Πολυτεχνείο Κρήτης
Τμήμα Μηχανικών Παραγωγής και Διοίκησης

Διπλωματική εργασία της ***Στυλιανής Κ. Παπαδάκη***

Διατριβή που υπεβλήθη για τη μερική ικανοποίηση των απαιτήσεων για την απόκτηση Μεταπτυχιακού Διπλώματος Ειδίκευσης

Επιβλέπων: Μουστάκης Βασίλειος

Χανιά 2009

Πολυτεχνείο Κρήτης
Τμήμα Μηχανικών Παραγωγής και Διοίκησης

Αξιολόγηση Στατιστικών Μεθοδολογιών για τη Διαχείριση Δεδομένων με Ελλιπή Στοιχεία

Assessment of Statistical Methods supporting Handling of Missing Data

Διπλωματική εργασία της Στυλιανής Κ. Παπαδάκη

Επιβλέπων: Μουστάκης Βασίλειος

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την .....

Β. Μουστάκης , Αναπληρωτής Καθηγητής

Ε. Γρηγορούδης, Επίκουρος Καθηγητής

Μ. Δούμπος, Επίκουρος Καθηγητής

Copyright © Παπαδάκη Στυλιανή, 2009
Με επιφύλαξη παντός δικαιώματος. All rights reserved.
Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Σύντομο Βιογραφικό Σημείωμα

Η Στυλιανή Κ. Παπαδάκη γεννήθηκε το 1983 στο Ηράκλειο της Κρήτης. Το 2001 ξεκίνησε της σπουδές της στη Σχολή Χημικών Μηχανικών του Εθνικού Μετσόβιου Πολυτεχνείου από όπου και αποφοίτησε το 2007. Το ίδιο έτος ξεκίνησε τις μεταπτυχιακές σπουδές της στο Πολυτεχνείο Κρήτης, στην Σχολή Μηχανικών Παραγωγής και διοίκησης, στον τομέα της Οργάνωσης και διοίκησης.

Ευχαριστίες

Με την ολοκλήρωση της μεταπτυχιακής μου εργασίας και των Μεταπτυχιακών σπουδών μου θα ήθελα να ευχαριστήσω όλα εκείνα τα άτομα που ήταν δίπλα μου και με βοήθησαν να επιτύχω τους στόχους μου.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κύριο Μουστάκη Βασίλειο για την εμπιστοσύνη, τη βοήθειά και την καθοδήγηση του καθόλη τη διάρκεια της διπλωματικής μου εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω τους καθηγητές κυρίους Δούμπο Μιχαήλ και Γρηγορούδη Ευάγγελο, για την τιμή που μου έκαναν να διαβάσουν την διπλωματική μου εργασία και να συμμετάσχουν στην επιτροπή εξέτασης αυτής.

Τέλος, οφείλω ένα μεγάλο ευχαριστώ στους γονείς μου, οι οποίοι στάθηκαν δίπλα μου από τα παιδικά μου χρόνια και με βοήθησαν με περίσσεια υπομονή και αγάπη σε όλη μου την πορεία μέχρι σήμερα, καθώς και στους φίλους μου και στους ξεχωριστούς ανθρώπους που βρίσκονται δίπλα μου, με στηρίζουν και με ανέχονται όλα αυτά τα χρόνια.

Περίληψη

Η στατιστική επεξεργασία και ανάλυση των δεδομένων χρησιμοποιείται καθημερινά σε πολλούς τομείς της επιστήμης. Αποτελεί αναπόσπαστο κομμάτι τόσο των χιλιάδων στατιστικών ερευνών που λαμβάνουν χώρα καθημερινά αναφορικά με όλους τους κλάδους της κοινωνίας, όσο και των ερευνητικών προγραμμάτων που εκπονούνται στους διάφορους επιστημονικούς κλάδους (ιατρική, βιοτεχνολογία, χημεία).

Κατά την στατιστική επεξεργασία και ανάλυση των δεδομένων οι ερευνητές έρχονται αντιμέτωποι με πολλά προβλήματα, τα οποία μπορεί να οφείλονται τόσο σε λάθος σχεδιασμό της έρευνας όσο και σε κάποιους αστάθμητους παράγοντες. Ένα από τα πιο συνήθη και σημαντικά προβλήματα που αντιμετωπίζουν είναι η έλλειψη κάποιων από τα στοιχεία της έρευνας.

Η έλλειψη αυτών των στοιχείων αποτελεί σημαντικό πρόβλημα για τους ερευνητές, καθώς μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα, τα οποία μπορεί να καταστρέψουν μία μακροχρόνια έρευνα ή ακόμα και να προκαλέσουν σοβαρά προβλήματα στην κοινωνία, παραδείγματος χάριν την δημιουργία ενός φαρμάκου που να έχει καταστρεπτικές συνέπειες για την υγεία του ανθρώπου.

Τις τελευταίες δεκαετίες, υπάρχει μία συστηματική προσπάθεια για τον σχεδιασμό μεθόδων, οι οποίες επιτρέπουν τον χειρισμό των δεδομένων που έχουν ελλιπή στοιχεία, χωρίς να επηρεάζονται τα συμπεράσματα των ερευνών.

Σκοπό αυτής της εργασίας αποτελεί η παρουσίαση των μεθόδων που χρησιμοποιούνται σήμερα για την αντιμετώπιση του προβλήματος του χειρισμού δεδομένων με ελλιπή στοιχεία και η σύγκρισή τους με σκοπό την καλύτερη αντιμετώπιση του προβλήματος.

Αναλυτικότερα, στο πρώτο κεφάλαιο θα γίνει μία ιστορική αναδρομή στο πρόβλημα των ελλιπών στοιχείων και των μεθόδων διαχείρισή του, θα τεθούν οι λόγοι για την ύπαρξη των ελλιπών στοιχείων, καθώς και οι συνέπειες που μπορεί να έχουν αυτές.

Στο δεύτερο κεφάλαιο θα τεθεί το πρόβλημα των ελλιπών δεδομένων, θα παρουσιαστούν οι μορφές (τύποι) έλλειψης των δεδομένων που βοηθούν στην επιλογή της μεθόδου και είναι σύμφωνα με τον Rubin και Little (2002): α) η Γενική μορφή ή μη μονότονη (General), α) η Μονότονη μορφή (Monotone), γ) το Μονομεταβλητό ελλείπον στοιχείο (Univariate nonresponse), δ) το Πολυμεταβλητό ελλείπον στοιχείο (Multivariate two patterns), ε) το File Matching και στ) το Factor Analysis. Τέλος παρουσιάζονται οι μηχανισμοί έλλειψης των στοιχείων όπως αυτοί εισήχθηκαν από τους Rubin και Little (1987) – MCAR, MAR και το MNAR – και οι οποίοι αποτελούν σημαντικό κριτήριο για την επιλογή της μεθόδου.

Στο τρίτο κεφάλαιο διαχωρίζονται οι μέθοδοι που χρησιμοποιούνται για την διαχείριση των δεδομένων με ελλιπή στοιχεία με βάση τον τρόπο που διαχειρίζονται τα ελλιπή στοιχεία και έτσι διακρίνονται σε τέσσερις κατηγορίες: α) μέθοδοι διαγραφής (Deletion Methods), β) μέθοδοι βαρών (Weighting Methods), γ) μέθοδοι καταλογισμού (Imputation Methods) και δ) μέθοδοι που στηρίζονται σε μοντέλα (Model Based Methods).

Στα επόμενα τέσσερα κεφάλαια παρουσιάζονται κάθε μία από αυτές τις τέσσερις κατηγορίες μεθόδων χωριστά. Αναπτύσσονται οι πιο αντιπροσωπευτικές μεθόδους κάθε κατηγορίας, παρουσιάζονται τα

πλεονεκτήματα και τα μειονεκτήματά τους και οι προϋποθέσεις κάτω από τις οποίες έχουν τα επιθυμητά αποτελέσματα.

Τέλος στο τελευταίο κεφάλαιο γίνεται μία σύγκριση μεταξύ αυτών των μεθόδων, με σκοπό να βρεθεί πότε κάθε μέθοδος υπερτερεί έναντι των άλλων και θα πρέπει να προτιμάται από τους ερευνητές, παραθέτοντας τα βασικά στοιχεία κάθε μεθόδου και τις προϋποθέσεις κάτω από τις οποίες χρησιμοποιείται.

Περιεχόμενα

ΠΕΡΙΛΗΨΗ 7

ΠΕΡΙΕΧΟΜΕΝΑ 10

ΕΙΣΑΓΩΓΙΚΕΣ ΕΝΝΟΙΕΣ 13

1.1 ΕΙΣΑΓΩΓΗ 13
1.2 ΙΣΤΟΡΙΚΗ ΕΞΕΛΙΞΗ 14
1.3 ΟΙ ΛΟΓΟΙ ΤΗΣ ΑΠΩΛΕΙΑΣ ΤΩΝ ΣΤΟΙΧΕΙΩΝ 16
1.4 ΟΙ ΣΥΝΕΠΕΙΕΣ ΤΗΣ ΕΛΛΕΙΨΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ 19

ΤΟ ΠΡΟΒΛΗΜΑ ΤΩΝ ΕΛΛΙΠΩΝ ΣΤΟΙΧΕΙΩΝ (MISSING DATA PROBLEM) 21

2.1 ΤΟ ΠΡΟΒΛΗΜΑ ΤΩΝ ΕΛΛΙΠΩΝ ΣΤΟΙΧΕΙΩΝ 21
2.2 ΟΙ ΜΟΡΦΕΣ ΤΩΝ ΕΛΛΙΠΩΝ ΔΕΔΟΜΕΝΩΝ (MISSING DATA PATTERNS) 23
2.2.1 Γενική μορφή ή μη μονότονη (General) 24
2.2.2 Μονότονη μορφή (Monotone) 25
2.2.3 Μονοεταβλητό ελλείπον στοιχείο (Univariate nonresponse) 26
2.2.4 Πολυμεταβλητό ελλείπον στοιχείο (Multivariate two patterns) 27
2.2.5 File Matching 29
2.2.6 Factor Analysis 30
2.3 ΜΗΧΑΝΙΣΜΟΙ ΤΩΝ ΕΛΛΙΠΩΝ ΣΤΟΙΧΕΙΩΝ 31
2.4 ΔΙΕΡΕΥΝΗΣΗ ΤΥΧΑΙΑΣ ΕΛΛΕΙΨΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ 37

ΜΕΘΟΔΟΙ ΔΙΑΧΕΙΡΙΣΗΣ ΔΕΔΟΜΕΝΩΝ ΜΕ ΕΛΛΙΠΗ ΣΤΟΙΧΕΙΑ – ΤΑΞΙΝΟΜΗΣΗ 39

3.1 ΠΑΡΑΔΟΣΙΑΚΕΣ – ΜΟΝΤΕΡΝΕΣ ΜΕΘΟΔΟΙ 40
3.2 ΤΑΞΙΝΟΜΗΣΗ ΜΕΘΟΔΩΝ ΜΕ ΒΑΣΗ ΤΟΝ ΧΕΙΡΙΣΜΟ ΤΩΝ ΕΛΛΙΠΩΝ ΣΤΟΙΧΕΙΩΝ 42
3.2.1 Μέθοδοι Διαγραφής (Deletion Methods) 42
3.2.2 Διαδικασίες Στάθμισης (Weigthing Procedures) 43
3.2.3 Μέθοδοι Απόδοσης (Imputation methods) 44
3.2.4 Μέθοδοι που βασίζονται σε μοντέλα (Model – Based Methods). 45

ΜΕΘΟΔΟΙ ΔΙΑΓΡΑΦΗΣ (DELETION PROCEDURES).....46

4.1 Η ΑΝΑΛΥΣΗ ΠΛΗΡΩΣ ΣΥΜΠΛΗΡΩΜΕΝΩΝ ΠΕΡΙΠΤΩΣΕΩΝ, COMPLETE CASE ANALYSIS Η LISTWISE DELETION 47

4.1.1 Ιδιότητες.....48

4.1.2 Πλεονεκτήματα – Μειονεκτήματα 48

4.1.3 Συμπεράσματα.....49

4.2 ΔΙΑΘΕΣΙΜΗ ΠΕΡΙΠΤΩΣΙΟΛΟΓΙΚΗ ΑΝΑΛΥΣΗ (AVAILABLE CASE ANALYSIS) Η PAIRWISE DELETION 50

4.2.1 Ιδιότητες.....50

4.2.2 Πλεονεκτήματα – Μειονεκτήματα 51

4.2.3 Συμπεράσματα.....52

ΔΙΑΔΙΚΑΣΙΕΣ ΣΤΑΘΜΙΣΗΣ (WEIGHTING PROCEDURES) 53

5.1 ΤΑ ΣΤΑΔΙΑ ΤΗΣ ΔΙΑΔΙΚΑΣΙΑΣ ΣΤΑΘΜΙΣΗΣ 54

5.2 ΠΡΟΣΑΡΜΟΣΜΕΝΕΣ ΟΜΑΔΕΣ ΣΤΑΘΜΙΣΗΣ (WEIGHTING CLASS ADJUSTMENTS) 57

5.3 POSTSTRATIFICATION ADJUSTMENT.....60

5.4 ΠΛΕΟΝΕΚΤΗΜΑΤΑ – ΜΕΙΟΝΕΚΤΗΜΑΤΑ.....61

5.5 ΣΥΜΠΕΡΑΣΜΑΤΑ 62

ΜΕΘΟΔΟΙ ΑΝΤΙΚΑΤΑΣΤΑΣΗΣ (REPLACEMENT PROCEDURES) 63

6.1 ΣΤΟΧΑΣΤΙΚΕΣ – ΑΙΤΙΟΚΡΑΤΙΚΕΣ ΜΕΘΟΔΟΙ 65

6.2 ΔΙΑΧΩΡΙΣΜΟΣ ΒΑΣΕΙ ΤΥΠΟΥ ΑΝΤΙΚΑΤΑΣΤΑΣΗΣ.....67

6.2.1 Μέθοδοι μέσης τιμής (mean based methods) 67

6.2.2 Μέθοδοι παλινδρόμησης (regression based methods).....68

6.2.3 Hot – deck Imputation 69

6.3 ΔΙΑΧΩΡΙΣΜΟΣ ΒΑΣΕΙ ΤΗΣ ΜΟΝΤΕΛΟΠΟΙΗΣΗΣ.....72

6.4 MULTIPLE IMPUTATION.....72

6.4.1 Γενικά.....72

6.4.2 Τα βήματα του Πολλαπλού Καταλογισμού (Multiple Imputation) 75

6.4.3 Οι προϋποθέσεις που απαιτούνται για τη χρησιμοποίηση του πολλαπλού καταλογισμού.....78

6.4.4 Πλεονεκτήματα και Μειονεκτήματα του Πολλαπλού καταλογισμού.....80

6.4.5 Προγράμματα που εκτελούν την μέθοδο MI 82

6.5 ΣΥΜΠΕΡΑΣΜΑΤΑ 82

ΜΕΘΟΔΟΙ ΒΑΣΙΣΜΕΝΕΣ ΣΕ ΜΟΝΤΕΛΑ (MODEL BASED PROCEDURES) **84**

7.1 ΕΙΣΑΓΩΓΗ 84
7.2 ΕΚΤΙΜΗΣΗ ΜΕΓΙΣΤΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ 85
7.2.1 Εισαγωγή 85
7.2.2 Εκτίμηση Μέγιστης πιθανότητας σε περιπτώσεις με ελλιπή στοιχεία 87
7.2.3 Ο αλγόριθμος EM (Expectation Maximization) 88
7.2.4 Η μέθοδος FIML 92
7.2.5 Λογισμικά που διαθέτουν την Μέγιστη πιθανοφάνεια 93
7.3 BAYESIAN METHODS 95
7.3.1 Γενικά 95
7.3.2 Ο αλγόριθμος DA 96
7.3.3 Η μέθοδος Gibbs' Sampler 99
7.3.4 Πλεονεκτήματα – Μειονεκτήματα των Μπεϋζιανών Μεθόδων 101
7.4 ΣΥΜΠΕΡΑΣΜΑΤΑ 102

ΣΥΓΚΡΙΣΗ ΤΩΝ ΜΕΘΟΔΩΝ ΓΙΑ ΤΟΝ ΧΕΙΡΙΣΜΟ ΠΡΟΒΛΗΜΑΤΩΝ ΜΕ ΕΛΛΙΠΗ ΔΕΔΟΜΕΝΑ **104**

8.1 ΤΑ ΕΡΩΤΗΜΑΤΑ ΤΩΝ ΕΡΕΥΝΗΤΩΝ ΠΟΥ ΑΝΤΙΜΕΤΩΠΙΖΟΥΝ ΤΟ ΠΡΟΒΛΗΜΑ ΤΩΝ ΕΛΛΙΠΩΝ ΣΤΟΙΧΕΙΩΝ 105
8.2 ΣΥΝΟΠΤΙΚΗ ΠΑΡΟΥΣΙΑΣΗ ΤΩΝ ΜΕΘΟΔΩΝ 107

ΕΠΙΛΟΓΟΣ – ΣΥΜΠΕΡΑΣΜΑΤΑ **117**

ΒΙΒΛΙΟΓΡΑΦΙΑ **120**

1

Εισαγωγικές έννοιες

1.1 Εισαγωγή

Η στατιστική επεξεργασία και ανάλυση των δεδομένων επιτρέπει την καλύτερη εικόνα και επεξήγηση των φαινομένων που απασχολούν την κοινωνία, ενώ μπορεί να παρέχει σημαντικά συμπεράσματα. Η χρήση των κλασσικών στατιστικών μεθόδων απαιτεί την πλήρη ύπαρξη δεδομένων, ενώ δεν λαμβάνει υπόψη της περιπτώσεις όπου κάποια από τα δεδομένα είναι ελλιπή.

Όμως οι περιπτώσεις στις οποίες οι αναλυτές διαθέτουν όλα τα δεδομένα είναι πολύ λίγες, καθώς σχεδόν πάντα υπάρχουν απώλειες κατά τη συλλογή

και την επεξεργασία των δεδομένων. Το πρόβλημα γίνεται μεγαλύτερο όσο πιο μεγάλος είναι ο όγκος των δεδομένων και η συλλογή τους μακροχρόνια.

Η απώλεια αυτών των δεδομένων αποτελεί ένα κοινώς αναγνωρισμένο πρόβλημα που εμφανίζεται κατά τη στατιστική επεξεργασία και ανάλυση των δεδομένων και αναφέρεται με τον όρο ελλιπή στοιχεία (missing data). Τα ελλιπή στοιχεία (missing data) μειώνουν την ικανότητα των αναλυτών να κατανοήσουν και να εξηγήσουν τα φαινόμενα με τα οποία ασχολούνται, καθώς σχεδόν όλες οι κλασσικές στατιστικές μέθοδοι δεν αναφέρουν το πώς μπορούν να αντιμετωπιστούν τέτοιες περιπτώσεις.

Για το λόγο αυτό αναπτύχθηκαν κατά καιρούς διάφορες μέθοδοι με σκοπό την επεξεργασία δεδομένων με ελλιπή στοιχεία, οι οποίες δεν επηρεάζονται σε μεγάλο βαθμό από την έλλειψη των δεδομένων και μπορούν να εξάγουν έγκυρα συμπεράσματα για τα φαινόμενα τα οποία μελετούνται.

1.2 Ιστορική Εξέλιξη

Τα ελλιπή στοιχεία (missing data), όπως αναφέρθηκε και παραπάνω είναι ένα συχνό φαινόμενο κατά την στατιστική επεξεργασία δεδομένων σε πολλούς τομείς της κοινωνίας. Συγκεκριμένα το πρόβλημα αυτό εμφανίζεται κυρίως στις επιστήμες που ασχολούνται με τον άνθρωπο και μπαίνει η υποκειμενικότητα του, όπως η ιατρική, η ψυχολογία, η κοινωνιολογία, η βιολογία κ.α. Επίσης, μπορεί να εμφανιστεί στην αρχαιολογία και την παλαιοντολογία, όπου ο αρχαιολόγοι έρχονται συχνά αντιμέτωποι με ευρήματα τα οποία δεν είναι πλήρη. Τέλος μπορεί να εμφανιστεί και σε επιστήμες που ασχολούνται με πειραματικά δεδομένα, καθώς συχνά εμφανίζονται προβλήματα τόσο στον εξοπλισμό που χρησιμοποιείται, όσο και στη διαδικασία μέτρησης.

Τα ελλιπή στοιχεία είναι ένα πρόβλημα που απασχολεί του επιστήμονες από τις πρώτες δεκαετίες του 20ου αιώνα, όπου έγιναν προσπάθειες για την αντιμετώπιση του, χρησιμοποιώντας αλγοριθμικές και υπολογιστικές λύσεις. Μέχρι και την δεκαετία του 70 τα ελλιπή στοιχεία αντιμετωπιζόταν με μεθόδους διόρθωσης, δηλαδή συμπλήρωση ή διαγραφή των περιπτώσεων που είχαν ελλιπή στοιχεία.

Το 1987 ο Donald B. Rubin έγραψε ένα βιβλίο για τα ελλιπή δεδομένα, στο οποίο συμπεριέλαβε όλα τα συμπεράσματά που υπήρχαν τα ημιτελή στοιχεία, πρότεινε ένα τρόπο ταξινόμησής των μεθόδων που χρησιμοποιούνται για το χειρισμό τους και ανέπτυξε τους μηχανισμούς των ελλιπών δεδομένων (MAR, MCAR, MNAR), τα οποία θα αναπτυχθούν παρακάτω. Το βιβλίο αυτό θεωρείται σταθμό για την αντιμετώπιση του προβλήματος των ελλιπών στοιχείων και χρησιμοποιείται έως σήμερα.

Το 1977 διατυπώθηκε ο αλγόριθμος EM (Expectation – maximization) από τους Dempster, Laird και Rubin, ο οποίος άνοιξε νέους δρόμους στον τρόπο με τον οποίο γινόταν ο χειρισμός των ελλιπών στοιχείων, καθώς επέτρεψε τον υπολογισμό της εκτίμησης της μέγιστης πιθανότητας (maximum likelihood) σε πολλά προβλήματα με ελλιπή στοιχεία. Ο αλγόριθμος αυτός χρησιμοποιήθηκε σε μεγάλο βαθμό και το 1987 ο Little και Rubin εκδίδουν ένα βιβλίο στο οποίο αναφέρονται πολλά παραδείγματα του αλγόριθμου EM, ενώ παρατίθενται και τεκμηριώνονται τα μειονεκτήματα των προγενέστερων μεθόδων (διαγραφή ή συμπλήρωση των περιπτώσεων με ελλιπή στοιχεία).

Την ίδια χρονική περίοδο ο Rubin εισάγει την ιδέα του πολλαπλού καταλογισμού (Multiple imputation). Η μέθοδος αυτή αναπτύχθηκε αργότερα με την ανάπτυξη των υπολογιστών, καθώς απαιτεί αρκετά μεγάλη υπολογιστική δύναμη. Στα τέλη της δεκαετίας του 80 αρχίζουν να αναπτύσσεται η Μπεϋσιανή προσομοίωση (Bayesian simulation), ενώ ο

πολλαπλός καταλογισμός και η μέγιστη πιθανότητα κερδίζουν έδαφος, εξαιτίας του δωρεάν λογισμικού που προσφέρεται.

Στην δεκαετία του 90 η μέθοδος του επαναπροσδιορισμού των βαρών (reweighting) ή διαδικασίες στάθμισης που χρησιμοποιούταν από πολλούς ερευνητές στις έρευνες χρησιμοποιήθηκε και για την επεξεργασία δεδομένων με ελλιπή στοιχεία σε μοντέλα παλινδρόμησης με ελλιπής συμμεταβλητές (covariates). Το 1994 οι έρευνες στρέφονται στο χειρισμό των ελλιπών μεταβλητών αποφεύγοντας τη χρησιμοποίηση ολοκληρωμένων παραμετρικών μοντέλων για όλο τον πληθυσμό.

Τα τελευταία χρόνια οι επιστήμονες έχουν στραφεί στην αξιολόγηση της ευαισθησίας των αποτελεσμάτων στις διάφορες εναλλακτικές υποθέσεις για τον χειρισμό των ελλιπών στοιχείων.

1.3 Οι λόγοι της απώλειας των στοιχείων

Η απώλεια των στοιχείων μπορεί να γίνει σε όλα τα στάδια μιας έρευνας ή ενός πειράματος, από τη συλλογή μέχρι και την επεξεργασία των δεδομένων. Οι λόγοι για τους οποίους ένας ερευνητής μπορεί να βρει ελλιπή στοιχεία στα δεδομένα του είναι ποικίλοι και μπορεί να οφείλονται τόσο σε λάθος χειρισμούς, όσο και τυχαία γεγονότα τα οποία επηρέασαν την έκβαση της συλλογής των δεδομένων.

Όταν ένας ερευνητής ασχολείται με πειραματικές μελέτες, η έλλειψη κάποιων δεδομένων οφείλεται κατά κύριο λόγο στον ίδιο τον ερευνητή. Συγκεκριμένα κατά τη διάρκεια ενός πειράματος μπορεί να χαθούν δεδομένα όταν:

- Ένας ερευνητής λαμβάνει λάθος μετρήσεις λόγω απροσεξίας
- Ξεχνά να πάρει κάποιες μετρήσεις λόγω παράλειψης

- Δεν υπάρχουν οι κατάλληλες συνθήκες, ώστε να ληφθούν σωστές μετρήσεις, π.χ. σε περίπτωση μόλυνσης του δείγματος, όπου οι μετρήσεις πρέπει να απορριφθούν
- Έχει χαλάσει ο απαιτούμενος εξοπλισμός και είτε παρέχει λανθασμένα δεδομένα που τελικά απορρίπτονται, είτε αποφασίζεται να μην γίνει καν η μέτρηση.

Από την άλλη, στις έρευνες οι απώλειες που παρουσιάζονται κατά τη συλλογή των δεδομένων μπορούν να ταξινομηθούν σε τρεις μεγάλες κατηγορίες: σε αυτές που οφείλονται α) στους συμμετέχοντες στην έρευνα, β) στο σχεδιασμό της έρευνας και γ) στην αλληλεπίδραση των συμμετεχόντων και του σχεδιασμού της έρευνας.

- *Απώλειες που οφείλονται στους συμμετέχοντες στην έρευνα:* Πολλές φορές τα άτομα που συμμετέχουν σε μία έρευνα μπορεί να μην είναι πρόθημα να απαντήσουν σε όλες τις ερωτήσεις, είτε επειδή αυτές αφορούν προσωπικά δεδομένα, είτε επειδή είναι ερωτήσεις που θεωρούνται ταμπού (ηλικία, οικονομική κατάσταση). Επίσης, ορισμένοι μπορεί να προσπεράσουν κάποιες ερωτήσεις χωρίς να το καταλάβουν.

- *Απώλειες που οφείλονται στο σχεδιασμό της έρευνας:* Σε αυτή την κατηγορία ανήκουν απώλειες που οφείλονται σε λάθος διατύπωση των ερωτήσεων ή σε ερωτήσεις που είναι διφορούμενες. Συγκεκριμένα ορισμένες έρευνες συμπεριλαμβάνουν ερωτήσεις που δεν μπορούν να απαντηθούν από όλους τους συμμετέχοντες. Παραδείγματος χάριν στην περίπτωση που ζητούνται τα χρόνια έγγαμου βίου, τα άτομα που δεν είναι παντρεμένα δεν μπορούν να απαντήσουν.

- *Απώλειες που οφείλονται τόσο στους συμμετέχοντες όσο και στον σχεδιασμό της έρευνας:* Σε αυτή την κατηγορία ανήκουν έρευνες που

είναι χρονοβόρες, περιέχουν πολλές ερωτήσεις και οι συμμετέχοντες κουράζονται κατά τη συμπλήρωσή τους και παραλείπουν κάποιες ερωτήσεις ή αφήνουν το ερωτηματολόγιο ημισυμπληρωμένο.

Μια ιδιαίτερη κατηγορία έρευνας αποτελούν οι μακροχρόνιες έρευνες στις οποίες τα ελλιπή στοιχεία οφείλονται στους παραπάνω λόγους, αλλά επιπλέον μπορεί να οφείλονται και σε άλλους λόγους, οι οποίοι είναι αποτέλεσμα των πολλών σταδίων που έχει μία τέτοια έρευνα σε διάφορες χρονικές περιόδους:

- Κάποιοι συμμετέχοντες αρχίζουν να συμμετέχουν σε μία έρευνα, αλλά στην πορεία φεύγουν καθώς είτε δεν τους ενδιαφέρει ποια αυτή η έρευνα, είτε μπορεί να υπάρχουν λόγοι υγείας ή θανάτου, όπως συμβαίνει στις περιπτώσεις ιατρικών ερευνών που αναφέρονται σε ανίατες αρρώστιες, όπως καρκίνο, ή όταν πρόκειται για άτομα μεγάλης ηλικίας.

- Άλλοι συμμετέχοντες αρχίζουν να ασχολούνται με μία έρευνα, αλλά αποφασίζουν να εγκαταλείψουν την έρευνα για προσωπικού λόγους και μετά αποφασίζουν να επανέλθουν σε αυτή.

- Ορισμένοι συμμετέχοντες μπορεί να αλλάξουν κατοικία και να αμελήσουν να ενημερώσουν τους υπεύθυνους ή για κάποιο άλλο λόγο να μην μπορούν να εντοπιστούν.

Επίσης δεδομένα μπορούν να χαθούν και κατά τα πρώτα στάδια της ανάλυσης, όπου δημιουργούνται πίνακες, οι οποίοι περιέχουν τα προς επεξεργασία δεδομένα λόγο απροσεξίας του αναλυτή ή επειδή αδυνατεί να καταλάβει τι εννοεί ο συμμετέχοντας.

1.4 Οι συνέπειες της έλλειψης των δεδομένων

Η στατιστική επεξεργασία επιτρέπει στους επιστήμονες να φτάσουν σε κάποια συμπεράσματα για το θέμα το οποίο τους απασχολεί. Όταν τα δεδομένα δεν είναι πλήρη, αλλά υπάρχουν και ελλιπή στοιχεία προκαλούνται δύο αρνητικές συνέπειες.

Η πρώτη αρνητική επίπτωση που συνεπάγεται η έλλειψη κάποιων δεδομένων είναι η μείωση της δύναμης της ανάλυσης. Κατά τους Verma και Goodale (1995), η στατιστική δύναμη οφείλεται σε τρεις παράγοντες, το επίπεδο σημαντικότητας του δείγματος, το αποτελεσματικό μέγεθος και το μέγεθος του δείγματος. Στην πράξη από τους τρεις παράγοντες χρησιμοποιείται μόνο ο ένας, το μέγεθος του δείγματος. Έτσι στην περίπτωση που αντί για *n* παρατηρήσεις που είχαν προγραμματιστεί να χρησιμοποιηθούν για τη στατιστική επεξεργασία, υπάρχουν λιγότερες, τότε η έλλειψη αυτών των παρατηρήσεων μειώνει το δείγμα, άρα και τη δύναμη της ανάλυσης και των αποτελεσμάτων.

Οι στατιστικές αποκλίσεις που μπορούν να προκληθούν από την έλλειψη κάποιων παρατηρήσεων αποτελούν τη δεύτερη αρνητική επίπτωση κατά τη στατιστική επεξεργασία των δεδομένων. Τα συστηματικά σφάλματα που μπορεί να προκύψουν από την παρουσία ελλιπών στοιχείων εμφανίζονται με διάφορους τρόπους.

- Αρχικά μπορεί να επηρεαστούν οι δείκτες τις κεντρικής τάσης (Measures of central tendency), είτε προς τα πάνω, είτε προς τα κάτω, ανάλογα με την κατανομή που εμφανίζουν τα ελλιπή στοιχεία.

- Οι δείκτες της διασποράς μπορεί, επίσης να επηρεαστούν ανάλογα με το μέρος της κατανομής στο οποίο παρατηρούνται τα ελλιπή δεδομένα

- Τέλος τα ελλιπή στοιχεία μπορεί να προκαλέσουν προς τα κάτω απόκλιση στους συντελεστές συσχέτισης. Η προς τα κάτω απόκλιση θεωρείται σχεδόν βέβαιη, καθώς οι υψηλές και χαμηλές τιμές που χάνονται περιορίζουν τη διακύμανση μίας μεταβλητής και μειώνουν το συσχετισμό της με μια άλλη μεταβλητή.

Από τα παραπάνω προκύπτει ότι τα ελλιπή δεδομένα είναι σχεδόν αδύνατο να μην εμφανιστούν κατά τη διάρκεια μιας έρευνας και δεν είναι δυνατόν να μην απασχολήσουν τους αναλυτές, καθώς θα προκαλέσουν τα προβλήματα που αναφέρθηκαν και τα αποτελέσματα δεν θα είναι έγκυρα.

2

Το Πρόβλημα των Ελλιπών Στοιχείων (Missing Data Problem)

2.1 Το πρόβλημα των ελλιπών στοιχείων

Τα ελλιπή δεδομένα αποτελούν συχνή περίπτωση κατά τη διαδικασία συλλογής των δεδομένων, ειδικά όταν η συλλογή αφορά μεγάλο όγκο δεδομένων. Για την επεξεργασία και την ανάλυση των συλλεχθέντων δεδομένων, οι περισσότερες στατιστικές μέθοδοι χρησιμοποιούν ορθογωνική διάταξη των δεδομένων σε πίνακες.

Η δημιουργία των ορθογώνιων πινάκων Y αποτελεί το πρώτο βήμα της στατιστικής επεξεργασίας. Οι πίνακες είναι διαστάσεων $n x p$, και οι σειρές

τους αναπαριστούν τις μονάδες (άτομα) και ονομάζονται περιπτώσεις (cases) και οι στήλες αναπαριστούν τις μεταβλητές που μετρούνται (variables) για κάθε περίπτωση.

Σε αυτό το στάδιο γίνεται εμφανές το μέγεθος του προβλήματος που αντιμετωπίζει ο αναλυτής, καθώς μπορεί να έχει μία πιο πλήρη εικόνα των δεδομένων που συλλέχθηκαν, του ποσοστού των ελλιπών στοιχείων και του πως κατανέμονται τα ελλιπή στοιχεία, π.χ. αν παρουσιάζονται σε συγκεκριμένες ερωτήσεις.

Ιδανική περίπτωση θα αποτελούσε ένας πίνακας που να διέθετε δεδομένα για όλες τις μεταβλητές, δηλαδή θα παρατηρούνταν όλα τα δεδομένα

$$Y = (Yobs)$$

Όμως αυτό είναι σχεδόν αδύνατο να συμβεί στις πραγματικές διαδικασίες συλλογής δεδομένων. Παρόλα αυτά οι πίνακες πρέπει να διαθέτουν τα πλήρη στοιχεία της έρευνας, δηλαδή να περιέχει τόσο τα παρατηρούμενα στοιχεία, όσο και τα ελλιπή στοιχεία. Άρα κάθε πίνακας Y (πλήρη στοιχεία) μπορεί να γραφεί:

$$Y = (Yobs, Ymis)$$

Κάθε στοιχείο Y του πίνακα μπορεί είτε να είναι διαθέσιμο, είτε να λείπει, δηλαδή το y_{ij} που είναι η τιμή της μεταβλητής j για το στοιχείο i. Ορίζεται ένας πίνακας R, ίδιων διαστάσεων με τον πίνακα Y, ο οποίος αναφέρεται ως πίνακας έλλειψης (missingness) και παίρνει την τιμή 1 όταν το y_{ij} παρατηρείται και την τιμή 0 όταν το y_{ij} λείπει:

$$R_{ij} = \begin{cases} 1, y_{ij} _ \pi\alpha\rho\alpha\tau\eta\rho\varepsilon\iota\tau\alpha\iota \\ 0, y_{ij} _ \delta\varepsilon\nu _ \pi\alpha\rho\alpha\tau\eta\rho\varepsilon\iota\tau\alpha\iota \end{cases}$$

Ο πίνακας έλλειψης (R) βοηθάει τους ερευνητές να δουν ευκολότερα το ποσοστό των ελλιπών δεδομένων, καθώς και αν τα ελλιπή δεδομένα παρουσιάζουν κάποια μορφή, δηλαδή αν συγκεντρώνονται σε κάποια ορισμένη μεταβλητή (univariate nonresponse) ή σε συγκεκριμένες μεταβλητές (multivariate nonreponse) ή σε κάποιο συγκεκριμένο άτομο (unit nonresponse). Επίσης μπορούν να ληφθούν συμπεράσματα για το αν τα ελλιπή στοιχεία σχετίζονται με κάποιο γεγονός που αφορά των μηχανισμό με τον οποίο παρουσιάζονται τα ελλιπή δεδομένα (MAR, MCAR, MNAR), π.χ. αν αφορούν κάποια συγκεκριμένη μεταβλητή μπορεί αυτή να είναι κάποια ερώτηση, την οποία δεν θέλουν να αποκαλύψουν οι συμμετέχοντες ή είναι τυχαία. Οι μορφές και οι μηχανισμοί των ελλιπών δεδομένων θα παρουσιαστούν αναλυτικά παρακάτω.

2.2 Οι μορφές των ελλιπών δεδομένων (missing data patterns)

Η δημιουργία των πινάκων επιτρέπει στους αναλυτές να εκτιμήσουν τα δεδομένα που έχουν και να αποφασίσουν με ποιο τρόπο να τα επεξεργαστούν. Όπως αναφέρθηκε και παραπάνω, είναι σχεδόν ακατόρθωτο σε μία έρευνα που περιλαμβάνει μεγάλο αριθμό δεδομένων να μην υπάρχουν ελλιπή δεδομένα.

Σε αυτό το σημείο κρίνεται ενδιαφέρον και απαραίτητο να παρουσιαστούν οι μορφές των ελλιπών στοιχείων, οι οποίες μπορούν να περιγράψουν ποια στοιχεία λείπουν και ποια στοιχεία υπάρχουν σε ένα πίνακα και τον μηχανισμό των ελλιπών δεδομένων (MCAR, MAR, MNAR) ο οποίος ασχολείται με τη σχέση της έλλειψης των δεδομένων και των μεταβλητών που εξετάζονται.

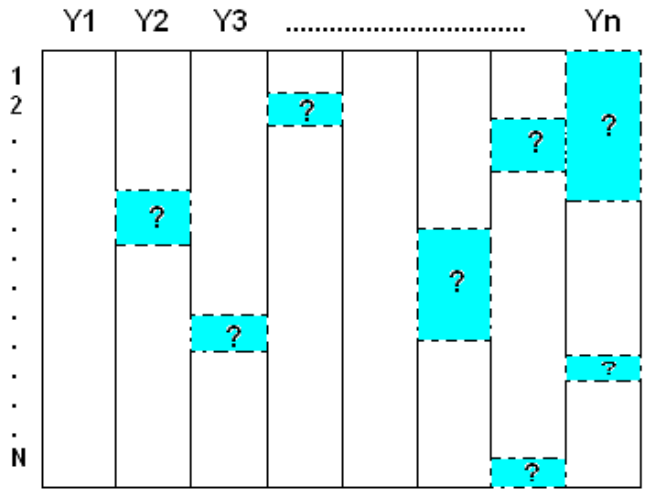
Οι μορφές των ελλιπών στοιχείων βοηθούν την επιλογή της μεθόδου που πρέπει να χρησιμοποιηθεί, ώστε η επεξεργασία των στοιχείων να γίνει με

σωστό τρόπο. Έτσι θα ήταν καλό να ταξινομούνται οι σειρές και οι στήλες των πινάκων που διαθέτουν δεδομένα της έρευνας με τέτοιο τρόπο, ώστε να ελεγχθεί αν υπάρχει κάποια συγκεκριμένη μορφή στα ελλιπή δεδομένα.

Οι κύριες μορφές των ελλιπών στοιχείων, σύμφωνα με τους Little και Rubin (2002) διακρίνονται σε έξι κατηγορίες οι οποίες είναι:

2.2.1 Γενική μορφή ή μη μονότονη (General)

Σε αυτήν την περίπτωση οι ελλιπής παρατηρήσεις είναι "τυχαίες" και δεν έχουν κάποια συγκεκριμένη μορφή. Δηλαδή, μπορεί να υπάρχουν σε ένα ή περισσότερα σημεία μίας ή περισσοτέρων μεταβλητών. Η διάταξη αυτής της μορφής φαίνεται στο σχήμα 1.



Σχήμα 1:
Γενική Μορφή ή μη μονότονη έλλειψης

Αυτή η μορφή των ελλιπών δεδομένων εμφανίζεται σε έρευνες όταν τα ερωτηματολόγια ενός ή περισσοτέρων συμμετεχόντων διαθέτουν

κάποιες μη συμπληρωμένες ερωτήσεις και ονομάζεται Item nonresponse.

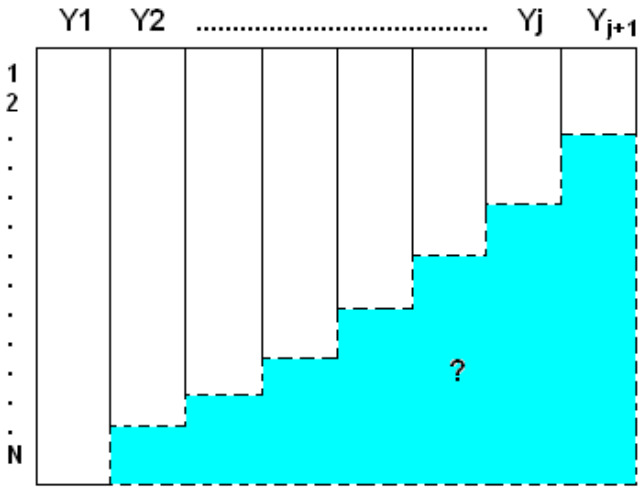
Η αντιμετώπιση αυτής της μορφής των ελλιπών στοιχείων γίνεται κυρίως με μεθόδους καταλογισμού (Imputation methods) ή αντικατάστασης, όπου κάθε στοιχείο που λείπει "αποδίδεται" (αντικαθίσταται) με μία λογική αξία, η οποία προκύπτει από τις διαθέσιμες πληροφορίες.

Επίσης, στην περίπτωση που τα ελλιπή δεδομένα αποτελούν μικρό ποσοστό των συνολικών δεδομένων μπορεί να χρησιμοποιηθούν και μέθοδοι διαγραφής (complete case analysis), οι οποίες μπορεί να εμφανίσουν και αποκλίσεις στην περίπτωση που τα ελλιπή στοιχεία σχετίζονται και με άλλα χαρακτηριστικά και δεν είναι "τυχαία".

2.2.2 Μονότονη μορφή (Monotone)

Η μονότονη μορφή των ελλιπών στοιχείων εμφανίζεται στις μακροχρόνιες έρευνες, όπου οι ερευνητές συλλέγουν στοιχεία από μία ομάδα συγκεκριμένων ανθρώπων για ένα μεγάλο χρονικό διάστημα, κάτω από διαφορετικές πειραματικές συνθήκες. Οι ελλιπής παρατηρήσεις οφείλονται σε συμμετέχοντες που αρχικά συμφώνησαν να συμμετέχουν στην έρευνα και στην πορεία είτε έχασαν το ενδιαφέρον τους και αποσύρθηκαν από αυτή είτε η ερευνητική ομάδα έχασε την επαφή μαζί τους λόγω αλλαγής κατοικίας ή ακόμα και θανάτου.

Στο σχήμα 2 φαίνεται η περίπτωση της μονότονης μορφής των ελλιπών στοιχείων, στην οποία οι μεταβλητές του πίνακα μπορούν να τακτοποιηθούν με τέτοιο τρόπο, ώστε όλα τα στοιχεία που λείπουν για την μεταβλητή Y_j να λείπουν και για την μεταβλητή Y_{j+1}.

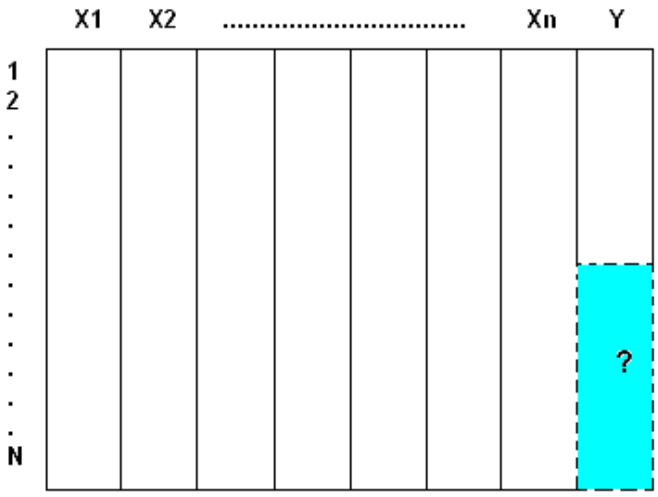


Σχήμα 2:
Μονότονη Μορφή έλλειψης

Στην πράξη είναι δύσκολο να διαθέτουμε δεδομένα με τέτοια μορφή, αλλά υπάρχουν αρκετές περιπτώσεις που τα ελλιπή στοιχεία πλησιάζουν αυτή την μορφή. Η διαχείριση ελλιπών δεδομένων μονότονης μορφής είναι πιο εύκολη από τα ελλιπή δεδομένα γενικής μορφής και τις τελευταίες δεκαετίες υπάρχει μεγάλο ενδιαφέρον για την διαχείριση αυτών των δεδομένων.

2.2.3 Μονομεταβλητό ελλείπον στοιχείο (Univariate nonresponse)

Σε αυτή την περίπτωση όλα τα ελλιπή δεδομένα εμφανίζονται σε μία και μόνο μεταβλητή (Y), όπως φαίνεται και στο σχήμα 3. Σε ένα σύνολο μεταβλητών, οι μεταβλητές $X_1,\ldots,X_n$ διαθέτουν όλες τις παρατηρήσεις, ενώ η μεταβλητή Y έχει και ελλιπή δεδομένα.



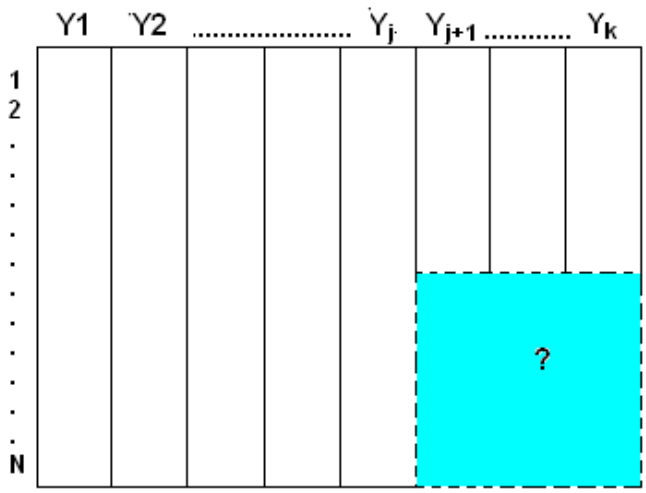
Σχήμα 3:
Μονο-μεταβλητή μορφή έλλειψης

Αυτή η μορφή έλλειψης δεδομένων εμφανίζεται κυρίως στα σχεδιασμένα πειράματα, όπου οι περισσότερες μεταβλητές διαθέτουν όλα τα δεδομένα, ενώ μία μπορεί να της λείπουν και κάποια δεδομένα λόγω λανθασμένων χειρισμών κατά την μέτρηση των στοιχείων της.

Αποτελεί την πιο εύκολη μορφή έλλειψης ως προς την αντιμετώπιση της. Συνήθως χρησιμοποιούνται κλασσικές μέθοδοι όπως οι μέθοδος των ελαχίστων τετραγώνων, ενώ μπορεί εύκολα να αντιμετωπιστεί και με τις μεθόδους που παρουσιάζονται παρακάτω.

2.2.4 Πολυμεταβλητό ελλείπον στοιχείο (Multivariate two patterns)

Η περίπτωση αυτή περιλαμβάνει περισσότερες από μία μεταβλητές, οι οποίες εμφανίζουν ελλιπή δεδομένα σε συγκεκριμένες περιπτώσεις (σειρές). Δηλαδή, όπως φαίνεται και στο σχήμα 4 οι μεταβλητές Y_1 – Y_j έχουν όλες της παρατηρήσεις, ενώ οι μεταβλητές Y_{j+1} – Y_k έχουν ημισμπληρωμένα στοιχεία.



Σχήμα 4:
Πολυμεταβλητή μορφή έλλειψης

Στις έρευνες αυτό ονομάζεται μοναδιαία μη ανταπόκριση (unit nonresponse) και οφείλεται στο γεγονός ότι ορισμένα άτομα που είναι σχεδιασμένο να λάβουν μέρος σε μία έρευνα είτε αρνούνται να συμμετέχουν, είτε δεν είναι δυνατόν να εντοπιστούν από τη ερευνητική ομάδα.

Η αντιμετώπιση του προβλήματος μπορεί να γίνει ακόμα και στη διαδικασία συλλογής των δεδομένων, όπου μπορεί να μειωθούν τα ελλιπή στοιχεία με επαναλαμβανόμενες προσπάθειες συλλογής δεδομένων (Madow et al, 1983), τη χρησιμοποίηση αντιπροσώπων των επιλεγμένων ατόμων (Moore, 1988), ή ακόμα και η σχεδιασμένη συμμετοχή περισσότερων ατόμων (Madow et al, 1983).

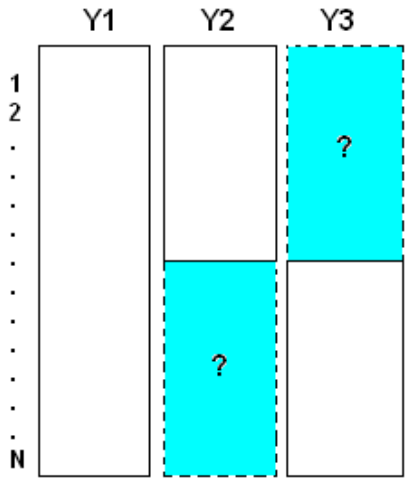
Επίσης μπορεί να οφείλεται και στον τρόπο σχεδίασης της έρευνας, όπου λόγω κάποιων πρακτικών λόγων να μην γίνουν κάποιες μετρήσεις, π.χ. για οικονομικούς λόγους. Σε αυτή την περίπτωση έχουν προταθεί αποτελεσματικές μέθοδοι για την εκτίμηση των παραμέτρων με συγκεκριμένα μοντέλα παλινδρόμησης, τα οποία χρησιμοποιούν τις διαθέσιμες σε κάθε στάδιο πληροφορίες. Ενώ για την αντιμετώπιση τέτοιας μορφής ελλιπή δεδομένα που δεν

οφείλονται στη σχεδίαση χρησιμοποιούνται κυρίως μέθοδοι Reweighting.

2.2.5 File Matching

Κάποιες φορές είναι επιθυμητό να συνδυάσουμε διαφορετικές ομάδες δεδομένων με σκοπό τη βελτίωση των αποτελεσμάτων της ανάλυσής μας. Για να επιτευχθεί αυτό πρέπει να συσχετιστούν με κάποιες κοινές μεταβλητές που διαθέτουν όλα τα δεδομένα. Ωστόσο κάποιες μεταβλητές δεν διαθέτουν όλα τα δεδομένα. Συγκρίνοντας λοιπόν κάποιες από τις μεταβλητές που παρουσιάζουν ελλιπή δεδομένα μπορεί να παρατηρήσουμε ότι δύο ή περισσότερες μεταβλητές δεν έχουν καμία κοινή παρατήρηση.

Όπως φαίνεται και στο σχήμα 5 η μεταβλητή Y_1 δεν έχει κανένα ελλιπές στοιχείο, η μεταβλητή Y_2 διαθέτει τις πρώτες παρατηρήσεις, ενώ η μεταβλητή Y_3 τις τελευταίες παρατηρήσεις και δεν έχουν κανένα κοινό σημείο.



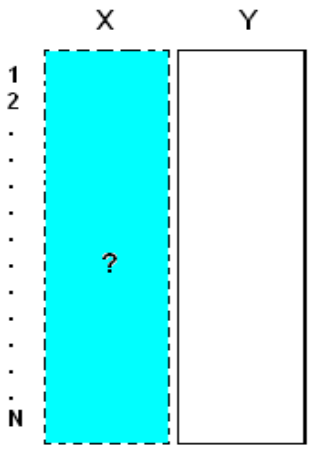
Σχήμα 5:
File Matching

Το file matching είναι μία τεχνική με βάση την οποία όταν υπάρχουν μεταβλητές που δεν έχουν κανένα κοινό στοιχείο (σχήμα 5, μεταβλητές Y2 και Y3) μπορούν να συνδυαστούν με μεταβλητές που έχουν κοινά στοιχεία (μεταβλητή Y1) και να συμπληρωθούν τα δεδομένα που λείπουν. Η τεχνική αυτή ονομάζεται και statistical matching.

2.2.6 Factor Analysis

Η παραγοντική ανάλυση (factor analysis) χρησιμοποιείται κυρίως από τους ερευνητές των κοινωνικών επιστημών και της ανθρώπινης συμπεριφοράς, σε προβλήματα όπου σημαντικές μεταβλητές δεν μπορούν να μετρηθούν απευθείας (εξυπνάδα, πολιτική τοποθέτηση, κοινωνικοοικονομική κατάσταση), με αποτέλεσμα να μην παρατηρούνται.

Με την παραγοντική ανάλυση γίνεται προσπάθεια να συνδεθούν αυτές οι μη παρατηρούμενες μεταβλητές (παράγοντες ή συνιστώσες), με μεταβλητές που παρατηρούνται και για τις οποίες υπάρχουν μετρήσεις, επιτυγχάνοντας με αυτόν τον τρόπο και μια ομαδοποίηση των παρατηρούμενων μεταβλητών σε κοινές συνιστώσες.



Σχήμα 6:
Factor Analysis

Στο σχήμα 6 φαίνεται αυτή η μορφή έλλειψης δεδομένων. Όπως φαίνεται η μεταβλητή Χ θεωρείται ως μεταβλητή που διαθέτει μόνο ελλιπή στοιχεία, ενώ η μεταβλητή Υ παρατηρείται πλήρως. Με την παραγοντική ανάλυση (Factor Analysis) πραγματοποιείται μία πολυμεταβλητή παλινδρόμηση μίας μεταβλητής Υ που παρατηρείται πλήρως, ως προς κάποιους μη παρατηρούμενους συντελεστές Χ που λείπουν πλήρως, για την διερεύνηση της σχέσης των δύο αυτών μεταβλητών, ώστε να μπορέσουν να συμπληρωθούν τα ελλιπή στοιχεία.

2.3 Μηχανισμοί των ελλιπών στοιχείων

Η κατανόηση των μηχανισμών των ελλιπών στοιχείων αποτελεί θέμα ζωτικής σημασίας για την επιλογή της μεθοδολογίας διαχείρισης των δεδομένων με ελλιπή στοιχεία, καθώς οι διάφορες μεθοδολογίες επιλύουν το πρόβλημα στηριζόμενες σε διαφορετικές προϋποθέσεις. Με άλλα λόγια ο προσδιορισμός του μηχανισμού των ελλιπών στοιχείων, μπορεί να βοηθήσει τους αναλυτές να επιλέξουν τη βέλτιστη μέθοδο για την αντιμετώπιση του προβλήματος των ελλιπών στοιχείων.

Ο όρος μηχανισμός ελλιπών στοιχείων έχει να κάνει με το αν η έλλειψη των στοιχείων που παρουσιάζεται σε ένα σύνολο δεδομένων έχει βαθύτερα αίτια ή όχι. Έτσι διακρίνουμε τρεις περιπτώσεις. Η έλλειψη μπορεί να είναι τελείως τυχαία, τυχαία υπό προϋποθέσεις ή μπορεί να οφείλεται σε συστηματικούς λόγους.

Σε κάθε σύνολο δεδομένων με ελλιπή στοιχεία είναι πιθανόν να υπάρχουν περισσότεροι από ένα μηχανισμό ελλιπών στοιχείων και γι' αυτό οι αναλυτές πρέπει να ερευνήσουν προσεκτικά τους πιθανούς λόγους έλλειψης

των δεδομένων πριν καταλήξουν στην επιλογή της βέλτιστης μεθόδου διαχείρισή τους.

Την έννοια του μηχανισμού των ελλιπών στοιχείων εισήγαγε για πρώτη φορά οι Rubin το 1987 και χρησιμοποιείται μέχρι και σήμερα. Σύμφωνα με αυτή την κατηγοριοποίηση υπάρχουν τρία είδη μηχανισμών, δεδομένα με α) εντελώς τυχαία έλλειψη (Missing Completely At Random, MCAR) β) τυχαία έλλειψη (Missing At Random, MAR) και γ) όχι τυχαία έλλειψη (Missing Not At Random, MNAR).

Ορίζεται ότι τα ολοκληρωμένα δεδομένα περιέχονται σε ένα πίνακα Y, με $Y=(Y_{ij})$ και τα ελλιπή δεδομένα περιέχονται σε ένα πίνακα $R=(R_{ij})$. Οι μηχανισμοί των ελλιπών στοιχείων χαρακτηρίζονται από την υποθετική διανομή του R στο Y, $f(R/Y,\varphi)$, όπου φ δηλώνει τις άγνωστες παραμέτρους. Επίσης ορίζεται ότι Y_{obs} είναι οι τιμές του Y οι οποίες παρατηρούνται και Ymis αυτές που λείπουν ($Y = (Y_{obs}, Y_{mis})$).

Ο Rubin (1987) ό ρισε ό τι αν η έλλειψη εξαρτάται μό νο από τις παρατηρούμενες τιμές και δεν εξαρτάται από τα ελλιπή στοιχεία, δηλαδή αν:

$$f(R / Y, \varphi) = f(R / Y_{obs}, \varphi), \text{ για όλα τα } Y_{mis}, \varphi \qquad (1)$$

τότε τα ελλιπή στοιχεία ακολουθούν των μηχανισμό MAR, Missing at random – η έλλειψή τους είναι τυχαία.

Μια ειδική περίπτωση του μηχανισμού MAR, είναι ο μηχανισμός MCAR (Missing Completely At Random), εντελώς τυχαία έλλειψη στοιχείων, η οποία εμφανίζεται όταν η έλλειψη δεν εξαρτάται, ούτε από τα παρατηρούμενα στοιχεία (Yobs), ούτε από τα ελλιπή στοιχεία (Ymis), αλλά οφείλονται σε τυχαία γεγονότα. Σε αυτή την περίπτωση ισχύει η σχέση:

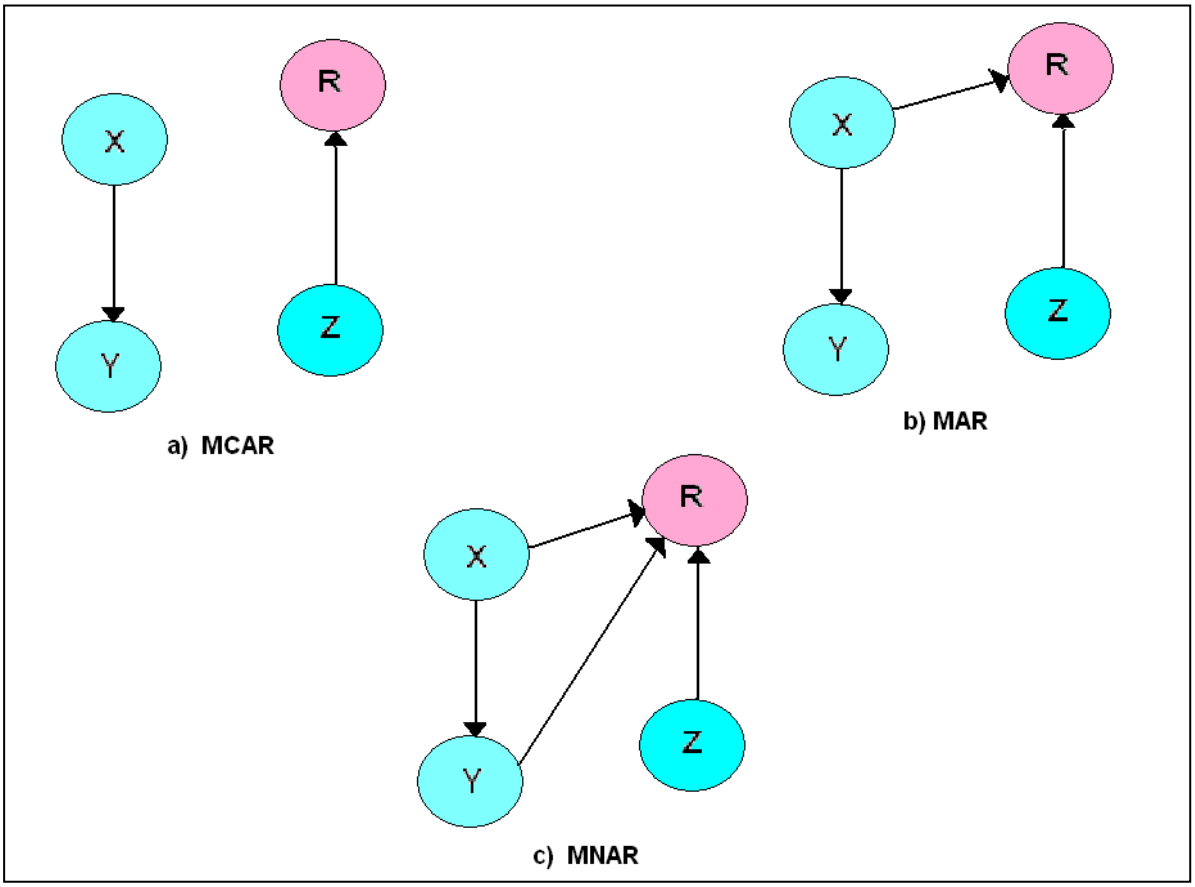
$$f(R / Y, \varphi) = f(R / \varphi), \text{ για όλα τα } Y, \varphi \qquad (2)$$

Τέλος, όταν παραβιάζεται η εξίσωση (1) και η διανομή εξαρτάται από τα ελλιπή στοιχεία Ymis, η έλλειψη θεωρείται ότι δεν είναι τυχαία, δηλαδή τα στοιχεία δεν λείπουν τυχαία και ακολουθούν τον μηχανισμό MNAR.

Οι τρεις αυτοί μηχανισμοί θα μπορούσαν να παρασταθούν και γραφικά, ώστε να γίνει πιο κατανοητή η ερμηνεία τους. Στην περίπτωση που μονο-μεταβλητή έλλειψη, έστω ότι X είναι μεταβλητές οι οποίες διαθέτουν πλήρη δεδομένα, Y η μεταβλητή που έχει και ελλιπή δεδομένα. Το R αναπαριστά την έλλειψη και το Z αναπαριστά αιτίες έλλειψης οι οποίες δεν σχετίζονται με τις μεταβλητές X, Y.

Στο σχήμα 7 που ακολουθεί φαίνεται η σχέση μεταξύ των X, Y, Z, R για τους τρεις μηχανισμούς έλλειψης.

Όπως φαίνεται λοιπόν όταν η έλλειψη των στοιχείων είναι τελείως τυχαία (MCAR) το R (έλλειψη) δεν σχετίζεται ούτε με τις μεταβλητές X, ούτε με την μεταβλητή Y, αλλά μόνο με τυχαίες αιτίες (Z). Όταν η έλλειψη των στοιχείων θεωρείται τυχαία (MAR), το R (έλλειψη) σχετίζεται και με το τόσο με τυχαίες αιτίες, όσο και με τις μεταβλητές που παρατηρούνται (X). Τέλος όταν η έλλειψη των στοιχείων δεν είναι τυχαία (MNAR), η έλλειψη σχετίζεται άμεσα με όλες της μεταβλητές, οφείλεται δηλαδή και στην ίδια την μεταβλητή που παρουσιάζει τα ελλιπή στοιχεία και στις άλλες μεταβλητές και σε τυχαίες αιτίες.



Σχήμα 7:
Γραφική Απεικόνιση Μηχανισμών Έλλειψης

Αναλυτικότερα, όταν τα στοιχεία είναι MAR, τα ημιτελή στοιχεία που προκύπτουν δεν προκύπτουν από τις ίδιες τις ελλιπείς αξίες, αλλά η έλλειψη των δεδομένων είναι μια λειτουργία μερικών άλλων παρατηρηθέντων μεταβλητών μέσα στο σύνολο των στοιχείων. Με άλλα λόγια η έλλειψη εξαρτάται από άλλες μεταβλητές για τις οποίες τα στοιχεία παρατηρούνται. Για να γίνει κατανοητό έστω ότι σε ένα ερωτηματολόγιο υπάρχει μία ερώτηση για την ηλικία των συμμετεχόντων. Οι γυναίκες είναι πιο πιθανό να αφήσουν αναπάντητη αυτή την ερώτηση σε σχέση με τους άντρες. Σε αυτή την περίπτωση λοιπόν τα ελλιπή στοιχεία εξαρτώνται από το φύλο του συμμετέχοντα, το οποίο έχει ερωτηθεί παραπάνω και έχει απαντηθεί.

Ομοίως, στις διαχρονικές μελέτες, η βασική γραμμή συνδιακύμανσης και οι εκβάσεις μπορούν συχνά να χρησιμοποιηθούν για να συναντήσουν τη MAR υπόθεση για τους συμμετέχοντες που χάνονται στην πορεία της έρευνας. Παραδείγματος χάριν, το γένος και το βάρος συμμετεχόντων που μετριέται στο πρώτο χρόνο μπορούν να συσχετιστούν με την επόμενη μέτρηση για το αν αυτή λείπει ή όχι, καθώς οι γυναίκες αποφεύγουν σε μεγαλύτερο βαθμό να αναφέρουν το βάρος τους.

Όπως ειπώθηκε και παραπάνω ο μηχανισμός MCAR είναι μία ειδική περίπτωση του μηχανισμού MAR. Ο μηχανισμός αυτός εμφανίζεται όταν η πιθανότητα της έλλειψης είναι ανεξάρτητη και από τις παρατηρηθείσες μεταβλητές, δηλαδή εκείνες τις μεταβλητές για τις οποίες είναι διαθέσιμα όλα τα στοιχεία και από τις μεταβλητές που παρουσιάζουν απώλεια τιμών, δηλαδή εκείνες οι μεταβλητές που είτε δεν έχουν κανένα στοιχείο, είτε έχουν ορισμένα ελλιπή στοιχεία.

Ένα παράδειγμα στο οποίο κυριαρχεί ο μηχανισμός MCAR εμφανίζεται όταν ένας από τους συμμετέχοντες μιας έρευνας δεν επιστρέφει για την συνέχιση της έρευνας για λόγους ανεξάρτητους από την μελέτη, δηλαδή για κάποιους τυχαίους λόγους, όπως ένα επαγγελματικό ταξίδι, ή κάποια αρρώστια, θανάτος ή ακόμα και λόγο κακοκαιρίας. Ένα άλλο παράδειγμα εμφανίζεται όταν μια δυσλειτουργία στους υπολογιστές αποκλείει την ανάκτηση των στοιχείων για ορισμένους συμμετέχοντες, αλλά όχι για κάποιους άλλους.

Όπως φαίνεται και από τα παραδείγματα αυτές οι περιστάσεις δεν εμφανίζονται εντελώς τυχαία, αλλά οι λόγοι της εμφάνισής τους δεν έχουν καμία σχέση με το θέμα της μελέτης ή με τις διαδικασίες που ακολουθούνται, οπότε θεωρείται ότι δεν επηρεάζουν την μελέτη.

Τέλος, τα στοιχεία που ακολουθούν τον μηχανισμό MNAR είναι στοιχεία που δεν λείπουν τυχαία, αλλά δεν συμπληρώνονται εξαιτίας κάποιων συγκεκριμένων λόγων, κάτι που μπορεί να δημιουργήσει προβλήματα στους ερευνητές κατά την επιλογή της μεθόδου επεξεργασία των στοιχείων. Αυτό συμβαίνει επειδή τα στοιχεία που λείπουν, λείπουν εσκεμμένα και αυτό μπορεί να επηρεάσει τα αποτελέσματα. Με άλλα λόγια η μη ανταπόκριση των συμμετεχόντων οφείλεται στην ίδια την ερώτηση την οποία δεν επιθυμούν να απαντήσουν.

Ένα συχνό φαινόμενο είναι η απάντηση σε ερωτήσεις που αφορούν ευαίσθητα προσωπικά δεδομένα, για παράδειγμα οι άνθρωποι που ζυγίζουν πολλά κιλά μπορεί να μην συμπληρώνουν μία ερώτηση σχετική με το βάρος τους. Κατά συνέπεια αν η μη αποκάλυψη του βάρους ενός συμμετέχοντα είναι σχετική με το βάρος του και δεν εξαρτάται από τις άλλες μεταβλητές που παρατηρούνται, η έλλειψη ακολουθεί τον μηχανισμό των μη τυχαίων ελλιπών στοιχείων.

Τα στοιχεία που είναι MNAR είναι δεν μπορούν να αγνοηθούν επειδή οι κλασσικές μέθοδοι ανάλυσης στοιχείων δεν μπορούν να λάβουν υπόψη τα ελλιπή στοιχεία που προκύπτουν από έναν μηχανισμό έλλειψης MNAR. Έτσι στις περιπτώσεις που τα ελλιπή στοιχεία θεωρούνται ότι μπορεί να οφείλονται σε μη τυχαία έλλειψη, οι αναλυτές θα πρέπει να ασχοληθούν με αυτά τα στοιχεία σαν να αποτελούν ειδικό τμήμα των στοιχείων που είναι για ανάλυση. Μία τέτοια διαμόρφωση στηρίζεται σε πολλές και αυστηρές υποθέσεις για το ποιος είναι ο λόγος που δημιουργήθηκαν τα ελλιπή στοιχεία κάτι που είναι δύσκολο και χρονοβόρο να γίνει. Έτσι αντί να γίνει αυτό είναι καλύτερο να ληφθούν πρόσθετα μέτρα κατά το σχεδιασμό, ώστε να περιοριστεί όσο το δυνατόν περισσότερο η πιθανότητα να παρουσιαστούν ελλιπή στοιχεία που να ακολουθούν μηχανισμό MNAR.

2.4 Διερεύνηση τυχαίας έλλειψης των δεδομένων

Όπως έχει αναφερθεί και παραπάνω η διερεύνηση της τυχαιότητας των ελλιπών στοιχείων αποτελεί ένα σημαντικό θέμα για την επιλογή της μεθόδου που θα χρησιμοποιηθεί. Σύμφωνα με τους Little και Rubin (1987) έχουν αναπτυχθεί δύο μέθοδοι για να διαγνωσθεί αν τα στοιχεία λείπουν τυχαία ή όχι.

Η πρώτη μέθοδος αξιολογεί τα ελλιπή στοιχεία για μια μεταβλητή με τη δημιουργία δύο ομάδων, στην μία από τις οποίες περιέχονται τα ελλιπή στοιχεία και στην άλλη όλες οι παρατηρούμενες τιμές για τη συγκεκριμένη μεταβλητή.

Εάν η μορφή των δύο ομάδων της συγκεκριμένης μεταβλητής έχει σημαντικές διαφορές σε σχέση με τις ομάδες των άλλων μεταβλητών ενδιαφέροντος, τότε θεωρείται ότι τα ελλιπή στοιχεία δεν λείπουν τυχαία. Σύμφωνα με αυτή την μέθοδο ο ερευνητής πρέπει να εξετάσει διάφορες μεταβλητές για να ελέγξει αν προκύπτει οποιοδήποτε συνεπές σχέδιο. Πολλές φορές μπορεί να εμφανιστούν και κάποιες τυχαίες διαφορές, όμως οποιαδήποτε διαφορά εμφανιστεί μπορεί να προειδοποιήσει για μη τυχαία έλλειψη των στοιχείων.

Η δεύτερη μέθοδος ελέγχει τον συσχετισμό της έλλειψης των στοιχείων για κάθε ζευγάρι μεταβλητών. Για κάθε μεταβλητή, τα παρατηρούμενα στοιχεία αντικαθίστανται από την τιμή, ενώ τα ελλιπή στοιχεία αντικαθίστανται από την μηδέν και δημιουργούνται οι πίνακες έλλειψης για κάθε μεταβλητή. Κατόπιν συσχετίζονται για κάθε ζευγάρι μεταβλητών οι παραπάνω πίνακες και προκύπτει ο βαθμός συσχέτισης μεταξύ των ελλιπών στοιχείων για κάθε ζευγάρι. Ο χαμηλός βαθμός συσχέτισης δείχνει τυχαία έλλειψη δεδομένων ανά ζευγάρι μεταβλητών.

Αν και δεν υπάρχει κάποιο όριο στον βαθμό συσχέτισης σύμφωνα με το οποίο να προσδιορίζεται αν τα δεδομένα λείπουν τυχαία ή όχι, οι στατιστικές δοκιμές συσχέτισης παρέχουν συντηρητικές εκτιμήσεις. Αν για όλα τα ζευγάρια μεταβλητών δεν υπάρχει μεγάλος βαθμός συσχέτισης τότε ο αναλυτής μπορεί να υποθέσει ότι τα δεδομένα λείπουν τυχαία, δηλαδή ακολουθούν μηχανισμό MCAR. Αν όμως μεταξύ ορισμένων ζευγαριών υπάρχουν ισχυρότεροι βαθμοί συσχετισμού μπορεί να υποτεθεί ότι τα δεδομένα ακολουθούν τον μηχανισμό MAR.

3

Μέθοδοι διαχείρισης δεδομένων με ελλιπή στοιχεία – Ταξινόμηση

Το πρόβλημα των ελλιπών στοιχείων απασχολεί όλο τον επιστημονικό κλάδο, καθώς καθημερινά όλοι έρχονται αντιμέτωποι με δεδομένα τα οποία δεν είναι πλήρη. Ο λάθος χειρισμός αυτών των δεδομένων οδηγεί σε λάθος αποτελέσματα και κατ' επέκταση στην εξαγωγή μη έγκυρων συμπερασμάτων.

Τις τελευταίες δεκαετίες έχει γίνει μία μεγάλη προσπάθεια να βρεθούν τρόποι αντιμετώπισης αυτού του προβλήματος και όλο και περισσότεροι επιστήμονες στρέφονται προς αυτή την κατεύθυνση. Ως αποτέλεσμα τις

τελευταίες δεκαετίες έχουν αναπτυχθεί πληθώρα μεθοδολογιών που προσπαθούν να χειριστούν δεδομένα τα οποία δεν είναι πλήρη.

Ο μεγάλος αριθμός των μεθοδολογιών που έχουν αναπτυχθεί οδήγησε στην ανάγκη για την ταξινόμηση τους σε ομάδες, ώστε να είναι πιο εύκολη η μελέτη τους. Σήμερα κυριαρχούν δύο τρόποι ταξινόμησης των μεθοδολογιών αυτών.

Ο πρώτος τρόπος έχει να κάνει με τον χρόνο που αναπτύχθηκαν και μπορούν να διακριθούν δύο ομάδες, α) οι "παραδοσιακές" μεθοδολογίες (Traditional Methods) και οι "μοντέρνες" μεθοδολογίες (Modern Methods). Ο δεύτερος τρόπος έχει να κάνει με τον τρόπο που χειρίζονται τα ελλιπή στοιχεία και σε αυτό το διαχωρισμό μπορούν να διακριθούν τέσσερις κατηγορίες, α) μέθοδοι διαγραφής (Deletion Methods), β) μέθοδοι βαρών (Weighting Methods), γ) μέθοδοι καταλογισμού (Imputation Methods) και δ) μέθοδοι που στηρίζονται σε μοντέλα (Model Based Methods).

3.1 Παραδοσιακές – Μοντέρνες Μέθοδοι

Οι παραδοσιακές τεχνικές έχουν ως σκοπό την "διόρθωση" των ελλιπών δεδομένων πριν από την ανάλυση τους, είτε με τη διαγραφή των ημιτελών δεδομένων, είτε με την συμπλήρωση των ελλιπών στοιχείων. Αυτές οι μέθοδοι, δεν θεωρούνται αποτελεσματικές καθώς έχουν τα επιθυμητά αποτελέσματα μόνο υπό αυστηρές προϋποθέσεις και τείνουν να εμφανίζουν συστηματικά σφάλματα όταν τα ελλιπή στοιχεία δεν ακολουθούν τον μηχανισμό της τυχαίας έλλειψης (MAR). Επίσης θεωρούνται λιγότερο ισχυρές από τις "μοντέρνες" μεθόδους (Modern Methods).

Οι "παραδοσιακές" μέθοδοι μπορεί να χωριστούν σε δύο επιμέρους κατηγορίες ανάλογα με τον τρόπο που προσπαθούν να "διορθώσουν" τα

ημιτελή στοιχεία, δηλαδή σε μεθόδους διαγραφής (Deletion Methods) και σε μεθόδους καταλογισμού (Imputation Methods) ή αντικατάστασης.

Οι κυριότερες μεθοδολογίες που ανήκουν σε αυτή την κατηγορία είναι:

- Mean Imputation Methods
- Regression Imputation Methods
- Hot deck Imputation Methods
- Complete case analysis
- Available case Analysis

Οι "μοντέρνες" μεθοδολογίες έχουν αναπτυχθεί τα τελευταία 30 χρόνια και θεωρούνται ότι έχουν ελκυστικό θεωρητικό υπόβαθρο, καθώς για να εφαρμοστούν απαιτούνται λιγότερες και πιο ρεαλιστικές προϋποθέσεις από ότι στις "παραδοσιακές" μεθοδολογίες. Ωστόσο αν δεν εφαρμοστούν με προσοχή ή αν δεν ικανοποιούνται κάποιες από τις απαιτούμενες προϋποθέσεις, ειδικά αυτές που έχουν σχέση με τους μηχανισμούς που ακολουθούν τα ελλιπή στοιχεία μπορεί τα αποτελέσματα να μην είναι τα επιθυμητά.

Σε αυτή την κατηγορία ανήκουν δύο μεθοδολογίες: ο πολλαπλός καταλογισμός (Multiple Imputation) και η μέγιστη πιθανότητα (Maximum Likelihood). Οι μεθοδολογίες αυτές δεν είναι καινούριες, αλλά έχουν λάβει μεγάλη προσοχή εξαιτίας του ελκυστικού θεωρητικού υπόβαθρου τους και της μεγάλης ποικιλίας λογισμικών που είναι διαθέσιμα (NORM, AMOS, LISREL 8.5).

Οι δύο αυτές μέθοδοι αναπτύχθηκαν περίπου την ίδια περίοδο και από κάποιους θεωρούνται ανταγωνίστηκες μέθοδοι. Όμως έχουν πολλά κοινά σημεία και μάλιστα υπό ορισμένες προϋποθέσεις παράγουν παρόμοια συμπεράσματα.

3.2 Ταξινόμηση μεθόδων με βάση τον χειρισμό των ελλιπών στοιχείων

Εκτός από τον παραπάνω τρόπο ομαδοποίησης των μεθοδολογιών που έχουν αναπτυχθεί για την αντιμετώπιση των δεδομένων με ελλιπή στοιχεία, στην βιβλιογραφία αναφέρεται ευρέως ο διαχωρισμός των μεθόδων με βάση τον τρόπο που χειρίζονται τα στοιχεία.

Σύμφωνα με αυτή την ομαδοποίηση διακρίνονται τέσσερις ομάδες μεθοδολογιών οι οποίες είναι: α) Μέθοδοι που βασίζονται σε πλήρη δεδομένα ή μέθοδοι διαγραφής (Deletion Methods), β) Διαδικασίες Στάθμισης (Weighting Procedures), γ) μέθοδοι αιτίασης (Imputation methods) ή διαδικασίες αντικατάστασης (replacement procedures) και δ) μέθοδοι που βασίζονται σε μοντέλα (model – Based Methods).

3.2.1 Μέθοδοι Διαγραφής (Deletion Methods)

Στις περιπτώσεις όπου κάποιες μεταβλητές δεν παρατηρούνται για κάποια άτομα του δείγματος, ένας απλός τρόπος αντιμετώπισης του προβλήματος είναι η διαγραφή αυτών των ατόμων και η ανάλυση μόνο των ατόμων που διαθέτουν στοιχεία για όλες τις παρατηρήσεις.

Αυτός ο τρόπος αντιμετώπισης θα συζητηθεί και θα αναλυθεί στο κεφάλαιο 4, όπου θα αναπτυχθούν οι μέθοδοι που ανήκουν σε αυτή την κατηγορία, καθώς και τα αρνητικά και τα θετικά κάθε μεθόδου.

Γενικά όμως οι μέθοδοι διαχείρισης που ανήκουν σε αυτή την κατηγορία πιστεύεται ότι είναι απλοί και εύκολοι να εφαρμοστούν, αλλά ότι παρουσιάζουν σημαντικές αποκλίσεις και μειώνουν τη δύναμη της στατιστικής επεξεργασίας, καθώς μειώνεται ο πληθυσμός

των δεδομένων που επεξεργάζονται. Τέτοιες μέθοδοι παρουσιάζουν ικανοποιητικά αποτελέσματα μόνο σε περιπτώσεις όπου ο αριθμός των ελλιπών στοιχείων δεν είναι μεγάλος.

3.2.2 Διαδικασίες Στάθμισης (Weighting Procedures)

Στις περισσότερες δειγματοληπτικές έρευνες, τα βάρη σχετίζονται με κάθε ένα άτομο ή ομάδα ατόμων του δείγματος και χρησιμοποιούνται κατά την ανάλυση για να παράγουν όσον το δυνατό εκτιμήσεις χωρίς αποκλίσεις για τις παραμέτρους του πληθυσμού στόχου.

Αυτά τα βάρη αντισταθμίζουν το γεγονός ότι τα στοιχεία του δείγματος μπορεί να έχουν επιλεγεί με διαφορετικά ποσοστά (π.χ. σε μια έρευνα να υπάρχουν περισσότεροι άντρες από ότι γυναίκες) ή ότι κάποια άτομα μπορεί να είναι πιο πιθανό να απαντήσουν σε μία ερώτηση από ότι άλλα. Έτσι η διαδικασία στάθμισης προσπαθεί τροποποιήσει τα βάρη του δείγματος, ώστε τα ελλιπή στοιχεία να αποτελούν μέρος του σχεδιασμένου δείγματος. Ο κύριος στόχος των βαρών είναι να μειώσουν τις αποκλίσεις των εκτιμήσεων των ερευνών, θεωρώντας ότι κάθε άτομο αντιπροσωπεύει ένα διαφορετικό μέρος του πληθυσμού στόχου.

Ωστόσο η ανάλυση των δεδομένων με τη χρησιμοποίηση των βαρών δημιουργεί και ορισμένα μειονεκτήματα. Πρώτον η στάθμιση μπορεί να περιπλέξει την ανάλυση των δεδομένων, καθώς πολλά τυποποιημένα στατιστικά πακέτα, είτε δεν αναγνωρίζουν τα βάρη, είτε τα αντιμετωπίζει ως τιμές των παρατηρήσεων. Έτσι στην τελευταία περίπτωση το άθροισμα των βαρών που αντιστοιχεί σε όλο το δείγμα, οδηγεί σε λάθος εκτιμήσεις στην ακρίβεια των εκτιμήσεων.

Επίσης η στάθμιση μπορεί να μεγαλώσει τις διαφορές των εκτιμήσεων. Συγκεκριμένα ο Kish δείχνει ότι όταν υπάρχουν διαφορικοί όροι η ρύθμιση της στάθμισης αυξάνει τη διασπορά των εκτιμήσεων κατά (1 + L), όπου L η σχετική διασπορά των βαρών. Στην ουσία η στάθμιση μειώνει την απόκλιση αυξάνοντας τη διασπορά. Λαμβάνοντας υπόψη την αντιστάθμιση απόκλισης – διασποράς, θα ήταν χρήσιμο να μπορούσαν να ελεγχθούν τα μέσα τετραγωνικά λάθη των εναλλακτικών εκτιμητών κάτι που μπορεί να γίνει σπάνια επειδή οι αποκλίσεις των στοιχείων είναι δύσκολο να υπολογιστούν.

3.2.3 Μέθοδοι Απόδοσης (Imputation methods)

Σε αυτή την κατηγορία ανήκουν οι μεθοδολογίες που έχουν ως στόχο την αντικατάσταση των ελλιπών στοιχείων με κατάλληλες τιμές που προέρχονται από τα άλλα στοιχεία (απαντήσεις) που έχουν καταγραφεί για το συγκεκριμένο άτομο ή άλλες πληροφορίες που μπορεί να είναι γνωστές και επιτρέπουν την συμπλήρωση του δεδομένου που λείπει. Οι μεθοδολογίες αυτής της κατηγορίας θα αναλυθούν παρακάτω στο κεφάλαιο 6.

Είναι σημαντικό να αναφερθεί ότι οι ερευνητές πρέπει να είναι πολύ προσεκτικοί με την αντικατάσταση των ελλιπών στοιχείων καθώς η λάθος αντικατάσταση μπορεί έχει ως αποτέλεσμα λάθος συμπεράσματα. Επίσης δεν πρέπει να γίνεται εκτεταμένη αντικατάσταση τιμών, καθώς η αντικατάσταση των ελλιπών στοιχείων δεν αντισταθμίζει μία ελλιπή σχεδιασμένη διαδικασία ή φτωχή συλλογή δεδομένων.

Γενικά οι διαδικασίες αντικατάστασης μπορούν να χρησιμοποιηθούν σε ορισμένες περιπτώσεις, εφόσον ο ερευνητής είναι σε θέση να το

κάνει. Είναι εύκολο να εκτελεσθούν, και μερικές συμπεριλαμβάνονται ως επιλογές στα στατιστικά πακέτα.

Το σημαντικότερο πλεονέκτημα αυτών των διαδικασιών είναι η διατήρηση του μεγέθους δειγμάτων και συνεπώς, της στατιστικής δύναμης και σε επόμενες αναλύσεις. Σε μεγαλύτερη ή μικρότερη έκταση, όλες οι διαδικασίες αντικατάστασης παρουσιάζουν αποκλίσεις όταν δεν υπάρχει τυχαία διανομή των ελλιπών στοιχείων (MAR μηχανισμός). Εντούτοις, η αντικατάσταση του ελλείποντος στοιχείου είναι κατάλληλη όταν δεν υπάρχει μεγάλος συσχετισμός μεταξύ των μεταβλητών.

3.2.4 Μέθοδοι που βασίζονται σε μοντέλα (Model – Based Methods).

Η κατηγορία αυτή περιλαμβάνει ένα μεγάλο αριθμό μεθόδων, οι οποίες δημιουργούνται από τον καθορισμό ενός μοντέλου για τα παρατηρούμενα δεδομένα και τα συμπεράσματα βασίζονται στην πιθανότητα ή σε μεταγενέστερη κατανομή αυτού του μοντέλου με παραμέτρους που εκτιμώνται με διαδικασίες όπως η μέγιστη πιθανότητα.

Το πλεονέκτημα αυτών των μεθόδων είναι η ευελιξία που δίνει στον αναλυτή. Η αποφυγή χρησιμοποίησης μεθόδων που εξειδικεύονται σε μία συγκεκριμένη περίπτωση στις προϋποθέσεις αυτού του μοντέλου τονίζει ότι το μοντέλο που παράγεται μπορεί να εκτιμηθεί.

4

Μέθοδοι Διαγραφής (Deletion Procedures)

Σε αυτό το κεφάλαιο θα παρουσιαστούν οι Μέθοδοι Διαγραφής (Deletion Procedures), οι οποίες αποτελούν το πιο απλό τρόπο διαχείρισης δεδομένων που διαθέτουν ελλιπή στοιχεία και χρησιμοποιούνται συχνά κατά την στατιστική επεξεργασία δεδομένων τα οποία δεν είναι πλήρη. Σε αυτή την κατηγορία ανήκουν δύο μέθοδοι, οι οποίες είναι:

- Η ανάλυση πλήρως συμπληρωμένων περιπτώσεων, Complete Case Analysis ή Listwise Deletion και
- Διαθέσιμη περιπτωσιολογική Ανάλυση, Available Case Analysis ή Pairwise Deletion

4.1 Η Ανάλυση πλήρως συμπληρωμένων περιπτώσεων, Complete Case Analysis ή Listwise Deletion

Η διαγραφή περίπτωσης (case deletion) ή ανάλυση πλήρως συμπληρωμένων περιπτώσεων (Complete Case Analysis) αποτελεί τη δημοφιλέστερη από τις παλαιότερες (παραδοσιακές) μεθόδους που χρησιμοποιούνται για την επεξεργασία δεδομένων με ελλιπή στοιχεία. Η μέθοδος αυτή χρησιμοποιείται σε πολλά στατιστικά πακέτα, καθώς θεωρείται βολική και απλή. Παρόλα αυτά η χρησιμοποίηση της είναι εξειδικευμένη και οι λεπτομέρειες διαφέρουν ανάλογα την περίπτωση.

Η μέθοδος αυτή ασχολείται μόνο με τις περιπτώσεις που διαθέτουν όλα τα δεδομένα για όλες τις μεταβλητές, διαγράφοντας τις περιπτώσεις που διαθέτουν ένα ή περισσότερα ελλιπή στοιχεία. Για παράδειγμα αν κάθε ερωτούμενος μίας έρευνας άφηνε από μία ερώτηση αναπάντητη, τότε το πακέτο λογισμικού θα απέρριπτε όλα τα στοιχεία και θα προειδοποιούσε τον αναλυτή ότι δεν υπάρχουν έγκυρες περιπτώσεις.

Όπως φαίνεται και από το παραπάνω παράδειγμα η μέθοδος αυτή θυσιάζει πολύ εύκολα μεγάλο αριθμό στοιχείων. Σύμφωνα με τους Kim και Carry (1977) η τυχαία διαγραφή του 10% των δεδομένων για κάθε μία από 5 μεταβλητές μπορεί εύκολα να οδηγήσει στη διαγραφή του 59% των περιπτώσεων του δείγματος. Επίσης ο Kauffmann (1988) αναφέρει περίπτωση όπου χρησιμοποιήθηκε αυτή η μέθοδος και από ένα δείγμα 624 έμειναν 201 περιπτώσεις.

Τέλος σύμφωνα με τους Rubin και Little (2002) η απώλεια των δεδομένων γίνεται ακόμα μεγαλύτερη στην περίπτωση όπου υπάρχουν πολλές μεταβλητές. Για παράδειγμα αν υπάρχουν 20 μεταβλητές και κάθε μεταβλητή παρουσιάζει έλλειψη δεδομένων στο 10% των περιπτώσεων, τότε αναμένεται να διατηρηθεί περίπου το 13% των παρατηρούμενων δεδομένων

κάτι που μειώνει πολύ την ακρίβεια των αποτελεσμάτων και τη στατιστική τους δύναμη.

4.1.1 Ιδιότητες

Η διαγραφή περίπτωσης (Case Deletion) δίνει έγκυρα αποτελέσματα όταν τα ελλιπή στοιχεία ακολουθούν τον μηχανισμό MCAR, ενώ σε ορισμένες περιπτώσεις δίνει ικανοποιητικά αποτελέσματα ακόμα και όταν τα ελλιπή στοιχεία ακολουθούν μηχανισμό MAR. Για παράδειγμα όταν εμφανίζεται Univariate μορφή έλλειψης (σχήμα 3) οι παράμετροι παλινδρόμησης του Y σε οποιαδήποτε υποσύνολο $X_1,\ldots,X_p$ μπορεί να υπολογιστεί με την μέθοδος των πλήρως συμπληρωμένων περιπτώσεων (Complete Case Analysis) και οι εκτιμήσεις να είναι έγκυρες ακόμα και για μηχανισμό MAR.

Ωστόσο αν τα ελλιπή στοιχεία ακολουθούν τον μηχανισμό MNAR τα αποτελέσματα που προκύπτουν από αυτή τη μέθοδο εμφανίζουν αποκλίσεις επειδή οι πλήρως συμπληρωμένες περιπτώσεις με βάση τις οποίες έγινε η ανάλυση μπορεί να μην αποτελούν αντιπροσωπευτικό δείγμα του πληθυσμού. Η απόκλιση όταν ακολουθείται ο μηχανισμός MNAR μπορεί να είναι μικρή αλλά μπορεί να είναι και ασήμαντη. Όμως δεν υπάρχει κάποιος τρόπος για να μετρηθεί η σημαντικότητα της.

4.1.2 Πλεονεκτήματα – Μειονεκτήματα

Όπως αναφέρθηκε και παραπάνω η ανάλυση των πλήρως συμπληρωμένων περιπτώσεων (Complete Case Analysis) ασχολείται μόνο με τις περιπτώσεις που παρατηρούνται όλες οι μεταβλητές (δεν παρουσιάζουν ελλιπή στοιχεία), διαγράφοντας τις περιπτώσεις που έχουν ελλιπή στοιχεία.

Τα πλεονεκτήματα αυτής της μεθόδου μπορούν να συνοψιστούν στα παρακάτω:

- Η μέθοδος είναι πολύ απλή και μπορούν να εφαρμοστούν όλες οι στατιστικές αναλύσεις, χωρίς καμία τροποποίηση των δεδομένων και
- Τα στατιστικά αποτελέσματα μπορούν να συγκριθούν, καθώς όλα είναι υπολογισμένα ως προς μία κοινή βάση δείγματος.

Όπως όλες οι μέθοδοι έτσι και αυτή παρουσιάζει και αρνητικά στοιχεία. Τα μειονεκτήματα της προέρχονται κυρίως από την απώλεια των δεδομένων που έχουν απορριφθεί. Αυτή η απώλεια των πληροφοριών που κρύβουν τα απορριφθέντα δεδομένα έχει δύο διαστάσεις. Πρώτον χάνεται η ακρίβεια της επεξεργασίας και δεύτερον εμφανίζεται στατιστική απόκλιση όταν τα ελλιπή δεδομένα δεν ακολουθούν τον μηχανισμό MCAR και οι πλήρως συμπληρωμένες περιπτώσεις δεν αποτελούν τυχαίο δείγμα όλων των περιπτώσεων.

4.1.3 Συμπεράσματα

Η ανάλυση των πλήρως συμπληρωμένων περιπτώσεων μπορεί να νομιμοποιηθεί στο όνομα της απλότητας της μόνο στην περίπτωση που η απώλεια της ακρίβειας και η στατιστική απόκλιση είναι ελάχιστη. Με άλλα λόγια στην περίπτωση που οι πληροφορίες που θα προέκυπταν από τις ημιτελείς περιπτώσεις είναι μηδαμινές και δεν επηρεάζουν τα συμπεράσματα της έρευνας. Αυτό συμβαίνει στην περίπτωση που η αναλογία των συμπληρωμένων περιπτώσεων είναι μεγάλη σε σχέση με τις ημι-συμπληρωμένες.

Για την εφαρμογή αυτής της μεθόδου πρέπει να έχει εξασφαλιστεί ότι τα ελλιπή δεδομένα ακολουθούν τον μηχανισμό MCAR, ή κατά περίπτωση τον μηχανισμό MAR, ενώ δεν προτείνεται όταν τα στοιχεία δεν λείπουν τυχαία δηλαδή ακολουθείται ο MNAR μηχανισμός.

Επίσης, είναι σημαντικό να εξασφαλιστεί ότι τα δεδομένα που λείπουν αποτελούν ένα μικρό ποσοστό και δεν θα επηρεάσουν τα αποτελέσματα της έρευνας. Είναι συνετό λοιπόν να χρησιμοποιείται μόνο όταν τα δεδομένα που λείπουν είναι ελάχιστα.

4.2 Διαθέσιμη Περιπτωσιολογική Ανάλυση (Available Case Analysis) ή Pairwise Deletion

Η Ανάλυση των πλήρως συμπληρωμένων περιπτώσεων έχει ως αποτέλεσμα την απώλεια μεγάλου αριθμού στοιχείων, μειώνοντας τη στατιστική δύναμη των αποτελεσμάτων. Έτσι αναπτύχθηκε μία εναλλακτική μέθοδος διαγραφής γνωστή ως διαθέσιμη περιπτωσιολογική ανάλυση (Available Case Analysis). Η μέθοδος αυτή όπως και η Listwise Deletion βρίσκεται σε πολλά στατιστικά πακέτα και χρησιμοποιείται αρκετά συχνά.

4.2.1 Ιδιότητες

Η μέθο δ ο αυτή, σε αντίθεση με την Complete Case Analysis, χρησιμοποιεί όλα τα διαθέσιμα στοιχεία (όλα τα στοιχεία που παρατηρούνται για κάθε μεταβλητή). Για να το επιτύχει αυτό κάνει επιμέρους συσχετισμούς μεταξύ των μεταβλητών, διαγράφοντας κάθε φορά μόνο την περίπτωση που παρουσιάζει ελλιπή στοιχεία για εκείνο τον συσχετισμό. Για παράδειγμα αν ένας συμμετέχοντας δεν έχει απαντήσει για τη μεταβλητή Α, δεν συμμετέχει στους συσχετισμούς της

μεταβλητής Α, αλλά συμμετέχει στους συσχετισμούς της μεταβλητής Β, Γ, κ.ο.κ.

Όπως γίνεται αντιληπτό από τα παραπάνω οι παράμετροι υπολογίζονται από διαφορετικά σύνολα του δείγματος. Τα διαφορετικά σύνολα του δείγματος που χρησιμοποιούνται δυσκολεύουν τον υπολογισμό των κοινών σφαλμάτων ή άλλων μέτρων αβεβαιότητας, καθώς οι αναλυτικές μέθοδοι θέλουν πολύ χρόνο και κόπο.

Επίσης επειδή οι συσχετισμοί, οι συνδιακυμάνσεις υπολογίζονται για διαφορετικά υποσύνολα του δείγματος μπορεί τα αποτελέσματα που θα προκύψουν να είναι ασυμβίβαστα από μαθηματικής άποψης (π.χ. συνδιακύμανση που να μην ανήκει στο διάστημα [-1, 1]). Αυτό μπορεί με τη σειρά του να έχει αρνητικά αποτελέσματα αν η επεξεργασία των στοιχείων γίνει με προγράμματα που στηρίζονται στη μέθοδο της μέγιστης πιθανότητας (maximum Likelihood) όπως τα προγράμματα LISREL, AMOS, EQS.

Τέλος πρέπει να αναφερθεί ότι, όπως και η Complete Case Analysis, δίνει ικανοποιητικά αποτελέσματα μόνο στην περίπτωση που ακολουθείται ο μηχανισμός MCAR, ενώ για τους άλλους μηχανισμούς τα αποτελέσματα είναι ανακριβή.

4.2.2 Πλεονεκτήματα – Μειονεκτήματα

Το κύριο πλεονέκτημα της Available Case Analysis είναι ότι χρησιμοποιεί όλα τα διαθέσιμα στοιχεία του δείγματος, αυξάνοντας τη στατιστική δύναμη των συμπερασμάτων.

Όμως για να μπορέσει να χρησιμοποιήσει όλα τα παρατηρούμενα δεδομένα αλλάζει τα υποσύνολα του δείγματος που χρησιμοποιεί για κάθε υπολογισμό, κάτι που αποτελεί το μεγάλο της μειονέκτημα. Η

μεταβλητότητα του υποσυνόλου του δείγματος ως προς το οποίο γίνονται οι υπολογισμοί μπορεί να δημιουργήσει πολλά πρακτικά προβλήματα όταν υπάρχουν μεγάλα σύνολα δεδομένων. Το πρόβλημα γίνεται ακόμα μεγαλύτερο όταν οι πίνακες έχουν υπολογιστεί για διάφορες κατηγορίες (όλες οι γυναίκες, οι παντρεμένες, οι ανύπαντρες κ.ο.κ.).

4.2.3 Συμπεράσματα

Η μέθοδος αυτή χρησιμοποιείται κυρίως για να αυξήσει τη στατιστική δύναμη των αποτελεσμάτων σε σχέση με την Complete Case Analysis καθώς χρησιμοποιεί όλα τα δεδομένα που παρατηρούνται. Θεωρείται πιο περίπλοκη στους υπολογισμούς, ενώ μπορεί να δημιουργήσει και σοβαρά προβλήματα στον αναλυτή κυρίως λόγω των διαφορετικών υποσυνόλων του δείγματος που χρησιμοποιούνται για τους υπολογισμούς. Η μέθοδος αυτή έχει ικανοποιητικά αποτελέσματα μόνο για ελλιπή στοιχεία που λείπουν τελείως τυχαία (MCAR), ενώ δεν προτείνεται όταν αυτά ακολουθούν τους άλλους μηχανισμούς (MAR, MNAR).

5

Διαδικασίες Στάθμισης (Weighting Procedures)

Οι διαδικασίες στάθμισης χρησιμοποιούν κάποια βάρη με σκοπό την μείωση της απόκλισης των εκτιμώμενων τιμών. Τα βάρη αντισταθμίζουν τη διαφορετική πιθανότητα παρατήρησης της τιμής για τις περιπτώσεις του δείγματος (σε έρευνες κάποια άτομα μπορεί να έχουν μεγαλύτερη πιθανότητα να απαντήσουν σε μία ερώτηση, ή σε πειράματα ανάλογα με τις συνθήκες κάποιες μετρήσεις μπορεί να είναι πιο πιθανό να παρατηρηθούν). Με άλλα λόγια, αυτή τη διαδικασία λαμβάνει υπόψη τα ελλιπή στοιχεία και τα αντιμετωπίζει ως μέρος του σχεδιασμένου πειράματος.

Οι διαδικασίες στάθμισης χρησιμοποιούνται συνήθως σε συνδυασμό με την διαγραφή περίπτωσης (case deletion) όταν τα ελλιπή δεδομένα δεν λείπουν

εντελώς τυχαία (MCAR). Μετά την απομάκρυνση των περιπτώσεων που παρουσιάζουν ελλιπή δεδομένα, οι περιπτώσεις που μένουν σταθμίζονται, ώστε η διανομή τους να πλησιάζει τη διανομή που θα είχαν αν δεν υπήρχαν ελλιπή στοιχεία.

5.1 Τα στάδια της Διαδικασίας Στάθμισης

Οι διαδικασίες στάθμισης περιλαμβάνουν τρία στάδια:

- ***Στάδιο 1ο:*** Στο πρώτο στάδιο παράγεται μία βάση βαρών, η οποία προκύπτει από τη διαφορετική πιθανότητα επιλογής των στοιχείων του δείγματος. Η βάση τίθεται ίση ή ανάλογη με το αντίστροφο της πιθανότητας επιλογής των στοιχείων: $w_i = \frac{1}{\pi_i}$, όπου π_i η πιθανότητα επιλογής του στοιχείου i.

- ***Στάδιο 2ο:*** Στο δεύτερο στάδιο η βάση των βαρών που έχει δημιουργηθεί προσαρμόζεται, ώστε να αντισταθμιστούν τα ελλιπή στοιχεία που υπάρχουν στο σύνολο των δεδομένων. Η πιο συνηθισμένη μέθοδος για την προσαρμογή των βαρών είναι η Weighting Class Adjustment (Προσαρμογή σταθμισμένων ομάδων). Με αυτή την μέθοδο το δείγμα χωρίζεται σε σταθμισμένες ομάδες και πολλαπλασιάζεται κάθε στοιχείο της βάση των βαρών με το αντίστροφο της πιθανότητα παρατήρησης αυτής της ομάδας.

- ***Στάδιο 3ο:*** Το τρίτο στάδιο ρυθμίζει περαιτέρω τα βάρη των παρατηρούμενων τιμών, έτσι ώστε να εκτιμηθούν ορισμένες τιμές για το σύνολο του πληθυσμού με την προσαρμογή κάποιων γνωστών τιμών για τα σύνολα αυτά. Ο κύριος στόχος αυτής της προσαρμογής είναι η μείωση της στατιστικής απόκλισης εξαιτίας της ημιτελής κάλυψης του πληθυσμού στόχου.

Αναλυτικότερα, το σημείο εκκίνησης της διαδικασίας στάθμισης είναι ο μη προσαρμοσμένος σταθμισμένος μέσος του δείγματος με τη χρήση μίας βάσης βαρών (w_i), τα οποία είναι αντιστρόφως ανάλογα της πιθανότητας επιλογής. Η απόκλιση που προκύπτει από τη χρήση αυτού του μέσου στην εκτίμηση του μέσου του πληθυσμού όταν υπάρχουν ελλιπή στοιχεία μπορεί να εκφραστεί με διάφορους τρόπους.

Μία μέθοδος που χρησιμοποιείται, υιοθετεί ένα προκαθορισμένο μοντέλο, σύμφωνα με το οποίο υπάρχουν δύο ομάδες (ομάδα ανταπόκρισης και ομάδα μη ανταπόκρισης) στις οποίες μπορεί να τοποθετηθεί κάθε ένα από τα στοιχεία του δείγματος. Αν παρατηρείται τοποθετείται στην ομάδα ανταπόκρισης και αν δεν υπάρχει στην ομάδα μη ανταπόκρισης.

Η απόκλιση του μη προσαρμοσμένου μέσου που υπολογίζεται μόνο από τα δεδομένα που παρατηρούνται μπορεί να εκφραστεί ως:

$$B(\bar{y}_r) = P_m \cdot (\bar{Y}_r - \bar{Y}_m)$$

όπου $\bar{y}_r$: ο μέσος όρος των συμμετεχόντων υπολογιζόμενος με τη βάση βαρών

P_m: Ο ρυθμός μη ανταπόκρισης

$\bar{Y}_r$: ο μέσος όρος των παρατηρούμενων τιμών

$\bar{Y}_m$: ο μέσος όρος των ελλιπών τιμών

Από την παραπάνω εξίσωση είναι εμφανές ότι η απόκλιση του $\bar{y}_r$ είναι μία εξίσωση τόσο του ρυθμού μη ανταπόκρισης, όσο και της διαφοράς των χαρακτηριστικών μεταξύ των παρατηρούμενων και των ελλιπών στοιχείων. Πετυχαίνοντας ένα υψηλό ρυθμό ανταπόκρισης είναι ο μόνος τρόπος να εξασφαλιστεί ότι η απόκλιση θα είναι μικρή καθώς δεν είναι δυνατόν να μην υπάρχουν διαφορές μεταξύ των χαρακτηριστικών των περιπτώσεων που διαθέτουν όλα τα στοιχεία και αυτών που έχουν ελλιπή στοιχεία.

Ένας άλλος τρόπος μοντελοποίησης του μηχανισμού μη ανταπόκρισης είναι να θεωρηθεί ότι τα στοιχεία του πληθυσμού έχουν διαφορετικές, μη μηδενικές πιθανότητες ανταπόκρισης στην έρευνα. Με αυτό το μοντέλο η απόκλιση του $\bar{y}_r$ δίνεται από τη σχέση:

$$B(\bar{y}_r) \approx \frac{1}{N \cdot \Phi} \cdot \sum (Y_i - \bar{Y}) \cdot (\Phi_i - \bar{\Phi})$$

όπου: Φ_i: η πιθανότητα το στοιχείο i να ανταποκριθεί στην έρευνα ($\Phi_i > 0$)

$\bar{\Phi} = \frac{\sum \Phi_i}{N}$ και

N: ο αριθμός των στοιχείων του πληθυσμού.

Αυτή η έκφραση δείχνει ότι η απόκλιση είναι μία εξίσωση της σχέσης μεταξύ των χαρακτηριστικών που μπορούν να εκτιμηθούν και της πιθανότητας ανταπόκρισης, με το $\bar{y}_r$ να μην παρουσιάζει απόκλιση μόνο όταν τα Y_i, Φ_i δεν συσχετίζονται, δηλαδή έχουμε μηχανισμό MCAR.

Αν η πιθανότητες ανταπόκρισης Φ_i ήταν γνωστές η απόκλιση του $\bar{y}_r$ θα μπορούσε να περιοριστεί αν εφαρμοζόταν το Φ_i^{-1} σαν προσαρμοσμένο βάρος μη ανταπόκρισης, οδηγώντας σε προσαρμοσμένα βάρη $w_i^* = \frac{w_i}{\Phi_i}$.

Δυστυχώς όμως η πιθανότητα ανταπόκρισης δεν είναι γνωστή . Μπορεί όμως να εκτιμηθεί με χρησιμοποίηση ενός μοντέλου του μηχανισμού ανταπόκρισης. Το πιο απλό μοντέλο που μπορεί να χρησιμοποιηθεί είναι το μοντέλο του μηχανισμού MCAR το οποίο θεωρεί ότι τα ελλιπή στοιχεία αποτελούν ένα τυχαίο υποσύνολο του δείγματος.

Αυτό σημαίνει ότι $\Phi_i = \bar{\Phi}$ για όλα τα i, άρα το $B(\bar{y}_r) = 0$ και το $\bar{y}_r$ θεωρείται ότι δεν έχει απόκλιση. Στις περισσότερες έρευνες αυτό το μοντέλο (μοντέλου

του μηχανισμού MCAR) θεωρείται ότι δεν ανταποκρίνεται στην πραγματικότητα, αφού υπάρχουν σημαντικές διαφορές μεταξύ των περιπτώσεων που διαθέτουν όλα τα στοιχεία και αυτών που έχουν ελλιπή δεδομένα.

Ένα πιο λογικό μοντέλο που μπορεί να χρησιμοποιηθεί για την εκτίμηση της πιθανότητας ανταπόκρισης είναι το μοντέλο του μηχανισμού MAR, το οποίο θεωρεί ότι ο πληθυσμός μπορεί να χωριστεί σε ομάδες, έτσι ώστε τα ελλιπή στοιχεία κάθε ομάδας να είναι ένα τυχαίο υποσύνολο της ομάδας. Ο μη προσαρμοσμένος μέσος παρουσιάζει κάποια απόκλιση με αυτό το μοντέλο.

Παρακάτω θα παρουσιαστούν τα βασικότερα μοντέλα πιθανότητας ανταπόκρισης και οι μέθοδοι που χρησιμοποιούνται για την προσαρμογή των βαρών.

5.2 Προσαρμοσμένες Ομάδες Στάθμισης (Weighting Class Adjustments)

Η μέθοδος αυτή χρησιμοποιείται στο δεύτερο στάδιο της διαδικασίας στάθμισης, στο οποίο τα βάρη που έχουν παραχθεί στο πρώτο στάδιο προσαρμόζονται, ώστε να αντισταθμιστεί η έλλειψη των στοιχείων που υπάρχει. Η μέθοδος αυτή ανήκει στην σταθμισμένη προσαρμογή δείγματος, η οποία προσπαθεί να ελαττώσει την απόκλιση της μη ανταπόκρισης, κατανέμοντας τη βάση βαρών των περιπτώσεων που έχουν ελλιπή στοιχεία σε αυτές που τα έχουν όλα. Μία απλή προσαρμογή που μπορεί να γίνει είναι η αύξηση της βάσης των βαρών των περιπτώσεων με πλήρη στοιχεία, πολλαπλασιάζοντας με το αντίστροφο ενός σταθμισμένου ρυθμού ανταπόκρισης:

$$\hat{\Phi} = \frac{\sum^{r} w_i}{\left(\sum^{r} w_i + \sum^{m} w_i\right)}$$

Όπου: $\sum^{r} w_i$ και $\sum^{m} w_i$ είναι τα αθροίσματα της βάσης των βαρών για τις περιπτώσεις που παρατηρούνται και αυτές που λείπουν αντίστοιχα.

Χρησιμοποιώντας τη βάση βαρών, w_i ή τα προσαρμοσμένα βάρη $w_i^* = \frac{w_i}{\hat{\Phi}}$, λαμβάνουμε την ίδια εκτιμώμενη τιμή για τα ποσά που εκφράζουν μέσο όρο (μέση τιμή, αναλογία, συντελεστές παλινδρόμησης). Όμως τα προσαρμοσμένα βάρη δίνουν καλύτερα αποτελέσματα στις εκτιμήσεις του συνόλου. Για να επιτευχθεί τιμή χωρίς καθόλου αποκλίσεις θα πρέπει τα ελλιπή στοιχεία να λείπουν τελείως τυχαία (MCAR).

Μία άλλη διαδικασία αυτής της κατηγορίας για την προσαρμογή των βαρών απαιτεί να υπάρχουν κάποια γνωστά χαρακτηριστικά για όλες τις περιπτώσεις του δείγματος (και αυτές που διαθέτουν όλα τα στοιχεία και αυτές που παρουσιάζουν κάποια ελλιπή στοιχεία), καθώς σύμφωνα με αυτά τα χαρακτηριστικά θα δημιουργηθούν οι προσαρμοσμένες ομάδες στάθμισης.

Σε αυτή την περίπτωση εκτιμάται η μέση τιμή $\bar{y}_{wc}$ κάθε σταθμισμένης ομάδας, η οποία προέρχεται προσαρμόζοντας τα βάρη των παρατηρούμενων δεδομένων κάθε ομάδας με το αντίστροφο του ρυθμού ανταπόκρισης της ομάδας: $w_{ci}^* = \frac{w_{ci}}{\hat{\Phi}_{ci}}$, όπου $\hat{\Phi}_{ci}$: ο σταθμισμένος ρυθμός ανταπόκρισης της ομάδας.

Με αυτό το μοντέλο πιθανότητας ανταπόκρισης, η κατά προσέγγιση απόκλιση της εκτιμώμενης μέσης τιμής δίνεται από τη σχέση:

$$B(\bar{y}_{wc}) = \frac{1}{N} \cdot \sum\sum \frac{1}{\bar{\Phi}_c} \cdot (y_{ci} - \bar{y}_c) \cdot (\Phi_{ci} - \bar{\Phi}_c)$$

Όπως φαίνεται από την παραπάνω σχέση, ο εκτιμητής της σταθμισμένης ομάδας δεν παρουσιάζει απόκλιση, υπό την προϋπόθεση ότι τα ελλιπή στοιχεία ακολουθούν τον μηχανισμό MAR στις ομάδες. Δηλαδή όταν το $\Phi_{ci} = \overline{\Phi}_c$ για όλες τις ομάδες ($B(\overline{y}_{wc})=0$).

Ο πιο συνηθισμένος τρόπος για την δημιουργία των ομάδων στάθμισης είναι η χρησιμοποίηση κάποιων βοηθητικών μεταβλητών που είναι διαθέσιμες τόσο για τις περιπτώσεις που έχουν όλα τα στοιχεία, όσο και για τις περιπτώσεις που παρουσιάζουν και ελλιπή στοιχεία. Ο Little ορίζει αυτή τη διαδικασία ως προδιάθεση διαστρωμάτωσης (Propensity Stratification).

Οι βοηθητικές μεταβλητές χρησιμοποιούνται για να χωρίσουν το δείγμα σε σταθμισμένες ομάδες με διαφορετικούς ρυθμούς ανταπόκρισης ώστε η ποικιλία των ρυθμών ανταπόκρισης να εξασφαλίζει ότι τα στοιχεία δεν καλύπτονται οι προϋποθέσεις του MAR μηχανισμού για οποιοδήποτε συνδυασμό αυτών των ομάδων και ελπίζεται να ικανοποιούνται αυτές οι προϋποθέσεις εντός των ομάδων.

Η επιλογή των βοηθητικών μεταβλητών σύμφωνα με τις οποίες θα γίνει ο διαχωρισμός των στοιχείων σε σταθμισμένες ομάδες άλλες φορές είναι εύκολο να γίνει και άλλες φορές παρουσιάζει κάποιες δυσκολίες. Σε πολλές περιπτώσεις υπάρχουν πολύ λίγα στοιχεία που παρατηρούνται για όλες τις περιπτώσεις του δείγματος. Συχνά, το μόνο που είναι γνωστό για τις περιπτώσεις που έχουν ελλιπή στοιχεία είναι μόνο κάποια βασικά χαρακτηριστικά τους και τότε είναι εύκολο να επιλεγούν οι βοηθητικές μεταβλητές που θα χρησιμοποιηθούν για την ομαδοποίηση των περιπτώσεων.

Σε άλλες πάλι περιπτώσεις υπάρχουν πολλά στοιχεία ακόμα και για τις περιπτώσεις που παρουσιάζουν ελλιπή στοιχεία (π.χ. σε μία μακροχρόνια έρευνα όπου διατηρείται ένα αρχείο με τους συμμετέχοντες, πολλοί από

αυτούς μπορεί να μην έχουν συμμετέχει σε κάποια στάδια της έρευνας αλλά υπάρχουν αρκετές πληροφορίες για αυτούς). Τότε το πρώτο στάδιο της ανάλυσης είναι η επιλογή των βοηθητικών μεταβλητών με τις οποίες θα γίνει η ομαδοποίηση των στοιχείων.

5.3 Poststratification adjustment

Η Poststratification adjustment ή εκτίμηση ποσοστού είναι μία από τις συχνότερα χρησιμοποιούμενες διαδικασίες για την σταθμισμένη προσαρμογή του πληθυσμού και χρησιμοποιείται στο τρίτο στάδιο της διαδικασίας στάθμισης, όπου εκτιμώνται κάποιες τιμές για το σύνολο του πληθυσμού με την προσαρμογή κάποιων γνωστών τιμών για τα σύνολα, με σκοπό τη μείωση της απόκλισης που οφείλεται στα ελλιπή στοιχεία.

Στη Poststratification adjustment τα βάρη για τα στοιχεία μίας ομάδας πολλαπλασιάζονται με ένα συντελεστή, έτσι ώστε το άθροισμα των βαρών για τους συμμετέχοντες της ομάδας να ισούται με τον συνολικό πληθυσμό της ομάδας.

Δηλαδή, $$w_{ci}^{*} = w_{ci} \cdot \frac{X_c}{\hat{X}_c}$$

Όπου X_c: ο γνωστός συνολικός πληθυσμός της ομάδας

$\hat{X}_c$: ο εκτιμητής του συνόλου που βασίζεται στα μη προσαρμοσμένα βάρη.

Η μόνη διαφορά μεταξύ του εκτιμητή αυτής της μεθόδου και του εκτιμητή της Weighting Class Adjustment είναι ότι ο προσαρμοσμένος συντελεστής βασίζεται σε γνωστά στοιχεία και όχι στο εκτιμώμενο σύνολο της ομάδας.

5.4 Πλεονεκτήματα – Μειονεκτήματα

Οι διαδικασίες στάθμισης χρησιμοποιούνται κυρίως σε συνδυασμό με την Complete Case Analysis και έχουν ως σκοπό να αντισταθμίσουν τα ελλιπή στοιχεία και κατ' επέκταση να μειώσουν τις αποκλίσεις που παρουσιάζουν οι εκτιμητές της έρευνας. Οι σταθμισμένοι εκτιμητές είναι συχνά εύκολο να υπολογιστούν και οι διαδικασίες αυτές θεωρείται ότι έχουν πολλά πλεονεκτήματα.

Ωστόσο οι διαδικασίες αυτές είναι ημιτελής καθώς στηρίζονται σε προϋποθέσεις, όπως ότι τα ελλιπή στοιχεία ακολουθούν μηχανισμό MAR, κάτι που δεν είναι εύκολο να αποδειχτεί. Επίσης, όπως και στην Complete Case Analysis, έτσι και εδώ δεν χρησιμοποιούνται όλα τα στοιχεία της έρευνας, καθώς οι ημιτελής περιπτώσεις διαγράφονται. Οι σταθμισμένοι εκτιμητές μπορεί να έχουν μεγάλη διακύμανση, όπως στις περιπτώσεις που υπάρχουν απομακρυσμένες τιμές για τις μεταβλητές, οι οποίες δίνουν μεγάλα βάρη. Τέλος, ενώ οι εκτιμητές μπορούν να υπολογιστούν αρκετά εύκολα οι αποκλίσεις του είναι δύσκολο να υπολογιστούν.

Η διαδικασία της στάθμισης είναι εύκολο να εφαρμοστεί όταν η μορφή των ελλιπών στοιχείων είναι μονότονη (σχήμα 2), ή μονο-μεταβλητή (σχήμα 3), αλλά είναι πολύ δύσκολο να εφαρμοστεί όταν είναι τυχαία (γενική, σχήμα 1), καθώς πρέπει να υπολογιστεί διαφορετικό σύνολο βαρών για κάθε μεταβλητή.

5.5 Συμπεράσματα

Οι διαδικασίες στάθμισης χρησιμοποιούνται για τη μείωση της στατιστικής απόκλισης των αποτελεσμάτων που προέρχεται από την έλλειψη κάποιων δεδομένων. Αυτό επιτυγχάνεται με την εισαγωγή κάποιων βαρών στα δεδομένα του προβλήματος με σκοπό να θεωρηθούν τα ελλιπή στοιχεία μέρος του. Χρησιμοποιούνται σε συνδυασμό με τις διαδικασίες διαγραφής, καθώς και σε αυτή την περίπτωση τα ελλιπή στοιχεία δεν λαμβάνονται υπόψη στην ανάλυση του προβλήματος. Τα αποτελέσματα των διαδικασιών αυτών είναι ικανοποιητικά όταν τα ελλιπή στοιχεία λείπουν τυχαία (MAR). Είναι εύκολο να εφαρμοστούν σε μορφές έλλειψης: μονότονη και μονο-μεταβλητή, ενώ είναι δύσκολο να εφαρμοστεί όταν τα δεδομένα έχουν τυχαία μορφή έλλειψης.

6

Μέθοδοι Αντικατάστασης (Replacement Procedures)

Οι μέθοδοι διαγραφής δεν κάνουν χρήση των περιπτώσεων για τις οποίες υπάρχουν ελλιπή στοιχεία, κάτι που μειώνει την στατιστική δύναμη των αποτελεσμάτων. Για την αντιμετώπιση αυτού του προβλήματος αναπτύχθηκε μία άλλη κατηγορία μεθόδων οι μέθοδοι αντικατάστασης (Replacement Procedures) ή μέθοδοι Καταλογισμού (Imputation methods).

Σκοπός των μεθόδων καταλογισμού είναι η συμπλήρωση των ελλιπών στοιχείων με κατάλληλες τιμές, οι οποίες προέρχονται από άλλα στοιχεία της έρευνας ή από άλλες πληροφορίες που μπορεί να διατίθενται για κάθε περίπτωση.

Οι μέθοδοι αντικατάστασης είναι σχετικά εύκολο να εφαρμοστούν και πολλές συμπεριλαμβάνονται στα στατιστικά πακέτα. Το μεγαλύτερο πλεονέκτημά τους είναι ότι διατηρούν το μέγεθος του δείγματος. Όμως η αντικατάσταση των ελλιπών τιμών πρέπει να γίνεται με μεγάλη προσοχή, καθώς μπορεί να επέλθουν μεγάλες αποκλίσεις. Η αντικατάσταση των ελλιπών στοιχείων θεωρείται κατάλληλη μόνο όταν η συσχέτιση μεταξύ των μεταβλητών είναι μικρή.

Σύμφωνα με τους Demster και Rubin (1983) η ιδέα του καταλογισμού είναι τόσο ελκυστική, όσο και επικίνδυνη. Ελκυστική επειδή μπορεί να προσφέρει ηρεμία στον αναλυτή και την απόλαυση ότι χρησιμοποιεί όλα τα δεδομένα του, τα οποία είναι πλήρως συμπληρωμένα. Επικίνδυνη επειδή συγκεντρώνει μαζί περιπτώσεις όπου το πρόβλημα είναι μικρότερης σημασίας και μπορεί να αντιμετωπιστεί με αυτό τον τρόπο και δύσκολες περιπτώσεις όπου εφαρμόζονται κοινοί εκτιμητές και τα δεδομένα που αντικαταστήθηκαν μπορεί να προκαλέσουν σημαντικές αποκλίσεις.

Υπάρχουν πολλές διαφορετικές διαδικασίες αντικατάστασης. Γενικά έχει αποδειχτεί ότι οι διαφορές μεταξύ αυτών των μεθόδων μπορεί να μειωθούν αν:

- Αυξηθεί το δείγμα
- Μειωθεί το ποσοστό των ελλιπών στοιχείων
- Μειωθούν οι ελλιπής μεταβλητές
- Μειωθεί το ποσοστό συσχέτισης μεταξύ των μεταβλητών

Οι μέθοδοι που ανήκουν σε αυτή την κατηγορία μπορούν να διαχωριστούν με διάφορους τρόπους. Αρχικά πρέπει να διαχωριστούν σε παραδοσιακές (traditional methods) και «μοντέρνες» μεθόδους (modern methods), καθώς σε αυτή την κατηγορία ανήκει μία από τις δύο «μοντέρνες» μεθόδους, ο πολλαπλός καταλογισμός (multiple Imputation). Εδώ αρχικά θα παρουσιαστούν οι παραδοσιακές μέθοδοι που θεωρούνται προάγγελοι του

πολλαπλού καταλογισμού (Multiple Imputation) και στην συνέχεια ο πολλαπλός καταλογισμός (Multiple Imputation).

Ένας βασικός τρόπος ταξινόμησης των μεθόδων αντικατάστασης είναι με βάση τη διαδικασία που ακολουθούν για την αντικατάσταση των ελλιπών στοιχείων. Έτσι προκύπτουν τρεις κατηγορίες α) μέθοδοι παλινδρόμησης (regression based methods), β) μέθοδοι μέσης τιμής (mean based methods) και γ) hot deck based methods.

Επίσης μπορούν να διαχωριστούν σε μεθόδους συνόλου και μεθόδους υποσυνόλων ή ομάδων. Στην πρώτη περίπτωση οι τιμές της αντικατάστασης προέρχονται λαμβάνοντας υπόψη όλο το σύνολο του δείγματος. Αντίθετα στη δεύτερη περίπτωση τα δεδομένα χωρίζονται σε κατηγορίες (με βάση κάποια κοινά στοιχεία) και λαμβάνονται τιμές αντικατάστασης για τα ελλιπή στοιχεία από κάθε κατηγορία χωριστά. Τέλος μπορούν να διαχωριστούν σε στοχαστικές και αιτιοκρατικές μέθοδοι.

6.1 Στοχαστικές – Αιτιοκρατικές Μέθοδοι

Αρχικά θεωρείται σκόπιμο να εισαχθούν οι όροι των στοχαστικών και αιτιοκρατικών μεθόδων, καθώς θα αναφερθούν και παρακάτω στις μεθόδους. Όπως αναφέρθηκε και παραπάνω ένας από τους τρόπους που μπορούν να χωριστούν οι μέθοδοι καταλογισμού είναι οι στοχαστικές και οι αιτιοκρατικές μέθοδοι. Η διαφορά αυτών των μεθόδων έγκειται σε ένα υπολειπόμενο όρο που διαθέτουν οι στοχαστικές μέθοδοι.

Αναλυτικότερα, έστω **y** η μεταβλητή η οποία έχει ελλιπή στοιχεία και για την οποία πρέπει να καταλογιστούν οι τιμές και $\mathbf{z} = (z_1, z_2, \ldots, z_p)$ το σύνολο των άλλων μεταβλητών (βοηθητικές μεταβλητές) με βάση τις οποίες θα αντικατασταθούν τα ελλιπή στοιχεία της μεταβλητής **y**. Επίσης, θεωρείται ότι το **y** είναι μία συνεχής μεταβλητή και ότι το **z** δεν έχει ελλιπή στοιχεία.

Επιπλέον έστω ότι y_{rk} είναι η τιμή του **y** για το στοιχείο **k** η οποία παρατηρείται και y_{mi} η τιμή του **y** για το στοιχείο **i** η οποία λείπει και $\hat{y}_{mi}$ η αποδοσμένη τιμή του **y** για το στοιχείο **i** που λείπει.

Πολλές μέθοδοι καταλογισμού μπορούν να αναπαρασταθούν με την εξίσωση παλινδρόμησης:

$$\hat{y}_{mi} = b_{r0} + \sum b_{ij} z_{mij} + \hat{e}_{mi}$$

Όπου: b_{r0}: ο παρεμποδιστής, b_{ij}: ο εκτιμώμενος συντελεστής παλινδρόμησης από την παλινδρόμηση του y στο z, για τα στοιχεία του y που παρατηρούνται, z_{mij}: η τιμή του z_i για το στοιχείο y να λείπει και $\hat{e}_{mi}$: ένας υπολειπόμενος όρος.

Η βασική διαφορά μεταξύ των στοχαστικών και των αιτιοκρατικών μεθόδων είναι ότι στις αιτιοκρατικές μεθόδους ο όρος $\hat{e}_{mi}$ είναι ίσος με μηδέν ($\hat{e}_{mi} = 0$), ενώ στις στοχαστικές δεν είναι. Οι αιτιοκρατικές μέθοδοι προτιμούνται στην περίπτωση που επιθυμείται καλύτερη τιμή πρόβλεψης, καθώς δεν έχουν των θόρυβο που παρουσιάζουν οι στοχαστικές μέθοδοι, εξαιτίας του όρου $\hat{e}_{mi}$. Όμως, οι μέθοδοι αυτές (αιτιοκρατικές) τοποθετούν όλες τις τιμές καταλογισμού στην γραμμή παλινδρόμησης με αποτέλεσμα να μην ανταποκρίνονται στην υπολογιστική μεταβλητότητα. Έτσι τα αποτελέσματα παρουσιάζουν μεγαλύτερες αποκλίσεις.

6.2 Διαχωρισμός βάσει τύπου αντικατάστασης

Οι μέθοδοι καταλογισμού μπορούν να διαχωριστούν και με βάση το πώς θα αντικατασταθούν τα ελλιπή στοιχεία. Με αυτόν τον τρόπο διακρίνονται τρεις ομάδες μεθόδων:

- Μέθοδοι μέσης τιμής (mean based methods)
- Μέθοδοι παλινδρόμησης (regression based methods)
- Hot Deck Based

6.2.1 Μέθοδοι μέσης τιμής (mean based methods)

Η κατηγορία αυτή όπως φαίνεται και από το όνομά της περιλαμβάνει μεθόδους που τα ελλιπή στοιχεία αντικαθίστανται με κάποια μέση τιμή που προκύπτει από τις παρατηρούμενες τιμές του δείγματος. Σε αυτή την κατηγορία μεθόδων μπορούν να διακριθούν τρεις υποκατηγορίες:

- Μέθοδοι Αντικατάστασης Μέσης Τιμής Συνόλου: Η τιμή που χρησιμοποιείται για να αντικαταστήσει τα ελλιπή δεδομένα για μία μεταβλητή, προέρχεται από την μέση τιμή όλων των παρατηρούμενων δεδομένων αυτής της μεταβλητής.

- Μέθοδοι Αντικατάστασης Μέσης Τιμής Υποσυνόλου: Σε αυτή την περίπτωση έχουμε μία μέθοδο υποσυνόλου ή ομάδας. Τα στοιχεία αρχικά χωρίζονται με την χρήση βοηθητικών μεταβλητών σε ομάδες και οι τιμές που χρησιμοποιούνται για την αντικατάσταση προκύπτουν από αυτές τις ομάδες. Οι ομάδες αυτές μπορούν να διαμορφωθούν από την πείρα του αναλυτή στο θέμα, ή κάποιες φορές σε ορισμένους αλγόριθμους, όπως το KAPPO, CHAID ή SEARCH χρησιμοποιούνται για να

καθορίσουν τις κατηγορίες καταλογισμού. Αφού χωριστούν οι ομάδες, υπολογίζεται η μέση τιμή κάθε ομάδας και κατόπιν κάθε ελλιπές στοιχείο παίρνει την μέση τιμή της ομάδας που ανήκει.

- Μέθοδοι Αντικατάστασης Μέσης Τιμής Περίπτωσης: Εδώ προκύπτει μία ατομική μέση τιμή για τον συγκεκριμένο συμμετέχοντα, με βάση τα παρατηρούμενα στοιχεία του ίδιου.

Η εφαρμογή των μεθόδων αυτής της κατηγορίας επιτρέπει την συμμετοχή όλων των δεδομένων στην τελική ανάλυση, ωστόσο μπορεί να οδηγήσει σε μη έγκυρα αποτελέσματα. Η χρησιμοποίηση της μέσης τιμής ως τιμή αντικατάστασης οδηγεί σε έγκυρες εκτιμήσεις μόνο όταν τα ελλιπή στοιχεία λείπουν τελείως τυχαία (MCAR). Όμως οι εκτιμήσεις συνδιακύμανσης και διασποράς δεν θεωρούνται έγκυρες, καθώς αυτή η μέθοδος υποεκτιμά την μεταβλητότητα των ελλιπών τιμών.

6.2.2 Μέθοδοι παλινδρόμησης (regression based methods)

Οι τιμές των ελλιπών δεδομένων αυτής της κατηγορίας αντικαθίστανται από τιμές που προέρχονται από μία ανάλυση παλινδρόμησης για κάθε μία μεταβλητή, η οποία στηρίζεται μόνο στις παρατηρούμενες τιμές της συγκεκριμένης μεταβλητής. Η διαδικασία αυτή περιλαμβάνει δύο στάδια.

Στο πρώτο στάδιο ο αναλυτής εκτιμά τη σχέση μεταξύ των μεταβλητών και στη συνέχεια χρησιμοποιεί τους συντελεστές παλινδρόμησης για να εκτιμήσει τις τιμές των ελλιπών στοιχείων. Για να χρησιμοποιηθεί αυτή η μέθοδος θα πρέπει να έχει εξασφαλιστεί ότι υπάρχει γραμμική σχέση μεταξύ των εκτιμώμενων και των ελλιπών στοιχείων.

Η μέθοδος αυτή δίνει σωστούς υπολογισμούς όταν τα στοιχεία λείπουν τελείως τυχαία (MCAR), και ικανοποιητικά αποτελέσματα αν τα στοιχεία λείπουν τυχαία (MAR). Ωστόσο, οι παράμετροι διασποράς συνεχίζουν να υποεκτιμούνται, καθώς αυτή η μέθοδος προϋποθέτει μηδενικά λάθη γύρω από την παλινδρόμηση.

Μία καλύτερη μέθοδος θεωρείται η στοχαστική παλινδρόμηση στην οποία κάθε ελλιπή τιμή του συνόλου των δεδομένων αντικαθίσταται από μία προβλεπόμενη τιμή που προέρχεται από μία ανάλυση παλινδρόμησης, η οποία στηρίζεται στα παρατηρούμενα στοιχεία και σε ένα υπολειπόμενο όρο (resiudal, $\hat{e}_{mi}$)

6.2.3 Hot – deck Imputation

Η κατηγορία αυτή περιλαμβάνει μεθόδους στις οποίες κάθε ελλιπής τιμή αντικαθίσταται από μία παρατηρούμενη τιμή με τυχαία επιλογή μίας περίπτωσης που είναι παρόμοια με την περίπτωση όπου εμφανίζεται η ελλιπής τιμή. Ως παρόμοιες περιπτώσεις μπορούν να ορισθούν οι περιπτώσεις που ταιριάζουν με την επιλεγμένη περίπτωση (με ελλιπή στοιχεία) από μία επιλεγμένη ομάδα περιπτώσεων.

Έστω μία περίπτωση Α, όπου όλα τα χαρακτηριστικά εκτός από ένα παρατηρούνται. Για να συμπληρωθεί το ελλιπές στοιχείο βρίσκονται όλες οι περιπτώσεις που διαθέτουν τιμή για αυτό. Κατόπιν διαχωρίζονται οι περιπτώσεις που διαθέτουν παρόμοια χαρακτηριστικά για τις υπόλοιπες μεταβλητές με την περίπτωση Α και λαμβάνεται τυχαία μία τιμή από αυτές για να συμπληρωθεί η ελλιπή τιμή.

Μια από τις μορφές του hot deck καταλογισμού (hot deck Imputation) είναι γνωστή και ως *διαδοχική hot deck μέθοδος* (sequential hot – deck method). Με αυτήν την μέθοδο καθορίζεται ένα σύνολο ομάδων με παρόμοιες περιπτώσεις, ώστε να συμπληρωθούν τα ελλιπή στοιχεία. Για κάθε μία κατηγορία δημιουργείται μία υπολογιστική θέση, στην οποία αποθηκεύεται η αξία της μεταβλητής που θα αντικαταστήσει την ελλιπή τιμή. Αν μία περίπτωση έχει τιμή για το y, αποθηκεύεται στη θέση της κατηγορίας. Κατόπιν ελέγχεται η επόμενη περίπτωση, αν έχει τιμή για το y, αποθηκεύεται στη θέση της κατηγορίας, αντικαθιστώντας την τιμή που υπάρχει μέχρι τότε. Αν όμως δεν έχει τιμή για το y, παίρνει την τιμή που είναι αποθηκευμένη στη θέση της κατηγορίας ως η τιμή του y. Η διαδικασία συνεχίζεται μέχρι να συμπληρωθούν όλα τα ελλιπή δεδομένα. Με αυτόν τον τρόπο τα αρχεία θεωρούνται διαδοχικά σε όλο το σύνολο των στοιχείων.

Τον καιρό που αναπτύχθηκε η *διαδοχική μέθοδο hot deck* (sequential hot – deck method) ήταν επιθυμητή η οικονομία στους υπολογισμούς, κάτι που έκανε αυτή την μέθοδο πολύ ελκυστική. Ωστόσο με την ανάπτυξη των υπολογιστών που έχει πραγματοποιηθεί τα τελευταία χρόνια η μέθοδος αυτή, σήμερα θεωρείται λιγότερο ελκυστική. Η μέθοδος αυτή έχει δύο κύρια μειονεκτήματα. Πρώτον, ο αριθμός κατηγοριών καταλογισμού πρέπει να περιοριστεί για να εξασφαλίσει ότι κάθε κατηγορία έχει τουλάχιστον μία περίπτωση που διαθέτει τιμή για το y, ώστε να μπορέσει να δώσει στις άλλες που δεν έχουν τιμή. Δεύτερον, όταν εμφανίζονται στη σειρά δύο ή περισσότερες περιπτώσεις με ελλιπή στοιχεία σε μια δεδομένη κατηγορία καταλογισμού, λαμβάνουν την ίδια τιμή για το y. Αυτή η "πολλαπλή χρήση ίδιων τιμών" προκαλεί απώλεια στην ακρίβεια των εκτιμήσεων της έρευνας σε σχέση με χρήση διαφορετικών τιμών για το y.

Κατά καιρούς έχουν δημιουργηθεί διάφορες εναλλακτικές hot deck μέθοδοι με σκοπό την εξάλειψη αυτών των μειονεκτημάτων. Ένας

σχετικά απλός τρόπος είναι να αποθηκευτούν, για παράδειγμα, τρεις τιμές περιπτώσεων που διαθέτουν τιμή για το y σε κάθε θέση κατηγορίας καταλογισμού, έτσι ώστε όταν παρουσιάζονται συνεχόμενες ελλιπής τιμές για το y, να λαμβάνεται μία από τις τρεις τιμές που είναι αποθηκευμένες και όχι πάντα η ίδια.

Μία άλλη περίπτωση hot – deck μεθόδου είναι η *ιεραρχική hot – deck μέθοδος (Hierarchical hot – deck method), η οποία έχει* μια πολύ σημαντική διαφορά. Με αυτήν την μέθοδο μπορούν να χρησιμοποιηθούν πολλές κατηγορίες καταλογισμού. Στο πρώτο βήμα, όλες οι περιπτώσεις της έρευνας, μέσα σε κάθε ομάδα, χωρίζονται σε περιπτώσεις που έχουν τιμή για το y και σε περιπτώσεις που δεν έχουν τιμή y, υπό την προϋπόθεση ότι υπάρχουν σε κάθε κατηγορία άτομα που έχουν απαντήσει και ότι η αναλογία ατόμων που δεν έχουν απαντήσει προς αυτούς που έχουν απαντήσει δεν είναι μεγάλη. (π.χ. όχι μεγαλύτερη από 3).

Κατόπιν επιλέγεται ένα δείγμα "χορηγών" από τις περιπτώσεις που έχουν τιμή για το y. Το δείγμα επιλέγεται με τέτοιο τρόπο, ώστε να εξασφαλίζεται ότι κάθε μία περίπτωση χρησιμοποιείται ίδιες φορές ή μία παραπάνω από τις άλλες. Αν υπάρχουν κατηγορίες καταλογισμού όπου υπάρχουν μόνο περιπτώσεις που δεν διαθέτουν τιμή για το y ή η αναλογία τους σε σχέση με αυτές που διαθέτουν τιμή για το y κρίνεται πολύ μεγάλη πρέπει να διαλυθούν αυτές οι κατηγορίες.

Αυτή η μέθοδος συντηρεί τα περιθώρια διανομής, ωστόσο μπορεί να διαστρεβλώσει τις σχέσεις μεταξύ των μεταβλητών. Επίσης, αν και θεωρείται εύκολο να εφαρμοστεί σε περιπτώσεις όπου παρατηρείται σε μια μεταβλητή, μια ελλείπουσα τιμή, μπορεί να είναι πολύ δύσκολο να εφαρμοστεί σε περιπτώσεις όπου πολλές μεταβλητές έχουν πολλά ελλιπή στοιχεία, καθώς μπορεί να είναι δύσκολο να καθοριστεί ποιες

περιπτώσεις μπορούν να θεωρηθούν παρόμοιες με αυτή που παρουσιάζει τα ελλιπή στοιχεία.

6.3 Διαχωρισμός βάσει της μοντελοποίησης

Ένας άλλος τρόπος με τον οποίο μπορεί να διακριθούν οι μέθοδοι καταλογισμού είναι η μοντελοποίηση της διανομής πρόβλεψης που χρησιμοποιείται. Σε αυτή την κατηγορία διακρίνουμε δύο ομάδες: την αναλυτική μοντελοποίηση (explicit modeling) και την πεπλεγμένη μοντελοποίηση (implicit modeling).

Στην αναλυτική μοντελοποίηση (explicit modeling) η διανομή πρόβλεψης βασίζεται σε ένα τυπικό στατιστικό μοντέλο, με αποτέλεσμα οι προϋποθέσεις για την εφαρμογή του μοντέλου να είναι σαφείς. Σε αυτή την κατηγορία ανήκουν οι μέθοδοι μέσης τιμής (mean Imputation methods) και οι μέθοδοι παλινδρόμησης (regression Imputation Methods).

Αντίθετα στη πεπλεγμένη μοντελοποίηση (implicit modeling) η διανομή πρόβλεψης οφείλεται σε ένα αλγόριθμο, ο οποίος συνεπάγεται ένα μοντέλο. Σε αυτή την περίπτωση οι προϋποθέσεις δεν είναι σαφής αλλά υποκρύπτονται και επιβάλλεται να καθοριστούν προσεκτικά, ώστε να είναι λογικές. Σε αυτή την κατηγορία ανήκουν οι hot – deck μέθοδοι (Hot – deck Imputation Methods).

6.4 Multiple Imputation

6.4.1 Γενικά

Ο πολλαπλός καταλογισμός αποτελεί συνδυασμό τόσο των μεθόδων αντικατάστασης, όσο και των μεθόδων που βασίζονται σε μοντέλα (κεφάλαιο 7), καθώς δημιουργείται από τον καθορισμό ενός μοντέλου για τα παρατηρούμενα δεδομένα (model – based methods), με βάση το οποίο

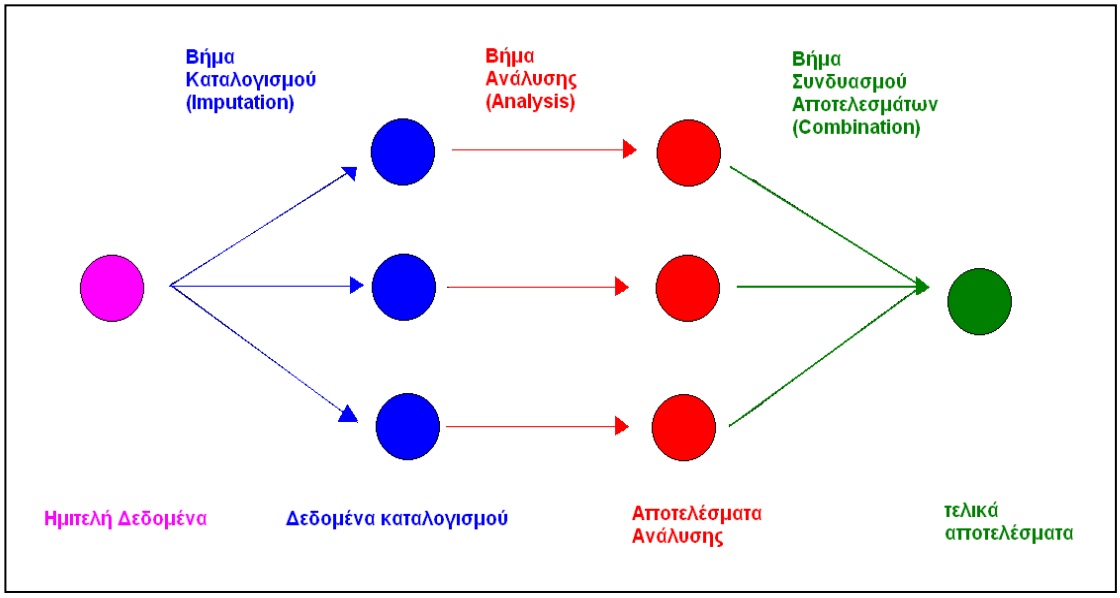
παράγονται τιμές που αποδίδονται στα ελλιπή δεδομένα (Imputation Methods) και δημιουργούνται σύνολα ολοκληρωμένων στοιχείων τα οποία κατόπιν αναλύονται για να παραχθούν οι τελικές εκτιμήσεις.

Ο πολλαπλός καταλογισμός (Multiple Imputation) είναι μία επέκταση της ιδέας του απλού καταλογισμού (Simple Imputation), η οποία έχει όλα τα θετικά του απλού καταλογισμού, επιλύοντας, επιπλέον, το πρόβλημα της κατανόησης της αβεβαιότητας (Schafer J. L. & Graham J. W., (2002)). Στον πολλαπλό καταλογισμό (Multiple Imputation) κάθε ελλιπές στοιχείο αντικαθίσταται με ένα σύνολο από $m>1$ αληθοφανείς τιμές με σκοπό τη δημιουργία m συνόλων με ολοκληρωμένα στοιχεία. Αυτά τα m σύνολα στοιχείων στη συνέχεια αναλύονται από κοινώς χρησιμοποιούμενα στατιστικά πακέτα Τα αποτελέσματα της ανάλυσης συνδυάζονται και παράγεται ένα αποτέλεσμα χρησιμοποιώντας απλούς κανόνες συνδυασμού, οι οποίες προτάθηκαν το 1987 από τον Rubin και Little και θα αναλυθούν παρακάτω. Οι τεχνικές αυτές δίνουν εκτιμήσεις για τις παραμέτρους και τα συστηματικά σφάλματα που πρέπει να ληφθούν υπόψη στην αβεβαιότητα λόγω των ελλιπών στοιχείων.

Ο πολλαπλός καταλογισμός έχει πολλά ελκυστικά χαρακτηριστικά και για αυτό κατατάσσεται στις ευρύτερα χρησιμοποιούμενες τεχνικές. Ένα από αυτά τα χαρακτηριστικά είναι ότι δεν χρειάζεται μεγάλος αριθμός επαναλήψεων για να παραχθούν ακριβείς επαναλήψεις, ανόμοια με άλλες μεθόδους που χρησιμοποιούνται για τη διαχείριση δεδομένων με ελλιπή στοιχεία.

Στις περισσότερες εφαρμογές 3 έως 5 τιμές απόδοσης είναι αποτελεσματικές για την παραγωγή έγκυρων αποτελεσμάτων. Ο Rubin το 1987 έδειξε ότι η αποτελεσματικότητα μίας εκτίμησης που βασίζεται σε m τιμές απόδοσης είναι περίπου: $\alpha = \frac{1}{1+\frac{\gamma}{m}}$, όπου γ είναι το όριο των ελλιπών πληροφοριών για την ποσότητα που εκτιμάται. Για 40% ελλιπή στοιχεία και m = 5 δίνει 93%

αποτελεσματικότητα, ενώ για m = 10 δίνει 96% αποτελεσματικότητα. Όπως είναι εμφανές από τα παραπάνω η αύξηση της αποτελεσματικότητας είναι πολύ μικρή με την αύξηση του m, ενώ για m = 5 θεωρείται αρκετά ικανοποιητική.



Σχήμα 8:
Γραφική Απεικόνιση των βημάτων του Πολλαπλού Καταλογισμού (Multiple Imputation Steps)

Ο Rubin το 1996 περιέγραψε τον πολλαπλό καταλογισμό (Multiple Imputation) ως μία διαδικασία τριών βημάτων, τα οποία φαίνονται στο σχήμα 8. Στο πρώτο βήμα ή στάδιο απόδοσης (Imputation step) δημιουργούνται σύνολα εύλογων τιμών που χρησιμοποιούνται για την αντικατάσταση των ελλιπών παρατηρήσεων. Κάθε ένα από αυτά τα σύνολα τιμών μπορεί να χρησιμοποιηθεί για την συμπλήρωση των ελλιπών στοιχείων, ώστε να δημιουργηθούν σύνολα ολοκληρωμένων στοιχείων. Στο δεύτερον στάδιο ή στάδιο Ανάλυσης (Analysis step) κάθε ένα από αυτά τα σύνολα μπορεί να αναλυθεί με χρήση τον κανονικών στατιστικών μεθόδων που χρησιμοποιείται όταν τα δεδομένα δεν παρουσιάζουν ελλιπή στοιχεία. Τέλος στο τελευταίο στάδιο (Combination /pooling) τα αποτελέσματα

συνδυάζονται κάτι που επιτρέπει να ληφθεί υπόψη η αβεβαιότητα σε σχέση με τη διαδικασία του καταλογισμού.

6.4.2 Τα βήματα του Πολλαπλού Καταλογισμού (Multiple Imputation)

Βήμα 1ο: Η απόδοση των τιμών

Η απόδοση των τιμών αποτελεί το πιο σημαντικό βήμα της μεθόδου του πολλαπλού καταλογισμού, καθώς στηρίζεται στις προϋποθέσεις των νόμων έλλειψης που παράγονται από τις παρατηρούμενες τιμές. Ο στόχος της απόδοσης είναι η απόδοση της σχέσης μεταξύ των μη παρατηρούμενων και των παρατηρούμενων μεταβλητών, ενώ λαμβάνει υπόψη την αβεβαιότητα που προέρχεται από τα ελλιπή στοιχεία.

Έστω Χ μία μεταβλητή που παρουσιάζει ελλιπή στοιχεία. Σε αυτό το βήμα αρχικά χρησιμοποιώντας τις περιπτώσεις που δεν παρουσιάζουν ελλιπή στοιχεία, εκτιμάται η πολλαπλή γραμμική παλινδρόμηση για την πρόβλεψη του Χ από τις άλλες σχετικές μεταβλητές. Κατόπιν συνδυάζονται οι συντελεστές παλινδρόμησης με τα παρατηρούμενα στοιχεία για την πρόβλεψη του Χ, ώστε να υπολογιστούν οι τιμές των περιπτώσεων του που δεν υπάρχουν. Στην συνέχεια χρησιμοποιείται ένας τυχαίος αριθμός γεννητόρων για να προσομοιωθεί ένα σύνολο από υπολειπόμενους όρους και να προστεθούν αυτοί οι υπολειπόμενοι όροι στη παλινδρόμηση πρόβλεψης για να παραχθεί ένα σύνολο από αποδομένες τιμές για το Χ. Η διαδικασία αυτή επαναλαμβάνεται για ορισμένες φορές ακόμα (5 – 10 φορές) και κάθε φορά προσομοιώνονται νέοι υπολειπόμενοι όροι για να δημιουργηθούν τα σύνολα του καταλογισμού.

Υπάρχει ένας μεγάλος αριθμός μοντέλων απόδοσης που χρησιμοποιούνται. Όταν η έλλειψη είναι μονότονη έχουν προταθεί για συνεχείς μεταβλητές απλές μέθοδοι όπως μέθοδοι τάσης (Propensity methods), και (predictive

Mean Matching) και για διακριτές μεταβλητές η λογιστική παλινδρόμηση (Logistic Regression) και η διακριτική ανάλυση (Discriminant analysis). Για πιο περίπλοκες μορφές έλλειψης χρησιμοποιούνται οι προσεγγίσεις Markov Chain Monte Carlo (MCMC).

Βήμα 2ο: Ανάλυση των συνόλων με τα ολοκληρωμένα στοιχεία

Το επόμενο στάδιο του καταλογισμού είναι να πραγματοποιηθεί η στατιστική ανάλυση ενδιαφέροντος για κάθε ένα από τα m σύνολα ολοκληρωμένων στοιχείων, αποθηκεύοντας τα διανύσματα των παραμέτρων και τις εκτιμήσεις των τυπικών σφαλμάτων. Για την ανάλυση των συνόλων χρησιμοποιούνται απλές στατιστικές μέθοδοι για πλήρη δεδομένα.

Βήμα 3ο: Ο συνδυασμός των εκτιμήσεων

Σε αυτό το βήμα όπως ειπώθηκε και παραπάνω γίνεται ο συνδυασμός των εκτιμήσεων που έχουν δημιουργηθεί από τη στατιστική ανάλυση των *m* συνόλων με ολοκληρωμένα δεδομένα με σκοπό την παραγωγή μίας κοινής εκτίμησης.

Η πιο απλή μέθοδος για τον συνδυασμό των αποτελεσμάτων των *m* αναλύσεων είναι η μέθοδος του Rubin που προτάθηκε το 1987 για μία μονοδιάστατη παράμετρο.

Έστω το Q αναπαριστά την ποσότητα ενός πληθυσμού που πρέπει να εκτιμηθεί. Έστω $\hat{Q}$ η εκτίμηση του Q και $\sqrt{U}$ το συστηματικό σφάλμα αν δεν υπήρχαν ελλιπή στοιχεία. Η μέθοδος θεωρεί ότι το δείγμα είναι αρκετά μεγάλο έτσι ώστε $\sqrt{U}\cdot\left(Q-\hat{Q}\right)$ είναι μία κανονική κατανομή, και $\hat{Q}\pm 1.96\cdot\sqrt{U}$ είναι περίπου το 95% της κάλυψης.

Βέβαια δεν μπορεί να υπολογιστεί το $\hat{Q}$ και $\sqrt{U}$. Επίσης υπάρχουν m διαφορετικές τιμές για αυτά $\left[\hat{Q}^{(j)}, U^{(j)}\right], j = 1, \ldots, m$.

Ο συνολικός εκτιμητής του Rubin (1987) είναι απλά ο μέσος όρος των m εκτιμητών

$$\overline{Q} = \frac{1}{m} \cdot \sum_{j=1}^{m} \hat{Q}^{(j)}$$

Η αβεβαιότητα του $\overline{Q}$ έχει δύο μέρη: τον μέσο όρο της εσωτερικής διασποράς της απόδοσης

$$\overline{U} = \frac{1}{m} \cdot \sum_{j=1}^{m} U^{(j)}$$

Και μεταξύ της διασποράς των αποδομένων τιμών

$$B = \frac{1}{m-1} \cdot \sum_{j=1}^{m} \left[\hat{Q}^{(j)} - \overline{Q}\right]^2$$

Η συνολική διασπορά είναι ίση με το άθροισμα των δύο αυτών στοιχείων

$$T = \overline{U} + (1 + m^{-1}) \cdot B$$

και η τετραγωνική ρίζα του T είναι το συνολικό συστηματικό σφάλμα.

Για τα διαστήματα εμπιστοσύνης ο Rubin (1987) πρότεινε τη χρήση μίας t προσέγγισης $T^{-\frac{1}{2}} \cdot \left(\overline{Q} - Q\right) \sim t_v$, όπου οι βαθμοί ελευθερίας δίνονται από

$$v = (m-1)\cdot\left[1+\frac{\overline{U}}{(1+m^{-1})\cdot B}\right]^2$$

Ο βαθμός ελευθερίας μπορεί να διαφέρει από m-1 έως ∞ εξαρτώμενος από τον ρυθμό των ελλιπών στοιχείων. Όταν οι βαθμοί ελευθερίας είναι μεγάλοι, η κατανομή t είναι σχεδόν κανονική, η συνολική διασπορά καλά εκτιμημένη και το κέρδος από την αύξηση του m θα είναι μικρό. Ο εκτιμώμενος ρυθμός των ελλιπών τιμών για το Q είναι κατά προσέγγιση $\frac{\tau}{\tau+1}$, όπου $\tau = \frac{(1+m^{-1})\cdot B}{\overline{U}}$ είναι η σχετική αύξηση της διασποράς εξαιτίας της μη ανταπόκρισης.

6.4.3 Οι προϋποθέσεις που απαιτούνται για τη χρησιμοποίηση του πολλαπλού καταλογισμού

Ο πολλαπλός καταλογισμός παράγεται υποθέτοντας ένα συγκεκριμένο μοντέλο καταλογισμού και η επιτυχία ή η αποτυχία του εξαρτάται από τις ιδιότητες και τις υποθέσεις αυτού του μοντέλου. Τρία είναι τα βασικά θέματα που καθορίζουν τις προϋποθέσεις της μεθόδου του πολλαπλού καταλογισμού

- Το μοντέλο των δεδομένων
- Η αρχική κατανομή για τις παραμέτρους του μοντέλου και
- Ο μηχανισμός που ακολουθούν τα ελλιπής στοιχεία

6.4.3.1 Το μοντέλο των δεδομένων

Το πρώτο και πιο σημαντικό βήμα για την ανάπτυξη του πολλαπλού καταλογισμού σε ένα σύνολο δεδομένων με ελλιπή στοιχεία είναι η υπόθεση ενός μοντέλου πιθανότητας που σχετίζει το Y ($Y = Y_{obs}, Y_{mis}$) με ένα σύνολο παραμέτρων. Χρησιμοποιώντας αυτό το μοντέλο πιθανότητας και μία αρχική κατανομή για τις παραμέτρους μπορεί να βρεθεί η κατανομή πρόβλεψης

$p(Y_{mis}/Y_{obs})$ για τα ελλιπή δεδομένα που εξαρτάται από τις παρατηρούμενες τιμές και κατόπιν να δημιουργηθούν οι τιμές απόδοσης από την κατανομή πρόβλεψης.

Το θεωρούμενο μοντέλο πρέπει να εκμεταλλευτεί όλη την γνώση που είναι διαθέσιμη για τη διαδικασία που παράγει τα δεδομένα. Το πιο εξυπηρετικό μοντέλο για συνεχείς παραμέτρους, είναι το κανονικό πολύ-μεταβλητό μοντέλο. Ένα από τα πλεονεκτήματά του είναι ότι το μοντέλο αυτό είναι εύκολο να διαχειριστεί υπολογιστικά. Το μοντέλο αυτό δίνει αρκετά καλά αποτελέσματα ακόμα και όταν οι μεταβλητές είναι δυαδικές ή ρητές, με την απόδοση να εκτελείται υποθέτοντας ένα κανονικό μοντέλο και κατόπιν οι καθοριζόμενες τιμές να στρογγυλοποιούνται στην πιο κοντινή κατηγορία. Αν υπάρχει μία μεταβλητή που δεν κατανέμεται ομαλά μπορεί να αλλάξει σε κανονική μεταβλητή και η αποδομένη τιμή να αλλάξει ξανά στην κανονική κλίμακα. Άλλα μοντέλα που χρησιμοποιούνται είναι το λογα-γραμμικό μοντέλο για ρητούς αριθμούς, ένα μείγμα των δύο μοντέλων για σύνολα μεικτών και το γραμμικό ιεραρχικό μοντέλο.

6.4.3.2 Αρχική Κατανομή

Η στατιστική διαδικασία που χρησιμοποιείται για την εφαρμογή της μεθόδου του πολλαπλού καταλογισμού είναι συνήθως η Μπευζιανή, η οποία απαιτεί μία αρχική κατανομή για να γίνει η ανάλυση. Η αρχική κατανομή και το μοντέλο των ολοκληρωμένων δεδομένων θα δώσουν την κατανομή πρόβλεψης $p(Y_{mis}/Y_{obs})$ για τα ελλιπή στοιχεία που εξαρτάται από τα παρατηρούμενα στοιχεία από τα οποία θα παραχθούν και οι τιμές αντικατάστασης των ελλιπών στοιχείων. Συνήθως για την διευκόλυνση, χρησιμοποιείται μία μη κατατοπιστική αρχική κατανομή. Επειδή υπάρχει υποκειμενικότητα στην επιλογή της αρχικής κατανομής, η μπεϋζιανή μέθοδος έχει κατά καιρούς κριτικαριστεί. Για μεγάλο αριθμό αναλύσεων, η αρχική κατανομή σχεδόν δεν έχει σημασία, καθώς όταν υπάρχει μεγάλο

αριθμός δείγματος κάθε λογική αρχική κατανομή δίνει παρόμοια αποτελέσματα. Αν το μέγεθος του δείγματος είναι μικρό η ανάλυση γίνεται με τη χρήση διαφορετικών αρχικών κατανομών και εξετάζεται κατοπιν αν τα αποτελέσματα αλλάζουν, ώστε να καταλήξουμε σε σωστά αποτελέσματα.

6.4.3.3 Ο μηχανισμός των ελλιπών δεδομένων

Ο πολλαπλός καταλογισμός που στηρίζεται σε μοντέλο προϋποθέτει ότι ο μηχανισμός που ακολουθούν τα ελλιπή στοιχεία είναι ο MAR. Η υπόθεση του MAR μηχανισμού επιτρέπει τη χρήση σχέσης μεταξύ των μεταβλητών, προερχόμενη από τις παρατηρούμενες τιμές που χρησιμοποιούνται για την δημιουργία των τιμών που θα αντικαταστήσουν τα ελλιπή δεδομένα. Για παράδειγμα σε μία μελέτη στην οποία υπάρχουν πληροφορίες για το X, με ελλιπή στοιχεία μόνο για τη μεταβλητή Y, η υπόθεση του MAR μηχανισμού (που βασικά σημαίνει ότι η έλλειψη εξαρτάται μόνο από το X) θα οδηγήσει στο τρόπο που το Y εξαρτάται από το X σε ολοκληρωμένα δεδομένα. Κατόπιν χρησιμοποιείται η σχέση εξάρτησης για την πρόβλεψη των ελλιπών τιμών του Y από τις σχετιζόμενες X μεταβλητές.

Παρόλο που είναι αλήθεια ότι τα αποτελέσματα του μπορεί να μην είναι έγκυρα αν δεν ισχύει πράγματι ο μηχανισμός MAR, αυτή η προϋπόθεση αποτελεί βολικό αρχικό σημείο για την ανάλυση. Θέτοντας την πιθανότητα να υπάρχουν στοιχεία που δεν μπορούν να αγνοηθούν οι υπολογισμοί θα γίνουν πολύ περίπλοκοι, ενώ το πρόβλημα θα εξειδικευθεί.

6.4.4 Πλεονεκτήματα και Μειονεκτήματα του Πολλαπλού καταλογισμού

Ο πολλαπλός καταλογισμός (Multiple Imputation) αναπτύχθηκε για την διαχείριση μεγάλων δειγμάτων που τα στοιχεία συλλέγονται μία φορά και χρησιμοποιούνται από μεγάλο αριθμό ερευνητών και για ένα αριθμό διαφορετικών αναλύσεων. Ο πολλαπλός καταλογισμός (Multiple

Imputation) είναι ιδανικός για αυτή την περίπτωση γιατί μπορεί να χρησιμοποιηθεί από το άτομο που μαζεύει τα στοιχεία (και έχει πρόσβαση σε μεγαλύτερο αριθμό πληροφοριών) με σκοπό τη δημιουργία ολοκληρωμένων συνόλων δεδομένων και κατόπιν τα ολοκληρωμένα στοιχεία που παράγονται (μία φορά) να χρησιμοποιηθούν από πολλά και διαφορετικά άτομα. Αυτό το πλεονέκτημα ισχύει και στον απλό καταλογισμό (Single Imputation). Επίσης όπως και ο απλός καταλογισμός (Single Imputation) (Single Imputation) έχει το πρακτικό πλεονέκτημα να επιτρέπει την ανάλυση των στοιχείων με τις κλασσικές μεθόδους που χρησιμοποιούνται για ολοκληρωμένα στοιχεία.

Όμως παρόλο που ο πολλαπλός καταλογισμός (Multiple Imputation) έχει τα ίδια πλεονεκτήματα με το απλό καταλογισμό (Single Imputation), αποκαθιστά τα αρνητικά του.

Το μειονέκτημα που εμφανίζεται στις μεθόδους του απλού καταλογισμού (Single Imputation) είναι ότι αποδίδουν μία απλή τιμή στα ελλιπή δεδομένα και με αυτό τον τρόπο τα αντιμετωπίζουν σαν να είναι γνωστά, χωρίς καμία αβεβαιότητα. Αυτό όμως δεν μπορεί να αντικατοπτρίσει την ποικιλομορφία του δείγματος.

Αντίθετα στον πολλαπλό καταλογισμό παράγονται m αποδομένες τιμές για κάθε ελλιπές στοιχείο άρα και m ολοκληρωμένα σύνολα δεδομένων που κατόπιν αναλύονται και στη συνέχεια μπορούν εύκολα να συνδυαστούν για να δημιουργήσουν μία έγκυρη εικόνα της ποικιλομορφίας του δείγματος εξαιτίας των ελλιπών στοιχείων.

Το μόνο μειονέκτημα του πολλαπλού καταλογισμού (Multiple Imputation) σε σχέση με την απλή απόδοση είναι ότι χρειάζεται περισσότερη δουλειά για να δημιουργηθούν τα πλήρη στοιχεία και να αναλυθούν τα αποτελέσματα. Η μεγαλύτερη δουλειά στην ανάλυση των δεδομένων, ωστόσο είναι μικρή

στη σημερινή εποχή με την ανάπτυξη της υπολογιστικής δύναμης των υπολογιστών.

6.4.5 Προγράμματα που εκτελούν την μέθοδο MI

Σήμερα υπάρχουν πληθώρα προγραμμάτων για τη μέθοδο του πολλαπλού καταλογισμού. Το NORM ένα ελεύθερο πρόγραμμα για τα Windows εφαρμόζει τη διαδικασία του πολλαπλού καταλογισμού υπό πολύ-μεταβλητό κανονικό μοντέλο.

Τέσσερα άλλα μοντέλα για τον πολλαπλό καταλογισμό σε S-PLUS είναι το NORM (Schafer, 1997), για πολύ-μεταβλητό κανονικό μοντέλο, για πολύ-μεταβλητά κατηγοριοποιημένα δεδομένα υπό το log-linear μοντέλο, το MIX για ποικίλα σύνολα δεδομένων που περιέχουν τόσο συνεχή όσο και κατηγοριοποιημένα δεδομένα και το PAN για πολυμεταβλητά ή σε συστάδες δεδομένα υπό το πολύ-μεταβλητό γραμμικό μοντέλο.

Το 2000 ο Y.C. Yuan με την διαδικασία PROC MI χρησιμοποίησε την μέθοδο του πολλαπλού καταλογισμού σε περιβάλλον SAS. Άλλες εκδόσεις του ίδιου αλγόριθμου έχουν εφαρμοστεί σε καινούριας κυκλοφορίας βιβλιοθήκες ελλιπών στοιχείων στο S-PLUS και LISREL.

Τέλος το SOLAS εφαρμόζει το MI σε Windows περιβάλλον και περιλαμβάνει μία διαδικασία πολλαπλού καταλογισμού βασισμένη σε μοντέλο, καθώς και παλαιότερες μεθόδους.

6.5 Συμπεράσματα

Σε αυτή την κατηγορία ανήκουν παραδοσιακές μέθοδοι, όπως ο απλός καταλογισμός (Single Imputation): Μέθοδοι μέσης τιμής (mean based methods) Μέθοδοι παλινδρόμησης (regression based methods) και Hot

Deck Based και σε μοντέρνες μέθοδοι στην οποία ανήκει ο πολλαπλός καταλογισμός (Multiple Imputation).

Το μεγαλύτερο πλεονέκτημα των μεθόδων αυτής της κατηγορίας και προτιμώνται από τις μεθόδους των προηγούμενων κεφαλαίων, καθώς χρησιμοποιούν όλα τα δεδομένα που έχουν συλλεχθεί με αποτέλεσμα να αυξάνουν την στατιστική δύναμη των αποτελεσμάτων.

Οι παραδοσιακές μέθοδοι αυτής της κατηγορίας απαιτούν πολλούς περιορισμούς για την εφαρμογή τους (συνήθως μηχανισμό έλλειψης MCAR) και μπορεί να παρουσιάσουν αποκλίσεις από τα πραγματικά αποτελέσματα, καθώς οι τιμές που αντικατέστησαν τα ελλιπή στοιχεία μπορεί να υπερεκτιμηθούν ή να υποεκτιμηθούν. Επίσης, κάθε ελλιπές στοιχείο αντικαθιστάται με μία τιμή κάτι που δεν αφήνει καμία αβεβαιότητα για την τιμή του ελλιπούς στοιχείου (θεωρείται ότι είναι η συγκεκριμένη).

Ο πολλαπλός καταλογισμός, παρουσιάζει και τα πλεονεκτήματα των βασισμένων σε μοντέλα μεθόδων (Model – Based Methods) και απαιτεί λιγότερους περιορισμούς. Δίνει περισσότερες από μία τιμές στα ελλιπή στοιχεία (m = 5 – 10) αποκαθιστώντας την αβεβαιότητα που πρέπει να υπάρχει για τις τιμές των ελλιπών στοιχείων.

7

Μέθοδοι βασισμένες σε μοντέλα (Model Based Procedures)

7.1 Εισαγωγή

Σε αυτή την κατηγορία μεθόδων ανήκουν η εκτίμηση μέγιστης πιθανότητας (Maximum Likelihood) και οι Μπεϋζιανές Μέθοδοι (Bayesian Methods). Η κύρια μέθοδος εκτίμησης της μέγιστης πιθανοφάνειας που χρησιμοποιείται για τον χειρισμό δεδομένων με ελλιπή στοιχεία είναι ο αλγόριθμος EM (Expectation Maximization), ενώ οι δύο πιο αντιπροσωπευτικές Bayesian μέθοδοι είναι το Data Augmentation και Gibbs' Sampler. Ο πολλαπλός καταλογισμό, όπως έχει ειπωθεί ανήκει τόσο στην κατηγορία των μεθόδων αντικατάστασης, όσο και στην κατηγορία των βασισμένων σε μοντέλα μεθόδων καθώς χρησιμοποιεί Bayesian μεθόδους (Data Augmentation) όσο

και τον αλγόριθμο EM για το βήμα του καταλογισμού (Imputation Step) της διαδικασίας.

7.2 Εκτίμηση Μέγιστης πιθανοφάνειας

7.2.1 Εισαγωγή

Η εκτίμηση της μέγιστης πιθανοφάνειας χρησιμοποιείται στη Στατιστική για την εκτίμηση των παραμέτρων των μοντέλων σε προβλήματα με ή χωρίς ελλιπή δεδομένα. Σε αυτή την μέθοδο στηρίζονται και πολλές άλλες στατιστικές διαδικασίες. Ο βασικός της στόχος είναι να προσδιορίσει τις τιμές των παραμέτρων του πληθυσμού, κυρίως παράγοντας ένα συγκεκριμένο δείγμα δεδομένων.

Η μέθοδος αυτή περιλαμβάνει μία επαναληπτική διαδικασία όπου δοκιμάζονται διαφορετικές τιμές για τις παραμέτρους που πρέπει να εκτιμηθούν με σκοπό να βρεθούν οι καταλληλότερες τιμές για το συγκεκριμένο δείγμα. Η προσαρμογή των δεδομένων σε ένα συγκεκριμένο σύνολο τιμών για τις παραμέτρους εκτιμάται από τον λογάριθμο πιθανότητας της τιμής που υπολογίζει τη σχετική πιθανότητα του συγκεκριμένου δείγματος, που δίνεται από τα δεδομένα που προέρχονται από ένα κανονικά κατανεμημένο πληθυσμό.

Αναλυτικότερα, έστω ότι το Y αντιπροσωπεύει τα δεδομένα, που μπορεί να είναι απλές τιμές, είτε διανυσματικές. Τα δεδομένα μπορεί να υποτεθεί ότι δίνονται από ένα μοντέλο που περιγράφεται από μία πιθανότητα $p(\theta/Y)$, που προσδιορίζεται από μία μονόμετρη ή διανυσματική παράμετρο θ, το οποίο ανήκει σε ένα διάστημα Ω_θ. Όταν δίνεται το μοντέλο και η παράμετρος θ, το $p(\theta/Y)$ είναι μία συνάρτηση του Y που δίνει την πιθανότητα ή την πυκνότητα για διάφορες τιμές του θ.

Δεδομένου ότι το Y αντιπροσωπεύει τις τιμές των δεδομένων, ορίζεται ως συνάρτηση πιθανότητας το L(θ/Y) κάθε συνάρτηση του $\theta \in \Omega_\theta$ ανάλογη της p(Y/θ), για την οποία ισχύει L(Y/θ) = 0 για κάθε $\theta \notin \Omega_\theta$. Η συνάρτηση πιθανότητας είναι μία συνάρτηση της παραμέτρου θ για σταθερά Y, όπου η πιθανότητα ή η πυκνότητα είναι μία συνάρτηση του Y για σταθερά θ.

Ως εκτίμηση της μέγιστης πιθανοφάνειας (MLE) του θ, ορίζεται η τιμή του θ για την οποία μεγιστοποιείται η πιθανότητα p(Y/θ). Για αναλυτικούς σκοπούς είναι καλύτερο και ευκολότερο να χρησιμοποιηθεί ο λογάριθμος της πιθανότητας, καθώς είναι αύξουσα μονότονη συνάρτηση, και η εκτιμώμενη τιμή του θ $(\hat{\theta})$ που μεγιστοποιεί το λογάριθμο της πιθανότητας, μεγιστοποιεί και την πιθανότητα. Ορίζεται λοιπόν ο φυσικός λογάριθμος της συνάρτησης πιθανότητας: $\ell(\theta / Y) = \ell n L(\theta / Y)$.

Αυτός ο ορισμός επιτρέπει την πιθανότητα περισσότερων από μία τιμών για την εκτίμηση της μέγιστης πιθανότητας. Ωστόσο για πολλά μοντέλα μπορεί η λύση να είναι μοναδική. Αν η πιθανότητα είναι διαφορίσιμη συνάρτηση του θ, η τιμή $\hat{\theta}$ που μεγιστοποιεί τη συνάρτηση μπορεί να βρεθεί διαφορίζοντας τη συνάρτηση ως προ το θ. Ένα σύνολο από απαιτούμενες συνθήκες για την μέγιστη εκτίμηση πιθανοφάνειας για το θ μπορεί να εξαχθεί μηδενίζοντας την παράγωγο:

$$D_\ell(\theta) = 0, \quad \text{όπου } D_\ell(\theta) = \frac{\partial \ell(\theta / Y)}{\partial \theta}$$

7.2.2 Εκτίμηση Μέγιστης πιθανότητας σε περιπτώσεις με ελλιπή στοιχεία

Όταν τα δεδομένα παρουσιάζουν ελλιπή στοιχεία η εκτίμηση μέγιστης πιθανότητας ακολουθεί την ίδια διαδικασία με αυτή που ακολουθεί όταν τα δεδομένα είναι πλήρη. Οι εκτιμήσεις της μέγιστης πιθανοφάνειας βρίσκονται επιλύοντας την εξίσωση πιθανότητας και κατόπιν δημιουργείται μία μεταγενέστερη κατανομή ενσωματώνοντας στην αρχική κατανομή τις απαραίτητες αλλαγές.

Αναλυτικότερα, το Y όπως και παραπάνω αντιπροσωπεύει τα δεδομένα, ενσωματώνοντας και τα ελλιπή δεδομένα, δηλαδή $Y = (Y_{obs}, Y_{mis})$, Yobs τα παρατηρούμενα δεδομένα και Ymis τα ελλιπή δεδομένα.

Έστω $p(Y/\theta) \equiv p(Y_{obs}, Y_{mis}/\theta)$ υποδηλώνει την πιθανότητα ή πυκνότητα της κοινής κατανομής των παρατηρούμενων και των ελλιπών τιμών. Όταν είναι δυνατόν να αγνοηθεί ο μηχανισμός έλλειψη, δηλαδή αν τα δεδομένα λείπουν τυχαία (MAR) τα περιθώρια κατανομής των παρατηρούμενων δεδομένων, $p(Y_{obs}/\theta) = \int p(Y_{obs}, Y_{mis}/\theta) dY_{mis}$, παρέχουν την σωστή πιθανότητα για την παράμετρο θ. Ο Little kai Rubin (1987) αναφέρουν αυτή τη συνάρτηση ως "the likelihood ignoring the missing data mechanism", ενώ οι Schafer και Graham (2002) ως "observed data likelihood". Ο λογάριθμος αυτής της συνάρτησης:

$$\ell(\theta / Y_{obs}) = \log L(\theta / Y_{obs})$$

παίζει σημαντικό ρόλο στην εκτίμηση του θ. Η εκτίμηση μέγιστης πιθανότητας του $\hat{\theta}$, δηλαδή η τιμή του θ για την οποία η παραπάνω εξίσωση γίνεται μέγιστη, έχει τις ίδιες ελκυστικές θεωρητικές ιδιότητες όπως και όταν τα δεδομένα είναι πλήρη. Αν τηρούνται οι γενικοί όροι τείνει να μην παρουσιάζει αποκλίσεις με την προϋπόθεση

το δείγμα να είναι μεγάλο. Επίσης είναι ιδιαίτερα ελκυστικό το γεγονός ότι όσο το μέγεθος του δείγματος αυξάνεται, η διασπορά πλησιάζει το θεωρητικά χαμηλότερο όριο που μπορεί να επιτευχθεί από οποιοδήποτε εκτιμητή χωρίς αποκλίσεις.

Η μεγιστοποίηση της παραπάνω εξίσωσης, για την εκτίμησης της μέγιστης πιθανοφάνειας είναι δύσκολο να υπολογιστεί άμεσα παρά μόνο σε λίγα προβλήματα, όπως το πρόβλημα της διμεταβλητής μονότονης κατανομής (bivariate normal monotone). Εκτός λοιπόν από ορισμένες ειδικές περιπτώσεις, οι εκφράσεις για την εκτίμηση της μέγιστης πιθανοφάνειας δε μπορούν να γραφτούν με απόλυτα και ο υπολογισμό τους απαιτεί μία επαναληπτική διαδικασία (Schafer και Graham 2002).

Για το λόγο αυτό αναπτύχθηκε ο EM αλγόριθμος (Expectation Maximization algorithm). Η βασική ιδέα του αλγόριθμου EM είναι η επίλυση των δύσκολων προβλημάτων εκτίμησης με ημιτελή στοιχεία, με επαναληπτικές επιλύσεις ενός πιο εύκολου προβλήματος που διαθέτει όλα τα στοιχεία.

7.2.3 Ο αλγόριθμος EM (Expectation Maximization)

Ο αλγόριθμος EM προτάθηκε το 1977 από τους Dempster, Laird και Rubin ως μία μέθοδος της εκτίμησης Μέγιστης Πιθανότητας (ML) για τον πίνακα συνδιακύμανσης και του διανύσματος της μέσης τιμής αλλά από τότε έχει χρησιμοποιηθεί σε ένα μεγάλο αριθμό στατιστικών πακέτων. Η τεχνική αυτή είναι μία επαναληπτική διαδικασία μέσα από την οποία εκτιμώνται τα ελλιπή δεδομένα και οι παράμετροι του μοντέλου και κερδίζει όλο και περισσότερο έδαφος λόγω των καλών αποτελεσμάτων που έχει και χρησιμοποιείται σε πολλά στατιστικά πακέτα.

Σε αυτό το σημείο θα πρέπει να γίνει διαχωρισμός μεταξύ των δύο φαινομενικά διαφορετικών προσεγγίσεων του αλγόριθμου EM. Η πρώτη προσέγγιση, η οποία θα μπορούσε να αναφερθεί και ως "ευθύς EM" ("direct EM") χρησιμοποιείται για να εκτιμήσει τις παραμέτρους ενός γραμμικού στατιστικού μοντέλου. Για παράδειγμα χρησιμοποιείται σε συνδυασμό με ένα ιεραρχικό γραμμικό μοντέλο και ένα μοντέλο δομικής εξίσωσης. Αυτή η προσέγγιση διαφέρει λίγο από την πρώτη (1997) που προτάθηκε επειδή το σύνολο των δεδομένων που χρησιμοποιείται για το μοντέλο μπορεί να μην έχει ελλιπή στοιχεία. Έτσι σε αυτή την περίπτωση ο αλγόριθμος EM θεωρεί ως ελλιπή στοιχεία τις παραμέτρους του μοντέλου που πρέπει να εκτιμηθούν. Αντίθετα ο "παραδοσιακός" αλγόριθμος EM εκτιμά τον πίνακα συνδιακύμανσης και το διάνυσμα μέσης τιμής με βάση ένα σύνολο δεδομένων με ελλιπή στοιχεία.

Ο αλγόριθμος EM αναπτύχθηκε, καθώς ήταν δύσκολο να μεγιστοποιηθεί η εξίσωση που χρησιμοποιείται στη μέθοδο της μέγιστης πιθανότητας.

Ο αλγόριθμος EM είναι μία επαναληπτική διαδικασία που διαχωρίζει το πρόβλημα σε δύο επιμέρους προβλήματα εκτίμησης και περιλαμβάνει δύο βήματα, το βήμα της μεγιστοποίησης ή βήμα M (Maximization) και το βήμα της πρόβλεψης ή βήμα E (Expectation), τα οποία εναλλάσσονται κυκλικά. Η διαδικασία σταματά όταν επέλθει σύγκλιση στις εκτιμώμενες τιμές.

Το βήμα M (Maximization) είναι αρκετά απλό στην περιγραφή. Σε αυτό το βήμα εκτελείται η εκτίμηση της μέγιστης πιθανότητας (ML) για την παράμετρο θ, όπως έχει περιγραφεί παραπάνω στην περίπτωση που δεν υπάρχουν ελλιπή στοιχεία. Δηλαδή σε αυτό το βήμα του EM

αλγόριθμου χρησιμοποιείται η ίδια υπολογιστική μέθοδος με το μοντέλο μέγιστης πιθανότητας ($\ell(\theta / Y)$).

Το βήμα Ε (Expectation) βρίσκει την υπό όρους πρόβλεψη για τα "ελλιπή στοιχεία" που δίνεται από τα παρατηρούμενα στοιχεία και τις τρέχουσες εκτιμώμενες παραμέτρους. Κατόπιν αντικαθιστά αυτές τις προβλέψεις για τα "ελλιπή στοιχεία". Ο όρος "ελλιπή στοιχεία" στον αλγόριθμο ΕΜ δεν υπονοεί απαραίτητα την αντικατάσταση των πραγματικών ελλιπών στοιχείων και γι' αυτό μπαίνει σε εισαγωγικά.

Η βασική ιδέα του αλγόριθμου ΕΜ σκιαγραφείται από μία ειδική ιδέα επαναληπτικής συμπλήρωσης των ελλιπών δεδομένων, όπου τα ελλιπή δεδομένα δεν είναι αυτά που πραγματικά λείπουν (Ymis), αλλά οι συναρτήσεις αυτών των δεδομένων που εμφανίζονται στην εξίσωση πιθανότητας των πλήρως συμπληρωμένων στοιχείων ($\ell(\theta / Y)$).

7.2.3.1 Τα βήματα του αλγόριθμου ΕΜ

Αναλυτικότερα, έστω ότι $\theta^{(t)}$ η τρέχουσα πρόβλεψη της παραμέτρου θ. Το *βήμα Ε* (Expectation) του αλγορίθμου ΕΜ βρίσκει την πιθανότητα των πλήρως συμπληρωμένων στοιχείων αν η τιμή της παραμέτρου θ ήταν $\theta^{(t)}$:

$$Q(\theta / \theta^{(t)}) = \int \ell(\theta / Y) \cdot f(Y_{mis} / Y_{obs}, \theta = \theta^{(t)}) dY_{mis}$$

Το *βήμα Μ* (Maximization) του αλγόριθμου ΕΜ καθορίζει την επόμενη εκτίμηση της παραμέτρου θ, δηλαδή το $\theta^{(t+1)}$ μεγιστοποιώντας την παραπάνω συνάρτηση πιθανότητας.

$$Q(\theta^{(t+1)} / \theta^{(t)}) \geq Q(\theta / \theta^{(t)}), \text{ για όλα τα } \theta$$

7.2.3.2 Τα βασικά προβλήματα του αλγόριθμου EM

Στη θεωρία οι μέθοδοι πιθανότητας είναι πιο ελκυστικές από τις εμπειρικές τεχνικές της διαγραφή περίπτωσης και της απλής απόδοσης. Ωστόσο έχουν και αυτές κάποιες σημαντικές προϋποθέσεις. Πρώτον, θεωρούν ότι το δείγμα είναι αρκετά μεγάλο, ώστε οι εκτιμήσεις που θα προέλθουν από το ML να μην παρουσιάζουν αποκλίσεις και να είναι κανονικά κατανεμημένες. Στα προβλήματα των ελλιπών στοιχείων το δείγμα μπορεί να πρέπει να είναι αρκετά μεγαλύτερο από το συνηθισμένο, επειδή οι ελλιπείς τιμές μειώνουν το μέγεθος του δείγματος.

Δεύτερον, η συνάρτηση πιθανότητας προέρχεται από ένα υποθετικό παραμετρικό μοντέλο για τα πλήρη στοιχεία p(Yobs,Ymis/θ). Στηριζόμενοι σε συγκεκριμένες εφαρμογές οι μέθοδοι πιθανότητας μπορεί ή μπορεί να μην είναι σταθερό οι προϋποθέσεις του μοντέλου. Μερικές φορές (δομική συνάρτηση του μοντέλου) η αποκλίσεις μπορεί να μην έχουν σημαντική επίδραση στις εκτιμήσεις αλλά μπορεί να προκαλέσουν τυπικά σφάλματα και οι στατιστικές δοκιμές να είναι λάθος. Αν κάποιος διανέμει με ολόκληρο το παραμετρικό μοντέλο, η διαδικασία εκτίμησης με τα ημιτελή στοιχεία είναι δυνατή, αλλά τυπικά προτιμάται τα ελλιπή στοιχεία να ακολουθούν τον μηχανισμό MCAR παρά τον μηχανισμό MAR (K. H. Yuan & Bentler, 2000; Zeger, Liang, & Albert, 1988).

Τέλος, οι μέθοδοι πιθανότητας επιθυμούν δεδομένα που να ακολουθούν τον μηχανισμό MAR. Όταν η έλλειψη δεν ρυθμίζεται από τους ερευνητές, είναι πιθανόν να μην ακολουθείται ο μηχανισμός MAR. Σε πολλές ρεαλιστικές

εφαρμογές, ωστόσο πιστεύεται ότι οι αποκλίσεις από τον μηχανισμό MAR δεν είναι αρκετά μεγάλες ώστε να ακυρώνουν τα αποτελέσματα μίας ανάλυσης που στηρίζεται σε MAR μηχανισμό (Collins et al., 2001). Όταν οι λόγοι της έλλειψης είναι ισχυρά συνδεμένοι με τα δεδομένα, μπορεί να διατυπωθεί μία πιθανοτική ή Μπευζιανή λύση ώστε να ληφθεί υπόψη.

Το σημαντικότερο πρόβλημα του αλγόριθμου EM είναι ότι δεν δίνει τυπικά σφάλματα για τις παραμέτρους που εκτιμάει. Αν εξαιρέσεις αυτό θεωρείται ότι δίνει τις καλύτερες εκτιμήσεις μέγιστης πιθανότητας όταν τα δεδομένα λείπουν τυχαία (MAR).

7.2.4 Η μέθοδος FIML

Η μέθοδος Πλήρους Πληροφορίας Μέγιστη Πιθανοφάνειας ή Η FIML / Full Information Maximum Likelihood αντιπροσωπεύει μία πρωταρχική μέθοδο για την εκτίμηση μέσης τιμής και της συνδιακύμανσης σε προβλήματα με ελλιπή στοιχεία όταν αυτά ακολουθούν τον μηχανισμό MAR.

Η μέθοδος FIML έχει τις ρίζες της στην εργασία των Hartley και Hocking (1971). Λαμβάνοντας υπόψη τις q ομάδες, μια για κάθε τύπο των χαμένων δεδομένων, το FIML υπολογίζει αρχικά την πιθανότητα για κάθε μια από τις q ομάδες. Εάν δεν υπάρχει κανένα ελλείπον στοιχείο, τότε θέτουμε $q=1$. Η πρόθεση του FIML είναι να χρησιμοποιήσει την πληροφορία από κάθε τύπο χαμένων δεδομένων για να εκτιμήσει τις παραμέτρους. Στις περιπτώσεις που ισχύει η υπόθεση της πολύ-μεταβλητής κανονικότητας, οι q πιθανότητες αθροίζονται. Το αποτέλεσμα της άθροισης των πιθανοτήτων χρησιμεύει ως η βάση για την εκτίμηση των παραμέτρων με χρήση μέγιστης πιθανοφάνειας.

Η FIML μέθοδος είναι εννοιολογικά παρόμοια με την μέθοδο των q-γκρουπ των Hartley και Hocking (1971), εκτός από το ότι η πιθανότητα υπολογίζεται για κάθε παρατήρηση, χρησιμοποιώντας οποιαδήποτε δεδομένα είναι διαθέσιμα για την συγκεκριμένη παρατήρηση. Στις περιπτώσεις της πολύ-μεταβλητής κανονικής κατανομής οι μέθοδοι MAR και FIML υπολογίζουν μια πιθανότητα για κάθε περίπτωση χρησιμοποιώντας όλα τα διαθέσιμα δεδομένα. Αφότου έχουν υπολογιστεί N πιθανότητες, αθροίζονται και ο εκτιμητής μέγιστης πιθανότητα χρησιμοποιείται για να υπολογίσει τους μέσους, τις διασπορές και τις συνδιακυμάνσεις βασισμένες στην αθροισμένη πιθανότητα. Αυτές οι εκτιμήσεις των παραμέτρων θα είναι αμερόληπτες και επαρκείς παρουσία ελλειπόντων δεδομένων υπό τις υποθέσεις MAR και πολύ-μεταβλητής κανονικότητας.

7.2.5 Λογισμικά που διαθέτουν την Μέγιστη πιθανοφάνεια

Εκτός από τον αλγόριθμο EM κατά καιρούς έχουν προταθεί και έχουν δοθεί στην αγορά διάφορες διαδικασίες μέγιστης πιθανότητας για τον χειρισμό προβλημάτων με ελλιπή στοιχεία. Μερικές από αυτές παρουσιάζονται παρακάτω.

Σύμφωνα με τους Collins, Schaffer και Kam στη δεκαετία του 80 η εκτίμηση μέγιστης πιθανοφάνειας για ένα πίνακα συνδιακύμανσης από ημιτελή πολυμεταβλητά δεδομένα εκτελείται με το BMDP (BMDP statistical software, 1992), ένα μεγάλης κλίμακας στατιστικό πακέτο, το οποίο το 2000 ενσωματώθηκε στο SPSS. Μια παρόμοια διαδικασία προτάθηκε το 1991 από τους Graham και Hofer, το οποίο διατέθηκε δωρεάν και είναι γνωστό ως EMCOV. Η διαδικασία είναι επίσης διαθέσιμη στο NORM, SAS (Y.C. Yuan, 2000), Amelia (King et al., 2001), SPLUS (Schimert, Schafer, Hesterbrg, Fraley & Clarkson, 2000), LISREL (jöreskog & Sörbom, 2001)και Mplus (L.K. Muthén & Muthén, 1998),

Επίσης η εκτίμηση μέγιστης πιθανότητας για τη δομική εξίσωση μοντελοποίησης (Stractual Equations Modelling, SEM) με ημιτελή δεδομένα είναι διαθέσιμα στο ελεύθερο λογισμικό MX (Neale, boker, Xie & Maes, 1999), AMOS (Arbuckle & Wothke, 1999), LISREL και Mplus, τα οποία υπολογίζουν τα συστηματικά σφάλματα της μέγιστης πιθανότητας για την δομική εξίσωση μοντελοποίησης από τα παρατηρούμενα ή τις αναμενόμενες πληροφορίες του πίνακα

Τα προγράμματα που περιλαμβάνουν την εκτίμηση μέγιστης πιθανότητας προσαρμόζουν τη συνήθη γραμμική παλινδρόμηση και ανάλυση της διακύμανσης (ANOVA), την επαληθευτική παραγοντική ανάλυση (Factor Analysis) και άλλα μοντέλα πολύπλοκων λανθανουσών μεταβλητών. Για παράδειγμα ένα μοντέλο παλινδρόμηση που καθορίζεται από ορισμένες ανεξάρτητες μεταβλητές και μία εξαρτημένη μεταβλητή παρέχεται στο AMOS. Όταν το μοντέλο είναι καθορισμένο, ο χρήστης συμπληρώνει ένα πίνακα με ελλιπή στοιχεία που υποδηλώνεται από ένα συγκεκριμένο κώδικα. Η διαδικασία της Μέγιστης Πιθανότητας τότε υπολογίζει τις εκτιμήσεις των παραμέτρων βασιζόμενη σε όλα τα διαθέσιμα στοιχεία, συμπεριλαμβανομένου και τις ημιτελείς περιπτώσεις. Τα συστηματικά σφάλματα από τα παρατηρούμενα ή τις αναμενόμενες πληροφορίες του πίνακα.

Με αυτά τα προγράμματα τα προβλήματα με ελλιπή στοιχεία λύνονται όπως και τα προβλήματα με πλήρη στοιχεία. Θα πρέπει όμως όσοι τα χρησιμοποιούν να είναι πολύ προσεκτικοί στους περιορισμούς αυτών των πακέτων και σίγουρα τα ελλιπή δεδομένα του δείγματος να ακολουθούν τον μηχανισμό έλλειψης MAR.

7.3 Bayesian Methods

7.3.1 Γενικά

Τα τελευταία δέκα έτη, οι ερευνητές στις κοινωνικές επιστήμες και στην ιατρική άρχισαν να χρησιμοποιούν όλο και περισσότερο τις Μπεϋζιανές στατιστικές μεθόδους. Οι κλασσικές στατιστικές μέθοδοι στηρίζονται στις συγκρίσεις των στατιστικών που παράγονται από τα στοιχεία των δειγμάτων στις υποτιθέμενες παραμέτρους του πληθυσμού, που κάθε μια υποτίθεται ότι είναι μια άγνωστη αλλά σταθερή αξία.

Σε αντίθεση με την κλασσική στατιστική θεωρία, η Μπεϋζιανή θεωρία στηρίζεται στην έννοια της υποκειμενικής πιθανότητας. Δηλαδή, οι Μπεϋζιανοί αναλυτές θεωρούν ότι οι παράμετροι (θ) είναι τυχαίες μεταβλητές και ενσωματώνουν τα στοιχεία και τις πληροφορίες που διαθέτουν για μια παράμετρο σε μία αρχική κατανομή των πιθανοτήτων.

Με άλλα λόγια, οι παράμετροι ενός πληθυσμού θα ποικίλουν ανάλογα με την υποκειμενική πιθανότητά τους. Μόλις συλλεχθούν τα στοιχεία, μια μαθηματική ρουτίνα που προέρχεται από το θεώρημα Baye s χρησιμοποιείται για να συνδυάσει την αρχική κατανομή με τα νέα στοιχεία για να παραγάγει μια ενημερωμένη μεταγενέστερη κατανομή, στην οποία τα συμπεράσματα στη συνέχεια απεικονίζονται με προσοχή.

Αν ο αναλυτής υποθέσει μια ομοιόμορφη (επίσης αναφερόμενος ως διάχυτος ή noninformative) αρχική κατανομή στην οποία κάθε αξία της κατανομής θεωρείται ότι έχει μια εξίσου πιθανή πιθανότητα, η μεταγενέστερη διανομή είναι ασυμπτωτικά ισοδύναμη με την πιθανότητα των παρατηρηθέντων στοιχείων. Στην Μπεϋζιανή ανάλυση, οι ελλιπείς τιμές στοιχείων αντιμετωπίζονται ως πρόσθετες παράμετροι που υπολογίζονται, με βάση τους περιορισμούς του μοντέλου που χρησιμοποιεί ο αναλυτής και τις επιλεγμένες αρχικές κατανομές των παραμέτρων του μοντέλου. Σύμφωνα με Burton et al (1998) "Οι βασισμένες στην Μπεϋζιανή θεωρία διαδικασίες

είναι διαισθητικές και μπορούν να οδηγήσουν σε πιο σωστές αποφάσεις" από την παραδοσιακή τεχνική της μηδενική υπόθεσης. Ένας λόγος για την αυξανόμενη χρήση των Μπεϋζιανών μεθόδων είναι σημαντικές βελτιώσεις στη δύναμη των υπολογιστών, καθώς ένα από τα μειονεκτήματά τους είναι ότι απαιτούν μεγάλη υπολογιστική δύναμη.

Οι δύο κύριες μέθοδοι που χρησιμοποιούν την Μπεϋζιανή θεωρία για τον χειρισμό δεδομένων με ελλιπή στοιχεία είναι ο αλγόριθμος Data Augmentation (DA) και η μέθοδος Gibbs' Sampler. Και οι δύο αυτές μέθοδοι είναι επαναληπτικές διαδικασίες, οι οποίες δίνουν ικανοποιητικά αποτελέσματα όταν η επανάληψη είναι μεταξύ παραμέτρων του μοντέλου και των ελλιπών στοιχείων.

7.3.2 Ο αλγόριθμος DA

Ο αλγόριθμος DA του Schafer περιλαμβάνει μια επαναληπτική διαδικασία δύο βημάτων. Η διαδικασία DA αρχίζει με μια αρχική εκτίμηση του πίνακα συνδιακύμανσης, δηλαδή μια εκτίμηση μέγιστης πιθανότητας (ML) που λαμβάνεται μέσω του αλγορίθμου EM.

Το πρώτο βήμα της διαδικασίας είναι το βήμα καταλογισμού ή βήμα I (step I / Imputation step) στο οποίο τα ελλιπή στοιχεία αντικαθιστούνται από κάποιες τιμές που έχουν προβλεφθεί μετά από μία σειρά πολλαπλών εξισώσεων παλινδρόμησης. Επειδή οι καθ' υπολογισμό τιμές αφορούν άμεσα μια επιφάνεια παλινδρόμησης, προστίθεται μια τυχαία διαταραχή σε κάθε αξία με τυχαία επιλογή ενός υπολειπόμενου όρου από μια κανονική κατανομή. Στη συνέχεια υπολογίζονται ένας νέος πίνακας συνδιακύμανσης και διάνυσμα μέσης τιμής από τα πλήρη στοιχεία χρησιμοποιώντας τυποποιημένους τύπους.

Προκειμένου να δημιουργηθούν τα καθ' υπολογισμό σύνολα στοιχείων με τις διαφορετικές εκτιμήσεις για τα ελλιπή στοιχεία, προστίθενται τυχαίες

διαταραχές και στις εξισώσεις παλινδρόμησης που χρησιμοποιούνται για να παραγάγουν τις εκτιμώμενες τιμές για τα ελλιπή στοιχεία. Αυτό εμφανίζεται στο «μεταγενέστερο βήμα» (Posterior Step) ή βήμα P (Step P), όπου ο νέος πίνακας συνδιακύμανσης και στοιχεία του διανύσματος της μέσης τιμής επιλέγονται με τυχαίο τρόπο από μια μεταγενέστερη κατανομή η οποία προκύπτει από τις τιμές των ολοκληρωμένων στοιχείων που λήφθηκαν στο προηγούμενο βήμα, βήμα I.

Οι καινούργιες εκτιμήσεις για τα Σ και μ χρησιμοποιούνται στο επόμενο βήμα, βήμα I (2η επανάληψη) για να υπολογιστούν οι νέες τιμές για τα ελλιπή στοιχεία με τη βοήθεια των νέων εξισώσεων παλινδρόμησης που διαφέρουν ελαφρώς από τις προηγούμενες. Αφού υπολογιστούν οι νέες τιμές χρησιμοποιούνται για την νέα κατανομή από την οποία επιλέγονται πάλι τυχαία ο πίνακας συνδιακύμανσης και το διάνυσμα μέσης τιμής, βήμα P.

Αναλυτικότερα σύμφωνα με τους Rubin και Little (2002) η μεταγενέστερη κατανομή για ένα μοντέλο με ελλιπή στοιχεία που ο μηχανισμός έλλειψής τους μπορεί να αγνοηθεί είναι:

$$p(\theta / Y_{obs}, M) \equiv p(\theta / Y_{obs}) = \text{constant} \times p(\theta) \times f(Y_{obs} / \theta)$$

όπου $p(\theta)$ είναι η αρχική κατανομή και $f(Y_{obs} / \theta)$ είναι η πυκνότητα για τα παρατηρούμενα στοιχεία. Συγκεκριμένα η πιθανότητα παραγοντοποιείται στα πλήρη στοιχεία,

$$L(\phi / Y_{obs}) = \prod_{q=1}^{Q} L_q(\phi_q / Y_{obs})$$

και θεωρώντας ότι οι παράμετροι $\Phi_1,\ldots\ldots,\Phi_Q$ είναι επίσης εκ των προτέρων ανεξάρτητοι, η μεταγενέστερη κατανομή παραγοντοποιείται με ανάλογο τρόπο με τα $\Phi_1,\ldots\ldots,\Phi_Q$ εκ των υστέρων ανεξάρτητα. Ως συνέπεια οι

απεικονίσεις $\phi^{(d)} = (\phi_1^{(d)}, \phi_2^{(d)}, ..., \phi_Q^{(d)})$ μπορούν να βρεθούν απευθείας από την παραγοντοποιημένη μεταγενέστερη κατανομή των ολοκληρωμένων δεδομένων. Οι απεικονίσεις του θ τότε θα ληφθούν τότε ως $\theta^{(d)} = \theta(\phi^{(d)})$, όπου θ(Φ) είναι ο αντίστροφος μετασχηματισμός από το Φ στο θ.

Ο αλγόριθμος DA όπως αναφέρθηκε και παραπάνω είναι μία επαναληπτική διαδικασία μεταξύ δύο βημάτων (Step I/Step P). Ξεκινάει με μία αρχική απεικόνιση του $\theta^{(0)}$ από μία εκτίμηση της εκ των υστέρων κατανομής του θ. Δίνει μία τιμή θ(t) του θ που απεικονίζεται στο βήμα t:

Βήμα Ι: Απεικόνιση $Y_{mis}^{(t+1)}$ με πυκνότητα $p(Y_{mis} / Y_{obs}, \theta^{(t)})$

Βήμα Ρ: Απεικόνιση $\theta^{(t+1)}$ με πυκνότητα $p(\theta / Y_{obs}, Y_{mis}^{(t+1)})$

Η διαδικασία υποκινείται από το γεγονός ότι η κατανομή σε αυτά τα δύο βήματα είναι πολλές φορές πιο εύκολο να απεικονισθεί από την μεταγενέστερες κατανομές $p(Y_{mis} / Y_{obs})$ και $p(\theta / Y_{obs})$, ή την συνδυασμένη μεταγενέστερη κατανομή $p(\theta, Y_{mis} / Y_{obs})$. Η επαναληπτική διαδικασία μπορεί να δειχτεί τελικά ότι αποδίδει μία απεικόνιση από την συνδυασμένη μεταγενέστερη κατανομή των θ, Y_{mis} δίνοντας το Y_{obs}, υπό την έννοια ότι όσο το t τείνει στο άπειρο, αυτή η ακολουθία συγκλίνει προς μία απεικόνιση από την μεταγενέστερη κατανομή για δοσμένο Y_{obs}.

Εδώ θα πρέπει να αναφερθεί ότι όταν αυτός ο αλγόριθμος χρησιμοποιείται στον πολλαπλό καταλογισμό τη δημιουργία των m συνόλων με πλήρη δεδομένα (Imputation βήμα) θα πρέπει να επαναληφθεί πολλές φορές, καθώς για τη δημιουργία των m συνόλων απαιτούνται περισσότερες τιμές για τα ελλιπή στοιχεία. Συγκεκριμένα οι τιμές που δίνονται στα ελλιπή στοιχεία σώζονται σε διευκρινισμένα διαστήματα στην ακολουθία. Για να δημιουργηθεί ο επιθυμητός αριθμός συνόλων με ολοκληρωμένα στοιχεία είναι σημαντικό η διαδικασία του DA να επαναληφτεί περισσότερες φορές από τον αριθμό των συνόλων με ολοκληρωμένα στοιχεία (m) που

επιθυμείται να ληφθούν, καθώς οι τιμές που θα ληφθούν για τα ελλιπή δεδομένα από δύο συνεχόμενα βήματα I του αλγόριθμου θα συσχετιστούν. Ένας αριθμός επαναλήψεων πρέπει να χωρίσει τις τιμές που υπολογίστηκαν για τα ελλιπή στοιχεία και θα χρησιμοποιηθούν στη φάση της ανάλυσης (αυτά αναφέρονται μερικές φορές ως *επαναλήψεις between imputation*). Επίσης θα πρέπει αρχικά να περάσουν κάποιοι κύκλοι βημάτων του αλγόριθμου για να ληφθεί το πρώτο σύνολο με ολοκληρωμένα στοιχεία που θα χρησιμοποιηθεί για το στάδιο της ανάλυσης (αυτοί αναφέρονται ως *επαναλήψεις καψίματος*). Ο καθορισμός του αριθμού επαναλήψεων καψίματος μπορεί να είναι σύνθετος και περιλαμβάνει αξιολόγηση της σύγκλισης (δηλ. σταθερότητα) των παραμέτρων που επιλέγονται στο βήμα P.

7.3.3 Η μέθοδος Gibbs' Sampler

Μία άλλη μέθοδος που χρησιμοποιεί την Μπεϋζιανή θεωρία για τον χειρισμό προβλημάτων με ελλιπή στοιχεία είναι η μέθοδος Gibbs' Sampler. Η μέθοδος Gibbs' Sampler είναι μία επαναληπτική μέθοδος παραγωγής τυχαίων μεταβλητών από μία (οριακή) κατανομή με άμεσο τρόπο, χωρίς να υπολογίζει τις πυκνότητες. Έτσι με τη χρήση αυτής της μεθόδου μπορούν να αποφευχθούν οι δύσκολοι υπολογισμοί, αντικαθιστώντας τους με διαδοχικούς απλούστερους υπολογισμούς. Τελικά με την μέθοδο Gibbs' Sampler λαμβάνεται μία απεικόνιση από τη συνδυασμένη κατανομή στην περίπτωση της γενικής (General) μορφής έλλειψης.

Συγκεκριμένα σύμφωνα με τους Rubin και Little (2002) η μέθοδος Gibbs' Sampler παράγει μία απεικόνιση από την κατανομή $P(x_1, x_2, ..., x_p)$ ενός συνόλου p τυχαίων μεταβλητών $X_1, X_2, ..., X_p$, όταν η απεικονίσεις της συνδυασμένης κατανομής είναι δύσκολο να υπολογιστούν, αλλά οι απεικονίσεις της υπό όρους κατανομής $p(x_j / x_1, ..., x_{j-1} / x_{j+1}, ..., x_p)$, j=1,….,p, είναι σχετικά εύκολο να υπολογιστούν.

Οι αρχικές τιμές των $x_1^{(0)}, x_2^{(0)}, ..., x_p^{(0)}$ επιλέγονται με κάποιο τρόπο. Κατόπιν για τις δοσμένες τιμές $x_1^{(t)}, x_2^{(t)}, ..., x_p^{(t)}$ της επανάληψης t, νέες τιμές βρίσκονται με την απεικόνιση από την επόμενη ακολουθία της p υπό όρους κατανομής:

$$
\begin{aligned}
&x_1^{(t+1)} \sim p(x_1 \setminus x_2^{(t)}, x_3^{(t)}, ..., x_p^{(t)}) \\
&x_2^{(t+1)} \sim p(x_2 \setminus x_1^{(t+1)}, x_3^{(t)}, ..., x_p^{(t)}) \\
&x_3^{(t+1)} \sim p(x_3 \setminus x_1^{(t+1)}, x_2^{(t+1)}, x_4^{(t)} ....., x_p^{(t)}) \\
&x_4^{(t+1)} \sim p(x_4 \setminus x_1^{(t+1)}, x_2^{(t+1)}, x_3^{(t+1)}, x_5^{(t)} ....., x_p^{(t)}) \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&x_p^{(t+1)} \sim p(x_p \setminus x_1^{(t+1)}, x_2^{(t+1)}, x_3^{(t+1)} ....., x_{p-1}^{(t+1)})
\end{aligned}
$$

Μπορεί να δειχτεί ότι, υπό σχετικά γενικούς περιορισμούς, η ακολουθία των επαναλήψεων $x_t = (x_1^{(t)}, x_2^{(t)}, ..., x_p^{(t)})$ συγκλίνει σε μία απεικόνιση από την συνδυασμένη κατανομή των $X_1^t, X_2^t, ..., X_p^t$. Σε αυτή τη μέθο δ οτα ανεξάρτητα στοιχεία X_j μπορεί να είναι σύνολα μεταβλητών και όχι μόνο μονόμετρες μεταβλητές.

Όταν το p=2, η μέθοδος Gibbs' Sampler είναι όμοια με τον αλγόριθμο του Data Augmentation αν το $X_1 = Y_{mis}$ και $X_2 = \theta$, και οι υπό όρους κατανομές στο Y_{obs}. Τότε, στο όριο, λαμβάνεται μία απεικόνιση από την συνδυασμένη κατανομή $(Y_{mis}, \theta / Y_{obs})$ εφαρμόζοντας την μέθοδο Gibbs' Sampler, όπου στην επανάληψη t από του νιοστού (*n*th) συνόλου δεδομένων είναι:

$$Y_{mis}^{(n,t+1)} \sim p(Y_{mis} \setminus Y_{obs}, \theta^{n,t}), \ \theta^{(n,t+1)} \sim p(\theta \setminus Y_{mis}^{(n,t)} Y_{obs})$$

Όπως και στη μέθοδο του Data Augmentation κάθε επανάληψη περιλαμβάνει μία απεικόνιση από την μεταγενέστερη προβλεπόμενη κατανομή των ελλιπών στοιχείων (Y_{mis}) και μία απεικόνιση από το θ και τα ελλιπή στοιχεία (Y_{mis}). Η μέθοδος Gibbs' Sampler μπορεί να τρέξει N ανεξάρτητες φορές και να παράγει N απεικονίσεις από την κατά προσέγγιση συνδυασμένη κατανομή του θ και των ελλιπών στοιχείων (Y_{mis}).

Η μέθοδος του Gibbs' Sampler χρησιμοποιείται για πιο περίπλοκα προβλήματα σε σχέση με τα προβλήματα που επιλύονται με την μέθοδο Data Augmentation.

7.3.4 Πλεονεκτήματα – Μειονεκτήματα των Μπεϋζιανών Μεθόδων

Οι διαδικασίες που βασίζονται στην Μπεϋζιανή θεωρία και παρουσιάστηκαν παραπάνω χρησιμοποιούνται κυρίως σε προβλήματα όπου τα δείγματα είναι σχετικά μικρά και δεν μπορεί να χρησιμοποιηθεί η διαδικασία της μέγιστης πιθανότητας.

Όπως όλες τις μεθόδους που βασίζονται σε μοντέλα δίνουν ευελιξία στον αναλυτή και προσφέρουν ικανοποιητικά αποτελέσματα. Θεωρούνται πιο αποτελεσματικές από τις παραδοσιακές τεχνικές καθώς αξιοποιούν όλη την πληροφορία που προέρχεται από τα δεδομένα.

Η διαδικασία του Gibbs' Sampler χρησιμοποιείται σε πιο περίπλοκες περιπτώσεις όπου το Data Augmentation δεν μπορεί να δώσει ικανοποιητικά αποτελέσματα. Ωστόσο σε κάποιες περιπτώσεις, όπου το πρόβλημα είναι πολύ περίπλοκο η διαδικασία Gibbs' Sampler να μην μπορεί να υπολογίσει τις απεικονίσεις που προκύπτουν από την ακολουθία της υπό όρους κατανομής. Τότε είναι απαραίτητο να χρησιμοποιηθούν άλλες μέθοδοι προσομοίωσης.

Επίσης οι διαδικασίες Data Augmentation και το Gibbs' Sampler σε περιπτώσεις που τα στοιχεία ανανεώνονται (όπως στις μακροχρόνιες έρευνες) μπορεί να είναι αναποτελεσματικές, καθώς θα πρέπει η διαδικασία να γίνει πάλι από τηη αρχή για να συμπεριληφθούν και τα νέα στοιχεία.

Μία μέθοδος που χρησιμοποιείται για να επιλύσει αυτά τα προβλήματα είναι ο σειριακός καταλογισμός (Sequential Imputation) που επιτρέπει την πρόσθεση των νέων στοιχείων. Η μέθοδος αυτή μπορεί να χρησιμοποιηθεί είτε ως εναλλακτική, είτε ως συμπληρωματική των παραπάνω μεθόδων. Επίσης, έχουν αναπτυχθεί και άλλες διαδικασίες που βασίζονται στη θεωρία των Markov Chain Monte Carlo (MCMC). Οι διαδικασίες αυτές χρησιμοποιούνται συνήθως σε συνδυασμό με τις παραπάνω μεθόδους.

7.4 Συμπεράσματα

Οι διαδικασίες που βασίζονται σε μοντέλα (Model – Based Methods) χρησιμοποιούνται, στις μέρες μας, σε προβλήματα με ελλιπή στοιχεία. Τις τελευταίες δεκαετίες έχουν αναπτυχθεί πληθώρα τέτοιων μεθόδων με επικρατέστερες τον πολλαπλό καταλογισμό (Multiple Imputation), μοντέλα που στηρίζονται στην εκτίμηση μέγιστης πιθανότητας (Maximum Likelihood) και μοντέλα που προέρχονται από την Μπεϋζιανή θεωρία (Bayesian Methods).

Οι διαδικασίες αυτής της κατηγορίας βασίζονται σε ένα μοντέλο που προσδιορίζεται από τα παρατηρούμενα στοιχεία του δείγματος και οι τελικές εκτιμήσεις βασίζονται στην πιθανότητα ή σε μία μεταγενέστερη κατανομή αυτού του μοντέλου μέσα από μία συγκεκριμένη διαδικασία.

Οι μέθοδοι αυτής της κατηγορίας θεωρούνται πιο ελκυστικές από τις εμπειρικές – παραδοσιακές τεχνικές, καθώς χρησιμοποιούν για την εκτίμηση των αποτελεσμάτων όλη τη δυνατή πληροφορία που δίνεται από τα

στοιχεία του δείγματος, αυξάνοντας την στατιστική δύναμη των αποτελεσμάτων. Επίσης, απαιτούν λιγότερους περιορισμούς και τα αποτελέσματα τους εφαρμόζονται σε ένα ευρύτερο φάσμα, κάτι που προσφέρει μεγάλη ευελιξία σους ερευνητές.

Ωστόσο για την εφαρμογή τους απαιτείται μεγάλη υπολογιστική ισχύς, μπορεί να είναι χρονοβόρες (αργή σύγκλιση) και πρέπει να δειχθεί μεγάλη προσοχή στους περιορισμούς που θέτουν, κυρίως για των μηχανισμό έλλειψης των στοιχείων που πρέπει να μπορεί να αγνοηθεί (Ignorable).

8

Σύγκριση των μεθόδων για τον χειρισμό προβλημάτων με ελλιπή δεδομένα

Η εργασία αυτή ασχολήθηκε με το πρόβλημα των ελλιπών στοιχείων που εμφανίζονται κατά την στατιστική ανάλυση δεδομένων, καθώς και με τις μεθόδους που χρησιμοποιούνται για να χειριστούν τέτοια προβλήματα. Σε αυτό το κεφάλαιο θα γίνει προσπάθεια να τονιστούν τα βασικά σημεία που πρέπει να προσέξει κάποιος που αντιμετωπίζει το πρόβλημα των ελλιπών στοιχείων. Επίσης, θα συνοψιστούν τα βασικά στοιχεία κάθε μίας από τις μεθόδους που χρησιμοποιήθηκαν, δίνοντας τα πλεονεκτήματα και τα μειονεκτήματα κάθε μεθόδου και σε ποιες περιπτώσεις προτείνεται η χρήση της.

8.1 Τα ερωτήματα των ερευνητών που αντιμετωπίζουν το πρόβλημα των ελλιπών στοιχείων

Τα ημιτελή στοιχεία αποτελούν ένα συχνό φαινόμενο κατά τη στατιστική επεξεργασία και ανάλυση δεδομένων που προέρχονται από έρευνες, καθώς και πειραματικές διαδικασίες. Έχει αποδειχτεί ότι όσο προσεκτικά και αν σχεδιάσει κάποιος το πρόβλημα που ερευνά δεν θα μπορέσει να προβλέψει όλους τους αστάθμητους παράγοντες που θα οδηγήσουν στην εμφάνιση ελλιπών δεδομένων. Το πρόβλημα γίνεται μεγαλύτερο στις περιπτώσεις όπου οι ερευνητές ασχολούνται με ανθρωποκεντρικές έρευνες και προστίθεται η εξάρτηση από την υποκειμενική κρίση του ανθρώπου.

Οι ερευνητές προσπάθησαν να επιλύσουν αυτό το πρόβλημα, δηλαδή την εμφάνιση ελλιπών στοιχείων στα δεδομένα, με την ανάπτυξη συγκεκριμένων μεθόδων. Οι μέθοδοι αυτές χωρίστηκαν σε τέσσερις κατηγορίες με βάση τον τρόπο που χειρίζονται τα ελλιπή στοιχεία και αυτές είναι: α) μέθοδοι διαγραφής (Deletion Methods), β) μέθοδοι βαρών (Weighting Methods), γ) μέθοδοι καταλογισμού (Imputation Methods) και δ) μέθοδοι που στηρίζονται σε μοντέλα (Model Based Methods).

Μετά την ανάπτυξη αυτών των μεθόδων οι ερευνητές βρέθηκαν αντιμέτωποι με το ερώτημα ποια από αυτές τις μεθόδους είναι η καλύτερη για την διαχείριση δεδομένων με ελλιπή στοιχεία. Η απάντηση σε αυτό το ερώτημα είναι ότι δεν υπάρχει βέλτιστη μέθοδος που να έχει τα επιθυμητά αποτελέσματα σε κάθε πρόβλημα με ελλιπή στοιχεία. Ως συνέπεια κάθε πρόβλημα θα πρέπει να αντιμετωπίζεται ξεχωριστά και ανάλογα με τις συνθήκες να επιλέγεται η κατάλληλη για τη συγκεκριμένη περίπτωση μέθοδος.

Το πιο σημαντικό στοιχείο στην επιλογή κάποιας μεθόδου είναι ο μηχανισμός έλλειψης (MCAR, MAR, MNAR), καθώς όλες οι τεχνικές που χρησιμοποιούνται για αυτά τα προβλήματα θέτουν συγκεκριμένους περιορισμούς για τους μηχανισμούς έλλειψης που πρέπει να ακολουθούν τα ελλιπή στοιχεία, ώστε να παράγουν αποτελέσματα που να ανταποκρίνονται στην πραγματικότητα.

Για παράδειγμα, οι μέθοδοι της ανάλυσης πλήρως συμπληρωμένων περιπτώσεων (Complete Case Analysis) ή Listwise Deletion, διαθέσιμης περιπτωσιολογικής ανάλυσης (Available Case Analysis) ή Pairwise Deletion και η μέθοδος της αντικατάστασης με την μέση τιμή (Mean Substitution) επιβάλλουν ότι τα ελλιπή δεδομένα λείπουν τελείως τυχαία (MCAR). Αυτό σημαίνει ότι αν εφαρμοστούν αυτές οι μέθοδοι σε ένα πρόβλημα που τα ελλιπή ακολουθούν τον μηχανισμό MCAR μπορεί τα οι εκτιμήσεις των παραμέτρων να μην παρουσιάζουν αποκλίσεις. Αν όμως δεν έχει εξασφαλιστεί ο μηχανισμός MCAR και τα στοιχεία μπορεί να ακολουθούν τους άλλους μηχανισμούς έλλειψης (MAR, MNAR) αυτές οι μέθοδοι θα δώσουν λάθος αποτελέσματα με μεγάλες αποκλίσεις από την πραγματικότητα και δεν θα πρέπει να χρησιμοποιούνται.

Οι μέθοδοι που βασίζονται σε μοντέλα (Model – Based Methods) και συγκεκριμένα η μέθοδος της μέγιστης πιθανότητας, του πολλαπλού καταλογισμού και της Μπεϋζιανής εκτίμησης δίνουν ικανοποιητικές εκτιμήσεις όταν τα δεδομένα ακολουθούν τους μηχανισμούς MCAR και MAR. Αν ισχύουν αυτοί οι μηχανισμοί οι παραπάνω μέθοδοι συνήθως παράγουν εκτιμήσεις χωρίς αποκλίσεις.

Όταν τα δεδομένα ακολουθούν τον μηχανισμό MNAR δεν υπάρχει κάποια συγκεκριμένη μέθοδος που να μπορεί να εξαλείψει απόλυτα την απόκλιση των εκτιμήσεων. Οι Muthén et al. υποστηρίζουν ότι μέθοδοι όπως η εκτίμηση μέγιστης πιθανότητας ή παρόμοιες μπορούν να δώσουν εκτιμήσεις

με μικρότερες αποκλίσεις από εμπειρικές μεθόδους όπως η Listwise Deletion και η Pairwise Deletion.

8.2 Συνοπτική παρουσίαση των μεθόδων

Οι μέθοδοι διαχείρισης δεδομένων με ελλιπή στοιχεία μπορούν να χωρισθούν, όπως έχει ήδη αναφερθεί σε τέσσερις κατηγορίες: α) μέθοδοι διαγραφής (Deletion Methods), β) μέθοδοι βαρών (Weighting Methods), γ) μέθοδοι καταλογισμού (Imputation Methods) και δ) μέθοδοι που στηρίζονται σε μοντέλα (Model Based Methods).

Οι μέθοδοι που ανήκουν στην πρώτη κατηγορία (Μέθοδοι Διαγραφής) είναι η ανάλυση πλήρως συμπληρωμένων περιπτώσεων, Complete Case Analysis ή Listwise Deletion η οποία διαγράφει όποια περίπτωση παρουσιάζει ελλιπή στοιχεία και η ανάλυση του δείγματος γίνεται μόνο με τις πλήρως συμπληρωμένες περιπτώσεις και η Διαθέσιμη περιπτωσιολογική Ανάλυση, Available Case Analysis ή Pairwise Deletion, η οποία κάνει επιμέρους συσχετισμούς μεταξύ των μεταβλητών, διαγράφοντας κάθε φορά μόνο την περίπτωση που παρουσιάζει ελλιπή στοιχεία για εκείνο τον συσχετισμό και κάνοντας την ανάλυση με τις υπόλοιπες περιπτώσεις.

Οι μέθοδοι αυτής της κατηγορίας θεωρούνται ως οι πιο εύκολα χρησιμοποιούμενες και γρήγορες μέθοδοι στο χειρισμό δεδομένων με ελλιπή στοιχεία. Έχουν ικανοποιητικά αποτελέσματα όταν ο μηχανισμός έλλειψης είναι MCAR και τα ελλιπή στοιχεία αποτελούν ένα πολύ μικρό μέρος του δείγματος. Στον Πίνακα 1, στο τέλος του κεφαλαίου συνοψίζονται τα βασικά στοιχεία κάθε μίας από τις δύο αυτές μεθόδους.

Η δεύτερη κατηγορία περιλαμβάνει τις διαδικασίες στάθμισης, οι οποίες χρησιμοποιούνται με σκοπό τη μείωση των αποκλίσεων των εκτιμήσεων από αυτά που θα προέκυπταν αν τα δεδομένα ήταν πλήρη. Όπως και οι

διαδικασίες διαγραφής έτσι και οι διαδικασίες στάθμισης διαγράφουν τι περιπτώσεις που έχουν ελλιπή στοιχεία, όμως τα λαμβάνουν υπόψη θεωρώντας το ως μέρους του σχεδιασμένου προβλήματος, δίνοντας τους μία πιθανότητα ύπαρξης.

Αυτό επιτυγχάνεται με την εισαγωγή βαρών ώστε να δειχθεί η διαφορετική πιθανότητα παρατήρησης των τιμών των δεδομένων. Δίνεται λοιπόν και στα ελλιπή στοιχεία μία πιθανότητα παρατήρησης ανάλογα με την έκτασή τους στο συγκεκριμένο σύνολο δεδομένων. Οι διαδικασίες αυτής της κατηγορίας περιλαμβάνουν τρία στάδια. Στο πρώτο στάδιο παράγεται μία βάση βαρών με βάση την πιθανότητα παρατήρησης κάθε τιμής των στοιχείων. Κατόπιν προσαρμόζεται αυτή η βάση ώστε να ληφθούν υπόψη και τα ελλιπή στοιχεία και στο τρίτο στάδιο τα βάρη ρυθμίζονται περαιτέρω ώστε να εκτιμηθούν κάποιες τιμές για το σύνολο του πληθυσμού μειώνοντας τη στατιστική απόκλιση.

Οι διαδικασίες αυτές χρησιμοποιούνται συνήθως σε συνδυασμό με τις μεθόδους διαγραφής για την μείωση των αποκλίσεων. Για να εφαρμοστούν θα πρέπει πρώτα να έχει διασφαλιστεί ότι ο μηχανισμός έλλειψης που ακολουθούν τα ελλιπή δεδομένα είναι MAR.

Η τρίτη κατηγορία περιλαμβάνει τις μεθόδους αντικατάστασης που όπως φαίνεται και από το όνομά τους αντικαθιστούν τα ελλιπή δεδομένα με κάποιες τιμές που παράγονται με αυτές τις διαδικασίες. Σε αυτή την κατηγορία μπορούν να διακριθούν πολλές υποκατηγορίες. Αρχικά θα πρέπει να διαχωριστούν στις “παραδοσιακές” και στις ”μοντέρνες” μεθόδους. Οι παραδοσιακές μέθοδοι είναι εμπειρικές, ενώ η μέθοδος της κατηγορίας των “μοντέρνων” μεθόδων ανήκει και στην κατηγορία των βασισμένων σε μοντέλα μεθόδων (Model Based Methods).

Οι παραδοσιακές μέθοδοι βάση του τύπου αντικατάστασης των ελλιπών στοιχείων χωρίζονται σε τρεις κατηγορίες: Μέθοδοι μέσης τιμής (mean

based methods), Μέθοδοι παλινδρόμησης (regression based methods) και στις Hot Deck Based Μεθόδους.

Οι μέθοδοι μέσης τιμής (mean based methods) αντικαθιστούν τα ελλιπή στοιχεία αντικαθιστώνται από κάποια μέση τιμή που προκύπτει από τα παρατηρούμενα δεδομένα και μπορούν να διακριθούν τρεις υποκατηγορίες: Μέθοδοι Αντικατάστασης Μέσης Τιμής Υποσυνόλου, Μέθοδοι Αντικατάστασης Μέσης Τιμής Συνόλου και Μέθοδοι Αντικατάστασης Μέσης Τιμής Περίπτωσης. Οι μέθοδοι αυτής της κατηγορίας έχουν το πλεονέκτημα ότι χρησιμοποιούν όλα τα δεδομένα που προσφέρονται από το δείγμα. Ωστόσο παρουσιάζουν αποκλίσεις από τις πραγματικές τιμές, καθώς η μέση τιμή μπορεί να υπερεκτιμήσει ή να υποεκτιμήσει την μεταβλητότητα των τιμών. Για να δώσουν ικανοποιητικά αποτελέσματα τα ελλιπή στοιχεία πρέπει να λείπουν τελείως τυχαία (MCAR).

Οι μέθοδοι παλινδρόμησης (regression based methods) αντικαθιστά τα ελλιπή δεδομένα με τιμές που προέρχονται από μία ανάλυση παλινδρόμησης για κάθε μία μεταβλητή, η οποία στηρίζεται μόνο στις παρατηρούμενες τιμές της συγκεκριμένης μεταβλητής. Περιλαμβάνει δύο στάδια, αρχικά εκτιμάται η σχέση μεταξύ των μεταβλητών και στη συνέχεια χρησιμοποιούνται οι συντελεστές παλινδρόμησης για να εκτιμηθούν οι τιμές των ελλιπών στοιχείων. Για να χρησιμοποιηθεί πρέπει να υπάρχει γραμμική συσχέτιση μεταξύ των εκτιμώμενων και των ελλιπών στοιχείων. Δίνει πολύ καλά αποτελέσματα αν τα ελλιπή στοιχεία ακολουθούν τον μηχανισμό MCAR και ικανοποιητικά αν ακολουθούν τον μηχανισμό MAR.

Οι Hot – Deck βασισμένες μέθοδοι αντικαθιστούν τα ελλιπή στοιχεία με μία παρατηρούμενη τιμή μίας περίπτωσης που είναι παρόμοια με την περίπτωση όπου εμφανίζεται η ελλιπής τιμή, με τυχαία επιλογή. Το πλεονέκτημα αυτής της μεθόδου είναι ότι αντικαθιστά τα ελλιπή στοιχεία με πραγματικές τιμές που παρατηρούνται σε άλλες περιπτώσεις του δείγματος και παρουσιάζουν κάποια σχέση με την περίπτωση που

εμφανίζεται το ελλιπές στοιχείο. Η δυσκολία εφαρμογής της επίκειται στη δυσκολία καθορισμού των παρόμοιων περιπτώσεων, ενώ δεν υπάρχουν πολλά στοιχεία για την ακρίβεια τέτοιων μεθόδων.

Ο πολλαπλός καταλογισμός (Multiple Imputation) είναι μία διαδικασία που μπορεί να ταξινομηθεί τόσο στις μεθόδους αντικατάστασης, καθώς συμπληρώνονται οι τιμές των ελλιπών στοιχείων, όσο και στις μεθόδους που βασίζονται σε μοντέλα καθώς δημιουργείται από τον καθορισμό ενός μοντέλου για τα παρατηρούμενα δεδομένα.

Ο πολλαπλός καταλογισμός είναι μία διαδικασία που περιλαμβάνει τρία βήματα. Στο πρώτο βήμα γίνεται η αντικατάσταση των ελλιπών δεδομένων (Imputation step). Δημιουργούνται σύνολα εύλογων τιμών που χρησιμοποιούνται για την αντικατάσταση των ελλιπών παρατηρήσεων. Η διαφορά του με τις παραδοσιακές μεθόδους είναι ότι δημιουργούνται περισσότερα από ένα σύνολα αντικατάστασης των ελλιπών δεδομένων και κατ' επέκταση και περισσότερα από ένα σύνολα με ολοκληρωμένα στοιχεία, διατηρώντας την αβεβαιότητα της τιμής αντικατάστασης. Στη συνέχεια τα ολοκληρωμένα σύνολα αναλύονται με κοινές στατιστικές μεθόδους που (Analysis Step) και τέλος τα αποτελέσματα των συνόλων συνδυάζονται και δίνουν μία συνδυασμένη εκτίμηση (Combination /pooling Step).

Η μέθοδος αυτή επιτρέπει τη δημιουργία ολοκληρωμένων συνόλων δεδομένων τα οποία είναι εύκολο να αναλυθούν με κοινές στατιστικές μεθόδους. Τα σύνολα αυτά μπορούν να δημιουργηθούν μία φορά και να αναλυθούν από διαφορετικούς αναλυτές και με διαφορετικές μεθόδους. Επίσης, τα διαφορετικά σύνολα δεδομένων που αναλύονται και συνδυάζονται δίνουν μία αντιπροσωπευτική εικόνα για την ποικιλομορφία του δείγματος. Από την άλλη είναι χρονοβόρα και απαιτεί μεγάλη υπολογιστική ισχύ.

Στην τελευταία κατηγορία ανήκουν μέθοδοι που βασίζονται σε μοντέλα (Model – Based Methods). Σε αυτή την κατηγορία μπορούν να διακριθούν μέθοδοι που χρησιμοποιούν την εκτίμηση μέγιστης πιθανότητας για την εκτίμηση των παραμέτρων του μοντέλου και μέθοδοι που στηρίζονται στην Μπεϋζιανή θεωρία.

Στην εκτίμηση της μέγιστης πιθανότητας (Maximum Likelihood Estimation) οι παράμετροι του μοντέλου εκτιμώνται από τα παρατηρούμενα δεδομένα και οι ελλιπής τιμές εκτιμώνται με βάση το μοντέλο των παρατηρούμενων δεδομένων για τις εκτιμώμενες παραμέτρους. Η μέθοδος αυτή αυξάνει την ακρίβεια των εκτιμήσεων, αλλά είναι δύσκολο να προσδιοριστεί το μοντέλο των παρατηρούμενων τιμών.

Μία εναλλακτική μέθοδος μέγιστης πιθανότητας είναι ο αλγόριθμος EM (Expectation Maximization Algorithm) που αποτελεί μία επαναληπτική διαδικασία δύο βημάτων – βήμα της μεγιστοποίησης ή βήμα M (Maximization) και βήμα πρόβλεψης ή βήμα E (Expectation), τα οποία εναλλάσσονται κυκλικά – που συνεχίζεται μέχρι να υπάρξει σύγκλιση στις εκτιμώμενες τιμές. Ο αλγόριθμος EM δίνει μεγάλη ακρίβεια στα αποτελέσματα, ωστόσο δεν υπολογίζει τα συστηματικά σφάλματα και για επέλθει σύγκλιση μπορεί να χρειαστεί μεγάλη διάρκεια.

Για να χρησιμοποιηθούν οι μέθοδοι που χρησιμοποιούν την εκτίμηση της μέγιστης πιθανότητας πρέπει τα δείγματα να είναι σχετικά μεγάλα, καθώς για μικρά δείγματα τα αποτελέσματα δεν είναι ικανοποιητικά. Όταν τα ελλιπή στοιχεία ακολουθούν τον μηχανισμό MCAR ή MAR δίνει πολύ καλές εκτιμήσεις, ενώ όταν ο μηχανισμός είναι MNAR οι εκτιμήσεις δεν είναι τόσο καλές.

Οι μέθοδοι που στηρίζονται στην Μπεϋζιανή θεωρία και παρουσιάστηκαν σε αυτή την εργασία είναι ο αλγόριθμος Data Augmentation και η διαδικασία Gibbs' Sampler. Οι μέθοδοι αυτές χρησιμοποιούνται όταν τα δείγματα είναι

μικρά και δεν μπορεί να χρησιμοποιηθεί η εκτίμηση μέγιστης πιθανότητας. Ο αλγόριθμος Data Augmentation μια επαναληπτική διαδικασία δύο βημάτων. Με το πρώτο βήμα, βήμα καταλογισμού ή βήμα I (step I / Imputation step) τα ελλιπή στοιχεία αντικαθιστώνται από κάποιες τιμές που έχουν δημιουργηθεί μετά από μία σειρά πολλαπλών εξισώσεων παλινδρόμησης. Κατόπιν ακολουθεί το μεταγενέστερο βήμα (Posterior Step) ή βήμα P (Step P), όπου ένας νέος πίνακας συνδιακύμανσης και τα στοιχεία του διανύσματος της μέσης τιμής επιλέγονται με τυχαίο τρόπο από τη μεταγενέστερη κατανομή που έχει προκύψει από τις τιμές των ολοκληρωμένων στοιχείων που λήφθηκαν στο προηγούμενο βήμα, βήμα I. Η διαδικασία επαναλαμβάνεται μέχρι να επέλθει σύγκλιση.

Όταν τα προβλήματα είναι περίπλοκα ο αλγόριθμος Data Augmentation δεν μπορεί να χρησιμοποιηθεί και τότε χρησιμοποιείται η διαδικασία Gibbs' Sampler. Η μέθοδος Gibbs' Sampler είναι μία επαναληπτική μέθοδος παραγωγής τυχαίων μεταβλητών από μία (οριακή) κατανομή με άμεσο τρόπο, χωρίς να υπολογίζει τις πυκνότητες. Έτσι με τη χρήση αυτής της μεθόδου μπορούν να αποφευχθούν οι δύσκολοι υπολογισμοί, αντικαθιστώντας τους με διαδοχικούς απλούστερους υπολογισμούς.

Οι μέθοδοι αυτές έχουν το μειονέκτημα ότι είναι αργές, η σύγκλιση θέλει χρόνο για να επέλθει και απαιτούν μεγάλη υπολογιστική ικανότητα. Από την άλλη δίνουν στον αναλυτή ευελιξία και τα αποτελέσματα τους είναι ακριβή.

Πίνακας 1: Μέθοδοι διαχείρισης δεδομένων με ελλιπή στοιχεία

Μέθοδος	***Περιγραφή***	***Χρήση***	***Πλεονεκτήματα***	***Μειονεκτήματα***
Μέθοδοι Διαγραφής				
Listwise Deletion	Διαγραφή όλων των περιπτώσεων που παρουσιάζουν ελλιπή δεδομένα		• Εύκολη Εφαρμογή • Υπάρχει σε πολλά στατιστικά πακέτα	• Μειώνεται η στατιστική δύναμη του δείγματος, μεγάλος αριθμός πληροφοριών δεν χρησιμοποιείται
Pairwise Deletion	Επιμέρους συσχετισμοί των μεταβλητών, διαγραφή μόνο των περιπτώσεων με ελλιπή στοιχεία για εκείνο τον συσχετισμό	• Όταν ο μηχανισμός είναι MCAR και • Τα ελλιπή στοιχεία αποτελούν μικρό ποσοστό των δεδομένων	• Εύκολη Εφαρμογή • Πιο συντηρητική από Listwise Deletion (μεγαλύτερη στατιστική δύναμη	• Η συσχέτιση μεταξύ των μεταβλητών μπορεί να οδηγήσει σε αποκλίσεις • Δημιουργία προβλημάτων από τα διαφορετικά υποσύνολα ανάλυσης, ειδικά σε μεγάλα δείγματα
Διαδικασίες Στάθμισης				
	Εισαγωγή βαρών της παρατήρησης των δεδομένων, τα ελλιπή στοιχεία θεωρούνται μέρος του σχεδιασμένου προβλήματος και	• Όταν τα δεδομένα ακολουθούν μηχανισμό MAR • Σε συνδυασμό με	• Μείωση των αποκλίσεων	• Δεν χρησιμοποιούν όλα τα δεδομένα του δείγματος

	λαμβάνουν βάρος παρατήρησης	μεθόδους διαγραφής		
Μέθοδοι Αντικατάστασης				
Μέθοδοι μέσης τιμής	Τα ελλιπή στοιχεία αντικαθίστανται από την μέση τιμή των παρατηρούμενων στοιχείων	Όταν η συσχέτιση μεταξύ των μεταβλητών είναι μικρή, μικρότερη από 10% των ελλιπών δεδομένων	• Εύκολη στη χρήση • Χρησιμοποιεί όλα τα δεδομένα του δείγματος	• Μειώνει την εκτίμηση διακύμανσης • Επιδρά αρνητικά στους βαθμούς ελευθερίας
Αντικατάσταση Μέσης Τιμής Συνόλου	Αντικατάσταση με την μέση τιμή όλων των παρατηρούμενων τιμών	Όταν υπάρχει μικρή συσχέτιση μεταξύ των ελλιπών μεταβλητών και των άλλων στοιχείων του συνόλου	• Εύκολη στη χρήση • Χρησιμοποιεί όλα τα δεδομένα του δείγματος	• Υποεκτίμηση της απόκλισης της διακύμανσης
Αντικατάσταση Μέσης Τιμής Υποσυνόλου	Αντικατάσταση με τη μέση τιμή των παρατηρούμενων τιμών μίας ομάδας περιπτώσεων	Όταν είναι εύκολο να χωρισθούν τα δεδομένα σε υποομάδες	• Δίνει καλύτερες • εκτιμήσεις από την μέθοδο μέση τιμής συνόλου	• Υποεκτίμηση της • απόκλισης της διακύμανσης • Δυσκολία διαχωρισμού υποομάδων
Αντικατάστασης Μέσης Τιμής Περίπτωσης	Αντικατάσταση με τη μέση τιμή των παρατηρούμενων στοιχείων μίας συγκεκριμένης περίπτωσης	Για τη δημιουργία κλίμακας τιμών	• Διατήρηση του δείγματος	• Υποθέτει ίσες μέσες τιμές και τυπικές αποκλίσεις μεταξύ των ελλιπών στοιχείων και των προβλέψεων
Μέθοδοι Παλινδρόμησης (Regression	Εκτίμηση της σχέσης μεταξύ των μεταβλητών, χρήση των συντελεστών παλινδρόμησης για	Όταν λείπουν πολλά δεδομένα (>20%) και οι μεταβλητές είναι	• Δίνει καλά αποτελέσματα για μηχανισμό MCAR και	• Αλλάζει τον αριθμό των βαθμών ελευθερίας

Imputation)	την εκτίμηση των ελλιπών στοιχείων	ισχυρά συσχετισμένες	ικανοποιητικά για MAR	
Hot – Deck Imputation	Αντικατάσταση των ελλιπών στοιχείων με πραγματικές τιμές άλλων παρόμοιων περιπτώσεων	Όταν η μορφή έλλειψης ακολουθεί ορισμένη μορφή	• Τα ελλιπή στοιχεία αντικαθιστώνται από πραγματικά δεδομένα • Δεν αλλάζει η κατανομή λόγω χρήσης μέσων τιμών	• Δύσκολο να προσδιοριστούν οι παρόμοιες ομάδες • Δεν μπορεί να εφαρμοστεί αν δεν υπάρχουν παρόμοιες περιπτώσεις • Δεν υπάρχουν πολλά στοιχεία για την ακρίβεια της
Μέθοδοι Βασισμένοι σε Μοντέλα				
Πολλαπλός Καταλογισμός (Multiple Imputation)	Είναι μία διαδικασία δημιουργίας m εύλογων συνόλων για την αντικατάσταση των ελλιπών στοιχείων, παράγοντας m σύνολα ολοκληρωμένων στοιχείων τα οποία αναλύονται και συνδυάζονται	• Όταν ο μηχανισμός έλλειψης είναι MAR • Τα δείγματα είναι μεγάλα • Τα στοιχεία επιθυμείται να αναλυθούν πολλές φορές και από πολλούς διαφορετικούς ερευνητές	• Δημιουργία m>1 συνόλων με ολοκληρωμένα δεδομένα • Χρήση απλών μεθόδων για ανάλυση • Διατηρείται η ποικιλομορφία του δείγματος • Τα ολοκληρωμένα χρησιμοποιούνται από διαφορετικούς ερευνητές	• Μεγάλη υπολογιστική ισχύ • Χρονοβόρα

Maximum Likelihood Estimation				
Maximum Likelihood	Εκτιμώνται οι παράμετροι του μοντέλου από τα παρατηρούμενα στοιχεία και τα ελλιπή δεδομένα εκτιμώνται από τις εκτιμώμενες παραμέτρους	Όταν το δείγμα είναι σχετικά μεγάλο και ικανοποιούνται η προϋποθέσεις της κατανομής	• Αυξάνει την ακρίβεια τον αποτελεσμάτων • Δίνει ευελιξία στον αναλυτή • Καλά αποτελέσματα για μηχανισμούς MCAR και MAR • Σχετικά ικανοποιητικά για μηχανισμό MNAR	• Δύσκολο να προσδιοριστεί το μοντέλο • Αυστηρές προϋποθέσεις
Ο αλγόριθμος EM	Εναλλακτική επαναληπτική μέθοδος που επαναλαμβάνεται μέχρι τη σύγκλιση των εκτιμήσεων			• Χρονοβόρα διαδικασία σύγκλισης
Bayesian Methods				
Data Augmentation	Επαναληπτική μέθοδος όπου τα ελλιπή στοιχεία αντικαθιστώνται από εκτιμήσεις που προέρχονται από μία σειρά πολλαπλών εξισώσεων και μετά από την νέα κατανομή επιλέγονται τυχαία νέα μ.τ. και πίνακας συνδιακύμανσης και η διαδικασία συνεχίζει μέχρι να συγκλίνουν	Σε περιπτώσεις που το δείγμα είναι μικρό και δεν μπορεί να χρησιμοποιηθεί η μέγιστη πιθανότητα	• Ακριβή αποτελέσματα • Αντιμετώπιση προβλήματος σε μικρά δείγματα • Καλά αποτελέσματα για μηχανισμούς MCAR και MAR • Σχετικά ικανοποιητικά για μηχανισμό MNAR	• Χρονοβόρα διαδικασία σύγκλισης • Μεγάλη υπολογιστική ισχύ
Gibbs' Sampler	Μία επαναληπτική μέθοδος παραγωγής τυχαίων μεταβλητών από μία (οριακή) κατανομή με άμεσο τρόπο, χωρίς να υπολογίζει τις πυκνότητες	Σε πιο πολύπλοκες περιπτώσεις με μικρό δείγμα δεδομένων		• Χρονοβόρα διαδικασία σύγκλισης • Μεγάλη υπολογιστική ισχύ

Επίλογος – Συμπεράσματα

Οι περισσότερες έρευνες που περιλαμβάνουν συλλογή δεδομένων, δυστυχώς, χαρακτηρίζονται από το πρόβλημα των ελλιπών δεδομένων. Τα ημιτελή αυτά δεδομένα μπορούν να δημιουργούν προβλήματα στη στατιστική επεξεργασία και ανάλυση και οδηγούν σε λανθασμένα συμπεράσματα. Για την αντιμετώπιση του προβλήματος αυτού έχουν επινοηθεί κάποιες μέθοδοι, οι οποίες διαχειρίζονται με συγκεκριμένο τρόπο τα ελλιπή στοιχεία, ώστε να μην επηρεάζουν την ποιότητα των αποτελεσμάτων της έρευνας.

Η παρούσα διπλωματική εργασία ασχολήθηκε το πρόβλημα των ελλιπών στοιχείων (Missing Data Problem) και τις μεθόδους που χρησιμοποιούνται για την αντιμετώπισής του. Συγκεκριμένα παρουσιάστηκαν οι μέθοδοι που χρησιμοποιούνται για την αντιμετώπιση σε τέτοιες περιπτώσεις, αναφέροντας τα βασικά σημεία κάθε μεθόδου, τον τρόπο αντιμετώπισης των ελλιπών στοιχείων, τα πλεονεκτήματα και τα μειονεκτήματα της.

Στη συνέχεια πραγματοποιήθηκε σύγκριση μεταξύ αυτών των μεθόδων με έμφαση στις προϋποθέσεις κάτω από τις οποίες κάθε μία μέθοδος επιλύει επιτυχώς το πρόβλημα των ελλιπών δεδομένων και θεωρείται επιθυμητή η χρησιμοποίησή της.

Το πρώτο ερώτημα που πρέπει να απαντηθεί για την επιλογή μίας μεθόδου διαχείρισης των ελλιπών δεδομένων είναι ποιο μηχανισμό (MCAR, MAR ή MNAR) ακολουθούν τα ελλιπή στοιχεία, καθώς κάθε μέθοδος εμπεριέχει κάποιους περιορισμούς για τον μηχανισμό έλλειψης.

Όταν τα δεδομένα λείπουν τελείως τυχαία (MCAR) οι μέθοδοι όλων των κατηγοριών μπορούν να δώσουν ικανοποιητικά αποτελέσματα. Όταν τα δεδομένα λείπουν τυχαία (MAR) οι μέθοδοι που βασίζονται σε μοντέλα δίνουν σωστά αποτελέσματα, ενώ οι διαδικασίες στάθμισης και οι μέθοδοι αντικατάστασης μπορούν να δώσουν ικανοποιητικά αποτελέσματα. Τέλος, όταν τα δεδομένα δεν λείπουν τυχαία (MNAR) δεν υπάρχει συγκεκριμένη μέθοδος που να μπορεί να εξαλείψει απόλυτα τις αποκλίσεις, ωστόσο οι μέθοδοι που βασίζονται σε μοντέλα δίνουν αποτελέσματα με μικρές αποκλίσεις.

Οι μέθοδοι διαγραφής θεωρούνται οι πιο εύκολα χρησιμοποιούμενες μέθοδοι στην αντιμετώπιση του προβλήματος των ελλιπών δεδομένων και υπάρχουν σε όλα τα στατιστικά πακέτα. Όμως παρουσιάζουν μεγάλες αποκλίσεις από τις πραγματικές τιμές και μειώνουν τη στατιστική δύναμη απορρίπτοντας μεγάλο αριθμό δεδομένων. Έτσι θα πρέπει χρησιμοποιούνται μόνο όταν τα ελλιπή δεδομένα αποτελούν μικρό ποσοστό των συνολικών στοιχείων και ο μηχανισμός έλλειψης είναι MCAR.

Οι διαδικασίες στάθμισης χρησιμοποιούνται σε συνδυασμό με τις μεθόδους διαγραφής βελτιώνοντας τις εκτιμήσεις (μικρότερες αποκλίσεις), ενώ μπορούν να χρησιμοποιηθούν και όταν τα δεδομένα λείπουν τυχαία (MAR). Όμως η μικρή στατιστική δύναμη των αποτελεσμάτων συνεχίζει να αποτελεί πρόβλημα.

Οι παραδοσιακές διαδικασίες αντικατάστασης αυξάνουν τη στατιστική δύναμη των αποτελεσμάτων καθώς χρησιμοποιούν όλα τα διαθέσιμα δεδομένα και επιτρέπουν την ανάλυση των δεδομένων με τις κλασσικές στατιστικές μεθόδους που χρησιμοποιούνται και για τα πλήρη στοιχεία διευκολύνοντας τη διαδικασία. Παρόλα αυτά υποεκτιμάται ή υπερεκτιμάται η απόκλιση διακύμανσης.

Ο πολλαπλός καταλογισμός, μία μέθοδος που εντάσσεται τόσο στις μεθόδου αντικατάστασης, όσο και στις μεθόδους που βασίζονται σε μοντέλα, διατηρεί όλα τα θετικά των μεθόδων αντικατάστασης, ενώ επιπλέον διατηρεί και την ποικιλομορφία του δείγματος (δημιουργεί m σύνολα ολοκληρωμένων δεδομένων, αφήνοντας την αβεβαιότητα της τιμής αντικατάστασης).

Οι μέθοδοι που βασίζονται σε μοντέλα, μέθοδοι μέγιστης πιθανότητας και Μπεϋζιανές μέθοδοι, δίνουν μεγαλύτερη ευελιξία στον αναλυτή. Τα αποτελέσματα τους έχουν μεγαλύτερη ακρίβεια από τις μεθόδους των άλλων κατηγοριών και μπορούν να χρησιμοποιηθούν ακόμα και όταν τα δεδομένα δεν λείπουν τυχαία με ικανοποιητικά αποτελέσματα.

Τέλος, θα ήταν συνετό να τεθούν κάποια θέματα που δεν καλύπτει η συγκεκριμένη διπλωματική και θα μπορούσαν να αποτελέσουν αντικείμενο μελέτης και αυτά είναι:

- Η ανάλυση νέων μεθόδων που στηρίζονται σε μοντέλα (Model – Based Methods) που αναπτύσσονται τα τελευταία χρόνια με σκοπό να βελτιώσουν τις εκτιμήσεις και αποτελούν παραλλαγές των μεθόδων που παρουσιάστηκαν.

- Η δημιουργία παραδειγμάτων με ελλιπή δεδομένα και η χρησιμοποίηση των παραπάνω μεθόδων με σκοπό να δειχθεί στη πράξη πότε κάθε μέθοδος έχει τα επιθυμητά αποτελέσματα.

Βιβλιογραφία

Acock A. C., (2005), Working with Missing Values, Journal of Marriage and Family, volume 67, 1012 – 1028.

Alison P. D. (2002) Missing Data, Series: Quantitative Applications in social Sciences, California, Sage Publications Inc.

Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. Sociological Methods & Research, 28, 301-309.

Arbuckle, J. L., & Wothke, W. (1999), AMOS 4.0 user's guide [Computer software manual], Chicago: Smallwaters.

Barnard, J., Rubin, D.B. (1999), Small-Sample Degrees of Freedom with Multiple Imputation, Biometrika, 86, 948–955.

Brick J. M., Kalton G. (1996), Handling missing data in survey research, Statistical Methods in Medical Research, volume 5, 215 – 238.

Buhi E. R., Goodson P., Neilands T. B., (2008), Out of Sight, Not Out of Mind: Strategies for Handling Missing Data, Am J Health Behav., volume 32, no 1, 83 – 92.

Burton PR, Gurrin LC, Campbell MJ (1998) Clinical significance not statistical significance: a simple Bayesian alternative to p values, J Epidemiol Community Health, volume 52, 318 – 323.

Casella George, Edward I. George (1992), Explaining the Gibbs Sampler, The American Statistician, volume 46, no 3, 167 – 174.

Cheung M. W. (2007), Comparison of Methods of Handling Missing Time – Invariant Covariates in Latent Growth Models Under the Assumption of Missing Completely at Random, Organizational Research Methods, volume 10, no 4, 609 – 634.

Collins L. M., Schafer J. L., Kam Chi Ming, (2001), A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures, American Psychological Association Inc, volume 6, no 4, 330 – 351.

Dempster, A. P., Laird, N., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), Journal of the Royal Statistical Society, Ser. B, 39, 1 – 38.

Donders A. R. T. et al. (2006), Journal of Clinical Epidemiology, volume 59, 1087 – 1091.

Enders C. K., (2006), A Primer on the Use of Modern Missing Data Methods in Phychosomatic Medicine Research, Phychosomatic Medicine, volume 68, 427 – 436.

Fichman, M., & Cummings, J. N. (2003), Multiple imputation for missing data: Making the most of what you know. Organizational Research Methods, 6, 282-308.

Graham, J. W., & Hofer, S. M. (1991), EMCOV.EXE users' guide [Computer software manual]. Unpublished manuscript, University of Southern California, Los Angeles.

Harmon R., Challenor P. (1997), A Markov Chain Monte Carlo methods for Estimation and assimilation into models, Ecological modeling, volume 101, 41 – 59.

Horton N. J., Kleinman K. P., (2007), Much Ado About Nothing: A Comparison Missing Data Methods and Software to Fit Incomplete Data Regression Models, American Statistical Association, The American Statistician, Volume61, no 1, 79 – 90.

Horton N. J., Lipsitz S. R., (2001), Multiple Imputation in Practice: Comparison of software packages of Regression Models with Missing Variables, The American Statistician, Volume 55, no 3, 244 – 254.

Husiman M., Imputation of missing network data: Some simple procedures, Dept. of Psychology, University of Groningen.

Ibrahim J. G., Chen M., Lipsitz S. R., Herring A. H. (2005), Missing Data Methods for Generalized Linear Models, A Comparison Review, Journal of American Statistical Association, volume 100, no 469, 332 – 345.

Jiang Yannan, (2004), Semi-parametric Maximum Likelihood for Multiphase Response – Selective Sampling and Missing Data Problems, PHD Thesis, Department of statistics, University of Auckland, 1 – 17.

Jöreskog, K. G., & Sörbom, D. (2001), LISREL (Version 8.5), [Computer software]. Chicago: Scientific Software International.

Kim, J. & Curry J, (1977), The treatment of missing data in multivariate analysis, Sociological Methods & Research, volume 6, 215 – 240.

King G., Joseph A., Kenneth S., (2001), Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation, American Political Science Review, volume 95, no 1, 49 – 68.

Kong A., Liu J. S., Wongs W. H., (1994), Sequential imputations and Bayesian Missing Data Problems, Journal of American Statistical Association, volume 89, no 425 Theory and Methods, 278 – 288.

Little R. J. & Rubin D. B. (2001), Statistical Data Missing, International Encyclopedia of Social and Behavioral Sciences, Elsevier Science Ltd., volume 1, 5019 – 5025.

Liu M., Wei L., Zhang J. (2006), Review of guidelines and literature for handling missing data in longitudinal clinical trials with a case, Pharmaceutical statistics, volume 5, 5 – 18.

Madow, W.G., Nisselson, H., Olkin, I. and Rubin, D.B. (1983), Incomplete Data in Sample Surveys. 1, 2, and 3, New York: Academic Press.

McKnight A. J., McKnight K.M., Sidali S., Figueredo A. J., (2007) Missing Data, A gentle Introduction, New York, John Wiley & sons Ltd, 1 – 17.

Molenberghs G. L., Kezward M., (2007), Missing Data in Clinical Studies, England, John Wiley & sons Ltd, 3 - 11

Moore J. C. (1988), Self/proxy response status and survey response quality, Journal of the official statistics, volume 4, 155 – 172.

Muthén, L. K., & Muthén, B. O. (1998), Mplus user's guide [Computer software manual]. Los Angeles: Muthén & Muthén.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (1999), Mx: Statistical modeling (5th ed.) [Computer software], Richmond: Virginia Commonwealth University, Department of Psychiatry.

Newman D. A. (2003), Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. Organizational Research Methods, 6, 328-362.

Olinsky A., Chen S., Harlow L., (2003), The comparative efficacy of imputation methods for missing data in structural equation modeling, European Journal of Operation Research, volume 151, 53 – 79.

Pearson R. K., The problem of Disguised Missing Data, SIGKDD Explorations, volume 8, Issue I, 83 – 92.

Peugh J. L., Enders Craig K., (2004), Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement, Review of Educational Research, volume 74, 525 – 556.

Pigott T. D., (2001), A Review of Methods for Missing Data, Educational Research and Evaluation, volume 7, no 4, 353 – 383.

Raghuanathan T. E., Lepkoswski J. M., Van Hoewyk J., Solenberger P. (2001), A Multivariate Technique for multiply Imputing Missing Values Using Sequence Regression Model, Survey Methology, volume 27, no 1, 85 – 95.

Rubin D. B & Little R. J.(2002), Statistical Analysis with missing data, second edition, New York, John Wiley & sons Ltd.

Rubin D. B & Little R. J.(1987), Statistical Analysis with missing data, New York, John Wiley & sons Ltd.

Schafer J. L. & Graham J. W., (2002), Missing Data: Our View of the State of the Art, American Psychological Association Inc, volume 7, 147 – 177.

Schafer J. L., & Olsen M. K. (1998), Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research, 33, 545-571.

Schafer J. L. (1997) Analysis of Incomplete Multivariate Data, New York, Chapman and Hall, chapter 3, 37 – 88.

Scheneider T. (2001), Analysis of Incomplete Climate Data: Estimation of Mean Value and covariance Matrices and Imputation Missing Values, American Meteorological Society, volume 14, 853 – 871.

Sinharay S., Stern H. S., Russell D., (2001), The Use o Multiple Imputation for the Analysis of Missing Data, American Psychological Association Inc, volume 6, no 4, 317 – 329.

Tsikriktsis Nikos, (2005), A review of techniques for treating missing data in OM survey research, Journal of Operations Management, volume 24, 53 – 62.

Verbeke Geert, Molenberghs Geert (2000), Linear Mixed Models for Longitudinal Data, New York, Springer, 221 – 229.

Witta L. E., (2001), Does Method of Handling Missing Data Affect Results of Statistical Equation Model?, paper presented at the Annual Meeting of Southwest, Education Research Association.

Walczak B., Massart D. L., (2001), Dealing with missing data: Part II, Chemometrics and Intelligent Laboratory Systems, volume 58, 29 – 42.

Yuan, K. H., & Bentler, P. M. (2000), Three likelihood-based methods for mean and covariance structure analysis with non normal missing data. Sociological Methodology, 30, 165–200.

Yuan, Y. C. (2000). Multiple imputation for missing data: Concepts and new development, In Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference (Paper No. 267). Cary, NC: SAS Institute

Zeger S. L., Liang K. Y., & Albert P. S. (1988), Models for longitudinal data: A generalized estimating equation approach. Biometrics, 44, 1049–1060.