# ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

Τομέας Μαθηματικών


# Επιλογή Στατιστικών Μοντέλων: Εφαρμογή σε Ψυχιατρικά Δεδομένα


*Συγγραφέας*
Αλεξάνδρα Πουλοπούλου

*Επιβλέπων Καθηγητής*
Φουσκάκης Δημήτρης, Αναπλ. Καθηγητής, ΣΕΜΦΕ

*Τριμελής Επιτροπή*
Φουσκάκης Δημήτρης, Αναπλ. Καθηγητής, ΣΕΜΦΕ
Λουλάκης Μιχάλης, Αναπλ. Καθηγητής, ΣΕΜΦΕ
Κολέτσος Ιωάννης, Επικ. Καθηγητής, ΣΕΜΦΕ



Αθήνα, Νοέμβριος 2017

*Στους γονείς μου...*
*που θυσίασαν τα πάντα για τα παιδιά τους.*

# ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά, θα ήθελα να ευχαριστήσω τον Αναπληρωτή Καθηγητή του Ε.Μ.Π., κο. Φουσκάκη Δημήτριο για τη δυνατότητα που μου προσέφερε να ασχοληθώ με το συγκεκριμένο θέμα, καθώς επίσης και για την πολύτιμη καθοδήγησή του, σε όλη τη διάρκεια εκπόνησης της παρούσας εργασίας.

Φυσικά, δε θα μπορούσα να παραλείψω την οικογένειά μου που με στήριξε με κάθε τρόπο, όλα αυτά τα χρόνια, και τον Πέτρο που υπήρξε κινητήρια δύναμη στην ολοκλήρωση αυτής της εργασίας.

Αλεξάνδρα Πουλοπούλου
Αθήνα, Οκτώβριος 2017

# ΠΕΡΙΛΗΨΗ

Την τελευταία δεκαετία, η χρήση των γενικευμένων γραμμικών μοντέλων έχει εξαπλωθεί σε πολλούς επιστημονικούς τομείς. Ο λόγος είναι, ότι βρίσκουν εφαρμογή σε δεδομένα διαφόρων φύσεων και αποδίδουν αξιόπιστα αποτελέσματα. Η ανάγκη μελέτης των σχέσεων εξάρτησης μεταξύ μεταβλητών, είναι το πιο συχνό ερώτημα στην επιστημονική κοινότητα. Με την πάροδο των χρόνων, τα κριτήρια επιλογής βέλτιστων μοντέλων και κατάλληλων μεταβλητών έχουν βελτιωθεί σημαντικά και αποτελούν ένα εξαιρετικά χρήσιμο εργαλείο.

Στην παρούσα διπλωματική εργασία, παρουσιάζεται η θεωρία των γενικευμένων γραμμικών μοντέλων παλινδρόμησης, με ιδιαίτερη έμφαση να δίνεται στο λογιστικό μοντέλο. Επιπλέον, μελετάται η ποινικοποιημένη μέθοδος επιλογής μεταβλητών lasso, η οποία αποτελεί την πιο διαδεδομένη τεχνική της κατηγορίας της. Στο τέλος, γίνεται εφαρμογή σε πραγματικά ψυχιατρικά δεδομένα, με τη χρήση του στατιστικού πακέτου της R.

Συγκεκριμένα, στο πρώτο κεφάλαιο, παρατήθεται ο ορισμός της κατάθλιψης και η επιδημιολογία της. Στην αρχή δίνεται μια σύντομη εισαγωγή στις διαστάσεις που έχει πάρει η ψυχική αυτή διαταραχή στον κόσμο με στατιστικά στοιχεία άλλων ερευνών. Στο πρώτο μέρος της εργασίας περιλαμβάνεται και η ιστορική αναδρομή της νόσου. Μετέπειτα, αποτυπώνεται η συμπτωματολογία της κατάθλιψης με βάση το σύστημα ταξινόμησης DSM-5 και οι μορφές της, όπως αυτές περιγράφονται στη Διεθνή Στατιστική Ταξινόμηση Νοσημάτων και Συναφών Προβλημάτων Υγείας (International Statistical Classification of Diseases and Related Health Problems - ICD-10). Τέλος, αναφέρονται οι τρόποι αντιμετώπισης της καταθλιπτικής διαταραχής στην σύγχρονη κοινωνία.

Στο δεύτερο κεφάλαιο, παρουσιάζεται η θεωρία γύρω από τα μοντέλα παλινδρόμησης για δίτιμες μεταβλητές απόκρισης. Ιδιαίτερη έμφαση δίνεται στο μοντέλο της λογιστικής παλινδρόμησης, καθώς είναι ευκολότερα ερμηνεύσιμο, για δίτιμα δεδομένα. Στη συνέχεια, γίνεται διαχωρισμός των γενικευμένων γραμμικών μοντέλων με βάση τη συνάρτηση σύνδεσης και έπειτα, παρουσιάζονται οι διαφορές μεταξύ δύο, αυτών της logit και probit συνάρτησης.

Στο τρίτο κεφάλαιο, γίνεται μία εισαγωγή στις μεθόδους επιλογής μεταβλητών. Στην αρχή του κεφαλαίου, παρουσιάζεται ένα από τα προβλήματα που καλούνται να λύσουν αυτές οι μέθοδοι, το φαινόμενο της πολυσυγγραμμικότητας. Η προσοχή στρέφεται στη ποινικοποιημένη μέθοδο lasso, η οποία παρουσιάζεται με κάθε λεπτομέρεια. Με σκοπό να εισάγουμε τη συγκεκριμένη μέθοδο, γίνεται επίσης, μία σύντομη περιγραφή των τεχνικών επιλογής υποσυνόλων, καθώς και της παλινδρόμησης κορυφογραμμής.

Στο τέταρτο και τελευταίο κεφάλαιο, εφαρμόζονται οι βασικές τεχνικές που ανα-

πτύχθηκαν στη θεωρία και διεξάγονται τα ανάλογα συμπεράσματα. Πιο συγκεκριμένα, προσαρμόζεται ένα μοντέλο λογιστικής παλινδρόμησης και ένα μοντέλο probit. Πριν από αυτό, έχει προηγηθεί ο καθαρισμός των διαθέσιμων δεδομένων και μία μικρή περιγραφή τους. Στη συνέχεια, πραγματοποιείται μία ανάλυση με τη μέθοδο lasso και τέλος, συγκρίνονται τα αποτελέσματα των μεθόδων που αναπτύχθηκαν.

# ABSTRACT

The last decade, the use of generalized linear models is spreading in many scientific areas. The reason of this phenomenon, is that they can be applied on different data and deliver reliable results. The need of building models to deliver and describe the relationship of dependence between the variables, is the most common scientific question. Over the years, model selection and variable selection criteria have been improved and constitute a set of extremely useful tools.

In this dissertation, we analyze the theory of generalized linear models, giving emphasis on logistic regression model. Furthermore, we study the variable selection method lasso, which is the most widespread technique in its category. In the end, the methods are applied to real psychiatric data, using the statistical package R.

In particular, the first chapter describes the definition of depression and its epidemiology. At the beginning, there is a brief introduction to the dimensions that this mental disorder have occupied in the world, using statistical results from other surveys. The first part of this study, includes the historical retrospective of the disease. Then, we describe the symptomatology of depression, based on the DSM-5 classification system and its forms, as described in the International Statistical Classification of Diseases and Related Health Problems (ICD-10). Finally, we mention the ways of dealing with the depressive disorder in modern society.

In the second chapter, we present the theory about regression models for binary data. In particular, we emphasize to the logistic regression model, as its interpretation is easier than other models. Then, we separate the generalized linear models based on the link function and we present the differences between the logit and probit function.

In the third chapter, we present the theory of variable selection methods. At the beginning, we mention the problem of multicollinearity, that these methods are made to solve. The main subject of this chapter, is the lasso method, which we present in every detail. In order to introduce this method, we begin with a brief description of the subset selection techniques and ridge regression.

In the fourth and final chapter, we provide some applications using the statistical package R and we report the final results. In particular, we fit a logistic regression model and a probit model. Finally, we perform a lasso analysis and we compare the results of the methods that have been applied.

# ΠΕΡΙΕΧΟΜΕΝΑ

**ΠΕΡΙΛΗΨΗ** **1**

**ABSTRACT** **3**

**1 ΕΙΣΑΓΩΓΗ ΣΤΟ ΠΡΟΒΛΗΜΑ** **9**
1.1 Στατιστικά στοιχεία της νόσου . . . . . 9
1.2 Ιστορική αναδρομή . . . . . 11
1.3 Σκοπός της μελέτης . . . . . 12
1.4 Ορισμός της κατάθλιψης . . . . . 12
1.5 Ορισμός της νοσηλείας . . . . . 12
1.6 Συμπτωματολογία . . . . . 12
1.7 Ταξινόμηση της νόσου . . . . . 14
1.8 Τρόποι αντιμετώπισης . . . . . 15

**2 ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΓΙΑ ΔΙΤΙΜΕΣ ΜΕΤΑΒΛΗΤΕΣ ΑΠΟΚΡΙΣΗΣ** **19**
2.1 Εισαγωγή . . . . . 19
2.2 Ιστορική αναδρομή . . . . . 19
2.3 Γενικευμένα γραμμικά μοντέλα (GLMs, Generalized Linear Models) . . 20
2.3.1 Εισαγωγή . . . . . 20
2.3.2 Το μοντελο . . . . . 23
2.3.3 Συνιστώσες του γενικευμένου γραμμικού μοντέλου . . . . . 24
2.3.4 Συνάρτηση σύνδεσης (*link function*) . . . . . 25
2.4 Γενικευμένα γραμμικά μοντέλα για δυαδικά δεδομένα . . . . . 26
2.4.1 Εισαγωγή . . . . . 26
2.4.2 Γραμμικό μοντέλο πιθανοτήτων (*LPM, Linear Probability Model*) 27
2.4.3 Ο μετασχηματισμός logit (*logit link*) . . . . . 27
2.4.4 Άλλοι μετασχηματισμοί για δυαδικά δεδομένα . . . . . 28
2.4.5 Το πρόβλημα της υπερμεταβλητότητας . . . . . 30
2.5 Πολλαπλή λογιστική παλινδρόμηση . . . . . 31
2.5.1 Εισαγωγή . . . . . 31
2.5.2 Το μοντέλο . . . . . 32
2.5.3 Εκτίμηση των συντελεστών του μοντέλου . . . . . 33
2.5.4 Ερμηνεία των συντελεστών του μοντέλου . . . . . 34
2.5.5 Έλεγχοι Υποθέσεων . . . . . 36

2.5.6 Διαστήματα εμπιστοσύνης . . . . . 37
2.5.7 Ελεγχοσυνάρτηση Deviance . . . . . 38
2.5.8 Πίνακες ταξινόμησης (*classification tables*) . . . . . 40
2.5.9 Καμπύλες ROC . . . . . 41
2.5.10 Το φαινόμενο της πολυσυγγραμμικότητας . . . . . 43
2.5.11 Κριτήρια επιλογής του μοντέλου . . . . . 44
2.6 Ο Μετασχηματισμός Probit . . . . . 47
2.6.1 Το μοντέλο . . . . . 48
2.6.2 Εκτίμηση των συντελεστών του μοντέλου . . . . . 49
2.6.3 Έλεγχοι καλής προσαρμογής του μοντέλου . . . . . 49
2.7 Επιλογή μεθόδων ανάμεσα σε Logit και Probit . . . . . 49

**3 ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΜΕ ΤΗ ΜΕΘΟΔΟ LASSO 55**
3.1 Εισαγωγή . . . . . 55
3.2 Το φαινόμενο της πολυσυγγραμμικότητας . . . . . 56
3.2.1 Διαγνωστικοί έλεγχοι για την πολυσυγγραμμικότητα . . . . . 57
3.2.2 Τρόποι αντιμετώπισης της πολυσυγγραμμικότητας . . . . . 58
3.3 Επιλογή υποσυνόλου (*Subset selection*) . . . . . 59
3.3.1 Επιλογή καλύτερου υποσυνόλου (*Best subset selection*) . . . . 59
3.3.2 Διαδικασίες επιλογής μοντέλου με βήματα (*Stepwise selection*) 60
3.4 Παλινδρόμηση κορυφογραμμής (*Ridge regression*) . . . . . 62
3.4.1 Το μοντέλο . . . . . 63
3.4.2 Εκτίμηση των συντελεστών του μοντέλου . . . . . 64
3.4.3 Γιατί παλινδρόμηση κορυφογραμμής; . . . . . 65
3.5 Η μέθοδος LASSO . . . . . 66
3.5.1 Το μοντέλο . . . . . 66
3.5.2 Η γεωμετρία της Lasso . . . . . 70
3.5.3 Τυπικά σφάλματα . . . . . 71
3.5.4 Εκτίμηση του συντελεστή $t$ . . . . . 72
3.5.5 Lasso στα γενικευμένα γραμμικά μοντέλα . . . . . 74
3.5.6 Cross-Validation (*cvl*) . . . . . 74

**4 ΕΦΑΡΜΟΓΗ ΜΕ ΧΡΗΣΗ ΤΟΥ ΣΤΑΤΙΣΤΙΚΟΥ ΠΑΚΕΤΟΥ R 77**
4.1 Εισαγωγή . . . . . 77
4.2 Εισαγωγή των δεδομένων στην R . . . . . 79
4.3 Καθαρισμός δεδομένων . . . . . 82
4.4 Περιγραφή των μεταβλητών . . . . . 87
4.4.1 Περιγραφικά στατιστικά . . . . . 87
4.4.2 Συσχετίσεις . . . . . 90
4.5 Επιλογή στατιστικού μοντέλου . . . . . 93
4.5.1 Μοντέλο λογιστικής παλινδρόμησης . . . . . 93
4.5.2 Μοντέλο probit . . . . . 97
4.5.3 Συμπεράσματα και επιλογή μοντέλου . . . . . 101
4.6 Προσαρμογή του στατιστικού μοντέλου παλινδρόμησης . . . . . 102
4.6.1 Ερμηνεία των αποτελεσμάτων . . . . . 105
4.7 Εφαρμογή της μεθόδου lasso στη λογιστική παλινδρόμηση . . . . . 107

**5 ΣΥΜΠΕΡΑΣΜΑΤΑ** **113**

**ΠΑΡΑΡΤΗΜΑ** **115**

**ΒΙΒΛΙΟΓΡΑΦΙΑ** **121**

# ΚΕΦΑΛΑΙΟ 1

# ΕΙΣΑΓΩΓΗ ΣΤΟ ΠΡΟΒΛΗΜΑ

> Before you diagnose yourself with depression or low self-esteem, first make sure you are not, in fact, just surrounded by idiots.
>
> *Sigmund Freud*

## 1.1 Στατιστικά στοιχεία της νόσου

Η κατάθλιψη αποτελεί μία από τις συχνότερες ψυχικές διαταραχές η οποία προσβάλλει ανθρώπους σε όλες τις κοινότητες σε ολόκληρο τον κόσμο. Μέχρι σήμερα, εκτιμάται ότι 350 εκατομμύρια άνθρωποι έχουν διαγνωσθεί με τη συγκεκριμένη νόσο. Με βάση έρευνα που διεξήχθη από τον Παγκόσμιο Οργανισμό Υγείας (Π.Ο.Υ) σε 17 χώρες, 1 στους 20 ανθρώπους αναφέρει ότι εμφάνισε ένα καταθλιπτικό επεισόδιο τον περασμένο χρόνο. Η καταθλιπτική διαταραχή εμφανίζεται συχνά σε νεαρή ηλικία, περίπου στα 15 έτη. Οι λόγοι εμφάνισης της διαταραχής μπορεί να ποικίλουν ανάλογα τον άνθρωπο και τους χαρακτηριστικούς τρόπους που αντιμετωπίζει τα πράγματα.

Η συνήθης συμπτωματολογία της είναι η μείωση των φυσιολογικών λειτουργιών του ατόμου και συχνά είναι επαναλαμβανόμενη. Για τους λόγους αυτούς, η κατάθλιψη είναι η κύρια αιτία ανικανότητας παγκοσμίως. Παρ'όλα αυτά, τα συμπτώματα εκδήλωσης του καταθλιπτικού επεισοδίου διαφέρουν από άτομο σε άτομο. Η ανάγκη για την αντιμετώπιση και τον περιορισμό της συγκεκριμένης νόσου αλλά, και άλλων νοσημάτων ψυχικής υγείας αυξάνεται συνεχώς. Σε τελικό στάδιο, η κατάθλιψη μπορεί να οδηγήσει σε θάνατο μέσω της αυτοκτονίας. Σχεδόν 1 εκατομμύριο ζωές χάνονται ανά χρόνο εξ'αιτίας των αυτοκτονιών. Ο αριθμός αυτός μεταφράζεται σε 3000 αυτοκτονίες κάθε μέρα. Η αυτοκτονία εκτιμάται ότι είναι η δεύτερη αιτία θανάτου σε ηλικίες 15-29 χρόνων.

Η εμφάνιση της καταθλιπτικής διαταραχής ποικίλει σημαντικά μεταξύ των πληθυσμών παγκοσμίως. Τα ποσοστά επικράτησης κυμαίνονται από 3% στην Ιαπωνία έως

16,9% στις Ηνωμένες Πολιτείες, με τις περισσότερες χώρες να βρίσκονται μεταξύ 8% και 12%. Παρ'ολα αυτά, η έλλειψη των διαγνωστικών εργαλείων για την συγκεκριμένη νόσο καθιστά δύσκολη τη σύγκριση των ποσοστών αυτών σε εθνικό επίπεδο. Επιπλέον, οι διάφορες πολιτισμικές διαφορές μεταξύ των κρατών και άλλοι παράγοντες επηρεάζουν την έκφραση των ψυχικών διαταραχών. Υπάρχουν πολλές σημαντικές διαφορές μεταξύ των πολιτισμών που μπορεί να καταστήσουν ένα άτομο πιο ευάλωτο, στο να νοσήσει, από κάποιο άλλο. Τέτοιοι παράγοντες μπορεί να είναι:

- **Το φύλο**. Μία γυναίκα φαίνεται να είναι 2 έως 3 φορές πιο πιθανό να εμφανίσει κατάθλιψη, από έναν άντρα.
- **Κοινωνικοί παράγοντες**, όπως χαμηλή εκπαίδευση.
- **Οικονομικοί παράγοντες**, όπως φτώχεια.
- **Κληρονομικότητα**. Σε άτομα που φέρουν οικογενειακό ιστορικό με τη νόσο, είναι 2 έως 3 φορές πιο πιθανό να εμφανίσουν κατάθλιψη σε κάποιο στάδιο της ζωής τους.
- **Έκθεση σε βίαιο περιβάλλον**.
- **Χωρισμός ή διαζύγιο**. Είναι πιο σύνηθες στους άντρες.
- κ.ά.

Υψηλή συσχέτιση, επίσης, φαίνεται να εμφανίζεται μεταξύ των διαταραχών της διάθεσης και της οικονομικής κρίσης. Παρ'όλο που αντίστοιχες έρευνες σπανίζουν, υπάρχουν ενδείξεις από προηγούμενες οικονομικές κρίσεις στις Ηνωμένες Πολιτείες, την Ασία και την πρώην Σοβιετική Ένωση, καθώς επίσης και από ορισμένα στοιχεία που έχουν προκύψει από την παρούσα οικονομική κρίση που συνδέουν τέτοιες καταστάσεις με την ψυχοπαθολογία. Τα περισσότερα κρούσματα τέτοιων περιπτώσεων αφορούν την καταθλιπτική διαταραχή και την αυτοκτονία. Συγκεντρωτικά, 1 στους 10 ανθρώπους πάσχουν από μείζον καταθλιπτικό επεισόδιο, και σχεδόν 1 στους 5 έχει υποφέρει από κατάθλιψη κατά τη διάρκεια της ζωής του. Ακόμα και πριν την παρούσα οικονομική κρίση, ο Παγκόσμιος Οργανισμός Υγείας εκτιμούσε ότι η κατάθλιψη θα παρουσιάζει συνεχή άνοδο στη λίστα των παγκόσμιων επικρατέστερων ασθενειών, ανεβαίνοντας από την τρίτη θέση το 2004 με ποσοστό 4,3% επι του συνόλου, στην πρώτη θέση μέχρι το 2030, με ποσοστό 6,2% επι του συνόλου - ακολουθούμενη από ισχαιμική καρδιακή νόσο, τροχαία ατυχήματα, εγκεφαλοαγγειακές παθήσεις και χρόνια αποφρακτική πνευμονοπάθεια. Ο Π.Ο.Υ. διαπίστωσε ότι η κατάθλιψη είναι ήδη η κύρια αιτία απώλειας ενός υγιούς τρόπου ζωής στις γυναίκες μεταξύ 15-44 ετών.

Μερικές φορές, στις χειρότερες εκφάνσεις της ασθένειας, οι άνθρωποι που πάσχουν από κατάθλιψη μπορεί να αποτελέσουν κίνδυνο για τους γύρω τους. Έχουν καταγραφεί περιστατικά ανθρωποκτονιών, ακολουθούμενων συνήθως από αυτοκτονία. Τα περιστατικά μπορεί να έχουν στόχο μέσα στο οικογενειακό περιβάλλον αλλά και σε πιο ευρύ κοινό. Οι τρόποι που ένα καταθλιπτικό άτομο μπορεί να διαπράξει ανθρωποκτονία ποικίλουν. Έχουν καταγραφεί τροχαία ατυχήματα, πρόκληση πυρκαγιάς, πρόκληση πυροβολισμών κ.ά. Αυτοί οι άνθρωποι συνήθως χαρακτηρίζονται από μείζον καταθλιπτικό επεισόδιο.

## 1.2 Ιστορική αναδρομή

Αρχαιολογικά ευρήματα δείχνουν ότι οι διαταραχές της διάθεσης προβλημάτισαν το ανθρώπινο είδος πολύ νωρίς στην ιστορία του. Φυσικά οι μεταφράσεις, οι αιτίες και οι τρόποι αντιμετώπισης διαφέρουν πολύ με αυτές του σήμερα. Η κατάθλιψη για πολλά χρόνια αποδίδοταν σε πνεύματα και δαίμονες. Στις κοινωνίες που προηγήθηκαν της γραφής (σε Ευρώπη και Νότιο Αφρική), οι ανθρωπολόγοι πιστεύουν, πως τρυπούσαν το κρανίο των ψυχικά ασθενών με σκοπό να απελευθερωθούν από τα κακά πνεύματα.

Τον 4ο αιώνα π.Χ., ο Ιπποκράτης ήταν ο μόνος που δεν ήταν ικανοποιημένος με την εξήγηση των 'υπερφυσικών δυνάμεων' και υποστήριζε ότι όλη η αρχή των νόσων οφείλεται σε κάποια βιολογική δυσλειτουργία του ανθρώπινου σώματος. Έτσι, άρχισε να παρακολουθεί πολύ στενά ανθρώπους με κάποια ψυχιατρική νόσο και να καταγράφει τη συμπεριφορά τους. Μέσω αυτού κατέληξε στην πρωτότυπη ιδέα ότι οι περισσότερες ψυχικές διαταραχές οφείλονται σε κάποια ανώμαλη έκκριση του σώματος. Προσδιόρισε τέσσερις απ'αυτές τις εκκρίσεις: το αίμα, τη λέμφο, την κίτρινη και μαύρη χολή. Κάθε μία θεωρούταν υπεύθυνη για κάποια ανωμαλία σε περίπτωση λανθασμένης ποσότητας, σύστασης ή θερμοκρασίας στο σώμα. Παραπάνω αίμα στον οργανισμό προκαλούσε απότομες αλλαγές στη διάθεση, περίσσεια λέμφου προκαλούσε αδιαφορία και νωθρότητα, η κίτρινη χολή δημιουργούσε εκνευρισμό και τέλος η μαύρη χολή ήταν υπεύθυνη για την κατάθλιψη. Από αυτήν την ερμηνεία, μάλιστα, ο Ιπποκράτης ονόμασε την κατάθλιψη 'μελαγχολία', διότι πίστευε ότι την προκαλεί η μέλαινα χολή στο μυαλό. Για την θεραπευτική αντιμετώπιση, ο Ιπποκράτης, πρότεινε ανάπαυση, σωστή διατροφή, αποφυγή κατανάλωσης αλκοόλ, υποστήριξη των ψυχιατρικά ασθενών, αποχή από τη σεξουαλική δραστηριότητα, καθώς και εμετικά, καθαρτικά, ιαματικά λουτρά, ταξίδια και μουσική.

Τον 5ο αιώνα και ύστερα από την πτώση της Ρωμαϊκής Αυτοκρατορίας η επιστημονική σκέψη όσον αφορά τις ψυχικές ασθένειες και την κατάθλιψη εξασθένησε και πάλι. Τον Μεσαίωνα, η εξάπλωση του Χριστιανισμού επηρέασε την εξήγηση της ψυχικής αυτής ασθένειας και την απέδιδε στον διάβολο. Μάλιστα οι ψυχικά ασθενείς θεωρούταν μολυσμένοι και ικανοί να μεταδώσουν την 'τρέλα' τους. Οι τρόποι αντιμετώπισης περιελάμβαναν τον εξορκισμό αλλά και πιο βάναυσες τεχνικές όπως ο πνιγμός και το κάψιμο. Πολλοί από τους ασθενείς εκείνης της εποχής ζούσαν δεμένοι και κακοποιημένοι στα τότε 'φρενοκομεία'.

Τα πράγματα άρχισαν να ξεκαθαρίζουν κατα τη δεκαετία του 1790, όπου ένας Γάλλος ιατρός, ο Phillippe Pinel μερίμνησε για την κάλυψη των αναγκών των ψυχικά ασθενών και προσπάθησε να κατανοήσει τη συμπεριφορά τους μέσω της παρατήρησης και της μελέτης. Μόλις στις αρχές τους 19ου αιώνα οι νοσούντες άρχισαν να αντιμετωπίζονται ως ασθενείς πλέον.

Πολλές διάσημες προσωπικότητες φαίνεται να υπέφεραν από κατάθλιψη. Ένα βιβλικό πρόσωπο της Παλαιάς Διαθήκης, ο Ιώβ, φέρεται να είναι από τους πιο ευρέως γνωστούς καταθλιπτικούς. Αλλά και άλλες προσωπικότητες όπως ο Michelangelo γνωστός Ιταλός γλύπτης, ποιητής, ζωγράφος και αρχιτέκτονας, ο Martin Luther γνωστός ως θεμελιωτής των χριστιανικών δογμάτων και πρακτικών του Προτεσταντισμού, ο Arthur Schopenhauer γνωστός Γερμανός φιλόσοφος, η Sylvia Plath αμερικανίδα ποιήτρια και νοβελίστρια και άλλοι. Η σχέση μεταξύ της δημιουργικότητας και των διαταραχών της διάθεσης έχει αποτελέσει αντικείμενο μελέτης και ανάλυσης στον τομέα της ψυχολο-

γίας.

## 1.3 Σκοπός της μελέτης

Η παρούσα εργασία αποτελεί μια αναδρομική μελέτη παραγόντων που επιδρούν στη διάρκεια νοσοκομειακής νοσηλείας ασθενών με κατάθλιψη. Μελετήθηκε το σύνολο των ασθενών που εισήχθησαν στην κλινική Asklepios Fachklinikum Lübben το έτος 2012 με κύρια διάγνωση την κατάθλιψη, οι οποίοι χωρίστηκαν σε δύο ομάδες ανάλογα με τη διάρκεια νοσηλείας τους (28 ημέρες).

Ο σκοπός είναι η ανεύρεση παραγόντων που επηρεάζουν το χρόνο νοσηλείας των ασθενών. Ως μεταβλητές επιλέχθηκαν δημογραφικοί παράγοντες, παράγοντες που σχετίζονται με τη νόσο όπως αυτοκτονικότητα ή απόπειρες αυτόκτονίας, ψυχιατρική ή σωματική συνοσηρότητα, ψυχιατρική αγωγή και εμφάνιση ανεπιθύμητων ενεργειών από τη φαρμακοθεραπεία.

## 1.4 Ορισμός της κατάθλιψης

Με βάση το Oxford Dictionary η κατάθλιψη ορίζεται ως:

**Καταθλιπτική Διαταραχή, η.** *Μία κατάσταση διάθεσης λύπης, κατήφειας και πεσιμιστικού ιδεασμού, με έλλειψη ενδιαφέροντος ή ευχαρίστησης σε φυσιολογικά διασκεδαστικές δραστηριότητες συνδυασμένη, σε σοβαρές περιπτώσεις, με ανορεξία και συνεπώς απώλεια βάρους, αϋπνία, υπερυπνία, εξάντληση, αίσθημα ανάξιου ή ενοχής, ελαττωμένη ικανότητα σκέψης ή συγκέντρωσης ή επαναλαμβανόμενες σκέψεις θανάτου ή αυτοκτονίας. Εμφανίζεται σαν σύμπτωμα πολλών ψυχικών διαταραχών.*

## 1.5 Ορισμός της νοσηλείας

Με βάση το Λεξικό της Νέας Ελληνικής Γλώσσας, του Γ. Μπαμπινιώτη, η νοσηλεία ορίζεται ως:

**Νοσηλεία, η.** *Η νοσηλεία ορίζεται με δύο τρόπους:*

1. *Η ιατροφαρμακευτική περίθαλψη ασθενούς στο σπίτι ή, συνηθέστερα, σε νοσοκομείο.*

2. *Το χρονικό διάστημα στο οποίο παρέχεται η παραπάνω περίθαλψη σε ασθενή.*

## 1.6 Συμπτωματολογία

Το καταθλιπτικό επεισόδιο δεν έχει αντίκτυπο μόνο στη διάθεση του ατόμου αλλά σε πολλούς τομείς της καθημερινότητάς του. Μπορεί να επηρεάσει τα κίνητρα συμπεριφοράς, τη σκέψη, τη βιολογική και κινητική λειτουργία του. Τα συμπτώματα αυτά μπορεί να είναι πρωτοεμφανιζόμενα, αλλά μπορεί και να αποτελούν επιδείνωση προϋπάρχοντων προβλημάτων.

Ο όρος 'καταθλιπτική διαταραχή' περιλαμβάνει μία ομάδα διαταραχών που εντάσσονται στην ίδια έννοια. Η καταθλιπτική διαταραχή περιλαμβάνει τη μείζων καταθλιπτική διαταραχή, επίμονη καταθλιπτική διαταραχή (δυσθυμία), προεμμηνορροϊκή δυσφορική διαταραχή, καταθλιπτική διαταραχή από χρήση ουσιών/φαρμάκων, καταθλιπτική διαταραχή που οφείλεται σε άλλη ιατρική πάθηση, άλλες καταθλιπτικές διαταραχές. Οι κοινές επιπτώσεις των προαναφερθέντων διαταραχών είναι η παρουσία λύπης, το αίσθημα του κενού, η ενοχλημένη διάθεση συνδυασμένη με σωματικές και γνωστικές αλλαγές που επηρεάζουν σημαντικά τη φυσιολογική λειτουργία του ατόμου. Αυτό που διαφέρει ανάμεσα σε αυτά είναι η διάρκεια του επεισοδίο, η χρονική στιγμή που προσβάλλεται ο ασθενής, κ.ά.

Η μείζων καταθλιπτική διαταραχή αποτελεί την πιο κλασική κατάσταση στο σύνολο των καταθλιπτικών διαταραχών. Χαρακτηρίζεται από διακριτά επεισόδια με διάρκεια τουλάχιστον δύο εβδομάδων που περιλαμβάνουν σαφείς αλλαγές στη γνωστική λειτουργία και τις νευροεγκεφαλικές λειτουργίες του ατόμου. Είναι δυνατή η διάγνωση με βάση ένα μόνο επεισόδιο, αν και η διαταραχή αυτή είναι επαναλαμβανόμενη στις περισσότερες περιπτώσεις. Ιδιαίτερη προσοχή πρέπει να δίνεται στην οριοθέτηση της φυσιολογικής θλίψης και της οδύνης από ένα μείζον καταθλιπτικό επεισόδιο. Υπάρχουν περιστάσεις στη ζωή ενός ατόμου που μπορεί να του προκαλέσουν θλίψη χωρίς αυτό να αποτελεί κάποιο επεισόδιο καταθλιπτικής διαταραχής. Το πένθος, για παράδειγμα, μπορεί να προκαλέσει μεγάλη ταλαιπωρία αλλά δεν προκαλεί τυπικά επεισόδιο μείζονος καταθλιπτικής διαταραχής. Όταν, όμως, τυχαίνει να συμβαίνουν μαζί, τα συμπτώματα της κατάθλιψης και η λειτουργική βλάβη, τείνουν να είναι πιο σοβαρά. Γενικά, η κατάθλιψη που σχετίζεται με το πένθος τείνει να προσβάλλει άτομα με άλλες ευπάθειες και για την αποκατάστασή τους προτείνεται φαρμακευτική αγωγή.

Η πιο χρόνια μορφή κατάθλιψης, η επίμονη καταθλιπτική διαταραχή (δυσθυμία) μπορεί να διαγνωστεί όταν η διαταραχή της διάθεσης συνεχίζεται για τουλάχιστον δύο χρόνια στους ενηλίκους και ένα χρόνο στα παιδιά.

I. Με βάση λοιπόν το DSM-5, έχουν καταγραφεί εννέα σημεία συμπτωματολογίας για το μείζον καταθλιπτικό επεισόδιο. Χρειάζονται τουλάχιστον πέντε από αυτά (ή περισσότερα) για να γίνει η διάγνωση. Τα συμπτώματα θα πρέπει να έχουν διάρκεια τουλάχιστον δύο εβδομάδων και να παρουσιάζουν εμφανή αλλαγή από την προηγούμενη λειτουργία του ατόμου. Απαραίτητη κρίνεται η παρουσία ενός εκ των (1) καταθλιπτική διάθεση ή (2) απώλεια ενδιαφέροντος ή ευχαρίστησης (ανηδονία).

   (1) Καταθλιπτική διάθεση κατά το μεγαλύτερο μέρος της ημέρας, σχεδόν κάθε μέρα όπως υποδεικνύεται είτε από υποκειμενική αναφορά (π.χ. αίσθημα λύπης, κενού, απελπισίας), ή από παρατήρηση τρίτων (π.χ. παρατήρηση του υποκειμένου δακρυσμένου). (Σημείωση: Σε παιδιά και εφήβους μπορεί να παρατηρηθεί ευαιρέθιστη διάθεση.)

   (2) Σημαντικά μειωμένο ενδιαφέρον ή ευχαρίστηση σε όλες, ή σχεδόν όλες, τις δραστηριότητες σχεδόν καθ'όλη τη διάρκεια της ημέρας, κάθε μέρα (όπως υποδεικνύεται είτε από το υποκείμενο, ή από παρατήρηση τρίτων).

   (3) Σημαντική απώλεια βάρους ενώ το υποκείμενο δεν βρίσκεται σε δίαιτα ή αύξηση βάρους (π.χ. αλλαγή μεγαλύτερη από το 5% του σωματικού βάρους

σε ένα μήνα), ή μείωση ή αύξηση στην όρεξη σχεδόν κάθε μέρα. (Σημείωση: Στα παιδιά θεωρείται η αποτυχία της αναμενόμενης αύξησης βάρους.)

(4) Αϋπνία ή υπερυπνία σχεδόν κάθε μέρα.

(5) Ψυχοκινητική διέγερση ή επιβράδυνση σχεδόν κάθε μέρα (όχι απλώς υποκειμενικά συναισθήματα ανησυχίας και επιβράδυνσης αλλά, είναι παρατηρήσιμο και από τρίτους).

(6) Κούραση ή απώλεια ενέργειας σχεδόν κάθε μέρα.

(7) Αισθήματα αναξιότητας ή υπερβολικής, ή αδικαιολόγητης ενοχής (που μπορεί να είναι παραπλανητική) σχεδόν κάθε μέρα (π.χ. ενοχή επειδή αρρώστησε).

(8) Μειωμένη ικανότητα σκέψης ή συγκέντρωσης, ή αναποφασιστικότητα, σχεδόν κάθε μέρα (είτε από υποκειμενική υπόδειξη, ή παρατήρηση από τρίτους).

(9) Επαναλαμβανόμενες σκέψεις για θάνατο (όχι απλώς ο φόβος να πεθάνουν), επαναλαμβανόμενος αυτοκτονικός ιδεασμός χωρίς συγκεκριμένο πλάνο, ή απόπειρα αυτοκτονίας ή ένα συγκρεκριμένο σχέδιο για αυτοκτονία.

II. Τα συμπτώματα προκαλούν κλινικά σημαντική ενόχληση ή έκπτωση της κοινωνικής, επαγγελματικής, ή άλλων σημαντικών περιοχών της λειτουργικότητας.

III. Το επεισόδιο δεν οφείλεται στις άμεσες φυσιολογικές επιδράσεις μίας ουσίας ή σε μία άλλη ιατρική κατάσταση.

IV. Η εμφάνιση του μείζονος καταθλιπτικού επεισοδίο δεν μεταφράζεται καλύτερα από τη σχιζοσυναισθηματική διαταραχή, τη σχιζοφρένεια, τη σχιζοφρενική διαταραχή, την παραληρητική διαταραχή ή άλλο καθορισμένο ή μη καθορισμένο φάσμα σχιζοφρένειας και άλλες ψυχωσικές διαταραχές.

V. Δεν υπήρξε ποτέ κάποιο μανιακό επεισόδιο ή επεισόδιο υπομανίας.

## 1.7 Ταξινόμηση της νόσου

Το μείζον καταθλιπτικό επεισόδιο με βάση το αν εμφανίζεται μία φορά ή είναι επαναλαμβανόμενο, με την τρέχουσα σοβαρότητα και με την παρουσία ή μη ψυχωτικών χαρακτηριστικών κωδικοποιείται αναλόγως. Έτσι η ταξινόμηση με βάση τη Διεθνή Στατιστική Ταξινόμηση Νοσημάτων και Συναφών Προβλημάτων Υγείας (International Statistical Classification of Diseases and Related Health Problems - ICD-10) είναι η ακόλουθη:

| **Βαρύτητα** | **Μοναδικό επεισοδιο** | **Επαναλαμβανόμενα επεισόδια** |
|---|---|---|
| Ήπιο | F32.0 | F33.0 |
| Μέτριο | F32.1 | F33.1 |
| Σοβαρό | F32.2 | F33.2 |
| Με ψυχωσικά χαρ/κα | F32.3 | F33.3 |
| Σε μερική ύφεση | F32.4 | F33.41 |
| Σε πλήρη ύφεση | F32.5 | F33.42 |
| Απροσδιόριστο | F32.9 | F33.9 |

Στην παρούσα εργασία η κύρια διάγνωση περιλαμβάνει τέσσερις από τις παραπάνω περιπτώσεις. Οι ασθενείς που μελετήθηκαν έχουν διαγνωσθεί με μέτριας (F32.1, F33.1) και σοβαρής βαρύτητας καταθλιπτικό επεισόδιο (F33.1, F33.2).

**F32.1 - Μέτριας βαρύτητας καταθλιπτικό επεισόδιο:** Το άτομο με μέτριας βαρύτητας καταθλιπτικό επεισόδιο παρουσιάζει συνήθως σημαντικές δυσκολίες στην εξακολούθηση των κοινωνικών, εργασιακών και φυσικών δραστηριοτήτων του.

**F32.2 - Βαρύ καταθλιπτικό επεισόδιο χωρίς ψυχωσικά συμπτώματα:** Στο βαρύ καταθλιπτικό επεισόδιο, ο πάσχων συνήθως παρουσιάζει σημαντική δυσφορία ή ανησυχία, εκτός εάν η επιβράδυνση αποτελεί προέχων χαρακτηριστικό. Είναι πιθανό να κυριαρχούν η απώλεια της αυτοεκτίμησης ή τα συναισθήματα μηδαμινότητας ή ενοχής, ενώ η αυτοκτονία αποτελεί σαφή κίνδυνο σε ιδιαίτερα βαριές περιπτώσεις.

**F33.1 - Υποτροπιάζουσα καταθλιπτική διαταραχή, παρόν επεισόδιο μέτριας βαρύτητας:** Το άτομο με υποτροπιάζουσα καταθλιπτική διαταραχή με παρόν επεισόδιο μέτριας βαρύτητας χαρακτηρίζεται από επαναλαμβανόμενα επεισόδια κατάθλιψης, με το παρόν επεισόδιο να είναι μέτριας βαρύτητας, όπως στο F32.1 και χωρίς κάποιο ιστορικό μανίας.

**F33.2 - Υποτροπιάζουσα καταθλιπτική διαταραχή, παρόν επεισόδιο βαρύ με ψυχωσικά συμπτώματα:** Στην υποτροπιάζουσα καταθλιπτική διαταραχή με παρόν βαρύ επεισόδιο με ψυχωσικά συμπτώματα χαρακτηρίζεται από επαναλαμβανόμενα επεισόδια κατάθλιψης, με το παρόν επεισόδιο να είναι βαριάς μορφής, όπως στο F32.2 και χωρίς κάποιο ιστορικό μανίας.

## 1.8 Τρόποι αντιμετώπισης

Η αντιμετώπιση της Μείζονος Καταθλιπτικής Διαταραχής (ΜΚΔ) περιλαμβάνει τη σωστή διάγνωση και την εφαρμογή των κατάλληλων θεραπευτικών μεθόδων. Η έγκυρη διάγνωση προκύπτει από τα διαγνωστικά κριτήρια, όπως αυτά περιγράφονται στην υπο-

ενότητα της συμπτωματολογίας. Το πρώτο βήμα για την θεραπεία του νοσούντος είναι ο ίδιος να αποδεχθεί την κρισιμότητα της κατάστασής του και να αναγνωρίσει την αναγκαιότητα της αντιμετώπισης.

Ανάλογα με τη σοβαρότητα του εκάστοτε περιστατικού, η αντιμετώπιση της κατάθλιψης μπορεί να γίνει σε εξωνοσοκομειακό ή ενδονοσοκομειακό πλαίσιο. Στις περιπτώσεις βαριάς καταθλιπτικής διαταραχής, ο ασθενής μπορεί να χρήζει νοσηλείας. Τέτοιες μορφές κατάθλιψης μπορεί να είναι απειλητικές για τη ζωή του νοσούντος και συνήθως χαρακτηρίζονται από απόσυρση η οποία οδηγεί σε αφυδάτωση και ασιτία. Η αντιμετώπιση των βιολογικών επιπτώσεων καθώς και περιπτώσεων αυτοκτονικού ιδεασμού απαιτούν νοσηλεία σε ενδονοσοκομειακό πλαίσιο.

Οι στόχοι της θεραπείας της κατάθλιψης είναι τρεις:

- Πλήρης ύφεση των καταθλιπτικών συμπτωμάτων.
- Αποκατάσταση της ποιότητας ζωής.
- Πρόληψη της έξαρσης ή της υποτροπής.

Στις βασικές αρχές αντιμετώπισης της κατάθλιψης είναι επίσης και η διάγνωση, η έρευνα και η θεραπεία των συνυπάρχοντων παθολογικών καταστάσεων. Η μείζων καταθλιπτική διαταραχή είναι πολύ συχνό φαινόμενο στον γενικό πληθυσμό, που καθιστά αναμενόμενη τη συνύπαρξή της με άλλες διαταραχές, ψυχιατρικής ή παθολογικής φύσεως. Σε αυτές τις περιπτώσεις απαιτείται ξεχωριστή θεραπεία.

Υπάρχουν τρεις κύριες μορφές θεραπείας που εφαρμόζονται ξεχωριστά ή συνδυαστικά ανάλογα με την βαρύτητα του καταθλιπτικού επεισοδίου. Αυτές είναι:

**Ψυχοθεραπεία**. Η ψυχοθεραπεία χωρίζεται στην γνωσιακή-συμπεριφορική και στην διαπροσωπική. Οι ενδείξεις αποτελεσματικότητάς της φαίνεται να είναι περισσότερες όταν γίνεται οργανομένα και σε χρονικά περιορισμένο πλαίσιο. Στις περιπτώσεις ήπιας έως μέτριας βαρύτητας καταθλιπτικής διαταραχής έχει τα ίδια αποτελέσματα με τα αντικαταθλιπτικά φάρμακα και μπορεί να εφαρμοστεί ως μόνη θεραπεία.

**Φαρμακευτική αγωγή**. Τα αντικαταθλιπτικά φάρμακα, σε επίπεδο πρωτοβάθμιας περίθαλψης, είναι συνήθως ή θεραπείας που επιλέγεται πρώτη για την αντιμετώπιση των καταθλιπτικών συμπτωμάτων.

**Ηλεκτροσπασμοθεραπεία**. Η ηλεκτροσπασμοθεραπεία, αν και είναι μία μέθοδος που διχάζει ηθικά, παραμένει ως ένα ουσιαστικό εργαλείο για τη θεραπεία της κατάθλιψης. Είναι ένα μέσο με ταχύτερα αποτελέσματα από τις άλλες μεθόδους, καθώς δημιουργεί νευροχημικές αλλαγές στη σύσταση του σώματος που στοχεύουν στην ελάττωση πολλών συμπτωμάτων της κατάθλιψης.

Υπάρχουν και άλλες σωματικές θεραπείες που η αποτελεσματικότητά τους ακόμα ερευνάται. Τέτοιες θεραπείες μπορεί να είναι:

**Φωτοθεραπεία**. Η φωτοθεραπεία πραγματοποιείται με έκθεση του ασθενή σε φθορίζοντα φωτισμό για 30 λεπτά ημερισίως κατα τους χειμερινούς μήνες. Πρόκειται

για μια αποτελεσματική θεραπεία για την κατάθλιψη με εποχιακό μοτίβο. Οι μελέτες που έχουν διεξαχθεί όμως, έχουν περιορισμένη ισχύ καθώς δεν υπάρχει επαρκές δείγμα.

**Διακρανική μαγνητική διέγερση**. Η μέθοδος αυτή πριλαμβάνει τη διέγερση των φλοιικών νευρώνων με μαγνητική επαγωγή, χρησιμοποιώντας ένα σύντομο, υψηλής έντασης μαγνητικό πεδίο. Από μικρές μελέτες υπάρχουν ενθαρρυντικά αποτελέσματα και παρουσιάζεται ως πιθανή εναλλακτική της ηλεκτροσπασμοθεραπείας.

**Χειρουργικές θεραπείες**. Συνήθως χρησιμοποιείται ως θεραπεία σε ασθενείς με ανθεκτική στα φάρμακα κατάθλιψη. Ουσιαστικά, με τη χειρουργική επέμβαση προκαλείται διέγερση του πνευμονογαστρικού με ένα ηλεκτρικό σύρμα που τοποθετείται γύρω από το δεξιό πνευμονογαστρικό νεύρο στον αυχένα το οποίο συνδέεται με ένα διεγέρτη που τοποθετείται στο θώρακα. Η μέθοδος αυτή αποτελεί την ύστατη λύση σε περίπτωση αποτυχίας των άλλων μεθόδων καθώς έχει περιορισμένες ενδείξεις αποτελεσματικότητας και είναι επίπονη.

**Στέρηση ύπνου**. Σε μερικές περιπτώσεις η ολική στέρηση του ύπνου έχει εντυπωσιακά αποτελέσματα στη μείωση των καταθλιπτικών συμπτωμάτων. Η αντικαταθλιπτική δράση αυτής της μεθόδου όμως φαίνεται να έχει παροδικά αποτελέσματα και είναι αρκετά δύσκολη στην εφαρμογή της για παραπάνω από μία εβδομάδα.

Εκτός των κύριων μορφών θεραπείας υπάρχουν και άλλες συμπληρωματικές θεραπείες που προτείνονται αλλά δε συστήνονται ως μονοθεραπείες. Το ενδιαφέρον στρέφεται στα φυσικά σκευάσματα και στα συμπληρώματα διατροφής που χορηγούνται χωρίς συνταγή καθώς και σε παρεμβάσεις στον τρόπο ζωής τους ασθενούς. Παρ'όλα αυτά, οι τεχνικές αυτές βασίζονται σε αδημοσίευτες ενδείξεις, με ελάχιστες γνώσεις σχετικά με τη δοσολογία και με αβεβαιότητα όσον αφορά διαδικασία παρασκευής τους. Οι εναλλακτικές αυτές θεραπείες περιλαμβάνουν:

**Το χόρτο του St John**. Πρόκειται για το εκχύλισμα ενός φυτού που λέγεται Χόρτο του St John (Hypericum perforatum), το οποίο φαίνεται ότι ανεβάζει την διάθεση και βρίσκεται κυρίως στην κεντρική Ευρώπη.

**Ωμέγα-3 λιπαρά οξέα και ινοσιτόλη**. Με βάση κάποιες έρευνες που έδειχναν ότι οι καταθλιπτικοί ασθενείς παρουσιάζουν διαταραχές στα φωσφολιπίδια, πολλοί ερευνητές πρόσθεσαν λιπαρά οξέα στα κλασικά αντικαταθλιπτικά φάρμακα, σε ασθενείς που δεν είχαν αποτέλεσμα στη θεραπεία και παρατήρησαν ότι είχαν θετικά αποτελέσματα. Επιπλέον, τα Ωμέγα-3 λιπαρά οξέα φαίνεται να μπορούν να βοηθήσουν στην προστασία έναντι της υποτροπής, όταν προστεθούν στους κλασικούς σταθεροποιητές της διάθεσης. Τέλος, κλινικές μελέτες έδειξαν ότι η ινοσιτόλη μπορεί να είναι χρήσιμη στην κατάθλιψη και στις αγχώδεις διαταραχές.

**S-αδενοσυλομεθειόνη**. Πρόκειται για ένα φυσικό αμινοξύ που βρίσκεται σε όλα τα κύτταρα. Σε μία μετα-ανάλυση βρέθηκε μία συγκρίσιμη αντικαταθλιπτική αποτελεσματικότητα και σε μερικές περιπτώσεις οι ασθενείς που έπαιρναν τη συγκεκριμένη ουσία ανέπτυξαν μανία.

**Τρυπτοφάνη**. Η τρυπτοφάνη είναι ένα συστατικό το οποίο παρουσιάζεται περισσότερο ως συμπληρωματική θεραπεία σε ασθενείς με κατάθληψη, παρά ως αυτούσιο αντικαταθλιπτικό. Παρ'όλα αυτά, έχει αποσυρθεί από τις ΗΠΑ και πολλές Ευρωπαϊκές χώρες καθώς ενοχοποιήθηκε για εμφάνιση συνδρόμουν ηωσινοφιλικής μυαλγίας.

**Δεϋδροεπιανδροστερόνη**. Πρόκειται για μία πρόδρομη ουσία της ανδροστενοδιόνης, η οπόια μετατρέπεται σε οιστρογόνα και ανδρογόνα. Αν και υπάρχουν αναφορές για την αντικαταθλιπτική δράση της, οι παρενέργειες που δημιουργεί την καθιστούν ακατάλληλη.

**Φυσική άσκηση**. Η άσκηση μπορεί να εφαρμοστεί σαν συμπληρωματική θεραπεία με όλες σχεδόν τις αντικαταθλιπτικές θεραπείες, αν και εφόσον ο ασθενής είναι ικανός. Από μελέτες φαίνεται πως η άσκηση σαν τρόπος αντιμετώπισης της κατάθλιψης παρουσιάζει μικρότερα ποσοστά υποτροπής από κάποια φάρμακα. Επιπλέον, η ένταξη της φυσικής δραστηριότητας μπορεί να αποτρέψει την αύξηση του βάρους που μπορεί να προκαλέσει η φαρμακευτική αγωγή που παρέχεται στον ασθενή.

Συμπερασματικά, για τις περιπτώσεις ήπιας έως μέτριας βαρύτητας καταθλιπτικής διαταραχής υπάρχουν αρκετές ενδείξεις ότι η ψυχοθεραπεία είναι αποτελεσματική μέθοδος θεραπείας. Όταν μάλιστα η ψυχοθεραπεία συνδυάζεται με την φαρμακοθεραπεία αυξάνεται η αποτελεσματικότητα σε ασθενείς με σοβαρή, υποτροπιάζουσα ή χρόνια κατάθλιψη. Γενικά, αυτός ο τρόπος αντιμετώπισης της κατάθλιψης αποτελεί πρώτης επιλογής θεραπευτική παρέμβαση.

Στις περιπτώσεις όμως που η ψυχοθεραπεία και η φαρμακοθεραπεία ως τρόπος αντιμετώπισης δεν παρουσιάζει τα θεμητά αποτελέσματα οδηγούμαστε στις σωματικές θεραπείες που προαναφέρθηκαν και γίνονται αυστηρά σε ενδονοσοκομειακό πλαίσιο από εξειδικευμένους ιατρούς. Για τον περιορισμό των ανεπιθύμητων ενεργειών και για την καλύτερη συμμόρφωση του ασθενή, συνήθως προτείνεται η μονοθεραπεία, παρ'όλο που ο συνδυασμός των θεραπειών μπορεί να παρουσιάσει πιο άμεσα αποτελέσματα.

# ΚΕΦΑΛΑΙΟ 2

# ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΓΙΑ ΔΙΤΙΜΕΣ ΜΕΤΑΒΛΗΤΕΣ ΑΠΟΚΡΙΣΗΣ

Statistics is the grammar of science.

*Karl Pearson*

## 2.1 Εισαγωγή

Η στατιστική εκμάθηση αναφέρεται σε ένα τεράστιο σύνολο εργαλείων που χρησιμοποιούνται με σκοπό την κατανόηση δεδομένων. Ανάλογα το πεδίο της επιστήμης, γίνεται χρήση διαφορετικών μεθόδων προσαρμοσμένες στη φύση του εκάστοτε προβλήματος. Τα εργαλεία αυτά βρίσκουν εφαρμογή σε πολλούς και διαφορετικούς τομείς, όπως για παράδειγμα οι επιχειρήσεις, η ιατρική και η αστροφυσική.

Ένα "μαθηματικό μοντέλο" αποτελεί την περιγραφή ενός συστήματος με τη χρήση μαθηματικών εννοιών. Τα μοντέλα χωρίζονται σε ντετερμινιστικά και στοχαστικά. Στην πρώτη περίπτωση, τα αποτελέσματα καθορίζονται με ακρίβεια ενώ, στην τελευταία περιλαμβάνεται αβεβαιότητα από άγνωστους παράγοντες. Τα μοντέλα που αποτελούνται από στοχαστικά μέρη ονομάζονται στατιστικά μοντέλα.

## 2.2 Ιστορική αναδρομή

Στις απαρχές του 19ου αιώνα, οι Legendre και Gauss δημοσίευσαν διάφορα άρθρα που παρουσίαζαν τη **μέθοδο των ελάχιστων τετραγώνων** (*method of least squares*), όπου υλοποιούσαν την προγενέστερη μορφή, αυτού που σήμερα αποκαλείται, **γραμμική παλινδρόμηση** (*linear regression*). Η μέθοδος αυτή λειτουργούσε αρκετά καλά όταν καλούταν να προβλέψει ποσοτικές τιμές. Όταν, όμως το πρόβλημα αποτελούταν από διακριτές τιμές, δεν είχε τα επιθυμητά αποτελέσματα. Το 1936, ο Fisher πρότεινε τη **γραμμική διακριτική ανάλυση** (*linear discriminant analysis*), που είχε σκοπό την πρόβλεψη διακριτών τιμών όπως, η επιβίωση ή όχι ενός ασθενή. Στη δεκαετία του 1940,

διάφοροι συγγραφείς πρότειναν μια εναλλακτική μέθοδο, γνωστή ως **λογιστική παλινδρόμηση** (*logistic regression*). Στις αρχές του 1970, οι Nelder και Wedderburn δημιούργησαν τον όρο **γενικευμένα γραμμικά μοντέλα** (*generalized linear models - GLMs*), ώστε να εντάξουν σε μία οικογένεια όλες τις μεθόδους που περιελάμβαναν τόσο τη γραμμική, όσο και τη λογιστική παλινδρόμηση.

Προς τα τέλη του 1970, αναπτύχθηκαν πολλές άλλες τεχνικές ανάλυσης δεδομένων. Παρ'όλα αυτά, οι περισσότερες μέθοδοι ήταν γραμμικές, καθώς τα τεχνολογικά εργαλεία της εποχής, δεν επέτρεπαν την επεξεργασία μη-γραμμικών μεθόδων. Το 1980, οι ηλεκτρονικοί υπολογιστές ήταν πλέον έτοιμοι να διεργαστούν μη-γραμμικές τεχνικές. Στα μέσα του 1980, οι Brieman, Friedman, Olshen και Stone σύστησαν στην επιστημονική κοινότητα, τη **μέθοδο της ταξινόμησης** (*classification*), καθώς και τα **δέντρα παλινδρόμησης** (*regression trees*). Η ίδια ομάδα ήταν και οι πρώτοι που ένταξαν πρακτικές μεθόδους επιλογής μοντέλων όπως είναι η **επικύρωση μέσω της διασταύρωσης** (*cross-validation*).

Από τότε και έπειτα, η στατιστική επιστήμη αποτελεί βασικό εργαλείο σε όλο και περισσότερους τομείς και προοδεύει συνεχώς. Επιπλέον σύμμαχος της ραγδαίας αυτής εξέλιξής της αποτελούν και τα διαθέσιμα τεχνολογικά εργαλεία που γίνονται όλο και πιο φιλικά προς τους χρήστες τους.

## 2.3 Γενικευμένα γραμμικά μοντέλα (GLMs, Generalized Linear Models)

### 2.3.1 Εισαγωγή

Τα **γενικευμένα γραμμικά μοντέλα** (*generalized linear models - GLMs*) αποτελούν την πιο σημαντική οικογένεια στατιστικών μοντέλων. Το όνομά τους οφείλεται στο γεγονός ότι γενικεύουν τα κλασικά γραμμικά μοντέλα, με βάση την κανονική κατανομή. Συνοπτικά, τα GLM μετατρέπουν τη γραμμική παλινδρόμηση επιτρέποντας στο γραμμικό μοντέλο να συσχετίζεται με τη μεταβλητή απόκρισης μέσω μιάς **συνάρτησης σύνδεσης** (*link function*).

Στη στατιστική μοντελοποίηση, το ενδιαφέρον στρέφεται στο τι πληροφορία μπορεί να αντληθεί για τα συστηματικά μέρη από εμπειρικά δεδομένα που περιέχουν τυχαίες συνιστώσες. Γενικά, ένα μοντέλο, χωρίζεται στο συστηματικό μέρος και το μεταβλητό. Το συστηματικό μέρος περιγράφει τα πρότυπα του φαινομένου που μας ενδιαφέρει.

Γενικά, έστω μία ποσοτική μεταβλητή απόκρισης Y και k διαφορετικές επεξηγηματικές μεταβλητές, $X_1, X_2, ..., X_k$. Η σχέση μεταξύ της Y και των $\boldsymbol{X} = (X_1, X_2, ..., X_k)$ μπορεί να περιγραφεί ως:

$$y = f(\boldsymbol{X}) + \varepsilon \tag{2.1}$$

όπου $f$ είναι κάποια προσαρμοσμένη, αλλά άγνωστη, συνάρτηση των $\boldsymbol{X} = (X_1, X_2, ..., X_k)$ και $\varepsilon$ ένα τυχαίο σφάλμα, το οποίο είναι ανεξάρτητο των $X$, και έχει μέση τιμή ίση με μηδέν, $\mathrm{E}(\varepsilon) = 0$. Σε αυτόν τον τύπο, η $f$ αποτελεί τη συστηματική πληροφορία που παρέχουν τα X για το Y. Το μοντέλο αυτό ονομάζεται **μοντέλο παλινδρόμησης** (*regression model*).

Το πιο απλό μοντέλο που μελετάει τη σχέση μεταξύ δύο τυχαίων μεταβλητών X και Y, ονομάζεται **απλό γραμμικό μοντέλο** (*simple linear model*). Το μοντέλο αυτό,

αποτελείται από μία μόνο επεξηγηματική τυχαία μεταβλητή , και περιγράφεται από τη σχέση:

$$y = \beta_0 + \beta_1 + \varepsilon \tag{2.2}$$

όπου με $\varepsilon$ συμβολίζεται το τυχαίο σφάλμα, το οποίο ακολουθεί την Κανονική κατανομή, $N(0, \sigma^2)$, με μέση τιμή $E(\varepsilon) = 0$ και διακύμανση $\sigma^2$. Η τυχαία μεταβλητή Y καλείται **μεταβλητή απόκρισης ή εξαρτημένη μεταβλητή** (*response or dependent variable*) και η μεταβλητή X καλείται **επεξηγηματική ή ανεξάρτητη μεταβλητή** (*predictor or independent variable*).

Οι επεξηγηματικές ή ανεξάρτητες μεταβλητές συνήθως μετριούνται σε μία από τις τρεις κατηγορίες που ακολουθούν:

- **Ονομαστική ταξινόμηση**. Στη **δυαδική ή διωνυμική** ταξινόμηση (*binary, dichotomous or binomial*) η μεταβλητή χωρίζεται σε δύο κατηγορίες, όπως για παράδειγμα: άντρας, γυναίκα. Εάν υπάρχουν παραπάνω από δύο κατηγορίες η μεταβλητή καλείται **πολύτιμη ή πολυωνιμική** (*polychotomous, polytomous or multinomial*).

- **Ταξινόμηση διάταξης**. Σε αυτήν την περίπτωση υπάρχει κάποια κατάταξη που ακολουθείται από τις τιμές της μεταβλητής. Με άλλα λόγια, η μεταβλητή χωρίζεται σε επίπεδα, όπως για παράδειγμα: νέος, μέση ηλικία, γέρος.

- **Ποσοτική**. Η μεταβλητή εμπίπτει σε ένα συνεχές φάσμα. Είναι άμεσα μετρήσιμη καθώς εκφράζει τη μεταβολή των τιμών της κατα αριθμητική ποσότητα, όπως για παράδειγμα: κιλά, ηλικία κτλ.

Οι μεταβλητές που ανήκουν στις δύο πρώτες κατηγορίες ονομάζονται **κατηγορικές μεταβλητές** (*categorical variables*) και οι τιμές αυτών **μετρήσεις ή συχνότητες** (*counts or frequencies*) της κάθε κατηγορίας. Μία κατηγορική επεξηγηματική μεταβλητή ονομάζεται **παράγοντας** (*factor*) και οι κατηγορίες της, **επίπεδα** (*levels*). Για τις συνεχείς μεταβλητές, οι τιμές τους αποτελούν ποσοτικά μεγέθη. Η μέθοδος, λοιπόν, που ακολουθείται για την εκάστοτε στατιστική ανάλυση εξαρτάται από τον τύπο της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών.

Παραπάνω παρουσιάστηκε το γραμμικό μοντέλο στην απλούστερη μορφή του, με τη χρήση, μόλις μίας επεξηγηματικής μεταβλητής x. Έστω ένα πρόβλημα με μεταβλητή απόκρισης $Y$ και $k$ επεξηγηματικές μεταβλητές $\boldsymbol{X} = (X_1, X_2, ..., X_k)$. Επιπλέον, υπό την υπόθεση τυχαίου δείγματος $n$, $\boldsymbol{y} = (y_1, y_2, ..., y_n)$ είναι οι τιμές τις μεταβλητής απόκρισης και $\boldsymbol{x_i} = (x_{i1}, x_{i2}, ..., x_{ik}), i = 1, ..., n$ είναι οι τιμές των επεξηγηματικών μεταβλητών.

Αρχικά, έστω ότι η μεταβλητή απόκρισης είναι ποσοτική. Το μοντέλο που εκφράζει τη σχέση μεταξύ των επεξηγηματικών μεταβλητών και της μεταβλητής απόκρισης ονομάζεται **πολλαπλό γραμμικό μοντέλο** (*multiple linear model*), με $k + 2$ άγνωστες παραμέτρους $\boldsymbol{\beta'} = (\beta_0, \beta_1, \beta_2, ..., \beta_k)$, το οποίο δίνεται από τη σχέση:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \tag{2.3}$$

όπου:

- $y_i$, $i = 1, 2, ..., n$, , οι τιμές της μεταβλητής απόκρισης $y$.
- $x_{ij}$, $i = 1, 2, ..., n$, $j = 1, 2, ..., k$, οι τιμές για την $i$-οστή παρατήρηση της επεξηγηματικής μεταβλητής $X_j$.
- $\beta_0, \beta_1, \beta_2, ..., \beta_k$, οι άγνωστοι συντελεστές του μοντέλου.
- $\varepsilon_i$, $i = 1, 2, ..., n$, τα τυχαία σφάλματα, τα οποία είναι ανεξάρτητα και ισόνομα, με:
  - $E(\varepsilon_i) = 0$, για κάθε $i$.
  - $V(\varepsilon_i) = \sigma^2$, για κάθε $i$, δηλαδή ικανοποιούν την υπόθεση της ομοσκεδαστικότητας.
  - $cov(\varepsilon_i, \varepsilon_j) = 0$, για $i \neq j$, δηλαδή τα τυχαία σφάλματα είναι ασυσχέτιστα μεταξύ τους.

Το πολλαπλό γραμμικό μοντέλο μπορεί να περιγραφεί και υπο την μορφή πινάκων ως εξής:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.4}$$

όπου:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

είναι ένα $n \times 1$ διάνυσμα με τιμές της μεταβλητής απόκρισης $Y$,

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

είναι ένας $n \times p$ πίνακας, με $p = k + 1$, ο οποίος ονομάζεται **πίνακας σχεδιασμού** (*design or model matrix*),

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

το $p \times 1$ διάνυσμα των συντελεστών, και:

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

το $n \times 1$ διάνυσμα των τυχαίων σφαλμάτων που ακολουθούν την πολυματάβλητη Κανονική κατανομή με:

- $E(\varepsilon) = \mathbf{0}$ και
- $V(\varepsilon) = \sigma^2 \boldsymbol{I}$, όπου $\mathbf{I}$ ο $n \times n$ μοναδιαίος πίνακας.

Τελικά, από τα παραπάνω καταλήγουμε στο συμπέρασμα ότι η αναμενόμενη τιμή της μεταβλητής απόκρισης $y$ είναι η:

$$E(\boldsymbol{y}) = E(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{X\beta} \tag{2.5}$$

με πίνακα διασποράς - συνδιασποράς τον:

$$V(\boldsymbol{y}) = V(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = V(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}. \tag{2.6}$$

Για την εφαρμογή του παραπάνω μοντέλου με σκοπό τη στατιστική ανάλυση δεδομένων είναι απαραίτητο να ισχύουν οι προϋποθέσεις που αναφέρθηκαν προηγουμένως. Δηλαδή, τα τυχαία σφάλματα να είναι ασυσχέτιστα μεταξύ τους και να ακολουθούν την Κανονική κατανομή. Αυτό όμως, στον πραγματικό κόσμο, δεν συμβαίνει πάντα. Για αυτόν τον λόγο, αναπτύχθηκαν τα **Γενικευμένα Γραμμικά Μοντέλα**, που αποτελούν ένα ενοποιητικό πλαίσιο των περισσοτέρων και πιο ευρέως γνωστών στατιστικών τεχνικών.

### 2.3.2 Το μοντελο

Οι πιο συνηθισμένες οικογένειες κατανομών έχουν πυκνότητες που μπορούν να γραφούν από κοινού σε μία ειδική μορφή, αυτήν της **εκθετικής οικογένειας κατανομών**. Μία τυχαία μεταβλητή $Y$ ανήκει στην εκθετική οικογένεια κατανομών αν η συνάρτηση πυκνότητας πιθανότητας (για συνεχή $Y$) ή συνάρτηση μάζας πιθανότητας (για διακριτή $Y$) μπορεί να γραφεί στη μορφή:

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = exp\big[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)\big], \tag{2.7}$$

όπου οι συναρτήσεις $b(\bullet)$, $c(\bullet)$, $d(\bullet)$ θεωρούνται γνωστές. Η παράμετρος $\boldsymbol{\theta}$ ονομάζεται **φυσική ή κανονική** (*natural or canonical*). Πολλές γνωστές κατανομές όπως είναι η διωνυμική, η Κανονική, η Poisson και η εκθετική κατανομή ανήκουν στην εκθετική οικογένεια. Από εδώ και πέρα, με $\boldsymbol{\theta}$ θα συμβολίζονται οι παραμέτροι του μοντέλου. Για παράδειγμα, στην περίπτωση της γραμμικής παλινδρόμησης: $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$.

Συνεπώς, όταν διατίθενται παρατηρήσεις $\boldsymbol{y} = (y_1, ..., y_n)$, με βάση τη σχέση (2.7), η πιθανοφάνεια γράφεται:

$$
\begin{aligned}
f(\boldsymbol{y};\boldsymbol{\theta}) &= \prod_{i=1}^{n} exp\big[y_i b(\theta_i) + c(\theta_i) + d(y_i)\big] \\
&= exp\left[\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)\right].
\end{aligned} \tag{2.8}
$$

Είναι γνωστό ότι η αναμενόμενη μέση τιμή $\mu_i$ της εξαρτημένης μεταβλητής $y_i$, επηρεάζεται από τις συμμεταβλητές $\boldsymbol{x}_i$. Η αναμενόμενη τιμή $\mu_i$ ανήκει στην εκθετική οικογένεια κατανομών και δίνεται από τη σχέση:

$$E(y_i) = \mu_i = b'(\theta_i). \tag{2.9}$$

Η **γραμμική προβλέπουσα** (*linear predictor*) είναι μία γραμμική συνάρτηση ενός συνόλου σταθερών παραμέτρων και επεξηγηματικών μεταβλητών, της οποίας η τιμή χρησιμοποιείται για να προβλέψει το αποτέλεσμα μίας εξαρτημένης μεταβλητής. Δηλαδή, ισχύει:

$$\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}. \tag{2.10}$$

Η συνάρτηση αυτή συνδέεται γραμμικά με την αναμενόμενη μέση τιμή $\mu_i$ και έτσι μέσω της **συνάρτησης σύνδεσης** (*link function*) έχουμε το εξής αποτέλεσμα:

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \eta_i. \tag{2.11}$$

Σημειώνεται ότι, στην πραγματικότητα, η σχέση μεταξύ της αναμενόμενης μέσης τιμής $\mu_i$ και της $\boldsymbol{x}_i^T \boldsymbol{\beta}$ είναι μη-γραμμική. Μέσω της συνάρτησης σύνδεσης $g(\bullet)$ όμως, παράγεται το επιθυμητό αποτέλεσμα. Έτσι, η παραπάνω σχέση περιλαμβάνει γραμμικά και μη-γραμμικά μέρη μαζί.

### 2.3.3 Συνιστώσες του γενικευμένου γραμμικού μοντέλου

Όλα τα Γενικευμένα Γραμμικά Μοντέλα αποτελούνται από τρεις συνιστώσες. Αυτές είναι:

- **Το στοχαστικό μέρος**, το οποίο αποτελείται από τις τυχαίες εξαρτημένες μεταβλητές $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_k)$ και καθορίζει τη δεσμευμένη κατανομή πιθανότητας $(\boldsymbol{Y}|\boldsymbol{X})$ για αυτές. Η κατανομή αυτή θα πρέπει να είναι ίδια για όλες και να περιέχεται στην εκθετική οικογένεια κατανομών.

- **Το συστηματικό μέρος**, το οποίο αποτελείται από τις ανεξάρτητες μεταβλητές. Οι επεξηγηματικές μεταβλητές εισάγονται γραμμικά, ως προγνωστικοί δείκτες στη δεξιά πλευρά του μοντέλου.

- **Η συνάρτηση σύνδεσης**, η οποία εξηγείται αναλυτικά στην παρακάτω υποενότητα.

### 2.3.4 Συνάρτηση σύνδεσης (*link function*)

Όπως αναφέρθηκε παραπάνω, μέσω της συνάρτησης σύνδεσης $g(\bullet)$ μπορεί να συσχετιστεί η γραμμική προβλέπουσα με την αναμενόμενη τιμή $\mu_i$. Η συνάρτηση σύνδεσης $g(\bullet)$ πρέπει να είναι 1-1, διαφορίσιμη και μονότονη συνάρτηση. Συνήθως, σε όλες τις παρατηρήσεις χρησιμοποιείται η ίδια συνάρτηση. Στην πραγματικότητα, η συνάρτηση σύνδεσης είναι απλά ένα τέχνασμα που σκοπό έχει την απλοποίηση των αριθμητικών μεθόδων εκτίμησης παραμέτρων όταν το μοντέλο περιλαμβάνει ένα γραμμικό μέρος. Με αυτήν την κανονικοποίηση, λοιπόν, όλοι οι άγνωστοι παράμετροι της γραμμικής κατασκευής αποκτούν επαρκή στατιστικά στοιχεία, εάν και εφόσον η κατανομής ανήκει στην εκθετική οικογένεια κατανομών.

Συγκεντρωτικά, η εκθετική οικογένεια κατανομών περιλαμβάνει τις εξής συνήθεις κατανομές:

- Κατανομή Poisson,
- Διωνυμική κατανομή (binomial),
- Κανονική κατανομή και Λογαριθμο-κανονική κατανομή (normal and log-normal),
- Γάμμα κατανομή, όπως επίσης Εκθετική κατανομή, log-γάμμα και Pareto (gamma, exponential, log-gamma and Pareto),
- Αντίστροφη Γκαουσιανή (inverse Gaussian).

Αντίστοιχα, οι συναρτήσεις σύνδεσης, κανονικές ή μη, είναι οι:

| **Κατανομή** | **Συνάρτηση σύνδεσης** |
|---|---|
| Κανονική, Λογαριθμο-Κανονική | $\eta_i = g(\mu_i) = \mu_i$ (identity link) |
| Γάμμα, Εκθετική | $\eta_i = g(\mu_i) = \frac{1}{\mu_i}$ (reciprocal link) |
| Αντίστροφη Γκαουσιανή | $\eta_i = g(\mu_i) = \frac{1}{\mu_i^2}$ (quadratic inverse link) |
| Γαμμα, Εκθετική, Poisson | $\eta_i = g(\mu_i) = log(\mu_i)$ (log link) |
| Διωνυμική | $\eta_i = g(\mu_i) = log\big(\frac{\mu_i}{n-\mu_i}\big)$ (logit link) |
| Διωνυμική | $\eta_i = g(\mu_i) = log\big[-log\big(\frac{\mu_i}{n}\big)\big]$ (c-log log link) |
| Διωνυμική | $\eta_i = g(\mu_i) = \Phi^{-1}\big(\frac{\mu_i}{n}\big)$ (probit link) |

Η συνάρτηση σύνδεσης *identity* χρησιμοποιείται για δεδομένα που ακολουθούν την Κανονική κατανομή. Αντίστοιχα, οι συναρτήσεις *logit*, *probit* και *complementary log log* χρησιμοποιούνται για διωνυμικά δεδομένα. Η Poisson κατανομή χρησιμοποιεί τη συνάρτηση *log*, ενώ η Εκθετική και Γάμμα κατανομή την συνάρτηση *reciprocal*.

## 2.4 Γενικευμένα γραμμικά μοντέλα για δυαδικά δεδομένα

### 2.4.1 Εισαγωγή

Όπως αναφέρθηκε και προηγουμένως, ορισμένες φορές η μεταβλητή απόκρισης μπορεί να είναι κατηγορική και να αποτελείται από δύο μόνο κατηγορίες (δίτιμη ή δυαδική). Για παράδειγμα, ο θάνατος ενός ασθενούς (Ναι, Όχι), η φθορά ενός ελαστικού (Ναι, Όχι) κ.ά. Έστω, λοιπόν, μία δυαδική μεταβλητή απόκρισης $Z$ και τα δύο πιθανά αποτελέσματά της 1 ("επιτυχία") και 0 ("αποτυχία").

$$Z = \begin{cases} 0, & \text{"αποτυχία"} \\ 1, & \text{"επιτυχία"} \end{cases}.$$

Οι τιμές τις μεταβλητής αποτελούν μία αυθαίρετη κωδικοποίηση των δύο ενδεχομένων. Τα ενδεχόμενα εκφράζονται μέσω πιθανοτήτων και είναι αντίστοιχα $P(Z = 1) = \pi$ για την επιτυχία και $P(Z = 0) = 1 - \pi$ για την αποτυχία. Η κατανομή που ακολουθεί η τυχαία μεταβλητή $Z$ είναι η Bernoulli, με μέση τιμή $E(Z) = \pi$ και διασπορά $V(Z) = \pi(1 - \pi)$.

Έστω, $m$ παρόμοιες τυχαίες μεταβλητές $Z_1, Z_2, ..., Z_m$ με πιθανότητα επιτυχίας $P(Z_j = 1) = \pi_j$, για κάθε $j = 1, 2, ..., m$. Κάθε $j-$παρατήρηση, έχει συνάρτηση μάζας πιθανότητας ίση με:

$$f(z_j) = \pi_j^{z_j}(1 - \pi_j)^{1-z_j}, \text{για κάθε} j = 1, 2, ..., m z_j = 0, 1.$$

Τότε η από κοινού συνάρτηση μάζας πιθανότητας, θα είναι:

$$\prod_{j=1}^{m} \pi_j^{z_j}(1 - \pi_j)^{z_j} = exp\left[\sum_{j=1}^{m} z_j log\Big(\frac{\pi_j}{1 - \pi_j}\Big) + \sum_{j=1}^{m} log\big(1 - \pi_j\big)\right], \tag{2.12}$$

όπου είναι φανερό ότι ανήκει στην εκθετική οικογένεια κατανομών. Εάν $Z_1 \sim b(n_1, \pi), ..., Z_m \sim b(n_m, \pi)$ (με την ίδια πιθανότητα επιτυχίας $\pi$) είναι ανεξάρτητες τυχαίες μεταβλητές, τότε:

$$\sum_{j=1}^{m} Z_j \sim b\Big(\sum_{j=1}^{m} n_j, \pi\Big)$$

και ορίζεται η τυχαία μεταβλητή:

$$Y = \sum_{j=1}^{m} Z_j$$

που εκφράζει τον αριθμό των επιτυχιών σε $\sum_j n_j$ δοκιμές. Με σκοπό η μεταβλητή $Y$ να ακολουθεί τη Διωνυμική (*binomial*) κατανομή, $Y \sim b(n, \pi)$, θα πρέπει η πιθανότητα επιτυχίας $\pi$ να είναι ίδια σε κάθε δοκιμή και οι δοκιμές να είναι ανεξάρτητες μεταξύ τους. Αν επιτευχθεί αυτό, η συνάρτηση πιθανότητας της τ.μ. $Y$ θα είναι:

$$f(y) = \binom{n}{y}\pi^y(1 - \pi)^{n-y}, y = 0, 1, 2, ..., n,$$

που περιγράφει κατάλληλα κάθε μεταβλητή αυτής της φύσης. Η μέση τιμή της μεταβλητής $Y$ είναι $E(y) = n\pi$ και η διασπορά $V(Y) = n\pi(1-\pi)$.

Τέλος, στην γενική περίπτωση, για $n$ ανεξάρτητες τυχαίες μεταβλητές $Y_1, Y_2, ..., Y_n$ που ακολουθούν την Διωνυμική κατανομή, $Y_i \sim b(n_i, \pi_i)$, και εκφράζουν τον αριθμό των επιτυχιών σε $n$ διαφορετικές υποομάδες του πληθυσμού, η συνάρτηση πιθανοφάνειας θα είναι:

$$l(\pi_1, ..., \pi_n; y_1, ..., y_n) = \left[ \sum_{i=1}^{n} y_i log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i log(1-\pi_i) + log\binom{n_i}{y_i} \right]. \quad (2.13)$$

### 2.4.2 Γραμμικό μοντέλο πιθανοτήτων (*LPM, Linear Probability Model*)

Έστω, οι $n$ ανεξάρτητες τυχαίες μεταβλητές $Y_1, Y_2, ..., Y_n$ και οι $n$ διαφορετικές υποομάδες του πληθυσμού, όπως ορίστηκαν παραπάνω. Σκοπός είναι να περιγραφεί το ποσοστό επιτυχίας σε κάθε υποομάδα, όσον αφορά τα επίπεδα των μεταβλητών απόκρισης και τις διάφορες επεξηγηματικές μεταβλητές που περιγράφουν την κάθε ομάδα. Έτσι, εάν $E(Y_i) = n_i\pi_i$, ορίζεται η μεταβλητή $P_i = Y_i/n_i$ με μέση τιμή $E(P_i) = \pi_i$. Τότε οι πιθανότητες $E(P_i) = \pi_i$ μοντελοποιούνται ως εξής:

$$g(\pi_i) = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

όπου $\boldsymbol{x}_i$ είναι ένα διάνυσμα επεξηγηματικών μεταβλητών, με ψευδομεταβλητές για κατηγορικές μεταβλητές και ποσοτικά μεγέθη για συμμεταβλητές. Επίσης με $\boldsymbol{\beta}$ συμβολίζεται το διάνυσμα των παραμέτρων και $g(\bullet)$ είναι η κατάλληλη συνάρτηση σύνδεσης για τα δεδομένα.

Ο πιο απλός τρόπος μοντελοποίησης των πιθανοτήτων είναι το **γραμμικό μοντέλο** (*linear model*) που διατυπώνεται ως εξής:

$$\pi = \boldsymbol{x}^T \boldsymbol{\beta}.$$

Το παραπάνω το μοντέλο έχει ένα βασικό μειονέκτημα. Παρ'όλο που το $\pi$ εκφράζει πιθανότητα, οι προσαρμοσμένες τιμές $\boldsymbol{x}^T\boldsymbol{\beta}$ μπορεί να παίρνουν τιμές μικρότερες του μηδενός ή/και μεγαλύτερες της μονάδας. Πράγμα που έρχεται σε αντιπαράθεση με το διάστημα τιμών των πιθανοτήτων που περικλύεται στο κλειστό διάστημα [0,1].

### 2.4.3 Ο μετασχηματισμός logit (*logit link*)

Η ερώτηση που τίθεται, λοιπόν, είναι η εξής: Πώς μπορούμε να μοντελοποιήσουμε τη σχέση μεταξύ των $\pi(X) = P(Y = 1|X)$ και της ανεξάρτητης μεταβλητής $X$; Όπως αναφέρθηκε παραπάνω, το γραμμικό μοντέλο έχει ένα πολύ σημαντικό μειονέκτημα, ότι δεν μπορεί να περιορίσει τις προβλεπόμενες τιμές μέσα στο επιθυμητό διάστημα [0,1]. Έτσι, για να αποφευχθεί αυτό, πρέπει να βρεθεί μία συνάρτηση που να μοντελοποιεί κατάλληλα τις πιθανότητες $\pi(X)$ έτσι ώστε να δίνει αποτελέσματα μεταξύ του 0 και του 1 για όλες τις τιμές του $X$.

Πολλές συναρτήσεις καθιστούν εφικτό αυτόν τον στόχο. Μία απ'αυτές είναι η **λογιστική συνάρτηση** (*logistic function*), που για μία επεξηγηματική μεταβλητή X δίνεται από τον τύπο:

$$\pi(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{2.14}$$

Για την προσαρμογή του παραπάνω μοντέλου χρησιμοποιείται η μέθοδος μέγιστης πιθανοφάνειας. Στο επόμενο βήμα και μετά τους κατάλληλους χειρισμούς, η παραπάνω συνάρτηση γίνεται:

$$\frac{\pi(X)}{1-\pi(X)} = e^{\beta_0+\beta_1 X}.$$

Η ποσότητα στο αριστερό μέρος της εξίσωσης $\frac{\pi(X)}{1-\pi(X)}$ καλείται **συμπληρωματικές (ή σχετικές) πιθανότητες** (*odds*) και λαμβάνει τιμές στο διάστημα [0,∞). Οι συμπληρωματικές πιθανότητες χρησιμοποιούνται σε αγώνες ιπποδρομιών, αντί για τις κλασικές πιθανότητες, καθώς σχετίζονται πιο φυσικά με τη σωστή στρατηγική στοιχημάτων.

Παίρνοντας τον λογάριθμο, και στις δύο πλευρές, της εξίσωσης, ο τύπος καταλήγει στη συνάρτηση σύνδεσης logit (*logit link*), όπως παρουσιάζεται παρακάτω:

$$\eta_X = g(\pi(X)) = log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 X. \tag{2.15}$$

### 2.4.4 Άλλοι μετασχηματισμοί για δυαδικά δεδομένα

Εκτός από την συνάρτηση σύνδεσης logit, όπως αναφέρθηκε παραπάνω, υπάρχουν και άλλες συναρτήσεις σύνδεσης που ταιριάζουν σε δυαδικά δεδομένα. Ένα από τα πιο κλασικά μοντέλα που χρησιμοποιείται για τέτοιους σκοπούς ονομάζεται **μοντέλο probit** (*probit model*), με συνάρτηση σύνδεσης:

$$\eta_X = g(\pi(X)) = \Phi^{-1}\left(\frac{\pi(X)}{n}\right) \tag{2.16}$$

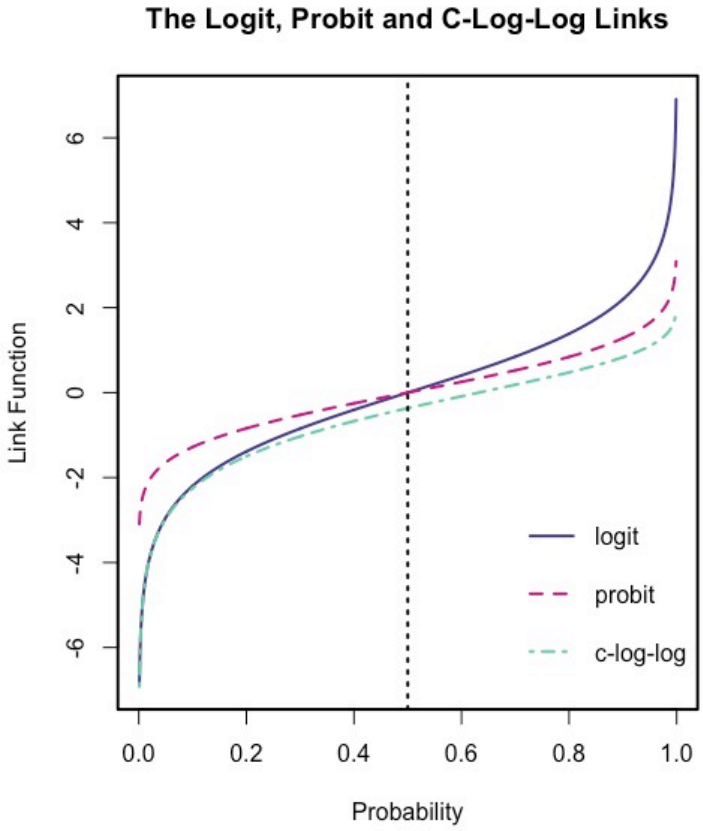
όπου με Φ συμβολίζεται η συνάρτηση κατανομής της τυποποιημένης Κανονικής κατανομής N(0,1). Το μοντέλο probit, λοιπόν, προέρχεται ύστερα από κατάλληλη μοντελοποίηση της Κανονικής κατανομής με αθροιστική συνάρτηση κατανομής (*cumulative probability function*):

$$\begin{aligned}\pi(X) &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{X} exp\left[-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2\right]ds \\ &= \Phi\left(\frac{X-\mu}{\sigma}\right).\end{aligned}$$

Έτσι,

$$\Phi^{-1}(\pi(X)) = \beta_0 + \beta_1 \tag{2.17}$$

όπου $\beta_0 = -\frac{\mu}{\sigma}$ και $\beta_1 = \frac{1}{\sigma}$. Επιπλέον, η συνάρτηση σύνδεσης $g(\bullet)$ είναι η αντίστροφη αθροιστική συνάρτηση Κανονικής κατανομής $\Phi^{-1}$. Το μοντέλο probit χρησιμοποιείται




Σχήμα 2.1: Σύγκριση των συναρτήσεων σύνδεσης logit, probit και complementary log-log.

σε διάφορους τομείς της βιολογίας και των κοινωνικών επιστημών, όπου έχει καλύτερη φυσική ερμηνεία.

Πολλά ακόμα μοντέλα χρησιμοποιούνται για τον ίδιο σκοπό. Ένα από τα πιο γνωστά αποτελεί και το **μοντέλο complementary log log** (*complementary log log model*) με συνάρτηση σύνδεσης:

$$\eta_X = g(\pi(X)) = log\big[-log(1-\pi(X))\big] \tag{2.18}$$

Το μοντέλο αυτό είναι παρόμοιο με αυτά των logit και probit, όταν το $\pi$ παίρνει τιμές κοντά στο 0.5, αλλά διαφέρει για τιμές του $\pi$ κοντά στο 0 και το 1. Έτσι, το complementary log log μοντέλο προτιμάται να χρησιμοποιείται όταν η πιθανότητα να συμβεί ένα γεγονός είναι πολύ μικρή ή πολύ μεγάλη.

Στο σχήμα (2.1), παρουσιάζονται οι τρεις συναρτήσεις σύνδεσης που αναφέρθηκαν παραπάνω. Πρόκειται για μία σύγκριση των μοντέλων σε σχέση με τις τιμές που μπορεί να πάρει η πιθανότητα $\pi$ για δυαδικά δεδομένα. Με μπλε χρώμα παριστάνεται η συνάρτηση σύνδεσης logit, με μωβ η συνάρτηση probit και τέλος με πράσινο χρώμα η συνάρτηση complementary log log. Όπως φαίνεται στο γράφημα, οι συναρτήσεις logit και probit παρουσιάζουν παρόμοια συμπεριφορά και είναι συμματρικές ως προς την τιμή πιθανότητας $\pi$= 0.5, όπου και ταυτίζονται. Αντιθέτως, το complementary log log

μοντέλο παρουσιάζει τελείως διαφορετική συμπεριφορά από τα άλλα δύο. Για τιμές κοντά στο 0 φαίνεται να ταυτίζεται με το μοντέλο logit ενώ, όσο η τιμή της πιθανότητας πλησιάζει στο 1, το μοντέλο complementary log log παρουσιάζει μία πιο αργή κίνηση προς το άπειρο ($\infty$) συγκριτικά με τα άλλα. Αυτή η ιδιόρυθμη συμπεριφορά είναι που καθιστά το μοντέλο, πολλές φορές, ακατάλληλο. Όλες οι συναρτήσεις είναι συνεχείς και αύξουσες.

### 2.4.5 Το πρόβλημα της υπερμεταβλητότητας

Όταν λοιπόν, στην εκάστοτε έρευνα υπάρχουν διαθέσιμα δυαδικά δεδομένα, πρέπει πρώτα να αντιμετωπιστούν κάποια θέματα που μπορεί να εμφανιστούν και έπειτα να ακολουθήσει η στατιστική τους ανάλυσή. Η πρώτη απόφαση που καλείται να πάρει ο ερευνητής είναι να επιλέξει την κατάλληλη συνάρτηση σύνδεσης από τις διαθέσιμες συναρτήσεις που αναφέρθηκαν στις προηγούμενες ενότητες. Ο Brown (1982) ανέπτυξε έναν έλεγχο για τη συνάρτηση σύνδεσης logit ο οποίος εφαρμόζεται σε κάποιο λογισμικό. Ο Aranda-Ordaz (1981) πρότεινε μία προσέγγιση η οποία εξετάζει μία πιο γενική οικογένεια συναρτήσεων σύνδεσης, που δίνονται από τον γενικό τύπο:

$$g(\pi, \lambda) = log\left[\frac{(1-\pi)^{-\lambda} - 1}{\lambda}\right]. \tag{2.19}$$

Εάν $\lambda = 1$, τότε $g(\pi) = log[\pi/(1-\pi)]$, που εκφράζει τη συνάρτηση σύνδεσης logit. Όσο το $\lambda \to 0$, τότε $g(\pi) \to log[-log(1-\pi)]$, το οποίο εκφράζει τη συνάρτηση σύνδεσης complementary log-log. Κατ'αρχήν, η ιδανική τιμή του $\lambda$ μπορεί να εκτιμηθεί από τα δεδομένα, αλλά η διαδικασία απαιτεί πολλά βήματα.

Το δεύτερο θέμα που συχνά αντιμετωπίζεται σε δεδομένα παρόμοιας φύσης αφορά τη μεταβλητότητα τους. Αν και το μοντέλο της λογιστικής παλινδρόμησης είναι το πιο δημοφιλές για δυαδικά δεδομένα, κατα την εφαρμογή του πολλές φορές παρουσιάζει προβλήματα. Το πρόβλημα που συχνά εμφανίζεται ονομάζεται **υπερμεταβλητότητα** ή **υπερδιασπορά** (*overdispersion*). Ενώ για το διωνυμικό μοντέλο καλείται και **επιπλέον διωνυμική μεταβλητότητα** (*extra binomial variation*). Σε αυτήν την περίπτωση, οι παρατηρήσεις $y_i$, οι οποίες φαίνεται να ακολουθούν τη Διωνυμική κατανομή, έχουν μεγαλύτερη διασπορά από τη θεωρητική $V(Y_i) = n_i\pi_i(1-\pi_i)$. Μια ένδειξη της παρουσίας της υπερμεταβλητότητας στα δεδομένα είναι όταν η ελεγχοσυνάρτηση deviance (βλ. ενότητα 2.5.7) είναι μεγαλύτερη των βαθμών ελευθερίας $df = n - p$, οι οποίοι εκφράζουν και την αναμενόμενη τιμή μιας τυχαίας μεταβλητής της κατανομής $\chi^2$. Αυτό μπορεί να οφείλεται σε ανεπαρκή προσδιορισμό του μοντέλου (π.χ. είτε έχουν παραλειφθεί επεξηγηματικές μεταβλητές, ή έχει επιλεχθεί λανθασμένη συνάρτηση σύνδεσης) ή κάποια πιο πολύπλοκη αιτία. Ωστόσο ο Lindsey (1999) θεωρεί καλύτερο κριτήριο για την ύπαρξη υπερμεταβλητότητας, η τιμή της ελεγχοσυνάρτησης deviance να είναι τουλάχιστον η διπλάσια των βαθμών ελευθερίας $df$, το οποίο ισοδυναμεί και με τη χρήση του κριτηρίου AIC.

Μερικοί από τους λόγους που οδηγούν σε αυτήν την συμπεριφορά, παρ'όλο που η επιλογή της κατανομής είναι σωστή, είναι οι εξής:

- Η επιλογή της συνάρτησης σύνδεσης δεν είναι σωστή.

- Δεν έχει καθοριστεί σωστά η γραμμική προβλέπουσα του μοντέλου (π.χ. κάποιες σημαντικές επεξηγηματικές μεταβλητές είναι εκτός του μοντέλου).
- Υπάρχουν ένα ή περισσότερα έκτροπα σημεία (outliers) στα δεδομένα.
- Ενδέχεται οι τ.μ. $Y_i$ να μην είναι ανεξάρτητες.

Υπάρχουν φυσικά και οι περιπτώσεις που δεν έχει επιλεχθεί η κατάλληλη κατανομή, με αποτέλεσμα να εμφανιστεί το πρόβλημα της υπερμεταβλητότητας. Η αιτία αυτή μπορεί να αντιμετωπιστεί αν χρησιμοποιηθεί, κάποια άλλη κατανομή, πιο κατάλληλη, η οποία να δίνει τη δυνατότητα για μεγαλύτερη μεταβλητότητα από την προεπιλεγμένη κατανομή. Για παράδειγμα, μία τέτοια κατανομή αποτελεί η Αρνητική Διωνυμική κατανομή, η οποία επιτρέπει μεγαλύτερη διακύμανση από την κατανομή Poisson. Ένα μοντέλο που έχει οριστεί ότι ακολουθεί την Διωνυμική κατανομή επεκτείνεται μέσω μίξης με την κατανομή Βήτα, δίνοντας την κατανομή Βήτα-Διωνυμική με διασπορά μεγαλύτερη της Διωνυμικής. Η κατανομή αυτή όμως, δεν ανήκει στην εκθετική οικογένεια και επομένως οι παράμετροι δεν εκτιμώνται με τη μέθοδο της μέγιστης πιθανοφάνειας στο πλαίσιο των γενικευμένων γραμμικών μοντέλων, αλλά με μία άλλη μέθοδο που ονομάζεται **Quasi-πιθανοφάνεια**. Εάν, λοιπόν, στο μοντέλο συμπεριληφθεί ένας επιπλέον παράγοντας $\phi$, η διασπορά της μίξης αυτής εκφράζεται από τη σχέση:

$$V(Y_i) = n_i\pi_i(1-\pi_i)\phi = n_i\pi_i(1-\pi_i)[1+(n_i-1)\psi] \tag{2.20}$$

όπου $\psi > 0$ είναι μια άγνωστη παράμετρος κλίμακας.

## 2.5 Πολλαπλή λογιστική παλινδρόμηση

### 2.5.1 Εισαγωγή

Σε αυτήν την ενότητα θα συζητηθούν προσεγγίσεις που χρησιμοποιούνται, υπό την υπόθεση ότι η μεταβλητή απόκρισης $Y$ είναι κατηγορική, και πιο συγκεκριμένα, δυαδική μεταβλητή. Η μεθοδολογία που χρησιμοποιείται για την πρόβλεψη τέτοιων μεταβλητών ονομάζεται **ταξινόμηση** (*classification*). Η διαδικασία πρόβλεψης της τιμής μίας κατηγορικής μεταβλητής για μία παρατήρηση μπορεί να χαρακτηριστεί και ως *ταξινόμηση* της συγκεκριμένης παρατήρησης σε κάποια κατηγορία ή κλάση της μεταβλητής απόκρισης. Οι μέθοδοι αυτοί, υπολογίζουν την πιθανότητα της κάθε κατηγορίας της εξαρτημένης μεταβλητής, και την χρησιμοποιούν ως βάση για την ταξινόμηση. Υπάρχουν πολλές τέτοιες τεχνικές που μπορούν να μοντελοποιήσουν μία κατηγορική μεταβλητή απόκρισης. Η συγκεκριμένη μελέτη ασχολείται με τον πιο διαδεδομένο τρόπο ανάλυσης, τη **λογιστική παλινδρόμηση** (*logistic regression*).

Το μοντέλο της λογιστικής παλινδρόμησης επικεντρώνει την προσοχή του στην μοντελοποίηση των συμπληρωματικών πιθανοτήτων (*odds*) ενός αποτελέσματος και αυτό είναι που το κάνει τόσο ελκυστικό. Οι πιθανότητες λαμβάνουν μέρος στην καθημερινότητα του ανθρώπου με πολλούς τρόπους, είτε συζητώντας για τυχερά παιχνίδια και αθλητικά, είτε αφορούν βιολογικούς και ιατρικούς σκοπούς, ή σχεδόν οτιδήποτε άλλο σχετίζεται με την καθημερινή ζωή. Η λογιστική παλινδρόμηση είναι ένα χρήσιμο εργαλείο για όλους αυτούς τους τομείς επειδή είναι ιδανικό για τον προσδιορισμό, τον

διαχωρισμό και την ομαδοποίηση διαφορετικών ειδών υποπληθυσμών. Το συγκεκριμένο μοντέλο χρησιμοποιείται στις περιπτώσεις που η μεταβλητή απόκρισης είναι δίτιμη, δηλαδή είναι αποτέλεσμα μιάς διαδικασίας *Bernoulli* που αποτελείται από αποτυχία/επιτυχία. Παραδείγματα τέτοιων μεταβλητών είναι: η αποτελεσματικότητα ενός φαρμάκου ή όχι (Ναι/Όχι), αν ένας ασθενής απεβίωσε ή ζει (Ναι/Όχι), κ.ά.

### 2.5.2 Το μοντέλο

Έστω $k$ ανεξάρτητες τυχαίες μεταβλητές, $\boldsymbol{X} = (X_1, X_2, ..., X_k)$ και έστω $Y$ μία δίτιμη μεταβλητή απόκρισης (0 ή 1). Η μεταβλητή $Y$ εξαρτάται από τις επεξηγηματικές μεταβλητές $\boldsymbol{X}$. Η σχέση αυτή εκφράζεται μέσω της εξαρτησης της πιθανότητας επιτυχίας $\pi$ απο τις ανεξάρτητες μεταβλητές $\boldsymbol{X}$. Έτσι, κατασκευάζεται το **μοντέλο της λογιστικής παλινδρόμησης** (*logistic regression model*) και εκφράζεται από τη σχέση:

$$\eta_{\boldsymbol{X}} = g(E(Y_{\boldsymbol{X}})) = g(\mu_{\boldsymbol{X}}) = \boldsymbol{X}^T\boldsymbol{\beta} \tag{2.21}$$

με δομή όπως περιγράφεται παρακάτω:

- $\eta_{\boldsymbol{X}} = g(\mu_{\boldsymbol{X}}) = log\left(\frac{\pi_{\boldsymbol{X}}}{1-\pi_{\boldsymbol{X}}}\right) = logit(\pi_{\boldsymbol{X}}) = \boldsymbol{X}^T\boldsymbol{\beta}$ (*συνάρτηση σύνδεσης logit*)
- $Y_{\boldsymbol{X}} \sim B(\mu_{\boldsymbol{X}})$, ($n_{\boldsymbol{X}} = 1$, για δυαδικά δεδομένα)
- ανεξαρτησία μεταξύ των παρατηρήσεων $Y_{\boldsymbol{X}}$.

Αντιστρέφοντας την παραπάνω συνάρτηση σύνδεσης, προκύπτει η σχέση:

$$\pi_{\boldsymbol{X}} = \frac{e^{\eta_{\boldsymbol{X}}}}{1 + e^{\eta_{\boldsymbol{X}}}}$$

όπου το πεδίο τιμών της κυμαίνεται μεταξύ του $0 < \pi_{\boldsymbol{X}} < 1$.

Συνδυάζοντας τα παραπάνω, το μοντέλο, για κάθε $i$-παρατήρηση, γράφεται ως εξής:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}, i = 1, 2, ..., n, \tag{2.22}$$

όπου η πιθανότητα επιτυχίας εκφράζεται από την σχέση:

$$\pi_i = \pi_{x_i} = \frac{e^{\beta_0+\beta_1 x_{i1}+...+\beta_k x_{ik}}}{1 + e^{\beta_0+\beta_1 x_{i1}+...+\beta_k x_{ik}}}$$

και η αναμενόμενη μέση τιμή, από τη σχέση:

$$E(Y_i) = \pi_i = \frac{e^{\boldsymbol{X}_i\boldsymbol{\beta}}}{1 + e^{\boldsymbol{X}_i\boldsymbol{\beta}}}$$

Όπως έχει αναφερθεί και σε προηγούμενη ενότητα, η συνάρτηση logit αποτελεί την κανονική συνάρτηση σύνδεσης της κατανομής Bernoulli και χρησιμοποιείται ευρέως για δεδομένα παρόμοιας φύσης.

### 2.5.3 Εκτίμηση των συντελεστών του μοντέλου

Στο μοντέλο της λογιστικής παλινδρόμησης που παρουσιάστηκε, οι συντελεστές $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_k)$ είναι άγνωστοι και πρέπει να εκτιμηθούν βάσει των δεδομένων. Όπως με όλα τα γενικευμένα γραμμικά μοντέλα, με σκοπό την προσαρμογή του μοντέλου στα δεδομένα, χρησιμοποιείται η **μέθοδος μέγιστης πιθανοφάνειας** (*method of maximum likelihood*). Η μέθοδος αυτή είναι η πιο συνήθης προσέγγιση που χρησιμοποιείται ώστε να προσαρμοστούν πολλά μη-γραμμικά μοντέλα. Στην περίπτωση της γραμμικής παλινδρόμησης εφαρμόζεται η μέθοδος ελάχιστων τετραγώνων, που στην πραγματικότητα αποτελεί μία ειδική περίπτωση της μέγιστης πιθανοφάνειας. Έτσι, η συνάρτηση πιθανοφάνειας για παρατηρήσεις $y_1, y_2, ..., y_n$ και επεξηγηματικές μεταβλητές $\boldsymbol{x}_i^T = (1, x_{i1}, ..., x_{ik})$, γράφεται ως:

$$L(\boldsymbol{\beta}) = L(\boldsymbol{\pi}; \boldsymbol{y}) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \tag{2.23}$$

.

Οι εκτιμήσεις για τους συντελεστές $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, .., \beta_k)$, προκύπτουν από τη μεγιστοποίηση του λογαρίθμου της συνάρτησης πιθανοφάνειας. Η πιθανοφάνεια εξαρτάται από τις, άγνωστες, πιθανότητες επιτυχίας $\pi_i$, οι οποίες με τη σειρά τους εξαρτώνται από τους συντελεστές $\boldsymbol{\beta}$ μέσω της σχέσης (2.23). Έτσι, για να προκύψει η συνάρτηση πιθανοφάνειας συναρτήσει των $\boldsymbol{\beta}$, ακολουθείται η παρακάτω διαδικασία:

$$\begin{aligned} l = logL(\boldsymbol{\beta}) = & \\ &= \sum_{i=1}^{k} \Big[ y_i log(\pi_i) + (1 - y_i) log(1 - \pi_i) \Big] \\ &= \sum_{i=1}^{k} \Big[ y_i log\Big(\frac{\pi_i}{1 - \pi_i}\Big) + log(1 - \pi_i) \Big] \\ &= \sum_{i=1}^{k} \Big[ y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - log(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}) \Big]. \end{aligned} \tag{2.24}$$

Παραγωγίζοντας τώρα την σχέση (2.24) ως προς τους συντελεστές $\boldsymbol{\beta}$, προκύπτει:

$$\begin{aligned} \frac{\partial logL(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{k} y_i x_{ij} - \sum_{i=1}^{k} x_{ij} e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} (1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}})^{-1}, j = 1, ..., k \\ &= \sum_{i=1}^{k} \Big[ y_i - e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} (1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}) \Big] x_{ij} \\ &= \sum_{i=1}^{n} (y_i - \pi_i) x_{ij}. \end{aligned} \tag{2.25}$$

Τελικά, για τον υπολογισμό των εκτιμητριων μέγιστης πιθανοφάνειας $\beta_j$ πρέπει να ικανοποιούνται οι εξισώσεις:

$$\begin{aligned} \sum_{i=1}^{k} (y_i - \hat{\pi}_i) x_{ij} = \sum_{i=1}^{k} (y_i - \hat{\mu}_i) x_{ij} = 0, j = 0, 1, 2, ..., k \\ \Rightarrow \boldsymbol{X}^T (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = \boldsymbol{0}, \end{aligned} \tag{2.26}$$

όπου:

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_n \end{bmatrix}$$

με $\mu_i = \pi_i$, $i = 1, 2, ..., n$ συμβολίζεται η μέση τιμή για κάθε συνιστώσα του τυχαίου δείγματος. Δηλαδή, εξισώνοντας τις μερικές παραγώγους με μηδέν προκύπτει ένα σύστημα από $k+1$ εξισώσεις το οποίο λύνεται με επαναληπτικές μεθόδους. Η διαδικασία επίλυσης δεν μπορεί να είναι διαφορετική διότι οι εξισώσεις είναι μη-γραμμικές ως προς τα $\hat{\boldsymbol{\beta}}$, εφόσον $log\hat{\mu}_i = exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$. Μία τέτοια διαδικασία μπορεί να είναι αυτή των **σταθμισμένων ελάχιστων τετραγώνων** (*WLS, Weighted Least Squares*). Η λύση του συστήματος αυτού δίνει τις τιμές των εκτιμητριών μέγιστης πιθανοφάνειας $\hat{\boldsymbol{\beta}}$ και η αντίστοιχη προσαρμοσμένη τιμή του αριθμού των επιτυχιών για την παρατήρηση $i$, είναι: $\hat{\mu}_i = \hat{\pi}_i$. Οι ποσότητες $e^{\hat{\beta}_j}$ εκφράζουν την αναμενόμενη πολλαπλασιαστική μεταβολή της $y$ για μία μονάδα αύξησης της αντίστοιχης επεξηγηματικής μεταβλητής $x_{ij}$, διατηρώντας τις υπόλοιπες συμμεταβλητές σταθερές.

### 2.5.4 Ερμηνεία των συντελεστών του μοντέλου

Σε σχέση με άλλα μοντέλα για δυαδικά δεδομένα, η λογιστική παλινδρόμηση μας δίνει τη δυνατότητα ερμηνείας των τιμών των συντελεστών $\hat{\boldsymbol{\beta}}$, καθώς και των διαστημάτων εμπιστοσύνης τους. Όπως έχει αναφερθεί και σε προηγούμενες ενότητες, η λογιστική παλινδρόμηση χρησιμοποιείται για την εξαγωγή συμπερασμάτων σε πολλές και διαφορετικές περιπτώσεις, όπου το αποτέλεσμα έχει δυαδική μορφή. Με σκοπό τη λήψη του βέλτιστου μοντέλου, εξετάζεται η σημασία της κάθε μεταβλητής.

Εφόσον πραγματοποιηθεί η εκτίμηση των παραμέτρων $\hat{\boldsymbol{\beta}}$, με τη χρήση της μεθόδου μέγιστης πιθανοφάνειας, οι παράμετροι αυτοί μπορούν να εκφραστούν μέσω των συμπληρωματικών πιθανοτήτων. Έτσι, η σχέση μεταξύ της προσαρμοσμένης πιθανότητας απόκρισης $\hat{\boldsymbol{\pi}}$ και των τιμών $x_0, x_1, x_2, ..., x_k$ των επεξηγηματικών μεταβλητών, εκφράζεται ως:

$$\hat{\pi} = \frac{e^{\boldsymbol{x}^T\hat{\boldsymbol{\beta}}}}{1 + e^{\boldsymbol{x}^T\hat{\boldsymbol{\beta}}}}$$

Η σχέση αυτή, μέσω του **λόγου των συμπληρωματικών ή σχετικών πιθανοτήτων** (*odds*) και της συνάρτησης σύνδεσης $logit(\hat{\pi})$, εκφράζεται ισοδύναμα ως:

$$\begin{aligned} log(odds) &= log\Big(\frac{\hat{\pi}}{1-\hat{\pi}}\Big) = \boldsymbol{x}^T\hat{\boldsymbol{\beta}} \\ \Rightarrow odds &= \frac{\hat{\pi}}{1-\hat{\pi}} = e^{\boldsymbol{x}^T\hat{\boldsymbol{\beta}}} = e^{\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+...+\hat{\beta}_kx_k}. \end{aligned} \tag{2.27}$$

Η ποσότητα $e^{\hat{\beta}_j}$ εκφράζει τον παράγοντα επί τον οποίον πολλαπλασιάζεται η σχετική πιθανότητα πραγματοποίησης του γεγονότος "επιτυχία", όταν η αντίστοιχη ανεξάρτητη μεταβλητή $X_j$ αυξηθεί κατά μία μονάδα, δεδομένου ότι οι υπόλοιπες συμμεταβλητές παραμένουν σταθερές. Από τα odds προκύπτει ότι αν ο εκτιμημένος συντελεστής $\hat{\beta}_j$ είναι θετικός ($\hat{\beta}_j > 0$), τότε ο παράγοντας $e^{\hat{\beta}_j}$ είναι μεγαλύτερος της μονάδας ( $e^{\hat{\beta}_j} > 1$ ). Ενώ, αν ο εκτιμημένος συντελεστής $\hat{\beta}_j$ είναι αρνητικός ($\hat{\beta}_j < 0$), τότε ο παράγοντας

$e^{\hat{\beta}_j}$ είναι μικρότερος της μονάδας ($e^{\hat{\beta}_j} < 1$). Το πρώτο γεγονός ($\hat{\beta}_j > 0$) σημαίνει ότι το $odds = \frac{\hat{\pi}}{1-\hat{\pi}}$ αυξάνεται με την αύξηση της $X_j$, αντίθετα αν $\hat{\beta}_j < 0$ τότε η σχετική πιθανότητα μειώνεται με την αύξηση της $X_j$.

Ο λόγος των συμπληρωματικών πιθανοτήτων (odds) για δυαδικά δεδομένα, ουσιαστικά εκφράζει την πιθανότητα πραγματοποίησης του γεγονότος "επιτυχία" ($Y = 1$) προς την πιθανότητα πραγματοποίησης του γεγονότος "αποτυχία" ($Y = 0$). Με άλλα λόγια, η πιθανότητα επιτυχίας εκφράζεται μέσω του γινομένου του λόγου των συμπληρωματικών πιθανοτήτων με την πιθανότητας αποτυχίας. Παραδείγματος χάρην, αν ο λόγος των συμπληρωματικών πιθανοτήτων πάρει την τιμή 2, $odds = \frac{\hat{\pi}}{1-\hat{\pi}} = 2$, τότε η αντίστοιχη ερμηνεία είναι ότι "η πιθανότητα η τυχαία μεταβλητή $Y$ να λάβει την τιμή 1, είναι δύο φορές μεγαλύτερη από την πιθανότητα να λάβει την τιμή 0". Αντίστοιχα, αν ο λόγος των συμπληρωματικών πιθανοτήτων είναι ίσος με 1, άμεσα προκύπτει ότι οι πιθανότητες πραγματοποίησης των δύο γεγονότων είναι ίσες, δηλαδή $\hat{\pi} = 1 - \hat{\pi} = 1/2$. Εάν, τώρα ο λόγος των συμπληρωματικών πιθανοτήτων είναι μικρότερος της μονάδας ($odds < 1$) η ερμηνεία διαφέρει. Για παράδειγμα, αν $odds = 0.7$, η αντίστοιχη ερμηνεία είναι ότι η πιθανότητα επιτυχίας ισούται με το 70% της πιθανότητας αποτυχίας. Επίσης, θα μπορούσε να ερμηνευθεί και ως ότι η πιθανότητα επιτυχίας είναι 30% μικρότερη από την πιθανότητα επιτυχίας.

Γενικεύοντας τα παραπάνω παραδείγματα, θεωρώντας $odds = \frac{\hat{\pi}}{1-\hat{\pi}} = \alpha$, οι ερμηνείες συνοψίζονται ως εξής:

- αν $\alpha > 1$, η πιθανότητα επιτυχίας είναι $(\alpha - 1)100\%$ φορές μεγαλύτερη από την πιθανότητα αποτυχίας,
- αν $\alpha < 1$, η πιθανότητα επιτυχίας είναι $(1 - \alpha)100\%$ φορές μικρότερη από την πιθανότητα αποτυχίας,

Οι παράμετροι της παλινδρόμησης μπορούν, επίσης, να εκφραστούν και μέσα από το **λόγο του λόγου των συμπληρωματικών πιθανοτήτων**, δηλαδή από τον **λόγο των odds** (*odds ratio*). Έστω ένα μοντέλο με δύο συμμεταβλητές $x_1$ και $x_2$. Γενικώς ο λόγος των odds ενός ατόμου με τιμές συμμεταβλητών $x_1$ σε σχέση με ένα άτομο με τιμές $x_2$ των ίδιων συμμεταβλητών προκύπτει ως:

$$\begin{aligned} \frac{odds_1}{odds_2} = \frac{\frac{\hat{\pi}_1}{1-\hat{\pi}_1}}{\frac{\hat{\pi}_2}{1-\hat{\pi}_2}} &= \frac{odds(y = 1 \mid x_1)}{odds(y = 1 \mid x_2)} \\ &= \frac{exp(\boldsymbol{x}_1^T \hat{\boldsymbol{\beta}})}{exp(\boldsymbol{x}_2^T \hat{\boldsymbol{\beta}})} \\ &= exp[(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \hat{\boldsymbol{\beta}}]. \end{aligned} \tag{2.28}$$

Εάν, στο μοντέλο με τις δύο συμμεταβλητές $x_1$ και $x_2$ υποτεθεί ότι η μία συμμεταβλητή, έστω η $x_2$, είναι ποσοτική και η $x_1$ είναι μία δείκτρια μεταβλητή με $x_1 = 0$ ή $x_1 = 1$, τότε:

$$odds\{y = 1 \mid x_1 = 0, x_2\} = exp(\hat{\beta}_0 + \hat{\beta}_2 x_2) \tag{2.29}$$

$$odds\{y = 1 \mid x_1 = 1, x_2\} = exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 x_2) \tag{2.30}$$

Τελικά, ο λόγος των odds είναι ίσος με:

$$\frac{odds\{y=1 \mid x_1=0, x_2\}}{odds\{y=1 \mid x_1=1, x_2\}} = e^{\hat{\beta}_1} \tag{2.31}$$

όπου είναι φανερό ότι δεν εξαρτάται από τη συμμεταβλητή $x_2$, που υποτέθηκε να είναι ποσοτική.

### 2.5.5 Έλεγχοι Υποθέσεων

Έπειτα από την εκτίμηση των συντελεστών, δηλαδή την προσαρμογή του μοντέλου της λογιστικής παλινδρόμησης, θα πρέπει να εξεταστεί η καταλληλότητα του μοντέλου. Με άλλα λόγια, θα πρέπει να ερευνηθεί η σχέση των παρατηρούμενων τιμών, με τις αντίστοιχες προβλεφθείσες του μοντέλου.

Ο σκοπός των στατιστικών ελέγχων είναι να διερευνηθεί η στατιστική σημαντικότητα των μεταβλητών του μοντέλου. Έτσι, μέσα από τις εκτιμημένες τιμές των παραμέτρων και τον έλεγχο Wald, καθίσταται εφικτός αυτός ο στόχος. Επιπλέον, εξίσου σημαντική διαδικασία, μετά την προσαρμογή του μοντέλου, είναι η σύγκριση των διάφορων υποψήφιων μοντέλων, ώστε να επιλεχθεί το καλύτερο για τα διαθέσιμα δεδομένα. Αυτό πραγματοποιείται με τη βοήθεια της ελεγχοσυνάρτησης deviance, η οποία θα παρουσιαστεί αναλυτικά παρακάτω.

#### Έλεγχος Wald

Από τη θεωρία της μεθόδου μέγιστης πιθανοφάνειας, για μεγάλα δείγματα, η εκτιμήτρια $\hat{\boldsymbol{\beta}}$ για την παράμετρο $\boldsymbol{\beta}$ ακολουθεί, ασυμπτωτικά, την πολυμετάβλητη Κανονική κατανομή:

$$\hat{\boldsymbol{\beta}} \dot{\sim} N_p(\boldsymbol{\beta}, \hat{V}(\hat{\boldsymbol{\beta}})), p = k+1,$$

όπου με $\hat{V}(\hat{\boldsymbol{\beta}}) = \boldsymbol{J}^{-1}(\hat{\boldsymbol{\beta}})$ συμβολίζεται η εκτιμήτρια του πίνακα διασποράς-συνδιασποράς της ασυμπτωτικής κατανομής της $\hat{\boldsymbol{\beta}}$ και με $\boldsymbol{J}(\hat{\boldsymbol{\beta}})$ συμβολίζεται ο παρατηρούμενος πίνακας πληροφορίας, με στοιχεία:

$$\boldsymbol{J}_{jr} = \sum_{i=1}^{n} \frac{E\Big[(y_i-\mu_i)^2\Big] x_{ij}x_{ir}}{[Var(Y_i)]^2} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2, j, r = 0, 1, ..., k.$$

Στα γενικευμένα γραμμικά μοντέλα, ο παρατηρούμενος πίνακας πληροφορίας είναι:

$$\boldsymbol{J}(\hat{\boldsymbol{\beta}}) = \boldsymbol{X^T}\hat{\boldsymbol{W}}\boldsymbol{X},$$

όπου με $\hat{\boldsymbol{W}}$ συμβολίζεται ο αντίστοιχος πίνακας $\boldsymbol{W} = diag\big(e^{\boldsymbol{x}^T\boldsymbol{\beta}}\big)$, μετά την αντικατάσταση των $\boldsymbol{\beta}$ από τα $\hat{\boldsymbol{\beta}}$. Από τη θεωρία της μεθόδου μέγιστης πιθανοφάνειας, η διασπορά του συντελεστή $\hat{\beta}_j$, $\hat{V}(\hat{\beta}_j)$, είναι το $j$-οστό διαγώνιο στοιχείο του πίνακα $\boldsymbol{J}^{-1}(\hat{\beta}_j)$, με αντίστοιχο τυπικό σφάλμα $se(\hat{\beta}_j) = \big(\hat{V}(\hat{\beta}_j)\big)^{1/2} = \big(\boldsymbol{J}^{-1}(\hat{\boldsymbol{\beta}})_{jj}\big)^{1/2}$.

Επειδή οι εκτιμήτριες της μέγιστης πιθανοφάνειας ακολουθούν ασυμπτωτικά την Κανονική κατανομή, προσεγγιστικά ισχύει ότι:

$$Z = \frac{\hat{\beta}_j - \beta_j}{\left(\boldsymbol{J}^{-1}(\hat{\boldsymbol{\beta}})_{jj}\right)^{1/2}} \dot{\sim} N(0,1), j = 0, 1, 2, ..., k, \tag{2.32}$$

όπου με $\boldsymbol{J}^{-1}(\hat{\boldsymbol{\beta}})_{jj}$ συμβολίζεται το $j$-οστό διαγώνιο στοιχείο του αντίστροφου, του παρατηρούμενου πίνακα πληροφορίας $\boldsymbol{J}(\hat{\boldsymbol{\beta}})$. Η σχέση αυτή μπορεί να χρησιμοποιηθεί για τον έλεγχο Wald μιας μηδενικής υπόθεσης, όπως για παράδειγμα της

$$H_0 : \beta_j = 0, (\text{ή κάποια άλλη συγκεκριμένη τιμή } \beta_{j0}),$$

με εναλλακτική,

$$H_1 : \beta_j \neq 0, (\beta_j \neq \beta_{j0}).$$

Όπως και στην περίπτωση του γραμμικού μοντέλου, $\beta_j = 0$ σημαίνει ότι, η μεταβλητή $X_j$ δε συμβάλλει στην πρόβλεψη της Y και μπορεί να αφαιρεθεί από το μοντέλο.

Η στατιστική συνάρτηση (2.32), μπορεί να χρησιμοποιηθεί για την κατασκευή ενός $100(1-\alpha)$%-διαστήματος εμπιστοσύνης, για την παράμετρο $\beta_j$, της μορφής:

$$\hat{\beta}_j \pm z_{\alpha/2} se(\hat{\beta}_j),$$

όπου με $z_{\alpha/2}$ συμβολίζεται το άνω $100(\alpha/2)$-ποσοστιαίο σημείο της $N(0,1)$ κατανομής και $se(\hat{\beta}_j) = \left(\hat{V}(\hat{\beta}_j)\right)^{1/2} = \left(\boldsymbol{J}^{-1}(\hat{\boldsymbol{\beta}})_{jj}\right)^{1/2}$.

### 2.5.6 Διαστήματα εμπιστοσύνης

Όπως προαναφέρθηκε, η στατιστική συνάρτηση του Wald χρησιμοποιείται για την κατασκευή διαστημάτων εμπιστοσύνης για τις παραμέτρους του μοντέλου. Ένα $100(1-\alpha)$%-διάστημα εμπιστοσύνης για την παράμετρο $\beta_j$ δίνεται από τη σχέση:

$$\hat{\beta}_j \pm z_{\alpha/2} se(\hat{\beta}_j), j = 0, 1, 2, ..., k. \tag{2.33}$$

Επιπλέον, από το διάστημα εμπιστοσύνης της παραμέτρου $\beta_j$ προσδιορίζεται και το αντίστοιχο $100(1-\alpha)$%-διάστημα εμπιστοσύνης για το λόγο των συμπληρωματικών πιθανοτήτων (odds ratio). Έτσι, η αντίστοιχη σχέση για το διάστημα εμπιστοσύνης του λόγου των odds είναι:

$$exp\left[\hat{\beta}_j \pm z_{\alpha/2} se(\hat{\beta}_j)\right] \tag{2.34}$$

Με $z_{\alpha/2}$ συμβολίζεται το $100(\alpha/2)$-ποσοστιαίο σημείο της τυπικής Κανονικής κατανομής με πιθανότητα ίση με $\alpha/2$.

Επιπλέον, είναι δυνατή η κατασκευή ενός $100(1-\alpha)$%-διαστήματος εμπιστοσύνης για τη γραμμική προβλέπουσα του μοντέλου, δοσμένων των τιμών των επεξηγηματικών μεταβλητών. Έστω, το σύνολο τιμών των επεξηγηματικών μεταβλητών $\boldsymbol{x}_0^T = (1, x_{01}, x_{02}, ..., x_{0k})$. Σε αυτήν την περίπτωση, η εκτιμήτρια της διασποράς της γραμμικής προβλέπουσας, θα είναι:

$$V(\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0^T V(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_0 = \boldsymbol{x}_0^T (\boldsymbol{X}^T \hat{\boldsymbol{V}} \boldsymbol{X})^{-1} \boldsymbol{x}_0.$$

Συνεπώς, ένα $100(1-\alpha)$%-διάστημα εμπιστοσύνης για τη γραμμική προβλέπουσα, θα είναι:

$$\boldsymbol{x}_0\hat{\boldsymbol{\beta}} - z_{\alpha/2}\sqrt{V(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}})} \leq \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} \leq \boldsymbol{x}_0\hat{\boldsymbol{\beta}} + z_{\alpha/2}\sqrt{V(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}})}.$$

Το διάστημα εμπιστοσύνης για τη γραμμική προβλέπουσα παρέχει τη δυνατότητα, να κατασκευαστεί ένα διάστημα εμπιστοσύνης για την πιθανότητα εμφάνισης επιτυχίας $\pi_0$, δεδομένου του ίδιου συνόλου τιμών των επεξηγηματικών μεταβλητών $\boldsymbol{x}_0^T = (1, x_{01}, x_{02}, ..., x_{0k})$. Θεωρώντας ως:

$$L(\boldsymbol{x}_0) = \boldsymbol{x}_0\hat{\boldsymbol{\beta}} - z_{\alpha/2}\sqrt{V(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}})},$$

και:

$$U(\boldsymbol{x}_0) = \boldsymbol{x}_0\hat{\boldsymbol{\beta}} + z_{\alpha/2}\sqrt{V(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}})}$$

το κάτω και άνω όριο του διαστήματος εμπιστοσύνης, αντίστοιχα, για τη γραμμική προβλέπουσα, τότε ένα $100(1-\alpha)$%-διάστημα εμπισοσύνης για την εκτίμηση της πιθανότητας εμφάνισης επιτυχίας $\pi_0$ είναι:

$$\frac{exp[L(\boldsymbol{x}_0)]}{1 + exp[L(\boldsymbol{x}_0)]} \leq \pi_0 \leq \frac{exp[U(\boldsymbol{x}_0)]}{1 + exp[U(\boldsymbol{x}_0)]} \tag{2.35}$$

Στην πράξη, για την κατασκευή των διαστημάτων εμπιστοσύνης πρώτα καθορίζεται η ποσότητα $1-\alpha$ (με το $\alpha$ να εκφράζει την ανοχή στο να μην περιέχεται στο διάστημα εμπιστοσύνης που θα χρησιμοποιήσουμε ο εκτιμώμενος συντελεστής $\beta_j$) και στη συνέχεια υπολογίζεται το διάστημα που έχει συντελεστή εμπιστοσύνης ίσο με αυτήν την ποσότητα. Το $0.95$ ($= 95\%$) αποτελεί τη συνηθέστερη επιλογή για την ποσότητα $1-\alpha$, ενώ άλλες συνηθισμένες επιλογές είναι το $0.90$ ($= 90\%$) και το $0.99$ ($= 99\%$).

### 2.5.7 Ελεγχοσυνάρτηση Deviance

Η **ελεγχοσυνάρτηση deviance** αποτελεί μια τεχνική που μπορεί να χρησιμοποιηθεί για την σύγκριση και ανάπτυξη των στατιστικών μοντέλων. Στην ανάλυση δεδομένων, σκοπός είναι να βρεθεί το βέλτιστο μοντέλο και όχι απλώς να ελεγχθεί η προσαρμογή ενός συγκεκριμένου μοντέλου. Για το σκοπό αυτό χρησιμοποιείται κυρίως η τεχνική του λόγου των πιθανοφανειών που ελέγχει τις υποθέσεις σχετικά με τις παραμέτρους του μοντέλου.

Από τη γενική θεωρία ισχύει ότι:

$$-2log\lambda = -2(log\hat{L}_0 - log\hat{L}_1) = -2(\hat{l}_0 - \hat{l}_1)\dot{\sim}\chi^2_d, \text{ ασυμπτωτικά,}$$

όπου $d = p_1 - p_0$, η διαφορά των διαστάσεων των παραμετρικών χώρων και $\hat{l}_0$, $\hat{l}_1$ οι μεγιστοποιημένες λογαριθμοποιημένες συναρτήσεις πιθανοφάνειας υπό την $H_0$ και $H_1$ αντίστοιχα.

Ένα μοντέλο ονομάζεται **πλήρες ή κορεσμένο** (*saturated*) εάν έχει τόσες παραμέτρους, όσες και παρατηρήσεις. Έστω δύο μοντέλα $M$ και $M^*$ με εκτιμήτριες μέγιστης πιθανοφάνειας $\hat{\boldsymbol{\beta}}$ και $\hat{\boldsymbol{\beta}}^*$, αντίστοιχα. Επιπλέον, έστω ότι $M^* \subset M$, δηλαδή όλες οι μεταβλητές του μοντέλου $M^*$ περιέχονται και στο $M$. Τότε ο έλεγχος της μηδενικής υπόθεσης είναι:

$$H_0 : \text{ισχύει το μοντέλο } M^*$$

με εναλλακτική,

$$H_1 : \text{ισχύει το μοντέλο } M.$$

Για να συγκρίνουμε τα εν λόγω μοντέλα χρησιμοποιούμε τον έλεγχο με βάση τον λόγο των μεγιστοποιημένων πιθανοφανειών:

$$-2\big(l(\hat{\boldsymbol{\beta}}^*) - l(\hat{\boldsymbol{\beta}})\big)\dot{\sim}\chi^2_d, \text{ ασυμπτωτικά}$$

όπου $\hat{\boldsymbol{\beta}}$ και $\hat{\boldsymbol{\beta}}^*$ είναι οι εκτιμήσεις στα μοντέλα $M$ και $M^*$ αντίστοιχα, και $d$ η διαφορά του πλήθους των παραμέτρων του ελαττωμένου από του πλήρες μοντέλου. Στην περίπτωση που η ελεγχοσυνάρτηση λάβει μεγάλες τιμές, το αποτέλεσμα οδηγεί σε απόρριψη της μηδενικής υπόθεσης $H_0$.

Εάν, τώρα, η εναλλακτική υπόθεση $H_1$ αντικατασταθεί από την υπόθεση:

$$H_S : \text{ισχύει το κορεσμένο μοντέλο, με } p_s = n,$$

δηλαδή το κορεσμένο μοντέλο έχει αριθμό παραμέτρων ίσο με τον αριθμό των παρατηρήσεων, και η $H_0$ αντικατασταθεί από την μηδενική υπόθεση:

$$H_0 : \text{ισχύει το υποψήφιο μοντέλο, με } p_0 = p < n,$$

τότε η ελεγχοσυνάρτηση deviance ορίζεται από τη σχέση:

$$2log\frac{L\{\text{κορεσμένο μοντέλο}\}}{L\{\text{υποψήφιο μοντέλο}\}} \tag{2.36}$$

και οι προβλεπόμενες τιμές $\tilde{\mu}_i$ ισούνται με τις παρατηρούμενες τιμές $y_i$, $\tilde{\mu}_i = y_i$. Η $H_0$ περιορίζεται στον $p = k + 1$-διάστατο χώρο, με δομή $\psi_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, με $\boldsymbol{x}_i^T = (x_{i0}, x_{i1}, ..., x_{ik})$ να είναι οι τιμές των $k$ συμμεταβλητών της στατιστικής μονάδας $i$ και $x_{i0} \equiv 1$.

Έστω $Y_1, Y_2, ..., Y_n$ ανεξάρτητες τυχαίες μεταβλητές, οι οποίες προέρχονται από την κατανομή Bernoulli, δηλαδή $Y_i \sim B(n_i, \pi_i)$ με μέση τιμή $E(Y_i) = \mu_i = \pi_i$, $\mu_i > 0$, και θα εκτιμάται από το $\tilde{\mu}_i = y_i$. Τότε η μεγιστοποιημένη τιμή του λογαρίθμου της συνάρτησης της πιθανοφάνειας, υπό την υπόθεση $H_S$, του κορεσμένου μοντέλου, θα είναι:

$$\tilde{l}_s = logL_s = \sum_{i=1}^{n}\Big\{y_i log\tilde{\pi}_i + (1 - y_i)log(1 - \tilde{\pi}_i)\Big\},$$

όπου $\tilde{\pi}_i = y_i$.

Για το υποψήφιο μοντέλο με αριθμό παραμέτρων $p_0 < n$, ισχύει ότι $\hat{\mu}_i = \hat{\pi}_i$, όπου $\hat{\pi}_i = \frac{exp\{\boldsymbol{x}^T\hat{\boldsymbol{\beta}}\}}{1+exp\{\boldsymbol{x}^T\hat{\boldsymbol{\beta}}\}}$, οι πιθανότητες απόκρισης που προκύπτουν από την προσαρμογή του μοντέλου. Σε αυτήν την περίπτωση, η μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας, υπό την υπόθεση $H_0$, θα είναι:

$$\hat{l}_0 = logL_0 = \sum_{i=1}^{n}\Big\{y_i log\hat{\pi}_i + (1 - y_i)log(1 - \hat{\pi}_i)\Big\}.$$

Η ελεγχοσυνάρτηση deviance, στην περίπτωση της λογιστικής παλινδρόμησης για δυαδικά δεδομένα, ορίζεται ως:

$$
\begin{aligned}
D(\hat{\boldsymbol{\beta}}) &= D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) \\
&= 2\{\tilde{l}_s - \hat{l}_0\} \\
&= 2\sum_{i=1}^{n} \Big\{ y_i log\tilde{\pi}_i + (1-y_i)log(1-\tilde{\pi}_i) - y_i log\hat{\pi}_i + (1-y_i)log(1-\hat{\pi}_i) \Big\} \\
&= 2\sum_{i=1}^{n} \left\{ y_i log\Big(\frac{\tilde{\pi}_i}{\hat{\pi}_i}\Big) + (1-y_i)log\Big(\frac{1-\tilde{\pi}_i}{1-\hat{\pi}_i}\Big) \right\} \\
&= 2\sum_{i=1}^{n} \left\{ y_i log\Big(\frac{y_i}{\hat{\mu}_i}\Big) + (1-y_i)log\Big(\frac{1-y_i}{1-\hat{\mu}_i}\Big) \right\}.
\end{aligned}
\tag{2.37}
$$

Όταν το μοντέλο της λογιστικής παλινδρόμησης είναι το καταλληλότερο για την περιγραφή των δεδομένων και το διαθέσιμο δείγμα τιμών είναι μεγάλο, η ελεγχοσυνάρτηση deviance ακολουθεί ασυμπτωτικά την κατανομή $X^2$, με βαθμούς ελευθερίας ίσους με τη διαφορά των παραμέτρων του κορεσμένου μοντέλου από το υποψήφιο.

Σημειώνεται ότι στην ειδική περίπτωση των δυαδικών δεδομένων, η ελεγχοσυνάρτηση deviance δε μας παρέχει πληροφορίες για την καταλληλότητα ενός μοντέλου. Παρ'όλα αυτά, η συνάρτησή της χρησιμοποιείται στο κριτήριο BIC σαν εναλλακτική περίπτωση για την επιλογή του κατάλληλου μοντέλου. Συνεπώς δεν μπορούμε να την απορρίψουμε από τους ελέγχους για τα γενικευμένα γραμμικά μοντέλα.

### 2.5.8 Πίνακες ταξινόμησης (*classification tables*)

Μερικές φορές καθίσταται χρήσιμο να συνοψισθεί η προβλεπτική ισχύς ενός προσαρμοσμένου μοντέλου λογιστικής παλινδρόμησης. Ένας, διαισθητικός, τρόπος να πραγματοποιηθεί αυτό είναι μέσω του λεγόμενου **πίνακα ταξιμόμησης** (*classification table*). Αυτός ο πίνακας περιλαμβάνει τις πραγματικές τιμές της μεταβλητής απόκρισης, $Y$, διασταυρωμένες με μία διχοτομημένη μεταβλητή, της οποίας οι τιμές προκύπτουν από το προσαρμοσμένο μοντέλο της λογιστικής παλινδρόμησης. Με άλλα λόγια, η μέθοδος αυτή ταξίνομεί τη δυαδική μεταβλητή $Y$ σε συνδυασμό με την αντίστοιχη τιμή του προβλεπτικού μοντέλου, δηλαδή αν $Y = 1$ ή $Y = 0$.

Για να πραγματοποιηθεί αυτή η διαδικασία, χρειάζεται πρώτα να οριστεί ένα όριο $\pi_0$, το οποίο θα συγκριθεί αργότερα με κάθε μία πιθανότητα που προβλέπεται από το μοντέλο. Εάν η προβλεφθείσα πιθανότητα υπερβαίνει το όριο που έχουμε ορίσει ($\hat{\pi}_i > \pi_0$), τότε η παραγόμενη μεταβλητή $Y$ παίρνει την τιμή 1 ($Y = 1$)·διαφορετικά, παίρνει την τιμή 0 ($Y = 0$). Συνήθως, η πιθανότητα $\pi_0$ ορίζεται να είναι ίση με 0.5. Η προτίμηση αυτής της μεθόδου αξιολόγησης μοντέλων προέρχεται από τη στενή σχέση μεταξύ των μεθόδων της λογιστικής παλινδρόμησης και της διακριτικής ανάλυσης (discriminant analysis), όταν η κατανομή των συμμεταβλητών είναι η πολυμετάβλητη Κανονική. Ωστόσο, δεν περιορίζεται μόνο σε αυτό το μοντέλο.

Πιο συγκεκριμένα, σε αυτήν την τεχνική, οι εκτιμηθείσες πιθανότητες χρησιμοποιούνται με σκοπό να προβλέψουν την ομαδοποίηση των παρατηρήσεων. Υποθετικά, αν το μοντέλο προβλέψει με ακρίβεια την κατηγορία που ανήκει η εκάστοτε παρατή-

ρηση, αυτό αποτελεί το αποδεικτικό στοιχείο που χρειάζεται, ώστε να θεωρηθεί ότι το μοντέλο προσαρμόζεται καλά. Δυστυχώς όμως, αυτό δε συμβαίνει συνήθως.

### 2.5.9 Καμπύλες ROC

Οι **ROC καμπύλες ή καμπύλες λειτουργικού χαρακτηριστικού δέκτη** (*ROC Curves, Receiver Operating Characteristic Curves*) αποτελούν μία τεχνική, όπου μέσω της γραφικής παράστασης ταξινομητών είναι δυνατή η οργάνωση, η επιλογή και η απεικόνησή αυτών. Η μέθοδος αυτή χρησιμοποιείται ευρέως στη Διαγνωστική Ιατρική, ενώ πρόσφατα έχει εισαχθεί και σε διάφορους τομείς της Μηχανικής Εκμάθησης. Αρχικά, η ευαισθησία και η ειδικότητα είναι δύο στατιστικά μέτρα που απεικονίζουν την απόδοση ενός δυαδικού δείκτη ταξινόμησης και ορίζονται ως εξής:

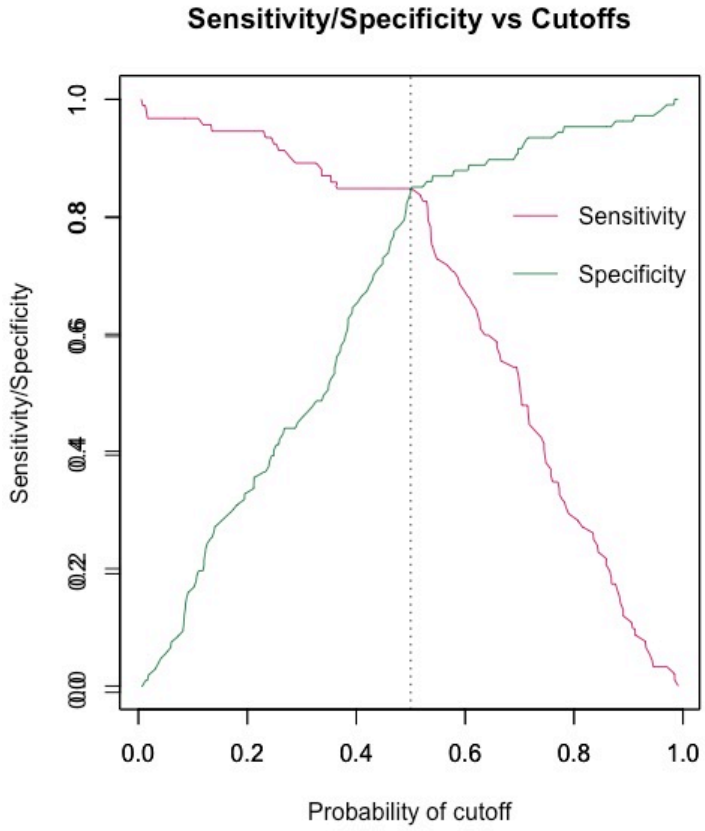
- **η "ευαισθησία"** (*Sensitivity*, επίσης γνωστή και ως *true positive rate*), υπολογίζει το ποσοστό των επιτυχιών που έχουν αναγνωριστεί σωστά ως τέτοια, και

- **η "ειδικότητα"** (*Specificity*, επίσης γνωστή και ως *true negative rate*), όπου υπολογίζει το ποσοστό των αποτυχιών που έχουν αναγνωριστεί σωστά ως τέτοια.

Οι δύο αυτές ποσότητες βασίζονται σε ένα προκαθορισμένο σημείο διαχωρισμού, $\pi_0$, ώστε να ταξινομηθεί ένα αποτέλεσμα του δοκιμαστικού δείγματος, ως θετικό. Έτσι, μία πιο ολοκληρωμένη περιγραφή της επάρκειας της ταξινόμησης αποτελεί η λεγόμενη **περιοχή κάτω από τη ROC καμπύλη** (*AUC, Area Under the ROC Curve*). Η καμπύλη αυτή, η οποία προέρχεται από τη θεωρία ανίχνευσης σημάτων, δείχνει πώς αντιμετωπίζει ο δέκτης την παρουσία σήματος υπό την ύπαρξη θορύβου. Ουσιαστικά, πρόκειται για μία γραφική παράσταση όπου απεικονίζεται η πιθανότητα του ανιχνευμένου αληθούς σήματος (*ευαισθησία*) συναρτήσει του λανθασμένου σήματος (*1 - ειδικότητα*) για ολόκληρη την εμβέλεια των πιθανών σημείων διαχωρισμού.

Η καμπύλη ROC προσφέρει περισσότερη πληροφορία από έναν πίνακα ταξινόμησης, καθώς συνοψίζει την προβλεπτική ισχύ του μοντέλου για όλα τα πιθανά σημεία διαχωρισμού, $\pi_0$. Όταν το $\pi_0$ πλησιάζει στο 0, σχεδόν όλες οι προβλέψεις είναι ίσες με 1, $\hat{y} = 1$, όπου η ευαισθησία είναι κοντά στη μονάδα και η ειδικότητα είναι κοντά στο 0. Αντιθέτως, όταν το $\pi_0$ πλησιάζει στο 1, σχεδόν όλες οι προβλέψεις είναι ίσες με 0, $\hat{y} = 0$, με την ευαισθησία και την ειδικότητα να παίρνουν τιμές κοντά στο 0 και το 1, αντίστοιχα. Η καμπύλη ROC συνήθως έχει καμπύλο σχήμα και ενώνει τα σημεία (0,0) και (1,1). Για δοθείσα ειδικότητα, η καλύτερη προβλεπτική ισχύς αντιστοιχεί σε υψηλότερη ευαισθησία. Συνεπώς, όσο καλύτερη είναι η προβλεπτική ισχύς, τόσο υψηλότερη είναι η τιμή της καμπύλη ROC.

Σαν πρώτο βήμα, σκοπός είναι να επιλεχθεί το βέλτιστο σημείο διαχωρισμού ώστε να ακολουθήσει η διαδικασία της ταξινόμησης. Έτσι, κάποιος θα μπορούσε να επιλέξει, για παράδειγμα, το σημείο εκείνο που μεγιστοποιεί ταυτόχρονα την ευαισθησία, αλλά και την ειδικότητα. Ανάλογα τα δεδομένα που υπάρχουν στη διάθεση του ερευνητή, το σημείο αυτό μπορεί να διαφέρει. Η επιλογή αυτή, λοιπόν, μπορεί να γίνει μέσω ενός γραφήματος, που έχει μορφή όπως το σχήμα (2.2). Στη συγκεκριμένη περίπτωση, ένα βέλτιστο σημείο διαχωρισμού μπορεί να είναι το $x = 0.5$, όπου είναι προσεγγιστικά το σημείο που οι δύο καμπύλες τέμνονται.

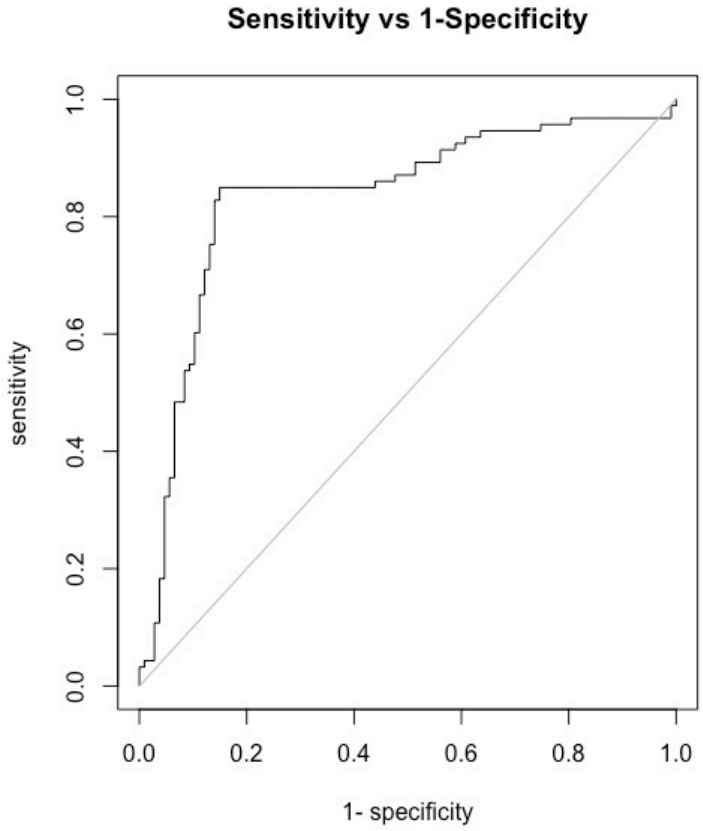



Σχήμα 2.2: Εύρεση σημείου τομής των "ευαισθησία", "ειδικότητα", το οποίο είναι ένα υποψήφιο βέλτιστο σημείο διαχωρισμού, για τη διαδικασία της ταξινόμησης.

Στο σχήμα (2.3) απεικονίζεται η ευαισθησία συναρτήσει της ποσότητας (1-ειδικότητα). Η καμπύλη που σχηματίζεται από τα σημεία αυτών των ποσοτήτων είναι η λεγόμενη καμπύλη ROC, και η περιοχή κάτω από την καμπύλη παρέχει ένα μέτρο διάκρισης. Στα δεδομένα που απεικονίζονται στο γράφημα η περιοχή κάτω από την καμπύλη είναι ίση με 0.8341875. Ένας γενικός κανόνας είναι ο εξής:

- Εάν $ROC = 0.5$, δεν υπάρχει ένδειξη για διαχωρισμό.
- Εάν $0.7 \leq ROC < 0.8$, θεωρείται ικανοποιητική ένδειξη για διαχωρισμό.
- Εάν $0.8 \leq ROC < 0.9$, θεωρείται άριστη ένδειξη για διαχωρισμό.
- Εάν $ROC \geq 0.9$, θεωρείται εξαιρετική ένδειξη για διαχωρισμό.

Στην πράξη, είναι σχεδόν αδύνατο να παρατηρηθούν τιμές που να ξεπερνούν το 0.9, για την περιοχή κάτω από την καμπύλη ROC. Στην πραγματικότητα, σε τέτοιες περιπτώσεις που ο διαχωρισμός είναι ξεκάθαρος είναι αδύνατο να εκτιμηθούν οι συντελεστές του λογιστικού μοντέλου παλινδρόμησης. Να σημειωθεί επίσης, ότι ένα φτωχά προσαρμοσμένο μοντέλο μπορεί και πάλι να παρουσιάζει καλό διαχωρισμό. Γενικά προτείνεται, η επίδοση ενός μοντέλου να αξιολογείται λαμβάνοντας υπόψιν τόσο το διαχωρισμό, όσο και τη διαμέτριση.




Σχήμα 2.3: Καμπύλη ROC, ή αλλιώς καμπύλη λειτουργικού χαρακτηριστικού δέκτη (*ROC Curve, Receiver Operating Characteristic Curve*)

### 2.5.10 Το φαινόμενο της πολυσυγγραμμικότητας

Το φαινόμενο της **πολυσυγγραμμικότητας** (*multicollinearity*) χαρακτηρίζεται ως μία τεχνική δυσκολία που αφορά τις επεξηγηματικές μεταβλητές. Ουσιαστικά, εμφανίζεται όταν υπάρχει έντονη συσχέτιση μεταξύ δύο ή περισσοτέρων ανεξάρτητων μεταβλητών. Γενικά, η παρουσία της πολυσυγγραμμικότητας στα δεδομένα, έχει ως αποτέλεσμα να εμφανίζονται αυξημένα τυπικά σφάλματα στις παραμέτρους $\hat{\boldsymbol{\beta}}$ και κατα συνέπεια να δυσκολεύει η εκτίμηση της επίδρασης της κάθε επεξηγηματικής μεταβλητής στην μεταβλητή απόκρισης. Αυτό συμβαίνει καθώς τα διαστήματα εμπιστοσύνης των αντίστοιχων συντελεστών θα παρουσιάζουν μεγάλο εύρος.

Όταν λοιπόν, εντοπίζεται αυτό το φαινόμενο είναι εξαιρετικά δύσκολο να επιλεχθεί το καλύτερο σύνολο επεξηγηματικών μεταβλητών που θα εισαχθεί στο προσαρμοσμένο μοντέλο. Για το λόγο αυτό, πριν την οποιαδήποτε διαδικασία παλινδρόμησης, καθίσταται απαραίτητο να ελεχθεί η παρουσία συσχέτισης μεταξύ των ανεξάρτητων μεταβλητών. Αυτή η διαδικασία γίνεται μέσω ενός στατιστικού εργαλείου που ονομάζεται **παράγοντας μεγέθυνσης διασποράς** (*VIF, Variance Inflation Factor*) και εφαρμόζεται σε κάθε επεξηγηματική μεταβλητή ξεχωριστά. Σημειώνεται ότι ο έλεγχος αυτός πραγματοποιείται μόνο στις μεταβλητές ποσοτικής φύσης.

Η ανάλυση του συντελεστή αυτού πραγματοποιείται $k$ φορές, όσες δηλαδή είναι

και οι επεξηγηματικές μεταβλητές, $X_i, i = 1, 2, \ldots, k$. Αρχικά, πραγματοποιείται παλινδρόμηση ελάχιστων τετραγώνων στη μεταβλητή $X_i$ συναρτήσω όλων των υπόλοιπων επεξηγηματικών μεταβλητών. Για παράδειγμα, για $i = 1$, η αντίστοιχη εξίσωση, θα είναι:

$$X_1 = \alpha_2 X_2 + \alpha_3 X_3 + ... + \alpha_k X_k + c_0 + e$$

όπου $c_0$ είναι ο σταθερός όρος και $e$ ο όρος τους σφάλματος. Στη συνέχεια, υπολογίζεται ο παράγοντας VIF για τον συντελεστή $\hat{\beta}_i$, από τη σχέση:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2.38}$$

όπου $R_i^2$ είναι ο συντελεστής προσδιορισμού της προηγούμενης εξίσωσης παλινδρόμησης της μεταβλητής $X_i$. Τέλος, ερμηνεύεται το αριθμητικό αποτέλεσμα του παράγοντα VIF και ανάλογα με το μέγεθός του προκύπτει το συμπέρασμα για την πολυσυγγραμμικότητα. Ένας γενικός κανόνας ερμηνείας του αποτελέσματος της παραπάνω εξίσωσης είναι:

- εάν $VIF_i = 1$, τότε οι επεξηγηματικές μεταβλητές είναι ασυσχέτιστες με την μεταβλητή $X_i$, ενώ
- εάν $VIF_i > 5$, τότε υπάρχουν ενδείξεις υψηλής πολυσυγγραμμικότητας μεταξύ των επεξηγηματικών μεταβλητών και της μεταβλητής $X_i$ (Montgomery and Peck, 1992).

Με άλλα λόγια, όσο αυξάνει η τιμή του παράγοντα VIF, τόσο αυξάνεται και η πολυσυγγραμμικότητα μεταξύ των επεξηγηματικών μεταβλητών.

### 2.5.11 Κριτήρια επιλογής του μοντέλου

Για την αξιολόγηση ενός μοντέλου, αλλά και για τη σύγκριση διαφορετικών μοντέλων ως προς τη σπουδαιότητά τους, γίνεται χρήση διάφορων εργαλείων γνωστά ως **μέτρα καταλληλότητας**. Πρόκειται για αριθμητικές ποσότητες, οι οποίες χρησιμοποιούνται, κυρίως, για την επιλογή του βέλτιστου μοντέλου. Για τις ανάγκες της παρούσας μελέτης, θα γίνει αναφορά στα κριτήρα AIC και BIC, καθώς και στους συντελεστές προσαρμογής ή συντελεστές προσδιορισμού (*coefficient of fit or coefficient of determination*), παρ'όλο που στη γενική θεωρία οι δείκτες αυτοί ποικίλουν ανάλογα με το μοντέλο.

#### Κριτήριο AIC (*Akaikes's Information Criterion*)

Το **κριτήριο AIC** (*Akaikes's Information Criterion*), παρουσιάστηκε το 1974 από τον Akaike, και εφαρμόζεται σε ένα μεγάλο σύνολο μοντέλων, που έχουν προσαρμοστεί με βάση τη μέθοδο της μέγιστης πιθανοφάνειας. Αποτελεί ένα κριτήριο επιλογής του βέλτιστου μοντέλου με το όσο το δυνατόν μικρότερο αριθμό παραμέτρων. Στη γενική περίπτωση, δίνεται από τη σχέση:

$$AIC = 2d - 2logL \tag{2.39}$$

όπου το $d$ εκφράζει το πλήθος των παραμέτρων του μοντέλου και το $L$ τη μεγιστοποιημένη τιμή της συνάρτησης πιθανοφάνειας για το εκτιμηθέν μοντέλο.

Συγκρίνοντας όλα τα υποψήφια μοντέλα, η μέθοδος αυτή καταλήγει στο βέλτιστο, με βάση το αριθμητικό αποτέλεσμα της σχέσης (2.39). Όσο μικρότερο το AIC, τόσο προτιμότερο το αντίστοιχο μοντέλο. Εάν εισαχθούν περισσότερες παράμετροι στο μοντέλο, αυτό τίνει να βελτιώσει την προσαρμογή του μοντέλου, ανεξάρτητα από το αν αυτές είναι στατιστικά σημαντικές ή όχι. Με άλλα λόγια, αν προστεθούν στο μοντέλο περισσότεροι παράμετροι, αυτό έχει ως αποτέλεσμα την αύξηση της ποσότητας $logL$ και κατα συνέπεια ο δεύτερος όρος του AIC μειώνεται. Αντιθέτως, αυξάνεται ο πρώτος όρος του AIC. Ο δεύτερος όρος $2d$ καλείται **ποινή** (*penalty*) και εκφράζει μία αύξουσα συνάρτηση του αριθμού των εκτιμηθέντων παραμέτρων. Ο ρόλος της ποινής αυτής είναι να προλαμβάνει το *overfitting*, το οποίο εμποδίζει το μοντέλο να προσαρμοστεί σωστά, λόγω της πολυπλοκότητάς του (παρουσία θορύβου).Τελικά, η εισαγωγή επιπλέον παραμέτρων στο μοντέλο μειώνει την τιμή του κριτηρίου AIC μόνο αν αυτές βελτιώνουν την προσαρμογή του μοντέλου.

Στην περίπτωση της λογιστικής παλινδρόμησης το κριτήριο AIC παίρνει τη μορφή:

$$\begin{aligned} AIC =& -2\sum_{i=1}^{n}\left[y_i log\hat{\pi}_i + (1-y_i)log(1-\hat{\pi}_i)\right] + 2p \\ =& 2\sum_{i=1}^{n}\left[log\big(1+e^{\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}\big) - y_i\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right] + 2p. \end{aligned} \tag{2.40}$$

**Κριτήριο BIC (*Bayesian Information Criterion*)**

Με παρόμοια λογική με το κριτήριο AIC, το 1978 προστέθηκε το **κριτήριο BIC** (*Bayesian Information Criterion*) από τον ισραηλινό μαθηματικό Gideon E. Schwarz. Σκοπός, λοιπόν, ήταν η επιλογή του βέλτιστου μοντέλου ανάμεσα σε μοντέλα με διαφορετικό αριθμό παραμέτρων. Αν και η αφετηρία του είναι διαφορετική από εκείνη του AIC, η λογική και οι χρήσεις των δύο κριτηρίων είναι παρόμοιες. Η βασική τους διαφορά είναι ότι η εισαγωγή επιπρόσθετων παραμέτρων αποθαρρύνεται σε μεγαλύτερο βαθμό από ότι στο AIC.

Στη γενική περίπτωση, το κριτήριο BIC ορίζεται από τη σχέση:

$$BIC = dlogn - 2logL \tag{2.41}$$

όπου $d$ είναι το πλήθος των παραμέτρων του μοντέλου και $L$ η μεγιστοποιημένη τιμή της συνάρτησης πιθανοφάνειας για το εκτιμηθέν μοντέλο.

Στην περίπτωση της λογιστικής παλινδρόμησης το κριτήριο BIC μετασχηματίζεται στην εξίσωση:

$$BIC = 2\sum_{i=1}^{n}\left[log\big(1+e^{\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}\big) - y_i\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right] + plogn \tag{2.42}$$

Σαν κριτήριο επιλογής του βέλτιστου μοντέλου, όπως και στο κριτήριο AIC, γίνεται σύγκριση μεταξύ των υποψήφιων μοντέλων και τελικά επιλέγεται το μοντέλο εκείνο που έχει τη μικρότερη τιμή BIC. Τέλος, και σε αυτήν την περίπτωση ισχύει η ποινή $d$ που έχει ως σκοπό την αποθάρρυνση του λεγόμενου overfitting.

### Κριτήρια $R^2$

Συνήθως, ο συντελεστής προσδιορισμού $R^2$ εκφράζει την επιτυχία του μοντέλου στη εξήγηση της συμπεριφοράς της μεταβλητής απόκρισης. Με άλλα λόγια, εκφράζεται το ποσοστό μεταβλητότητας της εξαρτημένης μεταβλητής, μέσα από το μοντέλο που προσαρμόζεται σε σύγκριση με ένα άλλο μοντέλο, το μηδενικό ή αυτό που περιλαμβάνει μόνο το σταθερό όρο $\beta_0$. Επομένως, είναι λογικό, να μπορεί να χρησιμοποιηθεί και για την εξερεύνηση του βέλτιστου μοντέλου ανάμεσα στα διάφορα εναλλακτικά μοντέλα. Γενικά, επειδή ο δείκτης αυτός βασίζεται σε αθροίσματα τετραγώνων της εξαρτημένης μεταβλητής, ενδέχεται η χρήση του να μην έχει νόημα στα γενικευμένα γραμμικά μοντέλα. Συνεπώς, η ίδια υπόθεση ισχύει και για το μοντέλο της λογιστικής παλινδρόμησης. Δεν παύει όμως να αποτελεί έναν χρησιμότατο δείκτη, ακόμα και για τα γενικευμένα γραμμικά μοντέλα, έπειτα από την κατάλληλη επέκτασή του.

Στο πολλαπλό γενικό γραμμικό μοντέλο, ο συντελεστής προσδιορισμού $R^2$ εκφράζεται μέσα από τη σχέση:

$$R^2 = 1 - \frac{SSE}{SST} \tag{2.43}$$

όπου $SSE$ είναι το άθροισμα τετραγώνων των υπολοίπων και $SST$ το συνολικό άθροισμα τετραγώνων. Επειδή το $SST$ προκύπτει από το γραμμικό μοντέλο που περιέχει μόνο ένα σταθερό όρο, ο McFadden (1974) πρότεινε ένα φυσικό ανάλογο του $R^2$, που ονομάζεται **ψευδό-$R^2$** (*pseudo*-$R^2$) και δίνεται από τον τύπο:

$$R_L^2 = 1 - \frac{l(\boldsymbol{\beta})}{\hat{l}_0} \tag{2.44}$$

όπου $l(\boldsymbol{\beta})$ είναι η μεγιστοποιημένη λογαριθμοποιημένη πιθανοφάνεια για το κατα προσαρμογή μοντέλο και $\hat{l}_0$ είναι η μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας για το μοντέλο που περιέχει μόνο το σταθερό όρο. Όταν το μοντέλο δεν περιέχει επεξηγηματικές μεταβλητές, το κριτήριο $R^2$ λαμβάνει την τιμή μηδέν. Όσο εισάγουμε μεταβλητές η τιμή τού δείκτη αυξάνεται και μπορεί να φτάσει έως την μέγιστη τιμή 1, που τη λαμβάνει στην περίπτωση του κορεσμένου μοντέλου. Το κριτήριο αυτό μπορεί να χρησιμοποιηθεί για οποιοδήποτε μοντέλο που έχει προσαρμοστεί με τη μέθοδο της μέγιστης πιθανοφάνειας.

Στην περίπτωση της λογιστικής παλινδρόμησης, για την επιλογή του καταλληλότερου μοντέλου, χρησιμοποιούνται τα δύο, ευρέως πιο γνωστά, κριτήρια αυτής της κατηγορίας. Το ψευδό-$R_M^2$ και ο διορθωμένος συντελεστής προσδιορισμού του προαναφερθέντα, $R_N^2$. Ο διορθωμένος συντελεστής προσδιορισμού $R_N^2$ χρησιμοποιείται για δεδομένα που ακολουθούν τη Διωνυμική κατανομή και η μελέτη του ξεφεύγει από τα πλαίσια αυτής της εργασίας. Η προσοχή θα επικεντρωθεί στο δείκτη ψευδό-$R_M^2$. Τη δεκαετία του ΄80, λοιπόν, οι Maddala (1983), Cox και Snell (1989) και Magee (1990), σύστησαν στον επιστημονικό κόσμο μία τροποποίηση του συντελεστή προσδιορισμού για το γενικό γραμμικό μοντέλο.

Έστω $M_0$ και $M_1$ δύο μοντέλα, με το $M_0$ να είναι το μοντέλο που περιέχει μόνο το σταθερό όρο. Εάν υποτεθεί ότι αυτά τα δύο είναι εμφωλευμένα το ένα στο άλλο,

$M_0 \subset M_1$, τότε η ελεγχοσυνάρτηση του λόγου των πιθανοφανειών είναι:

$$\begin{aligned} LR =& -2(\hat{l}_0 - \hat{l}_1) \\ =& -2\left(-\frac{n}{2}\right)(logSST - logSSE) \\ =& nlog\frac{SST}{SSE}, \end{aligned} \tag{2.45}$$

όπου $\hat{l}_0$ είναι η μεγιστοποιημένη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας του μοντέλου που περιέχει μόνο το σταθερό όρο, χωρίς επιπλέον συμμεταβλητές. Συγκρίνοντας τη σχέση (2.45) με αυτή του συντελεστή προσδιορισμού $R^2$ (σχ. 2.43), είναι φανερή η ομοιότητα στην ποσότητα $\frac{SSE}{SST}$. Έτσι, ορίζεται ο δείκτης:

$$R^2_{LR} = 1 - e^{-LR/n} \tag{2.46}$$

για τα γενικευμένα γραμμικά μοντέλα.

Με βάση τα παραπάνω, ισχύει:

$$\begin{aligned} \hat{l}_0 - \hat{l}_1 = log\left(\frac{\hat{L}_0}{\hat{L}_1}\right) =& -\frac{n}{2}\left(log\frac{SST}{SSE}\right) \\ =& log\left(\frac{SSE}{SST}\right)^{n/2}, \end{aligned} \tag{2.47}$$

από όπου και προκύπτει ο δείκτης ψευδό-$R^2_M$ για τα γενικευμένα γραμμικά μοντέλα, συνεπώς και για το μοντέλο της λογιστικής παλινδρόμησης, και δίνεται από τη σχέση:

$$R^2_M = 1 - \left(\frac{\hat{L}_0}{\hat{L}_1}\right)^{2/n} \tag{2.48}$$

Επίσης ο συγκεκριμένος συντελεστής προσδιορισμού μπορεί να χρησιμοποιηθεί για ένα οποιοδήποτε μοντέλο προσαρμοσμένο με τη μέθοδο της μέγιστης πιθανοφάνειας.

Το μέτρο αυτό, παρ'όλο που δημιουργήθηκε για τις ανάγκες δυαδικών δεδομένων, συχνά παρουσιάζει χαμηλές τιμές στη λογιστική παλινδρόμηση, ανεξαρτήτως της καλής προσαρμογής του εκάστοτε μοντέλου. Με άλλα λόγια, οι διάφοροι δείκτες καλής προσαρμογής και συμπεριφοράς ενός μοντέλου μπορεί να εμφανίζουν ικανοποιητικά αποτελέσματα, ενώ το αριθμητικό αποτέλεσμα του συντελεστή προσδιορισμού να παραμένει χαμηλό. Αυτό συμβαίνει, διότι το μοντέλο εξηγεί ή προβλέπει μόνο την πιθανότητα επιτυχίας $\pi = E(Y)$ και όχι τις ατομικές τιμές $y$ (0 ή 1). Δεδομένης της $\pi$, η επιτυχία ή αποτυχία είναι ένα τυχαίο γεγονός που το μοντέλο δεν μπορεί να προβλέψει. Συνεπώς, ένα μεγάλο μέρος της συνολικής μεταβλητότητας μένει ανεξήγητο, και άρα ένας τέτοιου τύπου δείκτης αναγκαστικά παίρνει χαμηλή τιμή.

## 2.6 Ο Μετασχηματισμός Probit

Η ιδέα της ανάλυσης Probit δημοσιεύθηκε αρχικά στο περιοδικό Science από το Chester Ittner Bliss, το 1934. Ο Bliss δούλευε ως εντομολόγος στο Connecticut πάνω

σε ένα πείραμα που αφορούσε την εξερεύνηση ενός αποτελεσματικού φυτοφαρμάκου για τον έλεγχο των εντόμων που τρέφονται με φύλλα σταφυλιών. Καταγράφοντας την αντίδραση των εντόμων σε διάφορες συγκεντρώσεις παρασιτοκτόνων, παρατήρησε ότι τα παρασιτοκτόνα επηρέαζαν τα έντομα σε διαφορετικές συγκεντρώσεις το καθένα. Ωστόσο, δεν διέθετε κάποια σωστή στατιστική μέθοδο που θα μπορούσε να συγκρίνει τις διαφορές. Η πιο λογική προσέγγιση ήταν να προσαρμόσει ένα μοντέλο παλινδρόμησης της αντίδρασης των εντόμων έναντι της συγκέντρωσης ή της δόσης, και να συγκρίνει τα αποτελέσματα ανάμεσα στα διαφορετικά παρασιτοκτόνα. Αυτό που παρατήρησε ήταν ότι η σχέση μεταξύ της αντίδρασης και της δόσης δημιουργούσαν γραφικά μία καμπύλη με C-σχήμα και κατα την τότε χρονική στιγμή, η παλινδρόμηση χρησιμοποιούταν μόνο για δεδομένα που ακολουθούσαν γραμμική συμπεριφορά. Ως εκ τούτου, ο Bliss ανέπτυξε την ιδέα του μετασχηματισμού της εν λόγω καμπύλης σε μία ευθεία γραμμή. Λίγο αργότερα, και βασιζόμενος στην ιδέα του Bliss, ένας καθηγητής στατιστικής στο πανεπιστήμιο του Εδιμβούργου, εν ονόματι David Finney (1952), έγραψε ένα βιβλίο με τίτλο "Probit Analysis". Σήμερα, η ανάλυση Probit είναι η πιο διαδεδομένη στατιστική μέθοδος όταν πρόκειται για προβλήματα δόσης/αντίδρασης.

Η ανάλυση Probit, λοιπόν, χρησιμοποιείται συνήθως στον τομέα της τοξικολογίας για τον προσδιορισμό της σχετικής τοξικότητας των χημικών ουσιών σε ζωντανούς οργανισμούς. Αυτό πραγματοποιείται ελέγχοντας την αντίδραση ενός οργανισμού κάτω από ποικίλες συγκεντρώσεις των εν λόγω χημικών ουσιών και κατόπιν συγκρίνοντας τα αποτελέσματα. Γενικά, η απόκριση είναι πάντα διωνυμικής φύσης (π.χ. θάνατος/επιβίωση) και η σχέση μεταξύ της αντίδρασης και των συγκεντρώσεων είναι πάντα σιγμοειδής (C-σχήμα). Η μέθοδος Probit λειτουργεί σαν μετασχηματισμός της καμπύλης σε γραμμή και έπειτα τρέχει την παλινδρόμηση στη σχέση. Μόλις εκτελεστεί η παλινδρόμηση, ο ερευνητής μπορεί να χρησιμοποιήσει το αποτέλεσμα της ανάλυσης Probit με σκοπό, να συγκρίνει την ποσότητα της χημικής ουσίας που απαιτείται, ώστε να δημιουργηθεί η ίδια αντίδραση σε κάθε μία από τις διάφορες χημικές ουσίες.

Ο σκοπός αυτού του μοντέλου είναι να εκτιμηθεί η πιθανότητα που μία παρατήρηση, με συγκεκριμένα χαρακτηριστικά, θα εμπίπτει σε μία από τις δύο κατηγορίες της εξαρτημένης μεταβλητής $Y$. Με άλλα λόγια, η μέθοδος Probit αποτελεί ένα τύπο δυαδικού μοντέλου ταξινόμησης με βάση τις προβλεπόμενες πιθανότητες. Το μοντέλο αυτό αποτελεί ένα δημοφιλή τρόπο ανάλυσης για δεδομένα με διατεταγμένη ή δυαδική μεταβλητή απόκρισης. Ουσιαστικά, χειρίζεται τα δεδομένα με τον ίδιο τρόπο που το κάνει και η λογιστική παλινδρόμηση, χρησιμοποιώντας τις ίδιες τεχνικές. Για να εκτιμηθεί το μοντέλο probit, το οποίο χρησιμοποιεί τη συνάρτηση σύνδεσης probit (*probit link function*), ακολουθείται η διαδικασία της μεθόδου μέγιστης πιθανοφάνειας.

### 2.6.1 Το μοντέλο

Έστω μία μεταβλητή απόκρισης $Y$, η οποία μπορεί να λάβει δύο πιθανά ενδεχόμενα, 0 και 1. Για παράδειγμα, η μεταβλητή $Y$ μπορεί να αφορά την παρουσία/απουσία κάποιας συγκεκριμένης κατάστασης, τον θάνατο/επιβίωση ενός πειραματόζωου κ.ά. Επιπλέον, έστω ένα διάνυσμα επεξηγηματικών μεταβλητών $\boldsymbol{X}$, οι οποίες υποτίθεται ότι επηρεάζουν το αποτέλεσμα $Y$. Το μοντέλο της μεθόδου probit, παίρνει τη μορφή:

$$\begin{aligned} \eta_X &= \boldsymbol{X}^T\boldsymbol{\beta} = \Phi^{-1}(P(Y=1 \mid X)) \\ \Rightarrow & P(Y=1 \mid X) = \Phi(\boldsymbol{X}^T\boldsymbol{\beta}), \end{aligned} \tag{2.49}$$

όπου $\Phi(\bullet)$ είναι η αθροιστική συνάρτηση κατανομής της τυποποιημένης Κανονικής κατανομής. Οι παράμετροι $\boldsymbol{\beta}$ εκτιμώνται, όπως και στη λογιστική παλινδρόμηση, μέσω της μεθόδου της μέγιστης πιθανοφάνειας.

### 2.6.2 Εκτίμηση των συντελεστών του μοντέλου

Στην παλινδρόμηση probit, όπως και στη λογιστική παλινδρόμηση, οι συντελεστές του μοντέλου εκτιμούνται μέσω της μεθόδου μέγιστης πιθανοφάνειας. Έτσι, έστω το σύνολο των παρατηρήσεων $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$. Τότε η από κοινού λογαριθμοποιημένη συνάρτηση πιθανοφάνειας θα είναι:

$$logL(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big( y_i log\Phi(\boldsymbol{x}_i^T\boldsymbol{\beta}) + (1-y_i)log\big(1-\Phi(\boldsymbol{x}_i^T\boldsymbol{\beta})\big)\Big) \tag{2.50}$$

Ο εκτιμητής $\hat{\boldsymbol{\beta}}$, ο οποίος μεγιστοποιεί αυτή τη συνάρτηση θα είναι σταθερός και ασυμπτωτικά κανονικός. Μπορεί, επίσης, να δειχθεί ότι αυτή η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας είναι κοίλη στο $\boldsymbol{\beta}$ και ως εκ τούτου οι τυπικοί αριθμητικοί αλγόριθμοι βελτιστοποίησης θα συγκλίνουν στο μέγιστο.

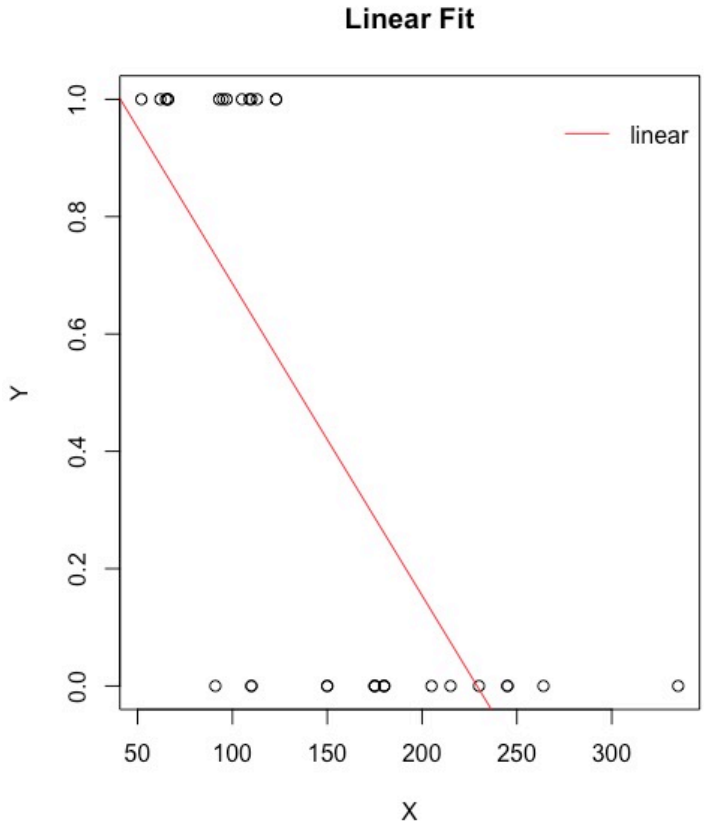
### 2.6.3 Έλεγχοι καλής προσαρμογής του μοντέλου

Ο τρόπος που μπορεί κάποιος να εκτιμήσει εάν το μοντέλο προσαρμόστηκε ικανοποιητικά για τα διαθέσιμα δεδομένα είναι αυτός του πίνακα ταξινόμησης. Δηλαδή, με άλλα λόγια, η μέθοδος αυτή βασίζεται στην απαρίθμηση των πραγματικών παρατηρήσεων που είναι κατηγοριοποιημένες στο κάθε γεγονός, σε σύγκριση με τις παρατηρήσεις που έχουν προβλεφθεί από το προσαρμοσμένο μοντέλο. Η ταξινόμηση γίνεται μέσω της προβλεφθείσας πιθανότητας όπου εξ'αρχής θέτεται ένα σημείο διαχωρισμού, συνήθως το $0.5$ και ανάλογα με αυτό η κάθε παρατήρηση κατηγοριοποιείται αντιστοίχως. Η μέθοδος αυτή περιγράφεται αναλυτικά στην παράγραφο 2.5.8 όπου πραγματοποιείται με βάση τη λογιστική παλινδρόμηση. Παρ'όλα αυτά, δεν αλλάζει η εφαρμογή της στο μοντέλο probit.

## 2.7 Επιλογή μεθόδων ανάμεσα σε Logit και Probit

Για να γίνει κατανοητή η διαφορά μεταξύ των συναρτήσεων σύνδεσης Logit και Probit χρειάζεται κάτι παραπάνω από μία ματιά στους μαθηματικούς τους τύπους. Η ερώτηση: "Ποιό μοντέλο ταιριάζει καταλληλότερα στα δεδομένα μου;", είναι η πιο συχνή όταν πρόκειται για διακριτές παρατηρήσεις. Οι συναρτήσεις σύνδεσης που χρησιμοποιούνται για μία διακριτή μεταβλητή απόκρισης, είναι τρεις:

- η συνάρτηση logit,
- η συνάρτηση probit και
- η συνάρτηση complementary log-log (cloglog).

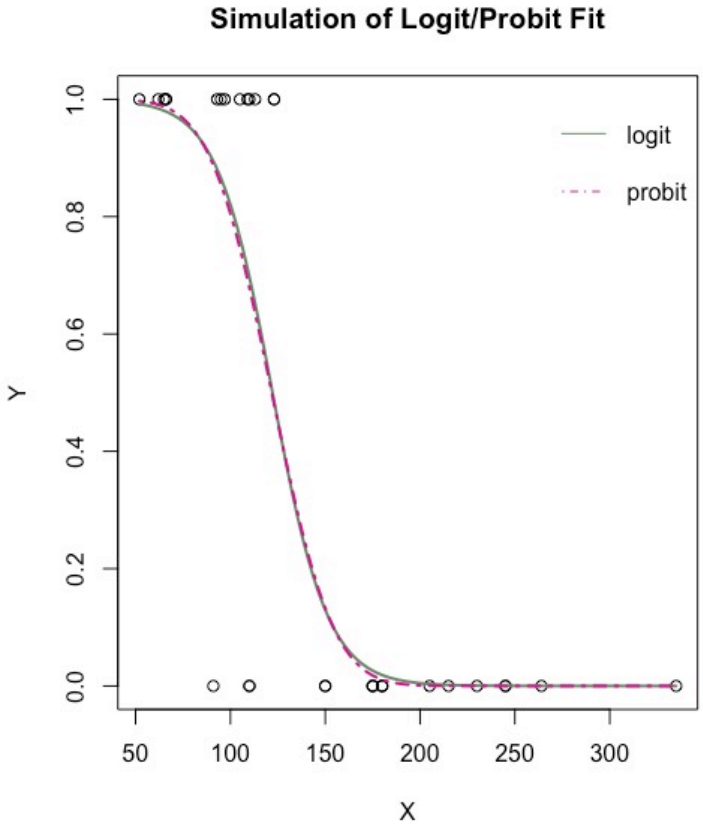



Σχήμα 2.4: Προσαρμογή γραμμικού μοντέλου σε διωνυμικά δεδομένα.

Από αυτές τις τρεις συναρτήσεις, στην παρούσα μελέτη, χρησιμοποιούνται οι Logit και Probit.

Το κοινό πρόβλημα που καλούνται να λύσουν οι δύο αυτοί μετασχηματισμοί είναι η προσαρμογή της θεωρητικής γραμμής της γραμμικής παλινδρόμησης σε διακριτά δεδομένα. Έτσι, αν παρθούν δεδομένα από την Διωνυμική κατανομή και πραγματοποιηθεί προσπάθεια να προσαρμοστεί ένα γραμμικό μοντέλο πάνω σε αυτά, έχει ως αποτέλεσμα την εικόνα που παρουσιάζεται στο σχήμα (2.4). Υπάρχουν πολλά προβλήματα που προκύπτουν με αυτήν την εικόνα. Μεταξύ άλλων, είναι ότι η γραμμή της παλινδρόμησης μπορεί να οδηγήσει σε προβλέψεις εκτός του διαστήματος [0,1]. Το πρόβλημα αυτό έρχονται να λύσουν οι συναρτήσεις σύνδεσης, οι οποίες προσαρμόζουν μία μη-γραμμική συνάρτηση στα δεδομένα που μοιάζει με το σχήμα (2.5). Δηλαδή, η ευθεία γραμμή έχει αντικατασταθεί από μία σιγμοειδούς μορφής καμπύλη, όπου:

- Σέβεται τα όρια της μεταβλητής απόκρισης $y$.
- Επιτρέπει διαφορετικούς ρυθμούς αλλαγής στα χαμηλά και υψηλά άκρα της κλιμακας $x$.
- Απομακρύνει το φαινόμενο της ετεροσκεδαστικότητας (υπό την υπόθεση ότι οι επεξηγηματικές μεταβλητές τηρούν τις κατάλληλες προδιαγραφές).




Σχήμα 2.5: Προσαρμογή Logit και Probit μοντέλων για διωνυμικά δεδομένα.

Οι συναρτήσεις σύνδεσης Logit και Probit, στην ουσία παίρνουν το γραμμικό μοντέλο και το μετασχηματίζουν σε μία συνάρτηση που απιδίδει μία μη-γραμμική σχέση. Με άλλα λόγια, εάν το προβλεπτικό γραμμικό μοντέλο έχει τη μορφή:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

το αντίστοιχο προβλεπτικό μοντέλο για τις συναρτήσεις σύνδεσης logit και probit θα είναι:

$$\hat{Y} = f(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)$$

Η διαφορά, τώρα, έγκειται στο πώς οι δύο αυτοί μετασχηματισμοί, ορίζουν την συνάρτηση $f(\bullet)$. Έτσι, για τον ορισμό της $f(\bullet)$, το μοντέλο logit χρησιμοποιεί την αθροιστική συνάρτηση κατανομής της λογιστικής κατανομής (ενότητα 2.4.3),

$$\eta_X = g(\pi(X)) = log\left(\frac{\pi(X)}{1-\pi(X)}\right)$$

ενώ αντιθέτως το μοντέλο probit χρησιμοποιεί την αθροιστική συνάρτηση κατανομής της τυποποιημένης Κανονικής κατανομής (ενότητα 2.6),

$$\eta_X = g(\pi(X)) = \Phi^{-1}(\frac{\pi(X)}{n})$$

Και οι δύο συναρτήσεις, παίρνουν οποιοδήποτε αριθμό και τον προσαρμόζουν κατάλληλα ώστε να εμπίπτει στο κλειστό διάστημα [0,1]. Επομένως, οτιδήποτε εκφράζεται από την ποσότητα $\alpha + \beta X$, μπορεί να μετασχηματιστεί έτσι, ώστε να υπολογίζεται μία προβλεπόμενη πιθανότητα.

Τι γίνεται όμως όταν καλούμαστε να επιλέξουμε τον καταλληλότερο μετασχηματισμό; Δηλαδή, ποια είναι η καλύτερη επιλογή ανάμεσα στην τυποποιημένη Κανονική κατανομή και την αντίστοιχη λογιστική; Η σύντομη απάντηση που δίνεται είναι ότι τα αποτελέσματα των δύο μοντέλων δε διαφέρουν πολύ. Αυτό είναι φανερό και από το σχήμα (2.5), όπου μπορεί κανείς να δει ότι οι δύο συναρτήσεις σχεδόν συμπίπτουν στην προσαρμογή του μοντέλου. Έτσι, μπορεί να θεωρηθεί ότι είναι δυνατή η τυχαία επιλογή ανάμεσα στις συναρτήσεις logit και probit. Ωστόσο, η απάντηση δεν είναι τόσο απλή.

Αρχικά, να σημειωθεί ότι, τα αποτελέσματα δε διαφέρουν σε μεγάλο βαθμό μόνο όταν πρόκειται για μεταβλητή απόκρισης δυαδικής φύσεως. Στην περίπτωση μίας πολυωνυμικής μεταβλητής (αποτελείται από τρεις και άνω κατηγορίες) τα πράγματα αλλάζουν θεμελιωδώς. Στην πράξη, το πολυωνυμικό μοντέλο της Logit είναι υπολογιστικά ευκολότερο, από αυτό της Probit, και αυτός μπορεί να είναι ένας λόγος επιλογής κάποιου συγκεκριμένου μετασχηματισμού από τους δύο.

Εάν, τώρα, η μελέτη επικεντρωθεί αυστηρά στο δυαδικό μοντέλο (δύο κατηγορίες: 0 και 1), τότε υπάρχουν κάποιες διαφορές μεταξύ των αποτελεσμάτων των Logit και Probit μοντέλων. Οι διαφορές αυτές είναι ορατές όταν η προσοχή στραφεί στις ουρές των υποκείμενων κατανομών. Μία από τις λίγες μελέτες που αφορούν, την αξιολόγηση κριτηρίων πληροφόρησης της αποτελεσματικότητας, μεταξύ των μοντέλων Logit και Probit, είναι αυτή των Chen και Tsurumi (2010). Τα εν λόγω κριτήρια είναι τα εξής:

- Το κριτήριο της διακύμανσης (*DIC, Deviance Information Criterion*)
- Το κριτήριο της προβλεπόμενης διακύμανσης (*PDIC, Predictive Deviance Information Criterion*)
- Το μη-σταθμισμένο άθροισμα τετραγώνων των σφαλμάτων (*USSE, Unweighted Sum of Squared Errors*)
- Το σταθμισμένο άθροισμα τετραγώνων των σφαλμάτων (*WSSE, Weighted Sum of Squared Errors*)
- Το κριτήριο AIC (*AIC, Akaike's Information Criterion*)
- Το κριτήριο BIC (*BIC, Bayesian Information Criterion*)

Τα βασικά συμπεράσματα που προκύπτουν από τη δημοσιευμένη μελέτη των Chen-Tsurumi, έχουν ως εξής:

- Εάν τα δυαδικά δεδομένα που μοντελοποιούνται είναι "ισορροπημένα" (δηλαδή, εάν είναι κατανεμημένα 50-50 μεταξύ του 0 και του 1), τότε κανένα από τα παραπάνω κριτήρια δεν είναι αποτελεσματικό στον διαχωρισμό μεταξύ των μοντέλων Logit και Probit.
- Εάν τα δεδομένα είναι ασύμμετρα κατανεμημένα μεταξύ του 0 και του 1, τότε τα καταλληλότερα κριτήρια επιλογής είναι αυτά της διακύμανσης και του Akaike (*DIC* και *AIC*).

- Είναι απαραίτητο το μέγεθος του δείγματος να ξεπερνάει τις 1000 παρατηρήσεις, ώστε να είναι δυνατή η επιλογή μεταξύ των μοντέλων Logit και Probit, χρησιμοποιώντας αυτές τις προσεγγίσεις.

Στην περίπτωση που τα παραπάνω κριτήρια δεν είναι βοηθητικά στην επιλογή του μοντέλου, υπάρχουν και άλλες τεχνικές που μπορεί να χρησιμοποιηθούν. Έτσι, ένας άλλος τρόπος είναι η κατασκευή ενός ελέγχου υποθέσεων (Chambers and Cox , 1967). Οι Chambers και Cox θεώρησαν ως μηδενική υπόθεση $H_0$ τη συνάρτηση σύνδεσης Logit και ως εναλλακτική, $H_1$, τη συνάρτηση Probit. Έπειτα κατασκεύασαν έναν έλεγχο με τις υποθέσεις $H_0$ και $H_1$ να εναλλάσονται. Δυστυχώς όμως, παρ'όλο που υπάρχουν ορισμένες μελέτες και εργαλεία με αυτόν τον σκοπό, κανένα δεν είναι απόλυτα έμπιστο, και επιπλέον, απαιτείται μεγάλος όγκος δεδομένων για να είναι αποτελεσματικά.

# ΚΕΦΑΛΑΙΟ 3

# ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΜΕ ΤΗ ΜΕΘΟΔΟ LASSO

Without data, you're just another person with an opinion.

*W. Edwards Deming, Data Scientist*

## 3.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα συζητήθεί και θα αναλυθεί το **πρόβλημα επιλογής μεταβλητών** (*variable selection problem*), που στην πραγματικότητα, πρόκειται για ένα **πρόβλημα επιλογής μοντέλων** (*model selection problem*). Στην πράξη, γίνεται σύγκριση μοντέλων που έχουν την ίδια δομή, αλλά διαφορετικές συμμεταβλητές. Αυτό είναι και το κύριο ερώτημα που καλούνται να απαντήσουν όλες οι μέθοδοι αυτής της κατηγορίας: *η επιλογή των συμμεταβλητών*.

Πότε όμως, ένα μοντέλο θεωρείται καλύτερο από ένα άλλο; Όπως και σε πολλές περιπτώσεις στη στατιστική επιστήμη, η επιλογή του καλύτερου μοντέλου είναι υποκειμενική. Είναι φυσικά αδύνατον να βρεθεί ένα μοναδικό μοντέλο που να ανταποκρίνεται και να αντικατοπτρίζει όλες τις ιδιότητες που απαιτούνται από ένα πρόβλημα. Έτσι, διαφορετικές διαδικασίες υποστηρίζονται από διαφορετικές επιστημονικές θεωρίες. Όποια μεθοδολογία, όμως και να επιλεχθεί πρέπει να τηρούνται δύο βασικές αρχές: *η καλή προσαρμογή του μοντέλου και η οικονομία*.

Για την καλύτερη κατανόηση της λειτουργίας της μεθόδου, θα αναπτυχθεί η θεωρία της, αρχικά στα γραμμικά μοντέλα και στο τέλος θα αναφερθούν οι προεκτάσεις που έχουν δημιουργηθεί και πως αυτή η ιδέα μπορεί να γενικευτεί στα γενικευμένα γραμμικά μοντέλα. Ξεκινώντας με την συνήθη υπόθεση: έστω $p+1$ ποσοτικές μεταβλητές, $\{\boldsymbol{x}^i, y_i\}$, $i = 1, 2, \ldots, n$, όπου $\boldsymbol{x^i} = (x_{i1}, \ldots, x_{ip})^T$ είναι οι επεξηγηματικές ή ανεξάρτητες συμμεταβλητές και $y_i$ είναι η μεταβλητή απόκρισης για την $i$-οστή παρατήρηση. Επιπλέον, έστω το προσαρμοσμένο μοντέλο της πολλαπλής γραμμικής παλινδρόμησης:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2).$$

Αρχικά, υπενθυμίζεται ότι η εκτίμηση των συντελεστών ενός μοντέλου παλινδρόμησης, μέσω της μεθόδου των ελάχιστων τετραγώνων (ε.τ.), επιτυγχάνεται ελαχιστοποιώντας το άθροισμα τετραγώνων των υπολοίπων, δηλαδή το:

$$SS = \sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}\boldsymbol{x}^i\right)^2$$

όπου $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ και $\boldsymbol{x}^i = (1, x_{i1}, \ldots, x_{ip}), i = 1, 2, \ldots, n$. Υπάρχουν δύο λόγοι που ένας στατιστικός δεν είναι συχνά ικανοποιημένος με τους συγκεκριμένους εκτιμητές. Ο πρώτος λόγος αφορά την ακρίβεια στην πρόβλεψη· δηλαδή οι εκτιμητές ελάχιστων τετραγώνων (ε.ε.τ.) συνήθως έχουν μικρή μεροληψία (bias) αλλά μεγάλη διασπορά, ενώ ο δεύτερος λόγος αφορά την ερμηνεία των αποτελεσμάτων. Το πρώτο πρόβλημα μπορεί να βελτιωθεί μέσω της **συρρίκνωσης** (*shrinking*) των συντελεστών του μοντέλου, ή θέτοντας κάποιους συντελεστές με μηδέν. Με αυτόν τον τρόπο, θυσιάζεται λίγο από τη μεροληψία, με σκοπό να μειωθεί η διασπορά των προβλεπόμενων τιμών και έτσι να βελτιωθεί συνολικά η ακρίβεια στην πρόβλεψη. Για το δεύτερο πρόβλημα, ο στατιστικός συχνά επιθυμεί, μέσα από ένα τεράστιο όγκο συμμεταβλητών, να επιλέξει το μικρότερο υποσύνολο, το οποίο παρουσιάζει τα ισχυρότερα αποτελέσματα.

Οι δύο συνηθέστερες τεχνικές για την βελτίωση των ε.ε.τ. είναι η **επιλογή υποσυνόλου** (*subset selection*) και η **παλινδρόμηση κορυφογραμμής** (*ridge regression*). Παρ'όλα αυτά, και οι δύο τεχνικές έχουν μειονεκτήματα. Η μέθοδος επιλογής υποσυνόλου είναι μία διακριτή διαδικασία, η οποία παρέχει πιο ερμηνεύσιμα αποτελέσματα, αλλά μπορεί να είναι υπερβολικά μεταβλητή. Δηλαδή οι συμμεταβλητές είτε διατηρούνται ή αφαιρούνται πλήρως από το μοντέλο. Μικρές αλλαγές πάνω στα δεδομένα μπορεί να οδηγήσουν στην επιλογή πολύ διαφορετικών μοντέλων και αυτό έχει σαν αποτέλεσμα να μειώνεται η ακρίβεια στην εκάστοτε πρόβλεψη. Από την άλλη μεριά, η παλινδρόμηση κορυφογραμμής είναι μία συνεχής διαδικασία, η οποία συρρικνώνει τους συντελεστές και για αυτόν τον λόγο είναι πιο σταθερή μέθοδος. Όμως, η τεχνική αυτή δε θέτει κανένα συντελεστή με το μηδέν και αυτό οδηγεί σε ένα μοντέλο που δεν ερμηνεύεται τόσο εύκολα όσο στην προηγούμενη διαδικασία.

Αυτά τα προβλήματα ήρθε να λύσει μία νέα μέθοδος που ονομάζεται μέθοδος **LASSO** (*Least Absolute Shrinkage and Selection Operator*). Το πλεονέκτημα αυτής της μεθόδου είναι ότι κάνει χρήση των θετικών στοιχείων των δύο παραπάνω τεχνικών, της μεθόδου επιλογής υποσυνόλου και της παλινδρόμησης κορυφογραμμής. Με άλλα λόγια, η lasso συρρικνώνει κάποιους συντελεστές, ενώ θέτει άλλους ίσους με μηδέν. Προτού όμως παρουσιαστούν αναλυτικά οι ιδιότητες της μεθόδου lasso, θα γίνει μία εισαγωγή στο φαινόμενο της πολυσυγγραμμικότητας, καθώς στις δύο αυτές τεχνικές.

## 3.2 Το φαινόμενο της πολυσυγγραμμικότητας

Όπως αναφέρθηκε σε προηγούμενο κεφάλαιο, **το φαινόμενο της πολυσυγγραμμικότητας** (*multicollinearity*) μπορεί να επηρεάσει αρνητικά τη σωστή εκτίμηση του μοντέλου. Στην ουσία, με τον όρο "πολυσυγγραμμικότητα" περιγράφεται η (στατιστικώς) υψηλή γραμμική συσχέτιση μεταξύ των επεξηγηματικών μεταβλητών. Η παρουσία αυτού του φαινομένου οδηγεί σε αυξημένα τυπικά σφάλματα των εκτιμημένων συντελεστών $\hat{\boldsymbol{\beta}}$. Αυτό έχει ως συνέπεια την αστάθεια των εκτιμήσεων και δυσκολεύει τον εντο-

πισμό των στατιστικά σημαντικών μεταβλητών. Σε πολλές αντίστοιχες περιπτώσεις, είναι πιθανό στατιστικά σημαντικοί παράγοντες να παρουσιάζονται σαν να μην επηρεάζουν το αποτέλεσμα. Έτσι, στις περιπτώσεις πολυσυγγραμμικότητας παρατηρείται αλλοίωση των αποτελεσμάτων και πολλές φορές δίνονται σημάδια εντελώς αντίθετων επιρροών.

Όταν δύο συμμεταβλητές έχουν υψηλή συσχέτιση, ουσιαστικά αυτό σημαίνει ότι φέρουν ίδιες πληροφορίες. Έτσι, γνωρίζοντας την τιμή μίας επεξηγηματικής μεταβλητής, μπορεί να προβλεφθεί με ακρίβεια η τιμή της άλλης. Αυτό έχει σαν αποτέλεσμα, τέτοιες μεταβλητές να μην προσθέτουν επιπλέον πληροφορία όσον αφορά την επίδρασή τους στη μεταβλητή απόκρισης $Y$. Το ίδιο φαινόμενο παρατηρείται και όταν μία συμμεταβλητή αποτελεί γραμμική συνάρτηση περισσοτέρων της μίας μεταβλητής. Έστω, για παράδειγμα, το γραμμικό μοντέλο:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

και έστω ότι η επεξηγηματική μεταβλητή $X_2$ συσχετίζεται γραμμικά με τη $X_1$ μέσω της σχέσης:

$$X_2 = a + bX_1.$$

Σε αυτήν την περίπτωση, δεν είναι δυνατό να ερμηνευτεί το μοντέλο με τον συνήθη τρόπο, καθώς οποιαδήποτε αλλαγή στη μεταβλητή $X_1$ έχει σαν αποτέλεσμα να μεταβληθεί και η τιμή της $X_2$.

Προσπαθώντας να προσεγγιστεί το πρόβλημα με μαθηματικό τρόπο, έστω η εκτίμηση των συντελεστών με τη μέθοδο των ελάχιστων τετραγώνων:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

όπου:

- $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ είναι το διάνυσμα των ε.ε.τ., διάστασης $(p+1) \times 1$.
- $\boldsymbol{X}$ είναι ο πίνακας σχεδιασμού, διάστασης $n \times (p+1)$.
- $\boldsymbol{y}$ είναι το διάνυσμα με τις τιμές της μεταβλητής απόκρισης, διάστασης $n+1$.

Εάν μία εκ των επεξηγηματικών μεταβλητών είναι γραμμικός συνδυασμός των υπολοίπων, ο αντίστροφος πίνακας $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ δεν υπάρχει. Στην πράξη, σπάνια θα παρατηρηθεί τέλεια γραμμική σχέση μεταξύ των επεξηγηματικών μεταβλητών. Εάν μία συμμεταβλητή είναι υψηλά συσχετισμένη με τις υπόλοιπες, αυτό θα έχει σαν αποτέλεσμα ασταθείς εκτιμήσεις και υψηλά τυπικά σφάλματα.

### 3.2.1 Διαγνωστικοί έλεγχοι για την πολυσυγγραμμικότητα

Υπάρχουν πολλοί τρόποι ώστε να εντοπισθεί εάν υπάρχει πρόβλημα πολυσυγγραμμικότητας. Ανάλογα με τη φύση των δεδομένων γίνεται χρήση του αντίστοιχου διαγνωστικού ελέγχου. Έτσι, τα εργαλεία που είναι διαθέσιμα για αυτόν το σκοπό είναι ο δείκτης συσχέτισης Pearson (Pearson's correlations), ο συντελεστής προσδιορισμού $R^2$, ο παράγοντας μεγέθυνσης διασποράς (VIF, Variance Inflation Factors) (βλ. 2.5.10) και τέλος ο έλεγχος των ιδιοτιμών του πίνακα $X^TX$.

**Δείκτης συσχέτισης Pearson**

Ο δείκτης συσχέτισης Pearson είναι δυνατό να εμφανίσει την υψηλή γραμμική συσχέτιση μεταξύ δύο συμμεταβλητών, αλλά αποτυγχάνει στην περίπτωση που ο γραμμικός συνδυασμός περιλαμβάνει περισσότερες μεταβλητές. (π.χ. $X_1 = X_2 + X_3 + X_4$).

**Παράγοντας μεγένθυνσης διασποράς (*VIF, Variance Inflation Factors*)**

Όπως έχει παρουσιαστεί και στο προηγούμενο κεφάλαιο, ο δείκτης αυτός, δίνεται από τον τύπο:

$$VIF_j = (1 - R_j^2)^{-1}$$

Με $R_j^2$ συμβολίζεται ο συντελεστής προσδιορισμού, όταν προσαρμόσουμε το μοντέλο παλινδρόμησης, με μεταβλητή απόκρισης τη συμμεταβλητή $X_j$ και επεξηγηματικές μεταβλητές, τις υπόλοιπες εκ των συμμεταβλητών. Στην περίπτωση που ο δείκτης $VIF_j$ είναι μεγαλύτερος του 10, ($VIF_k > 10$), τότε υπάρχει πιθανό πρόβλημα πολυσυγγραμμικότητας. Ουσιαστικά, η τετραγωνική ρίζα του δείκτη VIF περιγράφει το πόσο μεγαλύτερο είναι το τυπικό σφάλμα, σε σύγκριση με το τι θα μπορούσε να ήταν εάν η συγκεκριμένη μεταβλητή ήταν ασυσχέτιστη με τις υπόλοιπες συμμεταβλητές του μοντέλου.

**CI, Condition Indexes**

Η μεθοδολογία είναι σχετικά απλή. Στην αρχή υπολογίζονται οι ιδιοτιμές του πίνακα $\boldsymbol{X}^T\boldsymbol{X}$ και αυτές που είναι κοντά στο μηδέν υποδεικνύουν πρόβλημα. Ο μαθηματικός τύπος της μεθόδου είναι:

$$CI_i = \sqrt{\frac{\lambda_{max}}{\lambda_i}}, i = 1, ..., n,$$

όπου με $\lambda_{max}$ συμβολίζεται η μέγιστη ιδιοτιμή του πίνακα $\boldsymbol{X}^T\boldsymbol{X}$ και με $\lambda_i$ συμβολίζεται η αντίστοιχη ιδιοτιμή $i$. Τα αριθμητικά αποτελέσματα του δείκτη CI είναι και αυτά που υποδεικνύουν την πιθανή ύπαρξη του φαινομένου. Δηλαδή:

- Εάν $CI > 30$, τότε υπάρχει σημαντικό πρόβλημα πολυσυγγραμμικότητας.
- Ενώ, εάν $CI > 15$, τότε το πρόβλημα είναι πιθανό, αλλά όχι βέβαιο.

Όταν οι ιδιοτιμές έχουν χαμηλές τιμές, οι υψηλές τιμές στα αντίστοιχα ιδιοδιανύσματα υποδεικνύουν τις μεταβλητές που συμμετέχουν στους γραμμικούς συνδυασμούς.

### 3.2.2 Τρόποι αντιμετώπισης της πολυσυγγραμμικότητας

Στις περιπτώσεις αυτές λοιπόν, όπου η πολυσυγγραμμικότητα οφείλεται στην έντονη συσχέτιση των μεταβλητών, η ανάλυση παλινδρόμησης μπορεί να πραγματοποιηθεί αφού αφαιρεθεί μία μεταβλητή από το γραμμικά εξαρτημένο σύνολο. Αυτό δεν

είναι πάντα εύκολο, μιάς και συχνά μπορεί να παρουσιάζονται έντονες συσχετίσεις μεταξύ των μεταβλητών, χωρίς να είναι απόλυτα γραμμικά εξαρτημένες. Έτσι, σιγά σιγά αφαιρούνται οι μη-στατιστικά σημαντικές μεταβλητές από το μοντέλο ή γίνεται καλύτερη επιλογή συνδυασμού αυτών, με σκοπό την κατασκευή του καταλληλότερου στατιστικού μοντέλου. Αυτή η διαδικασία ονομάζεται **συρρίκνωση** (*shrinkage*) του μοντέλου.

Κατα καιρούς έχουν προταθεί πολλές τεχνικές αντιμετώπισης του φαινομένου της πολυσυγγραμμικότητας. Μεταξύ άλλων είναι και η χρήση εκτιμητικών μεθόδων, πέραν αυτής των ελάχιστων τετραγώνων. Οι Hoerl και Kennard (1970) ήταν από τους πρώτους που απάντησαν στο πρόβλημα αυτό, μέσω της παλινδρόμησης κορυφογραμμής (*ridge regression*). Στη συνέχεια, πιο πρόσφατες έρευνες προτείνουν διάφορες εναλλακτικές τεχνικές που σκοπό έχουν να μειώσουν τον αριθμό των παραμέτρων στο τελικό μοντέλο. Μία από αυτές τις τεχνικές αποτελεί και η μέθοδος LASSO (Tibshirani, 1996), η οποία χρησιμοποιεί μία μη-κυρτή ποινή με αποτέλεσμα η επιλογή των μεταβλητών να γίνεται αυτόματα. Αντιθέτως, η παλινδρόμηση κορυφογραμμής αυτό που κάνει είναι να μειώνει τους εκτιμητές κοντά στο μηδέν. Πιο αναλυτικά, οι μέθοδοι θα παρουσιαστούν στις επόμενες ενότητες.

## 3.3 Επιλογή υποσυνόλου (*Subset selection*)

### 3.3.1 Επιλογή καλύτερου υποσυνόλου (*Best subset selection*)

Για να εφαρμοστεί η μέθοδος **subset selection**, αρχικά προσαρμόζεται ένα ξεχωριστό μοντέλο παλινδρόμησης ελάχιστων τετραγώνων, για κάθε πιθανό συνδυασμό των $p$ συμμεταβλητών. Όταν αυτή η διαδικασία τελειώσει, πραγματοποιείται παρατήρηση όλων των μοντέλων που έχουν προκύψει. Σκοπός είναι να αναγνωριστεί αυτό που είναι το καλύτερο. Το να βρεθεί το καλύτερο μοντέλο ανάμεσα στα $2^p$ πιθανά ενδεχόμενα, δεν είναι ένα τετριμμένο πρόβλημα. Συνήθως, διασπάται σε δύο στάδια, όπως φαίνεται στον παρακάτω αλγόριθμο:

Στον παραπάνω αλγόριθμο, το βήμα (2) προσδιορίζει το καλύτερο μοντέλο για κάθε πιθανό μέγεθος υποσυνόλου, προκειμένου να μειωθεί το πρόβλημα από $2^p$ πιθανά μοντέλα, στα $p + 1$ μοντέλα.

Τώρα, αυτό που μένει είναι να γίνει η επιλογή ανάμεσα σε αυτά τα $p + 1$ μοντέλα. Αυτό το βήμα, πρέπει να γίνει με προσοχή διότι το RSS αυτών των $p + 1$ μοντέλων μειώνεται μονοτονικά και ο συντελεστής προσδιορισμού $R^2$ αυξάνεται σε αντίστοιχο ρυθμό, ανάλογα με την αύξηση του αριθμού των χαρακτηριστικών που περιλαμβάνονται στο μοντέλο. Επομένως, εάν, εν τέλει χρησιμοποιηθούν αυτά τα δύο στατιστικά μέτρα ώστε να επιλεχθεί το καλύτερο μοντέλο, τότε αυτό οδηγεί στο πλήρες μοντέλο. Έτσι, στο βήμα (3) γίνεται χρήση των κριτηρίων που αναφέρθηκαν με σκοπό την επιλογή του καταλληλότερου μοντέλου μεταξύ των $M_0, M_1, \ldots, M_p$.

Παρ'όλο που στον παραπάνω αλγόριθμο χρησιμοποιείται η μέθοδος των ελάχιστων τετραγώνων, η ίδια ιδέα μπορεί να εφαρμοστεί και σε άλλα είδη μοντέλων, όπως είναι αυτό της λογιστικής παλινδρόμησης. Σε αυτή την περίπτωση, αντί τα παραγόμενα μοντέλα να ταξινομηθούν με βάση το RSS (Βήμα (2)), χρησιμοποιείται η συνάρτηση *deviance* (απόκλιση). Η συνάρτηση *deviance* αποτελεί ένα μέτρο που παίζει το ρόλο

**Αλγόριθμος 1** Επιλογή Καλύτερου Υποσυνόλου (*Best Subset Selection*)

1: Έστω $M_0$ το μηδενικό μοντέλο (null model), το οποίο περιλαμβάνει μόνο το σταθερό όρο $\beta_0$. Το συγκεκριμένο μοντέλο απλά προβλέπει το δειγματικό μέσο κάθε παρατήρησης.

2: Για κάθε $k = 1, 2, \ldots, p$:

   i Προσάρμοσε όλα τα $\binom{p}{k}$ μοντέλα που περιλαμβάνουν ακριβώς $k$ συμμεταβλητές.

   ii Διάλεξε τα καλύτερα μεταξύ αυτών των $\binom{p}{k}$ μοντέλων και ονόμασέ τα $M_k$. Σε αυτήν την περίπτωση, ως καλύτερο μοντέλο ορίζεται αυτό που έχει το μικρότερο RSS, ή ισοδύναμα το μεγαλύτερο $R^2$.

3: Διάλεξε το καλύτερο μοντέλο ανάμεσα στα $M_0, M_1, \ldots, M_p$ χρησιμοποιώντας ένα κριτήριο ανάμεσα στα: AIC, BIC ή προσαρμοσμένο συντελεστή $R^2_{adj}$.

του RSS για μία ευρύτερη κατηγορία μοντέλων. Όπως έχει αναφερθεί, η συνάρτηση *deviance* είναι ίση με -2 φορές τη μέγιστη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας, όπου μικρές τιμές αυτής, υποδεικνύουν καλύτερη προσαρμογή του μοντέλου.

Αν και η μέθοδος αυτή αποτελεί μία πολύ ελκυστική πρόταση επιλογής μοντέλων, κυρίως λόγω της ευκολίας στην ερμηνεία της, πάσχει από υπολογιστικά προβλήματα. Ο αριθμός των πιθανών μοντέλων που πρέπει να ληφθούν υπ'οψιν, αναπτύσσεται ταχέως, καθώς το $p$ αυξάνει. Γενικά, υπάρχουν $2^p$ μοντέλα που περιλαμβάνουν υποσύνολα έως και $p$ συμμεταβλητών. Έτσι, εάν $p = 10$, τότε υπάρχουν περίπου 1000 πιθανά μοντέλα, και εάν $p = 20$, τότε υπάρχουν περισσότερες από 1000000 επιλογές, κ.ο.κ.

### 3.3.2 Διαδικασίες επιλογής μοντέλου με βήματα (*Stepwise selection*)

Για υπολογιστικούς λόγους, η μέθοδος επιλογής του καλύτερου υποσυνόλου δεν μπορεί να εφαρμοστεί όταν η τιμή του $p$ είναι πολύ μεγάλη. Επιπλέον, όσο μεγαλύτερο το διάστημα αναζήτησης, τόσο μειώνεται η προβλεπτική ισχύς του μοντέλου, παρ'όλο που βραχυπρόθεσμα μπορεί να δίνει ικανοποιητικά αποτελέσματα. Έτσι, ένα μεγάλο διάστημα αναζήτησης μπορεί να οδηγήσει στο λεγόμενο overfitting (παρουσία θορύβου) και σε υψηλή διασπορά των εκτιμημένων συντελεστών. Για αυτούς τους δύο λόγους, λοιπόν, οι **διαδικασίες επιλογής μοντέλου με βήματα** (*stepwise methods*) αποτελούν μία ελκυστική εναλλακτική τεχνική, η οποία περιορίζει πολύ περισσότερο το σύνολο των υποψήφιων μοντέλων προς επιλογή.

#### Διαδικασία της προς-τα-εμπρός επιλογής (*Forward stepwise selection*)

Η **διαδικασία της προς-τα-εμπρός επιλογής** (*forward stepwise selection*) είναι μία αποτελεσματική εναλλακτική της μεθόδου επιλογής καλύτερου υποσυνόλου. Αυτό διότι προσφέρει ευκολότερες υπολογιστικές διαδικασίες για μεγάλες τιμές του $p$. Ενώ η διαδικασία επιλογής του καλύτερου υποσυνόλου λαμβάνει υπόψιν όλα τα $2^p$ υποψήφια

μοντέλα, η προς τα εμπρός διαδικασία επιλογής εξετάζει ένα κατα πολύ μικρότερο, σύνολο μοντέλων. Η μέθοδος ξεκινάει από το μοντέλο $y = \beta_0$ και προσθέτει κάθε φορά από μία συμμεταβλητή στο μοντέλο έως ότου εισαχθούν όλες οι επεξηγηματικές μεταβλητές. Πιο συγκεκριμένα, σε κάθε βήμα εισάγεται η μεταβλητή εκείνη που δίνει τη μεγαλύτερη στατιστικά σημαντική τιμή για την προσαρμογή του μοντέλου. Τυπικά, η μέθοδος της προς τα εμπρός επιλογής παρουσιάζεται στον παρακάτω αλγόριθμο:

**Αλγόριθμος 2** Διαδικασία της προς-τα-εμπρός επιλογής (*Forward stepwise selection*)

1: Έστω $M_0$ το μηδενικό μοντέλο, το οποίο περιλαμβάνει μόνο το σταθερό όρο $\beta_0$ και καμία συμμεταβλητή.

2: Για $k = 0, 1, ..., p - 1$ :

i Εξέτασε όλα τα $p - k$ μοντέλα που αυξάνουν την προβλεπτική ισχύ του υποψήφιου μοντέλου $M_k$, το οποίο περιλαμβάνει μία επιπλέον συμμεταβλητή.

ii Διάλεξε το καλύτερο ανάμεσα σε αυτά τα $p - k$ μοντέλα και ονόμασέ το $M_{k+1}$. Σε αυτήν την περίπτωση ως καλύτερο μοντέλο ορίζεται αυτό που δίνει τη μεγαλύτερη μείωση του RSS ή έχει το μεγαλύτερο συντελεστή προσαρμογής $R^2$.

3: Διάλεξε το μοναδικό καλύτερο μοντέλο, ανάμεσα στα $M_0, ..., M_p$, χρησιμοποιώντας ένα κριτήριο ανάμεσα στα: διασταυρωμένο σφάλμα πρόβλεψης (*cross-validated prediction error*), $C_p(AIC), BIC$, προσαρμοσμένο συντελεστή $R^2_{adj}$.

Σε αντίθεση με τη διαδικασία επιλογής του καλύτερου υποσυνόλου, το οποίο αναλάμβανε να προσαρμόσει $2^p$ μοντέλα, η προς-τα-εμπρός διαδικασία επιλογής προσαρμόζει το μηδενικό μοντέλο, μαζί με $p - k$ μοντέλα κατα την $k$-οστή επανάληψη, για $k = 0, ..., p-1$. Συνολικά, όλη αυτή η διαδικασία, λοιπόν, προσαρμόζει $\sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ μοντέλα. Είναι φανερό ότι πρόκειται για ουσιαστική διαφορά καθώς, όταν για παράδειγμα $p = 20$, η διαδικασία επιλογής του καλύτερου υποσυνόλου απαιτεί να προσαρμοστούν $1.048.576$ μοντέλα, ενώ η προς-τα-εμπρός διαδικασία επιλογής απαιτεί να προσαρμοστούν μόλις $211$ μοντέλα.

Στο βήμα 2(β) του αλγορίθμου (2), πιστοποιείται το καλύτερο μοντέλο ανάμεσα σε αυτά τα $p - k$ μοντέλα που προσαυξάνουν το $M_k$ κατά μία επιπλέον συμμεταβλητή. Όπως αναφέρεται και στη διαδικασία, αυτό μπορεί να πραγματοποιηθεί μέσω του RSS ή του συντελεστή προσαρμογής $R^2$. Παρ'όλα αυτά, στο βήμα (3) του αλγορίθμου ζητείται να ευρεθεί το καλύτερο μοντέλο ανάμεσα σε όλο το σύνολο των μοντέλων με τις διαφορετικές μεταβλητές. Αυτή η διαδικασία είναι αρκετά απαιτητική και πραγματοποιείται μέσα από τον υπολογισμό των κατάλληλων μέτρων.

Η πλεονεκτική θέση της προς-τα-εμπρός διαδικασίας έναντι της πρώτης υπολογιστικά, είναι εμφανής. Όμως, ενώ αυτή η μέθοδος τείνει να δίνει καλά αποτελέσματα στην πράξη, δεν εγγυάται ότι θα βρει το καλύτερο δυνατό μοντέλο μέσα από το όλο σύνολο των $2^p$ μοντέλων.

**Διαδικασία της προς-τα-πίσω επιλογής (*Backward stepwise selection*)**

Με παρόμοιο σκεπτικό με αυτό της προηγούμενης μεθοδολογίας, η **προς-τα-πίσω διαδικασία επιλογής** (*backward stepwise selection*) αποτελεί μία αποτελεσματική εναλλακτική από αυτή της μεθόδου επιλογής του καλύτερου υποσυνόλου. Παρ'όλα αυτά, σε αντίθεση με την προς-τα-εμπρός διαδικασία, η εν λόγω τεχνική ξεκινάει με το ολοκληρωμένο μοντέλο των ελάχιστων τετραγώνων, το οποίο περιλαμβάνει όλες τις $p$ συμμεταβλητές και έπειτα απομακρύνει ανά μία, τις λιγότερο χρήσιμες. Η αλγοριθμική διαδικασία παρουσιάζεται παρακάτω:

**Αλγόριθμος 3** Διαδικασία της προς-τα-πίσω επιλογής (*Backward stepwise selection*)

1: Έστω $M_p$ να συμβολίζει το ολοκληρωμένο μοντέλο, το οποίο θα περιλαμβάνει όλες τις $p$ συμμεταβλητές.

2: Για $k = p, p-1, ..., 1$ :

   i Θεώρησε όλα τα $k$ μοντέλα, τα οποία περιλαμβάνουν όλες, εκτός μίας, εκ των συμμεταβλητών, στο μοντέλο $M_k$, για $k-1$ συμμεταβλητές συνολικά.

   ii Διάλεξε το καλύτερο μοντέλο μεταξύ αυτών των $k$ μοντέλων, και ονόμασέ το $M_{k-1}$. Στη συγκεκριμένη περίπτωση, το καλύτερο μοντέλο ορίζεται ως αυτό με το χαμηλότερο RSS, ή τον μεγαλύτερο συντελεστή $R^2$.

3: Διάλεξε το μοναδικό μοντέλο ανάμεσα από τα $M_0, ..., M_p$ χρησιμοποιώντας τους κατάλληλους δείκτες: διασταυρωμένο σφάλμα πρόβλεψης (*cross-validated prediction error*),$C_p(AIC), BIC$, προσαρμοσμένο συντελεστή $R^2_{adj}$.

Όπως και στην προς-τα-εμπρός διαδικασία, η προς-τα-πίσω προσέγγιση αναζητεί το καταλληλότερο μοντέλο ανάμεσα σε μόλις $1+p(p+1)/2$ μοντέλα και μπορεί επίσης να εφαρμοστεί σε περιπτώσεις που το $p$ παίρνει μεγάλες τιμές. Επιπλέον, όπως και πριν, η εν λόγω διαδικασία δεν εγγυάται ότι θα αποφέρει το καλύτερο μοντέλο που θα περιέχει ένα υποσύνολο των $p$ συμμεταβλητών.

Τέλος, για την εφαρμογή της προς-τα-πίσω τεχνικής επιλογής, απαιτείται ο αριθμός των δειγμάτων $n$ να είναι μεγαλύτερος από τον αριθμό των συμμεταβλητών $p$, ώστε να μπορεί να προσαρμοστεί το μοντέλο. Αντίθετα, η προς-τα-εμπός διαδικασία μπορεί να χρησιμοποιηθεί και όταν $n < p$, όπου και αποτελεί τη μοναδική βιώσιμη μέθοδο υποσυνόλων όταν η τιμή του $p$ είναι πολύ μεγάλη.

## 3.4 Παλινδρόμηση κορυφογραμμής (*Ridge regression*)

Όταν, λοιπόν, εμφανίζεται το φαινόμενο της πολυσυγγραμμικότητας στα δεδομένα, οι διακυμάνσεις των εκτιμήσεων τείνουν να είναι μεγάλες και μπορεί να διαφέρουν κατα πολύ από την πραγματική τους τιμή. Έτσι, η μέθοδος των ελαχίστων τετραγώνων αποτυγχάνει να εκτιμήσει σε ικανοποιητικό βαθμό τους συντελεστές παλινδρόμησης των μεταβλητών. Σαν εναλλακτική λύση, εφαρμόζεται μία τεχνική πάνω στο μοντέλο με τους $p$ εκτιμητές, όπου περιορίζει και τακτοποιεί τους εκτιμημένους συντελεστές, ή ισοδύναμα, *συρρικνώνει* τις εκτιμήσεις προς το μηδέν. Μπορεί να μην είναι άμεσα

ορατό γιατί ένας τέτοιου είδους περιορισμός μπορεί να βελτιστοποιήσει την προσαρμογή του μοντέλου, αλλά αποδεικνύεται ότι συρρικνώνοντας τους εκτιμημένους συντελεστές είναι δυνατό να μειωθεί σημαντικά η διασπορά.

### 3.4.1 Το μοντέλο

Η **Παλινδρόμηση κορυφογραμμής** (*Ridge regression*), ή $L_2$, λειτουργεί με παρόμοιο τρόπο με αυτό της μεθόδου των ελάχιστων τετραγώνων. Η διαφορά έγκειται στην ποσότητα που στοχεύουν να ελαχιστοποιήσουν οι δύο αυτές μέθοδοι. Η διαδικασία προσαρμογής των συντελεστών $\hat{\boldsymbol{\beta}}$, με τη μέθοδο των ελάχιστων τετραγώνων πραγματοποιείται μέσω της ελαχιστοποίησης του RSS (*residual sum of squares*):

$$RSS = \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

Με την παλινδρόμηση κορυφογραμμής, οι συντελεστές του μοντέλου εκτιμώνται μέσω της ελαχιστοποίησης μία ελαφρώς διαφορετικής συνάρτησης. Πιο συγκεκριμένα, οι εκτιμήσεις των συντελεστών της παλινδρόμησης κορυφογραμμής $\hat{\boldsymbol{\beta}}^R$ υπολογίζονται αν, ελαχιστοποιηθεί η παρακάτω συνάρτηση:

$$\sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3.1}$$

ή ισοδύναμα, η συνάρτηση:

$$\sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2, \text{υπό τον περιορισμό} \sum_{j=1}^{p} \beta_j^2 \leq t, \tag{3.2}$$

όπου $\lambda$ ή $t \geq 0$ είναι μία σταθερά και ονομάζεται **παράμετρος μεροληψίας ή παράμετρος συντονισμού** (*biasing parameter or tuning parameter*), αντίστοιχα. Ουσιαστικά, οι παράμετροι αυτοί ελέγχουν το βαθμό της συρρίκνωσης που θα υποστούν οι εκτιμήσεις. Όπως και στη μέθοδο των ελάχιστων τετραγώνων, η παλινδρόμηση κορυφογραμμής αναζητά εκτιμήσεις που προσαρμόζουν ικανοποιητικά τα δεδομένα, μειώνοντας το RSS. Ο δεύτερος όρος της (3.1), $\lambda \sum_{j=1}^{p} \beta_j^2$, καλείται **ποινή συρρίκνωσης** (*shrinkage penalty*) και παίρνει χαμηλές τιμές όταν οι συντελεστές $\beta_1, \ldots, \beta_p$ είναι κοντά στο μηδέν. Έτσι, αυτό έχει σαν αποτέλεσμα να συρρικνώνει μόνο τις εκτιμήσεις των $\beta_j$ πλησίον του μηδέν.

Η παράμετρος συντονισμού $\lambda$ χρησιμεύει για τον έλεγχο της σχετικής επίδρασης των δύο όρων της (3.1) στις εκτιμήσεις των συντελεστών της παλινδρόμησης. Όταν $\lambda = 0$, ο αντίστοιχος όρος της ποινής δεν έχει καμία επιρροή στο μοντέλο και συνεπώς η παλινδρόμηση κορυφογραμμής θα παράγει τις εκτιμήσεις των ελάχιστων τετραγώνων. Όμως, όσο $\lambda \to \infty$, η επιρροή της ποινής συρρίκνωσης αυξάνεται και οι εκτιμήσεις της παλινδρόμησης κορυφογραμμής προσεγγίζουν το μηδέν. Σε αντίθεση με τα ελάχιστα τετράγωνα, τα οποία παράγουν μόνο ένα σύνολο από εκτιμήσεις συντελεστών, η παλινδρόμηση κορυφογραμμής παράγει ένα διαφορετικό σύνολο εκτιμημένων συντελεστών, $\hat{\beta}_\lambda^R$, για κάθε τιμή του $\lambda$. Η επιλογή της καταλληλότερης τιμής για την παράμετρο συντονισμού $\lambda$ είναι πολύ κρίσιμη για την τελική διεξαγωγή των αποτελεσμάτων.

Στη σχέση (3.1) η ποινή συρρίκνωσης εφαρμόζεται μόνο στους συντελεστές $\beta_1, \beta_2, \ldots, \beta_p$, αλλά όχι στο σταθερό όρο $\beta_0$. Σκοπός της μεθόδου είναι να συρρικνωθεί η εκτιμώμενη συσχέτιση της κάθε συμμεταβλητής με τη μεταβλητή απόκρισης $y$. Αντιθέτως, δεν είναι επιθυμητό να συρρικνωθεί ο σταθερός όρος, καθώς αποτελεί απλά ένα μέτρο της μέσης τιμής της απόκρισης, όταν όλες οι επεξηγηματικές μεταβλητές είναι μηδέν. Με άλλα λόγια, όταν περιλαμβάνουμε τον σταθερό όρο $\beta_0$ στην παλινδρόμηση, τον αφήνουμε χωρίς ποινή. Έτσι, υπό την υπόθεση ότι οι στήλες του πίνακα $\boldsymbol{X}$, έχουν κεντραριστεί ώστε να έχουν μέση τιμή μηδέν, η εκτίμηση του σταθερού όρου θα πάρει τη μορφή, $\hat{\beta}_0 = \bar{y} = \sum_{i=1}^{n} y_i/n$.

### 3.4.2 Εκτίμηση των συντελεστών του μοντέλου

Με βάση τα παραπάνω, ελαχιστοποιώντας τη σχέση (3.1), υπολογίζεται η **εκτιμήτρια κορυφογραμμής** (*ridge estimator*), η οποία δίνεται από τον τύπο:

$$\hat{\boldsymbol{\beta}}_{\lambda}^{ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{3.3}$$

και συνήθως παρέχει καλύτερα υπολογιστικά σφάλματα πρόβλεψης από τις ε.ε.τ.. Η εκτιμήτρια $\hat{\boldsymbol{\beta}}_{\lambda}^{ridge}$, είναι μεροληπτική εκτιμήτρια, δηλαδή $E(\hat{\boldsymbol{\beta}}^{ridge}) = E(\boldsymbol{z}_p\boldsymbol{\beta}) = \boldsymbol{z}_p\boldsymbol{\beta}$, για κάθε $\lambda > 0$. Αποδεικνύεται εύκολα, ότι η εκτιμήτρια κορυφογραμμής είναι ένας γραμμικός συνδυασμός της εκτιμήτριας των ελαχίστων τετραγώνων, αφού:

$$\hat{\boldsymbol{\beta}}_{\lambda}^{ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{z}_p\boldsymbol{\beta}$$

με πίνακα συνδυασποράς:

$$Var(\hat{\boldsymbol{\beta}}_{\lambda}^{ridge}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}$$

και μέσο τετραγωνικό σφάλμα:

$$\begin{aligned} MSE(\hat{\boldsymbol{\beta}}_{\lambda}^{ridge}) &= Var(\hat{\boldsymbol{\beta}}_{\lambda}^{ridge}) + [bias(\hat{\boldsymbol{\beta}}_{\lambda}^{ridge})]^2 \\ &= \sigma^2 Tr[(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} \\ &\quad + \lambda \boldsymbol{I}_p)^{-1}] + \lambda^2\boldsymbol{\beta}^T(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\beta} \\ &= \sigma^2 \sum_{j=1}^{p} \frac{\mu_j}{(\mu_j + \lambda)^2} + \lambda^2\boldsymbol{\beta}^T(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-2}\boldsymbol{\beta}, \end{aligned} \tag{3.4}$$

όπου $\mu_1, \mu_2, \ldots, \mu_p$ συμβολίζονται οι ιδιοτιμές του πίνακα $\boldsymbol{X}^T\boldsymbol{X}$. Μέσα από τη σχέση αυτή, είναι εύκολο να παρατηρηθεί ότι, οποιαδήποτε αύξηση του όρου $\lambda$ προκαλεί αύξηση του δεύτερου όρου, δηλαδή της μεροληψίας, ενώ παράλληλα μειώνεται ο πρώτος όρος που αφορά τη διασπορά.

Σκοπός της μεθόδου κορυφογραμμής είναι να επιλεχθεί ένα $\lambda$ έτσι, ώστε η μείωση του όρου της διασποράς, στη σχέση (3.4), να είναι μεγαλύτερη από την αύξηση του όρου που εκφράζει τη μεροληψία. Οι Hoerl και Kennard απέδειξαν ότι αυτό μπορεί να επιτευχθεί, εάν για μία μη-μηδενική τιμή του $\lambda$, ικανοποιείται ο περιορισμός:

$$MSE(\hat{\boldsymbol{\beta}}_{\lambda}^{ridge}) < Var(\hat{\boldsymbol{\beta}}).$$

Οι τυπικές ε.ε.τ. των συντελεστών αποτελούν μία *εξισωτική κλίμακα*: πολλαπλάσιάζοντας το $X_j$ με έναν σταθερό όρο $c$, απλά οδηγεί σε μία κλιμάκωση των ε.ε.τ. από έναν παράγοντα $1/c$. Έτσι, ανεξαρτήτως, το πώς τροποποιείται η $j$-οστή συμμεταβλητή, ο παράγοντας $X_j\hat{\beta}_j$ θα παραμείνει ίδιος. Αντιθέτως, ο εκτιμητής κορυφογραμμής μπορεί να αλλάξει ουσιαστικά εάν πολλαπλασιαστεί μία συμμεταβλητή με κάποια σταθερά. Με άλλα λόγια, η ποσότητα $X_j\hat{\beta}_{j,\lambda}^{ridge}$ δεν εξαρτάται μόνο από την τιμή που παίρνει η παράμετρος συντονισμού $\lambda$, αλλά επίσης και από την κλιμάκωση της $j$-οστής ανεξάρτητης μεταβλητής. Στην πραγματικότητα, η τιμή του $X_j\hat{\beta}_{j,\lambda}^{ridge}$, μπορεί να εξαρτάται από την κλιμάκωση και άλλων συμμεταβλητών. Επομένως, είναι θεμητό, να εφαρμόζεται η παλινδρόμηση κορυφογραμμής, αφού οι ανεξάρτητες μεταβλητές έχουν πρώτα κανονικοποιηθεί έτσι, ώστε να βρίσκονται όλες στην ίδια κλίμακα. Συνεπώς, όλες οι κανονικοποιημένες συμμεταβλητές, θα έχουν τυπική απόκλιση ίση με τη μονάδα. Αυτό θα έχει σαν αποτέλεσμα, η τελική προσαρμογή του μοντέλου να μην εξαρτάται από την εκάστοτε κλίμακα.

### 3.4.3 Γιατί παλινδρόμηση κορυφογραμμής;

Το κύριο πλεονέκτημα της παλινδρόμησης κορυφογραμμής έναντι της μεθόδου ελαχίστων τετραγώνων εδράζεται στην ανταλλαγή σχέσεων μεταξύ μεροληψίας-διασποράς. Όσο το $\lambda$ αυξάνεται, η προσαρμογή του μοντέλου της παλινδρόμησης κορυφογραμμής μειώνεται, οδηγώντας σε μειωμένη διασπορά μεν, αλλά αυξημένη μεροληψία. Στις εκτιμήσεις των συντελεστών της μεθόδου των ελαχίστων τετραγώνων, παρατηρείται υψηλή διασπορά, αλλά δεν υπάρχει καθόλου μεροληψία. Το ίδιο συμβαίνει και στις εκτιμήσεις κορυφογραμμής για $\lambda = 0$, μιάς και όταν συμβαίνει αυτό, οι δύο τεχνικές συμπίπτουν. Παρ'όλα αυτά, όσο το $\lambda$ αυξάνεται, η συρρίκνωση των εκτιμημένων συντελεστών κορυφογραμμής οδηγούν σε μία ουσιαστική μείωση της διασποράς των προβλέψεων, με κόστος μία ελαφρά αύξηση της μεροληψίας.

Γενικεύοντας τα παραπάνω, σε περιπτώσεις όπου η σχέση μεταξύ της μεταβλητής απόκρισης και των συμμεταβλητών είναι σχετικά γραμμική, οι ε.ε.τ. θα έχουν χαμηλή μεροληψία, αλλά μεγάλη διασπορά. Αυτό σημαίνει ότι μία μικρή αλλαγή στα δεδομένα, μπορεί να προκαλέσει τεράστια αλλαγή στις εκτιμήσεις των συντελεστών. Συγκεκριμένα, όταν ο αριθμός των μεταβλητών $p$ είναι τόσο μεγάλος, όσο και ο αριθμός των παρατηρήσεων $n$, οι ε.ε.τ. θα είναι εξαιρετικά μεταβλητές. Επίσης, εάν $p > n$, τότε οι ε.ε.τ. δε θα έχουν καν μοναδική λύση, ενώ η παλινδρόμηση κορυφογραμμής μπορεί ακόμα να αποδόσει καλά, ανταλλάσσοντας μία μικρή αύξηση της μεροληψίας, για μία τεράστια μείωση της διασποράς. Ως εκ τούτου, η παλινδρόμηση κορυφογραμμής λειτουργεί καλύτερα σε περιπτώσεις όπου οι εκτιμήσεις των ελαχίστων τετραγώνων έχουν υψηλή διασπορά.

Η παλινδρόμηση κορυφογραμμής, επίσης, διαθέτει ουσιώδη υπολογιστικά πλεονεκτήματα έναντι της best subset selection, η οποία απαιτεί την αναζήτηση του καλύτερου μοντέλου μέσα από $2^p$ υποψήφια μοντέλα. Όπως προαναφέρθηκε, ακόμα και για μέτριες τιμές του $p$, τέτοιου είδους αναζητήσεις μπορεί να καθίστανται υπολογίστικά ακατόρθωτες. Αντιθέτως, για οποιαδήποτε σταθερή τιμή του $\lambda$, η παλινδρόμηση κορυφογραμμής προσαρμόζει μόνο ένα μοντέλο, και ολόκληρη η διαδικασία προσαρμογής

μπορεί να πραγματοποιηθεί αρκετά γρήγορα. Πραγματικά, είναι εύκολο να δειχθεί ότι, ο χρόνος υπολογισμών για τη λύση της εξίσωσης (3.1), ταυτόχρονα για όλες τις τιμές του $\lambda$, είναι σχεδόν πανομοιότυπος με αυτόν που απαιτείται για την προσαρμογή του μοντέλου χρησιμοποιώντας ελάχιστα τετράγωνα.

## 3.5 Η μέθοδος LASSO

Η παλινδρόμηση κορυφογραμμής έχει ένα φανερό μειονέκτημα. Σε αντίθεση με τις προηγούμενες τεχνικές που αναφέρθηκαν, οι οποίες θα επιλέξουν μοντέλα που περιλαμβάνουν μόνο ένα υποσύνολο των μεταβλητών, η παλινδρόμηση κορυφογραμμής θα συμπεριλάβει στο τελικό μοντέλο, όλες τις $p$ συμμεταβλητές. Η ποινή $\lambda \sum \beta_j^2$, στη σχέση (3.1), θα συρρικνώσει όλους τους συντελεστές προς το μηδέν, αλλά δε θα θέσει κάνεναν από αυτούς ακριβώς ίσο με μηδέν (εκτός εάν $\lambda = \infty$). Το γεγονός αυτό, ίσως δεν αποτελεί πρόβλημα όσον αφορά την ακρίβεια της πρόβλεψης, αλλά σίγουρα αποτελεί μία πρόκληση στην ερμηνεία του μοντέλου, ειδικά σε περιπτώσεις όπου ο αριθμός των μεταβλητών $p$ είναι αρκετά μεγάλος. Συνεπώς, αναπτύχθηκε η ανάγκη να δημιουργηθεί ένα μοντέλο που θα περιλαμβάνει μόνο τις περισσότερο στατιστικά σημαντικές συμμεταβλητές. Διότι, όσο και να αυξηθεί η παράμετρος συντονισμού $\lambda$, η μέθοδος κορυφογραμμής, θα μειώσει τα μεγέθη των συντελεστών αλλά, στο τέλος, δε θα αποκλείσει καμία από τις συμμεταβλητές.

Το 1996, λοιπόν, ο Tibshirani πρότεινε μία νέα τεχνική που περιορίζει το πρόβλημα της πολυσυγγραμμικότητας. Η μέθοδος **LASSO** (*Least Absolute Shrinkage and Selection Operator*), ή $L_1$. αποτελεί μία από τις καλύτερες τεχνικές επιλογής του καταλληλότερου μοντέλου, σήμερα. Ο σκοπός αυτής της μεθόδου είναι να κρατήσει τα καλά χαρακτηριστικά της επιλογής υποσυνόλου και της παλινδρόμησης κορυφογραμμής. Συγκεκριμένα, διακρίνεται για την ιδιότητά της να κάνει επιλογή μεταβλητών και συρρίκνωση του μοντέλου, ταυτόχρονα. Η τεχνική αυτή, την καθιστά πολύ χρήσιμη για μεγάλων διαστάσεων δεδομένα, ενώ παράλληλα προσφέρει εύκολα ερμηνεύσιμα αποτελέσματα.

Αξίζει να σημειωθεί ότι οι μέθοδοι $L_1$ και $L_2$ (lasso και ridge, αντίστοιχα) αποτελούν ειδικές περιπτώσεις της γενικότερης παλινδρόμησης **Bridge**, η οποία προτάθηκε από τους Frank και Friedman (1993). Οι δύο αυτές διαδικασίες κατηγοριοποιούνται στις μεθόδους με ποινή, που σημαίνει ότι τοποθετούν και οι δύο μία ποινή στον συντελεστή $\beta$. Έτσι, ο περιορισμός στη γενική του μορφή είναι:

$$\sum_{j=1}^{p} |\beta_j|^q \leq t$$

όπου $q \geq 0$. Η παλινδρόμηση κορυφογραμμής και η μέθοδος lasso ανταποκρίνονται στις τιμές $q = 2$, $q = 1$ αντίστοιχα.

### 3.5.1 Το μοντέλο

Έστω, $\{\boldsymbol{x}^i, y_i\}, i = 1, 2, ..., n$, το σύνολο των δεδομένων, όπου με $\boldsymbol{x}^i = (x_{i1}, ..., x_{ip})$ συμβολίζονται οι επεξηγηματικές μεταβλητές και $y_i$ οι αντίστοιχες αποκρίσεις. Ως συνήθως, η παλινδρόμηση γίνεται υπό την υπόθεση, ότι οι παρατηρήσεις είναι ανεξάρ-

τητες. Επιπλέον, για τον ίδιο λόγο με τη μέθοδο κορυφογραμμής, οι μεταβλητές $x_{ij}$ θεωρούνται κανονικοποιημένες έτσι, ώστε: $\mu = \sum_i x_{ij}/n = 0$, $\sigma^2 = \sum_i x_{ij}^2/n = 1$.

Με $\hat{\boldsymbol{\beta}}_\lambda^L = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$, οι lasso εκτιμήτριες $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_\lambda^L)$, ελαχιστοποιούν την ποσότητα:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j| \tag{3.5}$$

ή ισοδύναμα τη συνάρτηση:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2, \text{υπό τον περιορισμό} \sum_{j=1}^{p}|\beta_j| \leq t, \tag{3.6}$$

όπου $\lambda$ ή $t \geq 0$ αποτελεί την παράμετρο συντονισμού. Σημειώνεται ότι για όλες τις τιμές που μπορεί να λάβει η παράμετρος $t$, η λύση για το $\beta_0$ είναι: $\hat{\beta}_0 = \bar{y}$. Έτσι, χωρίς βλάβη της γενικότητας, είναι δυνατό να θεωρηθεί ότι $\bar{y} = 0$ και με αυτό τον τρόπο να παραλειφθεί ο παράγοντας $\beta_0$ από τις παραπάνω σχέσεις.

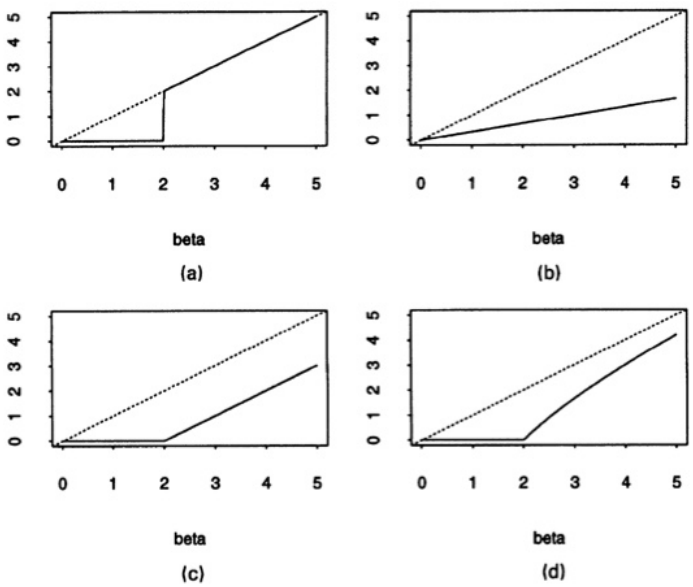
Όπως και στην παλινδρόμηση κορυφογραμμής, οι παράμετροι $\lambda$ ή $t$ ελέγχουν το ποσό της συρρίκνωσης. Έστω $\hat{\beta}_j^o$, οι ε.ε.τ. και έστω $t_0 = \sum|\hat{\beta}_j^o|$. Οι τιμές του $t$, οι οποίες είναι μικρότερες από αυτές του $t_0$, $t < t_0$, θα προκαλέσουν συρρίκνωση των λύσεων προς το μηδέν και κάποιοι εκ των συντελεστών μπορεί να είναι ακριβώς ίσοι με το μηδέν. Εάν, για παράδειγμα $t = t_0/2$, το αποτέλεσμα θα ήταν περίπου παρόμοιο με αυτό της μεθόδου βέλτιστου υποσυνόλου για μέγεθος μεταβλητών $p/2$. Επιπλέον, σημειώνεται ότι ο πίνακας σχεδιασμού δεν πρέπει να είναι κατ'ανάγκη πλήρους τάξης.

Η έμπνευση για τη lasso προήλθε από μία ενδιαφέρουσα πρόταση του Breiman (1993), η οποία καλείται non-negative garotte (*μη-αρνητικός στραγγαλισμός*). Σύμφωνα με αυτή τη μέθοδο, οι συντελεστές $\hat{\beta}_j$ έχουν τη μορφή $\hat{\beta}_j = c_j\hat{\beta}_j^o$, όπου οι ποσότητες $c_j$ υπολογίζονται έτσι, ώστε να ελαχιστοποιείται η συνάρτηση:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}c_j\hat{\beta}_j^o x_{ij}\right)^2, \text{υπό τον περιορισμό} \sum_{j=1}^{p}c_j \leq t, c_j \geq 0. \tag{3.7}$$

Η μέθοδος garotte ξεκινάει με τις ε.ε.τ. και τις συρρικνώνει με κάποιους μη-αρνητικούς παράγοντες των οποίων το άθροισμα είναι φραγμένο. Μετά από εξαντλητικές έρευνες, ο Breiman απέδειξε ότι η μέθοδος του έχει μικρότερο σφάλμα πρόβλεψης από αυτό των μεθόδων της επιλογής υποσυνόλου και της κορυφογραμμής. Υπάρχουν μόνο λίγες περιπτώσεις που δε συμβαίνει αυτό· όταν το πραγματικό μοντέλο έχει πολλούς μικρούς μη-μηδενικούς συντελεστές. Το μειονέκτημα της μεθόδου garotte είναι ότι η λύση της εξαρτάται τόσο από το πρόσημο, όσο και από το μέγεθος των ε.ε.τ.. Έτσι, σε περιπτώσεις υψηλών συσχετίσεων μεταξύ των επεξηγηματικών μεταβλητών, όπου οι ε.ε.τ. δε συμπεριφέρονται ικανοποιητικά, μπορεί και η μέθοδος garotte να αδυνατεί να ανταποκριθεί σωστά. Αντιθέτως, η lasso αποφεύγει τη χρήση των ε.ε.τ. και μαζί και τα μειονεκτήματά τους.

Για μια πιο πρακτική κατανόηση των μεθόδων συρρίκνωσης, ο Tibshirani παρουσιάζει τη μορφή που αποκτούν, στην περίπτωση που ο πίνακας σχεδιασμού είναι ορθο-




Σχήμα 3.1: Μορφή συρρίκνωσης συντελεστών στην περίπτωση ορθοκανονικού πίνακα σχεδιασμού: (a) Παλινδρόμηση Υποσυνόλου, (b) Παλινδρόμηση Κορυφογραμμής, (c) Lasso, (d) Garotte.

κανόνικός. Δηλαδή $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$, όπου $\boldsymbol{I}$ είναι ο ταυτοτικός πίνακας. Έτσι, εάν $\boldsymbol{X}$ είναι ο $n \times p$ πίνακας σχεδιασμού και $x_{ij}$ είναι το $ij$-οστό στοιχείο του πίνακα, τότε:

i Στη μέθοδο επιλογής υποσυνόλου, εάν $k$ το πλήθος των επιλεγμένων μεταβλητών, τότε η διαδικασία αυτή επιλέγει τις μεταβλητές εκείνες, οι οποίες αντιστοιχούν στους $k$ μεγαλύτερους, κατα απόλυτη τιμή, συντελεστές και θέτει τους υπόλοιπους ίσους με μηδέν. Διαφορετικά, για κάποια τιμή του $\lambda$, αυτό είναι ισοδύναμο με:

$$\hat{\beta}_j^{BS} = \begin{cases} \hat{\beta}_j^0, & |\hat{\beta}_j^0| > \lambda \\ 0, & \text{διαφορετικά.} \end{cases}$$

ii Στην παλινδρόμηση κορυφογραμμής ($L_2$), οι εκτιμήτριες των συντελεστών $\beta_j$ υπολογίζονται από τη λύση του συστήματος (3.2). Οι λύσεις αυτές είναι:

$$\hat{\beta}_j^{ridge} = \frac{1}{1+\gamma}\hat{\beta}_j^0$$

όπου το $\gamma$ εξαρτάται από το $\lambda$ ή το $t$.

iii Οι εκτιμήτριες της μεθόδου garotte είναι:

$$\hat{\beta}_j^{gar} = \left(1 - \frac{\gamma}{(\hat{\beta}_j^0)^2}\right)^+ \hat{\beta}_j^0.$$

Με το σύμβολο "+" στη θέση του εκθέτη, εκφράζεται η επιλογή μη-αρνητικών λύσεων.

iv Τέλος, στη μέθοδο lasso ($L_1$), οι λύσεις των εξισώσεων (3.6), όταν ο πίνακας σχεδιασμού είναι ορθοκανονικός, δίνονται από τον παρακάτω τύπο:

$$\hat{\beta}_j^L = sign(\hat{\beta}_j^0)(|\hat{\beta}_j^0 - \gamma|)^+ \tag{3.8}$$

όπου το $\gamma$ καθορίζεται από τη συνθήκη $\sum_j |\hat{\beta}_j| = t$. Παρόμοια με τη μέθοδο garotte, το σύμβολο "+" εκφράζει την επιλογή των, κατα απόλυτη τιμή, μεγαλύτερων συντελεστών, ενώ παράλληλα θέτει τους υπόλοιπους ίσους με μηδέν.

Στο σχήμα (3.1) φαίνεται η μορφή αυτών των συναρτήσεων. Η παλινδρόμηση κορυφογραμμής αλλάζει την κλίμακα των συντελεστών κατα ένα σταθερό παράγοντα. Αντίστοιχα, η μέθοδος lasso συρρικώνει τους συντελεστές κατα ένα σταθερό παράγοντα και άλλους τους αποκόπτει εξισώνοντάς τους με το μηδέν. Η διαφορά που έγκειται μεταξύ των μεθόδων garotte και lasso είναι ότι οι συντελεστές της πρώτης, υφίστανται μικρότερη συρρίκνωση όταν είναι μεγαλύτεροι. Μέσα από διάφορες προσομοιώσεις, αποδείχθηκε ότι οι διαφορές μεταξύ των μεθόδων lasso και garotte είναι μεγαλύτερες όταν ο πίνακας σχεδιασμού δεν είναι ορθογώνιος. Η διακεκομμένη γραμμή που παρουσιάζεται στα γραφήματα του σχήματος (3.1), αντιστοιχεί σε γωνία $45°$, δηλαδή στις ε.ε.τ..

Έστω, τώρα, ότι το μοντέλο αποτελείται από δύο επεξηγηματικές μεταβλητές και, έστω ότι οι δύο αντίστοιχες ε.ε.τ., $\hat{\beta}_j^0, j = 1, 2$, είναι θετικές. Εύκολα αποδεικνύεται, μέσω της (3.8), ότι οι εκτιμήτριες lasso είναι:

$$\hat{\beta}_j = (\hat{\beta}_j^o - \gamma)$$

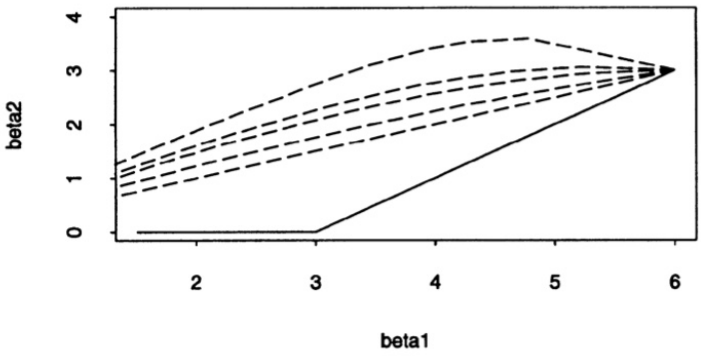
όπου $\gamma$ είναι επιλεγμένο έτσι, ώστε $\hat{\beta}_1 + \hat{\beta}_2 = t$. Ο παραπάνω τύπος ισχύει για $t \leq \hat{\beta}_1^o + \hat{\beta}_2^o$ και είναι έγκυρος ακόμα και αν οι συμμεταβλητές χαρακτηρίζονται από κάποια συσχέτιση. Λύνοντας ως προς $\gamma$:

$$\hat{\beta}_1 = \left(\frac{t}{2} + \frac{\hat{\beta}_1^o - \hat{\beta}_2^o}{2}\right)^+,$$

$$\hat{\beta}_2 = \left(\frac{t}{2} - \frac{\hat{\beta}_1^o - \hat{\beta}_2^o}{2}\right)^+.$$

Σε αντίθεση με τις εκτιμήσεις lasso, οι εκτιμήσεις κορυφογραμμής εξαρτώνται από τη συσχέτιση των επεξηγηματικών μεταβλητών. Στο σχήμα (3.2), οι γραμμές παρουσιάζουν τα ζεύγη $(\beta_1, \beta_2)$ όσο μεταβάλλεται το όριο των παραμέτρων lasso και κορυφογραμμής. Ξεκινώντας από την κάτω διακεκομμένη γραμμή και κινούμενοι προς τα πάνω, οι συσχετίσεις $\rho$ είναι 0, 0.23, 0.45, 0.68 και 0.90. Για όλες τις τιμές συσχέτισης $\rho$, οι εκτιμήσεις lasso ακολουθούν την ίδια πορεία (συνεχόμενη γραμμή στο σχήμα (3.2)), ενώ οι εκτιμήσεις κορυφογραμμής εξαρτώνται από την εκάστοτε τιμή της $\rho$ (διακεκομμένη γραμμή στο σχήμα (3.2)).

Συμπερασματικά, μετά από προσομοιώσεις που πραγματοποιήθηκαν από τον Tibshirani (1996), εξετάστηκε συγκριτικά η απόδοση των τεσσάρων μεθόδων, κάτω από τρία διαφορετικά σενάρια:




Σχήμα 3.2: Παράδειγμα εκτιμήσεων για μοντέλο παλινδρόμησης lasso και κορυφογραμμής με δύο επεξηγηματικές μεταβλητές

i *Μικρός αριθμός μεταβλητών, με σημαντική επίδραση στην εξαρτημένη μεταβλητή*: Στην περίπτωση αυτή, η επιλογή υποσυνόλου παρουσιάζει την καλύτερη συμπεριφορά, η lasso δεν ανταποκρίνεται τόσο καλά και η παλινδρόμηση κορυφογραμμής έδωσε τα χειρότερα αποτελέσματα.

ii *Μέτριος αριθμός μεταβλητών, με μέτρια επίδραση στην εξαρτημένη μεταβλητή*: Στην περίπτωση αυτή, η lasso παρουσιάζει τα καλύτερα αποτελέσματα και ακολουθούν η παλινδρόμηση κορυφογραμμής και η επιλογή υποσυνόλου.

iii *Μεγάλος αριθμός μεταβλητών, με μικρή επίδραση στην εξαρτημένη μεταβλητή*: Στην περίπτωση αυτή, καλύτερα αποτελέσματα, με διαφορά, είχε η παλινδρόμηση κορυφογραμμής ακολουθούμενη από τη lasso και την επιλογή υποσυνόλου.

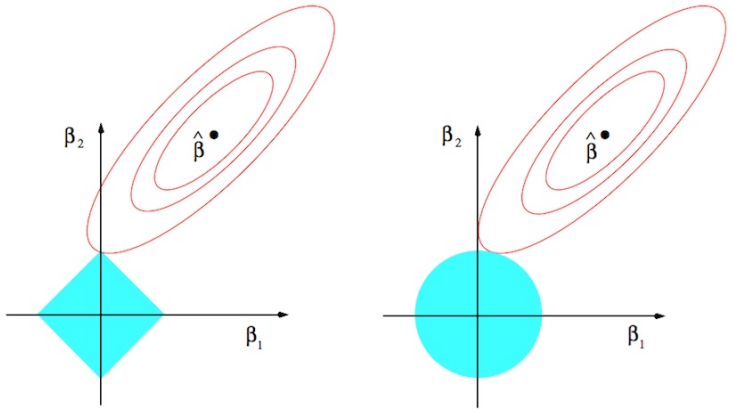
Η μέθοδος garotte ανταποκρίθηκε λίγο καλύτερα από τη lasso, στο πρώτο σενάριο και, λίγο χειρότερα στα επόμενα δύο. Αυτά τα αποτελέσματα αναφέρονται στην ικανότητα πρόβλεψης των μεθόδων και η μέτρησή τους πραγματοποιήθηκε μέσω της μεθόδου *Cross-Validation*, όπου και θα παρουσιαστεί παρακάτω. Οι μέθοδοι lasso, garotte και επιλογής υποσυνόλου έχουν το πλεονέκτημα, έναντι της κορυφογραμμής, ότι προσφέρουν εύκολα ερμηνεύσιμα υπομοντέλα.

### 3.5.2 Η γεωμετρία της Lasso

Στο σχήμα (3.1), παρουσιάστηκε η συμπεριφορά της μεθόδου lasso στην περίπτωση ορθοκανονικού πίνακα σχεδιασμού. Το ερώτημα που τίθεται, τώρα, είναι τι θα συμβεί σε μη-ορθοκανονικό περιβάλλον και γιατί η lasso είναι ικανή να παράγει μηδενικούς συντελεστές και η μέθοδος κορυφογραμμής όχι. Η διαφορά των δύο μεθόδων τοποθετείται στους περιορισμούς που επιβάλουν: $\sum \beta_j^2 \leq t$ για την παλινδρόμηση κορυφογραμμής και $\sum |\beta_j| \leq t$. Έστω το μοντέλο με τις δύο επεξηγηματικές μεταβλητές που παρουσιάστηκε πριν. Το σχήμα (3.3) παρουσιάζει την εικόνα εκτίμησης για τις δύο αυτές μεθόδους με βάση στους περιοριμούς τους.

Το κριτήριο $RSS = \sum_{i=1}^{n}(y_i - \sum_j \beta_j x_{ij})^2$ ισούται με την τετραγωνική συνάρτηση:

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^o)^T \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^o)$$




Σχήμα 3.3: Εικόνα εκτιμήσεων για: (a) Lasso και (b) Κορυφογραμμή.

Οι ελλείψεις στο σχήμα (3.3(a)) αντιστοιχούν στα περιγράμματα του RSS· η εσωτερική έλλειψη έχει μικρότερο RSS. Οι ελλείψεις αυτές είναι κεντραρισμένες με βάση τις αντίστοιχες ε.ε.τ., όπου και το RSS ελαχιστοποιείται. Το περιστραφόμενο τετράγωνο απεικονίζει την περιοχή περιορισμού. Σκοπός είναι να ελαχιστοποιηθεί το μέγεθος της έλλειψης.

Όπως φαίνεται στο σχήμα, λοιπόν, οι διαφορά των δύο μεθόδων παρουσιάζεται στη γεωμετρία της περιοχής περιορισμού. Για $p = 2$, ο περιορισμός της μεθόδου lasso παίρνει τη μορφή:

$$|\beta_1| + |\beta_2| \leq t.$$

Έτσι, η λύση της lasso περιγράφεται από το πρώτο σημείο όπου οι ελλείψεις θα ακουμπήσουν στο τετράγωνο. Μερικές φορές, αυτό το σημείο συμβαίνει να είναι η κορυφή του τετραγώνου, η οποία αντιστοιχεί σε μηδενικό συντελεστή. Από την άλλη, η εκτίμηση κορυφογραμμής δίνεται τη στιγμή που η έλλειψη ακουμπήσει την κυκλική περιοχή περιορισμού, όπου δεν υπάρχουν γωνίες, γεγονός που καθιστά σχεδόν αδύνατο να προκύψει μηδενικός συντελεστής. Όσο το $p$ αυξάνεται, το πολυδιάστατο τετράγωνο της περιοριστική περιοχής παρουσιάζει περισσότερες γωνίες και έτσι, είναι πολύ πιθανό μερικοί από τους συντελεστές να είναι ίσοι με το μηδέν. Με αυτόν τον τρόπο, η lasso πραγματοποιεί συρρίκνωση των συντελεστών και αποτελεί μία εξαιρετικά αποτελεσματική μέθοδο επιλογής μεταβλητών.

### 3.5.3 Τυπικά σφάλματα

Εφόσον η εκτίμηση lasso είναι μία μη-γραμμική και μη-διαφορίσιμη συνάρτηση των τιμών απόκρισης, είναι δύσκολο να εξασφαλιστεί η ακριβής εκτίμηση του τυπικού σφάλματος. Μία προσέγγιση που προτείνεται από τον Tibshirani, είναι μέσω της μεθόδου bootstrap [1], όπου είναι δυνατό είτε να καθοριστεί η τιμή του $t$, ή να βελτιστοποιηθεί

[1] Η μέθοδος Bootstrap είναι μία διαδικασία ανάθεσης μέτρων ακρίβειας (μεροληψία, διασπορά, διαστήματα εμπιστοσύνης, σφάλματα πρόβλεψης κ.ά.) σε δειγματικές εκτιμήτριες. Αυτή η τεχνική επιτρέπει την εκτίμηση κατανομής του δείγματος, χρησιμοποιώντας πολύ απλές μεθόδους. Γενικά, η μέθοδος Bootstrap ανήκει στις μεθόδους επαναδειγματοληψίας.

για κάθε δείγμα bootstrap. Ο καθορισμός του $t$ μέσω αυτής της διαδικασίας, είναι ανάλογος με την επιλογή του καλύτερου υποσυνόλου και έπειτα, με τον υπολογισμό των τυπικών σφαλμάτων των ελαχίστων τετραγώνων για αυτό το υποσύνολο.

Μία προσεγγιστική, κλειστού τύπου, εκτίμηση μπορεί να προκύψει γράφοντας την ποινή $\sum |\beta_j|$, ως $\sum \beta_j^2/|\beta_j|$. Ως εκ τούτου, είναι δυνατό να προσεγγιστούν οι εκτιμήσεις lasso $\tilde{\boldsymbol{\beta}}$ μέσω μιας παλινδρόμησης κορυφογραμμής της μορφής:

$$\boldsymbol{\beta}^\star = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{W}^-)^{(-1)}\boldsymbol{X}^T\boldsymbol{y}$$

όπου $\boldsymbol{W}$ είναι ένας διαγώνιος πίνακας, με διαγώνια στοιχεία τα $|\tilde{\beta}_j|$, $\boldsymbol{W}^-$ είναι ο γενικευμένος αντίστροφος του πίνακα $\boldsymbol{W}$ και $\lambda$ είναι επιλεγμένο έτσι, ώστε $\sum_j |\beta_j|^\star = t$. Ο πίνακας συνδιασποράς τότε, θα προσεγγίζεται από τη σχέση:

$$(\boldsymbol{X}\boldsymbol{X} + \lambda \boldsymbol{W}^-)^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{X}^-)^{-1}\hat{\sigma}^2,$$

όπου $\hat{\sigma}^2$ είναι μία εκτίμηση του σφάλματος διασποράς. Το πρόβλημα με αυτή τη διαδικασία είναι ότι, όταν $\hat{\beta}_j = 0$, τότε η εκτίμηση της διασποράς υπολογίζεται και αυτή ως μηδενική.

### 3.5.4 Εκτίμηση του συντελεστή $t$

Υπάρχουν τρεις μέθοδοι για την εκτίμηση της παραμέτρου lasso $t$: η cross-validation, η γενικευμένη cross-validation και η αναλυτική αμερόληπτη εκτιμήτρια του ρίσκου. Έστω, λοιπόν, το μοντέλο:

$$Y = \eta(\boldsymbol{X}) + \varepsilon$$

όπου $E(\varepsilon) = 0$ και $Var(\varepsilon) = \sigma^2$. Το **μέσο τετραγωνικό σφάλμα** (*mean-squared error*) μίας εκτιμήτριας $\hat{\eta}(\boldsymbol{X})$ ορίζεται από τη σχέση:

$$MSE = E\{\hat{\eta}(\boldsymbol{X}) - \eta(\boldsymbol{X})\} \tag{3.9}$$

που εκφράζει την αναμενόμενη τιμή που λαμβάνεται από την από κοινού κατανομή των $\boldsymbol{X}$ και $Y$, δεδομένου ότι η $\hat{\eta}(\boldsymbol{X})$ είναι σταθερή. Ένα παρόμοιο μέτρο αποτελεί και το **σφάλμα πρόβλεψης** (*prediction error*) της $\hat{\eta}(\boldsymbol{X})$, το οποίο δίνεται από τη σχέση:

$$PE = E\{Y - \hat{\eta}(\boldsymbol{X})\}^2 = MSE + \sigma^2 \tag{3.10}$$

Η εκτίμηση του σφάλματος πρόβλεψης πραγματοποιείται μέσω της διαδικασίας 5-fold cross-validation (Efron και Tibshirani, 1993). Η μέθοδος lasso αναπροσαρμόζεται υπό τους όρους μίας κανονικοποιημένης, πλέον, παραμέτρου:

$$s = \frac{t}{\sum_j \hat{\beta}_j^o}$$

και το σφάλμα πρόβλεψης υπολογίζεται σε ένα κλειστό διάστημα τιμών της $s$, μεταξύ του 0 και του 1. Η τιμή του $\hat{s}$ που δίνει το χαμηλότερο σφάλμα πρόβλεψης, είναι και αυτή που επιλέγεται τελικά.

Μία δεύτερη μέθοδος εκτίμησης της παραμέτρου $t$ μπορεί να προκύψει μέσω μίας γραμμικής προσέγγισης της εκτίμησης lasso. Έτσι, εάν ο περιορισμός $\sum_j |\beta_j| \leq t$ γραφεί ως, $\sum_j \beta_j^2/|\beta_j| \leq t$, αυτό είναι ισοδύναμο με το να προστεθεί μία ποινή Lagrangian, $\lambda \sum_j \beta_j^2/|\beta_j|$ στο RSS, με το $\lambda$ να εξαρτάται από το $t$. Έτσι, η υπό περιορισμόν λύση κορυφογραμμής $\tilde{\beta}$, γράφεται ως:

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{W}^-)^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{3.11}$$

όπου $\boldsymbol{W} = diag(|\tilde{\beta}_j|)$ και $\boldsymbol{W}^-$ συμβολίζει έναν γενικευμένο αντίστροφο. Ως εκ τούτου, ο αριθμός των αποτελεσματικών παραμέτρων στην περιορισμένη προσαρμογή του $\tilde{\boldsymbol{\beta}}$ προσεγγίζεται από:

$$p(t) = tr\{\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{W}^-)^{-1}\boldsymbol{X}^T\}. \tag{3.12}$$

Υπό την υπόθεση ότι, με $RSS(t)$ συμβολίζεται το άθροισμα τετραγώνων των υπολοίπων για την, υπό περιορισμόν, προσαρμογή, από την παράμετρο $t$, κατασκευάζεται η **γενικευμένη μορφή της μεθόδου cross-validation** (*generalized cross-validation*) και είναι:

$$GCV = \frac{1}{n}\frac{RSS(t)}{\{1 - p(t)/n\}^2}. \tag{3.13}$$

Τέλος, η τρίτη μέθοδος είναι βασισμένη στην αμερόληπτη εκτιμήτρια ρίσκου του Stein. Έστω ότι $\boldsymbol{z}$ είναι ένα πολυμετάβλητο τυχαίο διάνυσμα με μέση τιμή $\boldsymbol{\mu}$ και διασπορά τον ταυτοτικό πίνακα $\boldsymbol{I}$. Εάν $\hat{\boldsymbol{\mu}}$ είναι η εκτίμηση της μέσης τιμής $\boldsymbol{\mu}$, τότε $\hat{\boldsymbol{\mu}} = \boldsymbol{z} + \boldsymbol{g}(\boldsymbol{z})$, όπου $\boldsymbol{g} : \mathbb{R}^p \to \mathbb{R}^p$ είναι μία σχεδόν διαφορίσιμη συνάρτηση. Έπειτα, ο Stein έδειξε ότι:

$$E_\mu\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = p + E_\mu\left(\|\boldsymbol{g}(\boldsymbol{z})\|^2 + 2\sum_{i=1}^{p} dg_i/dz_i\right). \tag{3.14}$$

Εάν, τώρα, εφαρμοστεί αυτό το αποτέλεσμα στην εκτιμήτρια lasso (σχ. 3.8), υπολογίζεται το εκτιμημένο τυπικό σφάλμα του συντελεστή $\hat{\beta}_j^o$, από τη σχέση:

$$\hat{\tau} = \hat{\sigma}/\sqrt{n},$$

όπου $\hat{\sigma}^2 = \sum_i (y_i - \hat{y}_i)^2/(n-p)$. Έτσι, λοιπόν, από τη σχέση (3.14) και από το γεγονός ότι η ποσότητα $\hat{\beta}_j^o/\hat{\tau}$ απεικονίζει προσεγγιστικά, ανεξάρτητες μεταβλητές της τυποποιημένης Κανονικής κατανομής, ο τύπος:

$$R\{\hat{\beta}(\gamma)\} \approx \hat{\tau}^2\left\{p - 2\#(j; |\hat{\beta}_j^o/\hat{\tau}| < \gamma) + \sum_{j=1}^{p} max(|\hat{\beta}_j^o/\hat{\tau}|, \gamma)^2\right\}$$

δίνει κατά προσέγγιση μία αμερόληπτη εκτιμήτρια ρίσκου ή διαφορετικά, το τετραγωνικό σφάλμα $E\{\hat{\beta}(\gamma) - \beta\}^2$, όπου $\hat{\beta}_j(\gamma) = sign(\hat{\beta}_j^o)(|\hat{\beta}_j^o/\hat{\tau}| - \gamma)^+$. Επιπλέον, μία εκτίμηση για τον παράγοντα $\gamma$ μπορεί να είναι η:

$$\hat{\gamma} = argmin_{\gamma \geq 0}[R\{\hat{\beta}(\gamma)\}],$$

όπου αποτελεί την τιμή ελαχιστοποιεί τη συνάρτηση $R\{\hat{\beta}(\gamma)\}$. Από όλα τα παραπάνω, λοιπόν, υπολογιζεται μία εκτιμήτρια για την παράμετρο lasso, που δίνεται από τη σχέση:

$$\hat{t} = \sum_j (|\hat{\beta}_j^o| - \hat{\gamma})^+. \tag{3.15}$$

Αν και η παραγωγή του εκτιμητή $\hat{t}$ υποθέτει ένα ορθοκανονικό σχήμα, μπορεί ακόμα να χρησιμοποιηθεί στο σύνηθες μη-ορθοκανονικό σχήμα. Αξίζει να σημειωθεί ότι, η μέθοδος του Stein έχει σημαντικό υπολογιστικό προτέρημα για την εκτίμηση του συντελεστή $t$, έναντι της μεθόδου cross-validation.

### 3.5.5 Lasso στα γενικευμένα γραμμικά μοντέλα

Η μέθοδος lasso είναι δυνατό να εφαρμοστεί και στην περίπτωση των γενικευμένων γραμμικών μοντέλων. Έστω ένα οποιοδήποτε μοντέλο, το οποίο έχει αναπροσαρμοστεί από ένα διάνυσμα παραμέτρων $\boldsymbol{\beta}$. Επίσης, έστω ότι η εκτίμηση των συντελεστών αυτών, έχει διεξαχθεί μέσω της μεγιστοποίησης κάποιας συνάρτησης $l(\boldsymbol{\beta})$. Αυτή η συνάρτηση μπορεί να είναι η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας ή κάποιο άλλο μέτρο προσαρμογής.

Για να εφαρμοστεί η lasso, πρέπει να μεγιστοποιηθεί η συνάρτηση $l(\boldsymbol{\beta})$, υπό τον περιορισμό $\sum |\beta_j| \leq t$. Η μεγιστοποιήση της θα ήταν δυνατό να πραγματοποιηθεί με κάποια (μη-τετραγωνική) προγραμματιστική διαδικασία. Αντ'αυτού, θεωρούνται μοντέλα τα οποία μπορούν να εκτιμήσουν τους συντελεστές $\boldsymbol{\beta}$, μέσω μίας τετραγωνικής προσέγγισης της $l(\beta)$, που οδηγεί σε μία επαναληπτικά επανεκτεινόμενη διαδικασία ελάχιστων τετραγώνων. Μία τέτοια επαναληπτική διαδικασία είναι ισοδύναμη του αλγορίθμου Newton-Raphson. Χρησιμοποιώντας αυτή την προσέγγιση, είναι δυνατό να επιλυθεί το, υπο περιορισμόν, πρόβλημα μέσω επαναληπτικής εφαρμογής του αλγορίθμου της lasso, στα πλαίσια της επανεκτεινόμενης διαδικασίας ελάχιστων τετραγώνων. Ο Tibshirani τονίζει ότι η σύγκλιση αυτής της διαδικασίας, δεν είναι γενικά εξασφαλισμένη, αλλά βάσει της μέχρι στιγμής εμπειρίας έχει ανταποκριθεί αρκετά ικανοποιητικά.

Στην περίπτωση της λογιστικής παλινδρόμησης, ως συνάρτηση $l(\beta)$ χρησιμοποιείται η συνάρτηση πιθανοφάνειας, όπως παρουσιάζεται στη σχέση (2.23). Έπειτα, για την εφαρμογή της μεθόδου lasso, πραγματοποιείται η διαδικασία που περιγράφεται παραπάνω, όπως σε όλες τις περιπτώσεις των γενικευμένων γραμμικών μοντέλων.

### 3.5.6 Cross-Validation (*cvl*)

Στις τεχνικές με ποινή, καθοριστικό ρόλο παίζει η μέθοδος **cross-validation** (*cvl*). Με βάση αυτή, είναι δυνατό να βρεθούν οι βέλτιστες τιμές για την παράμετρο $\lambda$ έτσι, ώστε να καθοριστεί το καλύτερο μοντέλο. Η μέθοδος αυτή εισήχθηκε από τους Verweij και Van Houwelingen (1993), και ουσιαστικά υπολογίζει την *cross-validated συνάρτηση πιθανοφάνειας* για σταθερές τιμές των $\lambda$.

Η πιο απλή μορφή της μεθόδου είναι η *leave-one-out cvl*. Στην τεχνική αυτή, εξαιρείται μία παρατήρηση από την αναλυτική διαδικασία, σε κάθε επανάληψη και γίνεται πρόβλεψη της τιμής της μεταβλητής απόκρισης για αυτήν την παρατήρηση. Η πρόβλεψη

αυτή γίνεται κάνοντας χρήση του μοντέλου που έχει προκείψει από τις εναπομείνουσες $n-1$ παρατηρήσεις. Η διαδικασία επαναλαμβάνεται $n$ φορές και έτσι προκύπτει μια μέση ακρίβεια. Το 1983, ο Efron, υποστήριξε ότι η εφαρμογή της μεθόδου σε ομάδες (*fold*) δίνει πιο ακριβή αποτελέσματα. Με τον όρο $k-fold$ εννοείται μία ομάδα $k$ παρατηρήσεων. Στη $cvl$ διαδικασία, ολόκληρη η ομάδα, είναι αυτή που εξαιρείται σε κάθε επανάληψη. Για παράδειγμα, έστω $k=10$. Έτσι, κάθε φορά που θα τρέχει η μέθοδος, θα εξαιρείται μία ομάδα 10 παρατηρήσεων. Συνεπώς, το αρχικό σύνολο δεδομένων θα χωρίζεται σε 10 υποσύνολα, με τυχαίο τρόπο και κάθε υποσύνολο θα περιέχει ίδιο αριθμό παρατηρήσεων. Έπειτα, σε κάθε επανάληψη 9 στις 10 ομάδες θα χρησιμοποιούνται για την ανάπτυξη του μοντέλου. Στη συνέχεια, το μοντέλο που προκύπτει θα αξιολογείται ως προς την ακρίβεια, πάνω στην ομάδα που έχει εξαιρεθεί. Αυτή η διαδικασία επαναλαμβάνεται τουλάχιστον 10 φορές, από τις οποίες προκύπτουν τουλάχιστον 10 δείκτες (π.χ. $R^2$). Ένα μειονέκτημα της μεθόδου είναι η επιλογή του αριθμού των παρατηρήσεων που εξαιρούνται κάθε φορά, καθώς και ο αριθμός των επαναλήψεων που απαιτείται προκειμένου να αποκτήσουμε εκτιμήσεις για την ακρίβεια του μοντέλου.

Έστω $n$-παρατηρήσεις και ένα μοντέλο παλινδρόμησης που περιγράφει τα δεδομένα. Επίσης, έστω $l(\boldsymbol{\beta})$ η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας, όπου $\boldsymbol{\beta}$ το διάνυσμα των συντελεστών παλινδρόμησης. Η συμβολή της παρατήρησης $i$ στη λογαριθμοποιημένη πιθανοφάνεια, ορίζεται ως:

$$l_i(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - l_{(-i)}(\boldsymbol{\beta})$$

όπου $l_{(-i)}(\boldsymbol{\beta})$ είναι η λογαριθμοποιημένη πιθανοφάνεια αν αγνοηθεί η $i$-οστή παρατήρηση. Η τιμή του $\boldsymbol{\beta}$ που μεγιστοποιεί την $l_{(-i)}(\boldsymbol{\beta})$ συμβολίζεται με $\hat{\boldsymbol{\beta}}_{(-i)}$.

Εάν οι συνιστώσες της πιθανοφάνειας είναι ανεξάρτητες, όπως στα μοντέλα γραμμικής και λογιστικής παλινδρόμησης, η $l_i(\boldsymbol{\beta})$ είναι ίση με τη συμβολή της $i$-οστής συνιστώσας και $\sum_{i=1}^{n} l_i(\boldsymbol{\beta}) = l(\boldsymbol{\beta})$. Έτσι, ως *cross-validated συνάρτηση πιθανοφάνειας (cvl)* ορίζεται η:

$$cvl = \sum_{i=1}^{n} l_i(\hat{\boldsymbol{\beta}}_{(-i)}). \tag{3.16}$$

Για ένα μοντέλο, η $cvl$ μετρά πόσο καλά μπορεί να προβλεφθεί κάθε παρατήρηση $i$, χρησιμοποιώντας τις άλλες παρατηρήσεις. Για αυτόν τον λόγο, η συνάρτηση αυτή, εξυπηρετεί ως ένα μέτρο της ικανότητας πρόβλεψης του μοντέλου.

Εάν η μέθοδος αυτή χρησιμοποιηθεί κατάλληλα, είναι δυνατό, βελτιστοποιώντας τη συνάρτηση $cvl$, να υπολογιστούν, αριθμητικά και σχηματικά, οι βέλτιστες τιμές για τις παραμέτρους $\lambda$. Αυτές, θα χρησιμοποιηθούν στη συνέχεια για την προσαρμογή και εκτέλεση των μεθόδων με ποινή για την εύρεση του βέλτιστου στατιστικού μοντέλου.

# ΚΕΦΑΛΑΙΟ 4

# ΕΦΑΡΜΟΓΗ ΜΕ ΧΡΗΣΗ ΤΟΥ ΣΤΑΤΙΣΤΙΚΟΥ ΠΑΚΕΤΟΥ R

Torture numbers, and they'll confess to anything.

*Gregg Easterbrook*

## 4.1 Εισαγωγή

Στο κεφάλαιο αυτό θα γίνει εφαρμογή των μεθόδων που περιγράφηκαν στο κομμάτι της θεωρίας. Τα δεδομένα που θα χρησιμοποιηθούν προέρχονται από 381 ασθενείς που έχουν διαγνωσθεί με κλινική κατάθλιψη. Οι ασθενείς νοσηλεύθηκαν στην κλινική Asklepios Fachklinikum Lübben του Βερολίνου, το έτος 2012. Σκοπός της μελέτης είναι η ανεύρεση παραγόντων που επηρεάζουν τον χρόνο νοσηλείας των ασθενών. Το δείγμα συλλέχθηκε έπειτα από τη νοσοκομειακή περίθαλψη και χωρίστηκε σε δύο ομάδες ανάλογα με τη διάρκεια νοσηλείας (28 ημέρες). Οι στήλες των δεδομενων περιλαμβάνουν τα εξής:

- $X_1$ : Όνομα (*name*)
- $X_2$ : Επώνυμο (*surname*)
- $X_3$ : Ημερομηνία γέννησης (*birth*)
- $X_4$ : Ηλικία (*age*)
- $X_5$ : Διάρκεια νοσηλείας (*hosp*) - $0 :\leq 4$εβδομάδες, $1 :> 4$εβδομάδες
- $X_6$ : Αριθμός εισαγωγών στην κλινική (*adm*)
- $X_7$ : Φύλο (*gender*) - $0$ : Γυναίκα, $1$ : Άντρας

- $X_8$ : Οικογενειακή κατάσταση (*marital*) - $1$ : Άγαμος, $2$ : Σε σχέση, $3$ : Παντρεμένος, $4$ : Διαζευγμένος
- $X_9$ : Επαγγελματική κατάσταση (*profst*) - $1$ : Επαγγελματικά ενεργός, $2$ : Φοιτητής, $3$ : Άνεργος, $4$ : Συνταξιούχος
- $X_{10}$ : Αριθμός παιδιών (*child*)
- $X_{11}$ : Αυτοκτονικότητα πριν τη νοσηλεία (*suicbef*) - $0$ : Όχι, $1$ : Ναι
- $X_{12}$ : Αυτοκτονικότητα κατα τη διάρκεια της νοσηλείας (*suicdur*) - $0$ : Όχι, $1$ : Ναι
- $X_{13}$ : Απόπειρες αυτοκτονίας (*suicat*)
- $X_{14}$ : Οικογενειακό ιστορικό σε ψυχικές παθήσεις (*history*) - $0$ : Όχι, $1$ : Ναι
- $X_{15}$ : Κύρια διάγνωση (*diagn*) - $1 : F32.1, 2 : F32.2, 3 : F33.1, 4 : F33.2$
- $X_{16}$ : Συνυπάρχουσες ψυχικές παθήσεις (*coex*) - $0$ : Όχι, $1$ : Ναι
- $X_{17}$ : Σωματικές παθήσεις (*body*)
- $X_{18}$ : Διαβήτης (*diab*) - $0$ : Όχι, $1$ : Ναι
- $X_{19}$ : Καρδιαγγειακά προβλήματα (*khk*) - $0$ : Όχι, $1$ : Ναι
- $X_{20}$ : Ψυχοθεραπεία πριν τη νοσηλεία (*psychobef*) - $0$ : Όχι, $1$ : Ναι
- $X_{21}$ : Ψυχοθεραπεία κατα τη διάρκεια της νοσηλείας (*psycodur*) - $0$ : Όχι, $1$ : Ναι
- $X_{22}$ : Χρήση b-blockers (*bblock*) - $0$ : Όχι, $1$ : Ναι
- $X_{23}$ : Φαρμακευτική αγωγή (*meds*) - $0$ : Όχι, $1$ : Ναι
- $X_{24}$ : Αλλαγή φαρμακευτικής αγωγής (*changemed*) - $1$ : Όχι, $2$ : Ναι, αναποτελεσματικότητα, $3$ : Ναι, παρενέργειες
- $X_{25}$ : Υποθυρεοεϊδισμός (*hypoth*) - $0$ : Όχι, $1$ : Ναι
- $X_{26}$ : Εμφάνιση παρενεργειών από ψυχιατρική αγωγή (*uaw*) - $0$ : Όχι, $1$ : Ναι
- $X_{27}$ : Στρεσσογόνοι παράγοντες (*stress*) - $0$ : Όχι, $1$ : Ναι
- $X_{28}$ : Πρόβλημα αλκοολισμού (*alcohol*) - $0$ : Όχι, $1$ : Ναι

Μέσα στις παρενθέσεις αναφέρονται τα ονόματα τον μεταβλητών, όπως αυτά θα χρησιμοποιηθούν στη στατιστική ανάλυση. Επειδή τα δεδομένα είναι πραγματικά, περιέχουν ελλειπούσες τιμές (*missing values*), καθώς και στήλες δεδομένων οι οποίες δεν μπορούν να χρησιμοποιηθούν στην κατασκευή στατιστικού μοντέλου. Για το λόγο αυτό, αρχικά θα γίνει ένας καθαρισμός των δεδομένων. Αφού τα δεδομένα πλέον είναι κατάλληλα για ανάλυση, θα πραγματοποιηθεί περιγραφή των μεταβλητών με αριθμητικό

και γραφικό τρόπο. Στη συνέχεια, θα προσαρμοστούν δύο γενικευμένα γραμμικά μοντέλα με συναρτήσεις σύνδεσης, αυτές της logit και probit, με μεταβλητή απόκρισης την $X_5$, που αναφέρεται στη διάρκεια νοσηλείας των ασθενών ($0 :< 4$εβδομάδες, $1 :>$ 4εβδομάδες). Αφού επιλεχθεί το καταλληλότερο μοντέλο, σύμφωνα με διάφορα κριτήρια αλλά και την ερμηνευτική ευκολία, θα προσαρμοστεί το τελικό μοντέλο με το σύνολο των παρατηρήσεων. Τέλος, θα εφαρμοστεί η μέθοδος lasso, με σκοπό την επιλογή των καταλληλότερων μεταβλητών για την προσαρμογή του βέλτιστου μοντέλου.

## 4.2 Εισαγωγή των δεδομένων στην R

Τα δεδομένα έχουν αποθηκευτεί στο φάκελο *Thesis* σε αρχείο excel, τύπου .csv, το οποίο περιλαμβάνει 28 στήλες με τις αντίστοιχες μεταβλητές. Το αρχείο ονομάζεται *DataFinal.csv* και με σκοπό να περαστούν τα δεδομένα στο στατιστικό πακέτο της R, γίνεται χρήση της εντολής *read.csv()*. Η παράμετρος *header = TRUE* αναφέρεται στην πρώτη γραμμή των δεδομένων και επιβεβαιώνει ότι αποτελείται από τα ονόματα των μεταβλητών. Επιπλέον, επειδή οι τιμές της κάθε γραμμής του αρχείου χωρίζονται με το σύμβολο “;”, χρησιμοποιείται η παράμετρος *sep = ';'*, ώστε τα δεδομένα να ταξινομηθούν σε μορφή πίνακα. Τέλος, με το *na.strings = ''* τοποθετείται ένα λευκό κενό στις ελλειπούσες τιμές.

```
> thesis.data <- read.csv('/Users/Alexandra/Desktop/Thesis/Data/DataFinal.csv',
header = TRUE, sep = ';', na.strings = '')
```

Η επόμενη εντολή, σκοπό έχει την εξερεύνηση της δομής και της μορφής των δεδομένων που περάστηκαν. Έτσι, στην πρώτη γραμμή ζητείται από την R να επιστρέψει τις πρώτες 5 γραμμές από κάθε στήλη, ενώ η δεύτερη εντολή (str()) επιστρέφει τη δομή (*structure*) των μεταβλητών, συμπεριλαμβανομένων το όνομα της μεταβλητής, το είδος των τιμών, την κωδικοποίηση των επιπέδων της μεταβλητής (εάν αυτή είναι κατηγορική) και τέλος τις πρώτες τιμές της κάθε στήλης.

```
> head(thesis.data, 5)

     name   surname      birth   age  hosp  adm  gender
 Waltraud    Zillner   8/9/1938    75     1    2       0
     Rita    Zimmer  4/23/1935    78     1    1       0
   Ursula     Zenker 11/25/1940    73     1    1       0
   Marita    Zinkler  6/20/1962    51     1    2       0
  Gertrud  Ziemainz   4/3/1937    76     1    5

  marital    profst      child  suicbef  suicdur  suicat  history
        1         4          4        0        0       0        0
        1         4          4        0        0       0        0
        3         4          4        0        0       0        0
```

```
     1      3          3      0      1          1          0
     3      4          4      1      0          0          0

 diagn   coex       body   diab    khk  psychobef  psychodur
     2      1          3      1      0          0          1
     2      1          2      0      0          0          0
     4      1          2      0      0          0          0
     4      0          0      0      0          0          0
     3      1          1      0      1          0          1

bblock   meds  changemed  hypoth    uaw     stress    alcohol
     0      1          1      0      1          0          0
     1      1          2      0      0          0          0
     0      1          2      0      0          0          0
     0      1          1      0      0          0          0
     1      1          2      0      0          0          0
```

```
> str(thesis.data)
```

```
'data.frame': 381 obs. of 28 variables:
$ name : Factor w/ 217 levels " ","Adele","Adre",..: 213 175 207 146 78 202 69...
$ surname : Factor w/ 301 levels "Ahrens","Aland",..: 297 299 292 301 295 298...
$ birth : Factor w/ 335 levels " ","1/1/1938",..: 304 186 66 237 190 45 223 128
114...
$ age : int 75 78 73 51 76 54 31 54 74...
$ hosp : int 1 1 1 1 1 1 1 1 0 ...
$ adm : int 2 1 1 2 5 2 1 7 1 ...
$ gender : int 0 0 0 0 0 0 0 0 0 ...
$ marital : int 1 1 3 1 3 3 3 2 1 ...
$ profst : int 4 4 4 3 4 3 1 3 4 ...
$ child : int 4 4 4 3 4 3 1 3 4 ...
$ suicbef : int 0 0 0 0 1 0 0 1 0...
$ suicdur : int 0 0 0 1 0 0 0 0 0...
$ suicat : int 0 0 0 1 0 0 0 0 0...
$ history : int 0 0 0 0 0 0 0 0 0...
$ diagn : int 2 2 4 4 3 4 2 4 1...
$ coex : int 1 1 1 0 1 1 1 0 0...
$ body : int 3 2 2 0 1 5 4 4 3...
$ diab : int 1 0 0 0 0 0 0 0 0...
$ khk : int 0 0 0 0 1 0 0 0 1...
$ psychobef: int 0 0 0 0 0 1 0 1 0...
$ psychodur: int 1 0 0 0 1 1 1 1 0...
$ bblock : int 0 1 0 0 1 0 0 0 0 1...
```

```
$ meds : int 1 1 1 1 1 0 0 1 1 1...
$ changemed: int 1 2 2 1 2 1 1 1 1...
$ hypoth : int 0 0 0 0 0 0 0 0 0 1...
$ uaw : int 1 0 0 0 0 0 1 0 0...
$ stress : int 0 0 0 0 0 0 0 0 0...
$ alcohol : int 0 0 0 0 0 0 0 0 0...
```

Στη δομή των μεταβλητών φαίνεται ότι οι περισσότερες από αυτές, πλην των ονομάτων και της ημερομηνίας γέννησης, έχουν μεταφραστεί ως ακέραιοι αριθμοί. Η φύση των περισσοτέρων είναι κατηγορική και αυτό θα κάνουμε αμέσως μετά. Θα μετατρέψουμε, λοιπόν, τις μεταβλητές εκείνες, που είναι κατηγορικές.

```
> thesis.data$hosp <- as.factor(thesis.data$hosp)
```

```
> thesis.data$gender <- as.factor(thesis.data$gender)
```

```
> thesis.data$marital <- as.factor(thesis.data$marital)
```

```
> thesis.data$profst <- as.factor(thesis.data$profst)
```

```
> thesis.data$suicbef <- as.factor(thesis.data$suicbef)
```

```
> thesis.data$suicdur <- as.factor(thesis.data$suicdur)
```

```
> thesis.data$history <- as.factor(thesis.data$history)
```

```
> thesis.data$diagn <- as.factor(thesis.data$diagn)
```

```
> thesis.data$coex <- as.factor(thesis.data$coex)
```

```
> thesis.data$diab <- as.factor(thesis.data$diab)
```

```
> thesis.data$khk <- as.factor(thesis.data$khk)
```

```
> thesis.data$psychobef <- as.factor(thesis.data$psychobef)
```

```
> thesis.data$psychodur <- as.factor(thesis.data$psychodur)
```

```
> thesis.data$bblock <- as.factor(thesis.data$bblock)
```

```
> thesis.data$meds <- as.factor(thesis.data$meds)
```

```
> thesis.data$changemed <- as.factor(thesis.data$changemed)
```

```
> thesis.data$hypoth <- as.factor(thesis.data$hypoth)
```

```
> thesis.data$uaw <- as.factor(thesis.data$uaw)
```

```
> thesis.data$stress <- as.factor(thesis.data$stress)
```

```
> thesis.data$alcohol <- as.factor(thesis.data$alcohol)
```

Εάν, τώρα, καλέσουμε πάλι την εντολή *str()*, για τη δομή των μεταβλητών θα δούμε ότι οι κατηγορικές μεταβλητές μετατράπηκαν σε *factors*, όπως ήταν το επιθυμητό. Τα ονόματα των επιπέδων έχουν χαρακτηριστεί όπως φαίνεται στην πρώτη ενότητα.

## 4.3 Καθαρισμός δεδομένων

Όπως προαναφέρθηκε, λόγω του ότι τα δεδομένα είναι αληθινά, υπάρχουν κάποιες ελλειπούσες τιμές, όπως επίσης και κάποια στοιχεία που δεν θα φανούν χρήσιμα στην πορεία της ανάλυσης. Για να μπορέσουμε να αποφανθούμε για αυτά, θα τρέξουμε την εντολή *summary()*, η οποία θα δώσει μία σύντομη εικόνα των μεταβλητών.

```
> summary(thesis.data)

      name          surname          birth          age
 Ines : 7       Richter: 6   10/21/1928: 5   Min. :21.0
 Barbara: 6     Aland : 5    1/11/1964 : 3   1st Qu.:47.0
 Erika : 6      Noack : 5    1/30/1962 : 3   Median :53.0
 Gudrun : 6     Schulze: 5   10/29/1942: 3   Mean :54.4
 Andreas: 5     Lange : 4    11/5/1960 : 3   3rd Qu.:63.0
 (Other):350    Lehmann: 4   2/27/1957 : 3   Max. :92.0
 NA's : 1       (Other):352  (Other) :361

      hosp          adm          gender        merital
 0:123       Min. : 1.000     0: 265       1 :116
 1:258       1st Qu.: 1.000   1 :116       2 : 60
             Median : 1.000                3 :177
             Mean : 1.926                  4 : 24
             3rd Qu.: 2.000                NA's :4
             Max. :14.000
             NA's :1

      profst          child          suicbef        suicdur
 1:130       Min. :1.000      0 :213       0 :321
 2 : 10      1st Qu.:1.000    1 :168       1 : 60
 3 : 86      Median :3.000
 4 :151      Mean :2.703
 NA's : 4    3rd Qu.:4.000
             Max. :5.000
             NA's :1

      suicat          history        diagn          coex
 Min. :0.000      0 :316       1 : 28       0 :154
 1st Qu.:1.000    1: 65        2 : 61       1 :227
 Median :1.000                 3 : 49
 Mean :1.079                   4 :243
 3rd Qu.:1.000
 Max. :4.000

      body          diab          khk       psychobef
 Min. : 0.000     0 :336       0 :352       0 :198
 1st Qu.: 1.000   1: 45        1: 29        1:183
 Median : 2.000
 Mean : 2.194
 3rd Qu.: 3.000
 Max. :15.000
```

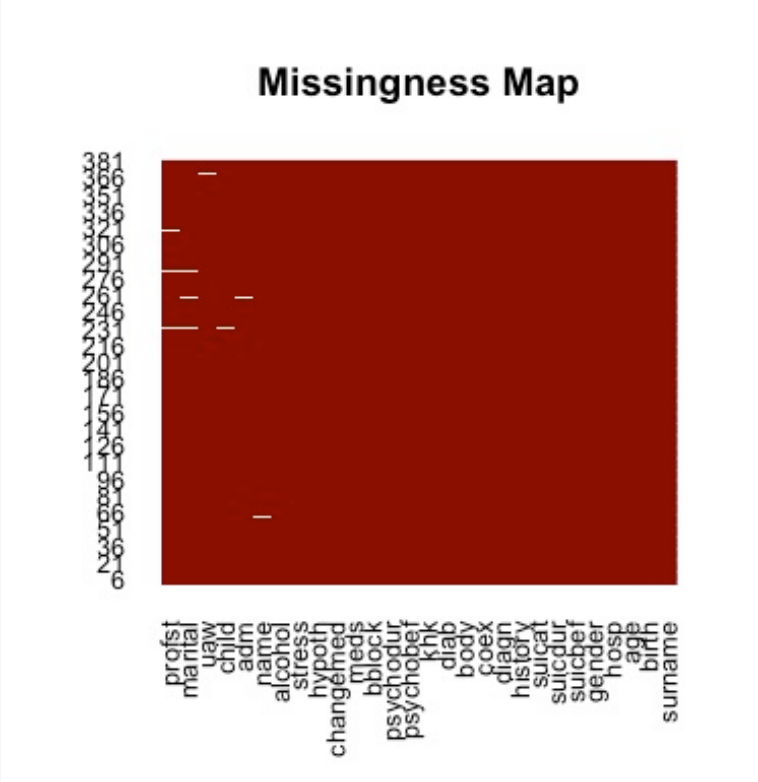
```
psychodur    bblock     meds    changemed
   0 : 55    0 :288    0 : 44     1 :290
   1:326      1 : 93   1:337       2 : 83
                                    3 : 8

    hypoth       uaw    stress     alcohol
   0 :344    0 :326   0 :278      0 :355
    1 : 37    1 : 54   1 :103      1 : 26
              NA's: 1
```

Όπως παρατηρούμε στα παραπάνω αποτελέσμα υπάρχουν ελλειπούσες τιμές. Για να απεικονίσουμε τις τιμές που λείπουν από τα δεδομένα θα χρησιμοποιήσουμε το πακέτο *Amelia*, που έχει δημιουργηθεί ειδικά γι'αυτόν τον λόγο. Έτσι, αφού φορτώσουμε τη βιβλιοθήκη στην R, θα δημιουργήσουμε το γράφημα απεικόνισης των ελλειπουσών τιμών με την εντολή (*missmap()*):

```
> library(Amelia)
```

```
> missmap(thesis.data, legend = F, col = c('beige', 'dark red'))
```




Με λίγα λόγια, ο λόγος που δημιουργήσαμε αυτό το γράφημα είναι για να μπορέσουμε, με εύκολο τρόπο, να δούμε εάν οι τιμές που λείπουν από τα δεδομένα είναι

πολλές και μπορούν να επηρεάσουν την ανάλυση. Όπως παρατηρούμε από το συγκεκριμένο διάγραμμα, ο αριθμός των ελλειπουσών τιμών είναι μικρός και για αυτόν τον λόγο είναι δυνατό να τις αφαιρέσουμε χωρίς να δημιουργηθεί ουσιαστική βλάβη στα τελικά αποτελέσματα. Έτσι, χρησιμοποιούμε την εντολή που παρουσιάζεται παρακάτω, η οποία επιστρέφει τις γραμμές της βάσης δεδομένων που περιέχουν τις τιμές που λείπουν:

```
> thesis.data[!complete.cases(thesis.data),]

     name   surname     birth  age  hosp   adm
    Karin  Wernitz   2/4/1972   40     0     1
   Marion  Steppan  4/15/1964   49     0     1
 Brigitte  Reimann  8/21/1953   59     1     4
     Anke   Roelle 10/10/1970   43     1     2
  Andreas   Müller  2/16/1961   52     1    NA
     Anne Ordowksy   1/2/1960   53     0     2
     <NA>   Dietze  9/20/1961   53     1     1

  gender  marital   profst  child  suicbef  suicdur
       0        4        3      3        0        0
       0        3     <NA>      5        0        0
       0     <NA>     <NA>      5        1        0
       0     <NA>     <NA>      5        0        0
       1     <NA>        4      4        0        0
       0     <NA>     <NA>     NA        0        0
       1        3        4      4        1        0

  suicat  history    diagn   coex   body     diab
       0        0        1      0      0        0
       1        0        2      1      0        0
       1        0        4      0      2        0
       1        0        3      1      3        0
       1        0        4      1      2        0
       1        0        4      1      2        0
       2        1        4      1      3        0
       1        0        4      1      4        1

     khk psychobef psychodur bblock   meds changemed
       0        0        0      0      0        1
       0        0        0      1      1        1
       0        1        1      1      1        1
       0        1        1      1      1        1
       0        0        1      0      1        2
       0        1        1      0      1        1
       0        0        1      1      1        1
```

```
 hypoth    uaw  stress  alcohol
      0   <NA>       0        0
      0      0       0        0
      0      0       0        0
      0      0       0        0
      0      0       0        0
      0      0       0        0
      0      0       0        0
```

Έτσι, το πρώτο που παρατηρούμε είναι ότι οι γραμμές που περιέχουν τουλάχιστον μία ελλειπούσα τιμή, είναι 7 σε αριθμό. Πρωτού, όμως, προβούμε σε διαγραφή αυτών, θα αφαιρέσουμε τις μεταβλητές οι οποίες δε μας προσφέρουν κάποια σημαντική πληροφορία για την ανάλυση. Αυτές είναι το όνομα και το επώνυμο του ασθενούς, καθώς και η ημερομηνία γέννησης. Για το σκοπό αυτό, θα χρησιμοποιήσουμε τη βιβλιοθήκη *dplyr*, και θα κάνουμε χρήση της εντολής *select()*.

```
> library(dplyr)
```

```
> thesis.data <- select(thesis.data, -name, -surname, -birth)
```

Έτσι, εάν καλέσουμε πάλι την προηγούμενη εντολή που επέστρεφε τις γραμμές με τις ελλειπούσες τιμές, θα παρατηρήσουμε ότι οι γραμμές πλέον μειώθηκαν σε 6. Αυτό συνέβει, όπως εύκολα μπορούμε να δούμε στον προηγούμενο πίνακα, διότι εμφανιζόταν μία γραμμή που δεν περιείχε ένα όνομα. Πράγμα που δε μας ενδιαφέρει στη συγκεκριμένη ανάλυση και θα ήταν λάθος να αφαιρέσουμε τα στοιχεία αυτής της γραμμής, πρωτού αφαιρέσουμε τη συγκεκριμένη μεταβλητή. Τώρα, μπορούμε να προχωρήσουμε στη διαγραφή των παραπάνω γραμμών με τις παρακάτω εντολές:

```
> thesis.data <- thesis.data[!is.na(thesis.data$marital),]
```

```
> thesis.data[!complete.cases(thesis.data),]
```

```
> thesis.data <- thesis.data[!is.na(thesis.data$profst),]
```

```
> thesis.data[!complete.cases(thesis.data),]
```

```
> thesis.data <- thesis.data[!is.na(thesis.data$uaw),]
```

```
> thesis.data[!complete.cases(thesis.data),]
```

Κάθε φορά που διαγράφουμε γραμμές, ελέγχουμε αυτές που έχουν απομείνει για να είμαστε σίγουροι ότι χρειάζεται να συνεχίσουμε. Αυτό διότι, διαγράφοντας μία γραμμή του πίνακα, είναι δυνατόν να διαγράφονται παραπάνω από μία ελλειπούσες τιμές. Επιπλέον, επειδή μαζί με τις γραμμές του πίνακα, διαγράφονται και οι δείκτες της κάθε μίας, θέλουμε να τους επαναφέρουμε. Αυτό το καταφέρνουμε με την παρακάτω εντολή:

```
> rownames(thesis.data) <- NULL
```

## 4.4 Περιγραφή των μεταβλητών

### 4.4.1 Περιγραφικά στατιστικά

Σε αυτήν την ενότητα, θα παρουσιάσουμε κάποια περιγραφικά στατιστικά χαρακτηριστικά για τις μεταβλητές του μοντέλου. Μία πρώτη εικόνα των δεδομένων παρουσιάστηκε στην προηγούμενη ενότητα μέσω της εντολής *summary()*. Θα δημιουργήσουμε μερικά γραφήματα, που πιστεύουμε ότι θα παρουσιάζουν ενδιαφέρον. Να υπενθυμίσουμε ότι, ως μεταβλητή απόκρισης θα χρησιμοποιηθεί η μεταβλητή που αφορά τη νοσηλεία των ασθενών και για αυτό το λόγο τα γραφήματα θα παρουσιαστούν συναρτήσει αυτής. Επιπλέον, θα δούμε πώς κατανέμονται αριθμητικά οι τιμές των μεταβλητών αυτών σε συνάρτηση με τη νοσηλεία. Για τα γραφήματα θα γίνει χρήση της βιβλιοθήκης *ggplot2*.

```
> library(ggplot2)
```

Για να διευκολυνθεί η διατύπωση των μεταβλητών στην παρακάτω διαδικασία, θα χρησιμοποιήσουμε την εντολή *attach()*, όπου μας επιτρέπει να καλέσουμε τις μεταβλητές μόνο με το όνομά τους.

```
> attach(thesis.data)
```

Καλώντας την εντολή *summary()* για τη μεταβλητή “age”, παρατηρούμε ότι εμφανίζονται κάποια στατιστικά στοιχεία αυτής. Έτσι, οι πληροφορίες που λαμβάνουμε είναι ότι οι ασθενείς του δείγματός μας έχουν μέση ηλικία τα $54.48$ έτη, ενώ η μέγιστη ηλικία που περιλαμβάνεται στο δείγμα είναι $92$ ετών. Η τυπική απόκλιση των τιμών της μεταβλητής είναι ίση με $15.70496$, η οποία μάς πληροφορεί ότι η διασπορά των τιμών, γύρω από την αναμενόμενη μέση τιμή, δεν είναι αρκετά μεγάλη.

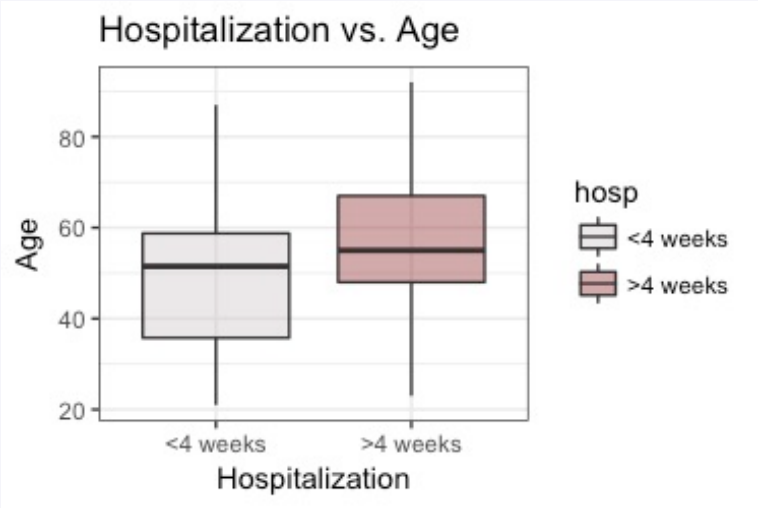
```
> summary(age)

   Min.   1st Qu.   Median    Mean   3rd Qu.    Max.
  21.00     47.00    54.00   54.48     64.00   92.00
```

```
> sd(age)

[1] 15.70496
```

Για να δημιουργήσουμε την εικόνα της μεταβλητής age, συναρτήσει της μεταβλητής απόκρισης hosp, χρησιμοποιούμε ένα θηκογράφημα που παίρνει την παρακάτω μορφή:



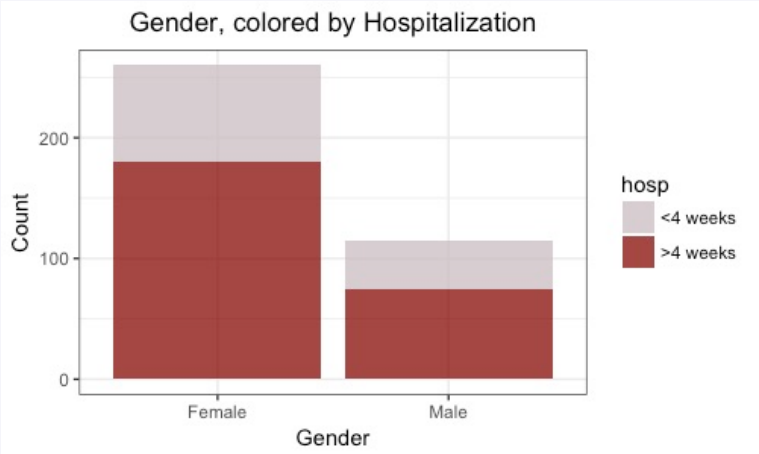

Στο παραπάνω θηκογράφημα, το δείγμα της μεταβλητής age χωρίζεται σε δύο περιπτώσεις, ανάλογα με τις ημέρες νοσηλείας. Έτσι, στο πρώτο σχήμα, που αντιστοιχεί στους ασθενείς που νοσηλεύθηκαν λιγότερες από 4 εβδομάδες, παρατηρούμε μεγαλύτερο ενδοτεταρτημοριακό εύρος, έναντι του θηκογραφήματος δεξιά. Επιπλέον, στο πρώτο δείγμα η διάμεσος πλησιάζει περισσότερο στην πάνω πλευρά του ορθογωνίου ($1^o$ τεταρτημόριο των παρατηρήσεων) και συνεπώς η κατανομή των δεδομένων έχει θετική λοξότητα. Αντίθετα, στο δεύτερο θηκογράφημα παρουσιάζεται αρνητική λοξότητα στην κατανομή των παρατηρήσεων, καθώς και μικρότερο ενδοτεταρτημοριακό εύρος.

Η επόμενη μεταβλητή που θα παρουσιάσουμε είναι αυτή που αφορά το φύλο των ασθενών. Θέλοντας να δούμε πώς κατανέμονται οι παρατηρήσεις ανάλογα με τον αριθμό ημερών νοσηλείας, έχουμε τα παρακάτω στατιστικά στοιχεία:

```
> table(hosp, gender)

      hosp  Female   Male
  <4 weeks      80     40
  >4 weeks     180     75
```

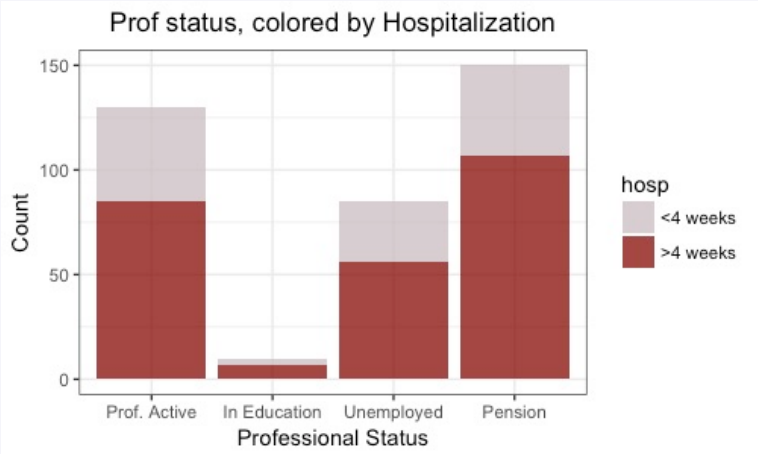
Είναι φανερό ότι το δείγμα αποτελείται από περισσότερες γυναίκες συγκριτικά, με τον αριθμό των αντρών. Η διαφορά, ωστόσο, δεν θεωρείται ότι είναι αρκετά μεγάλη ώστε να δημιουργήσει προβλήματατα στην ανάλυση. Το αντίστοιχο ραβδόγραμμα της μεταβλητής φαίνεται παρακάτω και είναι εμφανές ότι οι ασθενείς που νοσηλεύθηκαν παραπάνω από 4 εβδομάδες στην κλινική υπερτερούν σε αριθμό. Παρακάτω, θα δούμε πώς, και αν αυτή η μεταβλητή μπορεί να επηρεάσει στατιστικά το μοντέλο.




Μία ακόμα μεταβλητή που θεωρούμε ότι είναι ενδιαφέρον να παρουσιαστεί αφορά την επαγγελματική δραστηριότητα των ασθενών. Είναι εμφανές από τις συχνότητες που παρουσιάζονται στον πίνακα, ότι η καταθλιπτική διαταραχή σε επίπεδο κλινικής νοσηλείας επηρεάζει περισσότερο τους ασθενείς που είναι στο στάδιο σύνταξης. Επίσης, λιγότερη εμφάνιση της νόσου παρουσιάζεται σε ανθρώπους που βρίσκονται στην εκπαίδευση. Συγκεντρωτικά τα αποτελέσματα φαίνονται παρακάτω, μαζί με το αντίστοιχο ραβδόγραμμα.

```
> table(hosp, profst)
```

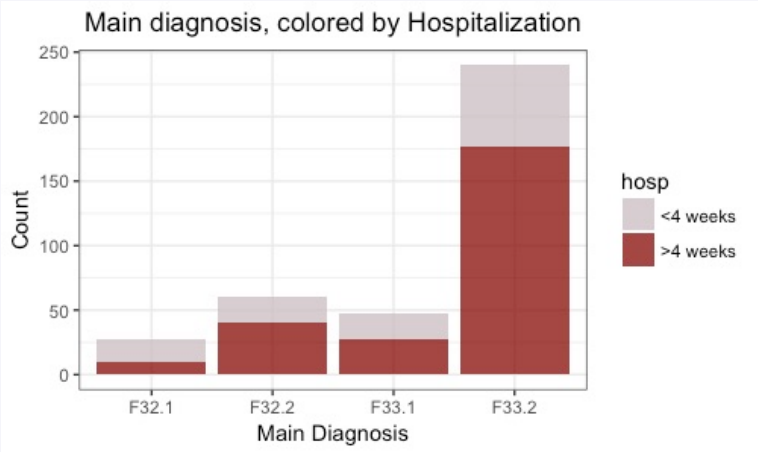
| hosp | Professionally Active | In Education | Unemployed | Pension |
|---|---|---|---|---|
| <4 weeks | 45 | 3 | 29 | 43 |
| >4 weeks | 85 | 7 | 56 | 107 |




Τέλος, παρατηρώντας τη μεταβλητή της κύριας διάγνωσης βλέπουμε αρχικά ότι χωρίζεται σε τέσσερις κατηγορίες. Οι κατηγορίες αυτές εκφράζουν τον τύπο και την βαρύτητα εμφάνισης της νόσου. Έτσι, με μία γρήγορη ματιά είναι εμφανές ότι η νοσηλεία των ασθενών είναι συχνότερη σε ασθενείς με, βαριάς μορφής, υποτροπιάζουσα καταθλιπτική διαταραχή. Όπως επίσης, είναι περισσότερο συχνό, ασθενείς αυτής της κατηγορίας, να νοσηλεύονται περισσότερο καιρό από άλλους.

```
> table(hosp, diagn)
```

| hosp | F32.1 | F32.2 | F33.1 | F33.2 |
|---|---|---|---|---|
| <4 weeks | 17 | 20 | 20 | 63 |
| >4 weeks | 10 | 40 | 28 | 177 |




### 4.4.2 Συσχετίσεις

Πριν προχωρήσουμε στην προσαρμογή του στατιστικού μοντέλου, κρίνεται απαραίτητο να ελεγχθούν πιθανές συσχετίσεις μεταξύ των μεταβλητών καθώς, ως

προβλεπτικοί παράγοντες έχουν χρησιμοποιηθεί και ποσοτικές μεταβλητές. Επομένως, αρχικά θα πρέπει να ξεχωρίσουμε αυτές τις μεταβλητές από το σύνολο το δεδομένων. Αυτό μπορεί να πραγματοποιηθεί με την εντολή *sapply()*, όπου μάς δίνει τη δυνατότητα να εφαρμόσουμε κάποια συνάρτηση της R σε όλο το μήκος των δεδομένων, χωρίς να χρειάζεται να το κάνουμε μεμονωμένα για κάθε μεταβλητή. Έτσι, εφόσον θέλουμε να ξεχωρίσουμε τις μεταβλητές που έχουν αριθμητικό χαρακτήρα, θα πρέπει να καλέσουμε τη συνάρτηση *is.numeric()*.

```
> numcols <- sapply(thesis.data, is.numeric)
```

Έπειτα, προχωρούμε στον έλεγχο της συσχέτισης με την εντολή *cor()*, που περιλαμβάνεται στις βασικές συναρτήσεις του στατιστικού πακέτου. Καλούμε την συνάρτηση συσχέτισης λοιπόν, πάνω στα αριθμητικά δεδομένα που ξεχωρίσαμε πριν και τα τυπώνουμε, λαμβάνοντας τα εξής αποτελέσματα:

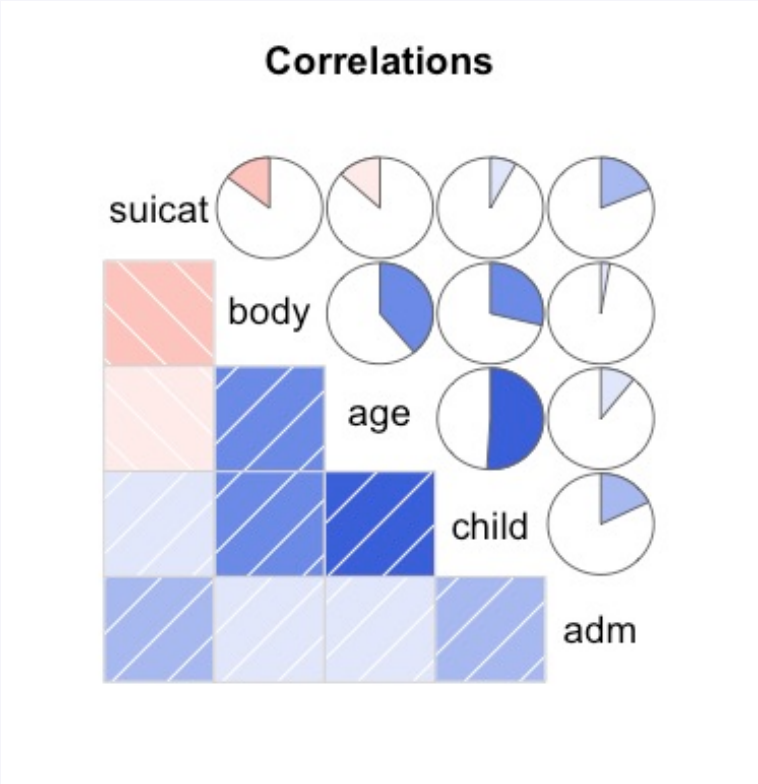
```
> cor.data <- cor(thesis.data[, numcols])
```

```
> print(cor.data)
```

| | age | adm | child | suicat | body |
|---|---|---|---|---|---|
| age | 1.0000000 | 0.10706077 | 0.5102297 | -0.1330360 | 0.38911752 |
| adm | 0.1070608 | 1.00000000 | 0.1785071 | 0.1874704 | 0.02764855 |
| child | 0.5102297 | 0.17850710 | 1.0000000 | 0.0799134 | 0.28795131 |
| suicat | -0.1330360 | 0.18747044 | 0.0799134 | 1.0000000 | -0.14570682 |
| body | 0.3891175 | 0.02764855 | 0.2879513 | -0.1457068 | 1.00000000 |

Για ευκολότερα ερμηνεύσιμα αποτελέσματα, θα δημιουργήσουμε και το γράφημα συσχετίσεων. Γι'αυτόν το σκοπό, θα πρέπει να κάνουμε χρήση της αντίστοιχης βιβλιοθήκης *corrgram* και έπειτα να επιχειρήσουμε να το δημιουργήσουμε. Συνεπώς, μετά από αυτά τα βήματα, λαμβάνουμε την εξής εικόνα:

```
> library(corrgram)
```




Αρχικά, θα πρέπει να αναφέρουμε ότι στα δεδομένα μας οι ποσοτικές μεταβλητές που τέθηκαν σε αυτή τη διαδικασία είναι οι: age, adm, child, suicat και body. Το διάγραμμα παρουσιάζει τις συσχετίσεις μεταξύ αυτών και χωρίζεται σε δύο μέρη. Έχουμε επιλέξει αυτόν τον τρόπο γιατί θεωρούμε ότι είναι ευκολότερα ερμηνεύσιμος. Έτσι, στο κάτω μισό του τετραγωνικού πίνακα, παρουσιάζονται οι συσχετίσεις μεταξύ των μεταβλητών με χρωματικό τρόπο. Όσο πιο σκούρο μπλε ή κόκκινο, χρωματίζεται το κάθε τετράγωνο, τόσο πιο έντονα συσχετίζονται οι αντίστοιχες μεταβλητές. Το μπλε χρώμα αναφέρεται σε συσχετίσεις που έχουν υπολογιστεί με θετικό πρόσημο, ενώ το κόκκινο απεικονίζει τα αρνητικά αποτελέσματα (βλέπε πίνακα συσχετίσεων). Από την άλλη μεριά, στο πάνω μισό του γραφήματος χρησιμοποιήθηκε η απεικόνιση μέσω "πίτας". Με αυτόν τον τρόπο, καθίσταται πιο εμφανές το ποσό συσχέτισης μεταξύ δύο μεταβλητών. Συνεπώς, στα γραφικά αποτελέσματα είναι αμέσως ορατή η έντονη συσχέτιση των μεταβλητών child και age ($= 0.5102$), καθώς επίσης και η συσχέτιση μεταξύ των μεταβλητών age και body ($= 0.3891$). Το γεγονός αυτό μάς οδηγεί στο συμπέρασμα της πολυσυγγραμμικότητας για το μοντέλο μας.

Λόγω της συσχέτισης που παρουσιάζει η μεταβλητή child, θα επιλέξουμε να την αφαιρέσουμε από το στατιστικό μοντέλο, καθώς είναι πολύ πιθανό να δημιουργηθούν προβλήματα κατα τη διάρκεια προσαρμογής του. Έτσι, με τη χρήση της εντολής *select()*, της βιβλιοθήκης *dplyr*, όπως κάναμε και παραπάνω, θα πραγματοποιήσουμε αυτήν την επιθυμία:

```
> thesis.data <- select(thesis.data, -child)
```

Πλέον, είμαστε έτοιμοι για περάσουμε στην προσαρμογή του λογιστικού μοντέλου παλινδρόμησης.

## 4.5 Επιλογή στατιστικού μοντέλου

Με σκοπό την αξιολόγηση του προβλεπτικού μοντέλου, θα χωρίσουμε το δείγμα σε δύο μέρη. Το *training sample*, το οποίο αποτελείται από το 70% των παρατηρήσεων και χρησιμοποιείται για την ανάπτυξη του μοντέλου και το *test sample*, το οποίο περιλαμβάνει το υπόλοιπο 30% των παρατηρήσεων και χρησιμοποιείται για την αξιολόγηση του μοντέλου. Ο χωρισμός του δείγματος γίνεται με τυχαίο τρόπο. Η διαδικασία αυτή θα πραγματοποιηθεί μέσω της βιβλιοθήκης *caTools* και με χρήση των εντολών *sample.split* και *subset*.

```
> library(caTools)
```

```
> split <- sample.split(thesis.data$hosp, SplitRatio = 0.7)
```

```
> train <- subset(thesis.data, split == T)
```

```
> test <- subset(thesis.data, split == F)
```

### 4.5.1 Μοντέλο λογιστικής παλινδρόμησης

Στη συνέχεια, προσαρμόζουμε το μοντέλο της λογιστικής παλινδρόμησης για τη δίτιμη μεταβλητή απόκρισης hosp και λαμβάνουμε τα παρακάτω αποτελέσματα.

```
> logit.model <- glm(hosp ~ ., family = binomial(link = 'logit'), data = train)
```

```
> summary(logit.model)

Call:
glm(formula = hosp ~ ., family = binomial(link = "logit"), data = train)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
 -2.4693  -0.6626   0.3478   0.6721   2.0578
```

```
Coefficients:
              Estimate   Std. Error  z value   Pr(>|z|)
(Intercept) -4.1504048  1.3020814   -3.188   0.00144  **
        age  0.0347391  0.0176482    1.968   0.04902  *
        adm -0.2344936  0.1432244   -1.637   0.10158
    gender1 -0.6294192  0.3912213   -1.609   0.10765
   marital2 -0.3280215  0.5446443   -0.602   0.54700
   marital3 -0.0739901  0.4583458   -0.161   0.87176
   marital4  0.5363207  0.7698973    0.697   0.48604
    profst2  1.8106680  1.1178583    1.620   0.10528
    profst3  0.0859081  0.4696398    0.183   0.85486
    profst4 -1.0200602  0.5662597   -1.801   0.07164  .
   suicbef1 -0.0407552  0.4133323   -0.099   0.92145
   suicdur1  0.8309733  0.5412034    1.535   0.12468
     suicat  0.0002098  0.3948129    0.001   0.99958
   history1  0.8277028  0.5335473    1.551   0.12082
     diagn2  0.7219104  0.7802918    0.925   0.35487
     diagn3  0.5713669  0.7557140    0.756   0.44961
     diagn4  1.4075028  0.6506331    2.163   0.03052  *
      coex1 -0.6894530  0.3809290   -1.810   0.07031  .
       body  0.1292276  0.1177983    1.097   0.27263
      diab1  0.0620951  0.7097664    0.087   0.93028
       khk1  2.0864946  1.3835020    1.508   0.13152
  psychobef1  0.0074718  0.3986250    0.019   0.98505
  psychodur1  2.5764256  0.6167536    4.177   2.95e-05 ***
    bblock1  1.8709973  0.5878704    3.183   0.00146  **
      meds1  0.0920771  0.5823557    0.158   0.87437
 changemed2  1.8065899  0.6265667    2.883   0.00394  **
 changemed3 -0.6889081  1.4651810   -0.470   0.63822
    hypoth1 -0.0198746  0.5897882   -0.034   0.97312
       uaw1  1.4993074  0.7849858    1.910   0.05614  .
     stress1 -0.6774048  0.4130158   -1.640   0.10098
    alcohol1  0.4080314  0.7550732    0.540   0.58893
—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 328.72 on 261 degrees of freedom
Residual deviance: 226.52 on 231 degrees of freedom
AIC: 288.52

Number of Fisher Scoring iterations: 6
```

Με σκοπό να συγκρίνουμε το μοντέλο της λογιστικής παλινδρόμησης, με αυτό της

μεθόδου probit, θα χρησιμοποιήσουμε τα κριτήρια AIC και BIC. Έτσι, τα μέτρα αυτά επιστρέφονται με τις εξής εντολές:

```
> AIC(logit.model)
[1] 288.5211
```

```
> BIC(logit.model)
[1] 399.1397
```

Για να μπορέσουμε να εκτιμήσουμε την προβλεπτική ικανότητα του μοντέλου logit, θα εφαρμόσουμε το προσαρμοσμένο μοντέλο στα δεδομένα *test*. Αυτό πραγματοποιείται με την εντολή *predict()*. Ως παράμετρο στην εντολή αυτή, θα χρησιμοποιήσουμε το *type = 'response'*, επειδή το πρόβλημα είναι ένα πρόβλημα ταξινόμησης.

```
> fit.prob.logit <- predict(logit.model, test, type = 'response')
```

Στη συνέχεια, θα πραγματοποιήσουμε μία ανάλυση ROC, ώστε να βρούμε το ιδανικό σημείο cut-off. Αυτό το σημείο θα το χρησιμοποιήσουμε, με σκοπό την ταξινόμηση των προβλεφθείσων τιμών. Για την ανάλυση αυτή θα χρησιμοποιηθεί η βιβλιοθήκη *pROC*.

```
> library(pROC)
```

Για την ROC ανάλυση θα χρησιμοποιηθεί η κύρια συνάρτηση αυτού του πακέτου, η οποία είναι η *roc()*. Η εντολή αυτή κάνει χρήση των πραγματικών τιμών της μεταβλητής απόκρισης *hosp*, στο δείγμα *test*, καθώς και των προσαρμοσμένων τιμών που υπολογίστηκαν με την εντολή *predict()*.

```
> analysis.logit <- roc(test$hosp, fit.prob.logit)
```

Στη συνέχεια, θα πραγματοποιήσουμε την ένωση δύο στηλών. Στην πρώτη στήλη θα περιλαμβάνονται τα κάτω όρια της καμπύλης ROC που υπολογίστηκαν και αποθηκεύτηκαν στη μεταβλητή *analysis.logit*, ενώ στη δεύτερη στήλη θα περιλαμβάνεται το άθροισμα των τιμών της "ευαισθησίας" και της "ειδικότητας", τα οποία έχουν περιγραφεί αναλυτικά στη θεωρία (βλ. ενότητα 2.5.9).

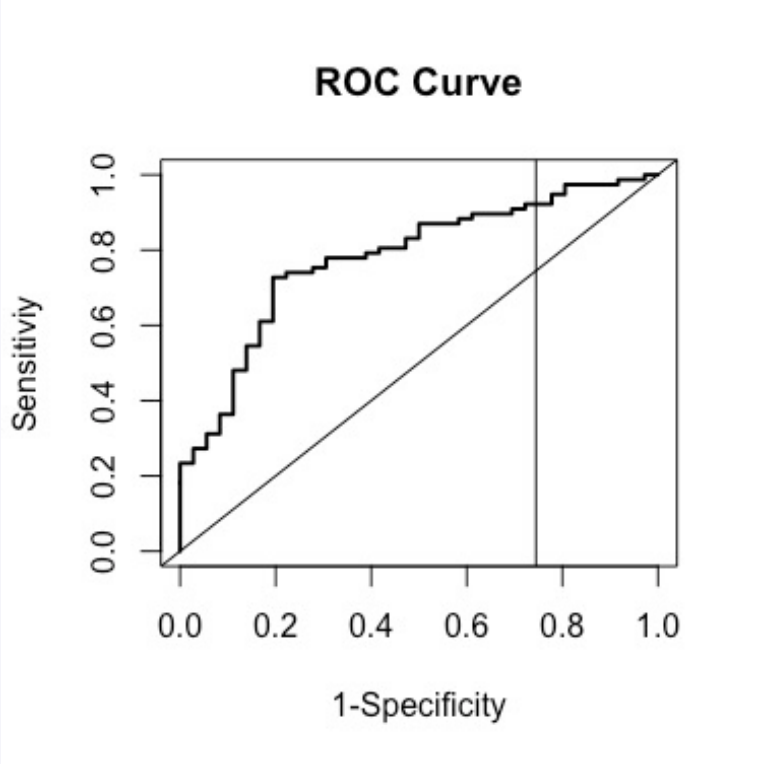
```
> e.l <- cbind(analysis.logit$thresholds, analysis.logit$sensitivities +
analysis.logit$specificities)
```

Με την επόμενη εντολή, απομονώνουμε τελικά την βέλτιστη τιμή της παραμέτρου *t*, την οποία και θα χρησιμοποιήσουμε ως σημείο cut-off, στην ταξινόμηση των

εκτιμημένων παρατηρήσεων.

```
> opt.t.l <- subset(e.l, e.l[,2]==max(e.l[,2]))[,1]
```

Με γραφικό τρόπο, η καμπύλη ROC για το μοντέλο logit που έχουμε προσαρμόσει, φαίνεται παρακάτω.




Τώρα, είμαστε σε θέση να εκτιμήσουμε τις προσαρμοσμένες τιμές με την εντολή *ifelse()*. Ως παράμετροι θα χρησιμοποιηθούν οι προβλεφθείσες πιθανότητες που υπολογίστηκαν με την εντολή *predict()* και αποθηκεύτηκαν με το όνομα *fit.prob.logit*. Επιπλέον, θα χρησιμοποιηθεί το βέλτιστο σημείο cut-off που υπολογίστηκε με τη διαδικασία ROC, με σκοπό την ταξινόμηση των παρατηρήσεων στην κάθε κατηγορία της μεταβλητής απόκρισης.

```
> opt.t.l
[1] 0.7446527
```

```
> fit.results.logit <- ifelse(fit.prob.logit > 0.7446527, 1, 0)
```

Ουσιαστικά, με την έκφραση αυτή εννοούμε: εάν η προσαρμοσμένη πιθανότητα είναι μεγαλύτερη από 0.7446527, τοποθέτησε την τιμή 1 ($>$ 4 εβδομάδες), διαφορε-

τικά βάλε 0 ($\leq$ 4 εβδομάδες). Θέλοντας να υπολογίσουμε το σφάλμα ταξινόμησης, χρησιμοποιούμε την εντολή *mean()*.

```
> missClassError.logit <- mean(fit.results.logit != test$hosp)
```

Το επόμενο αποτέλεσμα, προέρχεται αν αφαιρέσουμε το σφάλμα ταξινόμησης από τη μονάδα και αποτελεί την ακρίβεια πρόβλεψης του μοντέλου. Η ακρίβεια του μοντέλου logit που προσαρμόσαμε, εκτιμάται $\simeq$ 75%, ποσοστό που είναι αρκετά ικανοποιητικό.

```
> print(1 - missClassError.logit)
[1] 0.7522124
```

Τελικά, ο πίνακας ταξινόμησης, που παρουσιάζει το σύνολο των τιμών που τοποθετήθηκαν σωστά και αυτών που τοποθετήθηκαν λανθασμένα, φαίνεται παρακάτω.

```
> table(fit.prob.logit > 0.7446527)
 FALSE  TRUE
    50    63
```

Τέλος, για να συγκρίνουμε το υποψήφιο, με το κορεσμένο μοντέλο, υπολογίζουμε την *p-τιμή* του ελέγχου $\chi^2$:

```
> 1 - pchisq(logit.model$deviance, logit.model$df.residual)
[1] 0.5709008
```

Το συγκεκριμένο αποτέλεσμα, δείχνει ότι το προσαρμοσμένο μοντέλο είναι ένα καλό εναλλακτικό μοντέλο, έναντι του κορεσμένου. Ο λόγος που ισχύει αυτό, είναι διότι η *p-τιμή* του ελέγχου $\chi^2$ υπολογίζεται μεγαλύτερη του $0.05$. Το στατιστικό, αυτό, μέτρο αποτελεί μία ένδειξη της καλής προσαρμογής του μοντέλου στα δεδομένα. Συμπερασματικά, παρ'όλο που οι επεξηγηματικές μεταβλητές δεν είναι όλες στατιστικά σημαντικές, ο έλεγχος Wald μας επιβεβαιώνει ότι η παρουσία όλων βελτιώνει την απόδοση του προσαρμοσμένου μοντέλου. Έτσι, καταλήγουμε ότι το μοντέλο της λογιστικής παλινδρόμησης προσαρμόζεται ικανοποιητικά στα εν λόγω δεδομένα. Θα προχωρήσουμε στην ανάλυση του ίδιου δείγματος με τη μέθοδο probit, ώστε να συγκρίνουμε τα αποτελέσματα των δύο τεχνικών.

### 4.5.2 Μοντέλο probit

Η προσαρμογή του μοντέλου probit, θα πραγματοποιηθεί και πάλι στα δεδομένα *train* και *test*, τα οποία περιέχουν τις ίδιες ακριβώς παρατηρήσεις με πριν. Έτσι, θα συγκρίνουμε πιο αξιόπιστα τις δύο μεθόδους, με βάση τα τελικά τους αποτελέσματα.

Χρησιμοποιώντας ως μεταβλητή απόκρισης την hosp, λαμβάνουμε το μοντέλο probit ως εξής:

```
> probit.model <- glm(train$hosp ~ ., family = binomial(link = 'probit'), data =
train)
```

```
> summary(probit.model)

Call:
glm(formula = train$hosp ~ ., family = binomial(link = "probit"), data = train)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
 -2.4846  -0.6556   0.3293   0.6839   2.0341
Coefficients:
              Estimate  Std. Error  z value   Pr(>|z|)
(Intercept)  -2.483365   0.754413   -3.292   0.000996  ***
       age    0.021247   0.010248    2.073   0.038137  *
       adm   -0.134912   0.083359   -1.618   0.105565
   gender1   -0.359934   0.225817   -1.594   0.110954
  marital2   -0.199742   0.318395   -0.627   0.530436
  marital3   -0.054550   0.264241   -0.206   0.836448
  marital4    0.323827   0.443363    0.730   0.465154
   profst2    1.035595   0.649208    1.595   0.110675
   profst3    0.038459   0.275635    0.140   0.889033
   profst4   -0.620169   0.325915   -1.903   0.057060  .
  suicbef1   -0.008845   0.240372   -0.037   0.970645
  suicdur1    0.502080   0.313848    1.600   0.109653
    suicat   -0.022661   0.234589   -0.097   0.923046
  history1    0.497925   0.306229    1.626   0.103953
    diagn2    0.462811   0.448628    1.032   0.302252
    diagn3    0.365734   0.437588    0.836   0.403271
    diagn4    0.870246   0.375367    2.318   0.020428  *
     coex1   -0.418387   0.220028   -1.902   0.057234  .
      body    0.073131   0.066383    1.102   0.270610
     diab1    0.097294   0.402701    0.242   0.809087
      khk1    1.255137   0.789490    1.590   0.111878
 psychobef1   0.009072   0.231080    0.039   0.968684
 psychodur1   1.520440   0.348468    4.363   1.28e-05  ***
   bblock1    1.042838   0.317595    3.284   0.001025  **
     meds1    0.045742   0.341068    0.134   0.893313
 changemed2   1.028870   0.344789    2.984   0.002845  **
 changemed3  -0.373037   0.834169   -0.447   0.654734
   hypoth1    0.001932   0.346053    0.006   0.995545
```

```
      uaw1    0.904442   0.444195    2.036   0.041736   *
    stress1  -0.382066   0.240567   -1.588   0.112243
   alcohol1   0.228576   0.445072    0.514   0.607551
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 328.72 on 261 degrees of freedom
Residual deviance: 225.51 on 231 degrees of freedom
AIC: 287.51

Number of Fisher Scoring iterations: 7
```

Όπως παρατηρούμε στον πίνακα αποτελεσμάτων της ανάλυσης probit, μικρές διαφορές που παρατηρούνται μεταξύ των δύο μοντέλων βρίσκονται στις εκτιμήσεις των συντελεστών $\beta_j$ και στην τιμή του κριτηρίου $AIC$. Όσον αφορά τις εκτιμήσεις $\hat{\beta}_j$, η διαφορά που εντοπίζεται ήταν αναμενόμενη, λόγω της διαφορετικής συνάρτησης σύνδεσης που χρησιμοποιείται στη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας. Υπενθυμίζουμε ότι η συνάρτηση σύνδεσης της μεθόδου probit κάνει χρήση της συνάρτησης κατανομής πιθανότητας της τυποποιημένης Κανονικής κατανομής, $\Phi(\bullet)$.

Εκτός από το κριτήριο AIC, το οποίο παρουσιάζεται απευθείας στα αποτελέσματα της μεθόδου probit, θα καλέσουμε την αντίστοιχη εντολή για το κριτήριο BIC, όπως κάναμε και στο μοντέλο logit. Έτσι, έχουμε:

```
> AIC(probit.model)
[1] 287.5141
```

```
> BIC(probit.model)
[1] 398.1328
```

Στη συνέχεια, θα προχωρήσουμε στην εκτίμηση της προβλεπτικής ικανότητας του προσαρμοσμένου μοντέλου probit, ακολουθώντας την ίδια διαδικασία με πριν. Αρχικά, λοιπόν, θα εφαρμόσουμε το προσαρμοσμένο μοντέλο probit στα δεδομένα *test*, με την εντολή *predict()*.

```
> fit.prob.probit <- predict(probit.model, test, type = 'response')
```

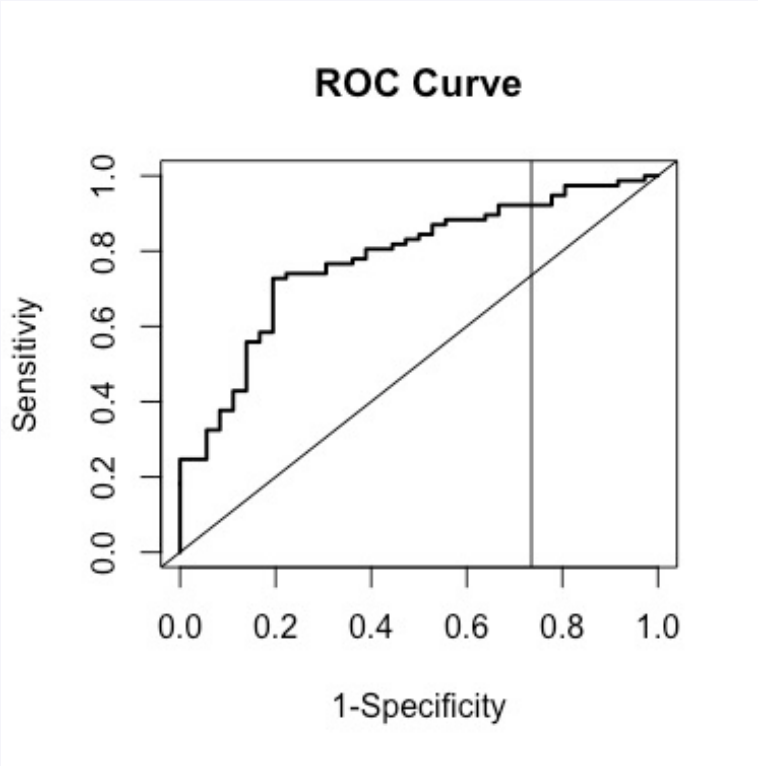
Θα προχωρήσουμε στην ανάλυση ROC, ώστε να εντοπίσουμε το βέλτιστο σημείο cut-off.

```
> analysis.probit <- roc(test$hosp, fit.prob.probit)
```

```
> e.p <- cbind(analysis.probit$thresholds, analysis.probit$sensitivities +
analysis.probit$specificities)
```

```
> opt.t.p <- subset(e.p, e.p[,2]==max(e.p[,2]))[,1]
```

Η σχηματική απεικόνιση της καμπύλης ROC, φαίνεται παρακάτω.




Τελικά, θα πραγματοποιήσουμε την ταξινόμηση των παρατηρήσεων, με βάση το σημείο cut-off, που υπολογίσαμε μέσω της μεθόδου ROC, για το μοντέλο probit και, θα υπολογίσουμε την ακρίβεια πρόβλεψης του μοντέλου probit.

```
> opt.t.p
[1] 0.7352399
```

```
> fit.results.probit <- ifelse(fit.prob.probit > 0.7352399, 1, 0)
```

```
> missClassError.probit <- mean(fit.results.probit != test$hosp)
```

```
> print(1 - missClassError.probit)
[1] 0.7522124
```

Παρατηρούμε ότι το ποσοστό ακρίβειας στην πρόβλεψη, μεταξύ τον δύο μεθόδων, εκτιμάται ότι είναι ίσο, δηλαδή $\simeq 75\%$. Τέλος, ο πίνακας ταξινόμησης των αληθών και ψευδών τοποθετήσεων, είναι:

```
> table(fit.prob.probit > 0.7352399)
 FALSE  TRUE
    50    63
```

Επίσης, η *p-τιμή* του στατιστικού ελέγχου $\chi^2$ του υποψήφιου μοντέλου παλινδρόμησης είναι:

```
> 1 - pchisq(probit.model$deviance, probit.model$df.residual)
[1] 0.5894735
```

Ομοίως με πριν, το υποψήφιο μοντέλο της μεθόδου probit είναι ένα καλό εναλλακτικό μοντέλο, έναντι του κορεσμένου.

### 4.5.3 Συμπεράσματα και επιλογή μοντέλου

Στις δύο παραπάνω ενότητες, προσαρμόσαμε δύο μοντέλα παλινδρόμησης, κάνοντας χρήση δύο διαφορετικών συναρτήσεων σύνδεσης. Λόγω της φύσης των διαθέσιμων δεδομένων, διαλέξαμε να πραγματοποιήσουμε παλινδρόμηση logit και probit, με σκοπό να συγκρίνουμε τα αποτελέσματα και να επιλέξουμε το καταλληλότερο μοντέλο.

Για την αξιολόγηση του μοντέλου, χωρίσαμε το δείγμα σε δύο μέρη, *train* και *test*, σε 70% και 30% των παρατηρήσεων, αντίστοιχα. Τα μοντέλα προσαρμόστηκαν πάνω στο δείγμα *train*. Υπολογίσαμε τους δείκτες AIC και BIC για το κάθε μοντέλο και πραγματοποιήσαμε ανάλυση ROC για την εύρεση του βέλτιστου σημείου cut-off. Στη συνέχεια, χρησιμοποιώντας το σημείο αυτό, εφαρμόσαμε το προσαρμοσμένο μοντέλο στα δεδομένα *test* και εκτιμήσαμε κάποια στατιστικά μέτρα που θα μας φανούν χρήσιμα για την επιλογή του μοντέλου. Ο λόγος που χρησιμοποιούμε το δείγμα *test* είναι ώστε να αξιολογήσουμε την προβλεπτική ικανότητα του μοντέλου σε δεδομένα, τα οποία δε "γνωρίζει".

Έπειτα από αυτή τη διαδικασία, είμαστε σε θέση να επιλέξουμε το μοντέλο που θεωρούμε καταλληλότερο για τα δεδομένα μας. Η αλήθεια είναι, ότι τα δύο μοντέλα δε διαφέρουν πολύ. Ο δείκτης AIC παρουσιάζει μόλις μία μονάδα διαφορά ($AIC_{log} = 288.5$, $AIC_{pr} = 287.5$) και το ίδιο συμβαίνει στο δείκτη BIC ($BIC_{log} = 399.1$, $BIC_{pr} = 398.1$). Συνεπώς, από αυτούς τους δύο δείκτες, δε μας δίνονται επαρκείς ενδείξεις για

να απορρίψουμε οποιαδήποτε από τις δύο μεθόδους. Το επόμενο μέτρο που θα συμβουλευτούμε είναι το ποσοστό ακρίβειας της ταξινόμησης. Ομοίως, το ποσοστό αυτό εκτιμάται το ίδιο και στα δύο μοντέλα ($\simeq 75\%$). Επιπλέον, οι πίνακες ταξινόμησης δεν παρουσιάζουν διαφορές και οι στατιστικά σημαντικές μεταβλητές παρουσιάζονται να είναι ίδιες με μικρές διαφορές στις εκτιμήσεις των συντελεστών και συνεπώς και στις *p-τιμές*.

Επομένως, η επιλογή ανάμεσα στα δύο μοντέλα είναι ελεύθερη, καθώς δε φαίνεται να επηρεάζονται, σημαντικά, τα αποτελέσματα από τη συνάρτηση σύνδεσης. Για λόγους ευκολίας στην ερμηνεία, θα επιλέξουμε το μοντέλο logit. Έτσι, θα προσαρμόσουμε ξανά το μοντέλο παλινδρόμησης, με συνάρτηση σύνδεσης τη "logit", στο σύνολο των δεδομένων και θα καταλήξουμε στα ανάλογα συμπεράσματα.

## 4.6 Προσαρμογή του στατιστικού μοντέλου παλινδρόμησης

Για να προβούμε σε συμπεράσματα πλέον, για την έρευνα, θα προσαρμόσουμε το μοντέλο της λογιστικής παλινδρόμησης στο σύνολο των δεδομένων. Να υπενθυμίσουμε ότι, σκοπός της έρευνας είναι η εύρεση παραγόντων που επηρεάζουν το χρόνο νοσηλείας των ασθενών με κατάθλιψη.

```
> final.model <- glm(hosp ~ ., family = binomial(link = 'logit'), data =
thesis.data)
```

```
> summary(final.model)

Call:
glm(formula = hosp ~ ., family = binomial(link = "logit"), data = thesis.data)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
 -2.3318  -0.7269   0.3565   0.7333   2.6296

Coefficients:
              Estimate  Std. Error  z value   Pr(>|z|)
(Intercept)   -4.02755    1.09231   -3.687   0.000227  ***
       age     0.02447    0.01390    1.761   0.078233  .
       adm    -0.11138    0.11066   -1.007   0.314162
   gender1    -0.06655    0.31048   -0.214   0.830268
  marital2    -0.01362    0.42596   -0.032   0.974501
  marital3     0.06967    0.34208    0.204   0.838606
  marital4    -0.32279    0.57957   -0.557   0.577559
    profst2    1.38233    0.91657    1.508   0.131515
    profst3    0.30144    0.37457    0.805   0.420962
    profst4   -0.68888    0.43100   -1.598   0.109972
```

```
      suicbef1  -0.25624  0.32377  -0.791  0.428685
      suicdur1   1.00710  0.43574   2.311  0.020820  *
        suicat  -0.08934  0.33151  -0.269  0.787553
      history1   0.42173  0.39977   1.055  0.291452
        diagn2   0.79427  0.60952   1.303  0.192539
        diagn3   0.31298  0.62120   0.504  0.614385
        diagn4   1.00889  0.52794   1.911  0.056005  .
         coex1  -0.33002  0.30631  -1.077  0.281300
          body   0.08085  0.09186   0.880  0.378772
         diab1  -0.08703  0.56222  -0.155  0.876986
          khk1   1.52268  0.97543   1.561  0.118517
   psychobef1    0.06769  0.32023   0.211  0.832583
   psychodur1    2.68119  0.51109   5.246   1.55e-07  ***
      bblock1    1.86777  0.48091   3.884  0.000103  ***
        meds1    0.14126  0.42910   0.329  0.741999
  changemed2     1.48713  0.48615   3.059  0.002221  **
  changemed3    -1.09957  1.36952  -0.803  0.422042
      hypoth1    0.18174  0.48134   0.378  0.705746
         uaw1    1.83525  0.70959   2.586  0.009699  **
      stress1   -0.45509  0.32075  -1.419  0.155946
     alcohol1   -0.21767  0.54579  -0.399  0.690031
—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 470.15 on 374 degrees of freedom
Residual deviance: 339.34 on 344 degrees of freedom
AIC: 401.34

Number of Fisher Scoring iterations: 6
```

Επιπλέον, θα υπολογίσουμε τις προσαρμοσμένες πιθανότητες με την εντολή *predict()*,

```
> fitted.probabilities <- predict(final.model, thesis.data, type = ’response’)
```

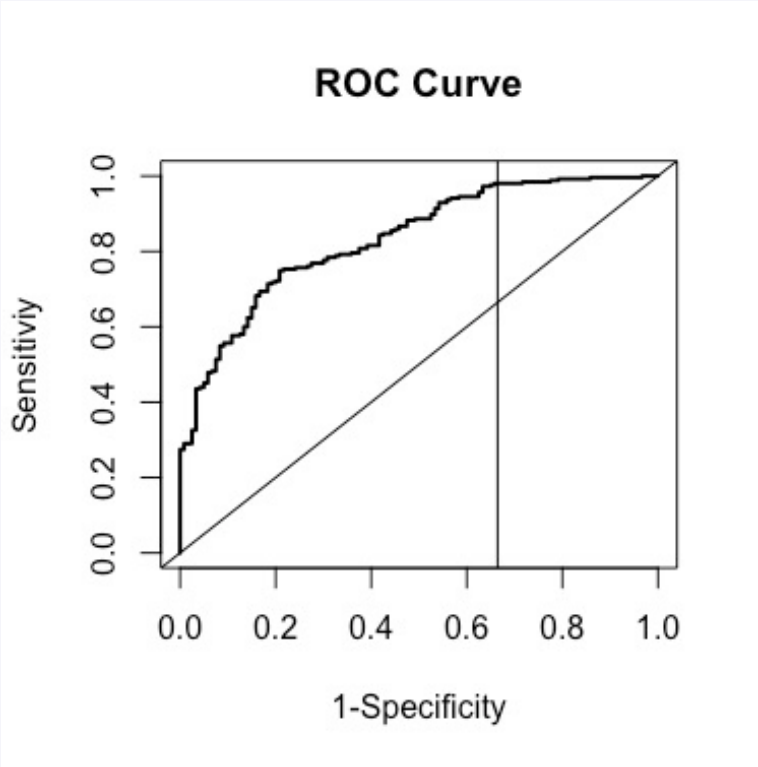
και θα βρούμε το βέλτιστο σημείο cut-off, για την ταξινόμηση των παρατηρήσεων στις δύο κατηγορίες της μεταβλητής απόκρισης, με την ανάλυση ROC.

```
> analysis <- roc(thesis.data$hosp, fitted.probabilities)
```

```
> e <- cbind(analysis$thresholds, analysis$sensitivities + analysis$specificities)
```

```
> opt_t <- subset(e, e[,2]==max(e[,2]))[,1]
```

Επίσης, η καμπύλη ROC με γραφικό τρόπο, φαίνεται παρακάτω.




Εάν, τώρα, χρησιμοποιήσουμε το σημείο $t$ που εκτίμησε η μέθοδος ROC, μπορούμε να βρούμε το σφάλμα ταξινόμησης, κάθώς και την ακρίβεια πρόβλεψης του μοντέλου.

```
> fitted.results <- ifelse(fitted.probabilities > 0.6644731, 1, 0)
```

```
missClassError <- mean(fitted.results != thesis.data$hosp)
```

```
> print(1 - missClassError)
[1] 0.7626667
```

Όπως βλέπουμε, το μοντέλο μπορεί να προβλέψει κατα $\simeq$ 76%, ποσοστό που είναι αρκετά ικανοποιητικό. Τέλος, θα παρουσιάσουμε τον πίνακα ταξινόμησης και θα

υπολογίσουμε την p-τιμή του ελέγχου $\chi^2$.

```
> table(fitted.probabilities > 0.6644731)

 FALSE  TRUE
   159   216
```

```
> 1 - pchisq(final.model$deviance, final.model$df.residual)

[1] 0.5608274
```

Τελικά, παρατηρούμε ότι η ταξινόμηση των σωστών και λανθασμένων παρατηρήσεων με βάση την πιθανότητα δεν είναι και τόσο ικανοποιητική. Παρ'όλα αυτά, με βάση την *p-τιμή* του ελέγχου $\chi^2$ συμπεραίνουμε ότι το υποψήφιο μοντέλο, είναι ένα καλό εναλλακτικό μοντέλο, έναντι του κορεσμένου.

### 4.6.1 Ερμηνεία των αποτελεσμάτων

Πριν περάσουμε στην ερμηνεία των αποτελεσμάτων που λάβαμε από το μοντέλο της λογιστικής παλινδρόμησης που προσαρμόσαμε παραπάνω, θα υπενθυμίσουμε τις έννοιες που θα χρησιμοποιήσουμε. Έτσι, με τον όρο *odds*, εννοούμε το λόγο της πιθανότητας επιτυχίας ($Y = 1$), προς τη πιθανότητα αποτυχίας ($Y = 0$). Ο λόγος αυτός, μας παρέχει τον αριθμό που χρειαζόμαστε για να πολλαπλασιάσουμε τη πιθανότητα αποτυχίας, με σκοπό να υπολογίσουμε την πιθανότητα επιτυχίας. Επιπλέον, ο λόγος των συμπληρωματικών πιθανοτήτων (*odds ratio*) αποτελεί το λόγο δύο σχετικών πιθανοτήτων (*odds*), δύο διαφορετικών κατηγοριών. Παρέχει, ουσιαστικά, τη σχετική αλλαγή των σχετικών πιθανοτήτων, κάτω από δύο διαφορετικές καταστάσεις. Να θυμίσουμε επίσης ότι, $exp\{\hat{\beta}_j\} =$ *odds ratio*. Επιπλέον, σημειώνουμε ότι όλες οι ερμηνείες γίνονται με την προϋπόθεση ότι όλοι οι υπόλοιποι παράγοντες παραμένουν σταθεροί.

Παρατηρώντας, τώρα, τα αποτελέσματα που πήραμε από το λογιστικό μοντέλο που προσαρμόσαμε, θα προβούμε στα αντίστοιχα συμπεράσματα. Αρχικά, θα αναφέρουμε ότι η μεταβλητή απόκρισης *hosp*, αποτελεί μία δυαδική μεταβλητή, η οποία ακολουθεί την κατανομή Bernoulli με πιθανότητα επιτυχίας $\pi$, $y \sim B(p)$. Έπειτα από την προσαρμογή του λογιστικού μοντέλου, ως στατιστικά σημαντικές κρίνονται οι εξής μεταβλητές: *suicdur, psychodur, bblock, changemed* και *uaw*. Στατιστικά σημαντικές ορίζονται οι μεταβλητές που έχουν *p-τιμή* μικρότερη του $0.05$. Παρ'όλα αυτά, δεν μπορούμε να παραβλέψουμε τις μεταβλητές *age* και *diagn*, οι οποίες παρουσιάζονται να επηρεάζουν, οριακά, το μοντέλο.

Πιο αναλυτικά, για τη μεταβλητή της αυτοκτονικότητας, κατά τη διάρκεια της νοσηλείας (*suicdur*), έχουμε υπολογίσει ότι: $exp\{1.00710\} = 2.73765$. Αυτό το αποτέλεσμα υποδηλώνει ότι, η σχετική πιθανότητα του χρόνου νοσηλείας μεγαλύτερου των $4$ εβδομάδων, είναι $2.74$ φορές μεγαλύτερη στους ασθενείς που παρουσιάζουν αυτοκτονικές τάσεις, σε σχέση με αυτούς που δεν έχουν.

Επιπλέον, στατιστικά σημαντική εμφανίζεται να είναι η μεταβλητή της ψυχοθεραπείας, κατά τη διάρκεια της νοσηλειάς (*psychodur*). Ο λόγος των συμπληρωματικών πιθανοτήτων της συγκεκριμένης μεταβλητής εκτιμάται ίσος με $exp\{2.68119\} =$

14.60246. Με άλλα λόγια, η σχετική πιθανότητα ο χρόνος νοσηλείας να είναι περισσότερος από 4 εβδομάδες, είναι 14.6 φορές μεγαλύτερη σε έναν ασθενή που υποβάλλεται σε ψυχοθεραπεία, συγκριτικά με κάποιον που δεν ακολουθεί αυτή τη διαδικασία.

Η χρήση b-blockers, φαίνεται να επηρεάζει σημαντικά τα αποτελέσματα του μοντέλου. Ο λόγος των συμπληρωματικών πιθανοτήτων εκτιμάται ίσος με $exp\{1.86777\} = 6.473844$. Πιο συγκεκριμένα, η σχετική πιθανότητα του χρόνου νοσηλείας μεγαλύτερου των 4 εβδομάδων είναι 6.5 φορές υψηλότερη για έναν ασθενή που κάνει χρήση b-blockers, σε σχέση με κάποιον ασθενή που δεν υποβάλλεται σε αυτή τη θεραπεία. Να σημειώσουμε εδώ, ότι, τα b-blockers είναι μία τάξη φαρμάκων που χρησιμοποιείται ιδιαιτέρως για τη διαχείρηση καρδιακών αρρυθμιών και προστατεύουν τον ασθενή από ένα δεύτερο καρδιακό επεισόδιο.

Η επόμενη μεταβλητή, αφορά την αλλαγή φαρμακευτικής αγωγής (*changemed*). Η μεταβλητή αυτή, χωρίζεται σε τρεις κατηγορίες. Έχουμε την πρώτη κατηγορία, η οποία είναι και η κατηγορία αναφοράς, που αφορά τους ασθενείς που δεν πραγματοποιήθηκε κάποια αλλαγή στα φάρμακά τους. Η δεύτερη και τρίτη κατηγορία αναφέρεται στην αλλαγή φαρμακευτικής αγωγής στις περιπτώσεις, αναποτελεσματικότητας του φαρμάκου, καθώς και της εμφάνισης παρενεργειών, αντίστοιχα. Ως στατιστικά σημαντική παρουσιάζεται η δεύτερη κατηγορία σε σχέση με την πρώτη. Έτσι, $exp\{1.48713\} = 4.424379$, που σημαίνει ότι η σχετική πιθανότητα νοσηλείας για περισσότερες από 4 εβδομάδες, ασθενών, που η φαρμακευτική τους αγωγή παρουσίασε αναποτελεσματικότητα, είναι κατα 4.42 φορές υψηλότερη, από τη σχετική πιθανότητα αυτών που δεν τροποιήθηκαν τα φάρμακά τους.

Τέλος, στατιστικά σημαντική είναι και η μεταβλητή των παρενεργειών από την ψυχιατρική φαρμακευτική αγωγή που παρέχεται στους ασθενείς (*uaw*). Πιο συγκεκριμένα, η σχετική πιθανότητα νοσηλείας για περισσότερες από 4 εβδομάδες ενός ασθενή που εμφανίζει ανεπιθύμητες ενέργειες από την ψυχιατρική φαρμακευτική αγωγή, είναι 6.37 ($exp\{1.83525\} = 6.266701$) φορές υψηλότερη, σε σχέση με έναν ασθενή που δεν αντιμετωπίζει προβλήματα με τα φάρμακά του.

Επιπλέον, όπως προαναφέραμε, δε θα έπρεπε να παραβλέψουμε τις μεταβλητές που είναι οριακά, στατιστικά σημαντικές (*age, diagn*). Έτσι, με βάση τα αποτελέσματα, όσον αφορά τη μεταβλητή της ηλικίας (*age*), έχουμε τα εξής αριθμητικά αποτελέσματα: $exp\{0.02447\} = 1.024772$. Πρακτικά, αυτό σημαίνει ότι για ένα χρόνο αύξησης της ηλικίας ενός ασθενή, αναμένεται περίπου 2.5% αύξηση στη σχετική πιθανότητα να νοσηλευτεί για περισσότερο των 4 εβδομάδων. Τέλος, για τη μεταβλητή της διάγνωσης (*diagn*), βλέπουμε ότι η σχετική πιθανότητα νοσηλείας μεγαλύτερη των 4 εβδομάδων, ενός ασθενή που διαγνώσθηκε με βαριά υποτροπιάζουσα καταθλιπτική διαταραχή (F33.2), είναι $exp\{1.00889\} = 2.742555$ φορές υψηλότερη, συγκριτικά με κάποιον που διαγνώσθηκε με μέτριας βαρύτητας καταθλιπτικό επεισόδιο (F32.1).

Για τις μη-στατιστικά σημαντικές μεταβλητές, δεν μπορούμε να απορρίψουμε την υπόθεση ότι, οι σχετικές πιθανότητες διαφοροποιούνται, συνεπώς για λόγους οικονομίας δεν προβαίνουμε σε πλήρη ερμηνεία των αποτελεσμάτων. Για το λόγο αυτό, περιοριστήκαμε στην ερμηνεία μόνο των στατιστικά σημαντικών μεταβλητών.

## 4.7 Εφαρμογή της μεθόδου lasso στη λογιστική παλινδρόμηση

Θα περάσουμε, τώρα, στην εφαρμογή μεθόδων επιλογής μεταβλητών με ποινή και πιο συγκεκριμένα, στη μέθοδο lasso. Το πακέτο *glmnet* είναι σχεδιασμένο για την πραγματοποίηση ποινικοποιημένης εκτίμησης στα γενικευμένα γραμμικά μοντέλα. Το συγκεκριμένο πακέτο υποστηρίζει τις μεθόδους lasso και elastic-net για μοντέλα γραμμικής παλινδρόμησης, λογιστικής παλινδρόμησης και πολυωνυμικής παλινδρόμησης. Συνοπτικά υπενθυμίζεται ότι, η μέθοδος lasso συρρικνώνει τις εκτιμήσεις των συντελεστών προς το μηδέν σε σχέση με τις ε.ε.τ.. Στόχος αυτής της συρρίκνωσης είναι, κατα κύριο λόγο, η επίλυση του προβλήματος της πολυσυγγραμμικότητας των συμμεταβλητών. Το ποσό της συρρίκνωσης που υφίστανται οι συντελεστές ρυθμίζεται μέσω της παραμέτρου συντονισμού $\lambda$.

Πλέον, υπάρχουν πολλά στατιστικά πακέτα που βοηθούν στην αναπαραγωγή της μεθόδου lasso. Επιλέγουμε το πακέτο *glmnet* καθώς είναι πιο φιλικό προς το χρήστη και περιέχει μεγαλύτερη γκάμα συναρτήσεων. Προτείνει άμεσα την ελάχιστη τιμή $\lambda$, καθώς και την τιμή του $\lambda$ που ισοδυναμεί με το μέσο τετραγωνικό σφάλμα της μεθόδου *cvl*. Θα ξεκινήσουμε φορτώνοντας την βιβλιοθήκη στην R.

```
> library(glmnet)
```

Η ιδιαιτερότητα της μεθόδου lasso είναι ότι απαιτεί όλοι οι συντελεστές της να είναι αριθμητικοί παράγοντες. Για το λόγο αυτό, θα κάνουμε χρήση των εντολών *sapply()* και *lapply()*, ώστε να μετατρέψουμε τις παρατηρήσεις κατάλληλα.

```
> indx <- sapply(thesis.data, is.factor)
```

```
> thesis.data[indx] <- lapply(thesis.data[indx], function(x)
as.numeric(as.character(x)))
```

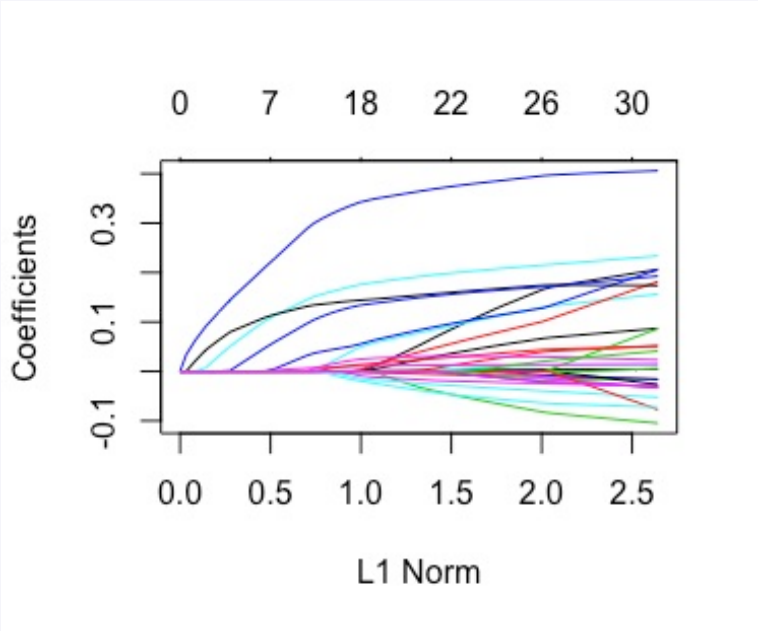
Έτσι, στην πρώτη εντολή, εντοπίζουμε τα σημεία του δείγματος *train*, στα οποία οι παρατηρήσεις είναι καταχωρημένες ως παράγοντες. Έπειτα, εφαρμόζουμε σε αυτά τη συνάρτηση *as.numeric()* και τα μετατράπουμε σε αριθμητικά δεδομένα. Συνεχίζουμε δημιουργώντας έναν πίνακα σχεδιασμού, μέσω της εντολής *model.matrix()*, με παράμετρο τις εκτιμημένες τιμές του μοντέλου της λογιστικής παλινδρόμησης, που προσαρμόσαμε στην προηγούμενη ενότητα. Για να μπορέσουμε να κάνουμε χρήση αυτού του πίνακα στη μέθοδο lasso, θα πρέπει να αφαιρέσουμε τη στήλη που αφορά το σταθερό όρο $\beta_0$. Γι'αυτό το λόγο, χρησιμοποιούμε και την κωδικοποίηση $[,-1]$.

```
> X <- model.matrix(final.model)[,-1]
```

Επόμενο βήμα, είναι η προσαρμογή του λογιστικού μοντέλου μέσω της ποινικοποιημένης μεγιστοποιημένης συνάρτησης πιθανοφάνειας. Η διαδικασία αυτή υπολογίζεται για την ποινή lasso σε ένα πλέγμα τιμών για την παράμετρο συντονισμού $\lambda$. Ως παράμετροι στην εντολή *glmnet()* χρησιμοποιούνται, ο πίνακας σχεδιασμού $X$, που δημιουργήσαμε παραπάνω και οι παρατηρήσεις της μεταβλητής απόκρισης, *hosp*, στο σύνολο του δείγματος.

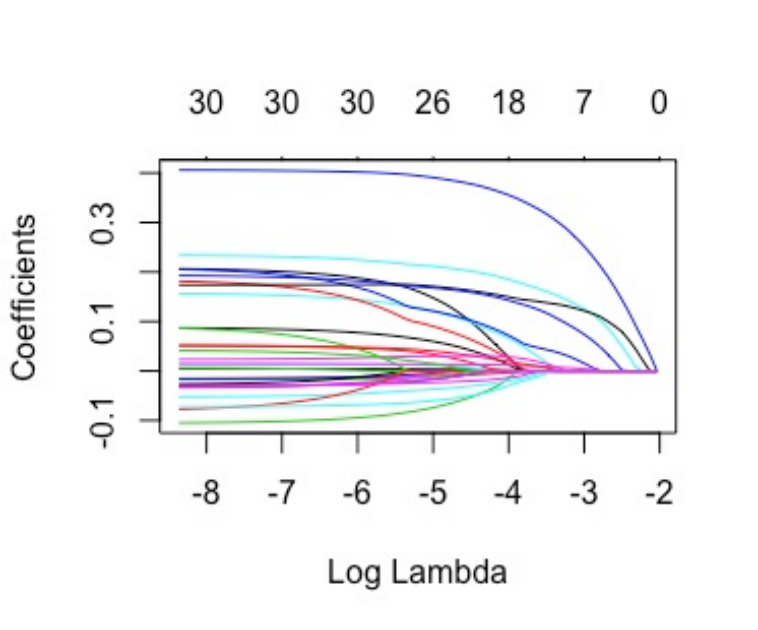
```
> lasso <- glmnet(X, thesis.data$hosp)
```

```
> plot(lasso, label = T)
```




Στο γράφημα που παρουσιάζεται παραπάνω, παρατηρούμε τη διαδρομή που ακολουθούν οι συντελεστές στην τεχνική lasso έως, ότου μηδενιστούν. Οι άξονες, πάνω στους οποίους έχει δημιουργηθεί αυτό το διάγραμμα αποτελούνται από τις τιμές των εκτιμημένων συντελεστών $\beta_j$ συναρτήσει της L1-norm των συντελεστών. Θα δημιουργήσουμε το αντίστοιχο γράφημα με τον $x-$άξονα να λαμβάνει τις τιμές της ακολουθίας *log-lambda*, όπου *lambda* είναι οι τιμές που υπολογίστηκαν και αποθηκεύτηκαν στην παράμετρο *lasso*, παραπάνω.

```
> plot(lasso, xvar = 'lambda', label = T)
```




Σκοπός είναι να βρεθεί η βέλτιστη τιμή του $\lambda$ έτσι, ώστε να παραμείνουν στο μοντέλο οι πιο χρήσιμοι συντελεστές. Οι υπόλοιποι, με βάση τη μέθοδο ποινικοποίησης lasso, θα συρρικνωθούν μέχρι να πάρουν την τιμή μηδέν και να αποκλειστούν από το μοντέλο. Θα εφαρμόσουμε τώρα τη μέθοδο *k-fold cross-validation*, ώστε να γίνει μια προσπάθεια να ευρεθούν οι βέλτιστες τιμές της παραμέτρου $\lambda$, όπως αναφέρθηκε στη θεωρία (βλέπε εν.3.5.6).

```
> lasso.cvl <- cv.glmnet(X, thesis.data$hosp)
```

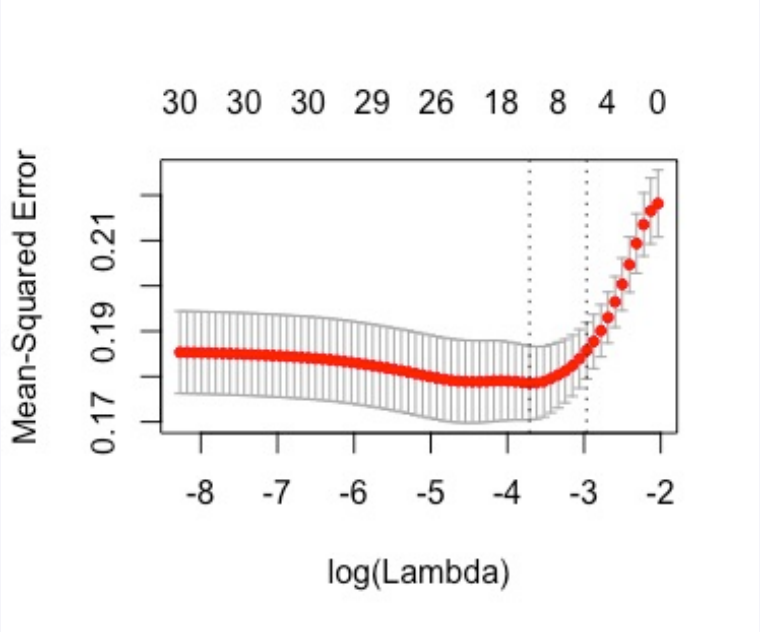
Από τις ποσότητες που υπολογίστηκαν μέσω της συνάρτησης *cv.glmnet()*, θέλουμε να λάβουμε τις τιμές για τη μικρότερη τιμή και τη μέγιστη τιμή του $\lambda$.

```
> lasso.cvl$lambda.min
[1] 0.02444168
```

```
> lasso.cvl$lambda.1se
[1] 0.05144739
```

Καλώντας την πρώτη εντολή, ζητάμε από την R να επιστρέψει την τιμή του $\lambda$, που αντιστοιχεί στο μικρότερο *mean cross-validated error*. Η δεύτερη εντολή, δίνει την μέγιστη τιμή του $\lambda$ έτσι, ώστε το σφάλμα να ακολουθεί τον κανόνα "one-standar-error". Ο κύριος λόγος του κανόνα αυτού είναι να επιλεχθεί το μικρότερο $\lambda$, του οποίου η ακρίβεια θα είναι συγκρίσιμη με το βέλτιστο.

```
> plot(lasso.cvl)
```




Στο παραπάνω σχήμα, η πρώτη κάθετη, διακεκομμένη γραμμή αντιστοιχεί στην τιμή της παραμέτρου *lambda.min* και η δεύτερη στην τιμή της παραμέτρου *lambda.1se*. Για την μικρότερη τιμή της παραμέτρου $\lambda$, οι συντελεστές του προσαρμοσμένου λογιστικού μοντέλου είναι αυτοί που παρουσιάζονται στον επόμενο πίνακα.

```
> coef(lasso.cvl, s = 'lambda.min')
31 x 1 sparse Matrix of class "dgCMatrix"

                    1
(Intercept)  0.177993783
age          0.001037358
adm                    .
gender1                .
marital2               .
marital3               .
marital4               .
profst2                .
profst3                .
profst4                .
suicbef1               .
suicdur1     0.039711464
suicat      -0.003928716
```

```
      history1                .
        diagn2                .
        diagn3                .
        diagn4    0.050618407
         coex1   -0.010290874
          body    0.010020578
         diab1                .
          khk1    0.011861139
   psychobef1                 .
   psychodur1    0.335472526
       bblock1    0.171429766
         meds1    0.021912877
   changemed2     0.142411350
   changemed3                 .
       hypoth1                .
          uaw1    0.129644460
       stress1   -0.015634800
      alcohol1                .
```

Παρατηρούμε ότι οι συντελεστές που δεν μηδενίστηκαν τελικά, είναι περισσότεροι από εκείνους που παρουσιάστηκαν σαν στατιστικά σημαντικοί όταν προσαρμόσαμε το μοντέλο της λογιστικής παλινδρόμησης. Επιπλέον, τα αποτελέσματα που αντιστοιχούν στην τιμή του *lambda.1se* υπολογίζονται εάν, στην παράμετρο "s", της εντολής *coef()* βάλουμε την αντίστοιχη τιμή του $\lambda$ που υπολογίσαμε παραπάνω.

```
> coef(lasso.cvl, s = 'lambda.1se')
31 x 1 sparse Matrix of class "dgCMatrix"
                          1
 (Intercept)   0.3727669322
         age   0.0002597777
         adm                .
     gender1                .
    marital2                .
    marital3                .
    marital4                .
      profst2               .
      profst3               .
      profst4               .
```

```
      suicbef1                .
      suicdur1                .
        suicat                .
      history1                .
        diagn2                .
        diagn3                .
        diagn4   0.0129597279
         coex1                .
          body   0.0035736934
         diab1                .
          khk1                .
    psychobef1                .
    psychodur1   0.2460178869
       bblock1   0.1221126438
         meds1                .
   changemed2   0.1199761813
   changemed3                .
      hypoth1                .
          uaw1   0.0703194989
       stress1                .
     alcohol1                 .
```

Οι τιμές αυτές των συντελεστών παρουσιάζουν την επίδραση των μεταβλητών στο μοντέλο, χωρίς αυτές να έχουν κανονικοποιηθεί. Ενδιαφέρον παρουσιάζει η επιλογή των μεταβλητών που έγινε για τις δύο αυτές τιμές της παραμέτρου $\lambda$. Είναι φανερό πλέον, ότι η επίδραση του $\lambda$ στις εκτιμήσεις των συντελεστών των μεταβλητών είναι καθοριστική. Τα αποτελέσματα που φαίνονται στον τελευταίο πίνακα, δε διαφέρουν σημαντικά με αυτά του μοντέλου της λογιστικής παλινδρόμησης που προσαρμόσαμε παραπάνω.

Συμπερασματικά, το προτεινόμενο μοντέλο, έπειτα από τη διαδικασία της μεθόδου lasso, περιλαμβάνει τις μεταβλητές: *age, diagn, body, psychodur, bblock, changemed* και *uaw*.

# ΚΕΦΑΛΑΙΟ 5

# ΣΥΜΠΕΡΑΣΜΑΤΑ

Enough research will tend to support your conclusions.

*Arthur Bloch*

Στην παρούσα μελέτη, ο σκοπός ήταν να εφαρμοστούν συγκεκριμένες μέθοδοι ανάλυσης, πάνω σε αληθινά ψυχιατρικά δεδομένα. Η εφαρμογή των τεχνικών αυτών έγινε με τη χρήση του στατιστικού πακέτου R. Τα δεδομένα, λόγω της πραγματικής φύσης τους, παρουσίαζαν προβλήματα όπως πολυσυγγραμμικότητα ή ελλειπούσες τιμές. Με τα κατάλληλα εργαλεία, καταφέραμε, εν μέρει, να αντιμετωπίσουμε τα εμπόδια και να εφαρμόσουμε τα μοντέλα που επιθυμούσαμε εξ'αρχής.

Αρχικά, έγινε μία εισαγωγή στο ερευνητικό ερώτημα που κληθήκαμε να απαντήσουμε. Το δείγμα αφορούσε ασθενείς που είχαν διαγνωσθεί με κλινική κατάθλιψη, το έτος 2012. Η λύση αυτή, θα προερχόταν από την προσαρμογή των στατιστικών μοντέλων της λογιστικής παλινδρόμησης και της probit ανάλυσης. Έτσι, έγινε μία εισαγωγή στη βασική θεωρία των γενικευμένων γραμμικών μοντέλων. Ιδιαίτερη έμφαση δόθηκε στα μοντέλα της λογιστικής παλινδρόμησης και της μεθόδου probit. Στη συνέχεια, παρουσιάστηκαν οι ποινικοποιημένες μέθοδοι επιλογής μεταβλητών και πιο συγκεκριμένα επικεντρωθήκαμε στη lasso. Τελικά, έγινε εφαρμογή όλων αυτών των τεχνικών στα διαθέσιμα δεδομένα.

Στην εφαρμογή που εκτελέσαμε στο τελευταίο κεφάλαιο, είχαμε ένα μικρό σχετικά δείγμα των $n = 381$ παρατηρήσεων. Οι επεξηγηματικές μεταβλητές που χρησιμοποιήθηκαν αποτελούνταν από κατηγορικής και ποσοτικής φύσης παρατηρήσεις. Τελικά, έπειτα από εκτεταμένο έλεγχο και καθαρισμό των δεδομένων, έγινε χρήση $p = 24$ συμμεταβλητών. Με το διαθέσιμο δείγμα, καταφέραμε να πραγματοποιήσουμε τις επιθυμητές αναλύσεις και να διεξάγουμε τα αντίστοιχα συμπεράσματα.

# ΠΑΡΑΡΤΗΜΑ

Στο κομμάτι αυτό της διπλωματικής εργασίας, θα παρουσιαστούν οι κώδικες που χρησιμοποιήθηκαν στην R για κάποια γραφήματα. Επιπλέον, για τα σχήματα που δε δημιουργήθηκαν από το μηδέν, λόγω έλλειψης γνώσεων ή εργαλείων, θα αναφερθούν οι αντίστοιχες πηγές.

## Κεφάλαιο 2

Για τη δημιουργία της καμπύλης ROC χρησιμοποιήθηκαν έτοιμα δεδομένα που συμπεριλαμβανόντουσαν στο πακέτο *ROCR*, το οποίο εγκαταστάθηκε ώστε να εφαρμοστεί η μέθοδος. Επιπλέον, στα σχήματα της προσαρμογής του γραμμικού μοντέλου και των μοντέλων logit και probit έγινε χρήση του *data frame*, με όνομα *mtcars* που προϋπάρχει στο στατιστικό πακέτο της R.

Σχήμα 2.1

```
> p <- (1:999)/1000

> logit <- log(p/(1-p))

> cloglog = log(-log(1-p))

> probit = qnorm(p)

> plot(p, logit, type = 'l', xlab = 'Probability', ylab = 'Link Function', pch = 1, col = 'darkslateblue', main = 'The Logit, Probit and C-Log-Log Links')

> lines(p, probit, col='mediumvioletred', lty = 2)

> lines(p, cloglog, col = 'aquamarine3', lty = 4)

> abline(v = 0.5, lty = 3)

> legend('bottomright', legend = c('logit', 'probit', 'c-log-log'), lty = c(1, 2, 4), col = c('darkslateblue', 'mediumvioletred', 'aquamarine3'), bty = 'n')
```

Σχήμα 2.2

```
> library(ROCR)

> data(ROCR.simple)

> pred <- prediction(ROCR.simple predictions, ROCR.simple labels) > perf1 <- performance(pred, ”sens”)

> perf2 <- performance(pred, ”spec”)

> plot(perf1, col = ’violetred3’, main = ’Sensitivity/Specificity vs Cutoffs’, xlab = ’Probability of cutoff’, ylab = ’Sensitivity/Specificity’, lty = 1)

> par(new=TRUE)

> plot(perf2, col = ’seagreen4’, xlab = ”, ylab = ”, lty = 1)

> abline(v = 0.5, lty = 3)

> legend(x = 0.65, y = 0.9, legend = c(’Sensitivity’, ’Specificity’), lty = c(1, 1), col = c(’violetred3’, ’seagreen4’), bty = ’n’)
```

Σχήμα 2.3

```
> library(AUC)

> ROCR.simple$labels <- as.factor(ROCR.simple$labels)

> plot(roc(ROCR.simple$predictions,ROCR.simple$labels), main = ’Sensitivity vs 1-Specificity’)
```

Σχήμα 2.4

```
> lmfit = lm(vs hp, data = mtcars)

> plot(vs hp, data = mtcars, main = ’Linear Fit’, xlab = ’X’, ylab = ’Y’)

> abline(lmfit, col = ’firebrick1’)

> legend(’topright’, legend = ’linear’, lty = 1, col = ’firebrick1’, bty = ’n’)
```

Σχήμα 2.5

```
> fit1 = glm(vs   hp, data=mtcars, family=binomial(logit))

> fit2 = glm(vs   hp, data=mtcars, family=binomial(probit))

> newdat <- data.frame(hp=seq(min(mtcars$hp), max(mtcars$hp),len=100))


> newdat$vs1 = predict(fit1, newdata=newdat, type="response")

> newdat$vs2 = predict(fit2, newdata=newdat, type="response")


> plot(vs hp, data=mtcars, main = "Simulation of Logit/Probit Fit", xlab = 'X',
ylab ='Y')

> lines(vs1   hp, newdat, col="darkseagreen4", lwd = 2)

> lines(vs2   hp, newdat, col="violetred", lwd = 2, lty = 4)

> legend('topright', legend = c('logit', 'probit'), lty = c(1, 4), col =
c("darkseagreen4", "violetred"), bty = 'n')
```

# Κεφάλαιο 3

- Σχήμα 3.1: το γράφημα έχει δημιουργηθεί για τη δημοσίευση "Regression shrinkage and selection via the lasso", του Robert Tibshirani.

- Σχήμα 3.2: το γράφημα έχει δημιουργηθεί για τη δημοσίευση "Regression shrinkage and selection via the lasso", του Robert Tibshirani.

- Σχήμα 3.3: το γράφημα έχει δημιουργηθεί για τη δημοσίευση "Regression shrinkage and selection via the lasso", του Robert Tibshirani.

# Κεφάλαιο 4

Θηκογράφημα, σελ. 88

```
> ggplot(thesis.data, aes(hosp, age))
+ geom_boxplot(aes(fill = hosp), alpha = 0.4) + theme_bw()
+ ggtitle(”Hospitalization vs. Age”)
+ xlab(”Hospitalization”)
+ ylab(”Age”)
+ labs(fill =”hosp”)
+ scale_fill_manual(values = c(”lavenderblush3”, ”darkred”))
```

Ραβδόγραμμα φύλου, σελ. 89

```
> ggplot(thesis.data, aes(gender))
+ geom_bar(aes(fill = hosp), alpha = 0.8)
+ scale_fill_manual(values = c(”lavenderblush3”, ”darkred”))
+ theme_bw()
+ ggtitle(”Gender, colored by Hospitalization”)
+ xlab(”Gender”)
+ ylab(”Count”)
+ labs(fill =”hosp”)
+ theme(plot.title = element_text(hjust = 0.5))
```

Ραβδόγραμμα επαγγελματικής κατάστασης, σελ. 90

```
> ggplot(thesis.data, aes(profst))
+ geom_bar(aes(fill = hosp), alpha = 0.8)
+ scale_fill_manual(values = c(”lavenderblush3”, ”darkred”))
+ theme_bw()
+ ggtitle(”Prof status, colored by Hospitalization”)
+ xlab(”Professional Status”)
+ ylab(”Count”)
+ labs(fill =”hosp”)
+ theme(plot.title = element_text(hjust = 0.5))
```

Ραβδόγραμμα κύριας διάγνωσης, σελ. 90

```
> ggplot(thesis.data, aes(diagn))
+ geom_bar(aes(fill = hosp), alpha = 0.8)
+ scale_fill_manual(values = c(”lavenderblush3”, ”darkred”))
+ theme_bw()
+ ggtitle(”Main diagnosis, colored by Hospitalization”)
+ xlab(”Main Diagnosis”)
+ ylab(”Count”) + labs(fill =”hosp”)
+ theme(plot.title = element_text(hjust = 0.5))
```

Διάγραμμα συσχετίσεων, σελ. 92

```
> corrgram(thesis.data, order = T, lower.panel = panel.shade, upper.panel = panel.pie, text.panel = panel.txt, main = ”Correlations”)
```

Διάγραμμα ROC, σελ. 96

```
> plot(1-analysis.logit$specificities, analysis.logit$sensitivities, type=”l”, ylab=”Sensitiviy”, xlab=”1-Specificity”, col=”black”, lwd=2, main = ”ROC Curve”) > abline(a=0,b=1) > abline(v = opt.t.l)
```

Διάγραμμα ROC, σελ. 100

```
> plot(1-analysis.probit$specificities, analysis.probit$sensitivities, type=”l”, ylab=”Sensitiviy”, xlab=”1-Specificity”, col=”black”, lwd=2, main = ”ROC Curve”) > abline(a=0,b=1) > abline(v = opt.t.p)
```

Διάγραμμα ROC, σελ. 104

```
> plot(1-analysis$specificities, analysis$sensitivities, type=”l”, ylab=”Sensitiviy”, xlab=”1-Specificity”, col=”black”, lwd=2, main = ”ROC Curve”) > abline(a=0,b=1) > abline(v = opt_t)
```

# ΒΙΒΛΙΟΓΡΑΦΙΑ

**Α) Διεθνής Βιβλιογραφία**


Agresti, A. (1990). *An introduction to categorical data analysis*. Wiley-Interscience. New York.

Cramer, J. S. (2003). *The origins and development of the logit model*. Campridge University.

Dobson, A. J. (2001). *An introduction to generalized linear models*. Chapman and Hall.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley-Interscience.

James, G. & Witten, D. & Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Katon, W, & Maj, M. & Sartorius, N. (2010). *Depression and Diabetes*. Willey-Blackwell.

Kennedy, S. H. & Lam, R.W. & Nut, D. J. & Thase, M. E. (2004) *Αποτελεσματική θεραπεία της κατάθλιψης*. Ιατρικές εκδόσεις Βαγιονάκης.

Lindsey, J. K. (1997). *Applying generalized linear models*. Springer.

Maindonald, J. & Braun, J. (2003). *Data analysis and graphics using R*. Cambridge University Press.

Global Health Organisation. (2012). *Depression: A global crisis*.

Persons, J.B. & Thase, M. E. & Crits-Christoph, P. (1996). The role of psychotherapy in the treatment of depression. *JAMA Psychiatry*.

Sherwoodandetal, A. (2007). Relationship of depression to death or hospitalization in patients with heart failure. *JAMA Internal Medicine*.

Sullivan, P.F. & Neale, M. C. & Kendler, K. S. (2000). Genetic epidemiology of major depression: review and meta-analysis. *The American Journal of Psychiatry*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistican Society*, 58(1):267–288.

Lieburg, M. J. (1988). *Famous Depressives*. Organon International bv.

Lieburg, M. J. (1989). *Depression and Music*. Organon International bv.

Wilson, J.R. & Lorenz, K.A. (2015). *Modeling Binary Correlated Responses using SAS, SPSS and R*. Springer.

Montgomery, D.C. & Peck, E. (1992). *Introduction to Linear Regression Analysis*. Wiley-Interscience.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Economics*. Cambridge University Press.

Cox, D. R. & Snell, E. J. (1989). *Analysis of binary data*. Second Edition.

Magee, L. (1990). $R^2$ measures based on Wald and likelihood ratio joint significance tests. *The American Statistician*.

Greenberg, B. G. (1980). ”Chester I. Bliss, 1899-1979.” International Statistical Review / Revue Internationale de Statistique 8(1): 135-136.

Finney, D. J. (1952). Probit Analysis. *Cambridge University Press*.

Chambers, E. A. & Cox, D. R. (1967). Discrimination between alternative binary response models. *Biometrika*, 573–578.

Brown, D.A. (1982). Reading diagnosis and remediation. *Englewood Cliffs*, 417–538.

Lindsey, J.K. (1999). On the use of Corrections for Overdispersion. Applied Statistics, 48, 553-561.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. Academic Press. New York.

**Β) Ελληνική Βιβλιογραφία**

Οικονόμου, Π. & Καρώνη, Χ. (2010). *Στατιστικά Μοντέλα Παλινδρόμησης*. Συμεών. Αθήνα.

Ηλιόπουλος, Γ. (2013). *Βασικές μέθοδοι εκτίμησης παραμέτρων*. Αθ. Σταμούλης.

Παγκόσμιος Οργανισμός Υγείας. (2011). *Ταξινόμηση ICD-10 ψυχικών διαταραχών και διαταραχών συμπεριφοράς*.

Χριστοπούλου, Α. (2008). *Εισαγωγή στην ψυχοπαθολογία του ενήλικα*. Τόπος.

Φουσκάκης, Δ. (2009). Παρουσίαση στο μάθημα Ανάλυση δεδομένων με Η/Υ - ΣΕΜΦΕ, (http://www.math.ntua.gr/ fouskakis/)