ΚΕΦΑΛΑΙΟ ΙΙΙ

ΠΟΛΛΑΠΛΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

ΕΝΟΤΗΤΕΣ

1. ΓΕΝΙΚΗ ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΠΟΛΛΑΠΛΗ ΠΑΛΙΝΔΡΟΜΗΣΗ
2. ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ ΜΕ ΤΗ ΜΕΘΟΔΟ ΤΟΥ ΑΠΟΚΛΕΙΣΜΟΥ ΜΕΤΑΒΛΗΤΩΝ
3. ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ ΜΕ ΤΗ ΜΕΘΟΔΟ ΤΗΣ ΠΡΟΟΔΕΥΤΙΚΗΣ ΠΡΟΣΘΗΚΗΣ ΜΕΤΑΒΛΗΤΩΝ
4. ΜΕΘΟΔΟΣ ΤΗΣ ΒΗΜΑΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
5. ΤΟ ΚΡΙΤΗΡΙΟ C_P ΤΟΥ MALLOWS
6. ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕ ΤΗ ΧΡΗΣΗ ΠΙΝΑΚΩΝ
7. ΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΠΟΛΥΣΥΓΓΡΑΜΜΙΚΟΤΗΤΑΣ

1. ΓΕΝΙΚΗ ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΠΟΛΛΑΠΛΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Σε πολλά πρακτικά προβλήματα είναι απαραίτητο να χρησιμοποιήσουμε δύο ή και περισσότερες ανεξάρτητες μεταβλητές προκειμένου να ερμηνεύσουμε με μεγαλύτερη ακρίβεια ένα φυσικό φαινόμενο ώστε να βγάλουμε σωστότερα συμπεράσματα.

Για παράδειγμα, προκειμένου να χρησιμοποιηθεί ένα μοντέλο παλινδρόμησης για να προβλεφθεί η ζήτηση ενός προιόντος μιάς εταιρίας σε 25 διαφορετικές πόλεις είναι ίσως σκόπιμο να χρησιμοποιηθούν κοινωνικοοικονομικές μεταβλητές (μέσο οικογενειακό εισόδημα, μόρφωση του αρχηγού της οικογένειας και μέσος αριθμός χρόνος εκπαίδευσης), δημογραφικές μεταβλητές (μέσο μέγεθος οικογενειών, ποσοστό συνταξιούχων) και περιβαλλοντολογικές μεταβλητές (μέση ημερήσια θερμοκρασία, δείκτης ατμοσφαιρικής ρύπανσης).

Ορισμός: Μοντέλα παλινδρόμησης που περιέχουν δύο ή περισσότερες ανεξάρτητες μεταβλητές ονομάζονται **μοντέλα πολλαπλής παλινδρόμησης** (*multiple regression models*).

1.1 Μοντέλο με Δύο Ανεξάρτητες Μεταβλητές

Το μοντέλο με δυο ανεξάρτητες μεταβλητές είναι η φυσική επέκταση της απλής ευθείας παλινδρόμησης ώστε να μελετώνται δυο ανεξάρτητες μεταβλητές X_1 και X_2 .
Ετσι θα έχουμε

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \quad i= 1,2,...,n \qquad (1)$$

όπου:

- Y_i είναι η τιμή της εξαρτημένης μεταβλητής στην i παρατήρηση.
- x_{i1} και x_{i2} είναι τιμές των ανεξαρτήτων μεταβλητών X_1 και X_2 στην i παρατήρηση οι οποίες υποτίθεται ότι είναι γνωστές σταθερές.
- α, β_1 και β_2 είναι οι παράμετροι του μοντέλου.
- Τα ε_i είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν $N(0, \sigma^2)$ κατανομές.

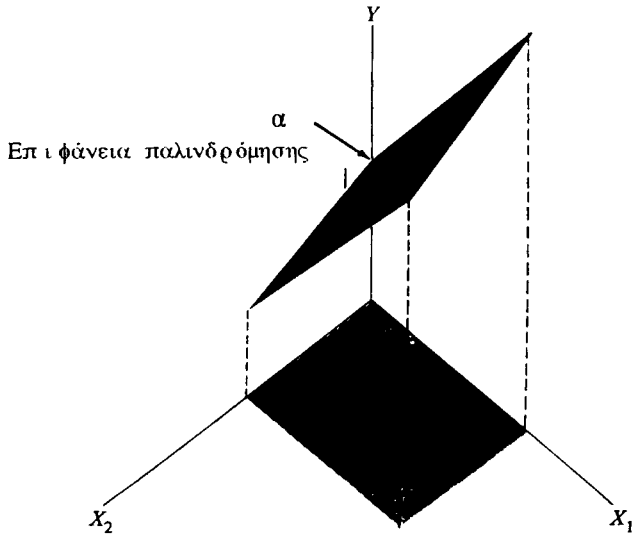
Η **συνάρτηση παλινδρόμησης** (*regression function*) ή αλλιώς **συνάρτηση ανταπόκρισης** (*response function*) του μοντέλου (1) είναι

$$E(Y \mid x_1, x_2) = \alpha + \beta_1 X_1 + \beta_2 X_2$$

Η συνάρτηση αυτή ονομάζεται αρκετές φορές **επιφάνεια παλινδρόμησης** (*regression surface*) ή **επιφάνεια ανταπόκρισης** (*response surface*).

Οι παράμετροι της πολλαπλής παλινδρόμησης έχουν ερμηνείες ανάλογες με αυτές της γραμμικής παλινδρόμησης. Ετσι στην επιφάνεια παλινδρόμησης:

i) Το α αντιστοιχεί στο σημείο τομής του άξονα του Y από την επιφάνεια (επίπεδο) παλινδρόμησης.
ii) Το β_1 δείχνει την μεταβολή της E(Y) όταν το X_1 μεταβάλλεται κατά μια μονάδα ενώ το X_2 παραμένει σταθερό
iii) Το β_2 δείχνει την μεταβολή της E(Y) όταν το X_2 μεταβάλλεται κατά μία μονάδα ενώ το X_1 παραμένει σταθερό.



Η επιφάνεια παλινδρόμησης στην περίπτωση που
τα α & β_1 είναι θετικά και το β_2 αρνητικό

1.2 Μοντέλο με k ανεξάρτητες μεταβλητές

Το μοντέλο παλινδρόμησης με k ανεξάρτητες μεταβλητές $X_1, X_2, ..., X_k$ θα έχει τη μορφή

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i \quad i=1,2,...n$$

όπου

- Y_i είναι η τιμή της εξαρτημένης μεταβλητής για την i παρατήρηση.
- X_{i1} , X_{i2} ,..., X_{ik} είναι οι τιμές των ανεξάρτητων μεταβλητών στην i παρατήρηση (υποτίθενται γνωστές σταθερές).
- Τα ε_i είναι ανεξάρτητα $N(0, \sigma^2)$.

Και στην περίπτωση αυτή α είναι η E(Y) για $X_1 = X_2 = ... = X_k = 0$ ενώ το β_i (i= 1,2,...,k) δείχνει την μεταβολή της E(Y) όταν η μεταβλητή X_i αυξηθεί κατά μια μονάδα ενώ όλες οι άλλες ανεξάρτητες μεταβλητές παραμένουν σταθερές.

1.3 Μοντέλο των Ελαχίστων Τετραγώνων

Η εκτιμήτρια της επιφάνειας παλινδρόμησης

$E(Y \mid x_1\ x_2) \equiv \mu_{Y|x_1,x_2} = \alpha + \beta_1 X_1 + \beta_2 X_2$

θα είναι η επιφάνεια

$$\hat{\mu}_{Y|x_1,x_2} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

Οι τιμές a , b_1 , b_2 των εκτιμητριών $\hat{\alpha}$, $\hat{\beta}_1$ και $\hat{\beta}_2$ προκύπτουν κατά τα γνωστά με την μέθοδο των ελαχίστων τετραγώνων.

Το σύστημα των κανονικών εξισώσεων για τα a, b_1 και b_2 είναι

$$\Sigma y_i = n\,a + (\Sigma x_{i1})b_1 + (\Sigma x_{i2})b_2$$

$$\Sigma x_{i1}\, y_i = (\Sigma x_{i1})a + (\Sigma x_{i1}^2)b_1 + (\Sigma x_{i1}\, x_{i2})b_2$$

$$\Sigma x_{i2}\, y_i = (\Sigma x_{i2})a + (\Sigma x_{i2}^2)b_2 + (\Sigma x_{i2}\, x_{i2})b_2$$

Η επίλυση των κανονικών αυτών εξισώσεων λόγω της πολυπλοκότητας των πράξεων γίνεται, συνήθως, στον υπολογιστή.

Εκτός από την προσαρμογή με την μέθοδο των ελαχίστων τετραγώνων κάποιου μοντέλου σε μιά σειρά από δεδομένα υπάρχει το

πρόβλημα στην πολλαπλή παλινδρόμηση το κατά πόσον μερικοί από τους όρους $\beta_i X_1$ στο μοντέλο έχουν σημαντική συνεισφορά στην εξήγηση της διακύμανσης που παρατηρείται στην εξαρτημένη μεταβλητή Y_i. Η πολλαπλή παλινδρόμηση παρέχει τη στατιστική συμπερασματολογία για τον καθορισμό του κατά πόσον μιά μεταβλητή είναι σημαντική με έλεγχο της μηδενικής υπόθεσης $H_0 : \beta_i = 0$ έναντι της εναλλακτικής $H_1 : \beta_i \neq 0$, i = 1,2,...,k. Αν η Η δεν απορριφθεί για κάποια τιμή του i συμπεραίνουμε ότι δεν υπάρχουν στοιχεία ικανά να μας πείσουν ότι η αντίστοιχη μεταβλητή έχει συνεισφορά σημαντική στο μοντέλο. Στην περίπτωση αυτή ο όρος $\beta_i X_i$ διαγράφεται από το μοντέλο απλοποιώντας έτσι τη διαδικασία.

Σημείωση: Οι έλεγχοι υποθέσεων είναι μέθοδοι που βοηθούν τον ερευνητή να καθορίσει τη σημαντικότητα μεταβλητών του μοντέλου. Θα πρέπει όμως να τονισθεί ότι η απόφαση για το κατά πόσον μιά μεταβλητή θα πρέπει να περιληφθεί στο μοντέλο ή όχι δεν θα πρέπει να ληφθεί με αποκλειστικό κριτήριο τον προηγηθέντα έλεγχο υποθέσεων. Οποιαδήποτε πρόσθετη πληροφορία είναι διαθέσιμη στον ερευνητή η οποία μπορεί να θεωρηθεί από αυτόν περισσότερο πειστική από ότι ο έλεγχος υποθέσεων δεν θα πρέπει να αγνοείται.

Ο έλεγχος της υποθέσεως που προαναφέραμε στηρίζεται στη στατιστική συνάρτηση

$$T = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$$

όπου $\hat{\beta}_1$ είναι η εκτιμήτρια ελαχίστων τετραγώνων του συντελεστή β_i της μεταβλητής $X_ι$ στο γενικό γραμμικό μοντέλο και $S_{\hat{\beta}_i}$ είναι η εκτιμούμενη τυπική απόκλιση της εκτιμήτριας $\hat{\beta}_1$. Οπως συνήθως η τιμή της στατιστικής συνάρτησης T συγκρίνεται με τα ποσοστιαία σημεία της κατανομής t με n-k-1 βαθμούς ελευθερίας. Οι υπολογισμοί αυτοί γίνονται συνήθως στους υπολογιστές.

Επειδή ο αριθμός των ελέγχων υποθέσεων της μορφής που προαναφέρθηκε σε ένα μοντέλο με πολλούς όρους είναι μεγάλος, αν κάνουμε έλεγχο υποθέσεων για κάθε ένα όρο ξεχωριστά η συνολική πιθανότητα λάθους τύπου Ι ίσως καταλήξει να είναι πολύ μεγάλη. Για το

λόγο αυτό μιά καλή στατιστική προσέγγιση είναι πριν γίνουν οι χωριστοί έλεγχοι υποθέσεων για κάθε μιά από τις παραμέτρους να γίνει ένα συνολικό τεστ του μοντέλου και στη συνέχεια να ακολουθήσουν οι ξεχωριστοί έλεγχοι μόνον αν το συνολικό τεστ δώσει στατιστικά σημαντικό αποτέλεσμα.

Παράδειγμα: Μια εταιρεία παροχής βιομηχανικών συμβουλών ανέλαβε να ερευνήσει το βαθμό ικανοποίησης των εργαζομένων Υ σε ένα ερευνητικό εργαστήριο. Στο εργαστήριο αυτό εργάζονται διοικητικοί υπάλληλοι, γραμματείς, βοηθοί εργαστηρίων, επαγγελματικό προσωπικό και διοίκηση. Οι γραμματικές γνώσεις του προσωπικού του εργαστηρίου αυτού ποικίλουν από απολυτήριο γυμνασίου μέχρι μεταπτυχιακούς τίτλους. Η εταιρεία συμβουλών επέλεξε τυχαία πενήντα εργαζομένους και κάθε έναν από αυτούς συνέλεξε πληροφορίες για τις παρακάτω ποσοτικές και ποιοτικές μεταβλητές.

X_1 : Ηλικία
X_2 : Φύλο (0 για γυναίκες και 1 άνδρες)
X_3 : Εβδομαδιαίος μισθός
X_4 : Αριθμός χρόνου απασχόλησης στο εργαστήριο
X_5 : Αριθμός χρόνου προυπηρεσίας σε παρόμοια δουλειά
X_6 : Επίπεδο εκπαίδευσης (σε χρόνια)
X_7 : Ποσοστό του συνολικού εισοδήματος που αντιπροσωπεύει ο μισθός
X_8 : Διοικητική θέση (1 αν υπάρχει και 0 αν δεν υπάρχει)

Εκτός από τα στοιχεία αυτά καθένας από τους εργαζόμενους που επελέγησαν στο δείγμα απαντά σε μιά σειρά από ερωτήσεις που έχουν σχεδιασθεί για να καθορίσουν το επίπεδο ικανοποίησης Υ από το αντικείμενο δουλειάς. Οι απαντήσεις στις ερωτήσεις αυτές χρησιμοποιούνται για να αντιστοιχίσουν μια μέτρηση της ικανοποίησης από τη δουλειά σε μιά κλίμακα από το 1, που αντιστοιχεί σε ελάχιστη ικανοποίηση, μέχρι το 15, που αντιστοιχεί σε πλήρη ικανοποίηση. Τα αποτελέσματα της διαδικασίας αυτής εμφανίζονται στο σχετικό πίνακα.

Στοιχεία για τους 50 εργαζόμενους

Αρ.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Y
1	23	1	35980	5	0	17	93	0	9.9
2	31	0	28420	11	0	14	99	0	7.8
3	64	1	59090	30	5	16	56	1	11.8
4	46	1	34480	12	7	16	94	0	10.4
5	34	1	23980	12	1	15	94	0	9.3
6	39	0	41560	7	2	17	93	0	8.8
7	31	1	18100	4	6	15	100	0	6.1
8	19	0	18510	1	0	15	100	0	7.0
9	33	0	13160	2	1	14	100	0	4.3
10	26	0	15410	2	1	14	92	0	5.1
11	62	1	57260	27	6	16	79	0	12.9
12	18	0	10310	1	0	13	95	0	5.2
13	21	1	13360	4	0	13	99	0	9.6
14	60	1	37770	16	7	16	98	0	8.6
15	26	1	21190	7	0	16	93	0	5.5
16	25	0	30740	1	0	16	83	0	6.2
17	18	0	10450	1	0	12	100	0	3.8
18	40	1	59860	6	2	19	90	1	11.7
19	30	1	10940	4	2	13	100	0	4.9
20	20	1	19770	2	0	15	97	0	11.2
21	61	0	33210	22	0	15	95	0	8.5
22	22	0	19930	2	2	16	92	0	5.4
23	35	1	46620	5	0	17	98	0	10.1
24	18	0	19690	1	0	13	97	0	9.9
25	40	1	38600	9	1	16	92	0	5.6
26	22	0	11370	2	1	12	91	0	5.7
27	43	0	41480	7	1	16	99	0	10.6
28	27	0	27580	2	0	16	100	0	9.3
29	70	0	50550	6	9	16	99	0	10.2
30	37	1	55520	16	0	19	93	1	6.8

Στοιχεία για τους 50 εργαζόμενους

Αρ.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Y
31	32	0	12000	5	3	12	99	0	6.2
32	30	0	19910	9	0	13	94	0	10.2
33	36	1	41750	5	2	18	93	0	6.4
34	69	0	21300	21	2	13	99	0	8.7
35	24	0	12540	2	0	13	99	0	5.1
36	29	0	17220	4	0	13	100	0	8.4
37	36	0	17480	9	0	12	91	0	6.8
38	31	0	19400	5	2	14	99	0	10.1
39	20	1	18480	2	0	15	99	0	8.3
40	46	1	35990	11	8	14	92	0	11.2
41	20	1	17710	3	0	14	97	0	9.0
43	62	0	30660	7	5	18	71	0	8.6
44	37	1	40960	14	0	16	92	0	11.9
45	34	0	15210	4	3	15	99	0	4.3
46	48	1	43930	22	0	16	92	0	10.5
47	34	1	39640	12	0	16	99	0	8.9
48	26	0	10700	5	2	12	93	0	5.6
49	29	1	25700	13	0	13	97	0	13.1
50	43	1	18900	18	2	13	95	0	9.3

Η εταιρεία παροχής συμβουλών ενδιαφέρεται να καθορίσει κατά πόσον υπάρχει κάποια σχέση των μεταβλητών X_1, X_2,...,X_8 και του βαθμού ικανοποίησης από την εργασία Y. Αν μιά τέτοια σχέση υπάρχει ο καθορισμός της θα είναι χρήσιμος γιατί μπορεί να βοηθήσει στην πρόβλεψη του βαθμού ικανοποίησης από την εργασία. Εκτός από αυτό, με τα στοιχεία που έχουν συγκεντρωθεί η εταιρεία παροχής συμβουλών θα είναι σε θέση να καθορίσει ποιές από τις μεταβλητές που προαναφέρθηκαν επηρεάζουν πραγματικά το βαθμό ικανοποίησης των εργαζομένων.

Με την προϋπόθεση ότι ισχύουν οι υποθέσεις που προαναφέρθηκαν, για το μοντέλο

$$E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_8 X_8$$

θα πρέπει, σύμφωνα με τα όσα προαναφέρθηκαν, να ελεγχθεί η μηδενική υπόθεση

$H_0 : \beta_1 = \beta_2 = ... = \beta_8 = 0$

$H_1 : \beta_i \neq 0$ για ένα τουλάχιστον i, i=1,2,...,8.

Τα αποτελέσματα του ελέγχου αυτού εμφανίζονται στον πίνακα που ακολουθεί και έχει προέλθει από εκτύπωση προγράμματος υπολογιστή.

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F
MODEL	8	150.91	18.86	5.39	.0001
ERROR	41	143.57	3.50		
TOTAL	49	294.49			
				r SQUARE 5125	
PARAMETER		ESTIMATE	T FOR H_0 PARAMETER = 0	PR > | T |	STD ERROR OF ESTIMATE
INTERCEPT		12.728	2.28	.0279	5.548
X_1		-6.461 E-02	-1.13	.2653	5.722 E-02
X_2		6.076 E-01	0.80	.4293	7.611 E-01
X_3		1.961 E-04	3.94	.0003	4.979 E-05
X_4		9.191 E-02	1.00	.3244	9.214 E-02
X_5		1.535 E-01	0.86	.3973	1.795 E-01
X_6		-7.678 E-01	-2.48	.0173	3.094 E-01
X_7		2.711 E-02	0.63	.5294	4.274 E-02
X_8		-1.894	-1.34	.1864	

1.4 Ερμηνεία του πίνακα πολλαπλής παλινδρόμησης

Η τιμή F στον παραπάνω πίνακα (F VALUE) μετρά το βαθμό συμφωνίας των δεδομένων με την μηδενική υπόθεση ότι όλοι οι συντελεστές β είναι ίση με το μηδέν. Η τιμή αυτή είναι ο λόγος των δύο μέσων τετραγώνων (MEAN SQUARE) στην στήλη 4 του πίνακα. Δηλαδή,

$$F = \frac{\text{MODEL MEAN SQUARE}}{\text{ERROR MEAN SQUARE}}.$$

Οι βαθμοί ελευθερίας στον αριθμητή και στον παρονομαστή είναι οιβαθμοί ελευθερίας του μοντέλου (MODEL DF) και οι βαθμοί ελευθερίας του σφάλματος (ERROR DF), αντίστοιχα, που εμφανίζονται στη στήλη

2. Το παρατηρούμενο επίπεδο σημαντικότητας (κρίσιμο επίπεδο, p-value), δηλαδή η πιθανότητα να καταλήξουμε σε μια τιμή της στατιστικής

συνάρτησης F τόσο μεγάλη ή μεγαλύτερη από την παρατηρηθείσα τιμή, όταν η H_0 ισχύει, εμφανίζεται στη στήλη 6 (PR > F).

Οι βαθμοί ελευθερίας του μοντέλου (MODEL DF) στη στήλη 2 αναφέρεται πάντοτε στον αριθμό των παραμέτρων β του μοντέλου. Ο συνολικός αριθμός βαθμών ελευθερίας (TOTAL DF) αντιστοιχεί στον αριθμό παρατηρήσεων μείον 1, ενώ οι βαθμοί ελευθερίας του λάθους (ERROR DF) είναι η διαφορά μεταξύ των δύο προαναφερθέντων βαθμών ελευθερίας.

MODEL DF = k
ERROR DF =n-k-1
TOTAL DF =n-1

Τα αθροίσματα των τετραγώνων καθορίζονται με τον ίδιο τρόπο όπως στην απλή γραμμική παλινδρόμηση. Δηλαδή,

$$\text{MODEL SUM OF SQUARES} = SST_r = \sum \left(\hat{Y}_i - \overline{Y}\right)^2$$

$$\text{ERROR SUM OF SQUARES} = SSE = \sum \left(Y_i - \hat{Y}_i\right)^2$$

$$\text{TOTAL SUM OF SQUARES} = SST = \sum \left(Y_i - \overline{Y}\right)^2$$

Προφανώς, και πάλι, το συνολικό άθροισμα τετραγώνων προκύπτει ως το άθροισμα των δύο άλλων αθροισμάτων τετραγώνων. Τα μέσα τετράγωνα (MEAN SQUARE) προκύπτουν ως τα πηλίκα των αθροισμάτων τετραγώνων διηρεμένα με τους αντίστοιχους βαθμούς ελευθερίας. r SQUARE είναι, όπως και προηγουμένως, η τιμή του συντελεστή προσδιορισμού του μοντέλου δηλαδή το ποσοστό της συνολικής διακύμανσης του Y που μπορεί να εξηγηθεί (να αποδοθεί) από το μοντέλο γραμμικής παλινδρόμησης που χρησιμοποιήθηκε.

$$r^2 = \frac{\text{MODEL SS}}{\text{TOTAL SS}} \equiv \frac{SST_r}{SST} = 1 - \frac{\text{ERROR SS}}{\text{TOTAL SS}} \equiv 1 - \frac{SSE}{SST}$$

Η στήλη που έχει επικεφαλίδα ESTIMATE δίνει τις τιμές b_i των στατιστικών συναρτήσεων (εκτιμητριών) $\hat{\beta}_1$ ενώ η στήλη με την επικεφαλίδα STD ERROR OF ESTIMATE δίνει τις τιμές των αντιστοίχων

εκτιμητριών των αντιστοίχων αποκλίσεων $S_{\hat{\beta}_i}$. Η τιμή της στατιστικής συνάρτησης T που αντιστοιχεί στο μοντέλο παλινδρόμησης που χρησιμοποιήθηκε εμφανίζεται στη στήλη που έχει επικεφαλίδα T FOR H_0. Τέλος το παρατηρούμενο επίπεδο σημαντικότητας ή, αλλιώς ελάχιστο επίπεδο σημαντικότητας (P value) που αντιστοιχεί στον έλεγχο H_0: $\beta_i=0$ έναντι της εναλλακτικής H_1: $\beta_i\neq0$ εμφανίζεται στη στήλη που έχει επικεφαλίδα PR > |T|. Επομένως, αν για παράδειγμα, χρησιμοποιήσουμε ως επίπεδο σημαντικότητας την τιμή 0.05 οποιαδήποτε τιμή του παρατηρούμενου επίπεδου σημαντικότητας (P value) μεγαλύτερη από 0.05 θα αποτελεί ένδειξη ότι η αντίστοιχη μεταβλητή είναι υποψήφια για να αποκλεισθεί από το γενικό γραμμικό μοντέλο που χρησιμοποιείται.

Στο παράδειγμα μας οι μεγάλες τιμές των παρατηρουμένων επιπέδων σημαντικότητας (P values) για τους εκτιμούμενους συντελεστές των X_1, X_2, X_4, X_5, X_7, και X_8 αποτελούν ένδειξη ότι οι έξη αυτοί παράγοντες είναι υποψήφιοι για απομάκρυνση από το μοντέλο.

1.5 Ενας μερικός έλεγχος του μοντέλου

Μέχρι τώρα μιλήσαμε για δύο μεθόδους ελέγχου υποθέσεων στην πολλαπλή παλινδρόμηση. Ο ένας αναφερόταν στον έλεγχο των συγκεκριμένων όρων του μοντέλου (όταν ελέγχεται η υπόθεση H_0: $\beta_i=0$ για κάποια συγκεκριμένη τιμή του i) χρησιμοποιώντας τη στατιστική συνάρτηση T και ο άλλος αναφερόταν στη μέθοδο όπου το συνολικό μοντέλο ελέγχεται με τη χρησιμοποίηση της στατιστικής συνάρτησης F όταν ελέγχεται η υπόθεση

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

που περιλαμβάνει όλους τους όρους του γενικού γραμμικού μοντέλου. Υπάρχει και μιά τρίτη στατιστική μεθοδολογία που βρίσκεται στο ενδιάμεσο των δύο προαναφερθεισών μεθοδολογιών. Η μεθοδολογία αυτή επιτρέπει τον ταυτόχρονο έλεγχο ενός αριθμού από τους όρους του μοντέλου χωρίς ταυτόχρονα να απαιτεί να ελεγχθούν όλοι οι όροι του μοντέλου. Η μεθοδολογία αυτή είναι χρήσιμη όταν ο ερευνητής ξέρει ότι κάποιοι από τους όρους πρέπει οπωσδήποτε να χρησιμοποιθούν αλλά είναι αβέβαιος για έναν αριθμό από τους υπόλοιπους όρους του μοντέλου και

θεωρεί ότι χρειάζεται έναν έλεγχο για να αποφασίσει για όλους αυτούς τους υπόλοιπους όρους ταυτόχρονα.

Πιο συγκεκριμένα αν β_1, β_2,..., β_q είναι οι συντελεστές των όρων για τους οποίους είμαστε βέβαιοι ότι θα πρέπει να περιλαμβάνονται στο μοντέλο και β_{q+1}, β_{q+2},..., β_k είναι οι συντελεστές των όρων που θέλουμε να ελεγχθούν ταυτόχρονα, η μηδενική υπόθεση που μας ενδιαφέρει είναι

$$H_0 : \beta_{q+1} = \beta_{q+2} = ... = \beta_k = 0.$$

Με την υπόθεση αυτή ελέγχουμε ουσιαστικά το κατά πόσον το πλήρες μοντέλο

$$E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_q X_q + \beta_{q+1} X_{q+2} + ... + \beta_k X_k$$

είναι κατάλληλο για να περιγράψει τα δεδομένα. Το άθροισμα των τετραγώνων των λαθών του μοντέλου αυτού συμβολίζεται με SSE_1 και οι βαθμοί ελευθερίας για το λάθος συμβολίζονται με DF_1. Το περιορισμένο μοντέλο που μας ενδιαφέρει να εξετάσουμε είναι

$$E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_q X_q \; q < k$$

Το άθροισμα τετραγώνων των λαθών για το περιορισμένο αυτό μοντέλο συμβολίζεται με SSE_2 και οι βαθμοί ελευθερίας του λάθους συμβολίζονται με DF_2.

Η στατιστική συνάρτηση F που χρησιμοποιείται για τον έλεγχο αυτής της υποθέσεως είναι η

$$F = \frac{\left[SSE_2 - SSE_1\right]/\left(DF_2 - DF_1\right)}{SSE_1/DF_1}$$

Η τιμή της στατιστικής αυτής συνάρτησης συγκρίνεται με τα εκατοστιαία σημεία της κατανομής F που δίνονται στους αντίστοιχους πίνακες με (DF_2-DF_1) βαθμούς ελευθερίας του αριθμητή και DF_1 βαθμούς ελευθερίας για τον παρονομαστή.

Παρατήρηση: Είναι ενδιαφέρον να παρατηρήσουμε ότι το άθροισμα τετραγώνων των λαθών SSE_1 για το πλήρες μοντέλο θα είναι πάντοτε μια ποσότητα μικρότερη από το άθροισμα των τετραγώνων των λαθών SSE_2 για το μερικό μοντέλο. Αυτό οφείλεται στο γεγονός ότι το πλήρες μοντέλο περιέχει όλους τους όρους του μερικού μοντέλου μαζί με k-q πρόσθετους

όρους. Επομένως ο αριθμητής της στατιστικής συνάρτησης F μετρά την ελάττωση στο άθροισμα τετραγώνων των λαθών που προκύπτει από τη χρησιμοποίηση k-q πρόσθετων όρων. Η μείωση αυτή του αθροίσματος των τετραγώνων των λαθών διαιρείται με τον αριθμό k-q των προσθέτων όρων. (Ο παρονομαστής της στατιστικής συνάρτησης F παραμένει ο ίδιος όπως προηγουμένως).

Μπορούμε επίσης να παρατηρήσουμε ότι η στατιστική συνάρτηση F που χρησιμοποιήθηκε στην προηγούμενη ενότητα είναι, στην πραγματικότητα, μια ειδική περίπτωση της στατιστικής συνάρτησης F που χρησιμοποιούμε εδώ. Αυτό συμβαίνει γιατί στην πραγματικότητα ο έλεγχος υποθέσεως του πλήρους μοντέλου που αναπτύξαμε στην προηγούμενη ενότητα είναι ουσιαστικά ένας έλεγχος του πλήρους μοντέλου με εναλλακτικό μοντέλο το

$$E(Y) = \alpha.$$

Το μοντέλο αυτό αντιστοιχεί σε μερικό μοντέλο με q=0. Το άθροισμα τετραγώνων των λαθών για το μερικό αυτό μοντέλο συμπίπτει με το συνολικό άθροισμα τετραγώνων SST (αφού δεν υπάρχουν μεταβλητές παλινδρόμησης στο μοντέλο αυτό). Επομένως η στατιστική συνάρτηση F για την περίπτωση αυτή γίνεται

$$F = \frac{(SST - SE)/[(n-1)-(n-k-1)]}{SSE/(n-k-1)} =$$

$$= \frac{SST_r/k}{SSE/(n-k-1)}$$

$$= \frac{\text{MODEL MEAN SQUARE}}{\text{ERROR MEAN SQUARE}} \equiv \left[\frac{MST_r}{MSE}\right].$$

Η τιμή αυτή της F ταυτίζεται με την τιμή της F της προηγουμένης ενότητας.

Παράδειγμα: Εστω ότι στο παράδειγμα μας ο ερευνητής θεωρεί ότι οι μεταβλητές για το φύλο (X_2), μισθό (X_3), επίπεδο εκπαίδευσης (X_6) και κατοχή διευθυντικής θέσης (X_8) είναι μεταβλητές που σχετίζονται πολύ με την ικανοποίηση στο χώρο δουλειάς και επομένως πρέπει να περιληφθούν οποσδήποτε στο μοντέλο.

Επομένως, ο ερευνητής θα ήθελε να ελέγξει τη στατιστική υπόθεση, σε επίπεδο στατιστικής σημαντικότητας a = 0.05

$$H_0 : \beta_1 = 0, \beta_4 = 0, \beta_5 = 0, \beta_7 = 0.$$

Αφού οι μεταβλητές X_1, X_4, X_5 και X_7 δεν περιλαμβάνονται στο μοντέλο κάτω από την μηδενική υπόθεση, το μερικό μοντέλο θα είναι

$$E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 + \beta_8 X_8$$

Παρατηρούμε ότι το μερικό μοντέλο στη περίπτωση μας έχει τέσσερις όρους και επομένως q=4 ενώ k=8 για το πλήρες μοντέλο. Το άθροισμα των τετραγώνων των λαθών SSE_2 υπολογίζεται με τη χρησιμοποίηση του τελευταίου αυτού μερικού μοντέλου από τα δεδομένα μας. Τα αποτελέσματα δίνονται στον πίνακα που ακολουθεί.

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F
MODEL	4	145.99	36.50	11.06	.0001
ERROR	45	148.50	3.30		
TOTAL	49	294.49			
				r SQUARE .4957	
PARAMETER		ESTIMATE	T FOR H_0 PARAMETER= 0	PR > | T |	STD ERROR OF ESTIMATE
INTERCEPT		14.867			
X_2		1.141	1.96	.0558	5.808 E-01
X_3		1.774 E-04	5.32	.0001	3.332 E-05
X_6		-8.088 E-01	-3.28	.0020	2.465 E-01
X_8		-1.665	-1.31	.1974	1.273

Από τον πίνακα αυτό βλέπουμε ότι $SSE_2 = 148.50$ και $DF_2 = 45$.
(Ενώ για το πλήρες μοντέλο είχαμε, από τον προηγούμενο πίνακα, $SSE_1 = 143.57$ και $DF_1 = 41$).

Επομένως η στατιστική μας συνάρτηση F γίνεται

$$F = \frac{\left[SSE_2 - SSE_1\right]/\left(DF_2 - DF_1\right)}{SSE_1/DF_1}$$

$$= \frac{[148.50 - 143.57]/(45 - 41)}{143.57/41}$$

$$= 0.35$$

Από τους πίνακες της F κατανομής προκύπτει ότι η τιμή της κατανομής F με 4 και 41 βαθμούς ελευθερίας σε επίπεδο σημαντικότητας 0.05 είναι $F_{4,41,0.95} = 2.606$. Αυτό μας οδηγεί στο συμπέρασμα ότι δεν υπάρχουν στατιστικές ενδείξεις ώστε οι μεταβλητές X_1, X_4, X_5 και X_7 να πρέπει να συμπεριληφθούν στο μοντέλο. Επομένως, με βάση το συγκεκριμένο δείγμα οδηγούμαστε στο συμπέρασμα ότι θα πρέπει να χρησιμοποιήσουμε το μερικό μοντέλο.

Παρατήρηση: Είναι ενδιαφέρον να παρατηρήσουμε ότι με βάση τις στατιστικές ενδείξεις που είναι διαθέσιμες ένας ερευνητής θα έπρεπε να αμφιβάλλει για το κατά πόσον η μεταβλητή X_8 θα έπρεπε να συμπεριληφθεί στο μοντέλο (αφού όπως προκύπτει από τον πίνακα το παρατηρούμενο επίπεδο σημαντικότητας (P value) που αντιστοιχεί στον εκτιμούμενη συντελεστή της X_8 είναι 0.1974, τιμή σημαντικά υψηλή. Παρ' όλα αυτά, ίσως υπάρχουν άλλοι, μη στατιστικοί, λόγοι για τους οποίους ο ερευνητής θα ήθελε να περιλάβει στο μοντέλο του μιά μεταβλητή που να αναφέρεται στο κατά πόσον ο κάθε υπάλληλος είναι διοικητικός στέλεχος ή όχι.

1.6 Διαστήματα εμπιστοσύνης για τα β

Με τους πίνακες που δίνουν οι υπολογιστές είναι πολύ εύκολο να κατασκευασθούν διαστήματα εμπιστοσύνης για κάθε ένα από τους συντελεστές β_i. Τα διαστήματα εμπιστοσύνης κατασκευάζονται με τον ίδιο ακριβώς τρόπο όπως αυτόν που είδαμε στην απλή γραμμική παλινδρόμηση. Η ερμηνεία επίσης των διαστημάτων εμπιστοσύνης για τα β_i είναι η ίδια όπως σε όλες τις περιπτώσεις των διαστημάτων εμπιστοσύνης. Ετσι το 100(1-a)% διάστημα εμπιστοσύνης για τον συντελεστή β_i της πολλαπλής παλινδρόμησης είναι

$$\hat{\beta}_i \pm t_{n-k-1,1-\alpha/2} S_{\hat{\beta}_i}$$

όπου $\hat{\beta}_1$ είναι η εκτιμήτρια ελαχίστων τετραγώνων του β_i, $t_{n-k-1,\ 1-a/2}$ είναι το (1-α/2) ποσοστιαίο σημείο της κατανομής t με n-k-1 βαθμούς ελευθερίας (ERROR DF στον πίνακα) και $S_{\hat{\beta}_i}$ είναι η εκτιμήτρια της τυπικής απόκλισης του $\hat{\beta}_1$ (STD ERROR OF ESTIMATE στον πίνακα).

Παράδειγμα: Να βρεθεί το 95% διάστημα εμπιστοσύνης στο μερικό μοντέλο του προβλήματος μας όπως αυτό δίνεται στην εξίσωση

$$E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 + \beta_8 X_8$$

Η σημειακή εκτίμηση του συντελεστού β_2, όπως προκύπτει από τον πίνακα, είναι 1.141 και η εκτίμηση της τυπικής απόκλισης για τον συντελεστή αυτό είναι 0.5808. Εξάλλου, από τους πίνακες της κατανομής t προκύπτει ότι $t_{45,0.975} = 2.0141$.
Επομένως, το ζητούμενο διάστημα εμπιστοσύνης είναι

$$\hat{\beta}_2 \pm t_{45,0.975}\ S_{\hat{\beta}_2} = 1.141 \pm 2.0141\ (0.5808)$$

$$= 1.141 \pm 1.170$$
$$= (-0.029\ ,\ 2.311).$$

Σημείωση: Επειδή η μεταβλητή X_2 είναι μιά εικονική μεταβλητή που δηλώνει φύλο, 0 για γυναίκες και 1 για άντρες, το διάστημα εμπιστοσύνης που υπολογίσαμε εκφράζει την εκτιμούμενη ποσότητα κατά την οποία το μέσο σκορ για την ικανοποίηση στο χώρο δουλειάς είναι υψηλότερο στους άντρες απότι στις γυναίκες.

1.7 Συντελεστής Μερικής Συσχέτισης (Partial Correlation Coefficient)

Στην απλή παλινδρόμηση, η σχέση μεταξύ Y και X_1 μπορεί να μετρηθεί από το δειγματικό συντελεστή συσχέτισης r_{YX_1}. Ομοίως, η ένταση της σχέσης της X_1 με το Y, λαμβάνοντας υπόψη την επίδραση της X_2 στην X_1 , μπορεί να συνοψισθεί από το δειγματικό συντελεστή

συσχέτισης των καταλοίπων της παλινδρόμησης του Y στην X_2. Ο τελευταίος αυτός συντελεστής συσχέτισης ονομάζεται *συντελεστής μερικής συσχέτισης (partial correlation coefficient)* και συμβολίζεται με $r_{YX_1|X_2}$. Πολλές φορές χρησιμοποιείται ο όρος μερικός συντελεστής συσχέτισης μεταξύ του Y και του X_1 προσαρμοσμένος για το X_2 *(partial correlation between Y and X_1 adjusted for X_2).*

1.8 Ορθογωνιότητα (Orthogonality)

Δύο μεταβλητές X_1 και X_2 λέγονται ορθογώνιες (orthogonal) αν η παλινδρόμηση της Y επί της X_1, προσαρμοσμένης για την X_2 ταυτίζεται με την παλινδρόμηση της Y επί της X_1 αγνοώντας την X_2. Η ευχάριστη αυτή κατάσταση συμβαίνει όταν ο δειγματικός συντελεστής συσχέτισης των X_1 και X_2 είναι ακριβώς μηδέν. Οταν η X_1 και η X_2 είναι ορθογώνιες, η επίδραση κάθε μιας από τις μεταβλητές είναι σαφώς καθορισμένη. Για το λόγο αυτό, όποτε είναι δυνατό, τα πειράματα σχεδιάζονται με τρόπο ώστε οι μεταβλητές να είναι ορθογώνιες.