2. ΧΡΗΣΗ ΣΤΑΤΙΣΤΙΚΩΝ ΠΑΚΕΤΩΝ ΣΤΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η χρησιμοποίηση των τεχνικών της παλινδρόμησης για την επίλυση πρακτικών προβλημάτων έχει διευκολύνει εξαιρετικά από την χρήση διαφόρων στατιστικών πακέτων. Τα πακέτα αυτά αφενός μεν διευκολύνουν την ανάλυση μεγάλου όγκου στοιχείων, αφετέρου δε επιτρέπουν, ως ένα βαθμό, την εξέταση του κατά πόσο οι υποθέσεις που απαιτείται να πληρούνται για την εφαρμογή των μεθόδων της γραμμικής παλινδρόμησης πληρούνται σε κάθε συγκεκριμένο πρόβλημα.

Στην ενότητα αυτή θα αναφερθούμε στις εντολές που απαιτούνται για την εφαρμογή των μεθόδων γραμμικής παλινδρόμησης σε δύο από τα πιο γνωστά στατιστικά πακέτα, το MINITAB και το SAS. (Για μια γενική εισαγωγή στα πακέτα αυτά, ο αναγνώστης παραπέμπεται στο βιβλίο: Ι. Πανάρετος και Ε. Ξεκαλάκη "*Εισαγωγή στη Στατιστική Σκέψη*", Αθήνα 1993).

Προκειμένου να δούμε πώς χρησιμοποιούνται τα δύο στατιστικά πακέτα που προαναφέρθηκαν για την γραμμική παλινδρόμηση, ας αναφέρουμε ένα παράδειγμα.

Παράδειγμα: Ένα μεσιτικό γραφείο ενδιαφέρεται να προβλέψει την τιμή πώλησης μονοκατοικιών. Μετά από προσεκτική μελέτη ο αναλυτής που εργάζεται για το γραφείο καταλήγει στο συμπέρασμα ότι η μεταβλητή που επηρεάζει περισσότερο την τιμή πώλησης των μονοκατοικιών είναι το εμβαδόν του σπιτιού. Για το λόγο αυτό, παίρνει ένα τυχαίο δείγμα από 15 μονοκατοικίες που πωλήθηκαν πρόσφατα και καταγράφει την τιμή πώλησης (σε εκατομμύρια δραχμές) και το εμβαδόν του σπιτιού (σε δεκάδες τ.μ.) Τα δεδομένα που καταγράφηκαν δίνονται στις μεταβλητές C1 και C2 που ακολουθούν. Να γίνει η ανάλυση γραμμικής παλινδρόμησης για το πρόβλημα αυτό με το MINITAB και το SAS.

2.1 Γραμμική Παλινδρόμηση με το MINITAB

Προκειμένου να εισαχθούν τα δεδομένα στις μεταβλητές C1 και C2 στο MINITAB χρησιμοποιούμε την εντολή

```
READ     C1       C2
          89.5     20.0
          79.9     14.8
          83.1     20.5
          56.9     12.5
          66.6     18.0
          82.5     14.3
         126.3     27.5
          79.3     16.5
         119.9     24.3
          87.6     20.2
         112.6     22.0
         120.8     19.0
          78.5     12.3
          74.3     14.0
          74.8     16.7
END
```

Η εντολή που χρειάζεται για να εκτελέσει το MINITAB τους υπολογισμούς είναι η

```
REGRESS C1 1 C2
```

Η εντολή αυτή υπολογίζει την ευθεία παλινδρόμησης ελαχίστων τετραγώνων όπου οι πρατηρηθείσες τιμές για το Y και X έχουν τοποθετηθεί στις στήλες 1 και 2 αντίστοιχα. Ο αριθμός 1 μεταξύ του C1 και του C2 στην βασική εντολή αναφέρεται στη χρησιμοποίηση μιας μόνο ανεξάρτητης μεταβλητής, στη συγκεκριμένη περίπτωση της μεταβλητής οι τιμές της οποίας έχουν τοποθετηθεί στη στήλη 2.

Σε μερικές περιπτώσεις της εντολής REGRESS είναι δυνατόν να προηγείται η εντολή

BRIEF K

όπου K μπορεί να είναι ίσο με 1, 2 ή 3. Όσο αυξάνει η τιμή του K , τόσο περισσότερα στατιστικά αποτελέσματα δίνει ο υπολογιστής. Γενικά, η τιμή K=1 δίνει τιμές για στατιστικές συναρτήσεις που είναι αρκετές για τις περισσότερες εφαρμογές. Εάν η εντολή BRIEF δεν δοθεί, τότε το MINITAB θα προχωρήσει θεωρώντας ότι η εντολή BRIEF 1 έχει δοθεί.

Προκειμένου να ερμηνευθεί ευκολότερα η απάντηση του υπολογιστή, χρησιμοποιούμε ονόματα για τις μεταβλητές

NAME C1 = ΄ y ΄

NAME C2 = ΄ x ΄

Με τον τρόπο αυτό οι σχετικές στατιστικές συναρτήσεις θα παραχθούν με την εντολή

REGRESS ΄ y ΄ 1 ΄x ΄

Το αποτέλεσμα που δίνει ο υπολογιστής στην εντολή αυτή είναι το εξής:

The regression equation is

y = 18.4 + 3.88x

Predictor	Coef	Stdev	t-ratio	p
Constant	18.35	14.81	1.24	0.237
x	3.8786	0.7936	4.89	0.000

s = 13.00 R-sq = 64.8% R-sq(adj) = 62.0%

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	1	403 4.4	4034.4	23.89	0.000

Error	13	219	5.8	168.9		
Total	14	623	0.2			

Unusual Observations

Obs.	x	y	Fit	Stdev.Fit	Residual	St. Resid.
12	19.0	120.80	92.05	3.42	28.75	2.29R

R denotes an obs. with large st. resid.

Ο πίνακας αυτός έχει τα εξής στοιχεία: η πρώτη γραμμή (the regression equation is) δίνει την εκτιμηθείσα με την μέθοδο των ελαχίστων τετραγώνων, ευθεία παλινδρόμησης. Στη συνέχεια, δίνονται στοιχεία για τις παραμέτρους του μοντέλου. Η πρώτη γραμμή δίνει τις επικεφαλίδες: predictor, coef (coefficientσυντελεστής), St.dev (standard deviation = τυπική απόκλιση t-ration (τιμή της στατιστικής συνάρτησης t κάτω από τη μηδενική υπόθεση $\alpha=0$ και $\beta=0$ αντίστοιχα) και p (p-value = παρατηρούμενο επίπεδο σημαντικότητας).

(Για τον ορισμό και τις ιδιότητες του παρατηρούμενου επιπέδου σημαντικότητας, ο αναγνώστης παραπέμπεται στο βιβλίο: Ι. Πανάρετος "*Εκτιμητική-Έλεγχοι Υποθέσεων*", Αθήνα 1993).

Η δεύτερη γραμμή δίνει τις τιμές των στατιστικών συναρτήσεων που εμφανίζονται στην πρώτη γραμμή για την παράμετρο α του μοντέλου, ενώ η τρίτη γραμμή δίνει τις τιμές των στατιστικών αυτών συναρτήσεων για την παράμετρο β του μοντέλου.

Η επόμενη γραμμή περιέχει την τιμή της εκτίμησης για την τυπική απόκλιση του σφάλματος (s), την τιμή του συντελεστή προσδιορισμού (R-sq) και την τιμή του προσαρμοσμένου συντελεστή προσδιορισμού (R-sq(adj)).

Η επόμενη ενότητα του πίνακα αναφέρεται στην ανάλυση της διακύμανσης (analysis of variance). Στην πρώτη γραμμή του υποπίνακα αυτού έχουμε την πηγή (SOURCE) της διακύμανσης, στη συνέχεια έχουμε τους βαθμούς ελευθερίας (DF = degree of freedom), το άθροισμα των τετραγώνων (MS = mean square), τον λόγο F της στατιστικής συνάρτησης ελέγχου και της p-τιμής. Οι επόμενες γραμμές δίνουν τις τιμές των στατιστικών αυτών συναρτήσεων για

την παλινδρόμηση (regression), το σφάλμα (error) και το σύνολο (total).

Η τελευταία ενότητα του πίνακα προσδιορίζει τις ακραίες τιμές (Unusual Observations). Στην επόμενη γραμμή έχουμε τη σειρά της παρατήρησης (Obs. = observation) η οποία είναι ασυνήθιστη, την τιμή χ, της παρατήρησης αυτής, την τιμή y, την τιμή του y που προκύπτει με την χρήση του μοντέλου για την συγκεκριμένη τιμή του χ (Fit = προσαρμογή), την τυπική απόκλιση της κατανομής της τυχαίας αυτής μεταβλητής (Stdev. Fit = standard deviation fit), το κατάλοιπο για την τιμή αυτή (residual) και την τυποποιημένη τιμή του καταλοίπου αυτού (St.resid. = standardized residual).

Η τελευταία γραμμή του πίνακα μας εξηγεί ότι το R που εμφανίζεται δίπλα στην τιμή 2.29 του τυποποιημένου καταλοίπου εμφανίζεται εκεί για να μας ειδοποιήσει ότι η συγκεκριμένη παρατήρηση έχει υπερβολικά μεγάλο τυποποιημένο κατάλοιπο *(R denotes an observation with a large standardized residual).*

Για να κατασκευάσουμε ένα 95% διάστημα εμπιστοσύνης για την αναμενόμενη τιμή του y για δοθέν χ και το 95% διάστημα πρόβλεψης του y για δοθέν x χρειάζεται να χρησιμοποιήσουμε μια εντολή και μια υποεντολή. Η εντολή και η υποεντολή είναι

```
REGRESS ´ y ´ 1 ´ x ´ ;
PREDICT    20.
```

Με τον τρόπο αυτό καθορίζουμε την τιμή x=20 στην υποεντολή PREDICT. Ο υπολογιστής θα δώσει την εξής απάντηση:

Fit	Stdev.Fit	95% C.I.	95% P.I.
95.92	3.66	(88.03 , 103.82)	(66.75 , 125.10)

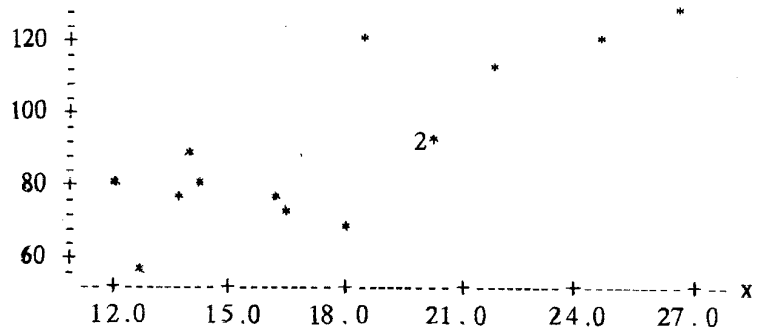
Ο πίνακας αυτός δίνει την προβλεπόμενη τιμή (fit) για την τιμή x=20, την τυπική απόκλιση της προβλεπόμενης αυτής τιμής (Stdev.Fit), το 95% διάστημα εμπιστοσύνης (95% C.I. = 95%,

confidence interval), και το 95% διάστημα πρόβλεψης (95% P.I. prediction interval).

Εάν θέλουμε να έχουμε ένα διάγραμμα σημείων (scattergram) χρησιμοποιούμε την εντολή

PLOT ΄y΄ ΄x΄

Τα σημεία στο διάγραμμα παρουσιάζονται με έναν αστερίσκο. Όπου στο διάγραμμα εμφανίζονται αριθμοί αυτό σημαίνει ότι στο συγκεκριμένο σημείο συμπίπτουν τόσα σημεία του διαγράμματος όσος είναι ο συγκεκριμένος αριθμός.



Αν μας ενδιαφέρει ο συντελεστής συσχέτισης των μεταβλητών x και y, η εντολή που θα πρέπει να χρησιμοποιήσουμε είναι η

CORRELATION ΄y΄ ΄x΄

ή, αν χρησιμοποιήσουμε τις δύο αρχικές μεταβλητές C1 και C2

CORRELATION C1 C2

2.2 Γραμμική Παλινδρόμηση με το SAS

Το στατιστικό πακέτο SAS χρησιμοποιεί την εντολή REG για να πραγματοποιήσει και απλή και πολλαπλή παλινδρόμηση (που θα δούμε στη συνέχεια). Η διαδικαστική αυτή εντολή (procedure statement) ακολουθείται από την εντολή MODEL η οποία καθορίζει την εξαρτημένη μεταβλητή και την ανεξάρτητη μεταβλητή (ή μεταβλητές αν πρόκειται για πολλαπλή παλινδρόμηση). Οι εντολές που προαναφέρθηκαν, βέβαια, ακολουθούν τις εντολές με τις οποίες εισάγονται τα δεδομένα στον υπολογιστή. Έτσι, στο παράδειγμά μας

με τις τιμές πώλησης των μονοκατοικιών χρειάζεται να χρησιμοποιήσουμε τις εξής εντολές για εισαγωγή δεδομένων.

```
DATA;
INPUT Y X;
CARDS;
 89.5  20.0
 79.9  14.8
 83.1  20.5
 56.9  12.5
 66.6  18.0
 82.5  14.3
126.3  27.5
 79.3  16.5
119.9  24.3
 87.6  20.2
112.6  22.0
120.8  19.0
 78.5  12.3
 74.3  14.0
 74.8  16.7
PROC  REG;
MODEL  Y=X;
```

Ο υπολογιστής δίνει τον εξής πίνακα ως απάντηση στις προηγηθείσες εντολές:

Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	4034.4144	4034.4144	23.885	0.0003
Error	13	2195.8215	168.9093		
C Total	14	6230.2360			

Root MSE	12.9965	R-square	0.6476
Dep Mean	88.8400	Adj. R-sq	0.6204
C.V.	14.6291		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H_0 Parameter=0	Prob>|T|
Intercep	1	18.3538	14.8077	1.239	0.2371
X	1	3.8785	0.7936	4.887	0.0003

Η πρώτη ενότητα του πίνακα δίνει στοιχεία αντίστοιχα με αυτά του αντίστοιχου πίνακα του MINITAB με την ισοδύναμη ορολογία. Η επόμενη ενότητα αναφέρεται στο τυπικό σφάλμα των λαθών (Root MSE =square root of mean square error = τετραγωνική ρίζα μέσου τετραγωνικού σφάλματος), στον αριθμητικό μέσο $\bar{y}$ των παρατηρήσεων στην εξαρτημένη μεταβλητή (Dep Mean = dependent mean), στον συντελεστή μεταβλητότητας C.V. (coefficient of variation). (Ο συντελεστής μεταβλητότητας στο SAS ορίζεται ως

$$\text{C.V.} = \left(\frac{S_{\varepsilon}^{*}}{\bar{y}} \right) \text{x} 100$$

Στην περίπτωσή μας, η τιμή αυτή είναι

$$C.V. = \left(\frac{12.995}{88.84} \right) x100 = 14.6291)$$

Η ίδια ενότητα του πίνακα δίνει τις τιμές του συντελεστή προσδιορισμού και του προσαρμοσμένου συντελεστή προσδιορισμού όπως αυτοί έχουν ορισθεί νωρίτερα. Τέλος, η τρίτη ενότητα αναφέρεται σε εκτιμήσεις των παραμέτρων του μοντέλου α (INTERCEP) και β, τις τιμές των αντιστοίχων στατιστικών συναρτήσεων T για τις υποθέσεις H_0: α=0 και H_0: β=0 και τα παρατηρούμενα επίπεδα σημαντικότητας.

Για να κατασκευάσουμε το 95% διάστημα πρόβλεψης για μια συγκεκριμένη τιμή του υ και το 95% διάστημα εμπιστοσύνης για την αναμενόμενη τιμή του Y|x, μπορούμε να χρησιμοποιήσουμε τις εντολές CLI και CLM, αντίστοιχα. Οι προβλέψεις αυτές όμως και οι εκτιμήσεις μπορούν να βασισθούν μόνο στις παρατηρηθείσες τιμές της ανεξάρτητης μεταβλητής x. Μπορούμε όμως να "ξεγελάσουμε" το SAS ώστε να δώσει διαστήματα πρόβλεψης και εμπιστοσύνης για άλλες τιμές του x. Για να το κάνουμε αυτό προσθέτουμε μια επιπλέον παρατήρηση που αναφέρεται στη δεδομένη τιμή του x (x_g) και μια τελεία (.) για το y. Το SAS θεωρεί την τελεία σαν μια χαμένη παρατήρηση (missing observation) και την αγνοεί όταν υπολογίζει την ευθεία παλινδρόμησης με την μέθοδο των ελαχίστων τετραγώνων. Παρ' όλα αυτά όμως κατασκευάζει το διάστημα πρόβλεψης και το διάστημα εμπιστοσύνης για την συγκεκριμένη τιμή του x όταν κάνουμε τις επιλογές CLI και CLM. Έτσι, για παράδειγμα, αν θέλουμε να βρούμε το 95% διάστημα πρόβλεψης για την τιμή μιας μονοκατοικίας 200 τ.μ. και ένα 95% διάστημα εμπιστοσύνης για την μέση τιμή μονοκατοικιών εμβαδού 200 τ.μ., θα προσθέσουμε στις ήδη υπάρχουσες παρατηρήσεις μια 16η παρατήρηση (y = . και x = 20). Δηλαδή

```
. 20.0
```

Επίσης, αλλάζουμε κατάλληλα την εντολή MODEL και την γράφουμε ως εξής:

```
MODEL Y=X/CLI CLM;
```

Η εντολή αυτή παρέχει τις εξής παραπέρα πληροφορίες ως απάντηση του υπολογιστή (πέραν αυτών που είδαμε μέχρι τώρα):

Obs	Dep Var Y	Predict Value	Std Err Predict
1	89.5000	95.9	3.655
2	79.9000	75.7564	4.293
3	83.1000	97.9	3.830
4	56.9000	66.8357	5.615
5	66.6000	88.1677	3.359
6	82.5000	73.8171	4.551
7	126.3	125.0	8.127
8	79.3000	82.3499	3.609
9	119.9	112.6	5.908
10	87.6000	96.7	3.721
11	112.6	103.7	4.526
12	120.8	92.0463	3.419
13	78.5000	66.0600	5.743
14	74.3000	72.6535	4.715
15	74.8000	83.1256	3.554
16		95.9	3.655

Lower 95% Mean	Upper 95% Mean	Lower 95% Predict	Upper 95% Predict
88.0278	103.8	66.7581	125.1
66.4825	85.0302	46.1872	105.3
89.8596	106.1	68.5930	127.1
54.7044	78.9669	36.2498	97.4
80.9121	95.4	59.1681	117.2
63.9857	83.6484	44.0684	103.6
107.5	142.6	91.8993	158.1
74.5534	90.1464	53.2103	111.5
99.8	125.4	81.7607	143.4
88.6613	104.7	67.4950	125.9
93.9044	113.5	73.9510	133.4
84.6595	99.4	63.0136	121.1
53.6521	78.4678	35.3633	96.8
62.4677	82.8393	42.7858	102.5
75.4486	90.8026	54.0177	112.2
88.0278	103.8	66.7581	125.1

Η πρώτη ενότητα αναφέρεται στις παρατηρήσεις (Obs. = observations) για την εξαρτημένη μεταβλητή (Dep.var. = dependent variable), την προβλεπόμενη, μέσω του μοντέλου, τιμή (predict value) και το τυπικό σφάλμα της πρόβλεψης (Std Err Predict = standard error of prediction).

Το δεύτερο μέρος του πίνακα δίνει τα κατώτερα και τα ανώτερα 95% σημεία των διαστημάτων εμπιστοσύνης για την μέση τιμή των μονοκατοικιών (αριστερά) και των διαστημάτων πρόβλεψης για την τιμή μονοκατοικιών (δεξιά) για τις διάφορες επιφάνειες μονοκατοικιών.

Αν θέλαμε να κατασκευάσουμε το διάγραμμα σημείων με το SAS, μετά την εντολή MODEL θα δώσουμε την εντολή

PLOT Y*X;

Στο διάγραμμα που δίνει ο υπολογιστής, τα διάφορα σημεία εμφανίζονται με τους αριθμούς 1 ή 2 ανάλογα με το αν είναι απλά ή διπλά σημεία στο διάγραμμα αυτό.