγούνται με την χρησιμοποίηση της παλινδρόμησης από το αντίστοιχο που εξηγείται μέσω της χρησιμοποίησης του απλού μοντέλου γραμμικής παλινδρόμησης.

Σύγκριση του μοντέλου τετραγωνικής παλινδρόμησης με το εκθετικό μοντέλο δίνει μία ελαφρά υπεροχή στο πρώτο, αφού ο συντελεστής προσδιορισμού $r^2 = 0.870$ σε σχέση με την τιμή $r^2 = 0.804$ που προσδιορίσαμε στο εκθετικό μοντέλο.

Φυσικά, η τιμή του συντελεστή προσδιορισμού r^2 δεν είναι το μόνο στοιχείο που θα χρησιμοποιήσει ο ερευνητής για να αποφασίσει ποιό μοντέλο θα χρησιμοποιήσει. Ενα στοιχείο που ο ερευνητής θα λάβει σίγουρα υπόψη του είναι το κατά πόσο τα χαρακτηριστικά του μοντέλου ανταποκρίνονται στα χαρακτηριστικά του υπό μελέτη φαινομένου.

Για παράδειγμα, στο εκθετικό μοντέλο όταν ο συντελεστής του X είναι αρνητικός, οι τιμές του Y τεινουν να ελαττώνονται προς τομηδέν όσο το X αυξάνεται. Θα ήταν πράγματι ρεαλιστικό να περιμένουμε ότι ο πληθυσμός Y ελαττώνεται προς το μηδέν όσο να μην του επιτρέπει να παίρνει τιμή μικρότερη από το μηδέν σε στιγμές που η οικονομία βρίσκεται σε κακή κατάσταση, στιγμές οι οποίες συνοδεύονται με υψηλούς ρυθμούς ανεργίας. Αφού το εκθετικό μοντέλο δεν επιτρέπει αρνητικές τιμές δεν θα ήταν ίσως το καταλληλότερο μοντέλο να χρησιμοποιηθεί για το συγκερκριμένο πρόβλημα. Παρόλα αυτά, το εκθετικό μοντέλο θα ήταν ίσως το καταλληλότερο από το αντίστοιχο τετραγωνικό σύμφωνα με το οποίο οι τιμές του Υ ελαττώνονται καθώς ο ρυθμός ανεργίας αυξάνεται μέχρι κάποιο συγκεκριμένο σημείο και στην συνέχεια αρχίζει να αυξάνεται και πάλι! Αυτό βέβαια δεν συμφωνεί με τις γενικές αρχές της Οικονομίας για το πώς συμπεριφέρεται ο πληθωρισμός. Επομένως ο ερευνητής ίσως καταλήξει στο συμπέρασμα ότι και τα δύο μοντέλα είναι ακατάλληλα και ίσως αρχίσει να ερευνά για κάποιο άλλο μοντέλο που θα ανταποκρίνεται καλύτερα στις καθιερωμένες οικονομικές θεωρίες για το πως θα πρέπει να συμπεριφέρεται ένα μοντέλο σε ακραίες τιμές της μεταβλητής Χ. Παρόλα αυτά, θα πρέπει να λεχθεί ότι για το σκοπό της προσαρμογής απλώς ενός μοντέλου στα δεδομένα για το **συγκεκριμένο** πεδίο των παρατηρηθεισών τιμών του Χ τόσο το τετραγωνικό όσο και το εκθετικό μοντέλο εξακολουθούν ίσως να είναι ικανοποιητικά. Αυτό, όμως, με την απαραίτητη διευκρίνηση ότι ο ερευνητής δεν θα χρησιμοποιήσει τα μοντέλα αυτά για να προβλέψει ρυθμούς πληθωρισμού για τιμές της ανεργίας πέρα από το πεδίο τιμών των παρατηρήσεων που χρησιμοποιήθηκαν για την εφαρμογή του μοντέλου.

Τέλος μπορεί να παρατηρήσει κανείς ότι, με βάση τις τιμές των συντελεστών προσδιορισμού r^2 το τελευταίο μοντέλο (που περιέχει τον όρο 1/Χ) παρέχει την καλύτερη προσαρμογή και το μικρότερο άθροισμα τετραγώνων λαθών (SSE) από όλα τα άλλα μοντέλα που χρησιμοποιήθηκαν εκτός από εκείνο της παλινδρόμησης τάξης μεγέθους.

5.2.4 Τελικά Συμπεράσματα

Όπως παρατηρούμε από τα αριθμητικά δεδομένα, η χρησιμοποίηση του πρόσθετου όρου $\beta_2 X^2$ στο μοντέλο της απλής γραμμικής παλινδρόμησης είχε σαν αποτέλεσμα μία αύξηση του r^2 από 0.722 σε 0.870. Η τιμή r^2 μπορεί να αυξηθεί ακόμη περισσότερο αν προσθέσουμε έναν τρίτο όρο $\beta_3 X^3$ στο τετραγωνικό μοντέλο που χρησιμοποιήσαμε. Εισαγωγή όρων υψηλότερης τάξης θα αυξήσει ακόμη περισσότερο την τιμή του r^2. Αυτό οφείλεται στη φύση της μεθόδου των ελαχίστων τετραγώνων για την προσαρμογή εξισώσεων σε δεδομένα. Ο αριθμός των παραμέτρων που είναι δυνατόν να χρησιμοποιηθούν σε ένα μοντέλο περιορίζεται μόνον από τον αριθμό των δεδομένων που είναι διαθέσιμα. Στην πραγματικότητα, είναι δυνατόν να εφαρμόσουμε με μία εξίσωση τέλεια όλα τα σημεία που αντιστοιχούν στο δείγμα μας, καταλήγοντας έτσι σε τιμή του r^2 ίση με το ένα, αν χρησιμοποιήσουμε ένα μοντέλο του οποίου ο αριθμός των παραμέτρων είναι ίσος με τον αριθμό των σημείων που είναι διαθέσιμα.

Μερικοί ερευνητές, που ασχολούνται με την προσαρμογή μοντέλων σε δεδομένα, έχουν την άποψη ότι όσο μεγαλύτερη είναι η τιμή του συντελεστή προσδιορισμού r^2 και όσο περισσότερους όρους περιλαμβάνει ένα μοντέλο τόσο καλύτερο είναι το μοντέλο. Η άποψη αυτή είναι λανθασμένη και είναι δυνατόν να οδηγήσει σε ακατάλληλα μοντέλα, παρά το γεγονός ότι τα μοντέλα αυτά έχουν πολύ μεγάλο συντελεστή προσδιορισμού r^2. Το πρόβλημα με την άποψη αυτή ότι τα μοντέλα τέτοιας μορφής ουσιαστικά παραμορφώνονται για να προσαρμόσουν όσο το δυνατόν καλύτερα τις παρατηρηθείσες τιμές αλλά στην ουσία προσεγγίζουν ανεπαρκώς τον πληθυσμό από τον οποίο προήλθαν τα συγκεκριμένα δεδομένα. Το χαρακτηριστικό αυτό ονομάζεται **υπερβολική προσαρμογή των δεδομένων** *(overfitting)*. Τα μοντέλα αυτά οδηγούν συνήθως σε πολύ ανεπαρκείς προβλέψεις για μελλοντικές παρατηρήσεις.