

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Βασικές τεχνικές μέτρησης πιστωτικού κινδύνου
Λογιστική παλινδρόμηση - διαχωριστική ανάλυση
Θεωρία – εφαρμογή

Γεώργιος Ιωάννη Τούντας

ΕΡΓΑΣΙΑ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος
Συμπληρωματικής Ειδίκευσης στη Στατιστική
Μερικής Παρακολούθησης (Part-time)

Αθήνα
Σεπτέμβριος 2011

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον καθηγητή μου κύριο Αθανάσιο Ν Γιαννακόπουλο για το ενδιαφέρον που έδειξε κατά τη διάρκεια της διατριβής.

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

Γεννήθηκα στην Αθήνα το 1982. Αποφοίτησα από το 14ο λύκειο Αθηνών το 2000. Το 2006 αποφοίτησα από το τμήμα Μαθηματικών του Εθνικού Καποδιστριακού Πανεπιστημίου Αθηνών με Κατεύθυνση στα Εφαρμοσμένα Μαθηματικά και Ειδίκευση στη Διδακτική. Το 2009 ολοκλήρωσα τον κύκλο σπουδών του Μεταπτυχιακού Προγράμματος Ειδίκευσης στη Στατιστική στο Οικονομικό Πανεπιστήμιο Αθηνών με Κατεύθυνση «Εφαρμοσμένη Στατιστική για Εκπαιδευτικούς και Στελέχη Επιχειρήσεων & Οργανισμών».

ABSTRACT

Georgios Tountas

Basic techniques for measuring credit risk
Logistic regression - discriminant analysis
Theory – Application

September 2011

This Thesis is consisted of two parts. The first part introduce the reader to the concept of financial risk and distinguish the types of risk. Particular reference is being made to the most important risk, credit risk, while analyzing the measuring techniques. The second part compares the two most important parametric techniques of measuring credit risk: the logistic regression and discriminant analysis. Chapters 3 and 4 analyze the mathematical background of both models while in 5 chapter the two models are used as methods of separating populations, based on real data using the statistical package spss. Finally it records the similarities and the differences between the two methods that measure credit risk.

ΠΕΡΙΛΗΨΗ

Γεώργιος Τούντας

Βασικές τεχνικές μέτρησης πιστωτικού κινδύνου
Λογιστική παλινδρόμηση- διαχωριστική ανάλυση
Θεωρία – εφαρμογή

Σεπτέμβριος 2011

Η παρούσα εργασία αποτελείται από δυο τμήματα. Στο πρώτο μέρος εισάγει τον αναγνώστη στην έννοια των χρηματοοικονομικών κινδύνων και διακρίνει τα είδη των κίνδυνων. Ιδιαίτερη μνεία γίνεται στον σημαντικότερο κίνδυνο, τον πιστωτικό ενώ παράλληλα αναλύονται οι τεχνικές μέτρησης του. Στο δεύτερο μέρος συγκρίνονται οι δυο σημαντικότερες παραμετρικές τεχνικές μέτρησης πιστωτικού κινδύνου: η λογιστική παλινδρόμηση και η διαχωριστική ανάλυση. Τα κεφάλαια 3 και 4 αναλύουν το μαθηματικό υπόβαθρο των δυο μοντέλων ενώ στο $5^{ο}$ κεφάλαιο τα δυο μοντέλα χρησιμοποιούνται ως μέθοδοι διαχωρισμού πληθυσμών σε πραγματικά δεδομένα με τη χρήση του στατιστικού πακέτου spss. Τέλος καταγράφονται οι ομοιότητες και οι διαφορές των δυο μεθόδων μέτρησης πιστωτικού κίνδυνου.

ΠΕΡΙΕΧΟΜΕΝΑ

1. Εισαγωγή

Εισαγωγή ……1

2. Οι κίνδυνοι στο χρηματοπιστωτικό σύστημα

2.1. Κίνδυνος……...3

2.2. Μέτρηση κινδύνου……..4

2.3. Διάκριση κινδύνων……..5

2.4. Πιστωτικός κίνδυνος……...7

2.5. Credit scoring – scorecards – τεχνικές μετρήσεις πιστωτικού κινδύνου……9

3. Διαχωριστική ανάλυση

3.1. Εισαγωγή.....15

3.2. Η διαχωριστική συνάρτηση του Fisher……..…….....17

3.3. Γενίκευση της διαχωριστικής ανάλυσης σε g πληθυσμούς.....20

4. Λογιστική παλινδρόμηση

4.1. Εισαγωγή.....27

4.2. Το μοντέλο…..…..….....……...……..…….....28

4.3. Ειδικά προβλήματα όταν η αποκρινόμενη μεταβλητή είναι δυαδική.....29

4.4. Απλή λογιστική παλινδρόμηση.....30

4.4.1. Απλό λογιστικό μοντέλο παλινδρόμησης.....30

4.4.2. Συνάρτηση πιθανοφάνειας.....30

4.4.3. Η ερμηνεία του $\mathbf{b_1}$.....32

4.4.4. Επαναλαμβανόμενες παρατηρήσεις.....32

4.5. Πολλαπλή λογιστική παλινδρόμηση.....33

4.5.1. Πολλαπλό λογιστικό μοντέλο παλινδρόμησης.....34

4.5.2. Προσαρμογή του μοντέλου.....35

4.5.3. Η ερμηνεία των συντελεστών.....35

5. Διαχωρισμός ομάδων με πραγματικά δεδομένα

5.1. Εισαγωγή.....37

5.2. Η παρουσίαση του προβλήματος.....38

5.3. Ανάλυση-παρουσίαση δεδομένων.....38

5.4. Επίλυση του προβλήματος με τη χρήση της διαχωριστικής ανάλυσης.....40

5.5. Επίλυση του προβλήματος με τη χρήση της λογιστικής παλινδρόμησης.....61

5.6. Ομοιότητες - Διαφορές διαχωριστικής ανάλυσης- λογιστικής παλινδρόμησης.....71

Βιβλιογραφία.....73

Κεφάλαιο 1

Εισαγωγή

Στη σημερινή εποχή, ο ρόλος των τραπεζών και των χρηματοπιστωτικών ιδρυμάτων είναι ιδιαίτερα ανεπτυγμένος . Αυτό οφείλεται αφ' ενός στην αύξηση της κατανάλωσης από πλευράς των νοικοκυριών και αφ' ετέρου στις νέες επενδυτικές ευκαιρίες που διαρκώς προκύπτουν για τις επιχειρήσεις. Επιπλέον, έχει αυξηθεί σημαντικά το διεθνές εμπόριο γεγονός που δημιουργεί νέες εργασίες και τομείς δραστηριοποίησης του τραπεζικού κλάδου. Όταν ξεκίνησε η λειτουργία των τραπεζών είχαν περιορισμένο αντικείμενο. Αρμοδιότητα τους ήταν η φύλαξη των χρημάτων των καταθετών τους και τα κύρια κέρδη τους προέρχονταν από ποσά που εισέπρατταν από τους πελάτες τους για την εργασία αυτή. Με την πάροδο των χρόνων το φάσμα των δραστηριοτήτων τους διευρύνθηκε με αποτέλεσμα σήμερα να καλύπτουν σχεδόν κάθε πιθανή ανάγκη των οικονομικών μονάδων. Στα πλαίσια αυτά, οι τράπεζες, πέραν του παραδοσιακού τους ρόλου ως φύλακες των αποταμιεύσεων ιδιωτών και επιχειρήσεων, διενεργούν διαμεσολαβητικές εργασίες για τις οικονομικές μονάδες – πελάτες τους, διευκολύνουν το εξωτερικό εμπόριο με την προσφορά συναλλάγματος, χορηγούν δάνεια καθώς και πλήθος άλλων προϊόντων, όπως πιστωτικές κάρτες, leasing, ασφάλειες κ.α. Αυτή η αλλαγή στη λειτουργία των τραπεζών σηματοδότησε νέες εισροές γι' αυτές, αλλά και κινδύνους οι οποίοι απορρέουν από το νέο τους ρόλο. Επιπλέον, το περιβάλλον στο οποίο λειτουργούν σήμερα χιλιάδες επιχειρήσεις είναι κάθε άλλο παρά στατικό. Η απελευθέρωση και διεθνοποίηση των χρηματοοικονομικών αγορών καθώς και η ανάπτυξη της τεχνολογίας είχε ως συνέπεια την ραγδαία αύξηση όχι μόνο του ανταγωνισμού μεταξύ των πιστωτικών ιδρυμάτων αλλά και των αναλαμβανόμενων από αυτά κινδύνων. Οικονομικές συγκυρίες, μεταβολές

στο πολιτικό σκηνικό πολλών χωρών, κλιματικές αλλαγές είναι μερικά από τα στοιχεία που επηρεάζουν καθημερινά τις λειτουργίες τους και καθιστούν το μέλλον αβέβαιο.

Συνεπώς, προκύπτει πλέον μία επιτακτική ανάγκη από πλευράς των τραπεζών να διασφαλίσουν τα συμφέροντά τους έναντι των νέων αυξανόμενων κινδύνων που καλούνται να αναλάβουν.

Στα κεφάλαια που ακλουθούν αναπτύξουμε τα αναλυτικά τα είδη κινδύνων με έμφαση στον πιστωτικό κίνδυνο, τον σημαντικότερο κίνδυνο που μπορεί να αντιμετωπίσει ένας τραπεζικός οργανισμός. Θα γίνει αναφορά στις τεχνικές μέτρησης του πιστωτικού κινδύνου και θα παρουσιαστούν σε θεωρητικό επίπεδο οι 2 σημαντικότερες τεχνικές. Στα τελευταία δυο κεφάλαια θα εφαρμόσουμε σε πραγματικά δεδομένα τις δυο προαναφερθείσες τεχνικές μέτρησης πιστωτικού κινδύνου με την χρήση του στατιστικού πακέτου spss.

Κεφάλαιο 2

Οι Κίνδυνοι στο χρηματοπιστωτικό σύστημα

2.1 Κίνδυνος

Κίνδυνος (risk) στην οικονομική ορολογία ορίζεται η μεταβλητότητα των δυνητικών αποτελεσμάτων μιας επένδυσης γύρω από την αναμενόμενη τιμή ή τον αριθμητικό τους μέσο.

Σε έρευνα που έγινε για λογαριασμό του Ινστιτούτου _διαχείρισης του Κινδύνου (IRM), ο κίνδυνος ορίζεται ως ο συνδυασμός της πιθανότητας ενός γεγονότος και των συνεπειών του. Οι συνέπειες, ωστόσο, μπορεί να είναι και θετικές και αρνητικές. Όταν ο κίνδυνος, συνδέεται με ευνοϊκά αποτελέσματα, χαρακτηρίζεται ως upside risk. Στην αντίθετη περίπτωση, ο κίνδυνος χαρακτηρίζεται downside. (Σε μεγάλο μέρος τη διεθνούς βιβλιογραφίας, μία επιχείρηση θεωρείται ότι αναλαμβάνει ρίσκο μόνο όταν έχει να αντιμετωπίσει πιθανές αρνητικές συνέπειες). Επομένως, γίνεται αντιληπτό ότι η έννοια του κινδύνου συνδέεται στενά με τις έννοιες της μεταβλητότητας (variability) και της αστάθειας (volatility).

2.2 Μέτρηση Κινδύνου

Ο κίνδυνος μετράται με την τυπική απόκλιση σ. Εάν ένα αξιόγραφο παρουσιάζει μεγάλες αποκλίσεις στην αξία του, τότε μεγαλύτερες είναι και οι διακυμάνσεις γύρω από τη μέση τιμή του, συνεπώς μεγαλύτερος είναι και ο κίνδυνος που εμπεριέχει και αντίστροφα. Ο τύπος της τυπικής απόκλισης δίνεται από τον τύπο:

$$\sigma = \left[\frac{\sum_{i=1}^{n}(R_{it} - E(R_{it})^2}{n-1}\right]^{1/2}$$

Για την εκτίμηση της διασποράς των δυνητικών αποτελεσμάτων μιας επένδυσης γύρω από την αναμενόμενη τιμή τους χρησιμοποιούμε κατανομή πιθανοτήτων. Για την κατανομή χρησιμοποιούμε την τυπική απόκλιση σ και για μια σχετική εκτίμηση του κινδύνου χρησιμοποιούμε τον παρακάτω τύπο, λαμβάνοντας υπόψη τις πιθανότητες P να συμβούν τα δυνητικά αποτελέσματα της επένδυσης:

$$\sigma = \left[\sum_{i=1}^{n} P_i * (R_{it} - E(R_{it})^2\right]^{1/2}$$

Ένα μέσο μέτρησης της διασποράς των δυνητικών αποδόσεων, στην περίπτωση που συγκρίνουμε τον κίνδυνο εναλλακτικών επενδύσεων με μεγάλες αποκλίσεις στις τιμές των αξιόγραφων, είναι ο συντελεστής μεταβλητότητας CV. Ο συντελεστής αυτός μετρά τον κίνδυνο σ ανά μονάδα αναμενόμενης απόδοσης E(R) και καθορίζεται από τον τύπο:

$$CV = \frac{\sigma}{E(R_{it})}$$

2.3 Διάκριση κινδύνων

Ο κίνδυνος που αντιμετωπίζει ένας πιστωτικός οργανισμός να διαφέρει το πραγματικό αποτέλεσμα μιας επένδυσης από το αναμενόμενο μπορεί να πηγάζει από επιμέρους κινδύνους. Παρόλο που τέτοια κατηγοριοποίηση είναι εξυπηρετική, είναι μόνο τυπική. Η χρήση και οι έννοιες διαφέρουν. Τα όρια μεταξύ των κατηγοριών είναι ασαφή. Στη διεθνή βιβλιογραφία δεν υπάρχει μια αυστηρή διάκριση μεταξύ των επιμέρους αυτών κινδύνων. Ενδεικτικά τα είδη των κινδύνων μπορούν να κατηγοριοποιηθούν ως εξής:

i) Κίνδυνος αγοράς (market risk):
Είναι ο κίνδυνος να παρουσιαστούν μη αναμενόμενες-απρόβλεπτες αλλαγές σε τιμές προϊόντων η επιτοκίων.
υποκατηγορίες στις οποίες μπορούμε να διακρίνουμε το market risk είναι:
α) Χρηματιστηριακός κίνδυνος ή κίνδυνος μεταβολής της μετοχής.
β) Χρηματοοικονομικός κίνδυνος (financial risk): ο κίνδυνος που προέρχεται από τη χρήση δανειακών κεφαλαίων από μια εταιρία.
γ) πληθωριστικός κίνδυνος (inflation risk): Ο πληθωρισμός μειώνει την αγοραστική δύναμη της επένδυσης και θα πρέπει το μέγεθος να αποπληθωριστεί για να βρούμε την πραγματική αξία.
δ) Επιτοκιακός κίνδυνος. ο κίνδυνος να αλλάξει η αξία μιας επένδυσης (όπως κατάθεση, δάνειο, ομόλογο κλπ.) εξαιτίας μεταβολών στο επίπεδο των επιτοκίων. Ο επιτοκιακός κίνδυνος επηρεάζει συνήθως τις τιμές των ομολόγων περισσότερο από των μετοχών.
ε) Συναλλαγματικός κίνδυνος (exchange rate risk): ο κίνδυνος απώλειας των αποδόσεων των κεφαλαίων από μια μείωση της συναλλαγματικής ισοτιμίας ή μιας υποτίμησης του νομίσματος
στ) Πολιτικός κίνδυνος (country risk, political risk): Οι επενδύσεις που γίνονται σε ξένες κεφαλαιαγορές μπορεί να συνοδεύονται από τον κίνδυνο απώλειας αποδόσεων λόγω μιας απρόσμενης πολιτικής αστάθειας ή μιας εμπόλεμης κατάστασης, καταστάσεις που χαρακτηρίζονται από μεγάλη μεταβλητότητα.

Στο μεγαλύτερο μέρος της διεθνούς βιβλιογραφίας η έννοια του market risk είναι ταυτόσημη με το "μη διαφοροποιήσιμο κίνδυνο" και τον συστημικό κίνδυνο:

«συστημικός κίνδυνος (or systemic risk): Ο συστημικός κίνδυνος είναι ο κίνδυνος κατάρρευσης ενός ολόκληρου οικονομικού συστήματος ή της αγοράς, και σε αντίθεση με τον μη συστημικό κίνδυνο, επηρεάζει ένα μεγάλο εύρος χρεογράφων. Τα επιτόκια, οι υφέσεις στην οικονομία και οι πόλεμοι αντιπροσωπεύουν παράγοντες συστημικού κινδύνου επειδή επηρεάζουν ολόκληρη την αγορά και δεν μπορούν να αποφευχθούν μέσω διαφοροποίησης χαρτοφυλακίου. Ο συστημικός κίνδυνος, σε αντίθεση με τον μη συστημικό, μπορεί να μετριαστεί μόνο μέσω του hedging και είναι έμφυτος στην αγορά.»

ii) Λειτουργικός / Επιχειρησιακός Κίνδυνος (operational risk):

Ο λειτουργικός ή επιχειρησιακός κίνδυνος συμπεριλαμβάνει όλους τους κινδύνους που αντιμετωπίζει μία επιχείρηση κατά την διάρκεια λειτουργίας της και οφείλονται στον ανθρώπινο παράγοντα. Ο επιχειρησιακός κίνδυνος περιλαμβάνει τα προβλήματα που παρουσιάζονται στις εσωτερικές διεργασίες, στις εργασιακές σχέσεις, στα συστήματα μιας επιχείρησης και γενικά αφορά κάθε δυσλειτουργία που οφείλεται στον άνθρωπο. Επιχειρήσεις με μικρότερες απαιτήσεις για ανθρώπινο κεφάλαιο θα έχουν και χαμηλότερο επιχειρησιακό κίνδυνο.

iii) Κίνδυνος ρευστότητας (liquidity risk):

Είναι ο κίνδυνος να αδυνατεί η επιχείρηση να προβεί σε άμεση ρευστοποίηση μιας επένδυσης στην τιμή ισορροπίας, λόγω έλλειψης αγοραστικού ενδιαφέροντος ή λόγω στενότητας οικονομικών πόρων

iv) Πιστωτικός κίνδυνος (credit risk):

Είναι ο πιο σημαντικός κίνδυνος για τα πιστωτικά ιδρύματα, για το όποιο θα ακολουθήσει εκτενής αναφορά.

Παρατηρήσεις:

Τα διάφορα είδη κινδύνων δεν είναι ανεξάρτητα μεταξύ τους. Δηλαδή είναι πιθανό η αύξηση του market risk εξαιτίας ενός πολέμου για παράδειγμα να έχει σαν αποτέλεσμα την αύξηση του credit risk μιας επιχείρησης και η αδυναμία αυτή να εκπληρώσει τις υποχρεώσεις της, να μετατραπεί σε κίνδυνο ρευστότητας για τον πιστοδοτικό της οργανισμό.

Όπως προαναφέρθηκε τα όρια μεταξύ των κινδύνων δεν είναι αυστηρά προκαθορισμένα. Γι'αυτό και σε αρκετές περιπτώσεις συγχέονται οι έννοιες ή επικαλύπτουν η μια την άλλη. Στο βιβλίο των Duffie και Singleton για παράδειγμα γίνεται διάκριση μεταξύ systemic και market risk.

2.4 Πιστωτικός κίνδυνος (credit risk)

Ο *πιστωτικός κίνδυνος* είναι ο σημαντικότερος και ουσιαστικότερος κίνδυνος τον οποίο έχει να αντιμετωπίσει μία επιχείρηση ή ένας οργανισμός. στο τραπεζικό σύστημα, το ποσοστιαίο μέγεθος του στο σύνολο των κινδύνων είναι στο 54% περίπου και ακολουθούν ο λειτουργικός κίνδυνος στο 24% και ο κίνδυνος αγοράς στο 19%.

Ορισμός:
Πιστωτικός κίνδυνος είναι ο κίνδυνος που προέρχεται από ενδεχόμενη αδυναμία του πιστούχου να εκπληρώσει τις υποχρεώσεις του που πηγάζουν από τις συμβατικές σχέσεις του με την τράπεζα η κάθε άλλο πιστοδοτικό οργανισμό.

Τρία βασικά στοιχεία του πιστωτικού κινδύνου είναι:
i) ο κίνδυνος αθέτησης
ii) ο κίνδυνος έκθεσης
iii) ο κίνδυνος ανάκτησης

Πιστοληπτική ικανότητα

Ως πιστοληπτική ικανότητα ορίζεται η ικανότητα των πιστούχων να ανταποκρίνονται στις οικονομικές τους υποχρεώσεις. Η ικανότητα αυτή προκύπτει από την αξιολόγηση συγκεκριμένων ποιοτικών και ποσοτικών χαρακτηριστικών των πιστούχων και αποτυπώνεται κλίμακα διαβάθμισης στην οποία αντιστοιχεί συγκεκριμένη πιθανότητα ασυνέπειας. συγκεκριμενα στην επιχειρηματική τραπεζική η αξιολόγηση της πιστοληπτικής ικανότητας των επιχειρήσεων αποτυπώνεται σε δεκάβαθμη κλίμακα διαβάθμισης: AA, A, BB, B, C, D, E, F, G, H (AA επιχείρηση με υψηλή πιστοληπτική ικανότητα). Στη λιανική τραπεζική οι πελάτες κατανέμονται σε ζώνες κινδύνου (low-medium-high risk).

Ο πιστωτικός κίνδυνος του πιστοδοτικού οργανισμού και πιστοληπτική ικανότητα του πελάτη είναι δυο αλληλένδετες έννοιες, αλλά απόλυτα διακριτές μεταξύ τους. Όπως αναφέρθηκε παραπάνω, ο πιστωτικός κίνδυνος δημιουργείται, όταν η πιστοληπτική ικανότητα της χρηματοδοτούμενης επιχείρησης είναι χαμηλή. Παράλληλα ενδεχόμενη χαμηλή πιστοληπτική ικανότητα ενός πελάτη ή μιας επιχείρησης έχει αποτέλεσμα να αυξήσει τον πιστωτικό κίνδυνο του πιστοδοτικού του οργανισμού.

Μέθοδοι μέτρησης πιστωτικού κινδύνου

Το νέο πλαίσιο κεφαλαιακής επάρκειας, γνωστό και ως κανονιστικό πλαίσιο της Βασιλείας ΙΙ, απαιτεί από τα πιστωτικά ιδρύματα να δημιουργήσουν τις κατάλληλες μεθοδολογίες για την εκτίμηση της Πιθανότητας Αθέτησης (Probability of Default ή PD), της Ζημιάς Δεδομένης της Αθέτησης (Loss Given Default ή LGD) και του Χρηματοδοτικού Ανοίγματος (Exposure At Default ή EAD). Η διαδικασία αυτή αφορά όλα τα χαρτοφυλάκια πιστώσεων, συμπεριλαμβανομένων και των χαρτοφυλακίων λιανικής τραπεζικής.

Η κλασική μέθοδος αποτίμησης του πιστωτικού κινδύνου, μέσω της εκτίμησης των παραπάνω παραμέτρων, βασίζεται στα υποδείγματα αποτίμησης του πιστωτικού κινδύνου που ισχύουν στα χαρτοφυλάκια επιχειρηματικών δανείων (Corporate Loans), με αποτέλεσμα να μην μπορούν να εφαρμοστούν άμεσα στα δάνεια λιανικής τραπεζικής. Αυτό συμβαίνει γιατί τα δάνεια λιανικής τραπεζικής και συγκεκριμένα η συμπεριφορά του τρόπου αποπληρωμής των ιδιωτών διαφέρει σε σημαντικό βαθμό με την αντίστοιχη συμπεριφορά των επιχειρήσεων.

Η παραδοσιακή διαδικασία βαθμοδότησης της πιστοληπτικής συμπεριφοράς των δανείων λιανικής τραπεζικής είναι το credit scoring και η κατασκευή scorecards.

2.5 Credit scoring – scorecards- τεχνικές μέτρησης πιστωτικού κινδύνου

Το credit scoring, όπως προαναφέρθηκε έχει σκοπό την αξιολόγηση της πιστοληπτικής ικανότητας των πελατών με την κατασκευή σκοροκαρτών.

«*η ανάπτυξη των scorecards είναι σημαντική στις λιανικές οικονομικές υπηρεσίες στον τομέα του πιστωτικού κινδύνου, το πλαίσιο της Βασιλείας ΙΙ και το marketing προϊόντων*.» αναφέρει μεταξύ άλλων ο Michael C.S. Wong

Η λογική της σκοροκάρτας είναι η συγκέντρωση χαρακτηριστικών του πελάτη, η εισαγωγή τους σε ένα μαθηματικό μοντέλο και η αντιστοίχιση του πελάτη με κάποιο score. Ουσιαστικά μια σκοροκάρτα είναι μια συνάρτηση f(x) : $R^n \rightarrow$ R όπου x_i είναι οι πληροφορίες του πελάτη και y είναι το score του πελάτη. Το score αυτό αντιστοιχεί σε μια πιθανότητα αποπληρωμής των υποχρεώσεων του πελάτη. Επομένως υψηλό score δείχνει υψηλή πιθανότητα αποπληρωμής των υποχρεώσεων σε προκαθορισμένο χρονικό διάστημα και χαμηλό score δείχνει μικρότερη πιθανότητα.

Τα στάδια για την υλοποίηση μιας scorecard είναι 3:

i) Model building: Η κατασκευή μοντέλων, δηλαδή στατιστικών υποδειγμάτων.
ii) Validation: Η επικύρωση του μοντέλου
iii) Implementation.

Στην κατασκευή στατιστικών υποδειγμάτων, τα πιστωτικά ιδρύματα διαφοροποιούν την μεθοδολογία και τα δεδομένα που αντλούν, ανάλογα με την κατηγορία των πελατών που θέλουν να μελετήσουν. Εν προκειμένω, τα στατιστικά υποδείγματα ανάλυσης πιστοληπτικής ικανότητας (credit scoring) διαχωρίζονται σε δύο μεγάλες κατηγορίες: στα application scoring και τα behavioral scoring υποδείγματα:

- Application scoring υποδείγματα τα οποία αξιολογούν αιτήσεις πελατών και χρησιμοποιούνται για την έγκριση ή μη της αίτησης. Τα κριτήρια που χρησιμοποιούνται είναι είναι προσωπικά στοιχεία του πελάτη οικονομικά στοιχεία και στοιχεία της συναλλακτικής του συμπεριφοράς με άλλα προϊόντα της τράπεζας που έχει ήδη στην κατοχή του.
- Behavioral scoring υποδείγματα αναφέρονται σε υφιστάμενους πελάτες και αξιολογούν την συμπεριφορά του πελάτη ως προς τα προϊόντα που έχει. Χρησιμοποιούνται για λόγους collections αλλά και πρόβλεψης της κινδύνου αθέτησης τους επόμενους μήνες.

Στην συνέχεια, παρουσιάζονται κάποιες βασικές κατηγορίες στατιστικών υποδειγμάτων που χρησιμοποιούνται κατά κύριο λόγο στην εκτίμηση του πιστωτικού κινδύνου και συγκεκριμένα στην εκτίμηση της πιθανότητας αθέτησης, συνδέοντας γνωστά στατιστικά υποδείγματα που χρησιμοποιούνται από σύγχρονα πιστωτικά ιδρύματα σε παγκόσμιο επίπεδο.

Παραμετρικές Τεχνικές

Οι τεχνικές αυτές περιλαμβάνουν στατιστικές και οικονομετρικές μεθόδους, που αποτελούν και τον παραδοσιακό τρόπο επίλυσης του προβλήματος.

Η πρώτη τεχνική καλείται Γραμμική Διακριτική Ανάλυση και υλοποιείται από τη σχέση: $Z = w_0 + w_1x_1 + w_2x_2w_3x_3 + w_nx_n$ όπου w_0 είναι ο σταθερός όρος, x_1, x_2, x_3,…,x_n είναι οι ανεξάρτητες μεταβλητές που αναφέρονται στα χαρακτηριστικά που μελετώνται (χρηματοοικονομικοί δείκτες) και w_1, w_2, w_3,…,w_n είναι οι συντελεστές των χαρακτηριστικών x για τις μεταβλητές. Το υπόδειγμα αυτό χρησιμοποιήθηκε κυρίως από τον Altman, όταν το 1968 ανέπτυξε το ZScore Model για την κατάταξη των επιχειρήσεων σε υγιείς και πτωχυμένες.

Όσον αφορά τις οικονομετρικές μεθόδους, οι πιο γνωστές είναι:

i) το γραμμικό υπόδειγμα πιθανότητας
ii) το λογιστικό υπόδειγμα πιθανότητας
iii) το κανονικό υπόδειγμα πιθανότητας

Η διαφορά των τεχνικών αυτών έγκειται κυρίως στη μορφή της συνάρτησης.

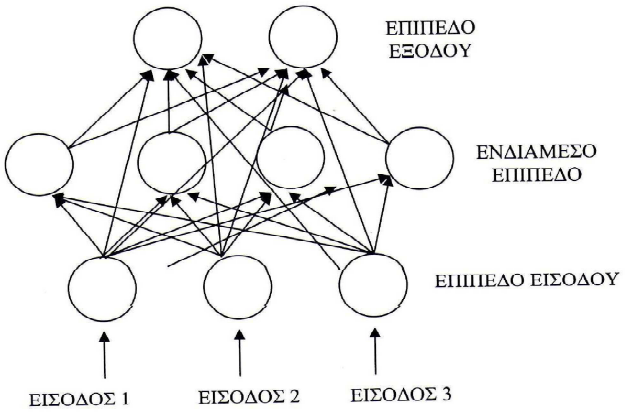
Μη Παραμετρικές Τεχνικές

Οι τεχνικές αυτές έχουν αναπτυχθεί κυρίως τα τελευταία έτη. Η σημαντικότητά τους άπτεται στο γεγονός ότι δεν υιοθετούν περιοριστικές στατιστικές υποθέσεις, αλλά μπορούν να προσαρμόζονται ικανοποιητικά είτε ως γραμμικά είτε ως μη γραμμικά υποδείγματα, ανάλογα με τα δεδομένα που χρησιμοποιούνται. Οι σημαντικότερες από αυτές τις τεχνικές είναι οι εξής:

1) Νευρωνικά δίκτυα (Neural Networks):

Η ανάπτυξη των Νευρωνικών δικτύων στηρίχτηκε κυρίως στην τεχνητή νοημοσύνη. Σκοπός τους είναι η προσομοίωση του τρόπου λειτουργίας του ανθρώπινου εγκεφάλου τη στιγμή που λαμβάνει εξωτερικά ερεθίσματα. Ένα τυπικό Νευρωνικό δίκτυο. απαρτίζεται από ένα επίπεδο εισόδου, οι κόμβοι του οποίου αντιστοιχούν σε κάθε χαρακτηριστικό (οικονομικοί δείκτες), ένα επίπεδο εξόδου, που αποτελείται από έναν αριθμό κόμβων ο οποίος είναι ίσος πάντα με τις κατηγορίες ταξινόμησης και μία σειρά ενδιάμεσων επιπέδων τα οποία αποτελούνται και αυτά από κόμβους.

Απεικόνιση Τεχνητού Νευρωνικού δικτύου



Πηγή: «διαχείριση Πιστωτικού Κινδύνου», Ζοπουνίδης, Λεμονάκης (2009)

2) Μηχανική Μάθηση:

Στη Μηχανική Μάθηση χρησιμοποιούνται τα δένδρα απόφασης και τα δένδρα ταξινόμησης και παλινδρόμησης. Στα πρώτα, κάθε κόμβος περιλαμβάνει τα κριτήρια αξιολόγησης, τα οποία ελέγχονται από τις συνθήκες που βρίσκονται στα «κλαδιά» του δένδρου, ενώ στα «φύλλα» βρίσκονται οι κατηγορίες ταξινόμησης. Τα δεύτερα είναι τα λεγόμενα CART (Classification and Regression Trees). Η τεχνική αυτή βασίζεται σε αλγόριθμους αναδρομικής διαφοροποίησης και παράγει ένα δένδρο παλινδρόμησης, αν η εξαρτημένη

μεταβλητή είναι συνεχής, ή ένα δένδρο ταξινόμησης, αν η εξαρτημένη μεταβλητή είναι διακριτή. Το βασικό πλεονέκτημα αυτών των συστημάτων είναι ότι μπορούν να επεξεργαστούν και ποιοτικά δεδομένα εξίσου ικανοποιητικά, καθώς και ότι η μορφή του τελικού υποδείγματος είναι κατανοητή.

3) Έμπειρα Συστήματα:

Αποτελούν το πιο γνωστό πεδίο εφαρμογής της τεχνητής νοημοσύνης. Είναι ένα αλληλεπιδραστικό πρόγραμμα υπολογιστή, που έχει τη δυνατότητα να χρησιμοποιεί «γνώσεις» για να βοηθήσει το χρήστη στην επίλυση ενός προβλήματος. Έχει τη δυνατότητα να δικαιολογεί ακόμη και τις μεθόδους που χρησιμοποίησε για να δώσει το τελικό αποτέλεσμα. Η εφαρμογή τους στον τραπεζικό κλάδο για την αξιολόγηση της πιστοληπτικής ικανότητας χρονολογείται από τις αρχές του ’90, ενώ για να παρακαμφθούν τα μειονεκτήματα που έχουν παρατηρηθεί, τα Ε.Σ. χρησιμοποιούνται σε συνδυασμό με άλλες τεχνικές, όπως τα νευρωνικά δίκτυα.

4) Ασαφής Λογική:

Η θεωρία αυτή αναπτύχθηκε από τον Zadeh (1983,2001) και
παρέχει τα κατάλληλα μέσα για τη μοντελοποίηση, αναπαράσταση και επίλυση προβλημάτων λήψης απόφασης, τα οποία εμπεριέχουν ένα βαθμό ασάφειας. Βασικά στοιχεία της θεωρίας αυτής είναι η έννοια του ασαφούς συνόλου και η συνάρτηση συμμετοχής.

5) Πολυκριτήρια Ανάλυση:

Τα στάδια που ακολουθούνται στην τεχνική αυτή είναι: ο καθορισμός του αντικειμένου της απόφασης, ο καθορισμός μίας έγκυρης «οικογένειας» κριτηρίων, η ανάπτυξη του υποδείγματος σύνθεσης κριτηρίων και η υποστήριξη της απόφασης (Ζοπουνίδης (2004)).

6) Μηχανές διανυσμάτων Υποστήριξης:

Η μέθοδος αυτή προτάθηκε αρχικά από τον Vapnik (1995) για την επίλυση προβλημάτων αξιολόγησης. Στόχος της μεθόδου είναι η δημιουργία μίας «βέλτιστης γραμμικής διαχωριστικής υπερεπιφάνειας» η οποία έχει τη μορφή

$f(x) = \tilde{w}\tilde{x} + b$, όπου w είναι το διάνυσμα των συντελεστών των μεταβλητών, x είναι το διάνυσμα των μεταβλητών και b ένας σταθερός όρος, και μεγιστοποιεί την απόσταση μεταξύ των δύο κατηγοριών.

Οι τεχνικές αυτές χρησιμοποιούνται κυρίως για την ανάπτυξη εξωτερικών συστημάτων διαβάθμισης της πιστοληπτικής ικανότητας των δανειοληπτών και αποτελούν τα παραδοσιακά μοντέλα. Οι τρεις τύποι στατιστικών αναλύσεων που τυχαίνουν ευρείας αποδοχής είναι η λογιστική παλινδρόμηση, η διακριτική ανάλυση και τα νευρωνικά δίκτυα.

Στο επόμενο κεφάλαιο θα αναπτύξουμε και θα συγκρίνουμε τις δύο από τις σημαντικότερες τεχνικές διαβάθμισης πελατών σε έναν τραπεζικό οργανισμό: τη διακριτή ανάλυση και τη λογιστική παλινδρόμηση.

Κεφάλαιο 3

Διαχωριστική ανάλυση

Ronald A. Fisher, 1936: «*Ο μηχανισμός επεξεργασίας που οικοδομήθηκε σε εφαρμογές απείρων δεδομένων, δεν είναι αρκετά ακριβής για απλά εργαστηριακά δεδομένα. Μόνο με συστηματική επιλογή προβλημάτων με λίγα δείγματα, ανάλογα με τα ιδιαίτερα χαρακτηριστικά τους, μπορούμε να έχουμε ακριβή τεστ σε πρακτικά δεδομένα.*»

3.1 Εισαγωγή

Η Γραμμική Διαχωριστική Ανάλυση ή Linear Discriminant Analysis ή LDA είναι μια τεχνική εξαγωγής χαρακτηριστικών που έχει εφαρμοστεί επιτυχώς σε πολλά στατιστικά προβλήματα αναγνώρισης. Εμπνευστής της ιδέας της διαχωριστικής ανάλυσης είναι ο Fisher (1935) και χρησιμοποιήθηκε ευρύτατα από τους Altman (1968), Altman et al (1977), Narayanan (1977), Dekin (1972), Libby (1975), Dambolena and Kgoury (1980), Casey and Bartczak (1985) κ.ά. Στην Ελλάδα, εφαρμόστηκε από τους Γραμματικό και Γκλούμπο (1984), με ικανοποιητικά αποτελέσματα.

Σκοπός της διαχωριστικής ανάλυσης είναι να χωρίσει δυο ή περισσότερα σύνολα με διαφορετικά χαρακτηριστικά σε ομάδες μεγιστοποιώντας τη μεταξύ κλάσεων διαχωριστικότητα και την εντός κλάσης μεταβλητότητα. Δημιουργεί μια συνάρτηση διάκρισης που αποτελεί τον γραμμικό συνδυασμό των προαναφερθέντων χαρακτηριστικών (ανεξάρτητες μεταβλητές) και βάσει της συνάρτησης υπολογίζεται ένα score. Χρησιμοποιείται κατά κανόνα σε τομείς risk management τραπεζικών οργανισμών με σκοπό την ταξινόμηση πελατών σε κατηγορίες. Έτσι αν ο πληθυσμός ήταν το πελατολόγιο του επιχειρηματικού κέντρου μιας τράπεζας, τα σύνολα στα οποία διακρίνει η διαχωριστική ανάλυση θα μπορούσαν να είναι υγιείς επιχειρήσεις και επιχειρήσεις πιθανές να οδηγηθούν σε χρεοκοπία το επόμενο τρίμηνο.

Η βασική ιδέα της διαχωριστικής ανάλυσης (ή διακριτικής ανάλυσης, discriminant analysis) είναι να κατατάξει παρατηρήσεις (συνήθως πολυδιάστατες) σε γνωστούς πληθυσμούς, με γνωστές κατανομές για κάθε πληθυσμό. Έστω ότι υπάρχουν k υποπληθυσμοί (ομάδες): $\Pi_1, \Pi_2, \Pi_3, \dots, \Pi_k$, με $k \geq 1$. Για τον κάθε υποπληθυσμό Π_i, υπάρχει μια κατανομή f_i. Η διαχωριστική αναλυση έχει 2 στοχους:

1) τη διαχώριση ενός πληθυσμού σε ευδιάκριτα σύνολα (υποπληθυσμούς) και
2) την ταξινόμηση των παρατηρήσεων στους προηγούμενους γνωστούς πληθυσμούς με γνωστές κατανομές για κάθε πληθυσμό, με την βοήθεια ενός κανόνα.

Αξιοσημείωτο είναι ότι αν και η διαχωριστική ανάλυση μοιάζει με την ανάλυση κατά ομάδες (συστάδες), έχει σημαντικές διαφορές από αυτή. Η πιο σημαντική από τις διαφορές είναι ότι ενώ στην διαχωριστική ανάλυση οι ομάδες είναι γνωστές , στην ανάλυση κατά ομάδες σκοπός είναι η εύρεση αυτών των ομάδων. Επιπλέον, στην διαχωριστική ανάλυση σκοπός μας είναι να δημιουργήσουμε ένα κανόνα που θα μας βοηθήσει να λάβουμε αποφάσεις στο μέλλον, ενώ στην ανάλυση κατά ομάδες ο κύριος στόχος είναι να δημιουργήσουμε ομοειδείς ομάδες με σκοπό την κατανόηση των ήδη υπαρχόντων στοιχείων και την μείωση της διασποράς σε επιμέρους ομάδες.

3.2 Η διαχωριστική συνάρτηση του Fisher

Ο Fisher προσπάθησε να διαχωρίσει δυο πληθυσμούς- ομάδες, χωρίς να κάνει καμία υπόθεση για την κατανομή των πληθυσμών. Το μόνο που υπέθεσε είναι ότι οι δυο πληθυσμοί έχουν ίσους πίνακες συνδιασποράς $\Sigma_1 = \Sigma_2 = \Sigma$, αφού χρησιμοποίησε την συνδυασμένη κοινή (pooled) εκτίμηση S_{pooled}.

Έστω $y_1 = l' * x_1$ και $y_2 = l' * x_2$ οι συναρτήσεις (score) που αντιστοιχούν στους πληθυσμούς Π_1 και Π_2 . Τα score των δύο πληθυσμών είναι όσο το δυνατόν πιο απομακρυσμένα, έτσι ώστε να μπορούμε να διαχωρίσουμε τους δύο πληθυσμούς και να κατατάξουμε κάθε καινούρια παρατήρηση σε έναν από τους πληθυσμούς. Ένα μέτρο του κατά πόσο κοντά είναι αυτά τα σκορ δίνεται από την σχέση:

$D = \frac{|\bar{y}_1 \bar{y}_2|}{s_y}$, όπου $s_y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j}\bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j}\bar{y}_2)^2}{n_1+n_2-2}$ είναι η συνολική εκτίμηση της διασποράς. Ο fisher δηλαδή πήρε την απόσταση των μέσων των δύο πληθυσμών διαιρούμενο με την τυπική απόκλιση, έτσι ώστε να απαλειφούν οι μονάδες μέτρησης. Ο σκοπός είναι η μεγιστοποίηση της απόστασης D γεγονός που συνεπάγεται την όσο γίνεται μεγαλύτερη διαφοροποίηση των scores των δύο πληθυσμών.

Θεώρημα

Η απόσταση D^2 μεγιστοποιείται όταν $D^2 = (\bar{x}_1 - \bar{x}_2)' * S_{pooled}^{-1} * (\bar{x}_1 - \bar{x}_2)$

Απόδειξη

$$D = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y} \Rightarrow D^2 = \frac{|\bar{y}_1 - \bar{y}_2|^2}{s_y^2} \Rightarrow D^2 = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}$$

$$\Rightarrow D^2 = \frac{(l' \cdot \bar{x}_1 - l' \cdot \bar{x}_2)^2}{s_y^2} \Rightarrow D^2 = \frac{[l' \cdot (\bar{x}_1 - \bar{x}_2)]^2}{l' \cdot S_{pooled} \cdot l}$$

Από την ισότητα Caushy – Schwarz έχουμε πως 2 τυχαία px1 διανύσματα α και β ισχύει η σχέση:

$$(a' \cdot b)^2 \le (a' \cdot a) \cdot (b' \cdot b)$$

Επειδή ο πίνακας διακυμάνσεων είναι θετικά ορισμένος, μπορούμε να θέσουμε:

$$a = S_p^{\frac{1}{2}} \cdot l \text{ και } b = S_p^{-\frac{1}{2}} \cdot (\bar{x}_1 - \bar{x}_2) \text{, όπου: } S_p = S_{pooled}$$

Άρα η προηγούμενη ισότητα γράφεται:

$$\left[\left(S_p^{\frac{1}{2}} \cdot l \right)' \cdot S_p^{-\frac{1}{2}} \cdot (\bar{x}_1 - \bar{x}_2) \right]^2 \le \left[\left(S_p^{\frac{1}{2}} \cdot l \right)' \cdot \left(S_p^{\frac{1}{2}} \cdot l \right) \right] \cdot \left[\left(S_p^{-\frac{1}{2}} \cdot (\bar{x}_1 - \bar{x}_2) \right)' \cdot \left(S_p^{-\frac{1}{2}} \cdot (\bar{x}_1 - \bar{x}_2) \right) \right]$$

$$\Rightarrow \left[l' \cdot S_p^{\frac{1}{2}} \cdot S_p^{-\frac{1}{2}} \cdot (\bar{x}_1 - \bar{x}_2) \right]^2 \le \left[l' \cdot S_p^{\frac{1}{2}} \cdot S_p^{\frac{1}{2}} \cdot l \right] \cdot \left[(\bar{x}_1 - \bar{x}_2)' \cdot S_p^{-\frac{1}{2}} \cdot S_p^{-\frac{1}{2}} \cdot (\bar{x}_1 - \bar{x}_2) \right]$$

$$\Rightarrow \left[l' \cdot (\bar{x}_1 - \bar{x}_2) \right]^2 \le \left[l' \cdot S_p \cdot l \right] \cdot \left[(\bar{x}_1 - \bar{x}_2)' \cdot S_p^{-1} \cdot (\bar{x}_1 - \bar{x}_2) \right]$$

$$\Rightarrow \frac{\left[l' \cdot (\bar{x}_1 - \bar{x}_2) \right]^2}{l' \cdot S_p \cdot l} \le \frac{\left[l' \cdot S_p \cdot l \right] \cdot \left[(\bar{x}_1 - \bar{x}_2)' \cdot S_p^{-1} \cdot (\bar{x}_1 - \bar{x}_2) \right]}{l' \cdot S_p \cdot l}$$

$$\Rightarrow \frac{\left[l' \cdot (\bar{x}_1 - \bar{x}_2) \right]^2}{l' \cdot S_p \cdot l} \le (\bar{x}_1 - \bar{x}_2)' \cdot S_p^{-1} \cdot (\bar{x}_1 - \bar{x}_2)$$

$$\Rightarrow D^2 \le (\bar{x}_1 - \bar{x}_2)' \cdot S_p^{-1} \cdot (\bar{x}_1 - \bar{x}_2)$$

Επομένως το D^2 μεγιστοποιείται για $D^2 = (\bar{x}_1 - \bar{x}_2)' * S_{pooled}^{-1} * (\bar{x}_1 - \bar{x}_2)$

Για αυτή την τιμή έχουμε την μέγιστη απόσταση των μέσων των δυο πληθυσμών και άρα την καλύτερη διάκριση.

Ο διαχωριστικός κανόνας ολοκληρώνεται ορίζοντας την κρίσιμη τιμή m, η οποία είναι η μέση τιμή των $\bar{y}_1$ και $\bar{y}_2$. δηλαδη:

$$m = \frac{\bar{y}_1 + \bar{y}_2}{2} = \frac{l' \cdot x_1 + l' \cdot x_2}{2} = \frac{l' \cdot (x_1 + x_2)}{2} = \frac{(\bar{x}_1 - \bar{x}_2) \cdot S_p^{-1} \cdot (x_1 + x_2)}{2}$$

Τελικά ο κανόνας παίρνει την ακόλουθη μορφή:

- η παρατήρηση x τοποθετείται στον πληθυσμό π_1 αν $y \geq m$
- η παρατήρηση x τοποθετείται στον πληθυσμό π_2 αν $y < m$

Παρατηρήσεις

1) το y στον παραπάνω κανόνα είναι εκείνη η τιμή που μεγιστοποιεί την απόσταση D^2.

2) η μέθοδος που περιγράφηκε χρησιμοποιείται περισσότερο στην πράξη συγκριτικά με άλλες μεθόδους γιατί δεν κάνει καμία υπόθεση για κανονικότητα των πληθυσμών. Παρόλα αυτά, υποθέτει γραμμικότητα της συνάρτησης που διαχωρίζει τους 2 πληθυσμούς.

3) σε μερικές περιπτώσεις ο υπολογισμός του D^2 είναι πολύ σημαντικός καθώς χρησιμοποιείται για να ελέγξουμε αν οι πληθυσμιακοί μέσοι είναι στατιστικά ίσοι. Επομένως, ένας έλεγχος για τις διαφορές των διανυσμάτων των μέσων θα μπορούσε να απεικονίσει ως ένας έλεγχος για την «σπουδαιότητα» του διαχωρισμού ο οποίος πραγματοποιήθηκε. Αν υποθέσουμε πως οι δυο πληθυσμοί ακλουθούν την κανονική κατανομή τότε ο έλεγχος υπόθεσης θα είναι ο ακόλουθος:

$H_0 : \mu 1 = \mu 2$

$H_1 : \mu 1 \neq \mu 2$

Αν η υπόθεση H_0 οδηγηθεί σε απόρριψη, τότε συμπεραίνουμε ότι ο διαχωρισμός ανάμεσα στους πληθυσμούς π_1 και π_2 είναι σωστός.

3.3 Γενίκευση της διαχωριστικής ανάλυσης σε g πληθυσμούς

Η μέθοδος διαχωρισμού που παρουσιάστηκε στην προηγούμενη ενότητα, αναφέρεται στην περίπτωση 2 πληθυσμών. Ας δούμε την περίπτωση που έχουμε g (με g≥3) πληθυσμούς.

Έστω ότι έχουμε g πληθυσμούς που έχουν ίσους πίνακες διασποράς $\Sigma_1=\Sigma_2=\Sigma_3=\ldots=\Sigma_g=\Sigma$ διάστασης pXp και διανύσματα μέσων τιμών $\mu_1, \mu_2, \mu_3, \ldots, \mu_g$ διάστασης pX1. τότε το διάνυσμα της μέσης τιμής όλων των πληθυσμών είναι:

$f_i(x)$ η συνάρτηση πυκνότητας του πληθυσμού π_i, i=1,2,3…k

$$\overline{\mu} = \frac{\mu_1 + \mu_2 + \ldots + \mu_g}{g} \Rightarrow \overline{\mu} = \frac{\sum_{i=1}^{g} \mu_i}{g}$$

Στην μέθοδο του fisher η προϋπόθεση της κανονικότητας της κατανομής των πληθυσμών δεν απαραίτητη. Έστω $x'=[x_1, x_2, x_3, \ldots, x_g]$ και $Y'=[Y_1, Y_2, Y_3, \ldots, Y_g]$ είναι οι μεταβλητές που προκύπτουν από τον γραμμικό μετασχηματισμό των x παρατηρήσεων. Αναλυτικότερα:

$$Y = l' \cdot x = \begin{bmatrix} l' \cdot x_1 \\ l' \cdot x_2 \\ \ldots \\ l' \cdot x_g \end{bmatrix}$$

Η μέση τιμή του Y δοθέντος ότι η παρατήρηση προέρχεται από τον π_i πληθυσμό, είναι:

$$E(Y) = l' \cdot E(X / \pi_i) = l' \cdot \mu_i$$

Και η διασπορά:

$$Var(Y) = l' \cdot Cov(X) \cdot l = l' \cdot \Sigma \cdot l$$

Η ολική μέση τιμή προκύπτει:

$$\overline{\mu_Y} = \frac{1}{g} \cdot \sum_{i=1}^{g} \mu_{iY} = \frac{1}{g} \cdot \sum_{i=1}^{g} l' \cdot \mu_i = l' \cdot \left(\frac{1}{g} \cdot \sum_{i=1}^{g} \mu_i \right) = l' \cdot \overline{\mu}$$

Σχηματίζουμε τον λόγο:

$$\frac{\sum_{i=1}^{g} (\mu_{iY} - \overline{\mu_Y})^2}{\sigma_Y^2} = \frac{\sum_{i=1}^{g} (l' \cdot \mu_i - l' \cdot \overline{\mu})^2}{l' \cdot \Sigma \cdot l} = \frac{l' \cdot \left(\sum_{i=1}^{g} (\mu_i - \overline{\mu}) \cdot (\mu_i - \overline{\mu})' \right) \cdot l}{l' \cdot \Sigma \cdot l} \Rightarrow \frac{\sum_{i=1}^{g} (\mu_{iY} - \overline{\mu_Y})^2}{\sigma_Y^2} = \frac{l' \cdot B_0 \cdot l}{l' \cdot \Sigma \cdot l}$$

Όπου:

$$B_0 = \sum_{i=1}^{g} (\mu_i - \overline{\mu}) \cdot (\mu_i - \overline{\mu})'$$

Ο παραπάνω λόγος αποτελεί ένα μέτρο μεταβλητότητας ανάμεσα στις ομάδες(variabilitity between groups) των Y τιμών σε σχέση νε την κοινή μεταβλητότητα εντός των ομάδων (variability within groups). Τότε μπορούμε να επιλέξουμε το l που μεγιστοποιεί αυτό τον λόγο, ώστε να επιτύχουμε τον καλύτερο διαχωρισμό. Στην πράξη τα μ_i και Σ είναι άγνωστα για αυτό τα εκτιμούμε με $\overline{x_i}$ και S_{pooled}. Έστω για κάθε πληθυσμό π_iαντιστοιχεί ένα δείγμα μεγέθους n_i, με i=1,2,3,…,g. Η δειγματική μέση τιμή για τον π_i πληθυσμό είναι: $\overline{x_i} = \frac{1}{n_i} * \sum_{j=1}^{n_i} x_{ij}$

Η ολική μέση τιμή είναι:

$$\overline{x} = \frac{n_1 \cdot x_1 + n_2 \cdot x_2 + \ldots + n_g \cdot x_g}{n_1 + n_2 + \ldots + n_g} = \frac{\sum_{i=1}^{g} n_i \cdot \overline{x_i}}{\sum_{i=1}^{g} n_i} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{g} n_i}$$

Το οποίο είναι ένα διάνυσμα διάστασης pX1.

Όμοια το B_0 εκτιμάται από το

$$\hat{B}_0 = \sum_{i=1}^{g} (\overline{x_i} - \overline{x}) \cdot (\overline{x_i} - \overline{x})'$$

Είναι γνωστό ότι:

$$S_{pooled} = \frac{(n_1-1)\cdot S_1 + (n_2-1)\cdot S_2 + \ldots + (n_g-1)\cdot S_g}{n_1+n_2+\ldots+n_g-g} = \frac{\sum_{i=1}^{g}(n_i-1)\cdot S_i}{n_1+n_2+\ldots+n_g-g}$$

$$\Rightarrow S_{pooled} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)\cdot(x_{ij}-\bar{x}_i)'}{n_1+n_2+\ldots+n_g-g} \quad \text{αν} \quad W = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)\cdot(x_{ij}-\bar{x}_i)'$$

$$\Rightarrow S_{pooled} = \frac{W}{n_1+n_2+\ldots+n_g-g}$$

$$\Rightarrow W = (n_1+n_2+\ldots+n_g-g)\cdot S_{pooled}$$

Επομένως το ίδιο i που μεγιστοποιεί το $\frac{\hat{\imath}'*\hat{B}_0*\hat{\imath}}{\hat{\imath}'*S_{pooled}*\hat{\imath}}$ μεγιστοποιεί επίσης και το $\frac{\hat{\imath}'*\hat{B}_0*\hat{\imath}}{\hat{\imath}'*W*\hat{\imath}}$

Εξάλλου μπορούμε να παρουσιάσουμε το βέλτιστο $\hat{\imath}$ στην περισσότερο συνηθισμένη μορφή ως ιδιοδιανύσματα $\hat{e}_i$ του πίνακα $W^{-1}*\widehat{B_o}$ επειδη αν $W^{-1}*\widehat{B_o}*\hat{e} = \hat{\lambda}*\hat{e}$ τότε:
$S_{pooled}^{-1}*\widehat{B_o}*\hat{e} = \hat{\lambda}*\left(\widehat{n_1}+\widehat{n_2}+\widehat{n_3}+\widehat{n_g}-\hat{g}\right)*\hat{e}.$

Έστω $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_s > 0$, όπου s≤min(g-1,p) είναι οι μη μηδενικές ιδιοτιμές του $W^{-1}*\widehat{B_o}*\hat{e}$ και $\widehat{e_1}, \widehat{e_2}, \widehat{e_3}, \ldots, \widehat{e_s}$ είναι τα ιδιοδιανύσματα τέτοια ώστε $\hat{e}'*S_{pooled}*\hat{e} = 1$.

Τότε το διάνυσμα των συντελεστών $\hat{\imath}$ το οποίο μεγιστοποιεί τον λόγο:

$\frac{\hat{\imath}'*\hat{B}_o*\hat{\imath}}{\hat{\imath}'*W*\hat{\imath}} = \frac{\hat{\imath}'*\left[\sum_{i=1}^{g}(\bar{x}_i-\bar{x})*(\bar{x}_i-\bar{x})'\right]*\hat{\imath}}{\hat{\imath}'*\left[\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\bar{x}_{ij}-\bar{x}_i)*(\bar{x}_{ij}-\bar{x}_i)'\right]*\hat{\imath}}$ (1) δίνεται από: $\hat{\imath}_1 = \hat{e}_1$.

Ο γραμμικός συνδυασμός $\widehat{\imath_1}'*x$ καλείται ως ο πρώτος δειγματικός διαχωριστής.

Η επιλογή $\hat{l}_2 = \hat{e}_2$ παράγει το δεύτερο δειγματικό διαχωριστή, $\hat{l}_2^{'} * x$ υπό την προϋπόθεση $Cov(\hat{l}_1^{'} * x, \hat{l}_2^{'} * x) = 0$.

Επαγωγικά το $\hat{l}_k^{'} * x = \widehat{e_k}^{'} * x$ είναι ο k οστός δειγματικός διαχωριστής, με $k \leq s$, υπό την υπόθεση $Cov(\hat{l}_k^{'} * x, \hat{l}_i^{'} * x) = 0$, με $i \leq k$.

Ισχυρισμός

$Var(\hat{l}_i^{'} * x)=1$, για i=1,2,3,…,s.

Απόδειξη

Έστω $\Sigma = P' * \Lambda * P$, όπου Λ ένας διαγώνιος πίνακας με θετικά στοιχεία λ_i.

Τότε $\Lambda^{1/2}$ είναι ο διαγωνιος πινακας με στοιχεία $\sqrt{\lambda_i}$. είναι $\Sigma^{1/2} = P' * \Lambda^{1/2} * P$ και

$\Sigma^{-1/2} = P' * \Lambda^{-1/2} * P$

Αν $\alpha = \Sigma^{1/2} * l$ τότε $\alpha' * \alpha = l' * \Sigma^{1/2} * \Sigma^{1/2} * l = l' * \Sigma * l$ και

$\alpha^{'} * \Sigma^{-1/2} * Bo * \Sigma^{-1/2} * \alpha = l^{'} * \Sigma^{1/2} * \Sigma^{-1/2} * Bo * \Sigma^{-1/2} * \Sigma^{1/2} * l = l^{'} * Bo * l$

άρα , αρκεί να μεγιστοποιήσουμε τον λόγο $\frac{\alpha' * \Sigma^{-1/2} * Bo * \Sigma^{-1/2} * \alpha}{(\alpha' * \alpha)}$ (πάνω στο διάνυσμα α). Αποδεικνύεται ότι το μέγιστο αυτού του λόγου είναι λ_1, η μεγαλύτερη ιδιοτιμή του $\Sigma^{-1/2} * Bo * \Sigma^{-1/2}$. Το μεγιστο συμβαίνει όταν $\alpha = e_1$, που είναι το κανονικοποιημένο διάνυσμα που σχετίζεται με το λ_1.

Επειδή $e_1 = a = \Sigma^{1/2} * l_1 => l_1 = \Sigma^{-1/2} * e_1$ και

$Var(l_1^{'} * x) = l_1^{'} * \Sigma * l_1 = e'_1 * \Sigma^{-1/2} * \Sigma * \Sigma^{-1/2} * e_1 = e'_1 * \Sigma^{-1/2} * \Sigma^{1/2} * \Sigma^{1/2} * \Sigma^{-1/2} * e_1 = e'_1 * e_1 = 1$

Επίσης το α κάθετο στο e_1, άρα το μέγιστο του παραπάνω λόγου είναι για $\alpha = e_2$, που είναι το κανονικοποιημένο ίδιο διάνυσμα που αντιστοιχεί στην ιδιοτιμή λ_2.

Όμοια βρίσκουμε $e_2 = \alpha = \Sigma^{1/2} * l_2 => \Sigma^{-1/2} * e_2$ και

$\mathrm{Cov}(l'_2 * x, l'_1 * x) = l'_2 * \Sigma * l_1 = e'_2 * \Sigma^{-1/2} * \Sigma * \Sigma^{-1/2} * e_1 = e'_2 * e_1 = 0,$ αφού το e_2 είναι καθετο στο e_1.

Ακόμα , $\mathrm{Var}(l'_2 * x) = l'_2 * \Sigma * l_2 = e'_2 * e_2 = 1.$

Με την ίδια λογική ακολουθούμε τη διαδικασία και για τους υπόλοιπους διαχωριστές.

Παρατήρηση

1) Τις ιδιοτιμές $\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \ldots, \hat{\lambda}_s$ τις βρίσκουμε αν λύσουμε την ορίζουσα $\left|W^{-1} * \hat{B}_o - \lambda * I\right| = 0,$ όπου I είναι ο μοναδιαίος πίνακας και τα ιδιοδιανύσματα τα βρίσκουμε λύνοντας την εξίσωση

$$\left(W^{-1} * \hat{B}_o - \lambda_i * I\right) * \hat{l}_i = 0$$

Έστω $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_s \end{bmatrix}$ το διάνυσμα των διαχωριστών (διάστασης sx1), που έχει μέση τιμή

$\mu_{iY} = \begin{bmatrix} \mu_{iY_1} \\ \vdots \\ \mu_{iY_s} \end{bmatrix} = \begin{bmatrix} l'_1 \mu_i \\ \vdots \\ l'_s \mu_i \end{bmatrix}$ για τον πληθυσμό π_i και πινακα διασπορας I, για ολους τους πληθυσμους.

Επειδή τα Y έχουν διασπορές μονάδες και μηδενικές συνδιασπορές, τότε η τετραγωνική απόσταση του Y=y από το μ_{iY} είναι:

$$(y - \mu_{iY})' * (y - \mu_{iY}) = \sum_{j=1}^{s} (y_j - \mu_{iY_j})^2$$

Ένας κανόνας κατάταξης είναι να τοποθετηθεί το y στον πληθυσμό π_κ αν η τετραγωνικη αποσταση του y από το $\mu_{\kappa Y}$ είναι μικρότερη από την τετραγωνική απόσταση του y από το μ_{iY} για κάθε $i \neq k$

Πρόταση

Για τον πίνακα $\Sigma^{-1/2} * \text{Bo} * \Sigma^{-1/2}$ υπάρχουν s μη μηδενικές ιδιοτιμές τέτοιες ώστε $s \leq \min\{g-1,p\}$

Απόδειξη

Ο $\Sigma^{-1/2} * \text{Bo}$ είναι ένας pxp πίνακας, έτσι πρέπει $s \leq p$.

Έχουμε g πληθυσμούς που θέλουμε να διαχωρίσουμε. Αν σχηματίσουμε τα g διανύσματα των διαφορών $\mu_1 - \bar{\mu}, \mu_2 - \bar{\mu}, \mu_3 - \bar{\mu}, \ldots, \mu_g - \bar{\mu}$ παρατηρούμε ότι:

$$(\mu_1 - \bar{\mu}) + (\mu_2 - \bar{\mu}) + (\mu_3 - \bar{\mu}) + \cdots + (\mu_g - \bar{\mu})$$
$$= (\mu_1 + \mu_2 + \mu_3 + \mu_g) - (\bar{\mu} + \bar{\mu} + \bar{\mu} + \cdots + \bar{\mu}) = g * \bar{\mu} - g * \bar{\mu}$$
$$= 0$$

Άρα η πρώτη διαφορά $\mu_1 - \bar{\mu}$ μπορει να γραφτει ως γραμμικος συνδιαμος των τελευταιωξν g-1 διαφορων. Οι γραμμικοι συνδιασμοι των g διανυσματων στην καθοριζουν ένα υπερεπιπεδο με διασταση $q \leq g-1$. Αν παρουμε οποιοδηποτε διανυσμα e κάθετο σε κάθε $\mu_i - \bar{\mu}$, τοτε το υπερεπιπεδο δινει:

$$B_o * e = \sum_{i=1}^{g} (\mu_i - \bar{\mu}) * (\mu_i - \bar{\mu})' * e = \sum_{i=1}^{g} (\mu_i - \bar{\mu}) * 0 = 0$$

Επομένως $\Sigma^{-1} * B_o * e = 0 * e$

Επομένως υπάρχουν p-q ορθογώνια ιδιοδιανύσματα που αντιστοιχούν στην μηδενική ιδιοτιμή.

Αυτό σημαίνει ότι υπάρχουν q ή λιγότερες μη μηδενικές ιδιοτιμές. Επειδή πάντοτε αληθεύει ότι $q \leq g-1$, ο αριθμός των μη μηδενικών ιδιοτιμών s πρέπει να ικανοποιεί την

$s \leq \min\{p,g-1\}$.

Επομένως δεν υπάρχει απώλεια πληροφορίας, αν σχεδιάσουμε τους διαχωριστές σε 2 διαστάσεις , αν οι ακόλουθες συνθήκες ισχύουν:

Αριθμός μεταβλητών	Αριθμός πληθυσμών	Μέγιστος αριθμός διαχωριστών
P	g=2	1
P	g=3	2
p=2	g	2

Παρατηρήσεις

1) Όταν οι πληθυσμιακοί παράμετροι είναι άγνωστοι, αυτοί εκτιμώνται από τις δειγματικές τους ποσότητες. Ο κανόνας του Fisher γίνεται:
   το x ανήκει στον π_καν
   $\sum_{j=1}^{r}(\hat{y}_j - \bar{y}_{kj})^2 = \sum_{j=1}^{r}[\hat{l}'_j(x - \bar{x}_i)]^2$ για κάθε i≠k, οπου το $\hat{l}'_j$ οριζεται από την (1), r≤s
2) Ο κανόνας του Fisher στηρίζεται στις ευκλείδειες αποστάσεις. Συγκρίνει τις τετραγωνικές αποστάσεις των παρατηρήσεων που θέλουμε να κατατάξουμε από τον μέσο όρο της κάθε ομάδας και τοποθετεί την κάθε παρατήρηση σε εκείνο πληθυσμό που η απόσταση είναι μικρότερη.

Κεφάλαιο 4

Λογιστική παλινδρόμηση

4.1 Εισαγωγή

Η Λογιστική παλινδρόμηση (logistic regression) αποτελεί μέρος της κατηγορίας των στατιστικών μοντέλων που ονομάζεται γενικευμένα γραμμικά μοντέλα. Αυτή η ευρεία κατηγορία των μοντέλων περιλαμβάνει την απλή παλινδρόμηση, την ANOVA, καθώς και πολυμεταβλητές στατιστικές όπως ANCOVA και loglinear παλινδρόμηση.

Η λογιστική παλινδρόμηση επιτρέπει σε κάποιον να προβλέψει ένα διακριτό αποτέλεσμα, από ένα σύνολο μεταβλητών που μπορεί να είναι συνεχής, διακριτές, διχοτομικές, ή ένα μείγμα αυτών. γενικότερη εξαρτημένη αποκρινόμενη μεταβλητή είναι δίτιμη , όπως η παρουσία ή απουσία ή επιτυχίας / αποτυχίας. Η διαφορά με την διαχωριστική ανάλυση έγκειται στο γεγονός ότι για να διαχωρίσουμε δυο ή περισσότερους πληθυσμούς με την μέθοδο της διαχωριστικής ανάλυσης πρέπει οι μεταβλητές που θα χρησιμοποιήσουμε να είναι ποσοτικές, δηλαδή να παίρνουν τιμές πραγματικούς αριθμούς. Όταν όμως υπάρχουν ανεξάρτητες μεταβλητές στο μοντέλο που είναι κατηγορικές, τότε απαιτείται η χρήση της λογιστικής παλινδρόμησης για το διαχωρισμό των πληθυσμών.

4.2 Το μοντέλο

Έστω ένα απλό γραμμικό μοντέλο παλινδρόμησης:
$Y_i = \beta_o + \beta_1 * x_i + \varepsilon_i$, όπου Y_i είναι μια δυαδική μεταβλητή, δηλαδή παίρνει τις τιμές 0 ή 1. Εφόσον, η Y_i είναι μια δυαδική μεταβλητή ακολουθεί την Bernoulli κατανομή.

Παρατήρηση:

$$Y \sim \text{Bernoulli}$$

$$P(Y = 1) = p_i \; και \;\; P(Y = 0) = q_i = 1 - p_i, \; p \; \epsilon \; (0,1)$$

$$\text{Με} \;\; E(Y) = p_i \; και \; V(Y) = p_i * (1 - p_i)$$

Επειδή $E\{\varepsilon_i\}=0$, ο αποκρινόμενος εξαρτημένος μέσος είναι:

$$E\{Y_i\} = \beta_o + \beta_1 * x_i \qquad (I)$$

Όμως αν πάρουμε τον ορισμό της μέσης τιμής βρίσκουμε ότι:

$$E\{Y_i\} = 1 * (p_i) + 0 * (1 - p_i) = p_i \qquad (II)$$

Από (I), (II) έχουμε:

$$E\{Y_i\} = \beta_o + \beta_1 * x_i = p_i \qquad (1)$$

Άρα ο αποκρινόμενος (response variable) μέσος $E\{Y_i\}$ δηλώνει την πιθανότητα ότι η Yi παίρνει την τιμή 1 , όταν η προβλέπουσα μεταβλητή (predictor variable) $x = x_i$.

4.3 Ειδικά προβλήματα όταν η αποκρινόμενη μεταβλητή είναι δυαδική

Ειδικά προβλήματα δημιουργούνται, όταν η αποκρινόμενη μεταβλητή είναι μια δυαδική μεταβλητή. Τα 3 πιο σημαντικά είναι:

- *Μη κανονικοί όροι σφαλμάτων*

Είναι: $Y_i = \beta_o + \beta_1 * x_i + \varepsilon_i = p_i => \varepsilon_i = Y_i - \beta_o - \beta_1 * x_i$

Αν $Y_i = 0 => -\beta_o - \beta_1 * x_i$

Επομένως τα κατάλοιπα δεν ακολουθούν την κανονική κατανομή.

- *Άνισες διασπορές σφαλμάτων*

Άλλο ένα πρόβλημα με τα κατάλοιπα είναι οι άνισες διασπορές όταν η αποκρινόμενη μεταβλητή είναι μια δυαδική μεταβλητή.

Απόδειξη:

$$\sigma^2\{Y_i\} = E\{(Y_i - E\{Y_i\})^2\} = (1 - p_i)^2 * p_i + (0 - p_i)^2 * (1 - p_i)$$
$$= (1 - p_i)^2 * p_i + {p_i}^2 * (1 - p_i) = p_i * (1 - p_i) * [(1 - p_i) + p_i]$$
$$= p_i * (1 - p_i) = E\{Y_i\} * [1 - E\{Y_i\}]$$

Ισχύει ότι $\varepsilon_i = Y_i - E\{Y_i\} = Y_i - p_i$

Επειδή $p_i =$ σταθερο συνεπαγεται ότι $\sigma^2\{Y_i\} = \sigma^2\{\varepsilon_i\}$

Οπότε:

$$\sigma^2\{\varepsilon_i\} = E\{Y_i\} * [1 - E\{Y_i\}] => \sigma^2\{\varepsilon_i\} = (\beta_o + \beta_1 * x_i) * (1 - \beta_o - \beta_1 * x_i)$$

Επομένως οι διασπορές των σφαλμάτων είναι συνάρτηση του x_i.

Για τον αποκρινόμενο μέσο ισχύει ο περιορισμός: $0 \leq E(Y) = p \leq 1$

4.4 απλή λογιστική παλινδρόμηση

Επειδή οι Y_i μεταβλητές είναι δυαδικές, παίρνουν μόνο την τιμή 1 ή μόνο την τιμή 0, θα χρησιμοποιήσουμε την μέθοδο μέγιστης πιθανοφάνειας για να εκτιμήσουμε τις παραμέτρους της λογιστικής αποκρινόμενης συνάρτησης.

4.4.1 Απλό λογιστικό μοντέλο παλινδρόμησης

Αν Y_i με i=1, 2, 3, …,n είναι n ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την κατανομή Bernoulli με αναμενόμενες τιμές $E\{Y_i\} = p_i$, τότε το λογιστικό μοντέλο παλινδρόμησης έχει την μορφή:

$$E\{Y_i\} = p_i = \frac{e^{(\beta_0+\beta_1 x_i)}}{1+e^{(\beta_0+\beta_1 x_i)}} \quad (2)$$

Οι x_i παρατηρήσεις υποτίθεται ότι είναι γνωστές σταθερές. Στην περίπτωση που οι x_i παρατηρήσεις είναι τυχαίες το $E\{Y_i\}$ θα είναι ο δεσμευμένος μέσος δοθέντος του x_i .

4.4.2 συνάρτηση πιθανοφάνειας

Επειδή κάθε Y_i παρατήρηση είναι μια τυχαία μεταβλητή Bernoulli, όπου : $P(Y_i = 1) = p_i$ και $P(Y_i = 0) = 1 - p_i$.
Οι συναρτήσεις πυκνότητας πιθανοτήτων θα είναι :

$$f_i(Y_i) = p_i^{Y_i} * (1 - p_i)^{1-Y_i} \; Y_i = 0,1 \text{ και } i=1,\dots,n \quad (3)$$

Παρατηρούμε ότι $f_i(1) = p_i$ και $f_i(0) = 1 - p_i$. Έτσι, $f_i(Y_i)$ απλά παριστάνει την πιθανότητα ότι $Y_i = 1$ ή $Y_i = 0$.

Αφού οι Y_i παρατηρήσεις είναι ανεξάρτητες, η από κοινού τους συνάρτηση πιθανότητας είναι:

$$g(Y_1, \dots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} p_i^{Y_i} * (1 - p_i)^{1-Y_i} \quad (4)$$

δουλεύοντας με τον λογάριθμο της από κοινού τους συνάρτησης πιθανότητας. Η (4) γίνεται: $\ln\big(g(Y_1,\dots,Y_n)\big) = \ln\big(\prod_{i=1}^{n} p_i^{Y_i} * (1-p_i)^{1-Y_i}\big)$.

Επομένως:

$$\ln\big(g(Y_1,\dots,Y_n)\big)\sum_{i=1}^{n} Y_i * \ln\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n}\ln(1-p_i) \quad (5)$$

Από την (1) και την (2) έχουμε: $\ln\left(\frac{p_i}{1-p_i}\right) = \beta_o + \beta_1 * x_i$ (6)

$$\text{και } 1-p_i = 1-\frac{e^{(\beta_o+\beta_1 x_i)}}{1+e^{(\beta_o+\beta_1 x_i)}} => 1-p_i = \frac{1+e^{(\beta_o+\beta_1 x_i)}}{1+e^{(\beta_o+\beta_1 x_i)}} - \frac{e^{(\beta_o+\beta_1 x_i)}}{1+e^{(\beta_o+\beta_1 x_i)}} =>$$

$$1-p_i = \frac{1}{1+e^{(\beta_o+\beta_1 x_i)}} => 1-p_i = (1+e^{(\beta_o+\beta_1 x_i)})^{-1} \quad (7)$$

Η (5) λόγω της (6) και της (7) γίνεται:

$$\ln L(\beta_o,\beta_1)\sum_{i=1}^{n} Y_i * (\beta_o+\beta_1 * x_i) - \sum_{i=1}^{n}\ln\big(1+e^{(\beta_o+\beta_1 x_i)}\big) \quad (8)$$

όπου το $L(\beta_o,\beta_1)$ αντικαθιστά το $g(Y_1,\dots,Y_n)$ για να δείξει σαφώς ότι τώρα απεικονίζουμε αυτή τη συνάρτηση ως συνάρτηση πιθανοοφάνειας των εκτιμώμενων παραμέτρων, δοθέντος των δειγματικών παρατηρήσεων .

Εκτίμηση μέγιστης πιθανοφάνειας

Οι εκτιμήσεις μέγιστης πιθανοφάνειας των β_oκαι β_1 στο απλό λογιστικό μοντέλο παλινδρόμησης είναι εκείνες οι τιμές των β_o και β_1οι οποίες μεγιστοποιούν την λογαριθμική συνάρτηση πιθανοφάνειας της (8). Δεν υπάρχουν λύσεις κλειστού τύπου, αντιθέτως με την χρήση αριθμητικών υπολογιστικών διαδικασιών βρίσκουμε τους εκτιμητές μέγιστης πιθανοφάνειας, τους οποίους τους συμβολίζουμε με b_oκαι b_1.

Η προσαρμοσμένη αποκρινόμενη συνάρτηση για την i παρατήρηση είναι:

$\hat{p}_i = \frac{e^{(b_o+b_1 x_i)}}{1+e^{(b_o+b_1 x_i)}}$ και για οποιαδήποτε παρατήρηση:

$$\hat{p} = \frac{e^{(b_o+b_1 x)}}{1+e^{(b_o+b_1 x)}} \quad (9)$$

αν χρησιμοποιήσουμε τον λογαριθμικό μετασχηματισμό της (1), η προσαρμοσμένη αποκρινόμενη συνάρτηση θα έχει την μορφή:

$$\hat{p}' = b_o + b_1 x \quad (10)$$

Όπου: $\hat{p}' = \ln\left(\frac{\hat{p}'}{1-\hat{p}'}\right)$ (11)

Η σχέση (11) λέγεται προσαρμοσμένη λογαριθμική αποκρινόμενη συνάρτηση. Μόλις η προσαρμοσμένη λογιστική αποκρινόμενη συνάρτηση αποκτηθεί, τα επόμενα βήματα είναι να εξετάσουμε την καταλληλότητα της προσαρμοσμένης αποκρινόμενης συνάρτησης, και εάν η προσαρμογή είναι καλή, να κάνουμε διάφορες εξαγωγές συμπερασμάτων και προβλέψεις. Θα αναβάλουμε την συζήτηση του πώς να εξετάσουμε την καταλληλότητα της προσαρμοσμένης αποκρινόμενης συνάρτησης και πως να κάνουμε εξαγωγές συμπερασμάτων και προβλέψεις μέχρις ότου μελετήσουμε το πολλαπλό λογιστικό μοντέλο παλινδρόμησης με έναν αριθμό από προβλέπουσες μεταβλητές.

4.4.3 Ερμηνεία του b_1

Ο εκτιμώμενος συντελεστής παλινδρόμησης b_1στην προσαρμοσμένη λογιστική αποκρινόμενη συνάρτηση (9) δεν δηλώνει την κλίση, όπως συμβαίνει στο γραμμικό μοντέλο παλινδρόμησης. Αυτό συμβαίνει γιατί η αύξηση μιας μονάδας στο x διαφέρει στην λογιστική παλινδρόμηση, σύμφωνα με την αρχική τοποθέτηση του x στην κλίμακα. Η ερμηνεία του b_1 βρίσκεται στην ιδιότητα ότι ο εκτιμώμενος λόγος πιθανοτήτων $\frac{\hat{p}}{1-\hat{p}}$ πολλαπλασιάζεται με e^{b_1} για κάθε μονάδα αύξησης του x .

Για να καταλάβουμε αυτό, παίρνουμε την τιμή της προσαρμοσμένης λογαριθμικής αποκρινόμενης συνάρτησης της (10) για $x = x_j$:

$$\hat{p}'(x_j + 1) = b_o + b_1(x_j + 1)$$

Τότε :

$$\hat{p}'(x_j + 1) - \hat{p}'(x_j) = b_o + b_1(x_j + 1) - b_o - b_1 x_j =>$$

$$\hat{p}'(x_j + 1) - \hat{p}'(x_j) = b_o + b_1 x_j + b_1 - b_o - b_1 x_j \;\; =>$$

$$\hat{p}'(x_j + 1) - \hat{p}'(x_j) = b_1$$

Όμως, από την (11) το $\hat{p}'(x_j)$ είναι ο λογάριθμος του εκτιμώμενου λόγου πιθανοτήτων όταν $x = x_j$ και θα το συμβολίζουμε $\hat{p}'(x_j) = \ln(\text{odds}_1)$. Όμοια, $\hat{p}'(x_j + 1) = \ln(\text{odds}_2)$. Άρα, η παραπάνω σχέση γράφεται:

$\ln(\text{odds}_2)+\ln(\text{odds}_1) = \ln\frac{\text{odds}_2}{\text{odds}_1} = b_1 => \frac{\text{odds}_2}{\text{odds}_1} = e^{b_1} => \widehat{OR} = e^{b_1}$ οπου το $\widehat{OR}$ είναι ο λόγος των εκτιμώμενων πιθανοτήτων (odds ratio).

4.4.4 Επαναλαμβανόμενες παρατηρήσεις

Σε πάρα πολλές εφαρμογές χρειάζεται να αποκτήσουμε επαναλαμβανόμενες παρατηρήσεις για συγκεκριμένες στάθμες της προβλέπουσας μεταβλητής. Για παράδειγμα, θέλουμε να κάνουμε μια μελέτη στο αν οι πελάτες μιας εταιρίας θα αγόραζαν ένα προϊόν και σε ποια τιμή. Πέντε τιμές μελετήθηκαν και 200 άτομα εκλέχθηκαν για κάθε τιμή. Η αποκρινόμενη μεταβλητή είναι δυαδική (θα αγόραζε, δεν θα αγόραζε το προϊόν), και η προβλέπουσα μεταβλητή είναι η τιμή που έχει πέντε στάθμες.

Έστω, $x_1, x_2, x_3, \ldots x_c$ οι c στάθμες της προβλέπουσας μεταβλητής x . Ο αριθμός των παρατηρήσεων στην στάθμη xj δηλώνεται με nj (j=1,2,3,…c) και το j Rj δηλώνει τον αριθμό των 1 στην στάθμη xj . Η αναλογία των μονάδων στην στάθμη x_j δηλώνεται με π_j και είναι $\pi_j = \frac{R_j}{n_j}$.

Η λογαριθμική συνάρτηση πιθανοφάνειας της (8) γίνεται:
$\ln L(\beta_o, \beta_1) \sum_{j=1}^{c}\{\ln\binom{n_j}{R_j} + R_j * (\beta_o + \beta_1 * x_j) - n_j * \ln\left(1 + e^{(\beta_o+\beta_1 x_j)}\right)\}$
Όπου $\binom{n_j}{R_j}$ ειναη η διαταξη των n_j ανα R_j.

4.5 Πολλαπλή λογιστική παλινδρόμηση

Σε πάρα πολλές εφαρμογές η αποκρινόμενη μεταβλητή εξαρτάται από περισσότερες από μια προβλέπουσες μεταβλητές. Σε αυτή την περίπτωση χρησιμοποιούμε την πολλαπλή λογιστική παλινδρόμηση.

4.5.1 Πολλαπλό λογιστικό μοντέλο παλινδρόμησης

Έστω τα διανύσματα β, X X_i διάστασης px1

$$\beta = \begin{bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \quad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_{p-1} \end{bmatrix}, \quad X_i = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{i,p-1} \end{bmatrix}$$

Τότε

$$\beta' * X = \beta_o + \beta_1 * x_1 + \cdots + \beta_{p-1} * x_{p-1}$$
$$\beta' * X_i = \beta_o + \beta_1 * x_{i1} + \cdots + \beta_{p-1} * x_{i,p-1}$$

Αν Y_i είναι ανεξάρτητες τυχαίες μεταβλητές Bernoulli με αναμενόμενες τιμές $E\{Y_i\} = p_i$ τότε το πολλαπλό λογιστικό μοντέλο παλινδρόμησης έχει τη μορφή: $E\{Y_i\} = \frac{e^{(\beta' * x_i)}}{1+e^{(\beta' * x_i)}}$ (13)

Οι X παρατηρήσεις υποτίθεται ότι είναι γνωστές σταθερές. Στην περίπτωση που οι X παρατηρήσεις είναι τυχαίες το $E\{Y_i\}$ θα είναι ο δεσμευμένος μέσος δοθέντος του $x_{i1}, x_{i2}, x_{i3}, \dots, x_{i,p-1}$

Παρατηρήσεις

Όμοια, με την απλή λογιστική αποκρινόμενη συνάρτηση, η πολλαπλή λογιστική συνάρτηση είναι μονοτονική, σχηματίζει μια καμπύλη με μορφή s ή αντίστροφο s σε σχέση με το β΄*X , και είναι σχεδόν γραμμική όταν το p παίρνει τιμές από 0.2 έως 0.8.

Οι X μεταβλητές ίσως είναι όλες διαφορετικές προβλέπουσες μεταβλητές, ή ίσως αναπαριστούν καμπυλώσεις ή και όρους αλληλεπιδράσεων. Επίσης, οι προβλέπουσες μεταβλητές μπορεί να είναι ποσοτικές, αλλά και ποιοτικές . Στην περίπτωση που είναι ποιοτικές παριστάνονται από δυαδικές μεταβλητές.

(Αν το λογιστικό μοντέλο περιέχει μόνο ποιοτικές μεταβλητές αυτό είναι γνωστό ως λογαριθμικό μοντέλο παλινδρόμησης, log- linear model)

Όμοια με το απλό λογιστικό μοντέλο παλινδρόμησης αποδεικνύεται ότι: $E\{Y_i\} = (1 + e^{-\beta' * X})^{-1}$

και αν κάνουμε τον μετασχηματισμό $p' = \ln \frac{p}{1-p}$

βρίσκουμε ότι: $p' = \beta' * X$.

4.5.2 Προσαρμογή του μοντέλου

Θα χρησιμοποιήσουμε την μέθοδο μέγιστης πιθανοφάνειας για να εκτιμήσουμε τις παραμέτρους της πολλαπλής λογιστικής αποκρινόμενης συνάρτησης. Η λογαριθμική συνάρτηση πιθανοφάνειας θα είναι:

$$\ln L(\beta) \sum_{i=1}^{n} Y_i * (\beta' * X_i) - \sum_{i=1}^{n} \ln\left(1 + e^{(\beta' * X_\iota)}\right) \quad (17)$$

Οι τιμές που μεγιστοποιούν την (17) δηλώνονται ως b , όπου: $b = \begin{bmatrix} b_o \\ b_1 \\ b_2 \\ \vdots \\ b_{p-1} \end{bmatrix}$

Η λογιστική αποκρινόμενη συνάρτηση έχει την μορφή :

$$\hat{p}_i = \frac{e^{(b' * X_i)}}{1+e^{(b' * X_i)}} = (1 + e^{-b' * X})^{-1}$$

όπου:

$$b' * X = b_o + b_1 * x_1 + \cdots + b_{p-1} * x_{p-1}$$

$$b' * X_i = b_o + b_1 * x_{i1} + \cdots + b_{p-1} * x_{i,p-1}$$

4.5.3 Η ερμηνεία των συντελεστών

Όταν το πολλαπλό μοντέλο παλινδρόμησης είναι ένα πρώτης τάξεως μοντέλο η ερμηνεία των εκτιμώμενων συντελεστών είναι ίδια με αυτή της απλής λογιστικής αποκρινόμενης συνάρτησης. Το e^{b_k}ξανά είναι ο εκτιμώμενος λόγος πιθανοτήτων για την προβλέπουσα μεταβλητή x_k , υπό την προϋπόθεση ότι οι άλλες προβλέπουσες μεταβλητές είναι σταθερές. Οι στάθμες στις οποίες οι προβλέπουσες μεταβλητές λαμβάνονται ως σταθερές δεν μας απασχολεί στο πρώτης τάξης μοντέλο. Όταν το πολλαπλό μοντέλο λογιστικής παλινδρόμησης δεν είναι ένα πρώτης τάξεως μοντέλο αλλά περιέχει τετραγωνικούς ή υψηλότερους σε βαθμούς όρους για τις προβλέπουσες μεταβλητές, ή και όρους αλληλεπιδράσεων, οι εκτιμώμενοι συντελεστές παλινδρόμησης δεν έχουν πλέον μια απλή ερμηνεία.

Κεφάλαιο 5

Διαχωρισμός ομάδων με πραγματικά δεδομένα

5.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα παρουσιάσουμε μια εφαρμογή τόσο της διαχωριστικής ανάλυσης όσο και της γραμμικής παλινδρόμησης στον τομέα της τραπεζικής και συγκεκριμένα στον τομέα του πιστωτικού κινδύνου.

Τα βήματα που θα ακολουθηθούν είναι τα εξής:

1) Η παρουσίαση του προβλήματος
2) Η παρουσίαση των δεδομένων, η ανάλυση και η εξαγωγή των αποτελεσμάτων για κάθε μια από τις δυο μεθόδους διαχωρισμού και
3) Η σύγκριση των δυο μεθόδων.

5.2 Η παρουσίαση του προβλήματος

Από το σύνολο του χαρτοφυλακίου της τράπεζας x, απομονώνουμε τις αιτήσεις που έγιναν για ένα συγκεκριμένο πιστωτικό προϊόν. Ως δειγματική περίοδο θεωρήσαμε την τριετία 2009-2011. Ο προς μελέτη πληθυσμός αποτελείται από 225899 αιτήσεις πελατών. Κάθε αίτηση έχει καταλήξει σε ένα αποτέλεσμα. Δηλαδή κάποιοι πελάτες έχουν κριθεί ικανοί να λάβουν το πιστωτικό προϊόν της τράπεζας και επομένως η αίτηση τους έχει εγκριθεί. Κάποιοι άλλοι πελάτες κρίθηκαν ότι δεν είναι ικανοί να ξεκινήσουν συνεργασία με την τράπεζα ως προς το συγκεκριμένο προϊόν, για αυτό το λόγο η αίτηση τους οδηγήθηκε σε απόρριψη.
Ο σκοπός μας είναι να κατασκευάσουμε μια σκοροκάρτα, δηλαδή ένα μοντέλο βάσει του οποίου θα αξιολογείται μια αίτηση στο μέλλον. Το μοντέλο αυτό θα είναι μια συνάρτηση $A \rightarrow B$ όπου A είναι το σύνολο των ανεξάρτητων μεταβλητών και B είναι το σύνολο της εξαρτημένης μεταβλητής. Το σύνολο A μπορεί να περιέχει ποσοτικές, ποιοτικές, διακριτές ή συνεχείς μεταβλητές. Το σύνολο B αποτελεί την εξαρτημένη μεταβλητή η οποία μπορεί να είναι συνεχής η διακριτή.

5.3 Ανάλυση-παρουσίαση δεδομένων

Τα δεδομένα που θα μελετήσουμε είναι στοιχεία αιτήσεων των πελατών. Στο συγκεκριμένο πρόβλημα κάθε πελάτης διακρίνεται από τον αριθμό αίτησης (application number) την οποία υποβάλλει σε ένα κατάστημα της τράπεζας. Τα υπόλοιπα στοιχεία της αίτησης αποτελούν τις ανεξάρτητες μεταβλητές του προβλήματος. Από αυτές, κάποιες είναι συνεχείς-ποσοτικές μεταβλητές όπως είναι το ετήσιο εισόδημα άλλες είναι διακριτές-ποσοτικές μεταβλητές και άλλες είναι ποιοτικές μεταβλητές. Τέλος υπάρχει ένα πεδίο που αποτελεί την εξαρτημένη μεταβλητή. Εδώ εξαρτημένη μεταβλητή είναι το application status η οποία παίρνει τιμές RJ=reject ή AC=accept. οι υπόλοιπες μεταβλητές είναι εξαρτημένες. Η μεταβλητή income είναι συνεχής μεταβλητή. Οι μεταβλητές age, depen, tad και tij είναι διακριτές μεταβλητές. Οι μεταβλητές

gender, area, phones, phones2, mphone, jphone, bank_ac, astatdel1, marstat1, mdep, residen, mres, occu, postj, home_phone είναι ψευδομεταβλητές.

Συμπερασματικά τα δυο σύνολα {rejected πελάτες} και {accepted πελάτες} αποτελούν τους 2 γνωστούς πληθυσμούς:

Π_1 = {πελατης | δεν εχει γινει αποδεκτη η αιτηση του} και

Π_2 = {πελατης | εχει γινει αποδεκτη η αιτηση του}

Σκοπός μας είναι βάσει της ιστορικότητας αυτής, να κατασκευάσουμε ένα μοντέλο με το οποίο θα αξιολογούμε οποιαδήποτε νέα αίτηση έρθει σε ένα κατάστημα και θα γίνεται αυτόματα δεκτή η θα απορρίπτεται.

Δεδομένα:

Στον παρακάτω πίνακα παρουσιάζονται οι μεταβλητές του προβλήματος και η ερμηνεία τους:

μεταβλητή	**ερμηνεία μεταβλητής**	**εξαρτημένη/ ανεξάρτητη**
application:	τελικο status αιτησης	εξαρτημένη
gender:	φυλο	ανεξάρτητη
age:	ηλικια	ανεξάρτητη
area:	περιοχη	Ανεξάρτητη
mphone:	υπαρξη κινητου	ανεξάρτητη
jphone:	τηλεφωνο εργασιας	ανεξάρτητη
cards:	Εχει άλλες καρτες ναι/ όχι	ανεξάρτητη
carown:	Κατοχος αυτοκινητου	ανεξάρτητη
depen:	εξαρτωμενα μελη	ανεξάρτητη
marstat1:	οικογενειακη καταταση	ανεξάρτητη
residen:	Κατοικια	ανεξάρτητη
tad:	ετη διαμονης στην υπαρχουσα κατοικια	ανεξάρτητη
tij:	ετη στην εργασια	ανεξάρτητη
occu:	Επαγγελμα	ανεξάρτητη
income:	ετησιο εισοδημα	ανεξάρτητη
postj:	υπαρξη ΤΚ εργασιας	ανεξάρτητη
home_phone:	υπαρξη τηλεφωνου κατοικιας	ανεξάρτητη

5.4 Επίλυση του προβλήματος με τη χρήση της διαχωριστικής ανάλυσης

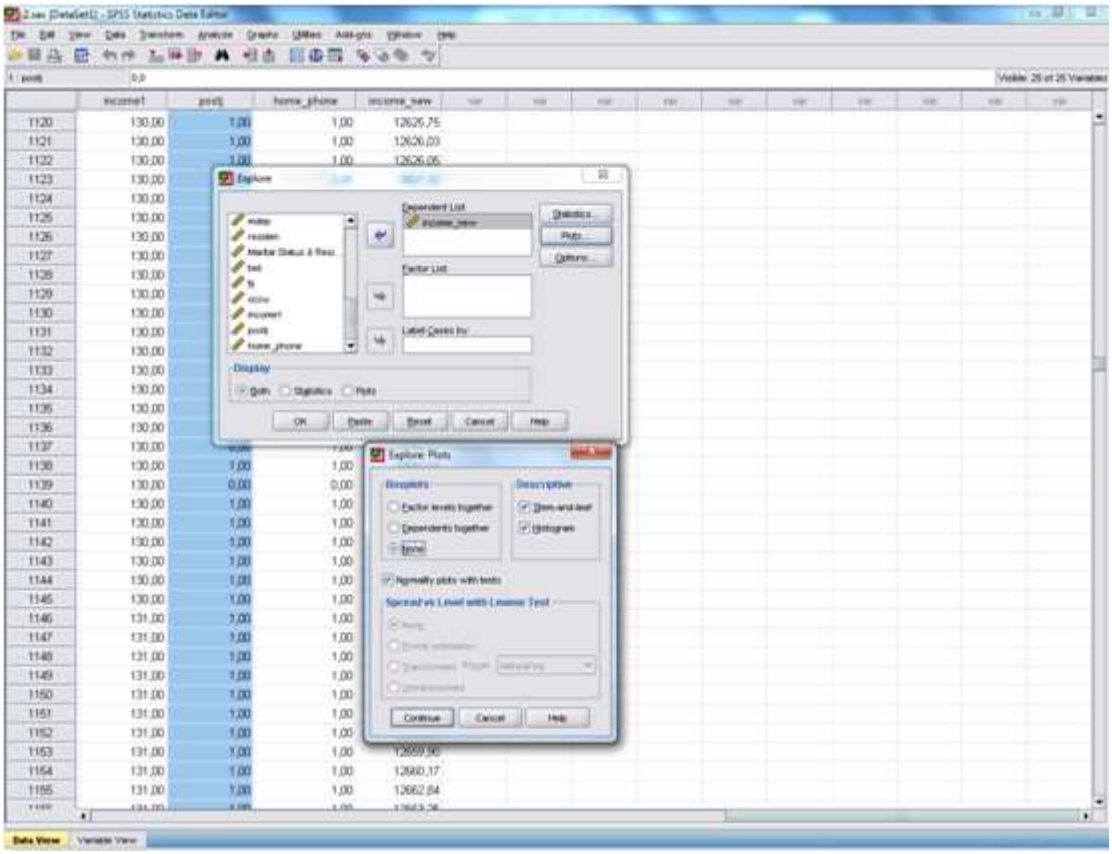
1ο στάδιο

Αρχικά και πριν κάνουμε οποιαδήποτε ανάλυση, καλό θα ήταν να ελέγξουμε την κανονικότητα των συνεχών μας μεταβλητών. Όπως έχει προαναφερθεί στην θεωρητική παρουσίαση της discriminant ananlysis είναι απαραίτητη η ισχύς της υπόθεσης της κανονικότητας των συνέχων μεταβλητών.

Analyze> Descriptive Statistics> Explore

Και τοποθετούμε τις συνεχείς μεταβλητές (εδώ μόνο η μεταβλητή income είναι συνεχής) ως εξής:

Στο πεδίο display ενεργοποιούμε την επιλογή plots και έπειτα πηγαίνουμε στο μενού plots και πραγματοποιούμε τις ακόλουθες ενέργειες Boxplot→none, Descriptive→histogram, normality, όπως φαίνεται και στην παρακάτω εικόνα:



Το syntax για να να πραγματοποιηθεί η παραπάνω εργασία είναι το ακόλουθο:

```
**********test κανονικότητας**********.
EXAMINE VARIABLES=income
  /PLOT HISTOGRAM NPPLOT
  /STATISTICS NONE
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
**************************************************************.
```

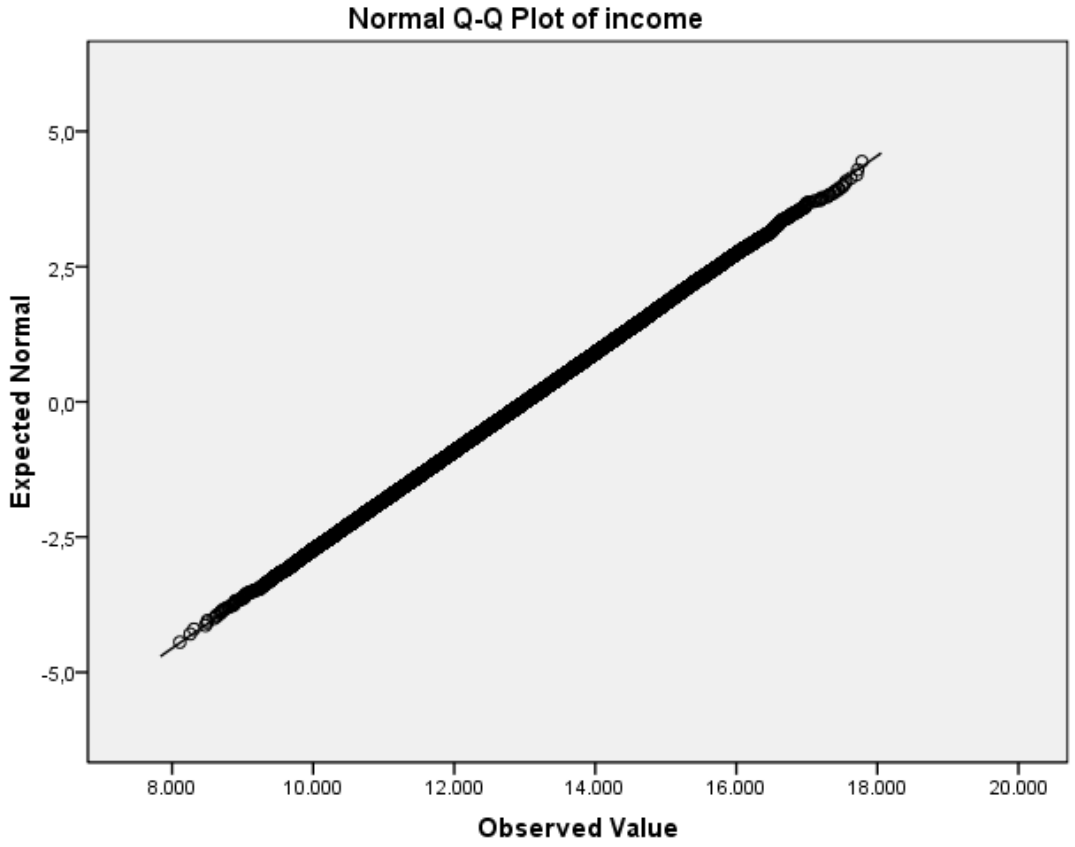
Κάνοντας τον παραπάνω έλεγχο, πρώτα λαμβάνουμε τον ακόλουθο πίνακα :

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
income	225899	100,0%	0	,0%	225899	100,0%

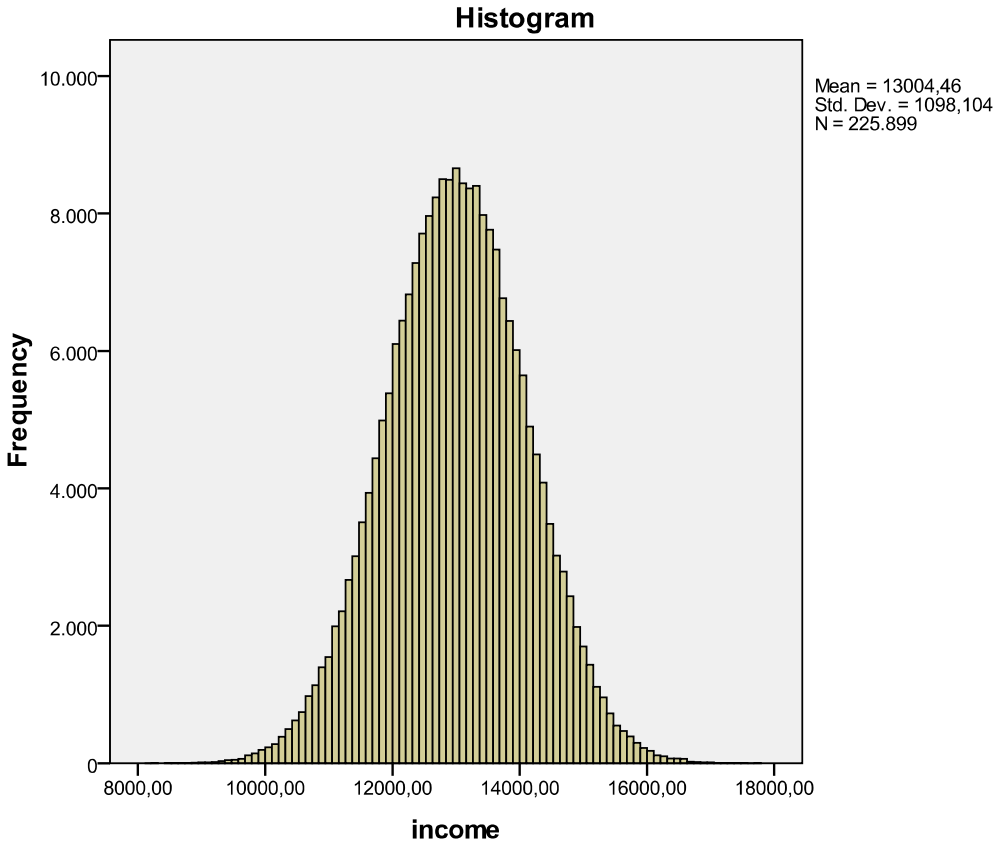
Όπως περιγράφεται, ο πληθυσμός αποτελείται από 225899 άτομα, για τα οποία δεν έχουμε κανένα missing value (αναμενόμενο, καθώς σε περίπτωση που δεν ήταν συμπληρωμένο το ετήσιο εισόδημα στην αίτηση, θα οδηγείτο σε αυτόματη απόρριψη και δεν θα την συμπεριλαμβάναμε στην αξιολόγηση).

Έπειτα λαμβάνουμε το ακόλουθο QQ-plot:



Με το test κανονικότητας φαίνεται πως η μεταβλητή income ακολουθεί την κανονική κατανομή.

Το ιστόγραμμα είναι το ακόλουθο:



Επιπλέον δίνονται και κάποια περιγραφικά μέτρα:
Όπως φαίνεται και στο ιστόγραμμα η μέση τιμή του υπό μελέτη πληθυσμού είναι mean = 13.004 και η τυπική απόκλιση stddeviation = 1.098

Η υπόθεση της κανονικότητας αποδεικνύεται ότι ισχύει και αλγεβρικά, καθώς το p-value=0,2>0,05 μέσω του test kolmogorov-smirnov

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
income	,002	225899	,200*

a. Lilliefors Significance Correction

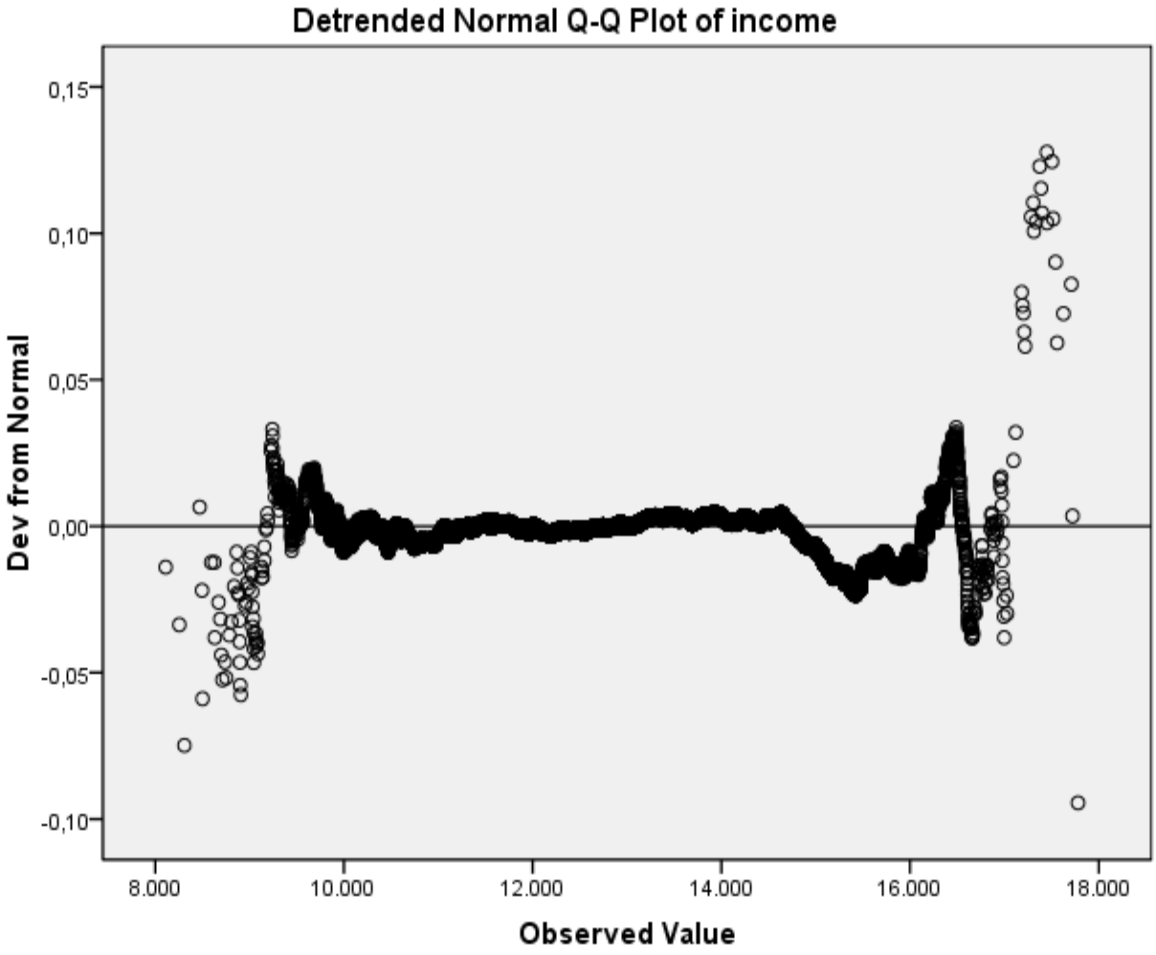
Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
income	,002	225899	,200*

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

Το detrended normal qq plot δείχνει τις διαφορές μεταξύ των παρατηρούμενες και αναμενόμενες τιμές της κανονικής κατανομής. Αν η κατανομή είναι κανονική, τα σημεία πρέπει να cluster σε μια οριζόντια γραμμή γύρω από το μηδέν.



Όπως φαίνεται από

1) το normal qq plot
2) το p-value
3) το detrended normal qq plot και
4) το ιστόγραμμα των συχνοτήτων

η υπόθεση της κανονικότητας της συνεχούς μεταβλητής income, δεν απορρίπτεται.

2ο στάδιο

Analyze→classify→Discriminant

Grouping Variable: Εδώ ζητείται η μεταβλητή που καθορίζει τις ομάδες. Εμείς τοποθετήσαμε την μεταβλητή application η οποία χωρίζει του πελάτες της τράπεζας σε άξιους να λάβουν κάποιου είδους πίστωσης από την τράπεζα (accepted) και σε μη άξιος για μια πίστωση (rejected). Επιπλέον πρέπει να ορίσουμε το εύρος τον ομάδων στην επιλογή Define Range. Εδώ ορίζουμε σαν ελάχιστη τιμή (minimum) το 0 και σαν μέγιστη τιμή (maximum) το 1.

Independents: Σε αυτή τη θέση τοποθετούμε τις ανεξάρτητες μεταβλητές με βάση τις οποίες θα γίνει η ταξινόμηση της κάθε παρατήρησης. Οι μεταβλητές πρέπει να ποσοτικές.

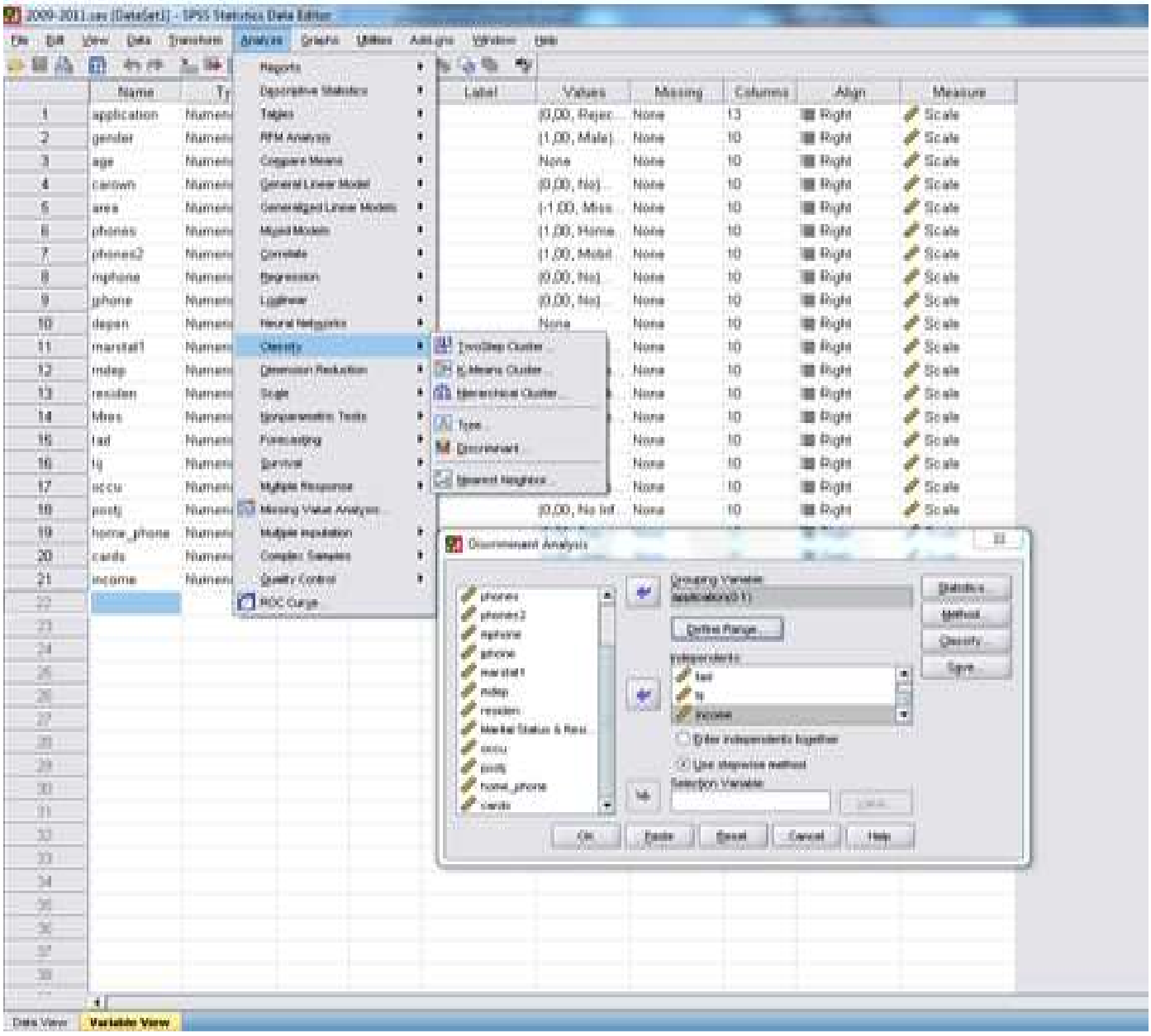
Οι επόμενες δυο επιλογές έχουν σχέση με τον τρόπο εισαγωγής των μεταβλητών στο μοντέλο:

1) χρήση όλων των ανεξάρτητων μεταβλητών (Enter Independents Together) στο μοντέλο ή εναλλακτικά

2) την χρήση κλιμακωτών μεθόδων επιλογής ανεξάρτητων μεταβλητών (Use Stepwise Method). Σε αυτή την περίπτωση, ξεκινάμε χωρίς καμία μεταβλητή στο μοντέλο και συνεχίζουμε προσθέτοντας τη μεταβλητή με τον καλύτερο δείκτη ανάλογα με την μέθοδο που διαλέγουμε στην υπό-ομάδα method ή το αντίστοιχο καλύτερο F, δεδομένου ότι ικανοποιείται το όριο

εισόδου της μεταβλητής που θέτουμε στο criteria. Παράλληλα σε κάθε βήμα ελέγχουμε αν κάποια από τις μεταβλητές που είναι ήδη στο μοντέλο πρέπει να αφαιρεθεί στις επιλογές criteria. Αν κάποια μεταβλητή πρέπει να αφαιρεθεί θα είναι αυτή με το μικρότερο F.

Η δεύτερη επιλογή είναι πολύ χρήσιμη στην πράξη γιατί εντοπίζει βήμα-βήμα τις ασήμαντες μεταβλητές για το διαχωρισμό και τις αφαιρεί από την διαχωριστική συνάρτηση. Επειδή εμείς εδώ έχουμε παραπάνω από δύο μεταβλητές καλό θα ήταν να χρησιμοποιήσουμε την επιλογή Use Stepwise Method.



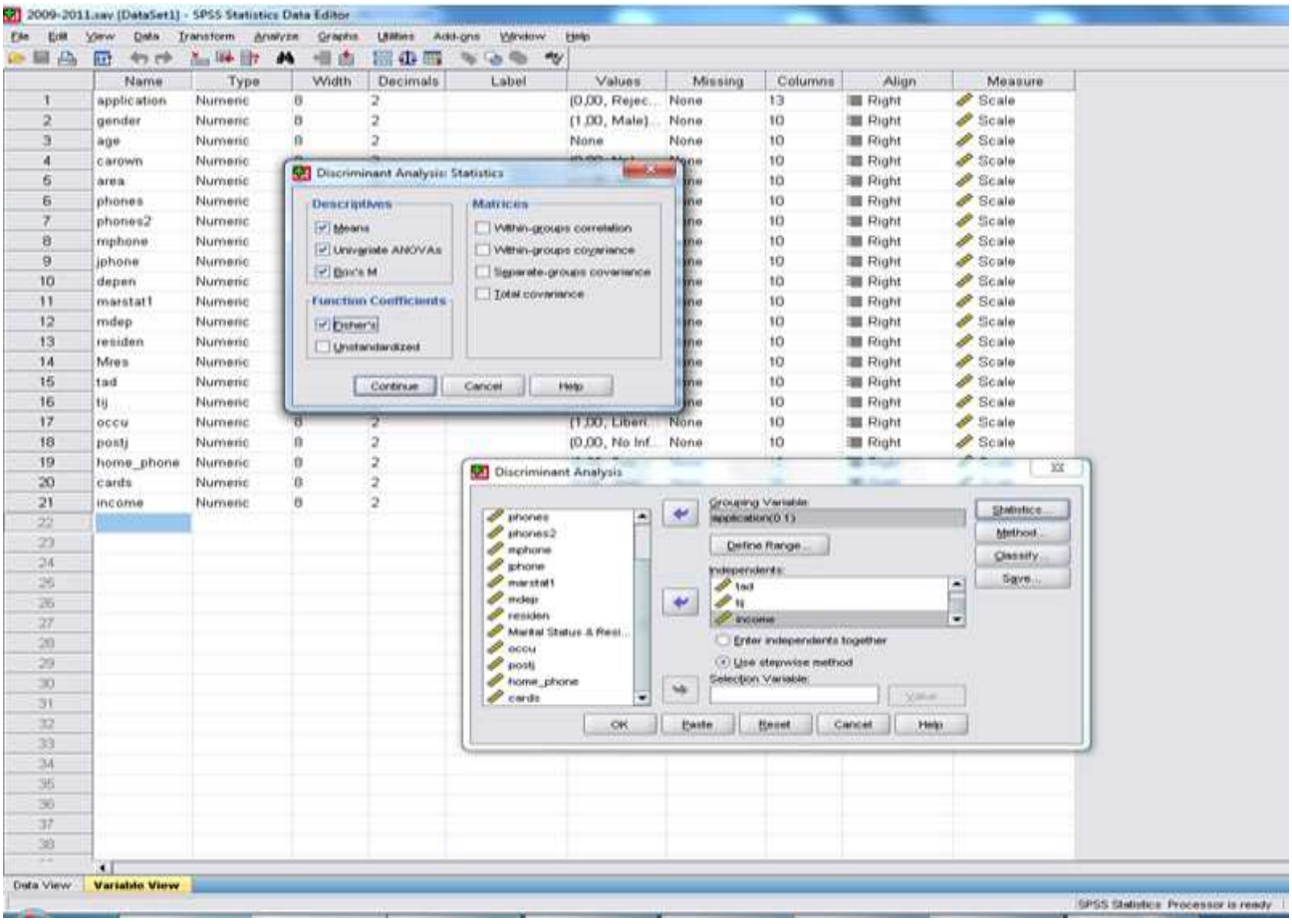
Αφού βάλουμε τις μεταβλητές όπως παραπάνω πηγαίνουμε στο μενού Statistics. Στην κλιμακωτή μέθοδο επιλογής μεταβλητών (Stepwise method ελέγχουμε αν κάποια από τις μεταβλητές που είναι ήδη στο μοντέλο πρέπει να αφαιρεθεί στις επιλογές criteria. Αν κάποια μεταβλητή πρέπει να αφαιρεθεί θα είναι αυτή με το μικρότερο F.

Υπομενού statistics:

Το υπομενού statistics μας είναι η επόμενη επιλογή μας, η οποία διακρίνεται σε 3 μέρη

1) Descriptives
2) Function coefficients
3) Matrices

Τσεκάρουμε τις παρακάτω επιλογές για την κάθε μια από τις 3 προαναφερθείσες κατηγορίες:



,Κατηγορία Descriptives:

Επιλέγοντας το «Means» λαμβάνουμε μέσες τιμές, επιλέγοντας το «Univariate ANOVAs» θα γίνει ανάλυση διακύμανσης κατά ένα παράγοντα, επιλέγοντας «Box's M» διεξάγουμε έλεγχο ισότητας πινάκων διακύμανσης συνδιακύμανσης. Συγκεκριμένα πραγματοποιεί τον έλεγχο:
$H_0 = \Sigma_1 = \Sigma_2 = \Sigma_3 = ... = \Sigma_K$

Αν τα P-value>0.05 τότε δεν απορρίπτεται η μηδενική υπόθεση. Εμείς θέλουμε να μην απορρίπτεται. Αξίζει να αναφερθεί ότι ο έλεγχος αυτός είναι ευαίσθητος στην κανονικότητα.
Βάσει των παραπάνω επιλογών λαμβάνουμε στο output τον ακόλουθο πίνακα, ο οποίος μας δείχνει τη μέση τιμή της κάθε μεταβλητής:

Group statistics

application		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
Rejected	age	44,2848	82,4591	111085	111085
	depen	0,4429	0,91491	111085	111085
	income	12768,255	1063,7069	111085	111085
	tad	11,952	12,30375	111085	111085
	tij	6,3009	7,94334	111085	111085
Accepted	age	44,48	53,64407	114811	114811
	depen	0,4554	0,88624	114811	114811
	income	13232,98	1082,4079	114811	114811
	tad	13,1683	12,43624	114811	114811
	tij	7,4695	8,80352	114811	114811
Total	age	44,384	69,32704	225896	225896
	depen	0,4493	0,90048	225896	225896
	income	13004,45	1098,109	225896	225896
	tad	12,5702	12,38617	225896	225896
	tij	6,8948	8,41185	225896	225896

Κάνοντας την ανάλυση διακύμανσης παίρνουμε τον ακόλουθο πίνακα:

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
age	1,000	,447	1	225894	,504
depen	1,000	10,890	1	225894	,001
income	,955	10585,694	1	225894	,000
tad	,998	545,692	1	225894	,000
tij	,995	1094,936	1	225894	,000

Παρατηρούμε πως για τη μεταβλητή age οι μέσοι για τα δύο επίπεδα της μεταβλητής application είναι ίδιοι αφού το p-value είναι μεγαλύτερο από το 0.05 επίπεδο σημαντικότητας που ορίσαμε στην αρχή της ανάλυσής μας, οπότε και δεν απορρίπτεται η μηδενική υπόθεση ότι οι μέσοι των δύο επιπέδων της μεταβλητής application για την εν λόγω μεταβλητή είναι ίσοι. Η μεταβλητή age πιθανότατα θα εξαιρεθεί από το τελικό μοντέλο.

Το Box's M το οποίο όπως προαναφέρθηκε, κάνει έλεγχο ισότητας των συνδιακυμάνσεων, δίνει τον παρακάτω πίνακα:

Test Results

Box's M		1601,743
F	Approx.	160,171
	df1	10
	df2	2,434E11
	Sig.	,000

Test Results

Box's M		1601,743
F	Approx.	160,171
	df1	10
	df2	2,434E11
	Sig.	,000

Tests null hypothesis of equal population covariance matrices.

Από τον παραπάνω πίνακα παρατηρούμε ότι το p-value=0<0,05. Δηλαδή δεν ισχύει η υπόθεση της ισότητας των συνδιακυμάνσεων. Γενικά ο έλεγχος και η ισότητα των συνδιακυμάνσεων έχει έμμεση σχέση με την κανονικότητα των μεταβλητών. Πάραυτα, επειδή τα δεδομένα προέρχονται από πραγματικές παρατηρήσεις ήταν πολύ πιθανό να μην ίσχυε η υπόθεση της κανονικότητας των συνδιακυμάνσεων και επομένως ο έλεγχος να μην αποδεχόταν τη μηδενική υπόθεση.

Εισαγωγή μεταβλητών στο μοντέλο

Η εισαγωγή μεταβλητών στο μοντέλο γίνεται με την μέθοδο stepwise
Βάσει της ανωτέρω μεθόδου εισάγονται τελικά οι παρακάτω μεταβλητές:

Variables Entered/Removed[a,b,c,d]

Step	Entered	Wilks' Lambda							
		Statistic	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	income	,955	1	1	225894,000	10585,694	1	225894,000	,000
2	tad	,954	2	1	225894,000	5481,245	2	225893,000	,000
3	depen	,954	3	1	225894,000	3669,476	3	225892,000	,000
4	tij	,953	4	1	225894,000	2757,273	4	225891,000	,000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

a. Maximum number of steps is 10.

b. Minimum partial F to enter is 3.84.

c. Maximum partial F to remove is 2.71.

d. F level, tolerance, or VIN insufficient for further computation.

Η εισαγωγή των μεταβλητών πραγματοποιείται σε 4 βήματα τα οποία παραθέτονται στον παρακάτω πίνακα:

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	income	1,000	10585,694	
2	income	,999	10391,697	,998
	tad	,999	359,973	,955
3	income	,991	10432,519	,998
	tad	,997	370,440	,955
	depen	,990	43,859	,954
4	income	,931	9541,685	,994
	tad	,970	333,128	,955
	depen	,960	53,238	,954
	tij	,878	19,749	,954

Οι μεταβλητές που σε κάθε βήμα δεν βρίσκονται στο μοντέλο φαίνονται στον ακόλουθο:

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0	age	1,000	1,000	,447	1,000
	depen	1,000	1,000	10,890	1,000
	income	1,000	1,000	10585,694	,955
	tad	1,000	1,000	545,692	,998
	tij	1,000	1,000	1094,936	,995
1	age	1,000	1,000	5,719	,955
	depen	,992	,992	33,406	,955
	tad	,999	,999	359,973	,954
	tij	,932	,932	40,857	,955

2	age	,999	,998	3,413	,954
	depen	,990	,990	43,859	,954
	tij	,905	,905	10,371	,954
3	age	,999	,990	3,774	,954
	tij	,878	,878	19,749	,953
4	age	,998	,877	3,375	,953

Όπως φαίνεται τελικά, η ηλικία δεν έχει καθοριστικό ρόλο στο διαχωρισμό των πελατών σε άξιους να λάβουν το προϊόν και μη.

Κατηγορία function coeffs

Σε αυτό το επίπεδο ορίζεται ο υπολογισμός για την εμφάνιση των συντελεστών της διαχωριστικής ανάλυσης (Function Coefficients).
Για κάθε ομάδα υπολογίζουμε ένα σκορ με βάση κάποια συνάρτηση. Στην περίπτωσή μας οι συναρτήσεις είναι γραμμικές ως προς τις ανεξάρτητες. Η επιλογή Fisher's υπολογίζει συντελεστές των γραμμικών συναρτήσεων των σκορ με την μέθοδο του Fisher. Τα αποτελέσματα που προέρχονται από τις παραπάνω επιλογές είναι τα ακόλουθα:

Classification Function Coefficients

	application	
	Rejected	Accepted
depen	-,174	-,209
income	,012	,012
tad	,077	,083
tij	-,316	-,314
(Constant)	-74,857	-80,117

Fisher's linear discriminant functions

Τώρα το μοντέλο για την πρώτη κατηγοριοποίηση θα είναι:

$w_1 = -0{,}174 * x_1 + 0{,}12 * x_2 + 0{,}77 * x_3 + (-0{,}316) * x_4 + (-74{,}857)$

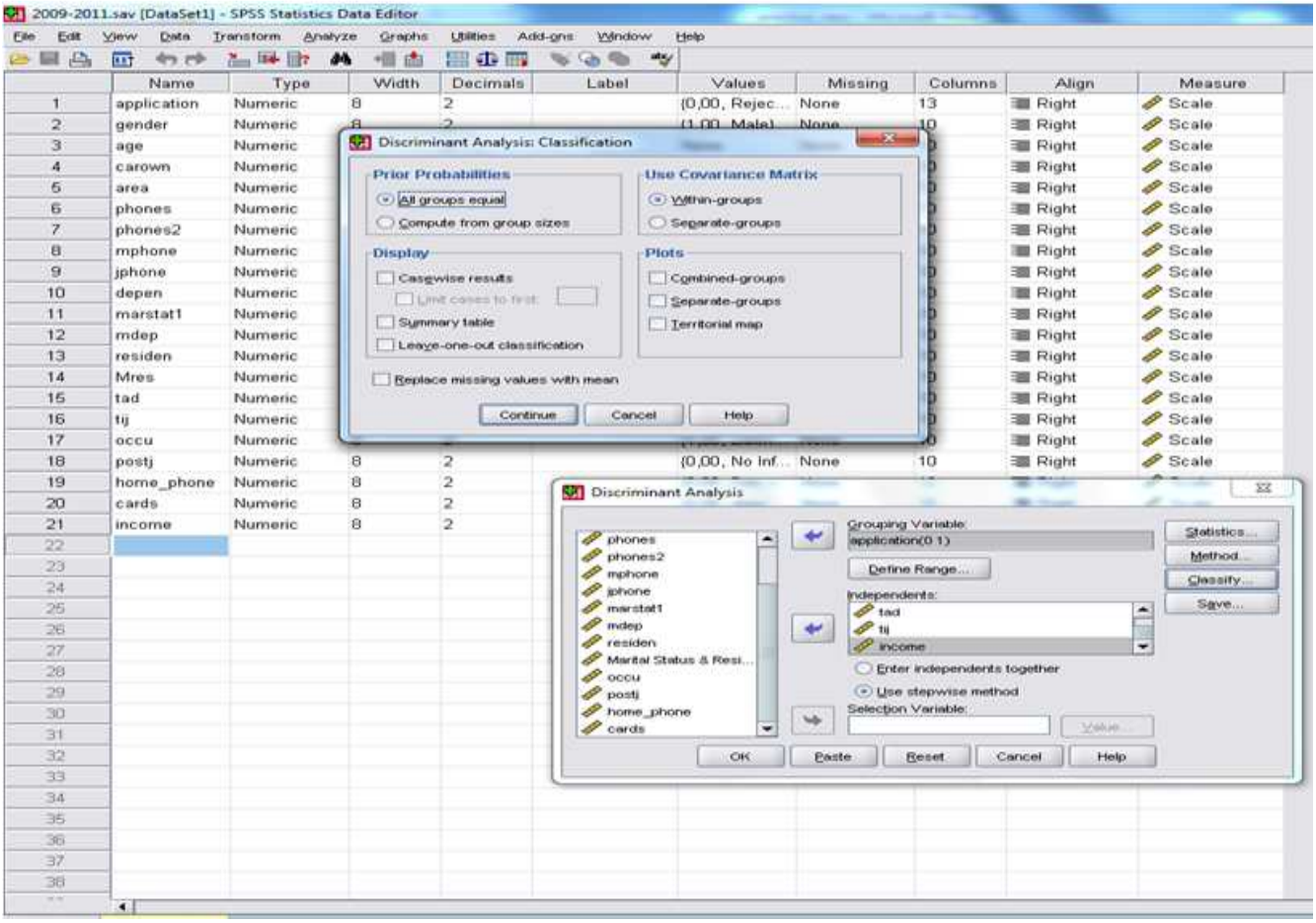
Όμοια εκφράζεται και το μοντέλο της δεύτερης κατηγορίας:

$$w_2 = -0{,}209 * x_1 + (0{,}12) * x_2 + 0{,}083 * x_3 + (-0{,}314) * x_4 + (-80{,}117)$$

Αυτό το πινακάκι μας βοηθάει να κατατάξουμε κάθε νέο πελάτη της τράπεζας που ζητάει κάποιου είδους πίστωση σε μία από τις δύο κατηγορίες της μεταβλητής application.

Αν $w_1 > w_2$ τότε κατατάσσουμε τον πελάτη στην πρώτη κατηγορία δηλαδή ότι δεν είναι άξιος για να του δώσει η τράπεζα κάποιου είδους πίστωσης, αλλιώς τον κατατάσσουμε στην 2η κατηγορία.

Υπομενού classify



Αποτελείται από 4 κατηγορίες

1) Prior probabilities

   Μπορούμε να διαλέξουμε ανάμεσα σε ίσες πιθανότητες ή σε υπολογισμό από το μέγεθος του δείγματος. Δυστυχώς αν έχουμε κάποια άλλου είδους πληροφόρηση δεν μπορούμε να την χρησιμοποιήσουμε στο SPSS. Για το λόγο αυτό, αν το SPSS είναι το μοναδικό στατιστικό πακέτο που διαθέτουμε, καλό θα ήταν όταν έχουμε κάποια πληροφόρηση για τα πραγματικά ποσοστά των ομάδων στον πραγματικό πληθυσμό τότε και στο δείγμα διατηρείται και η ίδια αναλογία έτσι ώστε να επιλέξουμε Compute from group sizes.

2) Use covariance matrix

   μπορούμε να επιλέξουμε αν η λύση θα γίνει με ίσες ή με άνισες διακυμάνσεις. Αυτό προϋποθέτει να κάνουμε πρώτα τον έλεγχο Box's M το οποίο στο δικό μας παράδειγμα απορρίπτει την υπόθεση της ισότητας πινάκων διακυμάνσεων-συνδιακυμάνσεων, οπότε επιλέγουμε χρήση του συνδυασμένου πίνακα συνδιακυμάνσεων (Within-groups).

3) plots

4) Diplay

   περιλαμβάνει εμφάνιση αποτελεσμάτων ανά παρατήρηση (casewise results). Αν έχουμε πολλές παρατηρήσεις μπορούμε να περιορίσουμε την εμφάνιση των αποτελεσμάτων στις πρώτες n παρατηρήσεις (limit cases to first..). Επιπλέον μπορούμε να δούμε πως κατατάσσεται η κάθε παρατήρηση αν κάνουμε διαχωριστική ανάλυση χωρίς συγκεκριμένη παρατήρηση και μετά να την κατατάξουμε με βάση την διαχωριστική συνάρτηση των άλλων παρατηρήσεων (Leave-one-out classification). Τέλος η επιλογή του Summary table μας δίνει ένα πολύ σημαντικό πίνακα:

Classification Results[b,c]

		application	Predicted Group Membership		
			Rejected	Accepted	Total
Original	Count	Rejected	66354	44731	111085
		Accepted	46355	68456	114811
		Ungrouped cases	1	2	3
	%	Rejected	59,7	40,3	100,0
		Accepted	40,4	59,6	100,0
		Ungrouped cases	33,3	66,7	100,0
Cross-validated[a]	Count	Rejected	66353	44732	111085
		Accepted	46361	68450	114811
	%	Rejected	59,7	40,3	100,0
		Accepted	40,4	59,6	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 59,7% of original grouped cases correctly classified.

c. 59,7% of cross-validated grouped cases correctly classified.

Ο οποίος είναι χρήσιμος για τον υπολογισμό της επιτυχίας της διαχωριστικής ανάλυσης μας. Πιο συγκεκριμένα εδώ βλέπουμε ότι έχουμε 59,6% ποσοστό σωστού διαχωρισμού για την συνολική διαχωριστική ανάλυση και 59,7% για την προσέγγιση της διασταυρωμένης επικύρωσης. Τα ποσοστά δεν είναι μεγάλα, είναι όμως ικανοποιητικά. Αυτό θα πρέπει να το περιμένουμε σε πραγματικά δεδομένα. Το ποσοστό του επιτυχημένου διαχωρισμού μπορούμε να το

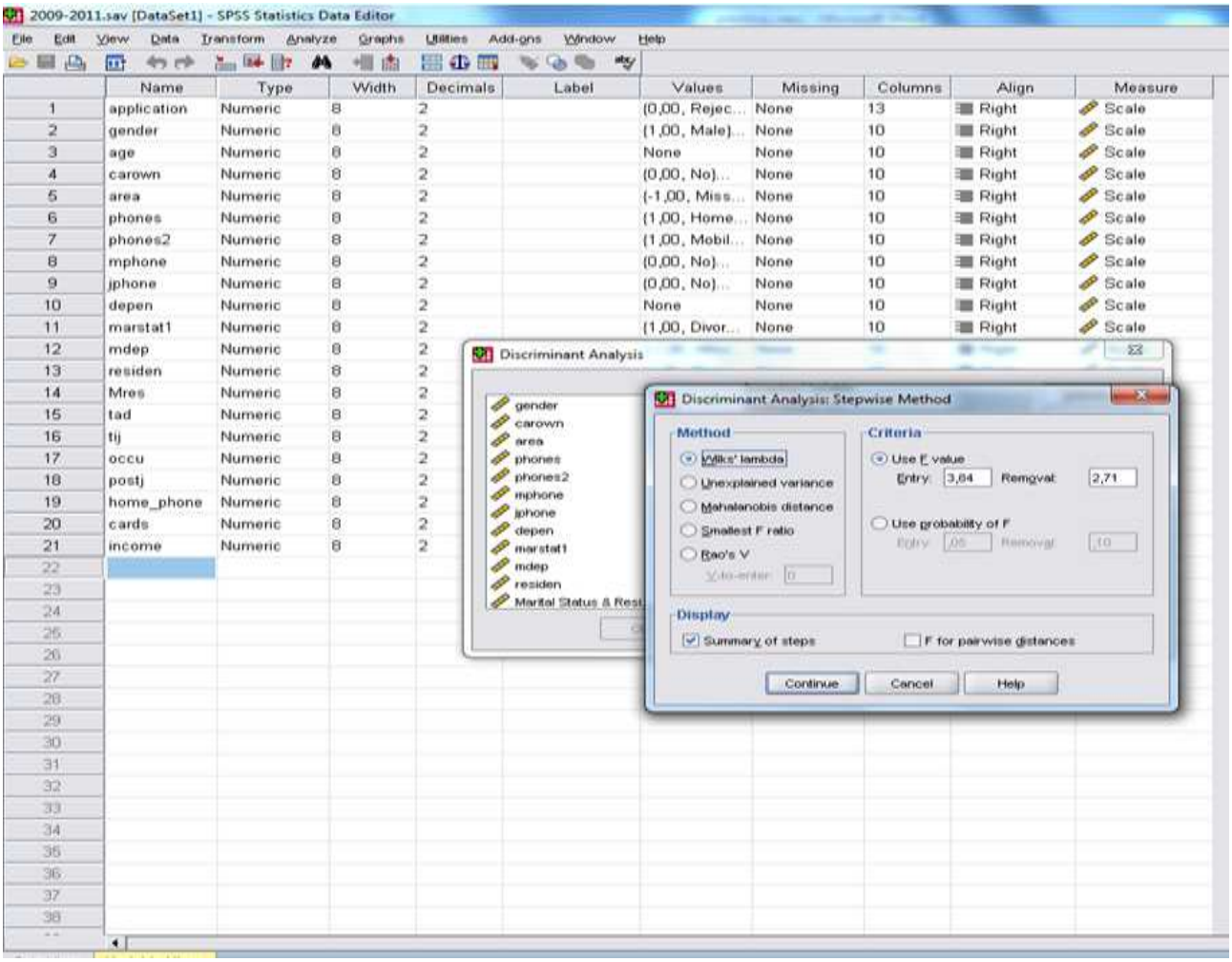
μετρήσουμε με το ποσοστό των σωστά καταχωρημένων παρατηρήσεων (59,6) όπως παραπάνω αλλά και με τον υπολογισμό του δείκτη συμφωνίας Kappa αφού αποθηκεύσουμε τις προβλεπόμενες ομάδες. Αυτό μπορούμε να το κάνουμε κάνοντας τον αντίστοιχο πίνακα συνάφειας (Crosstabs

Analyze> Descriptive Statistics> Crosstabs

μόνο που στο υπομενού Statistics θα επιλέξουμε μόνο την επιλογή Kappa. Αυτό που θέλουμε τιμές κοντά στο 1.

Υπομενού method

Αυτή η επιλογή είναι δυνατή μόνο στην περίπτωση που ορίσουμε την stepwise method ως μέθοδο επιλογής μεταβλητών.



Η πρώτη ομάδα των επιλογών αναφέρεται στην επιλογή της μεθόδου επιλογής των μεταβλητών (method):

1) Wilks' Lambda: Σε κάθε βήμα επιλέγουμε ποια μεταβλητή θα εισάγουμε στο μοντέλο μα βάση την μείωση του λάμδα του Wilks. Για κάθε μεταβλητής υπολογίζεται ένα F το οποίο βασίζεται στην διαφορά μεταξύ των λάμδα του Wilks για τα μοντέλα με και χωρίς την μεταβλητή. Μη ξεχνάμε ότι ο δείκτης Wilks μετράει το ποσοστό της μη ερμηνεύσιμης από το μοντέλο διακύμανσης.

2) Unexplained Variance: εδώ επιλέγουμε σαν δείκτη απόδοσης της κάθε μεταβλητής το άθροισμα της ερμηνεύσιμης διακύμανσης ανάμεσα στα ζευγάρια των ομάδων.

3) Mahalanobis Distance: η μέθοδος αυτή βασίζεται στον υπολογισμό της απόστασης Mahalanobis των δύο πιο κοντινών ομάδων. Σε κάθε βήμα εισάγεται η μεταβλητή που μεγαλώνει αυτή την απόσταση.

4) Smallest F-ratio: η μέθοδος αυτή βασίζεται στον υπολογισμό του F για όλα τα ζευγάρια των τιμών και από αυτές τις τιμές επιλέγουμε το μικρότερο F. Σε κάθε βήμα επιλέγεται εκείνη η μεταβλητή που μεγιστοποιεί αυτό το μικρότερο F.

5) Rao's V: η μέθοδος αυτή βασίζεται στον υπολογισμό της απόστασης Mahalanobis μεταξύ της κάθε ομάδας και του συνολικού δείγματος. Σε κάθε βήμα εισάγεται εκείνη η μεταβλητή που μεγιστοποιεί την απόσταση αυτή.

Η ομάδα επιλογών εμφάνισης λεπτομερειών (display) μας παρέχει τις εξής δυνατότητες:

1) δίνει την δυνατότητα να παρακολουθήσουμε περιληπτικά την εισαγωγή και απαλοιφή των μεταβλητών από το μοντέλο (summary of steps) και
2) να δούμε τους δείκτες F για όλες τις συγκρίσεις ανά ζεύγη (F for pairwise distances).

Σε αυτό το σημείο θα δώσουμε και την ερμηνεία κάποιων επιπλέον πινάκων που εξάγει η διαχωριστική ανάλυση.

- Ο πίνακας Wilks' Lambda

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,953	10768,347	4	,000

Ο παραπάνω πίνακας υπολογίζει το λάμδα του Wilks, το οποίο όπως είπαμε αντιπροσωπεύει το ποσοστό της διακύμανσης που δεν εξηγεί το μοντέλο. Μπορούμε να χρησιμοποιήσουμε το λάμδα για να ελέγξουμε την υπόθεση ότι οι μέσοι όλων των μεταβλητών ανά ομάδα είναι ίσοι. Αυτό το τεστ μπορεί να μας δώσει καλά αποτελέσματα σε περίπτωση που οι μεταβλητές μας δεν είναι κατάλληλες για διαχωρισμό. Εδώ απορρίπτουμε την ισότητα των μέσων (p-value=0.000<0.05), άρα δεν φαίνεται να υπάρχει πρόβλημα με την εφαρμογή της διαχωριστικής ανάλυσης μας.

- Ο πίνακας structure matrix

Ο πίνακας structure matrix μας δίνει τους δείκτες συσχέτισης κάθε ανεξάρτητης μεταβλητής με κάθε μια διαχωριστική συνάρτηση και μας βοηθάει στο να αξιολογήσουμε πόσο σημαντική είναι κάθε μεταβλητή για την κατασκευή της διαχωριστικής συνάρτησης. Θέλουμε τιμές πάνω από 0,5.

Structure Matrix

	Function
	1
income	,980
tij	,315
tad	,222
depen	,031
age[a]	-,012

Σε αυτό το dataset οι μεταβλητές δεν φαίνεται να είναι και οι καλύτερες για την εφαρμογή του εν λόγω διαχωρισμού σε άξιους και μη για χορήγηση πιστωτικής κάρτας, παρ' όλα αυτά είχαμε, όπως είδαμε παραπάνω, ένα καλό ποσοστό σωστού διαχωρισμού (59,6%).

- πίνακας κεντροειδών

Τέλος ο πίνακας κεντροειδών μας δίνει την μέση τιμή της κάθε κανονικοποιημένης διαχωριστικής συνάρτησης για κάθε μία ομάδα.

Functions at Group Centroids

application	Function
	1
Rejected	-,225
Accepted	,217

Εδώ έχουμε δύο ομάδες άρα μια συνάρτηση και μέσους ίσους με -0,225 για τους πελάτες που δεν αξίζουν να πάρουν πιστωτική κάρτα και επομένως θα απορριφτεί η αίτηση τους βάσει μοντέλου και 0,217 για αυτούς για τους οποίους θα γίνει έναρξη συνεργασίας με την τράπεζα για το συγκεκριμένο πιστωτικό προϊόν που έχουν αιτηθεί.

Ο διαχωρισμός του πληθυσμού με τη μέθοδο ανάλυσης διαχωρισμού έχει ολοκληρωθεί με ποσοστό σωστά καταχωρημένων παρατηρήσεων 59,6 %.

Το syntax για να πραγματοποιηθεί ο ανάλυση είναι το ακόλουθο:

**********discriminant analysis**********.

```
DISCRIMINANT
  /GROUPS=application(0 1)
  /VARIABLES=age depen income tad tij
  /ANALYSIS ALL
  /METHOD=WILKS
  /FIN=3.84
  /FOUT=2.71
  /PRIORS EQUAL
  /HISTORY
  /STATISTICS=MEAN STDDEV UNIVF BOXM COEFF TABLE CROSSVALID
  /PLOT=COMBINED SEPARATE  MAP
  /CLASSIFY=NONMISSING POOLED.
```

**********.

5.5 Επίλυση του προβλήματος με τη χρήση της λογιστικής παλινδρόμησης

Δεδομένου οτι η εξαρτημένη μεταβλητή «application» παίρνει μόνο δύο τιμές «RJ» και «AC» , μπορούμε να κάνουμε χρήση του μοντέλου της λογιστικής παλινδρόμησης ως μοντέλο διαχωρισμού.

Όπως έχει προαναφερθεί η λογική της λογιστικής Παλινδρόμησης στηρίζεται στα Odds που ισούνται με$\frac{\pi(x)}{1-\pi(x)}$, όπου π είναι η πιθανότητα εμφάνισης ενός ενδεχομένου. Επειδή όμως αυτός ο λόγος είναι θετικός και χωρίς πάνω όριο, όσο το π πηγαίνει στο 1 ο παραπάνω λόγος πηγαίνει στο άπειρο και όσο το π πηγαίνει στο 0 ο λόγος πηγαίνει στο μηδέν. Για να αντιμετωπίσουμε αυτό το πρόβλημα παίρνουμε το λογάριθμο του λόγου. Επομένως το μοντέλο της λογιστικής παλινδρόμησης παίρνει τη μορφή:

$$log\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \sum_{\kappa=1}^{p}\beta_0 + X_\kappa$$

Όπου $\pi(x)$ είναι η πιθανότητα ένα νέο δείγμα να ανήκει σε μια από τις δυο ομάδες.

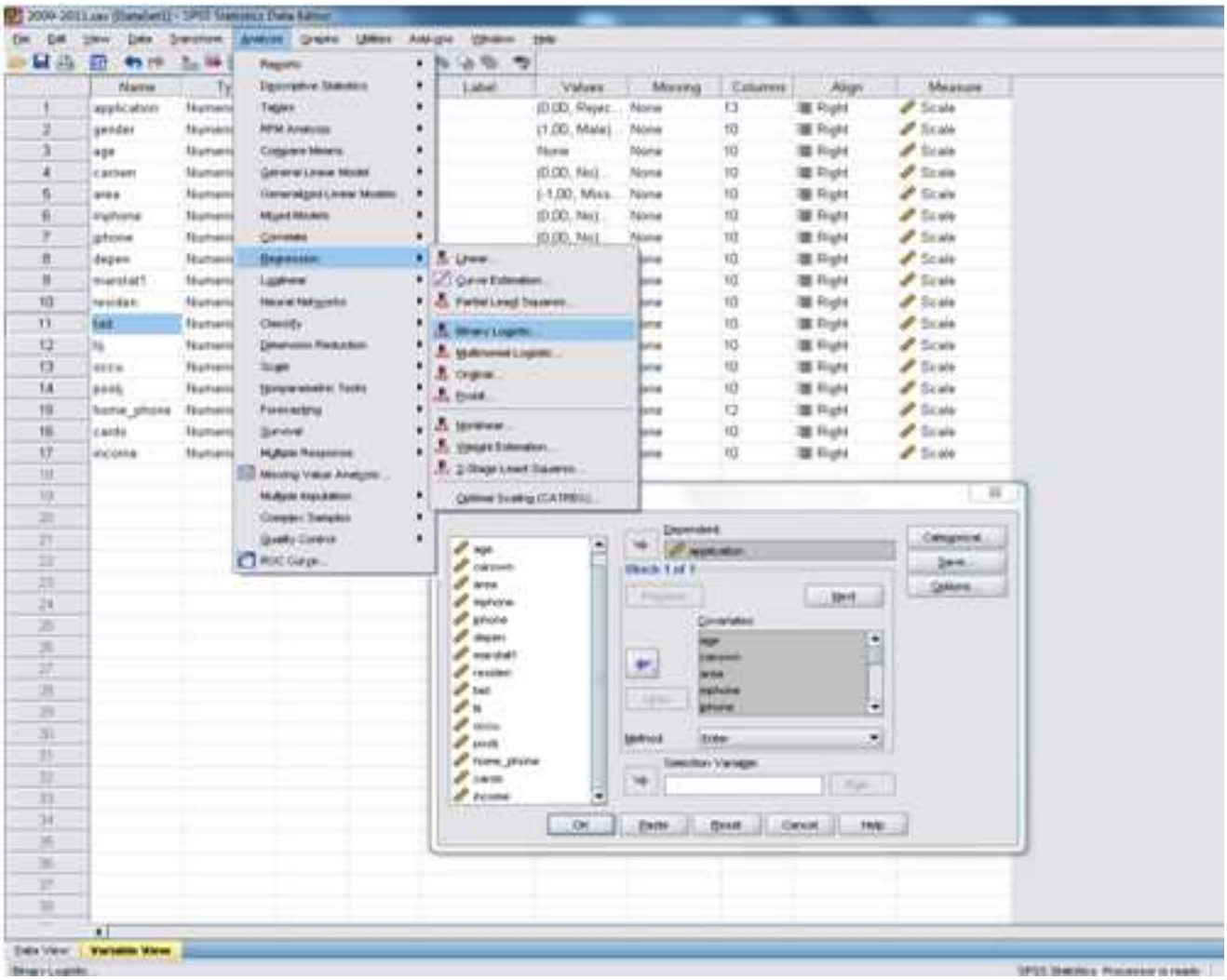
Τα στάδια που ακολουθούμε στο συγκεκριμένο πρόβλημα είναι δύο:

1^o στάδιο

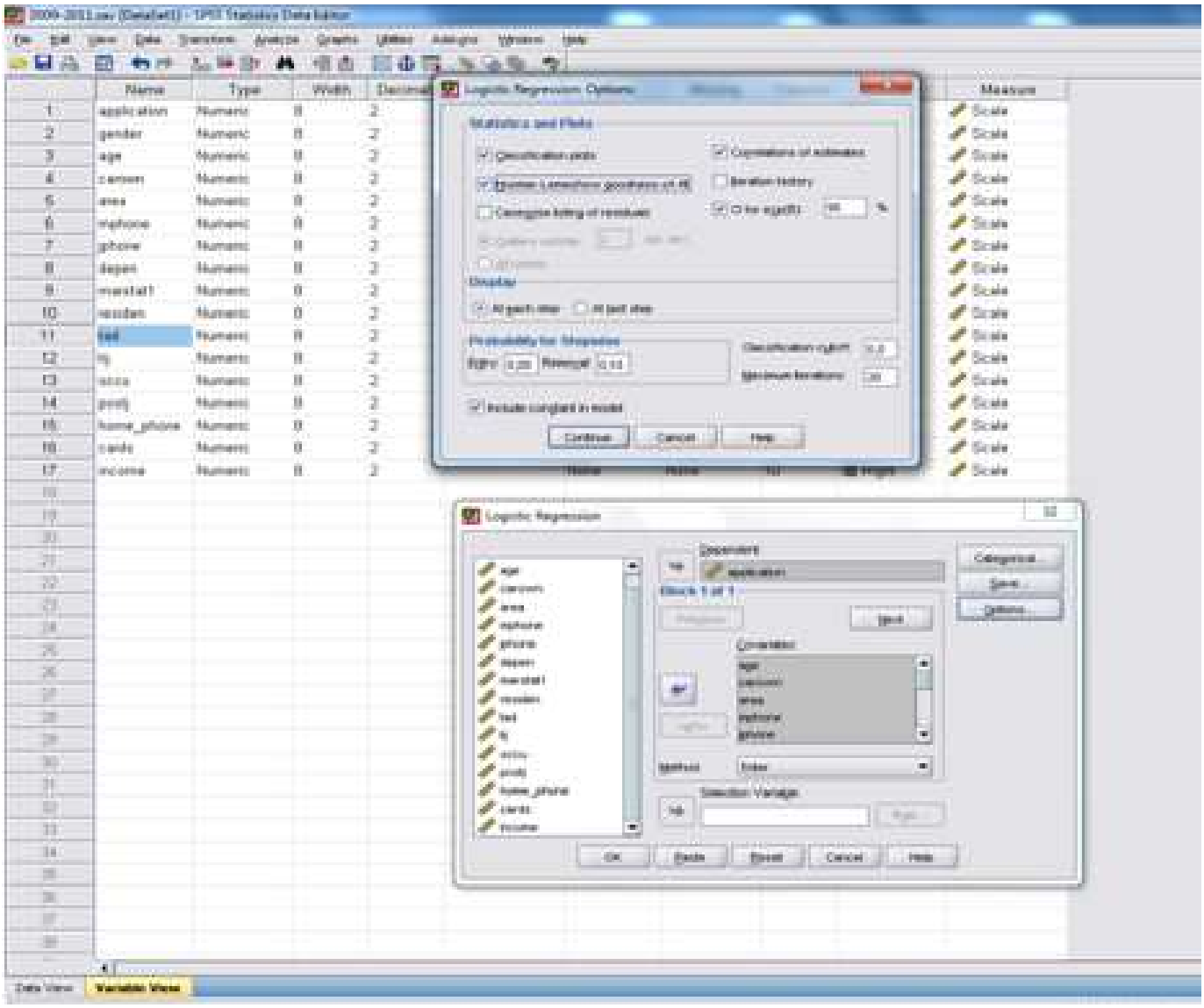
Ακολουθούμε τα παρακάτω βήματα:

Analyze> Regression> Binary Logistic

Και επιλέγουμε τις μεταβλητές τοποθετώντας στη θέση dependent τη μεταβλητή application και στη θέση covariates τις υπόλοιπες μεταβλητές όπως παρακάτω:



Πληκτρολογωντας το πεδιο «options» τσεκαρουμε τις παρακατω επιλογες:



Όπως έχει αναφερθεί αναλυτικά σε προηγούμενο κεφάλαιο η αρχική υπόθεση είναι:

Ho: $\beta_0 = \beta_1 = \cdots = \beta_\kappa = 0$

H1: αλλιως

Αν τα p-value για κάθε μεταβλητή είναι μικρότερα από 0.05 δηλαδή το 5% επίπεδο σημαντικότητας που έχουμε ορίσει τότε οι μεταβλητές είναι στατιστικά σημαντικές. Στην περίπτωση που p-value>0.05 τότε οι μεταβλητές

κρίνονται ακατάλληλες για το μοντέλο, οπότε εφαρμόζουμε μια μέθοδο επιλογής κατάλληλων μεταβλητών για το μοντέλο.
Αυτές οι μέθοδοι είναι οι Backward, Forward και Stepwise.

1) Στη **Backward** όλες οι μεταβλητές που διαθέτουμε που περιέχονται ήδη στην εξίσωση αξιολογούνται σύμφωνα με τα κριτήρια επιλογής για απαλοιφή (π.χ. ένα κριτήριο είναι η μεγιστοποίηση της πιθανοφάνειας). Εκείνες που είναι ακατάλληλες απαλείφονται μία-μία μέχρι να μην υπάρχει άλλη ακατάλληλη.
2) Στη **Forward** όλες οι μεταβλητές που διαθέτουμε που δεν περιέχονται στην εξίσωση αξιολογούνται σύμφωνα μα τα κριτήρια επιλογής για προσθήκη (π.χ. η μεγιστοποίηση πιθανοφάνειας). Εκείνες που είναι κατάλληλες προστίθενται μία-μία μέχρι να μην υπάρχει άλλη κατάλληλη.
3) Στη Stepwise, η οποία είναι και η πιο αξιόπιστη, η επιλογή των μεταβλητών στον σετ των μεταβλητών που διαθέτουμε προχωράει κατά βήματα. Σε κάθε βήμα αξιολογούνται οι μεταβλητές που περιλαμβάνονται ήδη στην εξίσωση σύμφωνα με τα κριτήρια επιλογής για απαλοιφή. Στην συνέχεια, οι μεταβλητές που δεν περιλαμβάνονται στην εξίσωση αξιολογούνται για προσθήκη. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να μην είναι κατάλληλη για προσθήκη ή απαλοιφή καμία μεταβλητή του σετ.

Θα ξεκινήσουμε την ανάλυση της λογιστικής παλινδρόμησης χωρίς να επιλέξουμε κάποια μέθοδο εντοπισμού στατιστικά σημαντικών μεταβλητών (στο πεδίο method επιλέγουμε το «enter»).

Πινακας 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	gender	,118	,009	173,000	1	,000	1,125	1,106	1,145
	age	,000	,000	,047	1	,828	1,000	1,000	1,000
	area	-,007	,001	34,006	1	,000	,993	,991	,996
	mphone	-,105	,013	69,696	1	,000	,901	,879	,923
	jphone	-,198	,015	172,851	1	,000	,821	,797	,845
	depen	-,088	,005	258,530	1	,000	,916	,906	,926
	marstat1	-,125	,005	563,830	1	,000	,882	,873	,892
	residen	-,159	,006	718,813	1	,000	,853	,843	,863
	tad	,004	,000	91,053	1	,000	1,004	1,003	1,004
	tij	-,003	,001	24,185	1	,000	,997	,996	,998
	occu	-,015	,001	224,227	1	,000	,985	,983	,987
	postj	,326	,013	605,196	1	,000	1,385	1,350	1,422
	home_phone	,597	,043	188,439	1	,000	1,816	1,668	1,978
	income	,000	,000	7777,154	1	,000	1,000	1,000	1,000
	Constant	-4,635	,083	3126,382	1	,000	,010		

Διεξάγοντας την ανάλυση με την μέθοδο Enter , παρατηρούμε στον πίνακα 1 ότι η μεταβλητή age που χρησιμοποιήσαμε στην ανάλυση δεν είναι στατιστικά σημαντική, αφού το p-value είναι μεγαλύτερα από το 0,05 επίπεδο στατιστικής σημαντικότητας που έχουμε ορίσει (επομένως επαληθεύει την Ho). Για να διορθώσουμε αυτό το πρόβλημα, θα χρησιμοποιήσουμε τη διαδικασία Backward (με την μέθοδο του Wald).

Το syntax για να πραγματοποιηθεί η ανωτέρω ανάλυση ειναι:

```
**********analyse-->regression-->binary**********.
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES application
  /METHOD=ENTER gender age carown area  mphone jphone depen marstat1
residen
     tad tij occu postj home_phone cards income
  /CLASSPLOT
  /PRINT=GOODFIT CORR CI(95)
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
**********************************************************.
```

2^{o} στάδιο

Επαναλαμβάνουμε τα προηγούμενα βήματα χρησιμοποιώντας τη μέθοδο Backward:

Analyze> Regression> Binary Logistic

Και τσεκάροντας ακριβώς τις ίδιες επιλογές. Επιπλέον στο πεδίο method επιλέγουμε τη δυνατότητα «Backward:Wald» και πατάμε ok.

Ο παρακάτω πίνακας 2 μας δείχνει πως το τελικό μοντέλο είναι στατιστικά σημαντικό αφού το p-value που αντιστοιχεί στο μοντέλο μας είναι 0.000<0.05 οπότε απορρίπτεται η μηδενική υπόθεση ότι όλοι οι συντελεστές β είναι ίσοι με μηδέν.

Πίνακας 2

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	13261,226	14	,000
	Block	13261,226	14	,000
	Model	13261,226	14	,000
Step 2[a]	Step	-,047	1	,828
	Block	13261,179	13	,000
	Model	13261,179	13	,000

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	13261,226	14	,000
	Block	13261,226	14	,000
	Model	13261,226	14	,000
Step 2[a]	Step	-,047	1	,828
	Block	13261,179	13	,000
	Model	13261,179	13	,000

Οι μεταβλητές που δεν παρέμεναν στο μοντέλο φαίνονται στον πίνακα 4

Πίνακας 4

Variables not in the Equation

		Score	df	Sig.
Step 2[a]	Variables age	,047	1	,828
	Overall Statistics	,047	1	,828

a. Variable(s) removed on step 2: age.

Παράλληλα ο πίνακας 3 δείχνει πώς βηματικά έγινε η επιλογή των μεταβλητών. Στο πρώτο βήμα υπάρχουν όλες οι μεταβλητές. Έπειτα στο βήμα 2 έμειναν στο μοντέλο μονάχα οι στατιστικά σημαντικές μεταβλητές. Στη στήλη Β φαίνονται οι συντελεστές β της κάθε μεταβλητής.

Πίνακας 3

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	gender	,118	,009	173,000	1	,000	1,125	1,106	1,145
	age	,000	,000	,047	1	,828	1,000	1,000	1,000
	area	-,007	,001	34,006	1	,000	,993	,991	,996
	mphone	-,105	,013	69,696	1	,000	,901	,879	,923
	jphone	-,198	,015	172,851	1	,000	,821	,797	,845
	depen	-,088	,005	258,530	1	,000	,916	,906	,926
	marstat1	-,125	,005	563,830	1	,000	,882	,873	,892
	residen	-,159	,006	718,813	1	,000	,853	,843	,863
	tad	,004	,000	91,053	1	,000	1,004	1,003	1,004
	tij	-,003	,001	24,185	1	,000	,997	,996	,998
	occu	-,015	,001	224,227	1	,000	,985	,983	,987
	postj	,326	,013	605,196	1	,000	1,385	1,350	1,422
	home_phone	,597	,043	188,439	1	,000	1,816	1,668	1,978
	income	,000	,000	7777,154	1	,000	1,000	1,000	1,000
	Constant	-4,635	,083	3126,382	1	,000	,010		
Step 2[a]	gender	,118	,009	173,164	1	,000	1,125	1,106	1,145
	area	-,007	,001	33,961	1	,000	,993	,991	,996
	mphone	-,105	,013	69,665	1	,000	,901	,879	,923
	jphone	-,197	,015	173,218	1	,000	,821	,797	,845
	depen	-,088	,005	258,520	1	,000	,916	,906	,926
	marstat1	-,125	,005	568,636	1	,000	,882	,873	,892
	residen	-,159	,006	718,966	1	,000	,853	,843	,863
	tad	,004	,000	91,015	1	,000	1,004	1,003	1,004
	tij	-,003	,001	24,288	1	,000	,997	,996	,998
	occu	-,015	,001	224,289	1	,000	,985	,983	,987
	postj	,326	,013	605,217	1	,000	1,385	1,350	1,422
	home_phone	,598	,043	190,922	1	,000	1,818	1,670	1,979
	income	,000	,000	7788,381	1	,000	1,000	1,000	1,000
	Constant	-4,638	,082	3203,598	1	,000	,010		

Παρατηρήσεις

1) Τελικά στο μοντέλο παρέμειναν οι gender, area, mphone, jphone, depen , marstat1, residen, tad, tij, occu, postj, home_phone, income. Αυτές οι μεταβλητές έχουν σημιτικό ρόλο στο διαχωρισμό των πελατών σε rejected και accepted.η age θεωρήθηκε στατιστικά μη σημαντική.

2) Το μοντέλο διαχωρισμού είναι:

$$\log\frac{\pi(x)}{1-\pi(x)} = 0{,}119 * x_1 + (-0{,}007) * x_2 + (-0{,}105) * x_3 + (-0{,}197) * x_4 + (-0{,}88) * x_5 + (-0{,}125) * x_6 + (-0{,}159) * x_7 + 0{,}04 * x_8 + (-0{,}03) * x_9 + (-0{,}15) * x_{10} + 0{,}326 * x_{11} + 0{,}598 * x_{12} + 0{,}0000000001 * x_{13} + (-4{,}638)$$

Το π(x) είναι η πιθανότητα κάθε νέος πελάτης του οποίου η αίτηση έγινε accepted, στο μέλλον να έχει καλή συναλλακτική συμπεριφορά. Δηλαδή να είναι Good πελάτης. Επομένως πρέπει να τεθεί ένα «αποκομμένο σημείο» cut off κάτω από το οποίο κάθε αίτηση θα γίνεται rejected. Αυτό εξαρτάται από την στρατηγική της κάθε τράπεζας: αν επιθυμεί να εφαρμόσει επιθετική πολιτική στην αγορά, θα έχει χαμηλό «αποκομμένο σημείο» επομένως θα εγκρίνει περισσότερες αιτήσεις με αποτέλεσμα να μεγαλώσει το χαρτοφυλάκιο της αλλά και παράλληλα να αυξηθεί ο κίνδυνος να συνεργαστεί με πελάτες που στο μέλλον θα αθετήσουν τις συμβατικές υποχρεώσεις τους με τον οργανισμό.

Τέλος ο πίνακας 5 δείχνει την ικανότητα διαχωρισμού του μοντέλου

Πίνακας 5

Classification Table[a]					
			Predicted		
			application		Percentage
	Observed		Rejected	Accepted	Correct
Step 1	application	Rejected	64472	46613	58
		Accepted	42487	72324	63
	Overall Percentage				60,6
Step 2	application	Rejected	64476	46609	58
		Accepted	42487	72324	63
	Overall Percentage				60,6
a. The cut value is ,500					

Ο πίνακας 5 δείχνει τον τελικό διαχωρισμό των πελατών σε rejected και accepted. Η ομάδα των πελατών που θα πάρει κάποια πίστωση αποτελείται από 114811(=42487+72324) άτομα, ενώ η ομάδα που αφορά αυτούς που δεν θα λάβουν πίστωση από 111085 (=64476+46609) άτομα. Αυτοί που φαίνεται ότι θα πάρουν πίστωση δεδομένου ότι δεν την αξίζουν ανέρχονται σε 42487, ενώ αυτοί που θα πάρουν κάποιου είδους πίστωση δεδομένου ότι την αξίζουν ανέρχονται σε 72324. Ανάλογα ερμηνεύουμε και τους πελάτες που δεν θα πάρουν κάποιου είδους πίστωση. Ο διαχωρισμός των πελατών έγινε με ποσοστό σωστού διαχωρισμού ίσο με 60,6% (αρκετά υψηλό).

Το syntax για την παραπάνω ανάλυση είναι το ακόλουθο:

```
**********analyse-->regression-->binary2**********.
LOGISTIC REGRESSION VARIABLES application
  /METHOD=BSTEP(WALD) gender age carown area mphone jphone depen
marstat1 residen
   tad tij occu postj home_phone cards income mres
  /CLASSPLOT
  /PRINT=GOODFIT CORR CI(95)
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
**************************************************.
```

5.6 Ομοιότητες - Διαφορές διαχωριστικής ανάλυσης- λογιστικής παλινδρόμησης

Οι δυο διαδικασίες που αναλύθηκαν στην εργασία έχουν ομοιότητες και διαφορές. Παραθέτουμε τις πιο σημαντικές από αυτές:

Ομοιότητες

1) Η διαχωριστική ανάλυση και η λογιστική παλινδρόμηση είναι δυο από τις σημαντικότερες πρωτογενείς μεθόδους κατάταξης παρατηρήσεων. Είναι οι πιο απλές και παράλληλα ουσιαστικές ως προς το αποτελέσματα που εξάγονται από αυτές, μέθοδοι διαχωρισμού ομάδων.
2) Όπως περιγράφεται στην εργασία, οι δυο μέθοδοι χρησιμοποιούνται για να διακρίνουν ένα πληθυσμό σε ομάδες αλλά και να κατατάξουν κάθε νέα παρατήρηση σε μια από τις δυο αυτές ομάδες.

Διαφορές

Θα παρουσιάσουμε τις διαφορές των δυο μεθόδων ως πλεονεκτήματα και μειονεκτήματα της μιας ως προς την άλλη:

<u>Πλεονεκτήματα λογιστικής παλινδρόμησης</u>

1) Τα μοντέλα λογιστικής παλινδρόμησης υπολογίζουν αυτόματα τις πιθανότητες κάθε ομάδας. Με τη διαχωριστική ανάλυση αυτό γινεται μόνο έμμεσα.
2) Υποθέσεις όπως:
   i) κανονικότητα συνέχων μεταβλητών
   ii) εξαρτημένες μεταβλητές πρέπει να είναι ποσοτικές,
   iii) οι εξαρτημένες μεταβλητές πρέπει να έχουν ίσες συνδιασπορές στον κάθε πληθυσμό

   είναι χαρακτηριστικά της διαχωριστικής ανάλυσης που δεν είναι απαραίτητα στην λογιστική παλινδρόμηση.

3) Δεν κάνει καμία υπόθεση για τις ανεξάρτητες μεταβλητές, οι οποίες είναι συνεχείς, διακριτές, διχοτομημένες ή και ένας συνδυασμός από αυτές

Πλεονεκτήματα διαχωριστικής ανάλυσης

1) Η διαχωριστική ανάλυση είναι υπολογιστικά απλούστερη μέθοδος και στηρίζεται σε ρεαλιστικότερες μεθόδους από τη λογιστική παλινδρόμηση.
2) Οι υποθέσεις για την διαχωριστική ανάλυση είναι πιο πολύπλοκες από τη λογιστική παλινδρόμηση. Αν όμως αν η υπόθεση της κανονικότητας ισχύει και για τον κάθε πληθυσμό και οι εξαρτημένες μεταβλητές έχουν τον ίδιο πίνακα διασποράς-συνδιασποράς, η διαχωριστική ανάλυση δίνει πολύ καλύτερα αποτελέσματα σε σχέση με την λογιστική παλινδρόμηση.
3) Η discriminant analysis απαιτεί πολύ μικρότερο μέγεθος δείγματος για να πραγματοποιηθεί.
4) Η discriminant analysis είναι ικανή να διαχωρίσει ένα πληθυσμό σε περισσότερες από μια ομάδες. Η λογιστική παλινδρόμηση δεν έχει τέτοια δυνατότητα, καθώς κάνει διαχωρισμό σε δυο μόνο ομάδες. Επομένως εάν μια τράπεζα θέλει να διαχωρίσει τους πελάτες σε 3 επίπεδα «Good», «Bad», «Grey» κατάλληλη μέθοδος είναι η discriminant analysis.

Για τους ανωτέρω λόγους στην πράξη προκρίνεται η λογιστική ανάλυση ως καλύτερο μοντέλο διαχωρισμού σε κάθε περίπτωση πάντως το κάθε μοντέλο έχει τόσο μειονεκτήματα που το κάνουν να αδυνατεί σε σχέση με το άλλο, όσο και πλεονεκτήματα που το καθιστούν ισχυρότερο στη σύγκριση των δυο μεθόδων διαχωρισμού ενός πληθυσμού.

Βιβλιογραφία

1) Hosmer, D. W., and S. Lemeshow, “Applied Logistic Regression”, New York: John Wiley & Sons, 1989.

2) Pregibon, D. “Logistic Regression Diagnostics”, Annals of Statistics 9, 1981

3) Murray, G. D., “ A cautionary Note on Selection of Variables in Discriminant Analysis”, Applied Statistics, 26, 1977.

4) Finney, D. J. Probit Analysis. 3rd edition. Cambridge, 1971.

5) Fisher, R. A., “The Statistical Utilization of Multiple Measurements”, Annals of Eugenics, 8, 1938.

6) Hand, D. J., Discrimination and classification, John Wiley, New York 1981

7) Naeem Siddiqi, Credit Risk Scorecards «Developing and Implementing Intelligent Credit Scoring»,Wiley and sas Istitute, Canada 2006.

8) Darrel Duffie-Keneth J Singleton, «Credit risk», Princeton series in finance, Princeton, New Jersey 2003

9) Golin Jonathan, «The Bank Credit Analysis Handbook: A Guide for Analysts, Bankers and Investors», Wiley, New York 2001

10) Γιώργος Λελεδάκης: "Ανάλυση και Διαχείριση Χαρτοφυλακίου", Εκδόσεις Οικονομικού Πανεπιστημίου Αθηνών, Αθήνα 2007.

11) Δημήτρης, Καρλής, Πολυμεταβλητή Στατιστική Ανάλυση, Εκδόσεις Αθ. Σταμούλη, Αθήνα 2005

12) Ευδοκία Ξεκαλάκη, «Μη Παραμετρική Στατιστική», Εκδόσεις Οικονομικό Πανεπιστήμιο Αθηνών, Αθήνα 2001

13) «Στοχαστική ανάλυση και εφαρμογές στη χρηματοοικονομική», Αθανάσιος Ν Γιαννακόπουλος, Φεβρουάριος 2003

14) Normal Draper – Harry Smith (μετάφραση: Ευσταθιος Χατζηκωνσταντινιδης-Αγλαΐα Καλαματιανου) «Εφαρμοσμένη Ανάλυση Παλινδρόμησης», εκδόσεις Παπαζηση, Αθηνα 1997

15) Ζοπουνιδης Κ., Λεμονακης Χ., «Διαχείριση πιστωτικού κινδύνου», εκδόσεις Κλειδάριθμος Αθήνα 2009

16) Έκθεση της ICAP «ΕΞΕΛΙΞΗ ΤΗΣ ΠΙΣΤΟΛΗΠΤΙΚΗΣ ΙΚΑΝΟΤΗΤΑΣ ΤΩΝ ΕΛΛΗΝΙΚΩΝ Α.Ε.& Ε.Π.Ε. ΕΠΙΧΕΙΡΗΣΕΩΝ 2005 – 2004»