4. ΜΕΘΟΔΟΣ ΤΗΣ ΒΗΜΑΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ (STEPWISE REGRESSION)

Η μέθοδος της **βηματικής παλινδρόμησης** (*stepwise regression*) είναι μιά άλλη μέθοδος επιλογής ενός "καλού" υποσυνόλου ανεξαρτήτων μεταβλητών. Η μέθοδος αυτή είναι παρόμοια με την μέθοδο της προοδευτικής προσθήκης μεταβλητών. Η διαφορά των δύο μεθόδων έγκειται στο γεγονός ότι για κάθε διαδοχικό βήμα η υπόθεση

$$H_0 : \beta_j = 0$$

ελέγχεται για **όλες** τις δυνατές ανεξάρτητες μεταβλητές ώστε να αποκλείονται εκείνες για τις οποίες οι τιμές της στατιστικής συνάρτησης $|T_j|$ είναι μικρότερες από ένα προκαθορισμένο κρίσιμο επίπεδο. Η επόμενη μεταβλητή προστίθεται στο υποσύνολο με την ίδια διαδικασία χρησιμοποίησης του κριτηρίου του μεγίστου συντελεστή συσχέτισης όπως στη μέθοδο της προοδευτικής προσθήκης μεταβλητών. Αυτή η βηματική επιλογή συνεχίζεται μέχρις ότου φθάσουμε σε ένα υποσύνολο μεταβλητών για το οποίο καμμιά από τις μεταβλητές που περιέχει το υποσύνολο αυτό δεν έχουν τιμή για τη στατιστική συνάρτηση $|T_j|$ μικρότερη από κάποια συγκεκριμένη κρίσιμη τιμή της μεταβλητής t και δεν υπάρχουν άλλες μεταβλητές που θα πρέπει να αξιολογηθούν για να περιληφθούν στο μοντέλο.

Οπως είναι φανερό η διαδικασία αυτή είναι πολύπλοκη, χρειάζεται πολλούς υπολογισμούς και για το λόγο αυτό γίνεται μόνο με στατιστικά πακέτα. Ολα σχεδόν τα στατιστικά πακέτα που κυκλοφορούν προβλέπουν τη δυνατότητα εφαρμογής της μεθόδου βηματικής παλινδρόμησης.

Η μέθοδος της βηματικής παλινδρόμησης εξηγείται καλύτερα με το παράδειγμα που ακολουθεί.

Παράδειγμα: Το Υπουργείο Υγείας μιάς χώρας ενδιαφέρεται να πραγματοποιήσει μια επιδημιολογική μελέτη για να εξακριβώσει αν είναι δυνατό να προβλεφθούν οι ρυθμοί θνησιμότητας (*mortality rates*) σε διαφορετικές περιοχές της συγκεκριμένης χώρας. Μελέτες της μορφής αυτής δεν περιλαμβάνουν, συνήθως, καλά σχεδιασμένα ή ελεγχομένα πειράματα και τα δεδομένα συνήθως, αποτελούνται από

μετρήσεις σε διάφορες μεταβλητές που αναφέρονται στον καιρό σε κοινωνικοοικονομικούς παράγοντες και τη ρύπανση της ατμόσφαιρας. Ο πίνακας που ακολουθεί παραθέτει μετρήσεις για οκτώ μεταβλητές από ένα τυχαίο δείγμα 30 τοποθεσιών από ολόκληρη την υπό μελέτη χώρα.

Οι μεταβλητές που χρησιμοποιήθηκαν είναι οι εξής:

X_1 : Μέση ετήσια βροχόπτωση (σε ίντσες).
X_2 : Μέση θερμοκρασία Ιανουαρίου (σε βαθμούς Fahrenheit).
X_3 : Μέση θερμοκρασία Ιουλίου (σε βαθμούς Fahrenheit).
X_4 : Διάμεσος του αριθμού των πλήρων χρόνων εκπαίδευσης για άτομα ηλικίας άνω των 25 ετών.
X_5 : Ποσοστό του αστικού πληθυσμού που δεν είναι λευκοί.
X_6 : Σχετικές τιμές διοξειδίου του θείου.
Y : Συνολικός δείκτης θνησιμότητας για κάθε είδους αιτία (εκφρασμένη ως αριθμός θανάτων ανά 100000 κατοίκους).

Είναι φανερό ότι ότι σε προβλήματα της μορφής αυτής ο αριθμός των δυνάμει σημαντικών ανεξαρτήτων μεταβλητών είναι πολύ εύκολο να γίνει ανεξέλεγκτος. Επομένως, το πρόβλημα συνίσταται στη συστηματική επεξεργασία ενός μεγάλου αριθμού μεταβλητών ώστε να καταλήξει κανείς σε ένα χρήσιμο και αξιόπιστο μοντέλο για την πρόβλεψη του ρυθμού θνησιμότητας.

Ρυθμοί θανάτων και 6 ενδεχόμενες ανεξάρτητες μεταβλητές σε 30 τυχαία επιλεγμένες τοποθεσίες

Αρ.Παρατ.	X_1	X_2	X_3	X_4	X_5	X_6	Y
1	32.2	21.2	70.9	11.1	1.9	94.6	942
2	17.6	37.0	69.3	12.3	1.1	13.1	852
3	37.6	36.9	79.3	10.1	10.0	9.8	924
4	31.7	36.4	μεταβλ3.1	11.0	2.0	28.4	928
5	33.0	34.5	73.1	10.2	5.4	243.0	949
6	36.0	32.2	73.2	11.2	0.9	1.4	901
7	39.7	36.7	71.3	10.7	1.7	11.0	938
8	42.8	35.4	78.8	11.3	2.4	18.5	901
9	40.0	25.5	75.9	10.7	4.0	56.4	942
10	40.1	31.2	72.2	9.8	5.6	87.6	1010
11	29.0	32.2	69.8	11.7	1.1	89.7	885
12	42.9	34.5	75.7	10.5	35.1	78.7	1039
13	37.1	41.2	77.2	12.1	15.0	129.0	988
14	54.6	34.5	75.5	9.0	4.0	74.1	981
15	37.3	38.4	74.4	11.3	3.6	54.7	944
16	33.0	36.4	74.8	9.8	22.2	36.3	998
17	33.7	28.7	69.3	11.8	0.8	1.9	892
18	39.3	42.7	76.4	9.9	2.9	3.9	918
19	52.6	38.1	74.1	10.9	24.5	1.9	960
20	28.5	34.1	71.0	11.4	1.9	16.9	908
21	31.8	25.9	73.8	10.8	3.7	21.6	913
22	40.5	51.7	76.8	9.1	29.6	84.7	1002
23	45.1	30.4	81.6	10.3	5.0	2.2	952
24	40.9	45.4	77.7	12.1	30.4	31.5	975
25	35.2	38.8	70.4	11.7	10.7	5.4	956
26	20.2	48.1	74.8	12.0	7.8	239.0	913
27	33.9	31.1	76.0	11.6	23.9	4.3	959
28	48.3	52.5	76.1	11.0	14.9	2.0	901
29	23.9	42.7	75.2	11.5	5.1	3.1	842
30	41.2	30.8	69.9	9.8	4.3	266.0	946

Σημείωση: Υπάρχουν δύο, βασικοί, λόγοι για τους οποίους η μέθοδος αποκλεισμού μεταβλητών δεν είναι, ίσως, κατάλληλη όταν ο αριθμός των υπό εξέταση μεταβλητών είναι μεγάλος. Ο πρώτος είναι ότι όταν ο αριθμός των μεταβλητών ξεπερνά τον αριθμό των παρατηρήσεων η μέθοδος αποκλεισμού μεταβλητών καθίσταται μαθηματικά αδύνατη. Ο δεύτερος είναι ότι, όπως παρατηρήσαμε και προηγουμένως, τόσο με την μέθοδο αποκλεισμού μεταβλητών όσο και με αυτήν της

προοδευτικής προσθήκης μεταβλητών δεν υπάρχει η δυνατότητα επανάκρισης μιάς μεταβλητής που έχει αποκλεισθεί (στη μέθοδο του αποκλεισμού μεταβλητών) ή αποκλεισμού μεταβλητής που έχει ήδη προστεθεί (στη μέθοδο της προοδευτικής προσθήκης μεταβλητών). Η μέθοδος της βηματικής παλινδρόμησης που θα χρησιμοποιήσουμε ξεπερνά τα προβλήματα αυτά.

4.1 Βήμα 1ο στη μέθοδο της βηματικής παλινδρόμησης

Η μέθοδος της βηματικής παλινδρόμησης ξεκινά με ανίχνευση όλων των ανεξαρτήτων μεταβλητών και στόχο τον προσδιορισμό της μεταβλητής εκείνης που έχει τον μεγαλύτερο συντελεστή συσχέτισης με την εξαρτημένη μεταβλητή. Η λίστα αυτή δεν περιλαμβάνει μόνον τις ανεξάρτητες μεταβλητές που θεωρήθηκαν αρχικά αλλά και άλλες μεταβλητές που ενδεχομένως ενδιαφέρουν και οι οποίες προκύπτουν ως συναρτήσεις των αρχικών μεταβλητών (π.χ. X_i^2 ή $X_i X_j$). Ετσι, στο παράδειγμά μας, εκτός από τις μεταβλητές X_1, ..., X_6 θεωρούμε και τους εξής όρους που αναφέρονται σε αλληλεπιδράσεις: $X_1 X_2$, $X_1 X_3$, $X_1 X_6$ και $X_4 X_5$. Οι δύο πρώτοι από τους όρους αυτούς αλληλεπιδράσεων αντιπροσωπεύουν την έλλειψη προσθετικότητας των επιρροών της βροχόπτωσης και της θερμοκρασίας ενώ οι άλλοι δύο άλλα ζευγάρια μεταβλητών που θεωρείται ότι επηρεάζουν τους ρυθμούς θνησιμότητας με μη προσθετικό τρόπο. Ετσι δημιουργήθηκαν δέκα συνολικά μεταβλητές από τις οποίες θα πρέπει να γίνει επιλογή στο πρώτο βήμα της βηματικής επιλογής. Οι απλοί συντελεστές συσχέτισης κάθε μιάς από τις δέκα αυτές μεταβλητές με την εξαρτημένη μεταβλητή Y δίνονται στον πίνακα που δίνεται στη συνέχεια.

Από τον πίνακα αυτό χρησιμοποιούνται οι απόλυτες τιμές των συντελεστών συσχέτισης ώστε να καθορισθεί η μεταβλητή εκείνη που έχει τον υψηλότερο συντελεστή συσχέτισης με την εξαρτημένη μεταβλητή Y. Οπως προκύπτει από τον πίνακα η μεταβλητή X_5 είναι εκείνη που έχει την ιδιότητα αυτή με $r_{X_5,Y} = 0.6445$. Επομένως το μοντέλο στο πρώτο βήμα της βηματικής επιλογής είναι το

$$\hat{\mu}_{Y|x} = \hat{\alpha} + \hat{\beta}_5 X_5$$

Αριθμ.	1	2	3	4	5	6	7	8	9	10
Μεταβλ.	X_1	X_2	X_3	X_4	X_5	X_6	X_1X_2	X_1X_3	X_1X_6	X_4X_5
Συσχέτ.	.5326	.0062	.2754	-.4924	.6445	.2362	.3816	.5258	.6150	.3429

Οπως προκύπτει από τον επόμενο πίνακα (που είναι εκτύπωση του υπολογιστή)

Πρώτο βήμα στη Βηματική Παλινδρόμηση

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR>F
MODEL	1	24497.80	24497.80	19.90	.0001
ERROR	28	34477.17	1231.33		
TOTAL	29	58974.97			

r SQUARE
.4154

PARAMETER	ESTIMATE	T FOR H_0 PARAMETER=0	PR>|T|	STD ERROR OF ESTIMATE
INTERCEPT	911.812			
X_5	2.858	4.46	.0001	0.641

Η εκτιμούμενη εξίσωση παλινδρόμησης που προκύπτει με την μέθοδο των ελαχίστων τετραγώνων είναι η

$$\hat{Y} = 911.812 + 2.858\, X_5 .$$

Ο συντελεστής προσδιορισμού που αντιστοιχεί στην εξίσωση αυτή παλινδρόμησης είναι $r^2 = (0.6445)^2 = 0.4154$. Αυτό σημαίνει ότι το 41.54% της συνολικής διακύμανσης των ρυθμών θνησιμότητας στις διαφορετικές περιοχές μπορεί να εξηγηθεί από το ποσοστό του αστικού πληθυσμού που δεν είναι λευκοί.

Μπορούμε να παρατηρήσουμε ότι στο πρώτο βήμα της διαδικασίας βηματικής παλινδρόμησης έχουμε ένα μοντέλο απλής γραμμικής παλινδρόμησης και επομένως, όπως γνωρίζουμε ήδη, το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που μπορεί να εξηγηθεί με τη χρησιμοποίηση της ανεξάρτητης μεταβλητής καθορίζεται ως το τετράγωνο του απλού συντελεστή συσχέτισης.

Από τον πίνακα επίσης παρατηρούμε ότι το παρατηρούμενο επίπεδο σημαντικότητας (P value) που αντιστοιχεί στην μεταβλητή X_5 είναι 0.0001.

Προκειμένου να προχωρήσουμε στη διαδικασία θα πρέπει η τιμή του παρατηρούμενου επίπεδου σημαντικότητας να είναι μικρότερη από κάποιο καθορισμένο επίπεδο σημαντικότητας για να προχωρήσουμε στο επόμενο βήμα. Στο παράδειγμα μας το επίπεδο σημαντικότητας είχε καθορισθεί στο 0.05.

Στην περίπτωση που το παρατηρούμενο επίπεδο σημαντικότητας υπερβαίνει την προκαθορισθείσα τιμή καμμιά μεταβλητή δεν περιλαμβάνεται στο μοντέλο και επομένως η ανάλυση σταματά στο σημείο αυτό. (Η απόφαση στο σημείο αυτό για το αν δηλαδή θα προχωρήσουμε ή όχι μπορεί, φυσικά, να ληφθεί και μέ βάση την παρατηρούμενη τιμή της στατιστικής συνάρτησης Τ για τη συγκεκριμένη παράμετρο (στην περίπτωσή μας η τιμή της στατιστικής συνάρτησης Τ που αναφέρεται στο $\hat{\beta}_5$ είναι 4.46) και σύγκριση αυτής με την τιμή της κατανομής t που προκύπτει από τους πίνακες για το προκαθορισμένο επίπεδο σημαντικότητας).

4.2 Βήμα 2_0 στη μέθοδο της βηματικής παλινδρόμδησης

Μετά τη συμπλήρωση ενός επιτυχούς πρώτου βήματος (με την έννοια ότι έχει ήδη επιλεγεί κάποια σημαντική μεταβλητή) εξετάζουμε το ενδεχόμενο να υπάρχει ένα μοντέλο κατάλληλο για το συγκεκριμένο πρόβλημα με δύο ανεξάρτητες μεταβλητές. Αυτό επιτυγχάνεται με το να θεωρήσει κανείς όλα τα μοντέλα που περιέχουν τη μεταβλητή X_5, διαδοχικά, με καθεμιά από τις υπόλοιπες μεταβλητές. Στην περίπτωση μας υπάρχουν 9 τέτοια δυνατά μοντέλα με δύο μεταβλητές εκ των οποίων η μία είναι η X_5. Για κάθε ένα από τα μοντέλα αυτά (που περιέχει τη μεταβλητή που είχε επιλεγεί στο πρώτο βήμα και μιά από τις υπόλοιπες μεταβλητές) ελέγχεται η υπόθεση

$$H_0 : \beta_i = 0$$

έναντι της

$$H_1 : \beta_i \neq 0$$

όπου β_i ο συνελεστής της μεταβλητής X_i η οποία, μαζί με την επιλεγείσα στο πρώτο βήμα μεταβλητή, αποτελούν τις δύο μεταβλητές του μοντέλου. Στη συνέχεια εξετάζουμε ποιά από τις μεταβλητές X_i στο αντίστοιχο μοντέλο δύο μεταβλητών που περιλαμβάνει την αρχικώς επιλεγείσα μεταβλητή και την X_i, έχει το μεγαλύτερο παρατηρούμενο επίπεδο σημαντικότητας (P value).

Εάν η τιμή του συγκεκριμένου αυτού παρατηρουμένου επιπέδου σημαντικότητας υπερβαίνει το προκαθορισθέν επίπεδο σημαντικότητας η ανάλυση σταματά στο σημείο αυτό με συμπέρασμα ότι το κατάλληλο μοντέλο είναι αυτό που έχει μιά μόνο μεταβλητή (αυτή που είχε επιλεγεί στο πρώτο βήμα).

Αντίθετα, αν η τιμή του παρατηρούμενου επίπεδου σημαντικότητας για τη συγκεκριμένη μεταβλητή είναι μικρότερη από το προκαθορισθέν επίπεδο σημαντικότητας α θεωρούμε ότι και η μεταβλητή αυτή στην οποία αντιστοιχεί η συγκεκριμένη τιμή του παρατηρούμενου επίπεδου σημαντικότητας πρέπει να περιληφθεί στο μοντέλο, οπότε οδηγούμεθα σε μοντέλο με δύο μεταβλητές.

Η εκτύπωση του προγράμματος για το δεύτερο βήμα στη διαδικασία αυτή για το πρόβλημά μας δίνεται από τον πίνακα που ακολουθεί.

Δεύτερο βήμα στη βηματική παλινδρόμηση

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR>F
MODEL	2	35388.09	17694.04	20.25	.0001
ERROR	27	23586.88	873.59		
TOTAL	29	58974.97			

r SQUARE
.6001

PARAMETER	ESTIMATE	T FOR H_0 PARAMETER=0	PR>|T|	STD ERROR OF ESTIMATE
INTERCEPT	1152.170			
X_4	-21.905	-3.53	.0015	6.204
X_5	2.666	4.91	.0001	0.543

Εχοντας οδηγηθεί στο συμπέρασμα ότι η μεταβλητή X_4, συνδυαζόμενη με τη μεταβλητή X_5, είναι εκείνη που έχει, σε ένα

μοντέλο δύο μεταβλητών, το μικρότερο παρατηρούμενο επίπεδο σημαντικότητας για το αντίστοιχο β (P value για την $\beta_4 = 0.0015$) καταλήγουμε στο εξής μοντέλο δύο μεταβλητών με τη μέθοδο της βηματικής παλινδρόμησης:

$$\hat{\mu}_{Y|x} = \hat{\alpha} + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5$$

Η αντίστοιχη εξίσωση για τα δεδομένα μας που βρίσκεται με τη μέθοδο των ελαχίστων τετραγώνων είναι η

$$\hat{\mu}_{Y|x} = 1152.170 + 21.905 X_4 + 2.666 X_5$$

όπως προκύπτει από τον προαναφερθέντα πίνακα.

Ο συντελεστής προσδιορισμού για το μοντέλο αυτό, όπως επίσης προκύπτει από τον πίνακα, είναι $r^2 = 0.6001$.

Από τον πίνακα επίσης παρατηρούμε ότι τα παρατηρούμενα επίπεδα σημαντικότητας (P values) για τις μεταβλητές X_4 και X_5 είναι 0.0015 και 0.0001 αντίστοιχα. Το παρατηρούμενο επίπεδο σημαντικότητας για την μεταβλητή X_5 δεν φαίνεται να έχει αλλάξει στην περίπτωσή μας. Εν γένει όμως, το παρατηρούμενο επίπεδο σημαντικότητας για την μεταβλητή η οποία παραμένει στη βηματική παλινδρόμηση είναι ενδεχόμενο να αλλάξει κάθε φορά που κάποια μεταβλητή προστίθεται, ή αφαιρείται, από το μοντέλο.

Σημείωση: Πρέπει να τονίσουμε ότι στη διαδικασία της βηματικής παλινδρόμησης από το δεύτερο βήμα και μετά, προκειμένου να εξετάσουμε την επόμενη μεταβλητή που θα περιληφθεί ενδεχομένως στο μοντέλο, αντί για την υιοθεσία της μεθόδου του ελάχιστου παρατηρούμενου επίπεδου σημαντικότητας που χρησιμοποιήσαμε προηγουμένως, είναι δυνατόν να χρησιμοποιηθεί η διαδικασία των μερικών συντελεστών συσχέτισης. Να θεωρήσουμε, δηλαδή, από τις υπόλοιπες προς εξέταση μεταβλητές τη μεταβλητή εκείνη που έχει τον μεγαλύτερο μερικό συντελεστή συσχέτισης με την εξαρτημένη μεταβλητή. Το αποτέλεσμα που θα προκύψει θα είναι το ίδιο ανεξάρτητα από τη λογική που θα υιοθετηθεί. Διαφορετικά στατιστικά πακέτα χρησιμοποιούν ή την μία ή την άλλη μέθοδο στη διαδικασία της βηματικής παλινδρόμησης.

4.3 Βήμα 3ο στη μέθοδο της βηματικής παλινδρόμησης

Μετά την ολοκλήρωση ενός επιτυχούς δευτέρου βήματος στη διαδικασία της βηματικής παλινδρόμησης συνεχίζεται η διαδικασία έρευνας για το κατά πόσον μιά από τις υπόλοιπες μεταβλητές που δεν έχουν ακόμα συμπεριληφθεί στο μοντέλο παρουσιάζει τα χαρακτηριστικά εκείνα που θα επιτρέψουν στον ερευνητή να τη συμπεριλάβει στο μοντέλο ώστε να κατασκευασθεί ένα μοντέλο με τρεις μεταβλητές.

Το βήμα αυτό της ανάλυσης είναι ακριβώς το ίδιο με το βήμα 2ο που έχουμε ήδη δει. Ελέγχουμε δηλαδή, ποιά είναι η μεταβλητή (αν υπάρχει τέτοια μεταβλητή) για την οποία η τιμή της στατιστικής συνάρτησης Τ που αντιστοιχεί στον συντελεστή της β_i και στον έλεγχο υποθέσεως

$$H_0 : \beta_i = 0$$

έναντι της

$$H_1 : \beta_i \neq 0$$

έχει το μικρότερο παρατηρούμενο επίπεδο σημαντικότητας. (Εναλλακτικά, επιλέγεται η μεταβλητή από εκείνες που έχουν απομείνει η οποία έχει τον μεγαλύτερο μερικό συντελεστή συσχέτισης με την εξαρτημένη μεταβλητή).

Το χαρακτηριστικό της διαδικασίας βηματικής παλινδρόμησης, που κάνει τη μέθοδο αυτή να διαφέρει από τις άλλες μεθόδους, είναι ότι καθε φορά που προστίθεται μία νέα μεταβλητή στο μοντέλο γίνονται έλεγχοι των υποθέσεων

$$H_0 : \beta_j = 0$$

έναντι της

$$H_1 : \beta_j \neq 0$$

για όλες τις μεταβλητές που έχουν ήδη περιληφθεί στο μοντέλο με την ίδια διαδικασία που είδαμε στη μέθοδο αποκλεισμού μεταβλητών. Εάν σε οποιονδήποτε από τους ελέγχους αυτούς, διαπιστώσουμε ότι η αντίστοιχη μεταβλητή δεν είναι πλέον στατιστικά σημαντική **διαγράφουμε** τη μεταβλητή αυτή από το μοντέλο.

Επομένως, είναι απαραίτητο στη μεθοδολογία αυτή να καθορίσουμε δύο επίπεδα σημαντικότητας α_1 και α_2 ως απαραίτητα στοιχεία σε ένα πρόγραμμα υπολογιστή της βηματικής παλινδρόμησης.

Η τιμή του α_1 χρησιμοποιείται για τη διαδικασία επιλογής μεταβλητών προκειμένου να περιληφθούν στο μοντέλο ενώ η τιμή του α_2 χρησιμοποιείται για τους ελέγχους που γίνονται για τις μεταβλητές που έχουν ήδη περιληφθεί στο μοντέλο. Προκειμένου η διαδικασία να γίνεται σωστά και το πρόγραμμα να προχωρά χωρίς άπειρες επαναλήψεις (loops) είναι απαραίτητο για τις δύο καθοριζόμενες τιμές α_1 και α_2 να ισχύει:

$$\alpha_2 \geq \alpha_1$$

Ετσι, αποφεύγεται ένας άπειρος κύκλος επαναλήψεων όποτε μια μεταβλητή προστίθεται στο μοντέλο και στη συνέχεια, σε κάποιο άλλο στάδιο διαγράφεται.

Στο παράδειγμά μας ορίσαμε $\alpha_1 = 0.05$ και $\alpha_2 = 0.10$.

Η διαδικασία του 3ου βήματος της βηματικής παλινδρόμησης στο πρόβλημά μας εμφανίζεται στην εκτύπωση που ακολουθεί.

Τρίτο βήμα στη βηματική παλινδρόμηση

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR>F
MODEL	3	39104.05	13034.68	17.06	.0001
ERROR	26	19870.92	764.27		
TOTAL	29	58974.97			

r SQUARE
.6631

PARAMETER	ESTIMATE	T FOR H_0 PARAMETER=0	PR>|T|	STD ERROR OF ESTIMATE
INTERCEPT	1196.165			
X_2	-1.732	-2.21	.0365	0.785
X_4	-20.649	-3.54	.0015	5.831
X_5	3.195	5.69	.0001	0561

Τα υπόλοιπα βήματα στη διαδικασία βηματικής παλινδρόμησης είναι παρόμοια με το βήμα 3. Οπως φαίνεται από τον τελευταίο

πίνακα από το βήμα 3 οδηγούμεθα στο συμπέρασμα ότι θα πρέπει να περιληφθούν στο μοντέλο οι μεταβλητές X_2, X_4 και X_5. Από τον Πίνακα αυτόν προκύπτει επίσης ότι η χρησιμοποίηση των τριών αυτών μεταβλητών οδηγεί στο να εξηγείται το 66.31% της διακύμανσης στους ρυθμούς θνησιμότητας (επειδή, όπως φαίνεται από τον Πίνακα, r^2=0.6631).

Αν συνεχίσουμε τη διαδικασία στη βηματική παλινδρόμηση για το παράδειγμά μας βλέπουμε ότι δεν υπάρχουν άλλες σημαντικές μεταβλητές που να πρέπει να περιληφθούν στο μοντέλο μετά το βήμα 3. Για το λόγο αυτό το βήμα 3 είναι και το τελευταίο της μεθόδου για το συγκεκριμένο πρόβλημα με συμπέρασμα ότι οι στατιστικά σημαντικές μεταβλητές για το πρόβλημα αυτό είναι η X_2, η X_4 και η X_5.

Σημείωση: Θα πρέπει να παρατηρήσουμε ότι ο έλεγχος της υπόθεσης

$$H_0 : \beta_i = 0$$

έναντι της

$$H_1 : \beta_i \neq 0$$

όπως εξηγήθηκε προηγουμένως για το βήμα 3 της βηματικής παλινδρόμησης γίνεται και για το βήμα 2. Ετσι, στο παράδειγμά μας δοθέντος ότι η μεταβλητή X_5 προστέθηκε στο μοντέλο στο βήμα 1 ο υπό συνθήκη έλεγχος

$$H_0 : \beta_5 = 0$$

έναντι της

$$H_1 : \beta_5 \neq 0$$

γίνεται στο βήμα 2 αφού ήδη έχει προστεθεί η μεταβλητή X_4 στο μοντέλο. Αυτό έχει σαν συνέπεια ότι είναι δυνατόν, μαθηματικά, να αφαιρεθεί μιά μεταβλητή στο βήμα 2 η οποία είχε αρχικά επιλεγεί στο βήμα 1 και έτσι το μοντέλο στο βήμα 2 να περιέχει μιά μόνο μεταβλητή. Εν γένει όμως η εμπειρία από τα πρακτικά προβλήματα έχει δείξει ότι δεν συμβαίνει κάτι τέτοιο ποτέ στην πράξη και επομένως ο υπό συνθήκη έλεγχος της υπόθεσης

$$H_0 : \beta_i = 0$$

έναντι της

$$H_1 : \beta_i \neq 0$$

γίνεται χρήσιμος και ουσιαστικός από το βήμα 3 της βηματικής παλινδρόμησης και μετά.

4.4 Σύγκριση της μεθόδου της βηματικής παλινδρόμησης με την μέθοδο αποκλεισμού μεταβλητών

Τόσο η μέθοδος αποκλεισμού μεταβλητών όσο και η μέθοδος της βηματικής παλινδρόμησης είναι διαδικασίες που χρησιμοποιούνται στη Στατιστική ώστε να καθορισθεί ένα υποσύνολο ανεξάρτητων μεταβλητών που είναι χρήσιμο στην πρόβλεψη της εξαρτημένης μεταβλητής. Με δεδομένο ότι οι δύο αυτές μέθοδοι είναι διαφορετικές (όπως επίσης είναι διαφορετικές από την μέθοδο της προοδευτικής προσθήκης μεταβλητών που επίσης εξετάσαμε) είναι φυσικό ότι δεν μπορεί να περιμένει κανείς να δώσουν τα ίδια ακριβώς υποσύνολα μεταβλητών. Παρ' όλα αυτά, εν γένει, τα δύο αυτά υποσύνολα δεν διαφέρουν σημαντικά.

Ας εξετάσουμε, για παράδειγμα, το τελευταίο πρόβλημα που χρησιμοποιήσαμε ως παράδειγμα στη βηματική παλινδρόμηση με την μέθοδο του αποκλεισμού μεταβλητών. Ξεκινώντας με τις δέκα μεταβλητές και εφαρμόζοντας τη διαδικασία της μεθόδου αποκλεισμού μεταβλητών που έχουμε ήδη περιγράψει θα καταλήξουμε ότι το τελευταίο (έβδομο) στάδιο στη διαδικασία αυτή δίνει τα δεδομένα που παρουσιάζονται στον πίνακα που ακολουθεί. Γιά κάθε στάδιο της ανάλυσης έχει χρησιμοποιηθεί το 0.05 σαν επίπεδο σημαντικότητας.

Οπως παρατηρούμε από τον πίνακα αυτό η μέθοδος του αποκλεισμού μεταβλητών δίνει ως στατιστικά σημαντικό υποσύνολο ανεξαρτήτων μεταβλητών αυτό που περιέχει τις μεταβλητές X_1, X_4, X_1X_2 και X_4X_5. (Η μέθοδος της βηματικής παλινδρόμησης, όπως είχαμε δει, είχε δώσει αντίστοιχα το υποσύνολο που περιείχε τις μεταβλητές X_2, X_4, και X_5).

Επομένως, παρότι και οι δύο οι διαδικασίες επέλεξαν τις μεταβλητές X_2, X_4, και X_5 για να συμπεριληφθούν στο μοντέλο οι μεταβλητές αυτές χρησιμοποιούνται διαφορετικά σε κάθε ένα από τα

μοντέλα. Πέραν τούτου παρατηρούμε ότι η διαδικασία αποκλεισμού μεταβλητών έχει περιλάβει και τη μεταβλητή X_1 στο τελικό μοντέλο. Εχουμε επομένως εδώ, μια περίπτωση όπου η διαδικασία αποκλεισμού μεταβλητών και η διαδικασία βηματικής παλινδρόμησης οδήγησαν σε δύο διαφορετικά υποσύνολα μεταβλητών.

Μέθοδος αποκλεισμού μεταβλητών

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR>F
MODEL	4	41260.62	10315.16	14.56	.0001
ERROR	25	17714.34	708.57		
TOTAL	29	58974.97			

r SQUARE .6996

PARAMETER	ESTIMATE	T FOR H_0 PARAMETER=0	PR>|T|	STD ERROR OF ESTIMATE
INTERCEPT	1105.682			
X_1	2.524	2.39	.0245	1.054
X_4	-4.887	-3.05	.0054	6.741
X_1*X_2	2.881 E-02	-2.44	.0221	2.002 E-02
X_4*X_5	2.881 E-01	5.45	.0001	5.290 E-02

Θα πρέπει επίσης να σημειωθεί ότι παρά το γεγονός πως όλες οι διαδικασίες που εξετάσαμε έχουν ως στόχο να επισημάνουν μόνο τις στατιστικά σημαντικές ανεξάρτητες μεταβλητές δεν υπάρχει καμμία ασφαλιστική δικλείδα που να εξασφαλίζει ότι αφενός όλες οι στατιστικά σημαντικές μεταβλητές έχουν προσδιορισθεί ή, από το άλλο μέρος, ότι το τελικό μοντέλο περιέχει μόνο στατιστικά σημαντικές μεταβλητές. Και οι τρεις διαδικασίες που εξετάσαμε χρησιμοποιούν δειγματικά δεδομένα και εξαρτώνται από πολλαπλά τεστ υποθέσεων τα οποία γίνονται στο ίδιο σύνολο δεδομένων. Είναι φυσικό ότι ένα διαφορετικό σύνολο στοιχείων για το ίδιο πρόβλημα ίσως καταλήξει σε διαφορετικά αποτελέσματα. Πέρα από αυτό τα αποτελέσματα είναι δυνατόν να επηρεασθούν από ακραίες τιμές (outliers) ή από την έλλειψη γραμμικότητας στα δεδομένα. Σε κάθε μιά από τις δύο περιπτώσεις, ένας μετασχηματισμός των δεδομένων, όπως π.χ. η θεώρηση λογαρίθμων των τιμών ή η θεώρηση των τάξεων μεγέθους των τιμών, πριν προχωρήσουμε στην ανάλυση

παλινδρόμησης αποδεικνύεται συχνά χρήσιμη στον προσδιορισμό των σημαντικών μεταβλητών.

4.5 Μέθοδοι Ολων των Πιθανών Παλινδρομήσεων

Είναι ενδεχόμενο, με την αυστηρή χρησιμοποίηση των μεθόδων που παρουσιάσαμε, να αγνοηθεί ένα καλό μικρό υποσύνολο μεταβλητών ή, να επιλεγεί ένα μικρό υποσύνολο ως το καλύτερο δυνατό ενώ υπάρχουν ενδεχομένως υποσύνολα με λίγες περισσότερες μεταβλητές μεταξύ αυτών που είναι υποψήφιες για να περιληφθούν στο μοντέλο με αρκετά υψηλότερη τιμή για το συντελεστή προσδιορισμού r^2. Μια λύση για το πρόβλημα αυτό είναι να εφαρμοσθούν όλα τα δυνατά μοντέλα παλινδρόμησης που είναι δυνατόν να κατασκευασθούν με υποσύνολα των ανεξαρτήτων μεταβλητών που είναι υποψήφιες να περιληφθούν στο μοντέλο. Αν υποθέσουμε, ότι υπάρχουν k ανεξάρτητες μεταβλητές που ενδεχομένως είναι κατάλληλες να περιληφθούν στο μοντέλο παλινδρόμησης τα δυνατά υποσύνολα των μεταβλητών αυτών (και επομένως τα αντίστοιχα μοντέλα που μπορούν να προκύψουν από αυτές) είναι

$$\binom{k}{1} + \binom{k}{2} + \ldots + \binom{k}{k} = 2^k - 1$$

Στη βιβλιογραφία υπάρχουν αρκετοί αποτελεσματικοί αλγόριθμοι που έχουν τη δυνατότητα υπολογισμού των αναλύσεων παλινδρόμησης για όλα αυτά τα 2^κ - 1 δυνατά μοντέλα (π.χ. το SAS).

Τέτοιοι αλγόριθμοι υπάρχουν στις εργασίες των Garside (1965), Schatgoff (1968), Furnival (1971), Furnival & Wilson (1974). Ο Frane έχει εξηγήσει και διατυπώσει τον τρόπο με τον οποίο η μέθοδος των Furnival & Wilson μπορεί να χρησιμοποιηθεί στο στατιστικό πακέτο BMDP.

4.6 Διαστήματα εμπιστοσύνης και πρόβλεψης με τη μέθοδο του αποκλεισμού μεταβλητών και της βηματικής παλινδόμησης

Χρησιμοποιώντας τις μεθόδους αναλύσεων παλινδρόμησης που αναπτύξαμε τελευταία είναι δυνατόν να κατασκευάσουμε διαστήματα

εμπιστοσύνης και πρόβλεψης για το μ και το Y χρησιμοποιώντας τους υπολογιστές. Στο τελευταίο παράδειγμα που εξετάσαμε που αναφέρεται σε ρυθμούς θνησιμότητας έστω ότι θέλουμε να κατασκευάσουμε ένα 95% διάστημα εμπιστοσύνης για το μέσο ρυθμό θνησιμότητας σε μιά περιοχή η οποία έχει τα εξής χαρακτηριστικά :

X_1 =43.3, X_2 =30.2, X_3 =76.5, X_4 =11.0, X_5 =18.8 και X_6 =166.0.

Για την ανάλυση της βηματικής παλινδρόμησης βλέπουμε, από το σχετικό πίνακα, ότι:

$\hat{\mu}_{Y|x}$ = 1196.165 - (1.732)(30.2) - (20.649)(11.0) + (3.195)(18.8) = 976.79

Το 95% διάστημα εμπιστοσύνης για το μέσο ρυθμό θνησιμότηταςδίνεται από το στατιστικό πρόγραμμα να έχει ακραία σημεία 956.39 και 997.20.

Αντίστοιχα, η μέθοδος του αποκλεισμού μεταβλητών δίνει από τον αντίστοιχο πίνακα

$\hat{\mu}_{Y|x}$ =1105.682+(2.524)(43.3)+(20.554)(11.0)-(0.04887)(43.3)(30.2)

+ (0.2881)(11.0)(18.8) = 984.55.

Το 95% διάστημα εμπιστοσύνης για το μέσο ρυθμό θνησιμότητας στην περίπτωση αυτή υπολογίζεται να είναι (962.19, 1006.94).

Αντίστοιχα, με την βοήθεια του υπολογιστή μπορούν να κατασκευασθούν διαστήματα πρόβλεψης για το Y.