

2. ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ ΜΕ ΤΗ ΜΕΘΟΔΟ ΤΟΥ ΑΠΟΚΛΕΙΣΜΟΥ ΜΕΤΑΒΛΗΤΩΝ (Backward Elimination Procedure)

Στην στατιστική βιβλιογραφία υπάρχουν πολλές μέθοδοι για τον καθορισμό του καλύτερου υποσυνόλου από ένα σύνολο μεταβλητών που είναι υποψήφιος να περιληφθούν σε ένα μοντέλο πολλαπλής παλινδρόμησης. Μια από τις μεθόδους αυτές που θα εξηγήσουμε αργότερα ξεκινά με μοντέλο που δεν έχει καμία μεταβλητή και στη συνέχεια προσθέτει κάθε φορά από μια μεταβλητή που έχει σημαντική συνεισφορά στο μοντέλο. Η μέθοδος αυτή ονομάζεται συνήθως **μέθοδος προοδευτικής προσθήκης μεταβλητών** (*forward procedure*). Η μέθοδος που θα συζητήσουμε εδώ χρησιμοποιεί την ακριβώς αντίθετη λογική: ξεκινά περιλαμβάνοντας όλες τις μεταβλητές στο μοντέλο και σε κάθε βήμα αποκλείει μια μεταβλητή που δεν έχει σημαντική συνεισφορά σε αυτό. Η μέθοδος αυτή ονομάζεται **μέθοδος αποκλεισμού μεταβλητών** (*backward elimination procedure*).

Όπως είδαμε σε προηγούμενη ενότητα έξι από τις οκτώ μεταβλητές του μοντέλου που χρησιμοποιήθηκε για το παράδειγμά μας είναι υποψήφιος για αποκλεισμό. Η μέθοδος αποκλεισμού μεταβλητών αποκλείει σε κάθε βήμα την μεταβλητή εκείνη που έχει το μεγαλύτερο παρατηρούμενο επίπεδο σημαντικότητας (*P value*) με την προϋπόθεση ότι το παρατηρούμενο αυτό επίπεδο σημαντικότητας υπερβαίνει το επίπεδο σημαντικότητας α που έχουμε προκαθορίσει. Έτσι, όπως βλέπουμε από τον πίνακα 2, αν χρησιμοποιήσουμε σαν επίπεδο σημαντικότητας το 0.05 στο πρώτο στάδιο της μεθόδου θα αποκλείσουμε την μεταβλητή X_7 γιατί έχει το μεγαλύτερο επίπεδο σημαντικότητας (*P value* = 0.5294) με τιμή μεγαλύτερη από το 0.05. Στη συνέχεια, χρησιμοποιώντας και πάλι τη μέθοδο των ελαχίστων τετραγώνων, προσαρμόζουμε το μοντέλο με τις υπόλοιπες επτά μεταβλητές στα δεδομένα. Τα αποτελέσματα δίνονται στον πίνακα που ακολουθεί. Παρατηρούμε, ότι, όπως είναι φυσικό, οι συντελεστές και τα παρατηρούμενα επίπεδα σημαντικότητας που αναφέρονται στις υπόλοιπες μεταβλητές στον πίνακα αυτού είναι όλα διαφορετικά από

τις αντίστοιχες τιμές του αντίστοιχου πίνακα, όταν χρησιμοποιήθηκε και η μεταβλητή X_7 .

Οι επόμενες 6 εκτυπώσεις από τον υπολογιστή εμφανίζουν τα υπόλοιπα βήματα στην μέθοδο αποκλεισμού μεταβλητών για το πρόβλημά μας. Όπως βλέπουμε στο δεύτερο βήμα αποκλείεται η μεταβλητή X_5 γιατί έχει παρατηρούμενο επίπεδο σημαντικότητας (P value) 0.5267. Στο τρίτο στάδιο αποκλείεται η μεταβλητή X_4 γιατί έχει παρατηρούμενο επίπεδο σημαντικότητας 0.5659. Η διαδικασία αυτή συνεχίζεται με τον ίδιο τρόπο μέχρι το στάδιο 7 όπου, όπως βλέπουμε, όλα τα παρατηρούμενα επίπεδα σημαντικότητας για τις μεταβλητές που έχουν μείνει έχουν τιμές μικρότερες από 0.05. Έτσι, καμία άλλη μεταβλητή δεν αποκλείεται από το μοντέλο και το τελικό μοντέλο στο οποίο καταλήγουμε με τη μέθοδο αυτή περιλαμβάνει μόνο τις μεταβλητές X_3 και X_6 .

Μοντέλο 7 μεταβλητών

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR>F
MODEL	7	149.51	21.36	6.79	.0001
ERROR	42	144.98	3.45		
TOTAL	49	294.49			

r SQUARE .5077				
PARAMETER	ESTIMATE	T FOR H ₀ PARAMETER=0	PR> T	STD ERROR OF ESTIMATE
INTERCEPT	15.313			
X_1	-5.319 E-02	-0.99	.3295	5.392E-02
X_2	7.246 E-01	0.99	.3286	7.331 E-01
X_3	1.902 E-04	3.92	.0003	4.856 E-05
X_4	7.182 E-02	0.84	.4079	8.591 E-02
X_5	1.009 E-01	0.64	.5267	1.580 E-01
X_6	-7.726 E-01	-2.52	.0158	3.071 E-01
X_8	-2.072	-1.51	.1384	1.372

Μοντέλο 6 Μεταβλητών

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR>F
MODEL	6	148.10	24.68	7.25	.0001
ERROR	43	146.39	3.40		
TOTAL	49	294.49			

r SQUARE				
5125				

PARAMETER	ESTIMATE	T FOR H ₀ PARAMETER=0	PR> T	STD ERROR OF ESTIMATE
INTERCEPT	15.080			
X ₁	-2.689 E-02	-0.78	.4406	3.455 E-02
X ₂	9.040 E-01	1.34	.1859	6.725 E-01
X ₃	1.874 E-04	3.90	.0003	4.802 E-05
X ₄	4.056 E-02	0.58	.5659	7.010 E-02
X ₆	-7.907 E-01	-2.60	.0126	3.037 E-01
X ₈	-1.986	-1.47	.1502	1.356

Μοντέλο 5 Μεταβλητών

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR>F
MODEL	5	146.96	29.39	8.77	.0001
ERROR	44	147.53	3.35		
TOTAL	49	294.49			

r SQUARE				
.4990				

PARAMETER	ESTIMATE	T FOR H ₀ PARAMETER=0	PR> T	STD ERROR OF ESTIMATE
INTERCEPT	15.811			
X ₁	-1.442 E-02	-0.54	.5932	2.680 E-02
X ₂	1.078	1.81	.0777	5.968 E-01
X ₃	1.944 E-04	4.22	.0001	4.611 E-05
X ₆	-8.673 E-01	-3.20	.0026	2.712 E-01
X ₈	-1.804	-1.38	.1750	1.308

Μοντέλο 4 Μεταβλητών

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR>F
MODEL	4	145.96	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

Μοντέλο 3 Μεταβλητών

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR>F
MODEL	3	140.34	46.78	13.96	.0001
ERROR	46	151.14	3.35		
TOTAL	49	294.49			

r SQUARE				
.4766				

PARAMETER	ESTIMATE	T FOR H ₀ PARAMETER=0	PR> T	STD ERROR OF ESTIMATE
INTERCEPT	15.165			
X ₂	1.125	1.92	.0608	5.851 E-01
X ₃	1.631 E-04	5.14	.0001	3.173 E-05
X ₆	-8.081 E-01	-3.25	.0021	2.483 E-01

Μοντέλο 2 Μεταβλητών

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR>F
MODEL	2	129.96	63.98	18.06	.0001
ERROR	47	166.53	3.54		
TOTAL	49	294.49			

		r SQUARE		
		.4345		
PARAMETER	ESTIMATE	T FOR H ₀ PARAMETER=0	PR> T	STD ERROR OF ESTIMATE
INTERCEPT	14.464			
X ₃	1.740 E-04	5.42	.0001	3.211 E-05
X ₆	-7.437 E-01	-2.94	.0051	2.530 E-01

Δηλαδή με την μέθοδο αυτή το μοντέλο που θα χρησιμοποιηθεί από την εταιρεία παροχής συμβουλών θα είναι το

$$E(Y) = \alpha + \beta_3 X_3 + \beta_6 X_6.$$

Παρατήρηση: Θα πρέπει να επισημάνουμε εδώ ότι, ενδεχομένως, μια διαφορετική μέθοδος επιλογής μεταβλητών για την πολλαπλή παλινδρόμηση θα καταλήξει σε διαφορετικό σύνολο μεταβλητών που θα πρέπει να περιληφθούν στο μοντέλο. Εν γένει όμως υπάρχει αρκετή συμφωνία μεταξύ των διαφόρων μεθόδων. Παρ' όλα αυτά συχνά υπάρχει σημαντική διαφορά μεταξύ των μεταβλητών του τελικού μοντέλου και αυτών που εμφανίζονται ως υποψήφιες για αποκλεισμό στο πρώτο στάδιο της μεθόδου αποκλεισμού μεταβλητών που περιγράψαμε. Αυτό συμβαίνει γιατί ο αποκλεισμός μιας μεταβλητής μεταβάλλει τα παρατηρούμενα επίπεδα σημαντικότητας που αντιστοιχούν στις υπόλοιπες μεταβλητές. Εξάλλου, από τη στιγμή που ο ερευνητής έχει πρόσβαση σε υπολογιστή είναι λογικό και φρόνιμο να αποκλείει μια μεταβλητή σε κάθε βήμα.

2.1 Η Πιθανότητα λάθους πρώτου είδους (Probability of a type I error)

Σε κάθε βήμα στην μέθοδο αποκλεισμού μεταβλητών (*backward elimination procedure*) γίνεται ένας έλεγχος υποθέσεως σε κάποιο συγκεκριμένο και προκαθορισμένο επίπεδο σημαντικότητας α . Η τιμή του α υποτίθεται ότι εκπροσωπεί την πιθανότητα να κάνουμε λάθος πρώτου είδους (να απορρίψουμε δηλαδή την μηδενική υπόθεση ενώ η υπόθεση αυτή είναι σωστή). Παρ' όλα αυτά η τιμή του α είναι πράγματι η πιθανότητα να κάνουμε λάθος πρώτου είδους μόνο στο πρώτο βήμα και μόνον αν οι υποθέσεις για την κατανομή των ϵ_i ισχύουν. Μετά το πρώτο βήμα, ο έλεγχος γίνεται στην πραγματικότητα δεσμευμένος έλεγχος (δεσμευμένος στο ενδεχόμενο ότι τουλάχιστον μια μεταβλητή έχει αφαιρεθεί από το μοντέλο). Για το λόγο αυτό η τιμή του α δεν υποδηλοί την πραγματική πιθανότητα λάθους πρώτου είδους. Το γεγονός επίσης ότι γίνονται μια σειρά από ελέγχους υποθέσεων με τη χρησιμοποίηση των ιδίων δεδομένων τείνει να αυξάνει την πιθανότητα να γίνει λάθος πρώτου είδους σε κάποιο σημείο της διαδικασίας.

Επομένως θα πρέπει να δίνεται ιδιαίτερη προσοχή στην ερμηνεία του α σε οποιαδήποτε διαδικασία όπως αυτή που περιγράφουμε (και όπως αυτή που θα περιγράψουμε στη συνέχεια) όπου γίνονται πολλαπλοί έλεγχοι υποθέσεων στο ίδιο σύνολο δεδομένων.

Σε τέτοιες περιπτώσεις το α είναι πράγματι η πιθανότητα να γίνει λάθος πρώτου είδους μόνο στο πρώτο βήμα της διαδικασίας και εφόσον βέβαια ισχύουν οι υποθέσεις για τις κατανομές των λαθών. Η πιθανότητα να γίνει λάθος πρώτου είδους στο επόμενο ή στα επόμενα βήματα είναι στην πραγματικότητα άγνωστη. Για τους ελέγχους αυτούς η τιμή του α χρησιμεύει μόνο για να μπορέσουμε να χρησιμοποιήσουμε τον έλεγχο και να αποφασίσουμε αν μια μεταβλητή πρέπει να μείνει στο μοντέλο ή όχι.

Είναι προφανές ότι η συνολική πιθανότητα να γίνει λάθος πρώτου είδους τουλάχιστον μια φορά κατά τη διάρκεια της διαδικασίας πολλαπλών αποφάσεων όπως αυτή που περιγράψαμε, είναι πάντοτε μεγαλύτερη από το α (αφού α είναι η πιθανότητα που

αντιστοιχεί στο πρώτο βήμα της διαδικασίας και, προφανώς, στα βήματα που ακολουθούν εμφανίζονται πρόσθετες ευκαιρίες για να γίνει λάθος πρώτου είδους).

Το λάθος που γίνεται από την αδυναμία υπολογισμού της συνολικής πιθανότητας λάθους πρώτου είδους στην διαδικασία που προαναφέραμε δεν αποτελεί σοβαρό πρόβλημα με την προϋπόθεση ότι ο αναλυτής γνωρίζει ότι το πρόβλημα αυτό υπάρχει. Αν το πρόβλημα αγνοηθεί μια αυξημένη πιθανότητα λάθους πρώτου είδους είναι ενδεχόμενο να έχει σοβαρές συνέπειες (όπως για παράδειγμα να καταλήξει κανείς στο συμπέρασμα ότι ένα φάρμακο είναι αποτελεσματικό μετά από μια εξαντλητική διαδικασία στην οποία έχουν χρησιμοποιηθεί πολλά φάρμακα).

2.2 Διάστημα εμπιστοσύνης για τον μέσο $\mu_{Y|x}$

Όπως και στην απλή γραμμική παλινδρόμηση το $100(1-\alpha)\%$ διάστημα εμπιστοσύνης τον μέσου $\mu_{Y|x}$ δίνεται από τον τύπο

$$\hat{\mu}_{Y|x} \pm t_{n-k-1, 1-\alpha/2} S_{\hat{\mu}_{Y|x}}$$

όπου $\hat{\mu}_{Y|x}$ είναι μια σημειακή εκτίμηση του $\mu_{Y|x}$ και υπολογίζεται με αντικατάσταση των γνωστών τιμών των μεταβλητών $X_1, \dots, X_k$ στην εκτιμήτρια ελαχίστων τετραγώνων της εξίσωσης παλινδρόμησης.

Το τυπικό λάθος του $\hat{\mu}_{Y|x}$ είναι δύσκολο να υπολογισθεί στην πολλαπλή παλινδρόμηση και για το λόγο αυτό συνήθως χρησιμοποιούνται οι τιμές που δίνει ο υπολογιστής.

Παράδειγμα: Όπως υπολογίσαμε στην προηγούμενη ενότητα η εκτιμώμενη επιφάνεια παλινδρόμησης, μετά το τέλος της διαδικασίας αποκλεισμού μεταβλητών είναι η

$$\hat{\mu}_{Y|x} = 14.46 + 0.0001740 X_3 - 0.7437 X_6$$

Να καθορισθεί το 95% διάστημα εμπιστοσύνης για την πραγματική μέση ικανοποίηση στο χώρο εργασίας εργαζομένων που έχουν εβδομαδιαίο μισθό 36000 και μεταπτυχιακό τίτλο Master's.

Λύση: Αντικαθιστώντας τις τιμές $X_3 = 36000$ και $X_6 = 17$ στην εκτιμηθείσα επιφάνεια παλινδρόμησης έχουμε

$$\hat{\mu}_{Y|x} = 14.46 + 0.0001740 (36000) - 0.7437 (17) = 8.085$$

Η τιμή αυτή είναι μια σημειακή εκτίμηση του $\mu_{Y|x}$.

Για τις τιμές αυτές του X_3 και X_6 το στατιστικό πακέτο SAS δίνει το παρακάτω 95% διάστημα εμπιστοσύνης για το $\mu_{Y|x}$:

LOWER 95% CL	UPPER 95% CL
FOR MEAN	FOR MEAN
7.184	8.986

Επομένως με βεβαιότητα 95% μπορούμε να πούμε ότι με βάση τα στοιχεία του δείγματος η πραγματική μέση ικανοποίηση από το χώρο δουλειάς για όλους τους εργαζομένους με εβδομαδιαίο εισόδημα 36000 και μεταπτυχιακό τίτλο Master's βρίσκεται μεταξύ των τιμών 7.184 και 8.986.

2.3 Διάστημα πρόβλεψης για το Y

Όπως έχουμε δει μία σημαντική χρησιμότητα της θεωρίας παλινδρόμησης είναι η πρόβλεψη αγνώστων τιμών της εξαρτημένης μεταβλητής Y για δοθείσες τιμές των εξαρτημένων μεταβλητών. Στην περίπτωση αυτή μια υψηλή τιμή του συντελεστή r^2 είναι σημαντική (αφού κάτι τέτοιο υποδηλώνει τόσο μια καλή προσαρμογή στα δεδομένα όσο και μια καλύτερη τιμή για την πρόβλεψη).

Το $100(1-\alpha)\%$ διάστημα πρόβλεψης για την τιμή Y με τη χρησιμοποίηση της πολλαπλής παλινδρόμησης για συγκεκριμένες τιμές των ανεξάρτητων μεταβλητών $X_1, X_2, \dots, X_k$ δίνεται από τον γνωστό τύπο

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

Επειδή και στην περίπτωση αυτή ο υπολογισμός του τυπικού λάθους $S_{\hat{Y}}$ είναι δύσκολος χρησιμοποιούνται στατιστικά πακέτα για τον προσδιορισμό του διαστήματος πρόβλεψης. Το στοιχείο που χρειάζεται για τον καθορισμό του είναι το μοντέλο που θα

χρησιμοποιηθεί για να δώσει την συγκεκριμένη πρόβλεψη. Το μοντέλο αυτό δεν είναι βέβαια υποχρεωτικά αυτό το οποίο προκύπτει ως τελευταίο βήμα στη μέθοδο αποκλεισμού μεταβλητών.

Παράδειγμα: Να υπολογισθεί ένα 95% διάστημα πρόβλεψης για το μοντέλο της προηγούμενης παραγράφου.

Λύση: Αντικαθιστώντας τις τιμές $X_3 = 36000$ και $X_6 = 17$ στην εκτιμηθείσα επιφάνεια παλινδρόμησης θα βρούμε, προφανώς, ως σημειακή εκτιμήτρια του Y την τιμή

$$\hat{Y} = 8.085$$

(Η σημειακή εκτιμήτρια για το Y είναι, προφανώς, η ίδια με τη σημειακή εκτιμήτρια του $\mu_{Y|x}$). Το 95% διάστημα πρόβλεψης για την πρόβλεψη Y που αντιστοιχεί στις παραπάνω τιμές του X_3 και X_6 δίνεται, από το στατιστικό πακέτο SAS ως εξής:

LOWER 95% CL	UPPER 95% CL
INDIVIDUAL	INDIVIDUAL
4.192	11.977

Σημείωση: Το διάστημα πρόβλεψης που καταλήξαμε για την ικανοποίηση στο χώρο δουλειάς εμφανίζεται μεγάλο σε σχέση με τα δεδομένα που έχουμε στον αρχικό πίνακα. Επομένως, είναι δύσκολο να εκτιμήσουμε με κάποια ακρίβεια, ποιά είναι η ακριβής ικανοποίηση από το επάγγελμα αυτό ενός συγκεκριμένου ατόμου με εβδομαδιαίο εισόδημα 36000 και μεταπτυχιακό τίτλο Master's. Το μέγεθος του διαστήματος πρόβλεψης εξαρτάται τόσο από το μέγεθος της διασποράς στα άτομα του δείγματος όσο και από το πόσο καλά το συγκεκριμένο μοντέλο εφαρμόζει στα δεδομένα.