ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΑΝΑΛΥΣΗ ΜΟΡΙΑΚΩΝ ΔΕΙΚΤΩΝ ΠΟΥ ΣΧΕΤΙΖΟΝΤΑΙ ΜΕ ΤΗΝ ΑΝΑΠΤΥΞΗ ΤΟΥ ΚΑΡΚΙΝΟΥ ΤΗΣ ΟΥΡΟΔΟΧΟΥ ΚΥΣΤΗΣ

Νικόλαος Η. Παπαδημητρίου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς

Ιούνιος 2012

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ..... συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών.

Τα μέλη της Επιτροπής ήσαν:

- Αναπλ. Καθ. Πολίτης Κωνσταντίνος (Επιβλέπων)

- Λεκτ. Ευαγγελάρας Χαράλαμπος

- Επίκ. Καθ. Κοτσίνας Αθανάσιος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

ANALYSIS OF MOLECULAR MARKERS ASSOCIATED WITH THE DEVELOPMENT OF BLADDER CANCER

By

Nick I. Papadimitriou

Thesis

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfillment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece

June 2012

Στους γονείς μου

Ηλία και Βασιλική

Ευχαριστίες

Ευχαριστώ πολύ τον επιβλέποντα καθηγητή μου κ. Πολίτη Κωνσταντίνο για τις συμβουλές του και την καθοδήγηση καθ'όλη την διάρκεια της συγγραφής της εργασίας και τον κ. Κοτσίνα για την διάθεση του δείγματος που πραγματεύτηκα.

Περίληψη

Ο καρκίνος της ουροδόχου κύστης είναι το είδος του καρκίνου πού αναπτύσσεται στα κύτταρα του βλεννογόνου της ουροδόχου κύστης. Είναι η δεύτερη σε συχνότητα νεοπλασία του ουροποιητικού συστήματος, με περισσότερες από 200.000 νέες διαγνώσεις και πάνω από 100.000 θανάτους κάθε χρόνο παγκοσμίως. Η αναλογία νέων περιπτώσεων πού παρουσιάζονται στα δύο φύλλα ανδρών-γυναικών είναι 3:1. Συνήθως παρουσιάζεται στα άτομα μεγαλύτερης ηλικίας , χωρίς αυτό να σημαίνει ότι δεν προσβάλλει και τα νεώτερα άτομα. Ο πιο συχνός ιστολογικός τύπος είναι το καρκίνωμα από μεταβατικό επιθήλιο (transitional cell carcinoma, TCC). Στο μεγαλύτερο ποσοστό των ατόμων με την νόσο η διάγνωση γίνεται στα πρώτα στάδια όπου και η θεραπεία είναι ουσιαστική και τα αποτελέσματα θετικά. Όμως είναι μια νόσος με συχνές υποτροπές, οπότε οι συστάσεις των ουρολόγων είναι η παρακολούθηση σε τακτικά χρονικά διαστήματα για αρκετά χρόνια μετά την πρώτη χειρουργική εξαίρεση τους ,για τον λόγο της έγκαιρης ανίχνευσης τους και βέβαια και της άμεσης αντιμετώπισης τους.

Τα δεδομένα της ανάλυσης μας τα προμηθευτήκαμε από την Ιατρική Σχολή Αθηνών. Αρχικά χρησιμοποιήσαμε κάποιες περιγραφικές στατιστικές μεθόδους που μας επέτρεψαν να πάρουμε μια γενική εικόνα των δεδομένων που είχαμε στην διάθεση μας. Στο κύριο κομμάτι της ανάλυσης μας πήγαμε ένα βήμα παραπάνω χρησιμοποιώντας την θεωρία της ανάλυσης παλινδρόμησης και της ανάλυσης διακριτών δεδομένων. Τέλος, προσπαθήσαμε να αντιμετωπίσουμε τα τυχόν προβλήματα που δημιουργούνται από την ύπαρξη ελλιπών δεδομένων.

Πέραν της ανάλυσης των δεδομένων, σκοπός της εργασίας ήταν και η παρουσίαση επιστημονικώς τεκμηριωμένων αποτελεσμάτων από διεθνείς μελέτες και η σύγκριση τους με τα δικά μας.

Abstract

The bladder cancer is a type of cancer that is developed in cells lining the bladder. It is the second most common malignancy of the urinary system, with more than 200.000 new diagnoses and more than 100.000 deaths each year. The ratio of new cases that arise in both sexes, men and women, is 3:1. Usually it affects older people but that does not mean that it does not affect younger people as well. The most common histologic type is transitional cell carcinoma, TCC. The majority of people with the disease are diagnosed at an early stage, where the treatment is effective and we have positive results. However, it is a disease with frequent relapses and so the urologists recommend regular monitoring for several years after the first surgery in order to detect it at an early stage and face it efficiently.

The data that we used in our analysis were collected by the Medical School of Athens. We originally used descriptive methods, which allowed us to get a general idea about our data. In the main part of our analysis we went even further by implementing regression analysis and categorical data analysis. Finally, we tried to deal with the missing values in our data and the potential problems they cause to our analysis.

Apart from the data analysis, purpose of this paper is also the presentation of scientifically proven results and the comparison with ours as well.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

Βιολογική προσέγγιση

Στο 1^{ο} κεφάλαιο αυτό που θα μας απασχολήσει είναι η κατανόηση από βιολογικής πλευράς του προβλήματος το οποίο έχουμε να αντιμετωπίσουμε. Για αυτό τον λόγο θα παρουσιάσουμε κάποια χρήσιμα στοιχεία που αφορούν στη δομή και λειτουργία των κυττάρων, στην ουροδόχο κύστη όπως και στον καρκίνο της ουροδόχου κύστης.

1.1 Τι είναι το κύτταρο

Το κύτταρο διαδραματίζει σπουδαίο ρόλο για την ανάλυση της παρούσας εργασίας μας καθώς αυτή γίνεται σε επίπεδο κυττάρου, όπως φανερώνουν και οι κύριες μεταβλητές που μας ενδιαφέρει να εξετάσουμε, οι οποίες είναι ο δείκτης κυτταρικού πολλαπλασιασμού Ki 67 και η Κυκλίνη Ε, ομοίως και για τους δείκτες E2F1 και E2F4 που έχουν τον ρόλο των ανεξάρτητων μεταβλητών μας. Επίσης ο καρκίνος της ουροδόχου κύστεως (όπως και κάθε είδος καρκίνου) είναι μια ασθένεια που προσβάλλει τα κύτταρα του εκάστοτε οργάνου. Για τον λόγο αυτό αλλά και για να γίνει πιο κατανοητή η έννοια και ο ρόλος των δεικτών μας θα γίνει αρχικά μια παρουσίαση όσον αφορά στο κύτταρο, μέσω της οποίας περιγράφονται κάποια βασικά στοιχεία για αυτό.

1.1.1 Δομή του κυττάρου

Κατά την Βιολογία, κύτταρο ονομάζεται η βασική δομική και λειτουργική μονάδα που εκδηλώνει το φαινόμενο της ζωής. Έτσι, ως κύτταρο νοείται το μικρότερο δομικό συστατικό της έμβιας ύλης, που αποτελείται από μια συστηματικά οργανωμένη ομάδα μορίων, που βρίσκονται σε δυναμική αλληλεπίδραση μεταξύ τους. Το κύτταρο διαθέτει μορφολογική, φυσική και χημική οργάνωση και την ικανότητα της αφομοίωσης, της ανάπτυξης και της αναπαραγωγής. Είναι μια μονάδα της ζωής ανεξάρτητη ως προς την αυτορρύθμιση και την προσαρμοστικότητά της σε σχέση με το περιβάλλον. Εκ του υφιστάμενου αριθμού αυτών οι οργανισμοί διακρίνονται σε μονοκύτταρους και πολυκύτταρους. Ο χώρος εντός του οποίου βιώνουν τα κύτταρα των πολυκυττάρων οργανισμών ονομάζεται μεσοκυττάριο υγρό. Μεγάλες ομάδες ομοειδών κυττάρων, κατά σύσταση και ορισμένη φυσιολογική λειτουργία, χαρακτηρίζονται ιστοί, (π.χ. μυϊκός ιστός), οι οποίοι και αποτελούν την μονάδα δεύτερης τάξης στον ανθρώπινο οργανισμό, μετά τα κύτταρα.

Τα κύτταρα παρουσιάζουν μεγάλη ποικιλία μεγεθών και διαστάσεων, αντιπροσωπευτικών της ικανότητάς τους για εξελικτική προσαρμογή σε διαφορετικά περιβάλλοντα και της διαφοροποίησής τους. Η διάμετρός τους ποικίλει από δέκατα του μικρομέτρου (ή χιλιοστά του χιλιοστομέτρου), όπως παρατηρείται σε βακτήρια, έως μερικά εκατοστόμετρα, σε θαλάσσια φύκη ή αυγά πτηνών. Τα ανθρώπινα κύτταρα είναι τάξης μεγέθους των 5 χιλιοστών του χιλιοστομέτρου μέχρι 1,5 χιλιοστόμετρο. Υπολογίζεται ότι το ανθρώπινο σώμα αποτελείται από εκατό τρισεκατομμύρια κύτταρα.

Ως οργανισμός, το κύτταρο διαθέτει την ικανότητα να ζει ακόμη και χωρίς την ύπαρξη άλλων κυττάρων. Η ιδιότητα αυτή προϋποθέτει την ύπαρξη μιας μεταβολικής μηχανής που μπορεί να αντλήσει ενέργεια από το περιβάλλον και να τη χρησιμοποιήσει σε ουσιώδεις βιοχημικές διεργασίες, που περιλαμβάνουν την κίνηση ουσιών, την εκλεκτική μεταφορά μορίων μέσα και έξω από το κύτταρο και την ικανότητα αλλαγής και διαμόρφωσής τους, δηλαδή της προσαρμογής τους στις περιβάλλουσες φυσικές και χημικές συνθήκες. Εκτός από τη μεταβολική μηχανή του το κύτταρο διαθέτει ομάδες γονιδίων που καθορίζουν τη σύνθεση ουσιών και μια διακριτή δομή την κυτταρική ή πλασματική μεμβράνη που τα απομονώνει από το εξωτερικό περιβάλλον. Προκειμένου να είναι βιώσιμο ένα κύτταρο, αρκούν 400 γονίδια ή και λιγότερα, ωστόσο τα περισσότερα κύτταρα περιέχουν αρκετά περισσότερα.

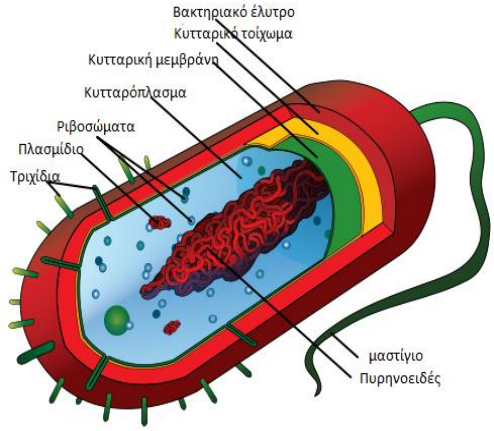
Τα ζωντανά κύτταρα αποτελούνται από περιορισμένο αριθμό χημικών στοιχείων. Ιδιαίτερο ρόλο παίζει ο Άνθρακας (C), το Υδρογόνο (H), το Οξυγόνο (O), το Άζωτο, ο Φωσφόρος (P) και το Θείο (S), που αποτελούν και το 99% περίπου του βάρους του. Τα χημικά συστατικά του είναι δυνατόν να ταξινομηθούν σε ανόργανα (Νερό (H_2O) + μεταλλικά ιόντα) και οργανικά (πρωτεΐνες, υδατάνθρακες, λίπη και νουκλεϊκά οξέα). Ένα ζωικό ή φυτικό κύτταρο αποτελείται κατά προσέγγιση (% κ.β.) από νερό 75-85%, πρωτεΐνες 10-20%, λιπίδια 2-3%, υδατάνθρακες 1% και ανόργανα υλικά (οξέα, βάσεις, άλατα) 1%. Τα τελευταία, αν και βρίσκονται σε πολύ μικρές συγκεντρώσεις, βοηθούν τις κυτταρικές λειτουργίες διατηρώντας σταθερό το pH (www.wikipedia.org).

1.1.2 Είδη κυττάρων

Τα κύτταρα διακρίνονται σε προκαρυωτικά και ευκαρυωτικά,

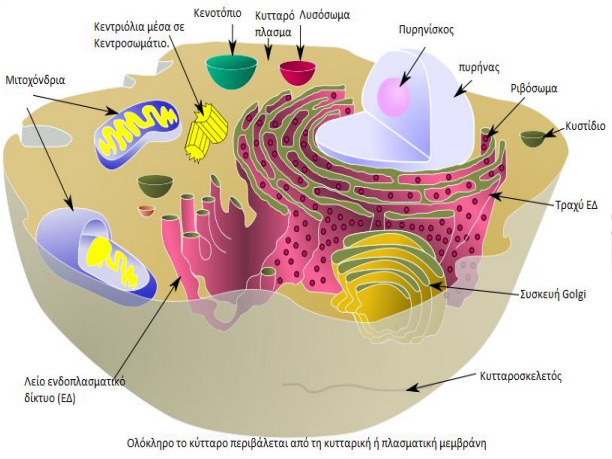
- προκαρυωτικά κύτταρα χαρακτηρίζονται τα κύτταρα εκείνα που παρουσιάζουν απλούστερη δομή από αυτή των ευκαρυωτικών όπως επίσης και μικρότερο μέγεθος. Εξωτερικά τα προκαρυωτικά κύτταρα διαχωρίζονται από το περιβάλλον με την κυτταρική μεμβράνη, πλην όμως εσωτερικά δεν έχουν οποιοδήποτε άλλο μεμβρανικό σχηματισμό, όπως π.χ. κυτταρικό πυρήνα. Αντ' αυτού το γενετικό τους υλικό που είναι σχεδόν πάντα ένα δίκλωνο κυκλικό μόριο DNA βρίσκεται σε μια περιοχή του κυττάρου που λέγεται πυρηνοειδές. Επίσης θεωρείται, λόγω της απλότητας της δομής τους, ότι τα προκαρυωτικά κύτταρα εμφανίστηκαν πριν από τα ευκαρυωτικά. Τέλος, ως επί το πλείστον οι προκαρυωτικοί οργανισμοί είναι μονοκύτταροι και ανήκουν στα Βακτήρια και στα Αρχαία (Εικ. 1.1).
- Ευκαρυωτικά ονομάζονται τα κύτταρα τα οποία έχουν σχηματισμένο πυρήνα. Από αυτό το είδος κυττάρων αποτελούνται ορισμένοι μονοκύτταροι οργανισμοί, όπως πρωτόζωα και φύκη, αλλά και όλοι οι πολυκύτταροι οργανισμοί, όπως τα φυτά και τα ζώα. Τα κύτταρα του ανθρώπου είναι επίσης ευκαρυωτικά. Εξετάζοντας τη δομή αυτών των κυττάρων, παρατηρείται ότι εξωτερικά περικλείονται από μία μεμβράνη και εσωτερικά ο πυρήνας διαχωρίζεται από το υπόλοιπο κύτταρο πάλι με μία μεμβράνη (πυρηνική μεμβράνη). Ανάμεσα στον πυρήνα και στην εξωτερική μεμβράνη, υπάρχουν οργανίδια, τα οποία είναι υπεύθυνα για τις διάφορες λειτουργίες του κυττάρου όπως τα μιτοχόνδρια, το λυσόσωμα, το ριβόσωμα κ.α. Σε αυτή την ταξινόμηση εξαίρεση αποτελούν οι ιοί και οι φάγοι, μια ιδιαίτερη κατηγορία «οργανισμών» με δυνατότητα παρέμβασης στις κυτταρικές λειτουργίες (Εικ. 1.2).

Άλλη ιδιόμορφη κατηγορία ύλης είναι τα μυκοπλάσματα (PPLO), μια ενδιάμεση μορφή ζωής ανάμεσα στους ιούς και τα βακτήρια. Μία ακόμη κατηγορία είναι τα απλοειδή και τα διπλοειδή κύτταρα που διακρίνονται σύμφωνα με τον αριθμό χρωμοσωμάτων που υπάρχουν στον πυρήνα: τα απλοειδή φέρουν περιττό αριθμό χρωμοσωμάτων, τα διπλοειδή άρτιο (www.wikipedia.org).



Εικόνα 1.1 προκαρυωτικό κύτταρο

(Πηγή:www.wikipedia.org)



εικόνα 1.2 τυπικό ζωικό ευκαρυωτικό κύτταρο

(Πηγή:www.wikipedia.org)

1.2 Κυτταρικός κύκλος

Ο κυτταρικός κύκλος είναι μια σειρά γεγονότων που πραγματοποιούνται στο κύτταρο και το οδηγούν στο διπλασιασμό του. Ο κυτταρικός κύκλος είναι μία διαδικασία ζωτικής σημασίας καθώς μέσω αυτού ένα γονιμοποιημένο ωάριο αναπτύσσεται σε έναν ώριμο οργανισμό, όπως επίσης τα μαλλιά, το δέρμα και κάποια όργανα του σώματος ανανεώνονται. Στα κύτταρα χωρίς πυρήνα (προκαρυωτικά) ο κυτταρικός κύκλος γίνεται μέσω μιας διαδικασίας που λέγεται δυαδική σχάση. Σε κύτταρα με πυρήνα (ευκαρυωτικά) ο κυτταρικός κύκλος γίνεται σε 4 φάσεις: G1, S (synthesis), G2 (οι 3 αυτές λέγονται από κοινού μεσοφάση) και M (μίτωση) (Εικ. 1.3).

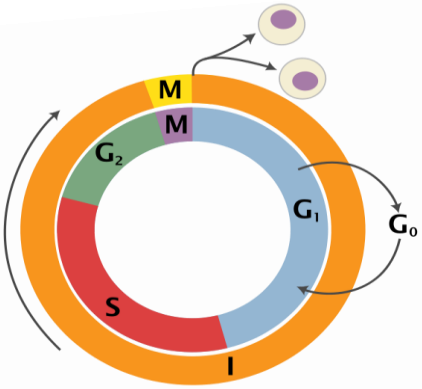
1.2.1 Στάδια του κύκλου

1. Φάση G_1 (gap 1): Η πρώτη φάση ύστερα από την προηγούμενη μίτωση και μέχρι την νέα έναρξη σύνθεσης του DNA ονομάζεται G_1 ή εναλλακτικά φάση ανάπτυξης. Κύριο χαρακτηριστικό της φάσης αυτής είναι οι διάφορες βιοχημικές διεργασίες που λαμβάνουν χώρα μέσα στο κύτταρο με σκοπό την παραγωγή ενζύμων που θα χρειαστούν στην φάση S για την σύνθεση του θυγατρικού DNA.
2. Φάση S (DNA synthesis, σύνθεση DNA): στην διάρκεια της φάσης αυτής το DNA αντιγράφει με εξαιρετική ακριβεία τον εαυτό του δημιουργώντας ένα πανομοιότυπο αντίγραφό του (θυγατρικό DNA).

3. Φάση G_2 (gap 2): κατά τη διάρκεια της φάσης αυτής το κύτταρο προετοιμάζεται για την είσοδό του στην επόμενη φάση, αυτή της μίτωσης. Πάλι σημαντικές βιοχημικές διεργασίες γίνονται για την παραγωγή ουσιών, οι οποίες είναι απαραίτητες για την διαδικασία της μίτωσης.
4. Φάση M (mitosis, μίτωση): το διπλασιασμένο DNA διαχωριζόμενο με ακρίβεια σε δύο ίσα μέρη συμπυκνώνεται σχηματίζοντας χρωμοσώματα ώστε καθένα από τα δύο θυγατρικά κύτταρα να διαθέτει πλήρες αντίγραφο του γενετικού υλικού του μητρικού κυττάρου. Σφάλματα κατά την φάση της μίτωσης μπορούν είτε να καταστρέψουν το κύτταρο μέσω του μηχανισμού της απόπτωσης είτε να προκαλέσουν μεταλλάξεις που μπορεί να οδηγήσουν σε καρκίνο.

Πέρα των τεσσάρων φάσεων που ως άνω αναφέρθηκαν υπάρχει και η φάση G_0 (φάση ηρεμίας), όπου το διηρημένο κύτταρο που προέκυψε από την μίτωση που μόλις ολοκληρώθηκε μπορεί, αντί της εισόδου στην φάση G_1 που το προπαρασκευάζει για νέα άμεση διαίρεση, να εισέλθει σε μια φάση ηρεμίας, η διάρκεια της οποίας ποικίλει ανάλογα με τις συνθήκες και τις ανάγκες του οργάνου, του ιστού, ή του οργανισμού.

Η επιλογή εισόδου του κυττάρου στην G_0 αντί της G_1 καθορίζεται από ποικίλα εξωτερικά ερεθίσματα (επάρκεια ή έλλειψη εξωγενών μιτογόνων, αυξητικών ή άλλων παραγόντων). Ανά πάσα στιγμή και με την επίδραση των μεταβολών των ως άνω εξωτερικών ερεθισμάτων, η φάση ηρεμίας (G_0) μετατρέπεται σε φάση G_1, φάση δηλαδή προετοιμασίας του κυττάρου για διαίρεση. Το καθοριστικό σημείο μετάπτωσης από την φάση ηρεμίας (G_0) στην φάση G_1 ονομάζεται restriction point (R – περιοριστικό σημείο «επαναλειτουργίας»). Μετά την δίοδο από το σημείο R το κύτταρο δεν υπακούει πλέον στους εξωγενείς παράγοντες και η πορεία του διά του κυτταρικού κύκλου μέχρι την ολοκλήρωση της μίτωσης είναι προδιαγεγραμμένη (http://www.onco.gr/documents/RigasAthanasiou.pdf).



Εικ. 1.3 απεικόνιση του κυτταρικού κύκλου.
Εξωτερικό δαχτυλίδι:
I = μεσοφάση,
M = μίτωση.
Εσωτερικό δαχτυλίδι:
M = μίτωση, G_1 = φάση 1,
G_2 = φάση 2, S = σύνθεση.
Εκτός δαχτυλιδιού: G_0 =φάση ηρεμίας.

Εικόνα 1.3 (Πηγή: www.wikipedia.org)

1.2.2 Η ρυθμιστική διαδικασία του κυτταρικού κύκλου

Η ρυθμιστική διαδικασία του κυτταρικού κύκλου εξασφαλίζεται μέσω:

1. των κυκλινών, ετεροδιμεροί συμπλεγμάτα πρωτεϊνών που διακρίνονται σε οικογένειες (A, B, D, E). Κάθε οικογένεια κυκλινών έχει την ευθύνη ελέγχου διαφορετικής φάσης του κυτταρικού κύκλου
2. των κυκλινοεξαρτωμένων κινασών (CDK). Κάθε κυκλινοεξαρτώμενη κινάση επάγει την δράση μιας κυκλίνης. Όταν ενεργοποιηθεί από μία κυκλίνη η κυκλινοεξαρτώμενη κινάση εκτελεί μία βιοχημική αντίδραση που ονομάζεται φωσφορυλίωση η οποία ενεργοποιεί ή απενεργοποιεί συγκεκριμένες πρωτεΐνες ώστε να πραγματοποιηθεί η είσοδος στην επόμενη φάση του κυτταρικού κύκλου.
3. των αναστολέων των συμπλεγμάτων κυκλινών που διακρίνονται σε δύο οικογένειες, την *cip/kip* (*CDK interacting protein/Kinase inhibitory protein*) και την INK4a/ARF (*In*hibitor of *K*inase 4/*A*lternative *R*eading *F*rame) οι οποίες έχουν ως σκοπό την διακοπή της συνέχισης του κυτταρικού κύκλου. Εφόσον τα γονίδια αυτά χρησιμοποιούνται ως μέσα για την αποτροπή δημιουργίας όγκων ονομάζονται επίσης ογκοκατασταλτικά (www.wikipedia.org).

1.2.3 Σημεία ελέγχου κυτταρικού κύκλου (checkpoints)

Τα σημεία ελέγχου χρησιμοποιούνται από το κύτταρο για τον έλεγχο και την ρύθμιση του κυτταρικού κύκλου. Επιπλέον αποτρέπουν την πρόοδο του κύκλου σε συγκεκριμένα σημεία, επιτρέποντας με τον τρόπο αυτό την επαλήθευση των απαραίτητων διαδικασιών που απαιτούνται και την διόρθωση τυχόν βλαβών του

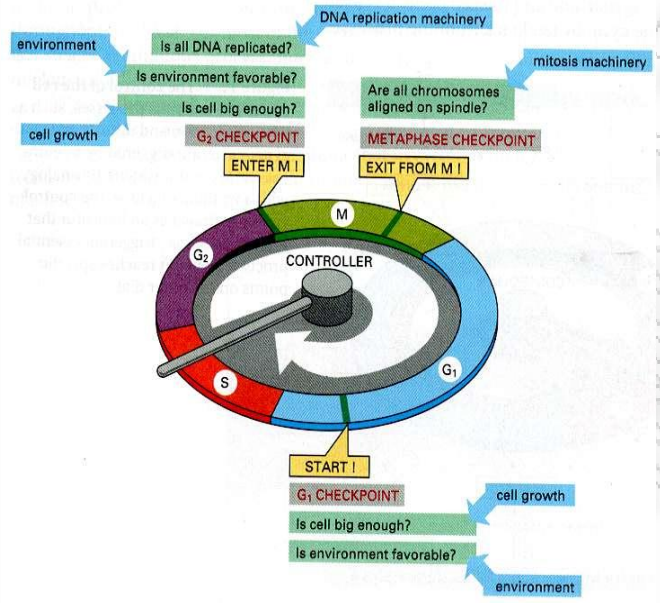
DNA. Ο κυτταρικός κύκλος δεν μπορεί να περάσει στην επόμενη φάση αν δεν πληρούνται τα κριτήρια του εκάστοτε σημείου ελέγχου. Συνολικά υπάρχουν 4 σημεία ελέγχου (Εικ. 1.4) :

1. Σημείο ελέγχου μεταξύ των φάσεων G_1 και S: Κατά την διάρκεια της μετάβασης από την φάση G_1 στην φάση S το σημείο ελέγχου συνίσταται στην ενεργοποίηση ή μη του ογκοκατασταλτικού γονιδίου p53, και πιθανώς των συγγενών του γονιδίων p63 και p73, που λειτουργεί ως φραγμός στην συνέχιση του κυτταρικού κύκλου με δύο τρόπους: 1) Βλάβη του DNA → ενεργοποίηση του p53 → ενεργοποίηση του $p21^{WAF1/CIP1}$CKI → αναστολή της δράσης του συμπλέγματος κυκλίνης/CDK → αναχαίτιση. 2) Ενεργοποίηση μέσω του p53 της διαδικασίας απόπτωσης (προγραμματισμένος κυτταρικός θάνατος). Είναι άγνωστο ποια από τις δύο διαδικασίες θα ακολουθηθεί σε κάθε περίπτωση.
2. Σημείο ελέγχου φάσης S: Το γονίδιο ATM, το οποίο πρόκειται για μια πυρηνική και κυτταροπλασματική κινάση που προάγει την φωσφορυλίωση συγκεκριμένων πρωτεϊνών που συμμετέχουν στα σημεία ελέγχου της βλάβης και της επισκευής του DNA.
3. Σημείο ελέγχου μεταξύ των φάσεων G_2 και M: Σημαντικό ρόλο στην αναχαίτιση του κυτταρικού κύκλου κατά την μετάβαση από την φάση G2 στην φάση M μετά από βλάβη του DNA παίζει η ενεργοποίηση του συμπλέγματος κυκλίνης B/CDK1 που υπό φυσιολογικές συνθήκες επιτρέπει στο κύτταρο να εισέλθει στη φάση της μίτωσης. Συγκεκριμένα, η βλάβη του DNA προκαλεί: ενεργοποίηση της κινάσης chk1 → φωσφορυλίωση του cdc25, γεγονός απαραίτητο για την ενεργοποίηση του συμπλέγματος κυκλίνη B/CDK1→ αναχαίτιση του κυτταρικού κύκλου.
4. Η φάση της μίτωσης, όπου το κύτταρο διαιρείται σε δύο πανομοιότυπα κύτταρα, ολοκληρώνεται σε τέσσερις διαδοχικές φάσεις:
   I. Πρόφαση
   II. Μετάφαση
   III. Ανάφαση
   IV. Τελόφαση

   Το σημείο ελέγχου για την φάση της μίτωσης βρίσκεται κατά την μετάβαση από την Μετάφαση στην Ανάφαση όπου τα χρωμοσώματα συνδέονται στην μιτωτική άτρακτο με σκοπό την σύνδεση όλων των

ζευγών χρωμοσωμάτων σε ενιαίο σύνολο. Τα γονίδια που εμπλέκονται στην διαδικασία αυτή είναι τα MAD1, MAD2, MAD3, BUB1, BUB2, BUB3, και πιθανώς το p53.

Προκειμένου να ξεκινήσει και να ολοκληρωθεί η κάθε φάση του κύκλου, θα πρέπει πρώτα να βεβαιωθεί ότι έχει ολοκληρωθεί επιτυχώς η προηγούμενη φάση. Τα σημεία ελέγχου ενεργοποιούνται από ποικιλία ενδογενών και εξωγενών παραγόντων (π.χ. ιοντίζουσα ακτινοβολία, διατροφικοί παράγοντες, θερμοκρασία, βλάβη του DNA, κ.α.) (http://www.onco.gr/documents/RigasAthanasiou.pdf).



Εικ. 1.4 Απεικόνιση των σημείων ελέγχου κατά την διάρκεια του κυτταρικού κύκλου

Εικόνα 1.4

(Πηγή:http://web.campbell.edu/faculty/garrett/PHAR%20408/cell%20cycle%20checkpoints.jpg)

1.2.4 Ο ρόλος του κυτταρικού κύκλου στην δημιουργία όγκων

Μια απορρύθμιση των συστατικών του κυτταρικού κύκλου μπορεί να οδηγήσει στην δημιουργία όγκων. Γονίδια όπως οι αναστολείς του κυτταρικού κύκλου κατά τη μετάλλαξή τους μπορεί να προκαλέσουν συνεχή και μη ελεγχόμενο πολλαπλασιασμό του κυττάρου με συνέπεια την δημιουργία όγκου. Αν και η διάρκεια του κυτταρικού κύκλου παραμένει η ίδια, το ποσοστό των κυττάρων που βρίσκονται σε διαδικασία διαίρεσης έναντι των κυττάρων σε κατάσταση ηρεμίας (G0) στην περίπτωση όγκου είναι πολύ μεγαλύτερο από ότι σε έναν φυσιολογικό ιστό. Κατά συνέπεια υπάρχει μια σημαντική αύξηση στον αριθμό των κυττάρων, καθώς ο αριθμός των κυττάρων που πεθαίνουν λόγω της απόπτωσης ή γήρατος παραμένει σταθερός (www.wikipedia.org).

1.3 Ουροδόχος κύστη

Η ουροδόχος κύστη αποτελεί μαζί με τους νεφρούς, τους ουρητήρες και την ουρήθρα το ουροποιητικό σύστημα του ανθρώπου, σκοπός του οποίου είναι η παραγωγή και η αποβολή των ούρων και μαζί με αυτά μιας σειράς άχρηστων συστατικών που παράγονται στον οργανισμό από τις καύσεις, καθώς και η διατήρηση του ισοζυγίου του νερού και των ηλεκτρολυτών στο ανθρώπινο σώμα. Ο ρόλος της ουροδόχου κύστης είναι η συλλογή των ούρων που εκκρίνουν οι νεφροί πριν την αποβολή τους μέσω της ούρησης. Τα ούρα εισέρχονται στην κύστη μέσω των ουρητήρων και εξέρχονται μέσω της ουρήθρας.

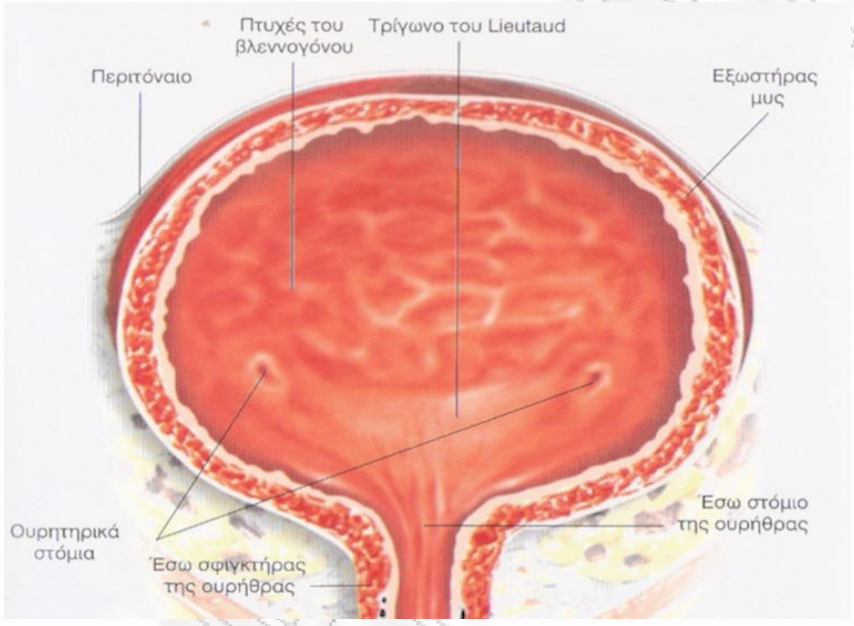
1.3.1 Δομή της κύστης

Είναι ένα κοίλο μυώδες όργανο μεταβλητών διαστάσεων (ανάλογα με το βαθμό πλήρωσής της). Βρίσκεται στο έδαφος της μικρής πυέλου πίσω από την ηβική σύμφυση και μοιάζει με μπαλόνι. Είναι υποπεριτοναϊκό όργανο και το επάνω μέρος της καλύπτεται από το περιτόναιο το οποίο την καθηλώνει στο έδαφος της πυέλου. Στηρίζεται και με διάφορους συνδέσμους που την συγκρατούν στο πρόσθιο τοίχωμα της κοιλιάς. Επάνω από την ουροδόχο κύστη βρίσκονται οι έλικες του ειλεού, πίσω της βρίσκονται στους άνδρες το ορθό και οι σπερματοδόχες λήκυθοι, και στις γυναίκες βρίσκεται η μήτρα. Κάτω από την ουροδόχο κύστη στις γυναίκες βρίσκεται ο κόλπος ενώ στους άνδρες ο προστάτης αδένας.

Στο εσωτερικό του οργάνου διακρίνουμε στο επάνω μέρος τον θόλο και στο κάτω το έδαφος. Στο έδαφος υπάρχει μια τριγωνική περιοχή, με τη βάση προς τα πίσω και την κορυφή προς τα εμπρός, η οποία λέγεται *κυστικό τρίγωνο*. Στις κορυφές του κυστικού τριγώνου υπάρχουν στόμια· τα δύο στόμια που βρίσκονται στο πίσω μέρος είναι τα στόμια εισόδου των ουρητήρων και το πρόσθιο (της κορυφής του τριγώνου) είναι το στόμιο εξόδου της ουρήθρας. Η επιφάνεια του κυστικού τριγώνου είναι πάντα λεία και ομαλή ενώ στο θόλο υπάρχουν πτυχές, όταν η κύστη είναι άδεια. Όταν γεμίσει, οι πτυχές αυτές εξαφανίζονται, καθώς το τοίχωμα της κύστης τεντώνει. Το τοίχωμα της ουροδόχου κύστης αποτελείται από έναν εξωτερικό λεπτό ινώδη ορογόνο χιτώνα, ένα μυϊκό χιτώνα από λείες μυϊκές ίνες, και τέλος, στο εσωτερικό, από τον βλεννογόνο, ο οποίος έχει μεταβατικό επιθήλιο. Ο μυϊκός χιτώνας σχηματίζει τον εξωστήρα μυ της κύστης ο οποίος, όταν συσπάται, εξωθεί τα περιεχόμενα ούρα προς την ουρήθρα. Στο στόμιο εξόδου της ουρήθρας υπάρχει ένας σφιγκτηρικός

μηχανισμός από λείες μυϊκές ίνες, του οποίου η λειτουργία είναι ακούσια. Ένας δεύτερος σφιγκτήρας από τους μύες του περινέου, πιο περιφερειακά από τον πρώτο, λειτουργεί με τη θέλησή μας.

Μπορούμε να κρατήσουμε τα ούρα στην κύστη μας μέχρι κάποιο όριο χωρίς πρόβλημα, ωστόσο όταν ο όγκος των ούρων που είναι μέσα στην ουροδόχο κύστη ξεπεράσει τα 400 cc η κύστη συσπάται κι αρχίζουμε να νιώθουμε ένα δυσάρεστο αίσθημα. Αν προσπαθήσουμε να κρατήσουμε τα ούρα περισσότερο, το αίσθημα αυτό επιδεινώνεται και όταν ο όγκος των ούρων φθάσει τα 650-700 cc η κύστη συσπάται μόνη της, οι σφιγκτήρες χαλαρώνουν και προκαλείται αυτόματη ούρηση, ανεξάρτητη από τη θέλησή μας, για λόγους προστασίας της ακεραιότητας της ουροδόχου κύστης (www.wikipedia.org).



Εικόνα 1.6 Απεικόνιση της ουροδόχου κύστης
(Πηγή:www.e-urology.gr)

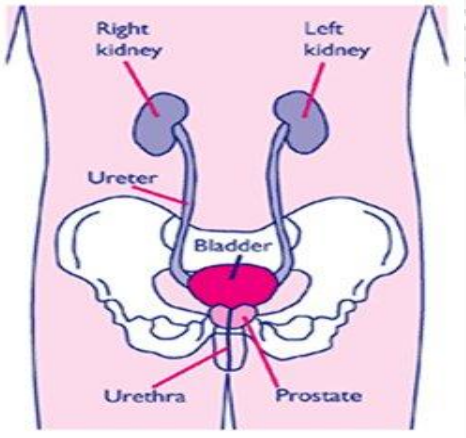
1.4 Καρκίνος της ουροδόχου κύστης

Στην παρούσα παράγραφο θα γίνει μια παρουσίαση της ασθένειας του καρκίνου της ουροδόχου κύστης. Πιο συγκεκριμένα αρχικά γίνεται μια γενική επισκόπηση της ασθένειας, στην συνέχεια παρουσιάζονται οι διάφοροι τύποι του καρκίνου, οι αιτίες και τα συμπτώματα της εμφάνισής του. Τέλος γίνεται παρουσίαση των μεθόδων θεραπείας και κάποιων στατιστικών στοιχείων. Οι πληροφορίες και τα δεδομένα αντλήθηκαν από την ιστοσελίδα www.emedicinehealth.com, η οποία περιλαμβάνει

πολλά ιατρικά άρθρα που αποσκοπούν στην ενημέρωση των ασθενών και απλών ανθρώπων γύρω από τα διάφορα είδη καρκίνων, και από την ιστοσελίδα του οργανισμού Cancer Research Uk, ο οποίος έχει ως στόχο την καταπολέμηση του καρκίνου μέσω της συνεχούς έρευνας, της ενημέρωσης των πολιτών και την ευαισθητοποίησης της πολιτείας (http:// info.cancerresearchuk.org /cancerstats/types/ bladder).

1.4.1 Γενική επισκόπηση

Ο καρκίνος της ουροδόχου κύστης εμφανίζεται όταν φυσιολογικά κύτταρα υποβάλλονται σε μεταλλάξεις και αρχίζουν να αναπτύσσονται και να πολλαπλασιάζονται ανεξέλεγκτα. Καθώς όλο και περισσότερα κύτταρα δημιουργούνται το μέγεθος του όγκου αυξάνει. Στην συνέχεια οι όγκοι κατακλύζουν τους γειτονικούς ιστούς καταλαμβάνοντας τον χώρο τους και δεσμεύοντας οξυγόνο και θρεπτικά συστατικά που χρειάζονται για την λειτουργία και επιβίωση τους. Οι όγκοι είναι καρκινικοί όταν είναι κακοήθεις και είναι πιθανό να ταξιδέψουν μέσω του αίματος ή του λεμφικού συστήματος σε απομονωμένα όργανα. Η διαδικασία με την οποία ο καρκίνος εισβάλει και επεκτείνεται σε άλλα όργανα ονομάζεται μετάσταση. Ο καρκίνος της ουροδόχου κύστης είναι πολύ πιο πιθανό να επεκταθεί σε γειτονικά όργανα και λεμφαδένες πριν να επεκταθεί μέσω της κυκλοφορίας του αίματος στους πνεύμονες, στο συκώτι, στα κόκκαλα ή σε άλλα όργανα.



Εικόνα 1.6

(Πηγή:http://info.cancerresearchuk.org/cancerstats/keyfacts/bladder-cancer/)

Εικ. 1.6 Αναπαράσταση του ουροποιητικού συστήματος στον άνδρα.

1.4.2 Τύποι του καρκίνου

Από τα διαφορετικά είδη κυττάρων που συνθέτουν την ουροδόχο κύστη εκείνα που βρίσκονται εσωτερικά του τοιχώματος της κύστης είναι το πιθανότερο να αναπτύξουν καρκίνο. Οι τύποι που αναφέρονται παρακάτω παίρνουν το όνομα τους από τα αντίστοιχα κύτταρα που μπορεί να γίνουν καρκινικά.

- **Ουροθηλιακό καρκίνωμα (Urothelial carcinoma)**: είναι με διαφορά ο πιο κοινός τύπος καρκίνου. Οφείλεται σε κύτταρα τα οποία συγκροτούν το εσώτατο στρώμα του τοιχώματος της κύστης τα οποία υποβάλλονται σε αλλαγές οι οποίες οδηγούν σε ανεξέλεγκτο πολλαπλασιασμό.
- **Πλακώδες καρκίνωμα (Squamous cell carcinoma)**: Είναι ένας τύπος καρκίνου που ξεκινάει στα πλακώδη κύτταρα (λεπτά, επίπεδα κύτταρα που μοιάζουν με λέπια ψαριού) ως αποτέλεσμα μιας λοίμωξης ή μιας ενόχλησης της κύστης η οποία έχει λάβει χώρα για πολλούς μήνες ή χρόνια.
- **Αδενοκαρκίνωμα (adenocarcinoma)**: Αυτοί οι καρκίνοι σχηματίζονται από κύτταρα που απαρτίζουν αδένες. Αδένες είναι εξειδικευμένες δομές που παράγουν και ελευθερώνουν υγρά, όπως βλέννα.

Στις Ηνωμένες Πολιτείες το ουροθηλιακό καρκίνωμα αντιστοιχεί σε ποσοστό μεγαλύτερο του 90% όλων των περιπτώσεων καρκίνου της ουροδόχου κύστης. Ακολουθούν το πλακώδες καρκίνωμα σε ποσοστό 3%-8% και το αδενοκαρκίνωμα με 1%-2% (http://www.emedicinehealth.com/bladder_cancer).

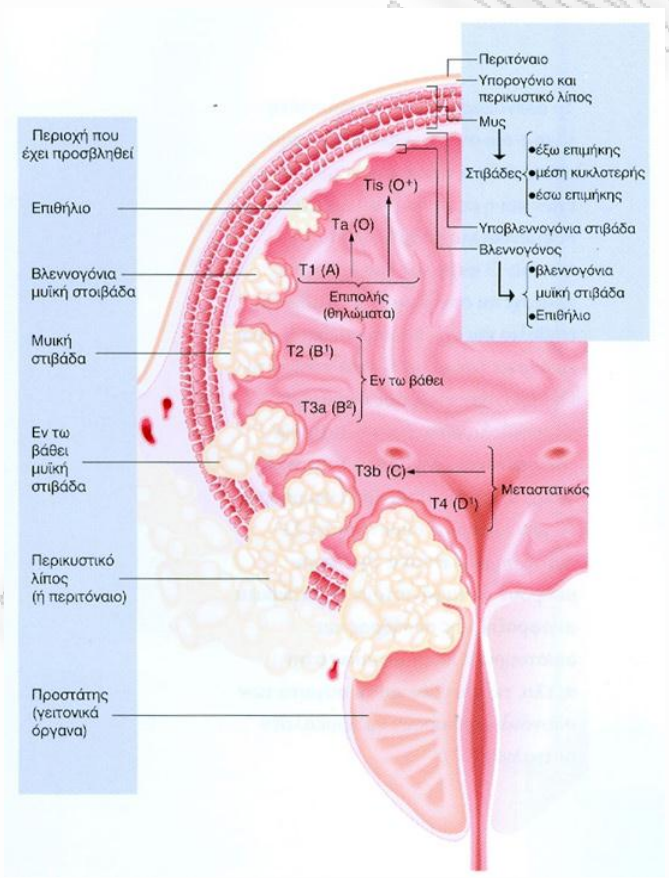
Οι καρκίνοι της ουροδόχου κύστης ταξινομούνται από το πόσο βαθιά έχουν εισβάλει στο τοίχωμα της ουροδόχου κύστης, το οποίο έχει πολλά στρώματα. Πολλοί γιατροί υποδιαιρούν τον καρκίνο της ουροδόχου κύστης σε επιφανειακό και διηθητικό. Ο επιφανειακός καρκίνος περιορίζεται στην εσωτερική επιφάνεια της ουροδόχου κύστης. Ο διηθητικός καρκίνος έχει διαπεράσει τουλάχιστον το μυϊκό στρώμα του τοιχώματος της ουροδόχου κύστης.

- Σχεδόν όλα τα αδενοκαρκινώματα και τα πλακώδη καρκινώματα είναι διηθητικά. Έτσι, από τη στιγμή που εντοπίζονται αυτές τις μορφές καρκίνου, συνήθως έχουν εισβάλει ήδη το τοίχωμα της ουροδόχου κύστης.
- Πολλά ουροθηλιακά καρκινώματα δεν είναι διηθητικά. Αυτό σημαίνει ότι δεν πάνε βαθύτερα από το επιφανειακό στρώμα της ουροδόχου κύστης.

Πέραν του σταδίου του καρκίνου (δηλαδή πόσο βαθιά έχει εισχωρήσει στο τοίχωμα της κύστης), ο βαθμός του καρκίνου (grade) παρέχει σημαντικές πληροφορίες που μπορούν να βοηθήσουν στην θεραπεία. Ο βαθμός σχετίζεται με το ύψος την ανωμαλίας που παρατηρείται κατά την μικροσκοπική εκτίμηση του όγκου. Κύτταρα από όγκο μεγάλου βαθμού έχουν μεγαλύτερη ανωμαλία από αντίστοιχα κύτταρα ενός όγκου μικρού βαθμού.

- Οι όγκοι χαμηλού βαθμού είναι λιγότεροι επιθετικοί.
- Όγκοι μεγάλου βαθμού είναι πιο επικίνδυνοι και έχουν την τάση να γίνονται διηθητικοί.

Από όλους τους τύπους καρκίνου, ο καρκίνος της ουροδόχου κύστης έχει μια ασυνήθιστα υψηλή τάση να επανεμφανίζεται μετά τη θεραπεία. Ο καρκίνος της ουροδόχου κύστης έχει ένα ποσοστό υποτροπής 50% με 80% και οι νέες περιπτώσεις είναι συνήθως, αλλά όχι πάντα, του ίδιου τύπου με την πρώτη, ενώ η ασθένεια αυτή μπορεί να προσβάλλει την κύστη ή άλλο τμήμα του ουροποιητικού συστήματος (νεφρά ή ουρητήρες).



Εικ. 1.7 Απεικόνιση των σταδίων εξέλιξης του καρκίνου.

Εικόνα 1.7

(Πηγή:http://www.andrologia.gr)

1.4.3 Αίτια

Δεν είναι ξεκάθαρο τι ακριβώς προκαλεί τον καρκίνο, αλλά γνωρίζουμε ότι οι ακόλουθοι παράγοντες αυξάνουν τον κίνδυνο να εμφανίσει κάποιος καρκίνο της ουροδόχου κύστης.

- **Κάπνισμα** : Το κάπνισμα αποτελεί τον σημαντικότερο παράγοντα για την εμφάνιση του καρκίνου ανεξαρτήτως φύλου. Στην Ευρώπη εκτιμάται ότι τα δύο τρίτα των περιπτώσεων στους άντρες και το ένα τρίτο στις γυναίκες οφείλεται στο κάπνισμα. Οι εν ενεργεία καπνιστές έχουν δύο με έξι φορές μεγαλύτερο κίνδυνο εμφάνισης του καρκίνου σε σύγκριση με όσους δεν έχουν καπνίσει ποτέ. Η διακοπή του καπνίσματος μειώνει τον κίνδυνο αυτό αλλά εξακολουθεί να είναι μεγαλύτερος σε σχέση με κάποιον που δεν έχει καπνίσει καθόλου για περισσότερα από είκοσι χρόνια.
- **Έκθεση σε χημικές ουσίες στο χώρο εργασίας**: Ο καρκίνος της ουροδόχου κύστης ήταν ένας από τους πρώτους καρκίνους που έχει αποδειχθεί ότι συνδέεται με τον βιομηχανικό τομέα. Τα πρώτα περιστατικά καταγράφηκαν στην Γερμανία το 1895 σε ένα εργοστάσιο βαφής ανιλίνης, αλλά δεν ήταν μέχρι την δεκαετία του 1950 που αποδείχτηκε ο κίνδυνος από τις αρωματικές αμίνες και ειδικότερα την βενζιδίνη και την α και β ναφθυλαμίνη. Η έκθεση σε πολυκυκλικούς αρωματικούς υδρογονάνθρακες (ΠΑΥ), οι οποίοι είναι υποπροϊόντα των διαδικασιών καύσης και κατά συνέπεια παρόντες σε ένα ευρύ φάσμα κλάδων, έχει επίσης διερευνηθεί. Υπολογίζεται ότι περίπου 4% των περιπτώσεων καρκίνου της ουροδόχου κύστης στους άνδρες στην Ευρώπη οφείλονται στην έκθεση σε ΠΑΥ. Το ποσοστό αυτό μπορεί να είναι υψηλότερη σε χώρες με λιγότερο οργανωμένες βιομηχανικές διεργασίες.
- **Διατροφικές συνήθειες**: Μια διατροφή πλούσια σε τηγανιτά κρέατα και ζωικά λίπη πιστεύεται ότι αυξάνει τον κίνδυνο εμφάνισης του καρκίνου. Επίσης, τα αποτελέσματα μιας συγκεντρωτικής ανάλυσης από 10 ευρωπαϊκές μελέτες δείχνουν ότι η μεγάλη κατανάλωση καφέ (πάνω από 10 φλιτζάνια την ημέρα) σχετίζεται με σημαντικά αυξημένο κίνδυνο καρκίνου της ουροδόχου κύστης στους άνδρες και τις γυναίκες, αλλά δεν υπάρχουν αποδείξεις για αύξηση του κινδύνου με μέτρια κατανάλωση.
- **Σχιστοσωμίαση**: Σχιστοσωμίαση είναι μια παρασιτική λοίμωξη που εμφανίζεται στην Αφρική και τη Μέση Ανατολή. Μια μορφή του

παρασίτου Schistosoma haematobium συνδέεται με τον καρκίνο της ουροδόχου κύστης και υπολογίζεται ότι προκάλεσε περίπου 10.600 περιπτώσεις της νόσου το 2002.

- **Η χρόνια φλεγμονή της ουροδόχου κύστης**: Οι συχνές λοιμώξεις της ουροδόχου κύστης, πέτρες της ουροδόχου κύστης, και άλλα προβλήματα του ουροποιητικού συστήματος που ερεθίζουν την κύστη αυξάνουν τον κίνδυνο εμφάνισης καρκίνου και πιο συχνά τον τύπο του πλακώδους καρκινώματος.
- **Οικογενειακό ιστορικό**: Πολλές μελέτες δείχνουν δύο έως έξι φορές μεγαλύτερο κίνδυνο καρκίνου της ουροδόχου κύστης σε συγγενείς πρώτου βαθμού των ασθενών με καρκίνο της ουροδόχου κύστης, με υψηλότερο κίνδυνο, εάν η σχετική διάγνωση για τον συγγενή γίνει πριν από την ηλικία των 45.

1.4.4 Συμπτώματα

Το πιο συνηθισμένο σύμπτωμα σε περισσότερες του 80% των περιπτώσεων είναι η ανώδυνη και συνήθως διαλείπουσα αιματουρία η οποία είναι συχνά σύμπτωμα και άλλων λιγότερο σοβαρών προβλημάτων. Παρόλα αυτά θα πρέπει πάντα να ερευνάται καθώς μπορεί με αυτό τον τρόπο να επιτευχτεί μια γρήγορη διάγνωση. Άλλα συμπτώματα είναι πόνος ή κάψιμο κατά την ούρηση χωρίς ενδείξεις ουρολοίμωξης όπως και η αλλαγή στις συνήθειες της ουροδόχου κύστης, όπως το να χρειάζεται να ουρήσει κάποιος συχνότερα ή να αισθάνεται την έντονη ανάγκη για ούρηση χωρίς να υπάρχουν πολλά ούρα. Πάλι όμως η ύπαρξη αυτών των συμπτωμάτων δεν σημαίνει ότι κάποιος πάσχει απαραίτητα από καρκίνο.

1.4.5 Έγκαιρη διάγνωση

Δυστυχώς ο καρκίνος της κύστεως στα αρχικά στάδια που είναι ιάσιμος είναι ασυμπτωματικός και μπορεί να παραμείνει για μεγάλο διάστημα χωρίς κανένα σύμπτωμα. Το μοναδικό σύμπτωμα είναι η αιματουρία που μπορεί να είναι μικροσκοπική ή μακροσκοπική. Για το λόγο αυτό πρέπει κάθε αιματουρία να ελέγχεται ακόμη και όταν εμφανίζεται για πρώτη φορά. Ο έλεγχος γίνεται με τις παρακάτω εξετάσεις:

- Υπερηχογράφημα νεφρών, κύστεως, προστάτη. Ελέγχεται όλο το ουροποιητικό σύστημα για την ύπαρξη νεοπλασίας.

- Κυτταρολογική ούρων. Εξετάζονται τα ούρα για ύπαρξη νεοπλασματικών κυττάρων. Πρόκειται για μία απλή εξέταση ενώ το ποσοστό επιτυχούς διάγνωσης είναι μεγάλο.
- Καρκινικοί δείκτες. Είναι βιοχημικές εξετάσεις με τις οποίες βρίσκουμε καρκινικές ουσίες στα ούρα που ελευθερώνονται από τα νεοπλασματικά κύτταρα. Δηλαδή ρίχνοντας λίγες σταγόνες από το αντιδραστήριο στα ούρα μπορούμε αμέσως με την αλλαγή του χρώματος να έχουμε πληροφορίες για την ύπαρξη νεοπλασματικών κυττάρων.
- Κυστεοσκόπηση. Είναι μία απλή εξέταση κατά την οποία περνώντας από την ουρήθρα ένα ειδικό εργαλείο που λέγεται κυστεοσκόπιο μπορούμε να δούμε όλη την κύστη και να διαπιστώσουμε την ύπαρξη ή μη όγκου. Εκτός από την διάγνωση σε περίπτωση που υπάρχει νεόπλασμα μπορούμε να έχουμε πληροφορίες όσον αφορά το στάδιο, το μέγεθος, και την θέση, πληροφορίες που μας βοηθούν στην επιλογή της θεραπείας.
- Αξονική τομογραφία. Συνήθως γίνεται για να γνωρίσουμε το στάδιο της νόσου και να καθορίσουμε το στάδιο της θεραπείας.

1.4.6 Θεραπεία

Οι θεραπείες που ακολουθούνται ενάντια στον καρκίνο της ουροδόχου κύστης είναι συγκεκριμένες και περιλαμβάνουν ακτινοθεραπεία, χημειοθεραπεία, ανοσοθεραπεία και χειρουργική επέμβαση.

- **Ακτινοθεραπεία**: Η ακτινοβολία είναι υψηλής ενέργειας ακτίνες που σκοτώνουν τα καρκινικά αλλά και φυσιολογικά κύτταρα στο πέρασμα τους. Ακτινοβολία μπορεί να δοθεί για μικρούς διηθητικούς καρκίνους της ουροδόχου κύστης. Χρησιμοποιούνται συνήθως ως μια εναλλακτική προσέγγιση στη χειρουργική επέμβαση. Ωστόσο, για μεγαλύτερη θεραπευτική αποτελεσματικότητα θα πρέπει να χορηγείται σε συνδυασμό με χημειοθεραπεία.
- **Χημειοθεραπεία**: Χημειοθεραπεία είναι η χρήση ισχυρών φαρμάκων που σκοπό έχουν να σκοτώσουν τον καρκίνο. Στην περίπτωση του καρκίνου της ουροδόχου κύστης, η χημειοθεραπεία μπορεί να χορηγηθεί μόνη της ή από κοινού με την χειρουργική επέμβαση ή την ακτινοθεραπεία ή και τα δύο. Μπορεί να χορηγείται πριν ή μετά των άλλων θεραπειών. Τα στάδια Ta, T1, και CIS του καρκίνου μπορούν να αντιμετωπιστούν με

ενδοκυστική χημειοθεραπεία. Μετά την αφαίρεση του όγκου ένα ή περισσότερα υγρά φάρμακα εισέρχονται στην ουροδόχο κύστη μέσω ενός καθετήρα. Τα φάρμακα παραμένουν στην κύστη για αρκετές ώρες και στη συνέχεια αποστραγγίζονται έξω, συνήθως κατά την ούρηση. Αυτή η θεραπεία συνήθως επαναλαμβάνεται μία φορά την εβδομάδα για αρκετές εβδομάδες. Καρκίνος που έχει εισβάλει βαθιά μέσα στην ουροδόχο κύστη, τους λεμφαδένες ή άλλα όργανα απαιτεί συστηματική ή ενδοφλέβια χημειοθεραπεία. Τα φάρμακα εισέρχονται με ένεση στην κυκλοφορία του αίματος μέσω μιας φλέβας. Με αυτό τον τρόπο, τα φάρμακα πάνε σχεδόν σε κάθε μέρος του σώματος και, στην ιδανική περίπτωση, σκοτώνουν τα καρκινικά κύτταρα όπου και αν βρίσκονται. Η χημειοθεραπεία είναι γνωστή για τις δυσάρεστες παρενέργειες της. Οι παρενέργειες εξαρτώνται από το ποια φάρμακα λαμβάνει κάποιος αλλά και από τον τρόπο χορήγησης. Παρόλα αυτά όμως είναι σχεδόν πάντα παροδικές και σταματάνε όταν τελειώσει η θεραπεία.

- **Ανοσοθεραπεία**: Η ανοσοθεραπεία εκμεταλλεύεται τη φυσική ικανότητα του οργανισμού να καταπολεμήσει τον καρκίνο. Το ανοσοποιητικό μας σύστημα δημιουργεί ουσίες που αντιμετωπίζουν τους «εισβολείς» όπως τα μη φυσιολογικά κύτταρα (τέτοια είναι τα καρκινικά). Μερικές φορές όμως το ανοσοποιητικό σύστημα δεν μπορεί να αντιμετωπίσει από τα πολύ επιθετικά καρκινικά κύτταρα και εδώ έρχεται η μέθοδος της ανοσοθεραπείας που στόχο έχει την ενδυνάμωση του ανοσοποιητικού συστήματος ενάντια στον καρκίνο. Η ανοσοθεραπεία συνήθως γίνεται μόνο σε στάδια Ta, T_1,CIS καρκίνων της ουροδόχου κύστης (πρόκειται για αρχικούς τύπου καρκίνου).
- **Χειρουργική επέμβαση**: Η χειρουργική επέμβαση είναι μακράν η πιο διαδεδομένη θεραπεία για τον καρκίνο της ουροδόχου κύστης. Χρησιμοποιείται για όλους τους τύπους και τα στάδια του καρκίνου της ουροδόχου κύστης ενώ υπάρχουν αρκετοί διαφορετικοί τρόποι χειρουργικής επέμβασης. Ποια προσέγγιση θα χρησιμοποιηθεί εξαρτάται από το στάδιο στο οποίο βρίσκεται ο καρκίνος. Οι μέθοδοι που ακολουθούνται είναι οι ακόλουθοι.
    1. **Διουρηθρική εκτομή με ηλεκτροπηξία**: Σε αυτή την περίπτωση μέσω της ουρήθρας φτάνει στην κύστη ένα όργανο

(ρεζεκτοσκόπειο) το οποίο στην άκρη του έχει ένα εργαλείο κοπής το οποίο κόβει ή καίει τον όγκο με την χρήση ηλεκτρικού ρεύματος. Η μέθοδος γίνεται συνήθως για την αρχική διάγνωση του καρκίνου της ουροδόχου κύστης και για την αντιμετώπιση των σταδίων Ta και T_1. Συχνά, μετά από διουρηθρική εκτομή, πρόσθετη θεραπεία είναι δεδομένη για να βοηθήσει στην αντιμετώπιση του καρκίνου.

2. **Ριζική κυστεκτομή**: Σε αυτή την περίπτωση, το σύνολο της ουροδόχου κύστης έχει αφαιρεθεί, καθώς και οι γύρω από αυτή λεμφαδένες όπως και άλλες δομές που μπορεί να περιέχουν τον καρκίνο. Αυτό γίνεται συνήθως για καρκίνους που έχουν τουλάχιστον εισβάλει στο μυϊκό στρώμα του τοιχώματος της ουροδόχου κύστης ή για επιφανειακές μορφές καρκίνου που εκτείνονται στο μεγαλύτερο μέρος της ουροδόχου κύστης ή που απέτυχαν να ανταποκριθούν σε πιο συντηρητικές θεραπείες.
3. **Τμηματική ή μερική κυστεκτομή**: Σε αυτή την περίπτωση, μέρος της ουροδόχου κύστης έχει αφαιρεθεί. Αυτό γίνεται συνήθως για χαμηλού βαθμού όγκους που έχουν εισβάλλει στο τοίχωμα της ουροδόχου κύστης, αλλά περιορίζονται σε μια μικρή περιοχή της ουροδόχου κύστης.

1.4.7 Αποτελέσματα

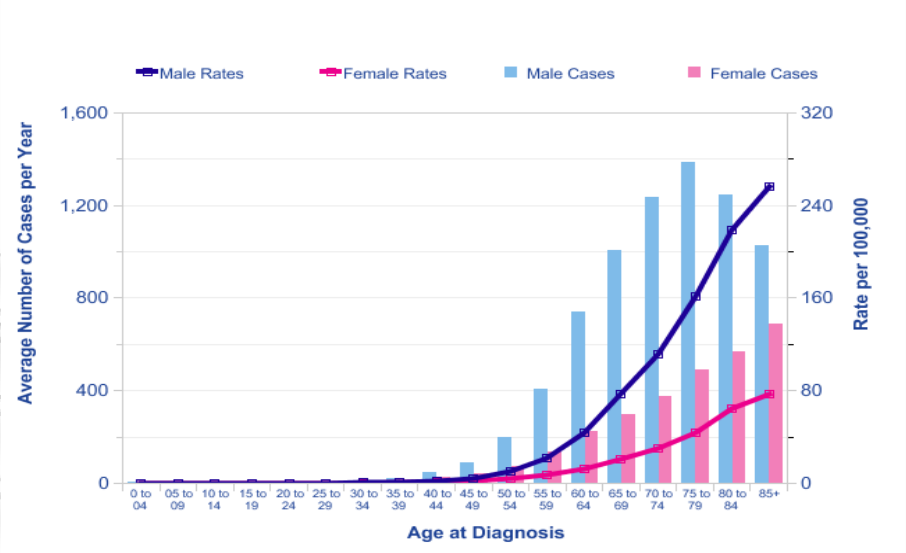
Το προφίλ των ασθενών με καρκίνο της ουροδόχου κύστης διαφέρει σημαντικά ανάλογα με το στάδιο του καρκίνου την στιγμή που έγινε η διάγνωση. Περίπου στο 90% των περιπτώσεων όπου διαγνώστηκε επιφανειακός καρκίνος (Ta, T1, CIS) οι ασθενείς ζήσανε για τουλάχιστον 5 χρόνια μετά την θεραπεία. Εάν ο καρκίνος είναι επιθετικός αλλά δεν έχει επεκταθεί έξω από την κύστη τότε η αντίστοιχη πιθανότητα πέφτει στο 75%. Εάν ο καρκίνος έχει προσβάλει τους λεμφαδένες ή κοντινά όργανα το ποσοστό πέφτει ακόμα περισσότερο στο 36% ενώ τέλος αν ο καρκίνος έχει προσβάλει μακρινά όργανα η πιθανότητα επιβίωσης για 5 χρόνια είναι μόλις 6%. Τέλος οι νέοι καρκίνοι που δημιουργούνται είναι πιο επιθετικοί από τους αρχικούς, οπότε οι πιθανότητα για μακροπρόθεσμη επιβίωση ασθενούς με καρκίνο μεγάλου βαθμού ή προχωρημένου σταδίου είναι περιορισμένη. Αντίθετα οι νέοι καρκίνοι που είναι επιφανειακοί και χαμηλού βαθμού σπάνια είναι απειλητικοί για την ζωή του ασθενούς (http://www.cancer.net, http://www.emedicinehealth.com)

1.4.8 Στατιστικά στοιχεία

Στο τελευταίο μέρος του πρώτου κεφαλαίου θα παρουσιάσουμε κάποια στατιστικά στοιχεία γύρω από τον καρκίνο της ουροδόχου κύστης. Οι πίνακες αναφέρονται σε περιστατικά στην Μεγάλη Βρετανία για το έτος 2008, πηγή αποτελεί η ιστοσελίδα http://info.cancerresearchuk.org, τα στοιχεία που αντλούμε όμως συμφωνούν με τα αποτελέσματα από αντίστοιχες έρευνες που έχουν γίνει και σε άλλες χώρες, συνεπώς μας βοηθούν στο να αποκτήσουμε μια καλύτερη εικόνα βασισμένη σε πραγματικά δεδομένα.

- Η πιθανότητα να αναπτύξει κάποιος καρκίνο της ουροδόχου κύστης είναι πάνω από 2 φορές μεγαλύτερη στους άντρες από ότι στις γυναίκες. Από την άλλη πλευρά είναι πιο πιθανό στις γυναίκες η διάγνωση να δείξει ότι πάσχουν από πιο προχωρημένη μορφή καρκίνου από ότι οι άντρες.
- Ο καρκίνος της ουροδόχου κύστης μπορεί να εμφανιστεί σε οποιαδήποτε ηλικία αλλά το πιο πιθανό είναι να εμφανιστεί σε άτομα άνω των 50 ετών. Η μέση ηλικία διάγνωσης είναι γύρω στα 60 έτη ενώ είναι πιθανό να εμφανίσουν καρκίνο και άτομα άνω των 80 ετών.

Πίνακας 1.4.8.1 (Μέσος αριθμός νέων κρουσμάτων ανά έτος και ειδικοί ηλικιακοί δείκτες ανά 100.000 άτομα, UK)



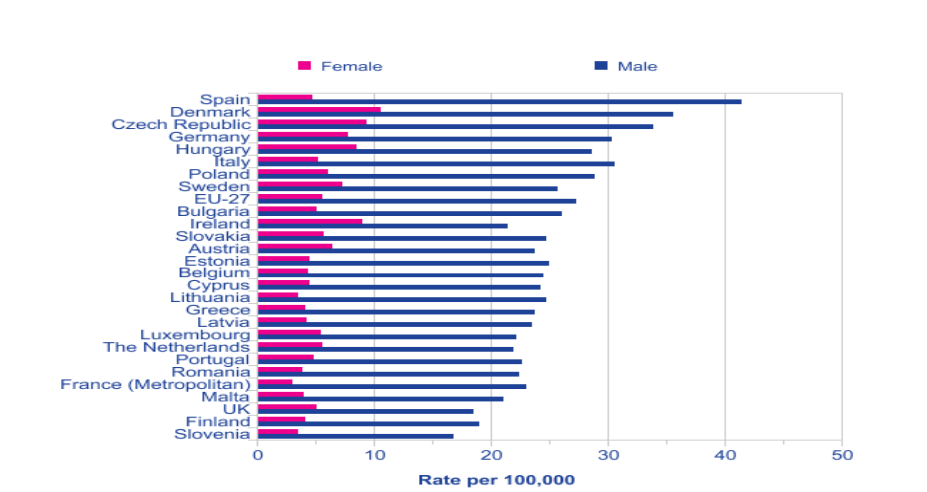
(Πηγή: http://info.cancerresearchuk.org)

- Μελέτες έχουν δείξει ότι ανεξαρτήτως φύλου οι λευκοί είναι πιο πιθανό να αναπτύξουν καρκίνο της ουροδόχου κύστης σε σχέση με όλες τις άλλες φυλές, μάλιστα η πιθανότητα είναι περίπου διπλάσια. Αυτές οι διαφορές

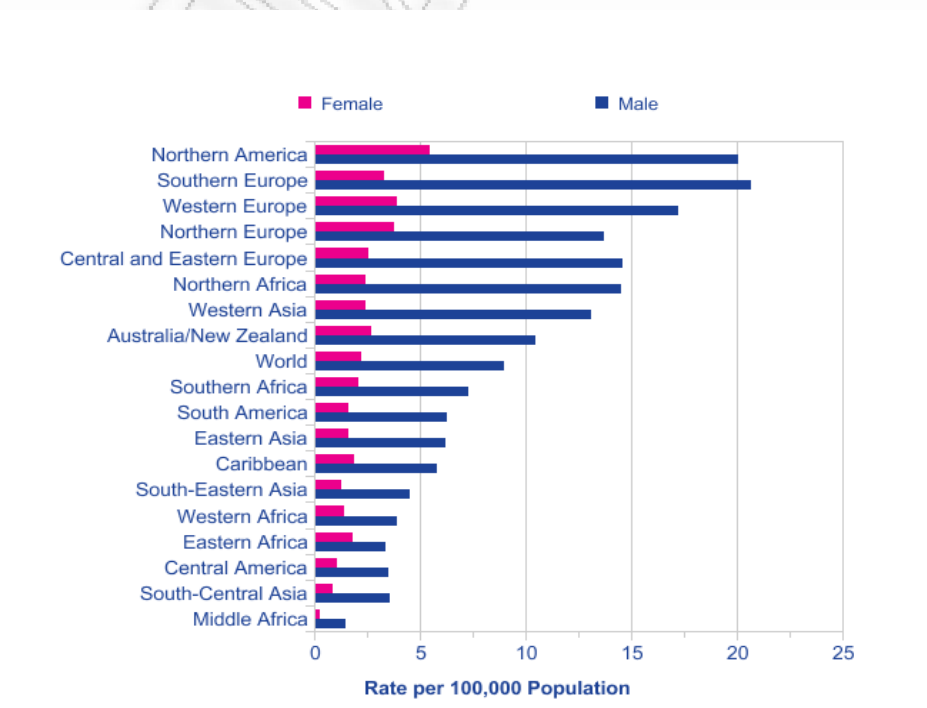
κατά πάσα πιθανότητα οφείλονται στα γονίδια καθώς έρευνες έχουν δείξει ότι διάφοροι γενετικοί παράγοντες που συνδέονται με τον καρκίνο της ουροδόχου κύστης, όπως για παράδειγμα το γονίδιο NAT2, υπάρχει σε μεγαλύτερες ποσότητες στους λευκούς ανθρώπους από ότι στους μη λευκούς.

- Στα παρακάτω διαγράμματα παρουσιάζονται οι εκτιμήσεις για τους ρυθμούς εμφάνισης του καρκίνου στην Ευρώπη των 27 αλλά και παγκόσμια για το 2008.

Πίνακας 1.4.8.2 (Τα ποσοστά εμφάνισης του καρκίνου στην Ευρώπη των 27)



Πίνακας 1.4.8.3 (Τα ποσοστά εμφάνισης του καρκίνου σε παγκόσμια κλίμακα)



(Πηγή: http://info.cancerresearchuk.org)

Εδώ φαίνεται καθαρά ότι τα περιστατικά είναι περισσότερα για τους άντρες από ότι για τις γυναίκες. Επίσης βλέπουμε ότι τα ποσοστά είναι χαμηλότερα για την Ελλάδα σε σχέση με τον μέσο όρο των 27 χωρών. Πιο συγκεκριμένα για τους άντρες ο μέσος ρυθμός εμφάνισης είναι 23.7 ανά 100.000 άτομα του πληθυσμού και για τις γυναίκες 4.1, ενώ οι αντίστοιχοι ρυθμοί για την Ευρώπη των 27 είναι 27.4 και 5.6 αντίστοιχα. Σχετικά με την παγκόσμια κατανομή που κρουσμάτων βλέπουμε ότι οι μεγαλύτεροι ρυθμοί εμφανίζονται στις βιομηχανικά ανεπτυγμένες χώρες και ειδικότερα στην βόρεια Αμερική, στην νότια και δυτική Ευρώπη αλλά και στις περιοχές της Ασίας και της Αφρικής που σχετίζονται με την λοίμωξη της σχιστοσωμίασης.

ΚΕΦΑΛΑΙΟ 2

Γραμμική παλινδρόμηση

Ένα από τα πιο συνηθισμένα εργαλεία για την εύρεση, εάν υπάρχει, και μοντελοποίηση της σχέσης μεταξύ της μεταβλητής απόκρισης και κάποιας ανεξάρτητης μεταβλητής είναι μέσω του εργαλείου της γραμμικής παλινδρόμησης. Η παρουσίαση έχει βασιστεί στις σημειώσεις του μαθήματος «Ανάλυση παλινδρόμησης και ανάλυση διακύμανσης» (Κούτρας ΠΜΣ 2011).

2.1 Απλή γραμμική παλινδρόμηση

Αρχικά θα παρουσιάσουμε την απλή περίπτωση, εδώ αυτό που έχουμε είναι η μεταβλητή απόκρισης και μία ανεξάρτητη μεταβλητή.

2.1.1 Το μοντέλο

Το μοντέλο που χρησιμοποιείται στην γραμμική παλινδρόμηση είναι το κανονικό γραμμικό μοντέλο το οποίο δίνεται από την παρακάτω συνάρτηση και ισχύουν τα εξής.

$$Y_i = \beta_0 + \beta_1 * X_i + \varepsilon_i$$

Y_i : Τιμή της εξαρτημένης μεταβλητής για το i ζεύγος παρατηρήσεων (τυχαία μεταβλητή).

β_0, β_1 : άγνωστες παράμετροι

X_i : Τιμή της ανεξάρτητης μεταβλητής (γνωστή – μη τυχαία μεταβλητή)

$\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n$: Τυχαία σφάλματα (τυχαίες μεταβλητές) που κατανέμονται σύμφωνα με την κανονική κατανομή με μέση τιμή 0 και σταθερή διακύμανση σ^2 (άγνωστη) και είναι ανεξάρτητα μεταξύ τους.

Από την στιγμή που $\varepsilon_i \sim N(0,\sigma^2)$ τότε ισχύει ότι $Y_i \sim N(\beta_0 + \beta_1 * X_i, \sigma^2)$ και επιπλέον τα Y_i θα είναι και ανεξάρτητα.

Οι εκτιμήτριες ελαχίστων τετραγώνων των παραμέτρων β_0, β_1 ξέρουμε ότι δίνονται από τις σχέσεις:

$$\hat{\beta}_1 = \sum_{i=1}^{n} k_i * Y_i, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 * \bar{X} = \sum_{i=1}^{n} \lambda_i * Y_i \ ,$$

όπου $k_i = \frac{X_i - \bar{X}}{S_{xx}}$ και $\lambda_i = \frac{1}{n} - k_i * \bar{X}$, οπότε προκύπτει ότι για τις εκτιμήσεις ισχύει ότι ακολουθούν την κανονική κατανομή με τα εξής χαρακτηριστικά.

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \quad \text{και} \quad \hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 * \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)\right),$$

όπου $S_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2$.

2.1.2 Μεταβλητότητα

Στο γραμμικό μοντέλο παλινδρόμησης η ολική μεταβλητότητα των παρατηρήσεων Y_i δίνεται από την σχέση:

$$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2, \qquad \text{(total sum of squares)}$$

Βλέπουμε δηλαδή ότι η ολική μεταβλητότητα αναλύεται σε 2 μέρη:
Αυτό που ερμηνεύεται από την ευθεία παλινδρόμησης

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \text{, (regression sum of squares)}$$

και ένα άλλο

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \text{, (error sum of squares)}$$

που μένει ανερμήνευτο και φανερώνει πόσο μεγάλη είναι η απόκλιση των πραγματικών δεδομένων από τα εκτιμώμενα που προκύπτουν σύμφωνα με το μοντέλο.

Επειδή στόχος μας είναι η όσο το δυνατόν μικρότερη ανερμήνευτη μεταβλητότητα, η ευθεία παλινδρόμησης θα είναι ικανοποιητική εάν το ποσοστό της ολικής μεταβλητότητας που ερμηνεύεται είναι όσο το δυνατόν μεγαλύτερο, δηλαδή ο λόγος

$$R^2 = \frac{SSR}{SSTO} = \frac{SSTO - SSE}{SSTO}$$

να είναι όσο το δυνατόν μεγαλύτερος. Η ποσότητα R^2 ονομάζεται συντελεστής προσδιορισμού (coefficient of determination) και παίρνει τιμές $0 \leq R^2 \leq 1$.

Πέρα από τα αθροίσματα τετραγώνων υπάρχουν και τα αντίστοιχα μέσα αθροίσματα τετραγώνων που προκύπτουν από την διαίρεση του κάθε αθροίσματος τετραγώνων με τους βαθμούς ελευθερίας που του αναλογούν. Έτσι έχουμε τα εξής:

$$\text{MSTO} = \frac{\text{SSTO}}{\text{n}-1}, \qquad \text{MSR} = \frac{\text{SSR}}{1} = \text{SSR}, \qquad \text{MSE} = \frac{\text{SSE}}{\text{n}-2}$$

ενώ το MSE είναι αμερόληπτη εκτιμήτρια του σ^2 (που στην πράξη συνήθως είναι άγνωστο).

2.1.3 Έλεγχος της υπόθεσης της μηδενικής κλίσης

Στα γραμμικά μοντέλα το βασικό μας ενδιαφέρον είναι η μελέτη για την ύπαρξη γραμμικής σχέσης μεταξύ της μεταβλητής απόκρισης (Y) και της ανεξάρτητης μεταβλητής (X). η έρευνα αυτή γίνεται μέσω των ελέγχων υποθέσεων σχετικά με την κλίση β_1, καθώς εάν η κλίση είναι 0 αυτό σημαίνει ότι οποιαδήποτε μεταβολή του X δεν επηρεάζει το Y. Πράγματι εάν $\beta_1=0$, το γραμμικό μοντέλο

$$\text{Y}_\text{i} = \beta_0 + \beta_1 * \text{X}_\text{i} + \varepsilon_\text{i}$$

γίνεται $Y_i = \beta_0 + \varepsilon_i$ όπου $\varepsilon_\text{i} \sim \text{N}(0,\sigma^2)$, δηλαδή $\text{Y}_\text{i} \sim \text{N}(\beta_0, \sigma^2)$ που σημαίνει ότι οι κατανομές των Y_i, i=1,2,…,n είναι ίδιες ανεξάρτητα από τις τιμές που παίρνει το X. Η υπόθεση που μας ενδιαφέρει να ελέγξουμε είναι η παρακάτω:

H_0: $\beta_1=0$ (δεν υπάρχει γραμμική σχέση μεταξύ των X και Y)

H_1: $\beta_1\neq0$ (υπάρχει γραμμική σχέση μεταξύ των X και Y –αμφίπλευρος έλεγχος)

Ο έλεγχος γίνεται σε επίπεδο σημαντικότητας α% και βασίζεται στην στατιστική συνάρτηση

$$\text{T} = \frac{\hat{\beta}_1}{\text{s}(\hat{\beta}_1)}$$

η οποία κάτω από την μηδενική υπόθεση ακολουθεί κατανομή t με n-2 βαθμούς ελευθερίας. Εάν $|T|>t_{n-2}(\alpha/2)$ τότε απορρίπτουμε την H_0 υπέρ της εναλλακτικής.

Ένας δεύτερος τρόπος για να ελέγξουμε αν υπάρχει γραμμική σχέση μεταξύ των X και Y είναι μέσω του F ελέγχου. Η στατιστική συνάρτηση που χρησιμοποιείται είναι η

$$\text{F} = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/1}{\text{SSE}/\text{n}-2}$$

και απορρίπτουμε την H_0 αν $F>F_{1,n-2}(\alpha)$.

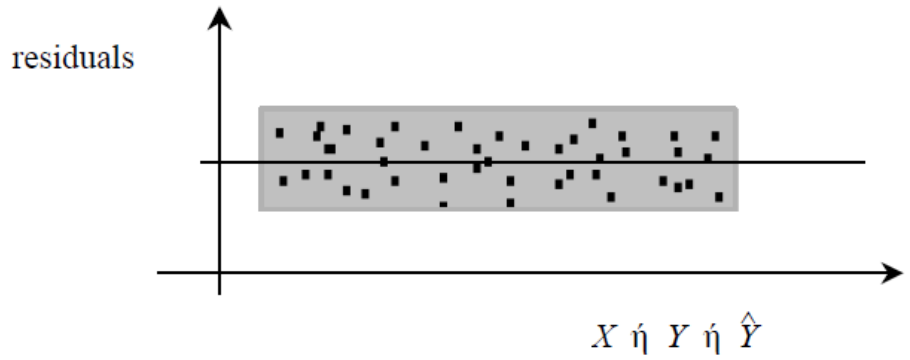
2.1.4 Αποκλίσεις από το απλό γραμμικό μοντέλο

Πολλές φορές μπορεί να συναντήσουμε περιπτώσεις όπου θα υπάρχουν αποκλίσεις από το γραμμικό. Οι πιο συνήθεις παρουσιάζονται στον παρακάτω πίνακα.

Πίνακας 2.1 (Αποκλίσεις από το γραμμικό μοντέλο)

	Υπόθεση που παραβιάζεται
Μη γραμμική παλινδρόμηση	$E(Y_i)= \beta_0 + \beta_1 * X_i$
Μη σταθερή διακύμανση των ε_i (ή ισοδύναμα των Y_i)	$V(\varepsilon_i)=V(Y_i)= \sigma^2$
Ύπαρξη εξάρτησης μεταξύ των σφαλμάτων ε_i (ή ισοδύναμα των Y_i)	$Cov(\varepsilon_i,\varepsilon_j)=0$ $i\neq j$ $Cov(Y_i,Y_j)=0$ $i\neq j$
Τα σφάλματα ε_i (ή ισοδύναμα των Y_i) δεν ακολουθούν κανονική κατανομή	$\varepsilon_i \sim N(0,\sigma^2)$ $Y_i \sim N(\beta_0 + \beta_1 * X_i, \sigma^2)$
Ύπαρξη εκτρόπων παρατηρήσεων (outliers)	
Παράλειψη σημαντικών ανεξάρτητων μεταβλητών.	

Ο έλεγχος γίνεται μέσω των καταλοίπων (residuals) $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ (τυχαίες μεταβλητές). Εάν το γραμμικό μοντέλο είναι σωστό τότε το διάγραμμα διασποράς των υπολοίπων έχει την παρακάτω μορφή.



Παρατηρούμε δηλαδή ότι το σχήμα μοιάζει με ένα παραλληλόγραμμο γύρω από το μηδέν. Οποιαδήποτε διαφοροποίηση από το σχήμα αυτό σημαίνει ότι κατά πάσα πιθανότητα υπάρχουν αποκλίσεις από το γραμμικό μοντέλο και κατά συνέπεια τα αποτελέσματα που παίρνουμε δεν είναι τα σωστά. Στην περίπτωση που καταλήξουμε στο ότι πράγματι υπάρχουν αποκλίσεις από το γραμμικό μοντέλο τότε προσπαθούμε μέσω μετασχηματισμών των δεδομένων να τα φέρουμε σε μια μορφή που θα ικανοποιούν τις προϋποθέσεις του γραμμικού μοντέλου.

2.2 Πολλαπλή γραμμική παλινδρόμηση

Αποτελεί ουσιαστικά την επέκταση της απλής παλινδρόμησης σε περιπτώσεις που μας ενδιαφέρει η εξέταση της σχέσης μεταξύ της μεταβλητής απόκρισης Y και κάποιων ανεξάρτητων μεταβλητών $X_1,X_2,\ldots,X_{\rho-1}$.

2.2.1 Το μοντέλο

Η εξίσωση που δίνει τον τύπο του μοντέλου είναι η ακόλουθη:

$$Y_i = \beta_0 * X_{i0} + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \cdots + \beta_{p-1} * X_{i,p-1} + \varepsilon_i, \qquad i = 1,2,\dots,n$$

Y_i : τυχαιές μεταβλητές

β_i : άγνωστες παράμετροι

X_{ij} : μη τυχαίες μεταβλητές (Η X_{i0} είναι μια πλασματική μεταβλητή που παίρνει πάντα την τιμή 1)

$\varepsilon_1,\varepsilon_2,\dots,\varepsilon_n$: Τυχαία σφάλματα (τυχαίες μεταβλητές) που κατανέμονται σύμφωνα με την κανονική κατανομή με μέση τιμή 0 και σταθερή διακύμανση σ^2 (άγνωστη) και είναι ανεξάρτητα μεταξύ τους.

Επίσης αφού $E(\varepsilon_i)=0$ και $V(\varepsilon_i)=\sigma^2$ θα ισχύει ότι:

$$E(Y_i)= \beta_0 * X_{i0} + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \cdots + \beta_{p-1} * X_{i,p-1}$$

$$V(Y_i)=\sigma^2$$

Η εξίσωση μπορεί να γραφτεί πιο απλά στην πινακική της μορφή ως εξής:

$$\boldsymbol{\Upsilon} = \mathbf{X} * \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Όπου $\boldsymbol{\Upsilon} = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_n \end{pmatrix}$, $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{pmatrix}$ και $\mathbf{X} = \begin{bmatrix} X_{10} & X_{11} & \cdots & X_{1,p-1} \\ X_{20} & X_{21} & & X_{2,p-1} \\ \vdots & & \ddots & \vdots \\ X_{n0} & X_{n1} & \cdots & X_{n,p-1} \end{bmatrix}$

Και $E(\boldsymbol{\varepsilon})=0$, $E(\mathbf{Y})=\mathbf{X}*\boldsymbol{\beta}$, $D(\boldsymbol{\varepsilon})=D(\mathbf{Y})=\sigma^2*\mathbf{I}$

Οι εκτιμήτριες ελαχίστων τετραγώνων των συντελεστών δίνονται από την σχέση $\widehat{\boldsymbol{\beta}}=(\mathbf{X'X})^{-1}*\mathbf{X'Y}$ και ισχύει ότι είναι αμερόληπτες εκτιμήτριες των αντίστοιχων παραμέτρων που εκτιμούν, δηλαδή $E(\widehat{\boldsymbol{\beta}})=\boldsymbol{\beta}$.

2.2.2 Αθροίσματα τετραγώνων και πρόσθετα αθροίσματα τετραγώνων

Όμοια με το απλό γραμμικό μοντέλο ορίζονται τα αθροίσματα τετραγώνων και τα αντίστοιχα μέσα αθροίσματα.

$$MST = \frac{SST}{n-1} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1} = \frac{\mathbf{Y'Y} - n * \overline{Y}^2}{n-1},$$

$$MSR = \frac{SSR}{p-1} = \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}{p-1} = \frac{\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n * \overline{Y}^2}{p-1},$$

$$MSE = \frac{SSE}{n-p} = \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n-p} = \frac{\mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}}{n-p}$$

Η τελευταία ποσότητα αποτελεί την αμερόληπτη εκτιμήτρια s^2 της άγνωστης διακύμανσης σ^2, είτε τα σφάλματα ακολουθούν την κανονική κατανομή είτε όχι.

Πέρα από τα αθροίσματα τετραγώνων στην πολλαπλή γραμμική παλινδρόμηση υπάρχει και η έννοια των πρόσθετων αθροισμάτων τετραγώνων. Για να γίνει εύκολα κατανοητή η έννοια αυτή ας θεωρήσουμε την περίπτωση που έχουμε 2 ανεξάρτητες μεταβλητές X_1 και X_2. Το πρόσθετο άθροισμα τετραγώνων $SSR(X_2|X_1)$ εκφράζει το επιπλέον τμήμα της συνολικής μεταβλητότητας που ερμηνεύεται με την χρήση της ανεξάρτητης μεταβλητής X_2 σε ένα μοντέλο που ήδη έχει την ανεξάρτητη μεταβλητή X_1 ή αντίστροφα εκφράζει την ελλάτωση της ανερμήνευτης μεταβλητότητας των Y που επιτυγχάνεται όταν χρησιμοποιήσουμε την X_2 σε ένα μοντέλο που ήδη έχει την μεταβλητή X_1. Μαθηματικά αυτό φαίνεται από την παρακάτω σχέση:

$$SSR(X_2|X_1) = SSR(X_1,X_2) - SSR(X_1) = SSE(X_1) - SSE(X_1,X_2)$$

Όπως και στην απλή περίπτωση έτσι και τώρα υπάρχει ο συντελεστής προσδιορισμού R^2 ενώ επιπλέον τώρα υπάρχουν και ο συντελεστής μερικού προσδιορισμού που αναφέρεται στα πρόσθετα αθροίσματα τετραγώνων και δίνεται από την σχέση

$$R^2_{Y2,1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$$

Ο συντελεστής αυτός εκφράζει το ποσοστό της μεταβλητότητας που έχει μείνει ανερμήνευτη μετά την χρήση της ανεξάρτητης μεταβλητής X_1, την οποία εξηγεί η προσθήκη στο μοντέλο και της μεταβλητής X_2. Σύμφωνα με αυτό ένα μικρό ποσοστό θα μπορούσε να μας βάλει σε σκέψεις κατά πόσο είναι αναγκαίο να προσθέσουμε την μεταβλητή X_2 σε ένα μοντέλο που έχει ήδη την ανεξάρτητη μεταβλητή X_1, ενώ το ντίθετο συμβαίνει όταν το ποσοστό αυτό είναι μεγάλο δηλαδή είναι απαραίτητο να μπει. Τέλος, η έννοια του πρόσθετου αθροίσματος τετραγώνων μπορεί εύκολα να γενικευτεί στην περίπτωση που έχουμε περισσότερες των δύο ανεξάρτητες μεταβλητές απλώς κάθε φορά πρέπει να ξέρουμε ποιες από αυτές αποτελούν την βάση (δηλαδή είναι ήδη μέσα στο μοντέλο) και ποιες σκοπεύουμε να προσθέσουμε.

2.2.3 Έλεγχοι υποθέσεων

Έστω ότι στο γενικό γραμμικό μοντέλο $\mathbf{Y} = \mathbf{X} * \boldsymbol{\beta} + \boldsymbol{\varepsilon}$ με κανονικά ανεξάρτητα σφάλματα $\boldsymbol{\varepsilon} \sim N_n(0,\sigma^2\mathbf{I_n})$, θεωρούμε έναν πίνακα A διάστασης r x p και ένα διάνυσμα

στήλη **c** διάστασης r x 1 (των οποίων τα στοιχεία είναι σταθεροί αριθμοί). Τότε ο έλεγχος της υπόθεσης H_0: $A\boldsymbol{\beta}=\mathbf{c}$ έναντι της H_1: $A\boldsymbol{\beta}\neq\mathbf{c}$ σε επίπεδο σημαντικότητας α γίνεται μέσω της συνάρτησης

$$F = \frac{n-p}{r} * \frac{\left(A\widehat{\boldsymbol{\beta}}-\mathbf{c}\right)'(A(X'X)^{-1}A')^{-1}(A\widehat{\boldsymbol{\beta}}-\mathbf{c})}{\left(\mathbf{Y}-X\widehat{\boldsymbol{\beta}}\right)'(\mathbf{Y}-X\widehat{\boldsymbol{\beta}})}$$

Απορρίπτουμε την μηδενική υπόθεση εάν ισχύει $F>F_{r,n-p}(\alpha)$. οι σημαντικότεροι έλεγχοι που χρησιμοποιούνται στην πράξη και αναφέρονται παρακάτω αποτελούν ειδική περίπτωση του παραπάνω αποτελέσματος.

1. Κρίσιμη περιοχή του ελέγχου της γραμμικής υπόθεσης H_0: $\mathbf{a'\beta}=\mathbf{c}$ έναντι της H_1: $\mathbf{a'\beta}\neq\mathbf{c}$ όπου $\mathbf{a}=(a_1,a_2,\ldots,a_{p-1})$

$$K: \left|\frac{\mathbf{a}'\widehat{\boldsymbol{\beta}}-\mathbf{c}}{s\sqrt{\mathbf{a'}(X'X)^{-1}\mathbf{a}}}\right| > t_{n-p}(\alpha/2)$$

2. Κρίσιμη περιοχή του ελέγχου της υπόθεσης H_0: $\beta_1=\beta_2=\ldots=\beta_{p-1}=0$ έναντι της H_1: κάποιο από τα β_i δεν είναι ίσο από με μηδέν

$$F^* = \left(\frac{n-p}{p-1}\right)\left(\frac{\widehat{\boldsymbol{\beta}}\mathbf{'X'Y}-n*\bar{Y}^2}{\mathbf{Y'Y}-\widehat{\boldsymbol{\beta}}\mathbf{'X'Y}}\right) = \frac{MSR}{MSE} > F_{p-1,n-p}(a)$$

Πέρα από τους δύο αυτούς ελέγχους έχουμε την δυνατότητα να πραγματοποιήσουμε ακόμα έναν με την βοήθεια των πρόσθετων αθροισμάτων τετραγώνων. Αυτός είναι ο έλεγχος της υπόθεσης H_0: $\beta_q=\beta_{q+1}=\ldots=\beta_{p-1}=0$ έναντι της H_1: $\beta_i\neq0$ για κάποιο i. Για να κάνουμε τον έλεγχο αυτό δημιουργούμε αρχικά το πλήρες μοντέλο

$$Y_i = \beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \cdots + \beta_{p-1} * X_{i,p-1} + \varepsilon_i, \qquad i = 1,2,\ldots,n$$

Και το περιορισμένο:

$$Y_i = \beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \cdots + \beta_{q-1} * X_{i,q-1} + \varepsilon_i, \qquad i = 1,2,\ldots,n$$

Στην συνέχεια υπολογίζουμε την στατιστική συνάρτηση:

$$F^* = \frac{\frac{SSR(X_q, X_{q+1}, \ldots, X_{p-1} | X_1, X_2, \ldots, X_{q-1})}{p-q}}{\frac{SSE(X_1, X_2, \ldots, X_{p-1})}{n-p}} =$$

$$= \frac{\frac{SSR(X_1, X_2, \ldots, X_{p-1}) - SSR(X_1, X_2, \ldots, X_{q-1})}{p-q}}{MSE(\text{Full Model})} \sim F_{p-q,n-p}$$

Απορρίπτουμε την H_0 εάν ισχύει $F^* > F_{p-q,n-p}(\alpha)$.

2.2.4 Πολυσυγγραμμικότητα

Πολύ συχνά παρατηρείται το φαινόμενο οι ανεξάρτητες μεταβλητές που χρησιμοποιούμε στα μοντέλα πολλαπλής παλινδρόμησης να εμφανίζουν μεγάλη συσχέτιση μεταξύ τους, αυτό το φαινόμενο ονομάζεται πολυσυγγραμικότητα και θα πρέπει να είμαστε πολύ προσεχτικοί στην εξαγωγή των συμπερασμάτων καθώς μπορεί να καταλήξουμε σε λάθος αποτελέσματα εξαιτίας των λανθασμένων τιμών που προκύπτουν από την ανάλυση. Σε περίπτωση που υπάρχει πολυσυγγραμμικότητα αυτή έχει κάποιες συνέπεις στα προσαρμοσμένα μοντέλα:

1. Οι προβλέψεις των τιμών της μεταβλητής απόκρισης Y με το πλήρες μοντέλο είναι περίπου ίδιες με τις προβλέψεις που προκύπτουν όταν αφαιρέσουμε μία ή περισσότερες εξαρτημένες μεταβλητές.
2. Το άθροισμα τετραγώνων των καταλοίπων (SSE) του πλήρους μοντέλου δεν διαφέρει σημαντικά από το άθροισμα τετραγώνων των καταλοίπων του μοντέλου που προκύπτει όταν αφαιρέσουμε μία ή περισσότερες εξαρτημένες μεταβλητές.
3. Τα πρόσθετα αθροίσματα τετραγώνων που αντιστοιχούν στην εισαγωγή κάποιας ανεξάρτητης μεταβλητής έχουν μικρές τιμές σε σχέση με την μεταβλητότητα που έμεινε ανερμήνευτη μετά τη διαμόρφωση ενός μη πλήρους μοντέλου.
4. Τα τυπικά σφάλματα των εκτιμητριών ελαχίστων τετραγώνων για το πλήρες μοντέλο είναι μεγαλύτερα από τα τυπικά σφάλματα των εκτιμητριών των ίδιων παραμέτρων σε ένα μη πλήρες μοντέλο.
5. Τα διαστήματα εμπιστοσύνης των παραμέτρων για το πλήρες μοντέλο είναι πολύ πλατύτερα από τα αντίστοιχα διαστήματα για τις ίδιες παραμέτρους σε ένα μη πλήρες μοντέλο.
6. Όταν από ένα μοντέλο πολλαπλής παλινδρόμησης αφαιρεθεί μια ανεξάρτητη μεταβλητή η οποία είναι υψηλά συσχετισμένη με κάποια ή κάποιες άλλες τότε μεταβάλλεται σημαντικά ο συντελεστής της μεταβλητής ή των μεταβλητών που είναι υψηλά συσχετισμένος με αυτήν που αφαιρέθηκε.

Ένας δείκτης που έχει προταθεί ως διαγνωστικό κριτήριο για την ύπαρξη πολυσυγγραμμικότητας είναι ο παράγοντας διόγκωσης διακύμανσης (variance

inflation factor). Ο δείκτης αυτός ορίζεται για κάθε ανεξάρτητη μεταβλητή X_k k=1,2,…,p-1 ενός μοντέλου πολλαπλής παλινδρόμησης με p παραμέτρους, μέσω του τύπου $\mathrm{VIF_k} = \frac{1}{1-\mathrm{R^2_k}}, \mathrm{k} = 1,2, \dots, \mathrm{p}-1$,

όπου R^2_k είναι ο συντελεστής προσδιορισμού του μοντέλου που χρησιμοποιεί ως εξαρτημένη μεταβλητή την X_k και ως ανεξάρτητες τις υπόλοιπες p-2 .

- Αν $VIF_k \cong 1$, η αντίστοιχη ανεξάρτητη μεταβλητή X_k δεν έχει πρόβλημα πολυσυγγραμμικότητας σε σχέση με τις υπόλοιπες.
- Αν VIF_k >10 τότε η $X_κ$ εμφανίζει πρόβλημα πολυσυγγραμμικότητας σε σχέση με τις υπόλοιπες ανεξάρτητες μεταβλητές.

Ως ένα ομοιόμορφο κριτήριο για την μη ύπαρξη πολυσυγγραμμικότητας στο σύνομο των δεδομένων έχει προταθεί η χρήση του μέσου όρου $\overline{\mathrm{VIF}} = \frac{1}{\mathrm{p}-1}\sum_1^{\mathrm{p}-1}\mathrm{VIF_k}$ και αν η ποσότητα $\overline{VIF}$ πάρει τιμή αρκετά πάνω από το 1 έχουμε ένδειξη πολυσυγγραμμικότητας (Κούτρας - Ευαγγελάρας 2010).

2.2.5 Επιλογή βέλτιστου συνόλου ανεξάρτητων μεταβλητών

Πολλές φορές υπάρχουν περιπτώσεις στις οποίες αν και έχουμε ή μπορούμε να έχουμε ένα μεγάλο πλήθος από ανεξάρτητες μεταβλητές θα μας ενδιέφερε η επιλογή ενός μικρού υποσυνόλου από αυτές για την δημιουργία ενός αποτελεσματικού μοντέλου πρόβλεψης. Στην συνέχεια της παραγράφου αυτής θα παρουσιαστούν διάφορες μέθοδοι για την επιλογή των ανεξάρτητων μεταβλητών.

Η $1^η$ μέθοδος που χρησιμοποιείται είναι η εξέταση όλων των μοντέλων και η απόφαση για το πιο κατάλληλο με την χρήση κάποιου ή κάποιων κριτηρίων. Τα πιο συνηθισμένα είναι τα εξής:

1. Το κριτήριο R^2: σύμφωνα με το κριτήριο αυτό καλύτερο είναι το μοντέλο που έχει τον μεγαλύτερο συντελεστή προσδιορισμού. Το μειονέκτημα του κριτηρίου αυτού είναι ότι όσο αυξάνει το πλήθος των ανεξάρτητων μεταβλητών που υπάρχουν στο μοντέλο το R^2 αυξάνει και αυτό άρα θα καταλήγαμε κάθε φορά ότι το καλύτερο μοντέλο θα είναι το πλήρες.
2. Το κριτήριο R^2_{adj}: για να αντιμετωπιστεί το παραπάνω μειονέκτημα έχει προταθεί ο τροποποιημένος συντελεστής προσδιορισμού που δίνεται από την σχέση:

$$R^2{}_{adj} = 1 - \frac{\frac{SSE}{n-p}}{\frac{SSTO}{n-1}} = 1 - \frac{MSE}{MSTO} = 1 - \frac{n-1}{n-p}(1 - R^2)$$

Σύμφωνα με την παραπάνω σχέση το βέλτιστο πλήθος μεταβλητών επιτυγχάνεται στο σημείο όπου ο δείκτης παίρνει την μέγιστη τιμή του.

3. Το κριτήριο MSE: σύμφωνα με το κριτήριο αυτό ως βέλτιστο αριθμό ανεξάρτητων μεταβλητών θεωρούμε αυτό που ελαχιστοποιεί το μέσο άθροισμα τετραγώνων των σφαλμάτων (MSE).
4. Το κριτήριο C_p του Mallows: όπως φανερώνει ο τίτλος το κριτήριο αυτό προτάθηκε από τον Mallows και δίνεται από τον τύπο.

$$Cp = \frac{SSE_p}{MSE} - (n - 2p)$$

όπου SSE_p είναι το άθροισμα τετραγώνων των υπολοίπων που αντιστοιχεί σε μοντέλο με p-1 ανεξάρτητες μεταβλητές και MSE είναι το μέσο τετραγωνικό σφάλμα του πλήρους μοντέλου (δηλαδή αυτού που χρησιμοποιεί όλες τις διαθέσιμες μεταβλητές). Για να επιλέξουμε πιο είναι το βέλτιστο πλήθος ανεξάρτητων μεταβλητών εργαζόμαστε ως εξής: αρχικά υπολογίζουμε την τιμή του C_p για όλα τα διαφορετικά υποσύνολα ανάλογα με το πλήθος των μεταβλητών, στην συνέχεια εντοπίζουμε την ελάχιστη τιμή του C_p για κάθε υποσύνολο και τέλος επιλέγουμε εκείνο το υποσύνολο των μεταβλητών για το οποίο ισχύει $Cp \cong p$. Εάν αυτό επιτυγχάνεται σε περισσότερα του ενός υποσύνολα τότε επιλέγουμε εκείνο με το μικρότερο C_p καθώς έτσι εξασφαλίζεται και μικρότερο άθροισμα τετραγώνων των υπολοίπων.

Ένα μειονέκτημα της μεθόδου αυτής είναι ότι σε περίπτωση πολλών ανεξάρτητων μεταβλητών υπάρχουν πολλοί συνδυασμοί που πρέπει να εξεταστούν. Για αυτό τον λόγο ακολουθείται μια διαδικασία που ονομάζεται διορθωτική παρέμβαση (t-directed search) σύμφωνα με την οποία προσαρμόζουμε αρχικά το πλήρες μοντέλο και ελέγχουμε για ποιες μεταβλητές ισχύει

$$T = \frac{\hat{\beta}_\kappa}{s(\hat{\beta}_\kappa)} : |T| < t_{n-p}(\alpha/2)$$

Για όσες ισχύει τις διώχνουμε από το μοντέλο και στην συνέχεια συνδυάζουμε όσες μείνανε με αυτές που διώξαμε προηγουμένως. Για παράδειγμα έστω ότι έχουμε

4 ανεξάρτητες μεταβλητές X_1, X_2, X_3, X_4 και μένουνε τελικά στο μοντέλο οι X_1, X_2 τότε να μοντέλα που θα πρέπει να συγκρίνουμε είναι τα X_1X_2, $X_1X_2X_3$, $X_1X_2X_4$, $X_1X_2X_3X_4$ και έτσι από τα 16 που θα έπρεπε να συγκρίνουμε αρχικά πέσαμε στα 4 τελικά.

Η 2^{η} μέθοδος για να δημιουργήσουμε ένα βέλτιστο αριθμό ανεξάρτητων μεταβλητών βασίζεται στην χρήση κάποιων επαναληπτικών διαδικασιών οι οποίες επιλέγουν με έναν συγκεκριμένο τρόπο ποιες μεταβλητές θα μπουν στο μοντέλο. Οι τρείς πιο δημοφιλείς είναι οι:

- Μέθοδος της προς τα εμπρός επιλογής (forward selection)
- Μέθοδος της προς τα πίσω απαλοιφής (backward elimination)
- Μέθοδος της κατά βήματα παλινδρόμησης (stepwise regression)

Α. Μέθοδος της προς τα εμπρός επιλογής

Η μέθοδος αυτή επιλέγει αρχικά την πιο σημαντική από τις ανεξάρτητες μεταβλητές, ξεκινά δηλαδή με την προσαρμογή όλων των απλών γραμμικών μοντέλων της μορφής

$$Y_i = \beta_0 + \beta_k * X_{i,k} + \varepsilon_i, \qquad i = 1,2,\dots,n$$

για κ=1,2,…,r-1, υπολογίζει τις τιμές των στατιστικών συναρτήσεων

$$F_{\kappa}^{*} = \frac{MSR(X_k)}{MSE(X_k)}$$

Και εντοπίζει για ποιον δείκτη k η ποσότητα αυτή μεγιστοποιείται, έστω X_1. Εάν $F_1^{*} > F_{1,n-p}(\alpha)$ Τότε η συγκεκριμένη μεταβλητή επιλέγεται ως η πρώτη που μπαίνει στο μοντέλο. Αν δεν υπάρχει τέτοια μεταβλητή τότε η διαδικασία σταματά εδώ και δεν μπαίνει καμία μεταβλητή στο μοντέλο. Στην συνέχεια προσαρμόζονται όλα τα γραμμικά μοντέλα της μορφής

$$Y_i = \beta_0 + \beta_1 * X_{i,1} + \beta_k * X_{i,k} + \varepsilon_i, \qquad i = 1,2,\dots,n$$

και επιλέγουμε την μεταβλητή για την οποία η ποσότητα

$$F_{\kappa,1}^{*} = \frac{MSR(X_k|X_{\kappa 1})}{MSE(X_k)}$$

μεγιστοποιείται έστω X_2. Εάν $F_{2,1}^{*} > F_{enter}$ τότε η X_2 μπαίνει στο μοντέλο αλλιώς η διαδικασία σταματά και μένουμε στο απλό γραμμικό μοντέλο του προηγούμενου βήματος. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να φτάσουμε σε ένα σημείο όπου $F < F_{enter}$ οπότε θα σταματήσει και το κατάλληλο μοντέλο θα αποτελείται από τις ανεξάρτητες μεταβλητές που θα έχουν μπει μέχρι τότε.

Β. Μέθοδος της προς τα πίσω απαλοιφής

Τώρα ξεκινάμε με το πλήρες μοντέλο

$$Y_i = \beta_0 * X_{i0} + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \cdots + \beta_{p-1} * X_{i,p-1} + \varepsilon_i$$

και επιλέγουμε τις μεταβλητές $X_κ$ για τις οποίες ισχύει

$$F^{**}{}_k = \frac{MSR(X_k|X_1, \dots, X_{k-1}, \dots, X_{p-1}}{MSE(X_1, \dots, X_{p-1})} < F_{remove}$$

και απορρίπτουμε από το μοντέλο την $X_κ$: $\min_κ F^{**}{}_k$. Στην συνέχεια επαναλαμβάνουμε το προηγούμενο βήμα μέχρι να μην γίνεται να απορρίψουμε καμία μεταβλητή.

Γ. Μέθοδος της κατά βήματα παλινδρόμησης

Οι δύο προηγούμενες μέθοδοι έχουν από ένα σοβαρό μειονέκτημα η καθεμία. Η πρώτη δεν επιτρέπει την απόρριψη μεταβλητής που έχει εισαχθεί στο μοντέλο και έχει καταστεί μη σημαντική λόγω της εισαγωγής κάποιας άλλης ανεξάρτητης μεταβλητής στο μοντέλο με την οποία έχει υψηλή συσχέτιση. Η δεύτερη δεν επιτρέπει το ακριβώς ανάποδο δηλαδή την εισαγωγή ανεξάρτητης μεταβλητής που είχε απορριφθεί σε προηγούμενο βήμα η οποία όμως έγινε σημαντική λόγω απόρριψης άλλης ανεξάρτητης μεταβλητής με την οποία είχε υψηλή συσχέτιση. Για τον λόγο έχει προταθεί η μέθοδος της κατά βήματα παλινδρόμησης η οποία ξεκινά με τα βήματα της προς τα εμπρός επιλογής και στην συνέχεια κάθε φορά που εισάγεται μια νέα μεταβλητή εξετάζει κατά πόσο μια από τις ήδη εισαγμένες γίνεται να απορριφθεί. Με τον τρόπο αυτό επιτυγχάνεται το να έχουμε ένα τελικό μοντέλο με μεταβλητές που δεν έχουν υψηλή συσχέτιση μεταξύ τους.

Τέλος, κάποιες παρατηρήσεις για τις μεθόδους. Είδαμε ότι για να λάβουμε την απόφαση για το αν θα εισαχθεί ή όχι μια ανεξάρτητη μεταβλητή στο μοντέλο ελέγχαμε τις τιμές των εκάστοτε ποσοτήτων F με τις F_{enter} και F_{remove} . Αυτές οι δύο ποσότητες καθορίζονται από την αρχή από τον χρήστη και συνήθως παίρνουνε την τιμή 4 ή κοντά σε αυτό. Επίσης πάντα πρέπει $F_{enter} \geq F_{remove}$ ώστε να μην κινδυνεύσει η μέθοδος να μπει σε έναν ατέρμονο κύκλο επανάληψης των βημάτων για την εισαγωγή και εξαγωγή των μεταβλητών (Κούτρας - Ευαγγελάρας 2010).

ΚΕΦΑΛΑΙΟ 3

Γενικευμένα γραμμικά μοντέλα

Στο προηγούμενο κεφάλαιο αναπτύξαμε την θεωρία γύρω από την γραμμική παλινδρόμηση. Όπως αναφέραμε στο προηγούμενο κεφάλαιο για να χρησιμοποιηθεί το γραμμικό μοντέλο πρέπει να ισχύει η υπόθεση της κανονικότητας. αν αυτή δεν ισχύει έχουμε διάφορες επιλογές όπως οι μετασχηματισμοί που κάναμε προηγουμένως. Υπάρχουν όμως περιπτώσεις όπου η υπόθεση της κανονικότητας δεν ισχύει ούτε προσεγγιστικά, οπότε τι μπορούμε να κάνουμε; Η πιο δημοφιλής πρακτική είναι η χρήση των γενικευμένων μοντέλων.

3.1 Γενικευμένα γραμμικά μοντέλα (ΓΓΜ)

Το 1972 οι Nelder & Wedderburn παρουσίασαν μια ενοποιημένη θεωρία για γραμμικά μοντέλα που δεν απαιτεί την υπόθεση της κανονικότητας για την μεταβλητή απόκρισης. Σύμφωνα με αυτήν, τα γραμμικά μοντέλα μπορούν να μελετηθούν ενιαία κάτω από την υπόθεση ότι η κατανομή της μεταβλητής απόκρισης ανήκει στην εκθετική οικογένεια κατανομών και για όλες τις κατανομές μέσα στην οικογένεια αυτή, οι εκτιμητές μεγίστης πιθανοφάνειας (ε.μ.π) των παραμέτρων του μοντέλου μπορούν να βρεθούν με τον ίδιο αλγόριθμο. Τα ΓΓΜ έχουν κάποια πλεονεκτήματα έναντι της συνήθους παλινδρόμησης.

- Πολύ μεγαλύτερο φάσμα εφαρμογών. Χρησιμοποιούνται και σε περιπτώσεις όπου δεν μπορεί να υποτεθεί ότι η κατανομή της Y είναι κανονική, ούτε καν προσεγγιστικά.
- Οι εκτιμητές των παραμέτρων προκύπτουν με την μέθοδο μέγιστης πιθανοφάνειας οπότε έχουν μια σειρά από επιθυμητές ιδιότητες όπως το ότι είναι αμερόληπτοι (ή τουλάχιστον ασυμπτωτικά) και έχουν την μικρότερη διακύμανση.
- Στις περισσότερες περιπτώσεις δεν χρειάζεται να υποθέσουμε σταθερή διακύμανση για τις τιμές της Y
- Στην περίπτωση που όλες οι μεταβλητές είναι κατηγορικές τα ΓΓΜ αποτελούν ένα βασικό τρόπο ανάλυσης σε πίνακες συνάφειας (Πολίτης 2011).

3.2 Εκθετική οικογένεια κατανομών

Μια κατανομή πιθανότητας λέμε ότι ανήκει στην εκθετική οικογένεια κατανομών όταν η συνάρτηση πιθανότητας (ή πυκνότητας αν η κατανομή είναι συνεχής) της κατανομής μπορεί να γραφεί στην μορφή

$$f_Y(y;\theta,\varphi) = \left[\exp\frac{y\theta - b(\theta)}{a(\varphi)} + c(y,\varphi)\right]$$

Όπου a, b, c είναι τρεις γνωστές συναρτήσεις, ενώ οι θ, φ είναι παράμετροι.

- Αν το φ είναι γνωστό τότε έχουμε την εκθετική οικογένεια με μία παράμετρο και το θ αναφέρεται ως η κανονική παράμετρος (canonical parameter) της κατανομής.
- Αν το φ δεν είναι γνωστό, τότε μπορούμε σε πολλές περιπτώσεις να το θεωρήσουμε σαν μια παράμετρο κλίμακας για την κατανομή, οπότε αποκαλείται παράγοντας όχλησης (nuisance factor) της κατανομής.

Πολλές γνωστές κατανομές ανήκουν στην παραπάνω οικογένεια όπως η κανονική, η διωνυμική, η Poisson και η Γάμμα. Στον παρακάτω πίνακα παρουσιάζονται συνοπτικά οι τύποι των συναρτήσεων για κάθε κατανομή (Πολίτης 2011).

Πίνακας 3.1 (Συναρτήσεις για διάφορα είδη κατανομών)

Κατανομή	θ	b(θ)	α(φ)	C(y,φ)
Κανονική $Y \sim N(\mu, \sigma^2)$	μ	θ^2	σ^2	$-\frac{1}{2}\left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]$
Διωνυμική $Y \sim Bi(n,p)$	$log\frac{p}{1-p}$	$nlog(1+e^{\theta})$	1	$log\binom{n}{y}$
Poisson $Y \sim Poi(\lambda)$	$log\lambda$	e^{θ}	1	$-\log(y!)$
Γάμμα $Y \sim Ga(n,a)$	-α	-nlog(-θ)	1	(n-1)log(y) – log(Γ(n))

3.3 Συναρτήσεις σύνδεσης

Σε ένα ΓΓΜ μια συνάρτηση σύνδεσης g είναι μια συνάρτηση που συνδέει το στοχαστικό τμήμα του μοντέλου (μέση τιμή της τ.μ Y) με το μη στοχαστικό τμήμα (γραμμικός συνδυασμός των ερμηνευτικών μεταβλητών X_i). Συγκεκριμένα έστω $\mu_i = E(Y_i)$ η μέση τιμή της μεταβλητής απόκρισης. Υποθέτουμε ότι αυτή εξαρτάται από τις τιμές των X_j j=1,2,…,k. Θεωρούμε την γραμμική συνάρτηση πρόβλεψης

$$\eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j * X_{ij} ,$$

όπου X_{ij} είναι η τιμή της μεταβλητής X_j για την παρατήρηση i. Τότε η συνάρτηση σύνδεσης συνδέει τη μέση τιμή της μεταβλητής απόκρισης με την παραπάνω συνάρτηση πρόβλεψης,

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j * X_{ij}$$

Η g θεωρούμε ότι είναι πάντα μια συνάρτηση μονότονη και διαφορίσιμη ενώ στην ειδική περίπτωση όπου $g = (b')^{-1}$ τότε η συνάρτηση σύνδεσης ονομάζεται κανονική συνάρτηση σύνδεσης (canonical link function). Στον παρακάτω πίνακα παρουσιάζονται οι κυριότερες κανονικές συναρτήσεις σύνδεσης.

Πίνακας 3.2 (Συναρτήσεις σύνδεσης για διάφορες κατανομές)

Κατανομή	b(θ)	b'(θ)=μ	$g(\mu)=(b')^{-1}(\mu)=\theta$
Κανονική	$\frac{\theta^2}{2}$	θ	Identity link $g(\mu)=\mu$
Poisson	e^{θ}	e^{θ}	Log link $g(\mu)=\log(\mu)$
Διωνυμική	$\log(1+e^{\theta})$	$\frac{e^{\theta}}{1+e^{\theta}}$	Logit link $g(\mu)=\log(\frac{\mu}{n-\mu})$
Γάμμα	$-\log(-\theta)$	$-\frac{1}{\theta}$	Inverse link $g(\mu)=-\frac{1}{\mu}$

Τέλος σε ένα ΓΓΜ η μεταβλητή απόκρισης εισέρχεται στο μοντέλο μέσω της μέσης της τιμής $g(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j * X_{ij}$. Με άλλα λόγια δεν γίνεται καμία υπόθεση για την κατανομή των σφαλμάτων στο μοντέλο παρά μόνο για την κατανομή της Y (Πολίτης 2011).

3.4 Λογιστική παλινδρόμηση

3.4.1 Εισαγωγικά στοιχεία

Το πιο διαδεδομένο ΓΓΜ για δίτιμα (ή διωνυμικά) δεδομένα είναι αυτό της λογιστικής παλινδρόμησης όπου ισχύει ότι

$$p(x) = \frac{e^{\alpha+\beta x}}{1+e^{a+\beta x}},$$

όπου αυτό προκύπτει λύνοντας ως προς p(x) την εξίσωση

$$logit[p(x)] := \log\left[\frac{p(x)}{1-p(x)}\right] = a + \beta x, \qquad p \in (0,1) \leftrightarrow \log\left[\frac{p}{1-p}\right] \in (-\infty, \infty)$$

Βλέπουμε ότι εδώ μοντελοποιείται γραμμικά ο λογάριθμος της σχετικής πιθανότητας (log odds) αντί η ίδια η πιθανότητα.

Το πρόσημο του συντελεστή β δίνει την μονοτονία της συνάρτησης:

- $\beta > 0 \Leftrightarrow$ η p(x) είναι γνησίως αύξουσα
- $\beta < 0 \Leftrightarrow$ η p(x) είναι γνησίως φθίνουσα
- $\beta = 0 \Leftrightarrow$ η p(x) είναι σταθερή (δεν εξαρτάται από το x)

Το μέγεθος του |β| δηλώνει την ταχύτητα με την οποία μεταβάλλεται η p(x).

- |β| μεγάλο ⇒ η p(x) αυξάνει ή φθίνει γρήγορα
- |β| μικρό ⇒ η p(x) αυξάνει ή φθίνει αργά.

(Ηλιόπουλος 2011)

Ο ρυθμός μεταβολής της p(x) καθορίζεται από το μέγεθος του β:

$$\frac{dp(x)}{dx} = \beta * p(x)[1 - p(x)]$$

Επομένως τοπικά, η p(x) προσεγγίζεται από μια ευθεία με αντίστοιχη κλίση:

Κοντά στο x_0: $p(x) \approx p(x_0) + \beta * p(x_0)[1 - p(x_0)](x - x_0)$

Η ευθεία με την μέγιστη (κατ' απόλυτη τιμή) κλίση β/4 αντιστοιχεί στο $x_0 = -\alpha/\beta$ αφού $p(-\alpha/\beta)=1/2$. Το σημείο $-\alpha/\beta$ αναφέρεται ως διάμεσο επίπεδο αποτελεσματικότητας (median effective level) και συμβολίζεται EL_{50}.

3.4.2 Εκτίμηση και συμπερασματολογία για τις παραμέτρους

Έστω ένα logit μοντέλο με k ερμηνευτικές μεταβλητές. Η εκτίμηση του διανύσματος **β'**=(β_0, β_1, β_2, … , β_κ) των παραμέτρων προκύπτει μέσω της μεγιστοποίησης της συνάρτησης πιθανοφάνειας

$$L(\mathbf{p}, y) = \exp\left[\sum_{i=1}^{n} y_i \log\frac{p_i}{1-p_i} + \sum_{i=1}^{n} \log(1-p_i)\right] + c_n(\mathbf{y})$$

όπου **p**=(p_1,…,p_n)' και **y**=(y_1,…,y_n)' και η συνάρτηση c_n δεν εξαρτάται από τις (άγνωστες) παραμέτρους p_i, ή ισοδύναμα του λογαρίθμου της

$$\log L(\mathbf{p}, y) = \sum_{i=1}^{n} y_i \log\frac{p_i}{1-p_i} + \sum_{i=1}^{n} \log(1-p_i)$$

(Πολίτης 2011)

Ο λογάριθμος της πιθανοφάνειας είναι μια γνησίως κοίλη συνάρτηση και επομένως έχει ένα μοναδικό μέγιστο. Αυτό είναι πεπερασμένο αν και μόνο αν τα (x) δεν είναι διαχωρισμένα, δηλαδή αν:

$$[\min\{x_i, y=0\}, max\{x_i, y=0\}] \cap [\min\{x_i, y=1\}, max\{x_i, y=1\}] \neq \emptyset.$$

Έστω το μοντέλο $\text{logit}[p(x)] := \log\left[\frac{p(x)}{1-p(x)}\right] = a + \beta x$, ο έλεγχος που μας ενδιαφέρει είναι ο H_o:β=0 έναντι H_1:β≠0. Δεδομένου του μοντέλου β=0 σημαίνει ότι δεν υπάρχει εξάρτηση από την επεξηγηματική μεταβλητή ενώ β≠0 σημαίνει ότι υπάρχει. Την απόφαση μπορούμε να την λάβουμε μέσω δύο ελέγχων

- Ο έλεγχος Wald: Ο έλεγχος αυτός βασίζεται στην στατιστική συνάρτηση $\frac{\hat{\beta}}{SE(\hat{\beta})}$ και υπό την H_0 έχει ασυμπτωτικά την κανονική κατανομή. Η H_0 απορρίπτεται για μεγάλες τιμές (κατ' απόλυτη τιμή) αυτής της στατιστικής συνάρτησης.
- Ο έλεγχος λόγου πιθανοφανειών: Ο έλεγχος αυτός βασίζεται στην στατιστική συνάρτηση $2(\log L_1 - \log L_0)$ όπου L_0 είναι η μέγιστη πιθανοφάνεια υπό την H_0 και L_1 είναι η μέγιστη πιθανοφάνεια υπό την H_1. Υπό την H_0 έχει ασυμπτωτικά κατανομή χι-τετράγωνο με βαθμούς ελευθερίας ίσους με:

  $$(\text{πλήθος παραμέτρων υπό την } H_0) - (\text{πλήθος παραμέτρων υπό την } H_1)$$

  Η H_0 απορρίπτεται για μεγάλες τιμές της στατιστικής συνάρτησης. Να σημειωθεί ότι η παραπάνω στατιστική συνάρτηση έχει πάντοτε μη αρνητικές τιμές αφού $L_0 \leq L_1$: Η L_1 είναι η μέγιστη πιθανοφάνεια ενώ L_0 είναι η μέγιστη πιθανοφάνεια περιορίζοντας όμως το β να είναι ίσο με το μηδέν

Τα αντίστοιχα διαστήματα εμπιστοσύνης μπορούμε να τα πάρουμε αντιστρέφοντας τους δύο ελέγχους. Για τον έλεγχο Wald το αντίστοιχο 100(1-α)% διάστημα εμπιστοσύνης για το β είναι $\hat{\beta} \pm z_{a/2}\text{SE}(\hat{\beta})$.

Για τον έλεγχο λόγου πιθανοφανειών το αντίστοιχο 100(1-α)% διάστημα εμπιστοσύνης περιέχει τις τιμές β_0 για τις οποίες $2[\log L_1 - \log L(\beta_0)] \leq X^2_{1,\alpha}$ (Ηλιόπουλος 2011).

Μια έννοια που χρησιμοποιείται συχνά και σχετίζεται με την προσαρμογή ενός μοντέλου είναι αυτή της απόκλισης (deviance). Πρόκειται για γενίκευση της έννοιας του αθροίσματος των τετραγώνων των καταλοίπων και προκύπτει από τον έλεγχο λόγου πιθανοφανειών. Έστω L_M η μέγιστη πιθανοφάνεια υπό το μοντέλο M και L_S η μέγιστη πιθανοφάνεια υπό το κορεσμένο μοντέλο (δηλαδή το μοντέλο που έχει τόσες παραμέτρους όσα και τα δεδομένα) και ταιριάζει τέλεια στα δεδομένα. Η απόκλιση του μοντέλου M είναι

$$\text{Deviance}_M = 2(\log L_S - \log L_m)$$

Στην περίπτωση μας η απόκλιση ασυμπτωτικά έχει κατανομή χι-τετράγωνο με βαθμούς ελευθερίας ίσους με (πλήθος παρατηρήσεων)-(πλήθος παραμέτρων μοντέλου Μ). Για δύο μοντέλα M_0, M_1 με το δεύτερο να είναι ειδική περίπτωση του πρώτου, ο έλεγχος λόγου πιθανοφανειών για τον έλεγχο H_0: ισχύει το M_0 κατά H_1: ισχύει το M_1 είναι ουσιαστικά η διαφορά των αποκλίσεων

$$2(\log L_1 - \log L_0) = 2(\log L_S - \log L_0) - 2(\log L_S - \log L_1)$$
$$= \text{Deviance}_0 - \text{Deviance}_1$$

Διαισθητικά όσο πιο μικρή είναι η απόκλιση ενός μοντέλου, τόσο πιο κοντά είναι στο κορεσμένο και αυτό παρέχει ένδειξη καλής προσαρμογής. Ένα διαγνωστικό κριτήριο για την προσαρμογή του μοντέλου είναι το Deviance/(βαθμοί ελευθερίας) μεγάλες τιμές του οποίου υποδεικνύουν κακή προσαρμογή (Πολίτης, Ηλιόπουλος 2011).

Τέλος μία υποσημείωση, συνήθως ο έλεγχος Wald και ο έλεγχος με τον λόγο πιθανοφανειών δίνουν το ίδιο αποτέλεσμα- για μικρά δείγματα είναι πιθανό να διαφέρουνε. Στην περίπτωση αυτή συνίσταται η χρήση του λόγου πιθανοφανειών.

3.4.3 Ερμηνεία των παραμέτρων σε σχέση με την σχετική πιθανότητα (odds) και τον λόγο σχετικών πιθανοτήτων (odds ratio)

Η έννοια της σχετικής πιθανότητας είναι σημαντική για την ερμηνεία των παραμέτρων σε ένα μοντέλο λογιστικής παλινδρόμησης. Γενικά, η σχετική πιθανότητα (odds) ενός ενδεχομένου Α ορίζεται ως ο λόγος

$$\text{odds} = \frac{P(A)}{1 - P(A)} \ ,$$

όπου P(A) δηλώνει την πιθανότητα να συμβεί το ενδεχόμενο Α. Τιμή της σχετικής πιθανότητας μεγαλύτερη του 1 δηλώνει ότι το ενδεχόμενο στον αριθμητή είναι πιο πιθανό να συμβεί από αυτό στον παρονομαστή.

Ένας από τους βασικούς λόγους που το μοντέλο της λογιστικής παλινδρόμησης προτιμάται από τα άλλα ΓΓΜ είναι η ευκολότερη διαισθητική ερμηνεία των αποτελεσμάτων με βάση τη σχετική πιθανότητα. Υπενθυμίζουμε ότι

$$\text{logit}[p(x)] := \log\left[\frac{p(x)}{1 - p(x)}\right] = a + \beta x \ \acute{\eta} \ \text{odds} = \frac{p(x)}{1 - p(x)} = \exp(a + \beta x)$$

Άρα αύξηση της τιμής του x κατά μία μονάδα μεταβάλλει την τιμή του λογαρίθμου σχετικής πιθανότητας κατά β:

$$\log\left[\frac{p(x+1)}{1-p(x+1)}\right] = \alpha + \beta(x+1) = \alpha + \beta x + \beta = \log\left[\frac{p(x)}{1-p(x)}\right] + \beta$$

Ισοδύναμα, αύξηση της τιμής του x κατά μία μονάδα, προκαλεί πολλαπλασιαστική αύξηση της σχετικής πιθανότητας κατά e^{β}:

$$\frac{p(x+1)}{1-p(x+1)} = e^{\alpha+\beta(x+1)} = e^{\beta}\left(e^{\alpha+\beta x}\right) = e^{\beta}\frac{p(x)}{1-p(x)}$$

Είναι $e^{\beta} > 1 \Leftrightarrow \beta > 0$ ενώ $e^{\beta} < 1 \Leftrightarrow \beta < 0$. (Πολίτης 2011)

Έστω ότι έχουμε δύο ενδεχόμενα Α, Β τότε ο λόγος των σχετικών πιθανοτήτων (odds ratio) του Α ως προς το Β είναι:

$$\text{odds ratio} = \Theta_{AB} = \frac{\frac{P(A)}{1-P(A)}}{\frac{P(B)}{1-P(B)}} = \frac{\text{odds A}}{\text{odds B}} = \frac{P(A) * [1-P(B)]}{P(B) * [1-PA)]}$$

και δείχνει πόσες φορές η σχετική πιθανότητα του ενδεχομένου Α είναι μεγαλύτερη του ενδεχομένου Β. Ας επιστρέψουμε πάλι στην περίπτωση ενός logit μοντέλου με μία μεταβλητή (δίτιμη) και σύμφωνα με την κλασική κωδικοποίηση δηλαδή

$$X_i = \begin{cases} 0, & \mu\eta\ \acute{\upsilon}\pi\alpha\rho\xi\eta\ \tau\sigma\upsilon\ \pi\alpha\rho\acute{\alpha}\gamma\sigma\nu\tau\alpha \\ 1, & \acute{\upsilon}\pi\alpha\rho\xi\eta\ \tau\sigma\upsilon\ \pi\alpha\rho\acute{\alpha}\gamma\sigma\nu\tau\alpha \end{cases}$$

Το μοντέλο μας έχει την μορφή $log\left[\frac{p(x)}{1-p(x)}\right] = a + \beta x$, οπότε είναι εύκολο να δούμε πλέον ότι:

$$\log(\text{odds; μη ύπαρξη παράγοντα}) = \log(\text{odds}; x_i = 0) = \alpha$$
$$\log(\text{odds; ύπαρξη παράγοντα}) = \log(\text{odds}; x_i = 1) = \alpha + \beta$$

και κατά συνέπεια:

$$\text{odds ratio} = \theta = \frac{\text{odds; ύπαρξη παράγοντα}}{\text{odds; } \mu\eta \text{ ύπαρξη παράγοντα}} = e^{\beta}$$

που μας δίνει τον σημειακό εκτιμητή του odds ratio μέσω του εκτιμητή του συντελεστή της λογιστικής παλινδρόμησης, ενώ το αντίστοιχο διάστημα εμπιστοσύνης είναι το:

$$\exp(\hat{\beta} \pm z_{a/2}SE(\hat{\beta})).$$

Να τονίσουμε ότι σε περίπτωση που χρησιμοποιήσουμε κάποια άλλη κωδικοποίηση τότε αλλάζει η εκτίμηση του β αλλά η εκτίμηση του odds ratio

παραμένει η ίδια. Αυτό όμως δεν ισχύει στην περίπτωση συνεχούς ανεξάρτητης μεταβλητής X, εδώ ισχύει:

$$e^{\beta} = \frac{odds; X = x + 1}{odds; X = x}$$

Δηλαδή ο συντελεστής β ισούται με τον λογάριθμο του odds ratio όταν οι συγκρινόμενοι «ασθενείς» έχουν μια μονάδα διαφορά στο επίπεδο της ανεξάρτητης μεταβλητής. Για διαφορά m μονάδων στο επίπεδο της ανεξάρτητης μεταβλητής το αντίστοιχο odds ratio θα ισούται με:

$$\mathrm{e}^{\mathrm{m}\beta} = \frac{\mathrm{odds; X = x + m}}{\mathrm{odds; X = x}}$$

(Κατέρη 2010)

Μια σημείωση σχετικά με αυτό. Σε περίπτωση που έχουμε περισσότερες από μία ανεξάρτητες μεταβλητές η ερμηνεία των παραμέτρων ενός μοντέλου λογιστικής παλινδρόμησης σε σχέση με τους λόγους σχετικών πιθανοτήτων ισχύει όταν στα μοντέλα δεν υπάρχουν όροι αλληλεπίδρασης. (Πολίτης 2011)

Τέλος, στο σημείο αυτό θα αναφέρουμε ένα ακόμα πλεονέκτημα της λογιστικής παλινδρόμησης. Η λογιστική παλινδρόμηση μπορεί να χρησιμοποιηθεί για συμπερασματολογία και να ερμηνευτεί ακόμα και σε περιπτώσεις που η απόκριση Y δεν είναι τυχαία αλλά σταθερή. Δηλαδή ακόμα και αν η X είναι τυχαία και η Y σταθερή, μπορούμε να εφαρμόσουμε την λογιστική παλινδρόμηση και να εκτιμήσουμε κατά πόσο η X είναι επεξηγηματική μεταβλητή για την Y. στην περίπτωση αυτή δεν μπορούμε να εκτιμήσουμε τις πιθανότητες p(x) αλλά μπορούμε να εκτιμήσουμε τα odds ratio αφού το β έχει την ίδια ερμηνεία όπως και στην γενική περίπτωση. Επομένως μπορούμε να εκτιμήσουμε ότι όταν η X μετακινηθεί από το επίπεδο x_1 στο επίπεδο x_2 η σχετική πιθανότητα του ενδεχομένου Y=1 πολλαπλασιάζεται με $e^{\widehat{\beta}(x_2 - x_1)}$ ενώ ο λογάριθμος μεταβάλλεται κατά $\hat{\beta}(x_2 - x_1)$ (Ηλιόπουλος 2011).

3.4.4 Συμπερασματολογία για την πιθανότητα επιτυχίας p(x)

Όπως αναφέρθηκε και πιο πάνω στην γενική περίπτωση όπου η μεταβλητή απόκρισης Y είναι μια δίτιμη τυχαία μεταβλητή και έχουμε μία ανεξάρτητη μεταβλητή X το μοντέλο της λογιστικής παλινδρόμησης δίνεται από την σχέση

$$\text{logit}[p(x)] := \log\left[\frac{p(x)}{1-p(x)}\right] = a + \beta x$$

με αντίστοιχη πιθανότητα επιτυχίας

$$p(x) = \frac{e^{\alpha+\beta x}}{1+e^{a+\beta x}}$$

Τα αντίστοιχα 100(1-α)% διαστήματα εμπιστοσύνης για τις παραπάνω εκτιμώμενες ποσότητες δίνονται από τις σχέσεις:

$$\text{logit}[\hat{p}(x)] \pm z_{a/2} * SE(\text{logit}[\hat{p}(x)])$$

και

$$\hat{p}(x) \pm z_{a/2} * SE(\hat{p}(x))$$

Αυτά είναι διαστήματα τύπου Wald και βασίζονται στις ασυμπτωτικές κατανομές των εκτιμητών. Για να υπολογιστούν τα παραπάνω διαστήματα εμπιστοσύνης πρέπει να υπολογίσουμε τα αντίστοιχα τυπικά σφάλματα, τα οποία υπολογίζονται μέσω της μεθόδου Δέλτα (Ηλιόπουλος 2011).

3.4.5 Λογιστική παλινδρόμηση με κατηγορικές ανεξάρτητες μεταβλητές και η σχέση με τους πίνακες συνάφειας

Έστω Y μια δίτιμη μεταβλητή απόκρισης η οποία έχει δύο επεξηγηματικές μεταβλητές X, Z με τιμές 0 και 1 η καθεμία. Δεδομένων αυτών μπορούν να προκύψουν τέσσερα διαφορετικά μοντέλα λογιστικής παλινδρόμησης ανάλογα με την σχέση που υπάρχει ανάμεσα στις τρεις μεταβλητές.

1. Μοντέλο της ομοιόμορφης συνάφειας των X,Y (και των Y,Z).

$$\text{logit}[P(Y=1)] = \alpha + \beta_1 X + \beta_2 Z$$

   Σε αυτό το μοντέλο οι X, Z δεν αλληλεπιδρούν, δηλαδή ανεξάρτητα από την κατηγορία της Z η διαφορά του logit για X=1 από το logit για X=0 είναι η ίδια και ισούται με $(\alpha + \beta_1 * 1 + \beta_2 * Z) - (\alpha + \beta_1 * 0 + \beta_2 * Z) = \beta_1$ που είναι ο λογάριθμος του λόγου σχετικών πιθανοτήτων (log odds ratio). Εξαιτίας του γεγονότος ότι το odds ratio παραμένει σταθερό δεδομένου του επιπέδου της τρίτης μεταβλητής το συγκεκριμένο μοντέλο ονομάζεται μοντέλο της ομοιόμορφης συνάφειας.

2. Μοντέλο της δεσμευμένης ανεξαρτησίας των X,Y δεδομένου του Z (και των Y,Z δεδομένου του X).

$$\text{logit}[P(Y=1)] = a + \beta_2 Z$$

Αποτελεί ειδική περίπτωση του μοντέλου της ομοιόμορφης συνάφειας όταν όλα τα odds ratio ισούνται με την μονάδα. Δεδομένου ότι $\text{odds ratio} = e^{\beta_1}$ πρέπει $\beta_1=0$ οπότε προκύπτει και ο παραπάνω τύπος.

3. Κορεσμένο μοντέλο.

$$\text{logit}[P(Y=1)] = \alpha + \beta_1 X + \beta_2 Z + \beta_3 XZ$$

Η ομοιόμορφη συνάφεια χάνεται όταν προσθέσουμε τον όρο της αλληλεπίδρασης. Πλέον για κάθε παρατήρηση έχουμε και μια διαφορετική παράμετρο.

4. Μοντέλο της ανεξαρτησίας της Y από τις X,Z.

$$\text{logit}[P(Y=1)] = \alpha$$

Σε αυτήν την περίπτωση το odds ratio παραμένει σταθερό ανεξαρτήτου του επιπέδου των δύο μεταβλητών και για να ισχύει αυτό πρέπει και οι δύο συντελεστές αναγκαστικά να ισούνται με το μηδέν. (Ηλιόπουλος 2011)

Όλα τα παραπάνω ισχύουν με τον ίδιο τρόπο αν οι ανεξάρτητες μεταβλητές έχουν περισσότερα των δύο επιπέδων. Ας υποθέσουμε ότι για την δίτιμη μεταβλητή απόκρισης Y έχουμε δύο ανεξάρτητες μεταβλητές X,Z με περισσότερα των δύο επίπεδα. Το μοντέλο της ομοιόμορφης συνάφειας είναι το

$$\text{logit}[P(Y=1)] = \alpha + \beta_i{}^X + \beta_k{}^Z$$

που παριστάνει την επίδραση της X μέσω των παραμέτρων $\beta_i{}^X$ και της Z μέσω των $\beta_\kappa{}^Z$. Στην περίπτωση αυτή τα log odds ratio που δημιουργούνται από δύο οποιεσδήποτε κατηγορίες i,j της X και τις δύο (αναγκαστικά) της Y ισούνται με

$$(\alpha + \beta_i{}^X + \beta_k{}^Z) - (\alpha + \beta_j{}^X + \beta_k{}^Z) = \beta_i{}^X - \beta_j{}^X$$

δηλαδή ανεξάρτητα από το k.

Η δεσμευμένη ανεξαρτησία των X και Y δοθέντος του Z εκφράζεται μέσω του μοντέλου

$$\text{logit}[P(Y=1)] = a + \beta_k{}^Z$$

όπου ισχύει $\beta_1{}^X = \cdots = \beta_k{}^X = 0$

Το κορεσμένο μοντέλο εκφράζεται μέσω του μοντέλου

$$\text{logit}[P(Y=1)] = \alpha + \beta_i{}^{X} + \beta_k{}^{Z} + \beta_{ik}{}^{XZ},$$

όπου οι παράμετροι $\beta_{ik}{}^{XZ}$ παριστάνουν τις αλληλεπιδράσεις των X,Z

Τέλος παρόμοια είναι η ερμηνεία αν έχουμε περισσότερες των δύο ανεξάρτητες μεταβλητές (Ηλιόπουλος 2011).

Η λογιστική παλινδρόμηση με κατηγορικές ανεξάρτητες μεταβλητές είναι από τα συνηθέστερα εργαλεία για την ανάλυση πινάκων συνάφειας και για αυτό τον λόγο θα παρουσιάσουμε κάποια βασικά στοιχεία της ανάλυσης πινάκων συνάφειας. Θα ξεκινήσουμε με την απλή περίπτωση ενός 2X2 πίνακα συνάφειας που θα μας βοηθήσει στην παρουσίαση κάποιων βασικών ποσοτήτων. Έστω ο παρακάτω πίνακας συχνοτήτων:

Πίνακας 3.3 (Πίνακας συνάφειας)

Ανεξάρτητη μεταβλητή	Μεταβλητή απόκρισης	
	Επιτυχία	Αποτυχία
ομάδα Α	n_{11}	n_{12}
ομάδα Β	n_{21}	n_{22}

Θεωρούμε ότι τα δεδομένα μας σε κάθε γραμμή προέρχονται από μια διωνυμική κατανομή, ενώ η πιθανότητα επιτυχίας για την ομάδα Α είναι p_1 και για την ομάδα Β p_2.

Ο δειγματικός λόγος σχετικών πιθανοτήτων (odds ratio) είναι:

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

και αποτελεί τον ΕΜΠ για τον πραγματικό λόγω σχετικών πιθανοτήτων για την ομάδα Α έναντι της ομάδας Β. επειδή η συνάρτηση $p \to p/(1-p)$ είναι γνησίως αύξουσα στο (0,1), ισχύει:

- $\theta < 1 \Leftrightarrow p_1 < p_2$
- $\theta = 1 \Leftrightarrow p_1 = p_2$
- $\theta > 1 \Leftrightarrow p_1 > p_2$

Για την εύρεση διαστημάτων εμπιστοσύνης δεν χρησιμοποιούμε τον παραπάνω εκτιμητή αλλά τον φυσικό λογάριθμο αυτού, του οποίου η δειγματική κατανομή είναι

η κανονική για μεγάλα δείγματα. Το τυπικό σφάλμα για τον εκτιμητή $\ln\hat{\theta}$ είναι ασυμπτωτικά ίσο με:

$$\mathrm{ASE}\left(\ln\hat{\theta}\right) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Οπότε ένα διάστημα εμπιστοσύνης για την παράμετρο ln(θ) είναι το

$$\ln\hat{\theta} \pm z_{a/2} * \mathrm{ASE}\left(\ln\hat{\theta}\right)$$

ενώ με αντιλογαρίθμηση παίρνουμε ένα διάστημα εμπιστοσύνης για το θ.

Σε περίπτωση που κάποιο από τα κελιά σε έναν πίνακα συχνοτήτων είναι κενό, δηλαδή $n_{ij}=0$, τότε η σημειακή εκτίμηση του odds ratio δεν γίνεται καθώς θα ισούται με 0 ή ∞ . Αντ' αυτού χρησιμοποιείται ο παρακάτω εκτιμητής

$$\hat{\theta} = \frac{(n_{11} + 0.5)(n_{22} + 0.5)}{(n_{12}+0.5)(n_{21} + 0.5)}$$

Μια άλλη ποσότητα που χρησιμοποιείται είναι ο σχετικός κίνδυνος (relative risk) της ομάδας Α ως προς την ομάδα Β και εκτιμάται μέσω της σχέσης

$$\mathrm{RR} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{n_{11}/(n_{11} + n_{12})}{n_{21}/(n_{21} + n_{22})}$$

Το τυπικό σφάλμα του παραπάνω εκτιμητή δεν μπορεί να υπολογιστεί εύκολα και για αυτό υπολογίζουμε το τυπικό σφάλμα για τον λογάριθμο του παραπάνω εκτιμητή που δίνεται από την σχέση:

$$\mathrm{SE}(\ln\mathrm{RR}) = \frac{1 - \hat{p}_1}{n_{1+}\hat{p}_1} + \frac{1 - \hat{p}_2}{n_{2+}\hat{p}_2},$$

όπου n_{1+} και n_{2+} είναι τα αθροίσματα γραμμών του 2X2 πίνακα.

Οπότε ένα διάστημα εμπιστοσύνης για την παράμετρο ln(RR) είναι το

$$\ln\mathrm{RR} \pm z_{a/2} * \mathrm{ASE}(\ln\mathrm{RR})$$

Πάλι αν θέλουμε το διάστημα εμπιστοσύνης για τον πραγματικό σχετικό κίνδυνο $\frac{p_1}{p_2}$ το μόνο που έχουμε να κάνουμε είναι να αντιλογαριθμήσουμε.

Γενικά σε έναν 2X2 πίνακα συνάφειας μας ενδιαφέρει ο έλεγχος για την ισότητα των δύο ποσοστών. Αυτός μπορεί να γίνει με πολλούς τρόπους όπως ο έλεγχος X^2 ή ο

ακριβής έλεγχος του Fisher για μικρά δείγματα. Στα πλαίσια ενός ΓΓΜ ο έλεγχος αυτός γίνεται μέσω των δειγματικών λόγων σχετικών πιθανοτήτων οι οποίοι έχουν κάποια πλεονεκτήματα όπως του ότι μπορούν να χρησιμοποιηθούν για διατάξιμα δεδομένα και μας βοηθούν να καταλάβουμε που οφείλεται η τυχόν απόκλιση από την ανεξαρτησία. Σε περίπτωση που ισχύει η ανεξαρτησία γραμμών στηλών τότε ο λόγος σχετικών πιθανοτήτων ισούται με την μονάδα.

Σε περίπτωση που έχουμε έναν r x c πίνακα συνάφειας με $\mathrm{r}, \mathrm{c} > 2$ τότε όσον αφορά τον έλεγχο της ανεξαρτησίας χρησιμοποιούμε το κελί στην πρώτη γραμμή και στην πρώτη στήλη ως βάση και υπολογίζουμε όλους τους λόγους σε σχέση με αυτό. Για να ισχύει η ανεξαρτησία πρέπει όλοι οι λόγοι να ισούνται με την μονάδα.

Συγκεντρωτικά ο έλεγχος που μας ενδιαφέρει είναι ο $H_0: p_1 = p_2 \; vs \; H_1: p_1 \neq p_2$ ο οποίος ανάγεται στον $H_0: \theta = 1 \; vs \; H_1: \theta \neq 1$. Η ανεξαρτησία απορρίπτεται εάν ο εκτιμητής θ είναι έξω από το διάστημα εμπιστοσύνης που παρουσιάσαμε παραπάνω ενώ δεν απορρίπτεται αν είναι μέσα σε αυτό.

Το τελευταίο κομμάτι της θεωρίας σχετικά με τους πίνακες συνάφειας αφορά τους πίνακες τριπλής εισόδου και το παράδοξο του Simpson. Συνήθως σε έναν πίνακα τριών διαστάσεων θεωρούμε μία μεταβλητή ελέγχου (έστω Z) και μελετάμε την σχέση των μεταβλητών X και Y στα διάφορα επίπεδα της Z. υπάρχουν δύο κύρια είδη συνάφειας που μας ενδιαφέρουν:

- Η μερική συνάφεια (partial association) ανάμεσα σε X και Y, δηλαδή η συσχέτιση για κάθε επίπεδο της Z ξεχωριστά.
- Η περιθώρια συνάφεια(marginal association) των X και Y με την χρήση ενός περιθώριου πίνακα που προκύπτει από την συγχώνευση των δεδομένων για τα διάφορα επίπεδα της μεταβλητής πλαισίου Z.

(Πολίτης 2011)

Μερικές φορές παρατηρείται το φαινόμενο η κατεύθυνση της περιθωριακής συνάφειας να είναι πολύ διαφορετική από αυτή των αντίστοιχων δεσμευμένων και μάλιστα μπορεί να είναι αντίστροφη. Ένα τέτοιο φαινόμενο ονομάζεται παράδοξο του Simpson και εμφανίζεται όταν η μεταβλητή πλαισίου παρουσιάζει ισχυρή συνάφεια και με τις δύο άλλες μεταβλητές. Για αυτό το λόγο θα πρέπει να είμαστε προσεχτικοί κατά την ερμηνεία ενός πίνακα τριπλής εισόδου.

3.4.6 Πολλαπλή λογιστική παλινδρόμηση

Στην περίπτωση περισσότερων από μιας επεξηγηματικών μεταβλητών $X_1,\ldots, X_p$ η πιθανότητα απόκρισης Y=1 εκφράζεται μέσω της σχέσης

$$p(x_1, \ldots, x_p) = \frac{\exp(x_1, \ldots, x_p)}{1 + \exp(x_1, \ldots, x_p)} ,$$

πλέον οι επεξηγηματικές μεταβλητές μπορούν να είναι είτε συνεχείς, είτε κατηγορικές (καλό να τις δηλώνουμε ως παράγοντες) είτε συνδυασμός και των δύο. Η ανάλυση είναι ακριβώς η ίδια όπως στην απλή περίπτωση οπότε δεν θα επεκταθούμε περισσότερο.

Το μόνο που θα προσθέσουμε είναι κάποια γενικά στοιχεία που πρέπει να λαμβάνουμε υπόψη κατά την ανάλυση. Αν τα δεδομένα μας δεν είναι ισορροπημένα, δηλαδή αν η απόκριση Y=1 εμφανίζεται πολύ λίγες ή πάρα πολλές φορές σε σχέση με την Y=0, τότε δεν είναι σωστό να χρησιμοποιούμε πολλές επεξηγηματικές μεταβλητές. Ένας χονδρικός κανόνας είναι να χρησιμοποιούμε μια επεξηγηματική μεταβλητή για τουλάχιστον 10 αποκρίσεις από την κάθε κατηγορία. Βέβαια ακόμα και αν παραβιάσουμε τον κανόνα η διαδικασία εκτίμησης μπορεί να γίνει, αλλά τότε οι εκτιμητές μπορεί να είναι αρκετά μεροληπτικοί. Επίσης, υπάρχει η πιθανότητα οι επεξηγηματικές μεταβλητές να εμφανίζουν πολυσυγγραμικότητα, όπως ακριβώς συμβαίνει και στην γραμμική παλινδρόμηση. Για αυτό τον λόγο πρέπει να είμαστε προσεκτικοί στην επιλογή των επεξηγηματικών μεταβλητών ώστε να αποφεύγονται περιπτώσεις όπου εμφανίζονται ισχυρές συσχετίσεις μεταξύ τους (Ηλιόπουλος 2011).

Τέλος, στην περίπτωση που έχουμε πολλές επεξηγηματικές μεταβλητές η σειρά με την οποία εισέρχονται στο μοντέλο επηρεάζει γενικά την σημαντικότητα τους. Συνεπώς είναι απαραίτητο να αναφέρουμε ποιες μεταβλητές υπάρχουν ήδη στο μοντέλο πριν εξετάσουμε την σημαντικότητα μιας νέας. Επίσης αν έχουμε σχετικά λίγες μεταβλητές τότε θα μπορούσαμε να τις εισάγουμε με διαφορετική σειρά στο μοντέλο και αν τα αποτελέσματα παραμένουν τα ίδια τότε προχωράμε κανονικά με την ανάλυση μας. Σε περίπτωση όμως που έχουμε διαφορετικά αποτελέσματα τότε την απόφαση για το ποιες θα κρατήσουμε θα την πάρουμε αφού λάβουμε υπόψη και άλλους παράγοντες όπως την σημαντικότητα κάθε μεταβλητής στην ανάλυση μας (Πολίτης 2011).

3.4.7 Λογιστική παλινδρόμηση με περισσότερες κατηγορίες

Όταν η απόκριση έχει περισσότερες από δύο κατηγορίες τότε η ανάλυση διαφέρει στην περίπτωση που είναι ονοματική ή διατακτική. (βέβαια αν είναι διατακτική μπορούμε να την αντιμετωπίσουμε ως ονοματική αλλά χάνουμε σε αποδοτικότητα).

3.4.7.1 Μοντέλα logit για ονοματικές (nominal) μεταβλητές

Έστω J>2 το πλήθος των κατηγοριών και p_1, p_2,…, p_J οι αντίστοιχες πιθανότητες που ικανοποιούν την σχέση $\sum_{j=1}^{J} p_j = 1$. Για να πραγματοποιήσουμε την ανάλυση επιλέγουμε μία κατηγορία αναφοράς (baseline category), χωρίς βλάβη της γενικότητας υποθέτουμε ότι αυτή είναι η τελευταία (η κατηγορία J). Ως logit ως προς την κατηγορία αναφοράς J ορίζονται οι ποσότητες

$$\log\left(\frac{p_j}{p_J}\right) = \alpha_j + \beta_j x, \qquad j = 1,2,\dots,J-1$$

για περισσότερες επεξηγηματικές μεταβλητές απλώς αλλάζει το δεξί μέλος της σχέσης σε $\alpha_j + \beta_{1j}x_1 + \cdots + \beta_{pj}x_p$. Ενώ οι αντίστοιχες πιθανότητες επιτυχίες δίνονται από τις σχέσεις:

$$p_j = \frac{e^{\alpha_j+\beta_j x}}{1 + e^{\alpha_1+\beta_1 x} + \cdots + e^{\alpha_{J-1}+\beta_{J-1} x}}, \qquad j = 1,2,\dots,J-1$$

$$p_J = \frac{1}{1 + e^{\alpha_1+\beta_1 x} + \cdots + e^{\alpha_{J-1}+\beta_{J-1} x}}$$

Για i≠j ο λογάριθμος της σχετικής πιθανότητας της i κατηγορίας ως προς την j ισούται με:

$$\log\left(\frac{p_i}{p_j}\right) = \log\left(\frac{p_i/p_J}{p_j/p_J}\right) = \log\left(\frac{p_i}{p_J}\right) - \log\left(\frac{p_j}{P_J}\right) = (\alpha_i - \alpha_j) + (\beta_i - \beta_j)x$$

(Ηλιόπουλος 2011)

3.4.7.2 Μοντέλα logit για διατακτικές (ordinal) μεταβλητές

Το κλασσικό μοντέλο λογιστικής παλινδρόμησης για διατακτικές μεταβλητές είναι το cumulative logit model. Αν η απόκριση Y έχει J διατεταγμένες κατηγορίες με αντίστοιχες πιθανότητες p_1,…,p_J και υπάρχει μία επεξηγηματική μεταβλητή X (όμοια ισχύει και για περισσότερες μεταβλητές) θέτουμε:

$$\text{logit}[P(Y \le j)] = \log\left[\frac{P(Y \le j)}{P(Y > j)}\right] = \log\left(\frac{p_1 + \cdots + p_j}{p_{j+1} + \cdots + p_J}\right) = a_j + \beta x,$$
$$j = 1, \ldots, J-1$$

όπου βλέπουμε ότι ο συντελεστής κλίσης είναι σταθερός και ίσος με β.
Οι πιθανότητες των κατηγοριών δίνονται από τις σχέσεις

$$p_1 = \frac{e^{\alpha_1+\beta x}}{1 + e^{\alpha_1+\beta x}}$$
$$p_j = \frac{e^{\alpha_j+\beta x}}{1 + e^{\alpha_j+\beta x}} - \frac{e^{\alpha_{j-1}+\beta x}}{1 + e^{\alpha_{j-1}+\beta x}}, \qquad j = 2, \ldots, J-1$$
$$p_J = 1 - \frac{e^{\alpha_{J-1}+\beta x}}{1 + e^{\alpha_{J-1}+\beta x}}$$

Εδώ ο λογάριθμος του λόγου σχετικών πιθανοτήτων για δύο διαφορετικές τιμές της X, έστω x_1 και x_2 είναι:

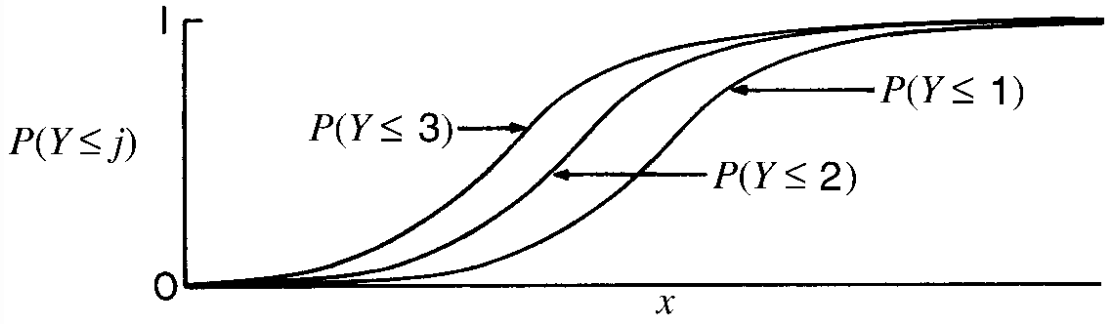
$$\log\left[\frac{P(Y \le j|X = x_2)/P(Y > j|X = x_2)}{P(Y \le j|X = x_1)/P(Y > j|X = x_1)}\right] = \beta(x_2 - x_1),$$

δηλαδή είναι ανάλογος της διαφοράς $x_2 - x_1$ ενώ η σταθερά αναλογίας β είναι ανεξάρτητη από το j (Ηλιόπουλος 2011).

Βασική υπόθεση του μοντέλου αυτού είναι η υπόθεση των αναλογικών odds (proportional odds assumption), δηλαδή αυτό που έχουμε γράψει παραπάνω ότι ο συντελεστής β είναι κοινός για όλα τα logit. Με απλά λόγια η υπόθεση αυτή σημαίνει ότι η σχέση ανάμεσα σε κάθε ζευγάρι αποτελεσμάτων είναι η ίδια. Για παράδειγμα ο συντελεστής β για την χαμηλότερη κατηγορία της εξαρτημένης μεταβλητής έναντι όλων των παραπάνω κατηγοριών είναι ο ίδιος με αυτόν που περιγράφει την σχέση της αμέσως μεγαλύτερης κατηγορίας έναντι όλων των παραπάνω. Επειδή η σχέση όλων των γκρουπ είναι η ίδια χρειαζόμαστε μόνο ένα σετ συντελεστών (http://www.ats.ucla.edu/stat/r/dae/ologit.htm).

Γραφικά η παραπάνω υπόθεση φαίνεται από το παρακάτω σχήμα που παρουσιάζει την κατανομή των logit για μια διατακτική μεταβλητή τεσσάρων επιπέδων και για μια ανεξάρτητη μεταβλητή x.

Διάγραμμα 3.4 (Υπόθεση των αναλογικών odds)



Όπου βλέπουμε πως για συγκεκριμένο j η καμπύλη απόκρισης? (response curve) είναι η καμπύλη της λογιστικής παλινδρόμησης έχοντας ως κατηγορίες απόκρισης τα αποτελέσματα $Y \le j$ *και* $Y > j$. Βλέπουμε πως οι καμπύλες για j=1,2,3 έχουν ίδιο σχήμα αλλά απέχουν η μια από την άλλη στον οριζόντιο άξονα (Agresti 2007).

ΚΕΦΑΛΑΙΟ 4

Παρουσίαση των μεταβλητών

Σκοπός του κεφαλαίου αυτού είναι αρχικά η παρουσίαση των μεταβλητών που έχουμε στην διάθεση μας και στην συνέχεια θα ακολουθήσει μια αρχική διερευνητική ανάλυση με την εξαγωγή κάποιων αρχικών αποτελεσμάτων.

4.1 Προέλευση των δεδομένων και περιγραφή

Τα δεδομένα μας χορηγήθηκαν από την Ιατρική Σχολή Αθηνών και σε αυτά περιλαμβάνονται οι μετρήσεις που έγιναν σε 45 άτομα και των δύο φύλων. Πιο συγκεκριμένα για κάθε άτομο που μετέχει στην έρευνα έχει καταγραφεί η ηλικία του, το φύλο του, στην συνέχεια καταγράφεται το στάδιο του καρκίνου του ασθενή μέσω της ιστολογικής έκθεσης καθώς και ο βαθμός του μέσω της μεταβλητής grade. Επίσης καταγράφονται οι τιμές του δείκτη κυτταρικού πολλαπλασιασμού Ki 67 και της Κυκλίνης Ε όπως και των πρωτεϊνών E2F1 και E2F4. Τέλος υπάρχει η μεταβλητή Risk η οποία αποτελεί έναν συμψηφισμό των μεταβλητών grade και stage και δείχνει εάν ο ασθενής διατρέχει χαμηλό ή υψηλό κίνδυνο.

Πριν προχωρήσουμε στην αρχική ανάλυση θα κάνουμε μια σύντομη παρουσίαση των μεταβλητών που έχουμε στην διάθεση μας δίνοντας μεγαλύτερη έμφαση στους κυτταρικούς δείκτες.

- Ο κυτταρικός δείκτης πολλαπλασιασμού Ki 67 είναι μια πρωτεΐνη στενά συνδεδεμένη με τον πολλαπλασιασμό των κυττάρων. Κατά την διάρκεια της μεσοφάσης το Ki 67 ανιχνεύεται αποκλειστικά στον πυρήνα του κυττάρου ενώ κατά την διάρκεια της μίτωσης μεταφέρεται στην επιφάνεια των χρωμοσωμάτων. Το Ki 67 είναι παρόν καθ'όλη την διάρκεια του κυτταρικού κύκλου (G_1,S,G_2 και μίτωση) με εξαίρεση την φάση ηρεμίας G_0. Για αυτό τον λόγο το Ki 67 αποτελεί έναν εξαιρετικό δείκτη του κλάσματος πολλαπλασιασμού ενός πληθυσμού κυττάρων (www.wikipedia.org).
- Η Κυκλίνη Ε είναι μια πρωτεΐνη μέλος της οικογένειας των κυκλίνων, η οποία αυξάνεται σε ποσότητα στο τέλος της G_1 φάσης και αρχίζει να μειώνεται στην αρχή της S φάσης. Η Κυκλίνη αλληλεπιδρά με την Cdk2

(είδος κινάσης) και από κοινού ξεκινάνε τις διαδικασίες για τον διπλασιασμό του DNA κυρίως μέσω της δέσμευσης ουσιών που εμποδίζουν την μετάβαση στην φάση S. Υπερέκφραση της Κυκλίνης Ε έχει παρατηρηθεί σε αρκετές κακοήθειες και σχετίζεται με υψηλό πολλαπλασιασμό (http://carcin.oxfordjournals.org/content/25/3/375.short).

- Τα E2F1 και E2F4 είναι πρωτεΐνες που ανήκουν στην οικογένεια μεταγραφικών παραγόντων E2F η οποία επηρεάζει σημαντικά στοιχεία του μηχανισμού του διπλασιασμού όπως τις Κυκλίνες Α και Ε αλλά και στοιχεία όπως την πρωτεΐνη p73 που επηρεάζει την διαδικασία της απόπτωσης. Η οικογένεια E2F χωρίζεται σε δύο ομάδες, στην πρώτη περιλαμβάνονται οι E2F1-3 οι οποίοι προάγουν την μεταγραφή (transcriptional activators) και στην δεύτερη οι E2F4-5 κυρίως και σε μικρότερο βαθμό η E2F6 που την εμποδίζουν (transcriptional repressors). Η συντονισμένη ενεργοποίηση των δύο ομάδων διασφαλίζει την σωστή πρόοδο του κυτταρικού κύκλου κάτω από τις κατάλληλες συνθήκες ή την απαγορεύει και ενισχύει αντίθετα τον μηχανισμό της απόπτωσης ως μέτρο ανταπόκρισης σε διάφορες ανωμαλίες κατά την διαδικασία της μεταγραφής ή σε περίπτωση ζημιάς του DNA. Πάντως πρέπει να τονίσουμε ότι ο ρόλος των E2F ως προωθητές ή αναστολείς του κυτταρικού κύκλου δεν είναι σταθερός αλλά μεταβάλλεται. Ειδικότερα ο E2F1 αν και ανήκει στην πρώτη ομάδα που προωθεί τον κυτταρικό κύκλο, εντούτοις σχετίζεται με την απόπτωση ενώ αποτελεί και μέρος των μέτρων που λαμβάνονται σε περίπτωση βλάβης του DNA. Η απορύθμιση της λειτουργίας των παραγόντων E2F είναι κάτι που συμβαίνει πάντα στην περίπτωση ενός καρκίνου. Πράγματι ο εκτεταμένος πολλαπλασιασμός, η αντοχή στην απόπτωση και η ενεργοποίηση των μέτρων που λαμβάνονται στην περίπτωση ζημιάς του DNA είναι χαρακτηριστικά του καρκίνου και σχετίζονται άμεσα ή έμμεσα με τους παράγοντες E2F. Σε σχέση με τα διαφορετικά είδη καρκίνων η επίδραση των E2F ως ογκογενείς ή ογκοκατασταλτικοί εξαρτάται από τον τύπο των κυττάρων και του καρκίνου (Yoshida,2008).
- Ο βαθμός κακοήθειας του καρκίνου (grade) είναι ένα σύστημα που μας βοηθά να ταξινομήσουμε τους διάφορους καρκίνους ανάλογα με το μέγεθος της ανωμαλίας τους και το πόσο γρήγορα μπορεί να αναπτυχθεί

και να εξαπλωθεί. Οι παθολόγοι ανάλογα με την μικροσκοπική εικόνα του καρκίνου που θα πάρουν θα τον τοποθετήσουν σε μία εκ των τριών βαθμίδων. Τα καρκινικά κύτταρα 1ου βαθμού δεν διαφοροποιούνται πολύ από τα φυσιολογικά κύτταρα και έχουν την τάση να αναπτύσσονται και να πολλαπλασιάζονται αργά. Για αυτό το λόγο οι καρκίνοι 1ου βαθμού θεωρούνται οι λιγότερο επιθετικοί. Αντίθετα τα καρκινικά κύτταρα δεν μοιάζουν με τα αντίστοιχα φυσιολογικά κύτταρα και οι αντίστοιχοι καρκίνοι μεγαλώνουν και επεκτείνονται ποιο γρήγορα από τους αντίστοιχους καρκίνους χαμηλότερου βαθμού. Αυτά φαίνονται συνοπτικά στον παρακάτω πίνακα

Πίνακας 4.1 (βαθμός κακοήθειας του καρκίνου)

G1	Καλά διαφοροποιούμενος (χαμηλός βαθμός)
G2	Μέτρια διαφοροποιούμενος (ενδιάμεσος βαθμός)
G3	Φτωχά διαφοροποιούμενος (υψηλός βαθμός)

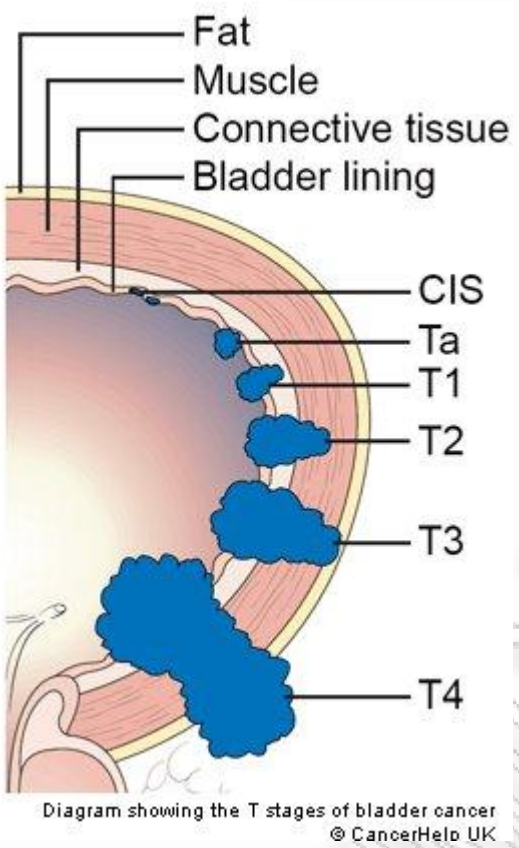
(Πηγή:http://cancerhelp.cancerresearchuk.org/type/bladder-cancer/treatment/bladder-cancer-stage-and-grade)

- Το στάδιο του καρκίνου περιγράφει την σοβαρότητα του καρκίνου βασιζόμενο στο μέγεθος του αρχικού καρκίνου και στο κατά πόσο έχει επεκταθεί στο υπόλοιπο σώμα. Από τα πιο διαδεδομένα συστήματα για τον υπολογισμό του σταδίου του καρκίνου είναι το σύστημα TNM, το οποίο βασίζεται στο μέγεθος του όγκου (T), την έκταση της εξάπλωσης στους λεμφαδένες (N) και την παρουσία μετάστασης (M). σχετικά με το μέγεθος του όγκου ο καρκίνος ταξινομείται σε μία από τις παρακάτω κατηγορίες.

Πίνακας 4.2 (στάδιο του καρκίνου)

Cis ή Tis	Πολύ αρχικά, αν και μεγάλου βαθμού, καρκινικά κύτταρα που έχουν εντοπιστεί μόνο στο εσωτερικό στρώμα του τοιχώματος της κύστης
Ta	Ο καρκίνος βρίσκεται μόνο στο εσωτερικό στρώμα του τοιχώματος της κύστης
T1	ο καρκίνος έχει αρχίσει να αναπτύσσεται στο συνδετικό ιστό κάτω από τα τοιχώματα της ουροδόχου κύστης
T2	ο καρκίνος έχει επεκταθεί μέσω του συνδετικού ιστού στο μυ
T3	ο καρκίνος έχει επεκταθεί μέσω του μυ στο στρώμα λίπους
T4	ο καρκίνος έχει εξαπλωθεί έξω από την ουροδόχο κύστη

(Πηγή:http://cancerhelp.cancerresearchuk.org/type/bladdercancer/treatment/bladder-cancer-stage-and-grade)



Εικ. 4.1 Απεικόνιση των σταδίων του καρκίνου

(Πηγή:http://cancerhelp.cancerresearchuk.org/prod_consump/groups/cr_common/@cah/@gen/documents/image/crukmig_1000img-12151.jpg)

Αν ένας καρκίνος βρίσκεται σε ένα από τα τρία πρώτα στάδια (Tis, Ta, T1) τότε αυτός ονομάζεται επιφανειακός (superficial). Οι καρκίνοι με στάδιο T2 και T3 ονομάζονται διηθητικοί (invasive). Τέλος καρκίνος με στάδιο T4 ονομάζεται και προχωρημένος (advanced).

4.2 Περιγραφικά στοιχεία

Στην συγκεκριμένη ενότητα θα κάνουμε μια περιγραφική ανάλυση των δεδομένων έτσι ώστε να πάρουμε μια γενική εικόνα. Πιο συγκεκριμένα, μέσω πινάκων και διαγραμμάτων θα παρουσιάσουμε κάποια πρώτα στοιχεία για τις μεταβλητές, για παράδειγμα για τις συνεχής μεταβλητές θα καταγράψουμε κάποια περιγραφικά μέτρα όπως την μέση τιμή, την τυπική απόκλιση κτλ, όπως επίσης θα κάνουμε και κάποιους ελέγχους κανονικότητας. Σχετικά με τις κατηγορικές μεταβλητές πέραν της παρουσίασης της κατανομής τους θα γίνουν και κάποιοι X^2 έλεγχοι ή τα ακριβή τεστ του Fisher ώστε να δούμε αν υπάρχει κάποια σχέση μεταξύ μεταβλητών που πιστεύουμε ότι θα έχει ενδιαφέρον. Στις συνεχείς μεταβλητές θα ελέγξουμε την κανονικότητα των δεδομένων μέσω των κατάλληλων διαγραμμάτων όπως και μέσω των αντίστοιχων ελέγχων (Shapiro-Wilk).

4.2.1 Φύλο των ασθενών

Η πρώτη μεταβλητή που έχουμε στην διάθεση μας είναι το φύλο των ασθενών όπου σύμφωνα με τα δεδομένα έχουμε ότι οι άντρες ασθενείς είναι αρκετοί περισσότεροι από τις γυναίκες. Από εδώ φαίνεται πως επιβεβαιώνεται αυτό που αναφέρουμε στο 1^{ο} κεφάλαιο ότι δηλαδή ο συγκεκριμένος τύπος καρκίνου φαίνεται να προσβάλλει πιο συχνά τους άντρες σε σχέση με τις γυναίκες.

Πίνακας 4.3 (Κατανομή του φύλου των ασθενών)

Φύλο	Απόλυτες συχνότητες	Σχετικές Συχνότητες
Άντρας	37	0.82
Γυναίκα	8	0.18

4.2.2 Ηλικία των ασθενών

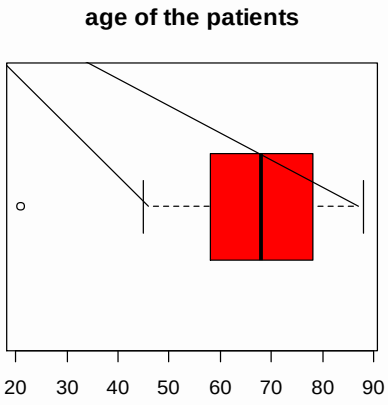
Επόμενη μεταβλητή που έχουμε στη διάθεση μας και με την οποία θα ασχοληθούμε είναι η ηλικία των ασθενών όπου φαίνεται ότι ο καρκίνος προσβάλει τις μεγαλύτερες ηλικιακά ομάδες.

Πίνακας 4.4 (Περιγραφικά μέτρα για την ηλικία των ασθενών)

Ηλικία	
Ελάχιστο	21
1^{ο} τεταρτημόριο	58
Διάμεσος	68
Μέση τιμή	67.19
3^{ο} τεταρτημόριο	78
Μέγιστο	88
Ελλείπουσες τιμές	2

Διάγραμμα 4.1 (Ιστόγραμμα συχνοτήτων)

Διάγραμμα 4.2 (Θηκόγραμμα)



Παρατηρούμε ότι αν εξαιρέσουμε την ελάχιστη τιμή που ανήκει σε ασθενή ηλικίας 21 χρονών και ίσως αποτελεί μια σπάνια περίπτωση εμφάνισης της ασθενείας σε τόσο μικρή ηλικία παρατηρούμε ότι η συντριπτική πλειοψηφία των κρουσμάτων

εμφανίζεται από το διάστημα των ηλικιών 50-60 και άνω, επιβεβαιώνοντας τις μελέτες που παρουσιάστηκαν στο 1^{ο} κεφάλαιο.

4.2.3 Βαθμός κακοήθειας (grade) του καρκίνου

Μια σημαντική μεταβλητή είναι ο βαθμός κακοήθεια του καρκίνου. Στα δεδομένα μας εμφανίζονται περιπτώσεις και των τριών βαθμίδων της κλίμακας grade, ενώ όπως βλέπουμε και από τον πίνακα αλλά και το διάγραμμα η πλειοψηφία τους αφορά φτωχά διαφοροποιημένους καρκίνους (G3).

Πίνακας 4.5 (Κατανομή του βαθμού κακοήθειας του καρκίνου)

Grade	Απόλυτες συχνότητες	Σχετικές Συχνότητες
G1	11	24.4%
G2	16	35.6%
G3	18	40%

4.2.4 Το στάδιο (stage) του καρκίνου

Πέρα από τον βαθμό κακοήθειας ένα άλλο στοιχείο που χαρακτηρίζει τον καρκίνο είναι το στάδιο του. Εδώ παρατηρούμε ότι η πλειοψηφία των περιπτώσεων αφορά επιφανειακούς καρκίνους, δηλαδή καρκίνους με στάδιο pTa ή pT1. Επίσης η συγκεκριμένη μεταβλητή είναι αυτή με το μεγαλύτερο πρόβλημα όσον αφορά τις ελλείπουσες τιμές καθώς από τις 45 περιπτώσεις έχουμε 10 κενές.

Πίνακας 4.6 (Κατανομή του σταδίου του καρκίνου)

Stage	Απόλυτες συχνότητες	Σχετικές Συχνότητες
pTa	21	46.7%
pT1	6	13.3%
pT2	8	17.8%
Ελλείπουσες τιμές	10	22.2%

4.2.5 Ο κίνδυνος (risk) που αντιμετωπίζει ο ασθενής

Εδώ παρατηρούμε ότι ο αριθμός των ασθενών που αντιμετωπίζουν χαμηλό σχετικά κίνδυνο είναι ελαφρά μεγαλύτερος από όσους αντιμετωπίζουν υψηλό.

Πίνακας 4.7 (Κατανομή για τον κίνδυνο των ασθενών)

Risk	Απόλυτες συχνότητες	Σχετικές Συχνότητες
Low	25	55.6%
High	20	44.4%

4.2.6 Σχέσεις ανάμεσα σε grade, stage και risk

Στο σημείο αυτό και πριν προχωρήσουμε στην παρουσίαση των υπόλοιπων μεταβλητών μας είναι ενδιαφέρων να δούμε τι σχέσεις αναπτύσσονται ανάμεσα στις τρεις παραπάνω μεταβλητές. Αρχικά μας ενδιαφέρει η σχέση που υπάρχει μεταξύ του βαθμού κακοήθειας και του σταδίου του καρκίνου και στην συνέχεια η σχέση αυτών των δύο με τον κίνδυνο του ασθενή.

4.2.6.1 Σχέση ανάμεσα σε stage και grade

Ο έλεγχος για τις δύο μεταβλητές θα γίνει με δύο τρόπους, αρχικά με το ακριβές τεστ του Fisher θα ελέγξουμε την ανεξαρτησία των δύο μεταβλητών και στην συνέχεια με το γ των Goodman-Kruskal θα ελέγξουμε την συνάφεια των δύο μεταβλητών.

Πίνακας 4.8 (Πίνακας συνάφειας: grade-stage)

		Grade		
		G1	G2	G3
Stage	Pta	9	10	2
	Pt1	0	0	6
	Pt2	0	1	7

Πίνακας 4.9 (Exact Fisher Test)

Data	Πίνακας 4.8
p-value	1.006e-05
alternative hypothesis	two.sided

Πίνακας 4.10 (Goodman – Kruskal γ)

Gamma	0.939
Std.error	0.061
CI	0.821

Από τον πρώτο πίνακα βλέπουμε ότι απορρίπτεται ξεκάθαρα η υπόθεση της ανεξαρτησίας των δύο μεταβλητών (p-value<0.05) ενώ από το γ των Goodman και Kruskal (το οποίο είναι ένα μέτρο συνάφειας για διατακτικές μεταβλητές) βλέπουμε ότι μεταξύ των δύο μεταβλητών υπάρχει μια πολύ ισχυρή (σχεδόν τέλεια γραμμική) συνάφεια. Μια μικρή υποσημείωση, σε περίπτωση ανεξαρτησίας ο δείκτης γ ισούται με το μηδέν, το αντίστροφο όμως (δηλαδή γ=0 να συνεπάγεται ανεξαρτησία) δεν ισχύει πάντα εκτός από την περίπτωση των 2x2 πινάκων (Κατέρη, 2010).

4.2.6.2 Σχέση ανάμεσα σε stage και risk

Όπως έχει ειπωθεί προηγουμένως (βλ. 4.1) η μεταβλητή risk αποτελεί έναν συμψηφισμό των μεταβλητών stage και grade είναι ενδιαφέρον να δούμε πια η σχέση της μεταβλητής risk με τις άλλες δύο. Πρώτη περίπτωση την οποία θα ερευνήσουμε είναι μεταξύ των μεταβλητών stage και risk. Ο έλεγχος θα γίνει και εδώ με δύο

τρόπους, αρχικά με το ακριβές τεστ του Fisher και στην συνέχεια με ένα τεστ γραμμικής τάσης (linear trend test) δεδομένης της διατακτικής φύσης των δεδομένων. Από το τεστ της γραμμικής τάσης θα πάρουμε και έναν συντελεστή συσχέτισης για τις δύο μεταβλητές (Ηλιόπουλος, 2011).

Πίνακας 4.11 (Πίνακας συνάφειας: risk-stage)

		Risk	
		Low	High
Stage	Pta	19	2
	Pt1	0	6
	Pt2	0	8

Πίνακας 4.12 (Exact Fisher Test)

Data	Πίνακας 4.11
p-value	5.173e-08
alternative hypothesis	two.sided

Πίνακας 4.13 (linear trend test)

data	risk.low out of risk.total
X-squared	23.7781
p-value	1.081e-06
correlation	0.84

Από τον έλεγχο του Fisher απορρίπτουμε πάλι την ανεξαρτησία των μεταβλητών (p-value<0.05) κάτι το οποίο επιβεβαιώνεται και από τον έλεγχο γραμμικής τάσης. Επιπλέον έχουμε έναν αρκετά υψηλό συντελεστή συσχέτισης ρ=0.84.

4.2.6.3 Σχέση ανάμεσα σε grade και risk

Την ίδια διαδικασία που ακολουθήσαμε πριν θα ακολουθήσουμε και τώρα όσον αφορά τις μεταβλητές grade και risk.

Πίνακας 4.14 (Πίνακας συνάφειας: risk-grade)

		Risk	
		Low	High
grade	G1	11	0
	G2	14	2
	G3	0	18

Πίνακας 4.15 (Exact Fisher Test)

Data	Πίνακας 4.14
p-value	5.521e-11
alternative hypothesis	two.sided

Πίνακας 4.16 (linear trend test)

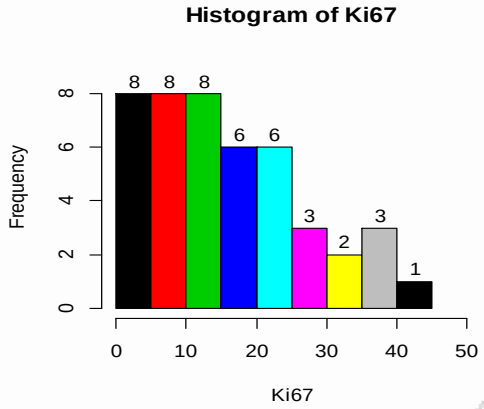
data	risk.low out of risk.total
X-squared	32.1664
p-value	1.415e-08
correlation	0.97

Τα αποτελέσματα που έχουμε είναι παρόμοια με την προηγούμενη περίπτωση και πιο συγκεκριμένα πάλι απορρίπτεται ξεκάθαρα η υπόθεση της ανεξαρτησίας (p-value<0.05) ενώ βλέπουμε ότι ο συντελεστής συσχέτισης είναι πάρα πολύ υψηλός (ρ=0.97).

4.2.7 Κυτταρικός δείκτης πολλαπλασιασμού Ki 67

Ο πρώτος από τους κυτταρικούς δείκτες που μας ενδιαφέρουν είναι η πρωτεΐνη Ki 67. Οι πληροφορίες που παίρνουμε για την πρωτεΐνη αυτή παρουσιάζονται στον παρακάτω πίνακα και το αντίστοιχο ιστόγραμμα.

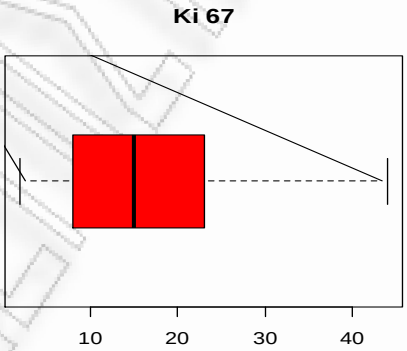
Διάγραμμα 4.3 (Ιστόγραμμα συχνοτήτων)



Πίνακας 4.17 (Περιγραφικά μέτρα για το Ki67)

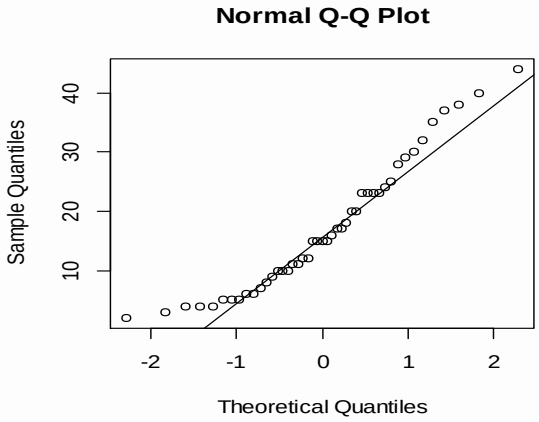
Ki 67	
Ελάχιστο	2
1^{ο} τεταρτημόριο	8
Διάμεσος	15
Μέση τιμή	17.02
3^{ο} τεταρτημόριο	23
Μέγιστο	44

Διάγραμμα 4.4 (Θηκόγραμμα)



Από το ιστόγραμμα παρατηρούμε ότι οι περισσότερες τιμές είναι συγκεντρωμένες στο αριστερό άκρο και για αυτό τον λόγο το ιστόγραμμα παρουσιάζει αυτό το σχήμα. Ενώ και από το θηκόγραμμα παρατηρούμε ότι το σχήμα κλείνει προς τα αριστερά χωρίς όμως να υπάρχει κάποια ακραία τιμή. Σχετικά με την κανονικότητα των δεδομένων έχω ότι τα δεδομένα μου δεν ακολουθούν την κανονική κατανομή όπως δείχνουν το παρακάτω διάγραμμα (qq-plot) και ο αντίστοιχος έλεγχος.

Διάγραμμα 4.5 (qq-plot)



Πίνακας 4.18 (Έλεγχος Shapiro-Wilk)

Data	Ki 67
W	0.9332
p-value	0.0122

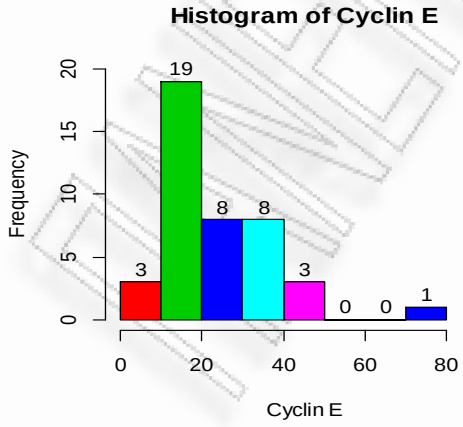
4.2.8 Κυκλίνη Ε

Ο δεύτερος κυτταρικός δείκτης που μας ενδιαφέρει είναι η πρωτεΐνη κυκλίνη Ε. σχετικά με αυτήν έχουμε τα παρακάτω αποτελέσματα.

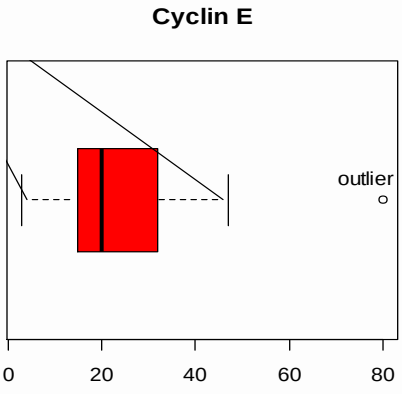
Πίνακας 4.19 (Περιγραφικά μέτρα για την Κυκλίνη Ε)

Κυκλίνη Ε	
Ελάχιστο	3
$1^{ο}$ τεταρτημόριο	15
Διάμεσος	20
Μέση τιμή	24.1
$3^{ο}$ τεταρτημόριο	32
Μέγιστο	80
Ελλείπουσες τιμές	3

Διάγραμμα 4.6 (Ιστόγραμμα συχνοτήτων)



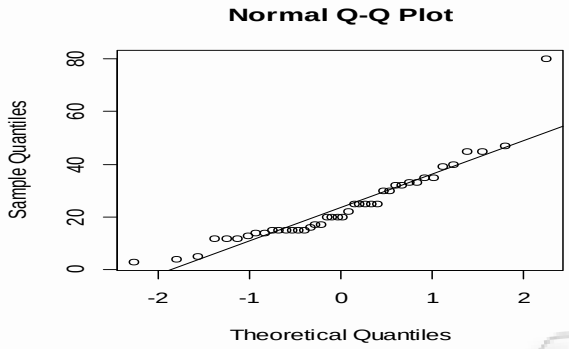
Διάγραμμα 4.7 (Θηκόγραμμα)



Από το ιστόγραμμα παρατηρούμε ότι η πλειοψηφία των παρατηρήσεων βρίσκεται μεταξύ του 10-20% και επίσης ότι υπάρχει και μια ακραία τιμή που αντιστοιχεί στην μέγιστη τιμή 80%. Σχετικά με την κανονικότητα της Κυκλίνης Ε αυτή απορρίπτεται

και η αιτία για αυτό είναι η ακραία παρατήρηση καθώς αν την απομακρύνουμε τότε δεν έχουμε πρόβλημα όπως βλέπουμε από τα επόμενα διαγράμματα και τους αντίστοιχους ελέγχους. Στην πρώτη περίπτωση έχουμε τα αποτελέσματα για όλα τα δεδομένα ενώ στην δεύτερη αφού διώξουμε την ακραία παρατήρηση.

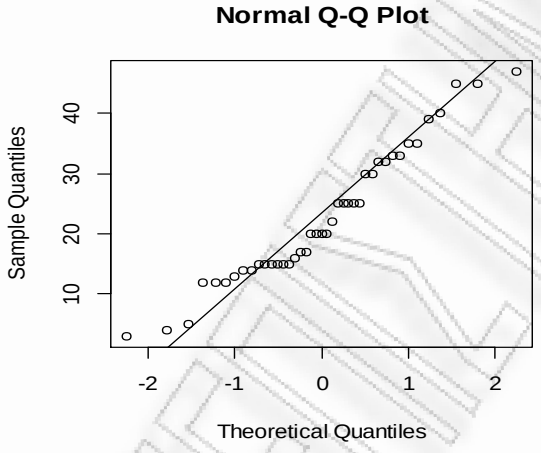
Διάγραμμα 4.8 (qq-plot)



Πίνακας 4.20 (Έλεγχος Shapiro-Wilk)

Data	Cyclin E
W	0.8819
p-value	0.0004

Διάγραμμα 4.9 (qq-plot)



Πίνακας 4.21 (Έλεγχος Shapiro-Wilk)

Data	Cyclin E
W	0.9518
p-value	0.081

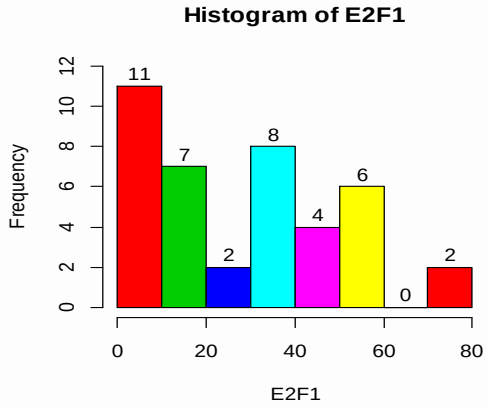
4.2.9 Οι πρωτεΐνες E2F1 και E2F4

Οι τελευταίες ποσότητες τις οποίες έχουμε στην διάθεση μας και θα μας απασχολήσουν είναι οι πρωτεΐνες E2F1 και E2F4 για τις οποίες έχουμε.

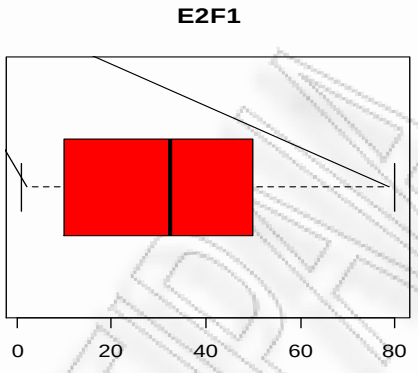
Πίνακας 4.22 (Περιγραφικά μέτρα για τα E2F1 και E2F4)

	E2F1	E2F4
Ελάχιστο	1	5
$1^{ο}$ τεταρτημόριο	10	31.25
Διάμεσος	32.5	52.5
Μέση τιμή	31.95	50.45
$3^{ο}$ τεταρτημόριο	50	70
Μέγιστο	80	85
Ελλείπουσες τιμές	5	3

Διάγραμμα 4.10 (Ιστόγραμμα συχνοτήτων)

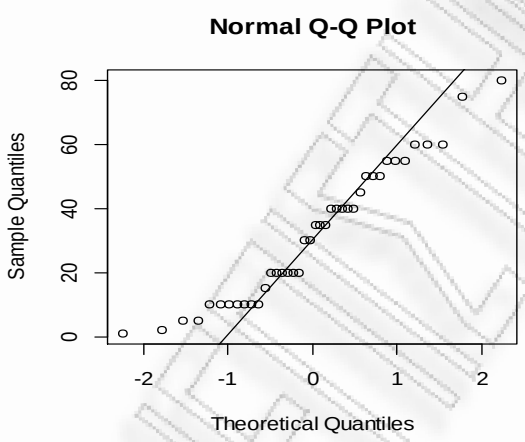


Διάγραμμα 4.11 (Θηκόγραμμα)



Από τα διαγράμματα δεν παρατηρούμε να υπάρχει κάποια ακραία τιμή ενώ η πλειοψηφία των περιπτώσεων έχει τιμή για το E2F1 μέχρι 20%. Ενώ σχετικά με την κανονικότητα των δεδομένων έχουμε ότι αυτή απορρίπτεται οριακά αν και από το σχήμα βλέπουμε ότι οι παρατηρήσεις στα άκρα ξεφεύγουν αρκετά από την ευθεία γραμμή .

Διάγραμμα 4.12 (qq-plot)

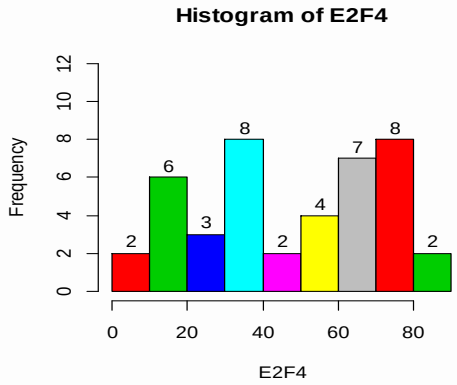


Πίνακας 4.23 (Έλεγχος Shapiro-Wilk)

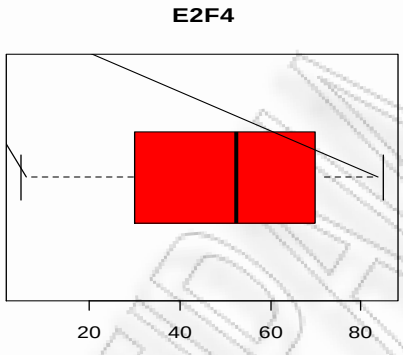
Data	E2F1
W	0.9435
p-value	0.045

Σχετικά με την E2F4 παρατηρούμε ότι υπάρχει μια πιο ισορροπημένη κατανομή των τιμών με μια ελαφριά κλίση προς τις μεγαλύτερες τιμές ενώ και πάλι δεν υπάρχει κάποια τιμή που να θεωρηθεί ακραία. Σχετικά με την κανονικότητα παρατηρούμε ότι πάλι απορρίπτεται καθώς από το σχήμα βλέπουμε ότι οι παρατηρήσεις στις άκρες απομακρύνονται από την ευθεία γραμμή, αυτό επιβεβαιώνεται και από τον έλεγχο (p-value<0.05).

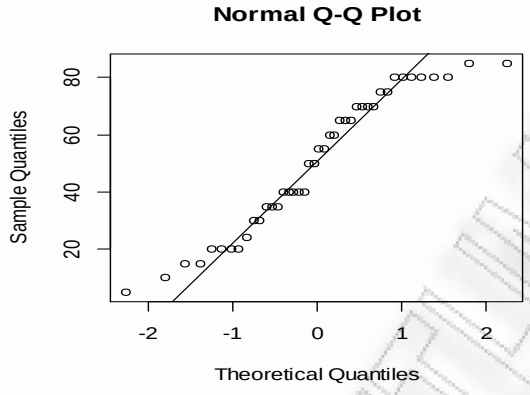
Διάγραμμα 4.13 (Ιστόγραμμα συχνοτήτων)



Διάγραμμα 4.14 (Θηκόγραμμα)



Διάγραμμα 4.15 (qq-plot)



Πίνακας 4.24 (Έλεγχος Shapiro-Wilk)

Data	E2F4
W	0.9307
p-value	0.013

Κεφάλαιο 5

Ανάλυση σε επίπεδο κυττάρου

Αφού πήραμε μια γενική εικόνα για τις μεταβλητές μας και τα δεδομένα μας συνολικότερα, τώρα θα περάσουμε στο κύριο μέρος της ανάλυσης. Στο συγκεκριμένο κεφάλαιο θα μας απασχολήσει το τι γίνεται σε επίπεδο κυττάρου και πιο συγκεκριμένα μας ενδιαφέρουν οι σχέσεις που αναπτύσσονται μεταξύ των κυτταρικών δεικτών αλλά θα ελέγξουμε επίσης εάν οι δείκτες Ki67 και κυκλίνη Ε επηρεάζονται από το φύλο και την ηλικία των ασθενών.

5.1 Σχέσεις των δεικτών με τα χαρακτηριστικά των ασθενών

Αρχικά θα ελέγξουμε αν και κατά πόσο η ηλικία και το φύλο των ασθενών επηρεάζουν τις τιμές των Ki67 και κυκλίνης Ε. Αυτό θα το ελέγξουμε μέσω παραμετρικών ή μη ελέγχων ανάλογα με το αν τα δείγματα στα γκρουπ που δημιουργούνται ακολουθούν την κανονική κατανομή.

5.1.1 Σχέση Ki67 με το φύλο και την ηλικία

Σχετικά με τα δύο δείγματα ως προς το φύλο έχουμε:

Πίνακας 5.1 (Μέση τιμή και τυπική απόκλιση για Ki67 ανά φύλο)

Φύλο	Ki67	Ασθενείς
Άντρες	16.76(±11.51)	37
Γυναίκες	18.25(±10.01)	8

Όπως βλέπουμε ο αριθμός των γυναικών είναι πολύ μικρός οπότε δεν έχει νόημα να κάνουμε έλεγχο κανονικότητας σε αυτήν την ομάδα οπότε την απόφαση για να δούμε αν υπάρχει διαφορά μεταξύ των δύο δειγμάτων θα την πάρουμε μέσω του μη παραμετρικού ελέγχου του Wilcoxon.

Πίνακας 5.2 (Wilcoxon rank sum test)

W	129.5
p-value	0.5926

Άρα δεν μπορούμε να απορρίψουμε την υπόθεση μου ότι δεν υπάρχει διαφορά μεταξύ των φύλων όσον αφορά το Ki67 (p-value>0.05)

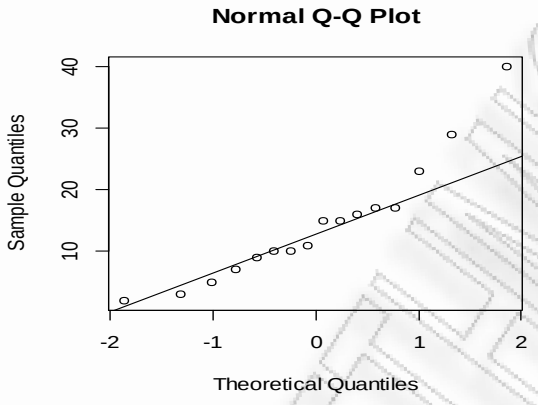
Στην συνέχεια όσον αφορά την επίδραση της ηλικίας έχουμε:

Πίνακας 5.3 (Μέση τιμή και τυπική απόκλιση για Ki67 ανά ηλικία)

Ηλικία	Ki67	Ασθενείς
<65	14.31 ± 9.88	16
≥65	18.15 ± 12.07	27

Όπου 65 είναι το κατώφλι της ηλικίας (Khan, 2003). Για να ελέγξουμε κατά πόσο οι διαφορές που παρατηρούνται είναι στατιστικά σημαντικές θα ακολουθήσουμε την ίδια διαδικασία με πριν. Πρώτα θα δούμε τι γίνεται με την κανονικότητα στο δείγμα των ασθενών κάτω των 65 ετών.

Διάγραμμα 5.1 (qqplot)



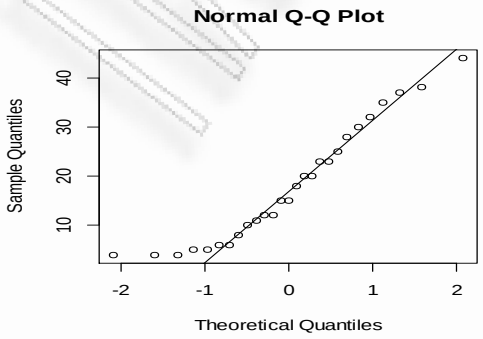
Πίνακας 5.4 (Ελεγχος Shapiro-Wilk)

Data	<65
W	0.9032
p-value	0.09049

Από το διάγραμμα μάλλον καταλήγουμε στο συμπέρασμα ότι δεν ισχύει η κανονικότητα καθώς στην πάνω άκρη φαίνεται ότι οι παρατηρήσεις φεύγουνε από την ευθεία γραμμή. Σύμφωνα με το τεστ όμως δεν μπορούμε να απορρίψουμε ότι τα δεδομένα προέρχονται από την κανονική κατανομή (p-value>0.05). Ο λόγος που συμβαίνει αυτό είναι ίσως λόγο της διαδικασίας που ακολουθεί ο έλεγχος Shapiro-Wilk

Στην συνέχεια σχετικά με τους ασθενείς άνω των 65 έχουμε τα εξής:

Διάγραμμα 5.2 (qqplot)



Πίνακας 5.5 (Έλεγχος Shapiro-Wilk)

Data	>65
W	0.9215
p-value	0.04299

Εδώ βλέπουμε ότι οριακά απορρίπτεται η υπόθεση της κανονικότητας (p-value<0.05), οπότε θα πρέπει πάλι να στραφούμε στον έλεγχο του Wilcoxon.

Πίνακας 5.6 (Wilcoxon rank sum test)

W	253.5
p-value	0.352

Άρα δεν μπορώ να απορρίψω την υπόθεση μου ότι δεν υπάρχει διαφορά μεταξύ των ηλικιακών όσον αφορά το Ki67 (p-value>0.05).

Οπότε συγκεντρωτικά μπορούμε να πούμε ότι ούτε το φύλο ούτε η ηλικία επηρεάζουν τον δείκτη Ki67.

5.1.2 Σχέση Κυκλίνης Ε με το φύλο και την ηλικία

Με παρόμοιο τρόπο όπως πριν θα δουλέψουμε και στην περίπτωση της Κυκλίνης Ε. Σχετικά με τα δύο δείγματα ως προς το φύλο έχουμε:

Πίνακας 5.7 (Μέση τιμή και τυπική απόκλιση για την Κυκλίνη Ε ανά φύλο)

Φύλο	Κυκλίνη Ε	Ασθενείς
Άντρες	24.65(±15.16)	34
Γυναίκες	21.75(±9.7)	8

Όπως και για τον δείκτη Ki 67 έτσι και τώρα θα στραφούμε απευθείας στο μη παραμετρικό τεστ για τον έλεγχο της ισότητας των δύο γκρουπ.

Πίνακας 5.8 (Wilcoxon rank sum test)

W	152
p-value	0.6185

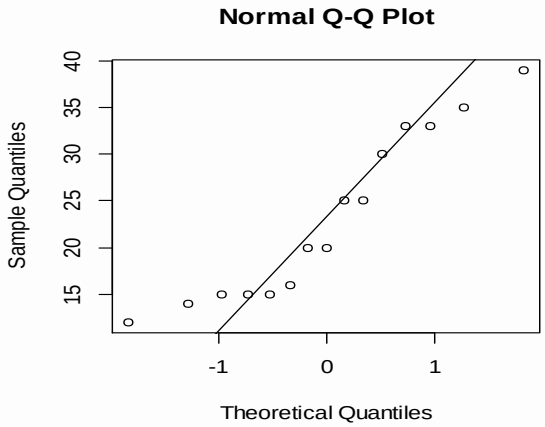
Παρατηρούμε ότι πράγματι δεν υπάρχει διαφορά μεταξύ των δύο γκρουπ, άρα το φύλο του ασθενούς δεν επηρεάζει την Κυκλίνη Ε.

Σχετικά με την επίδραση της ηλικίας έχουμε ότι:

Πίνακας 5.9 (Μέση τιμή και τυπική απόκλιση για την Κυκλίνη Ε ανά ηλικία)

Ηλικία	Κυκλίνη Ε	Ασθενείς
<65	23.13±8.93	15
≥65	25.44±16.99	25

Διάγραμμα 5.3 (qqplot) Πίνακας 5.10 (Ελεγχος Shapiro-Wilk)

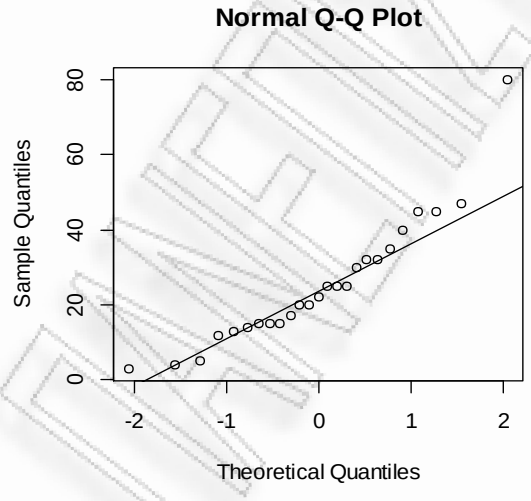


Data	<65
W	0.9024
p-value	0.1037

Από το διάγραμμα για τους ασθενείς κάτω των 65 δεν είναι καθαρό το τι συμβαίνει αν και οι παρατηρήσεις στο αριστερό άκρο του σχήματος φαίνεται να απομακρύνονται αρκετά από την ευθεία. Σύμφωνα με το τεστ όμως δεν μπορούμε να απορρίψουμε ότι τα δεδομένα προέρχονται από την κανονική κατανομή (p-value>0.05). Αυτό πρέπει να οφείλεται τόσο στον τρόπο υπολογισμού που χρησιμοποιεί το συγκεκριμένο τεστ όσο και στο σχετικά μικρό δείγμα κάτι που οδηγεί στο να απορρίπτεται σχετικά πιο δύσκολα η υπόθεση της κανονικότητας.

Στην συνέχεια σχετικά με τους ασθενείς άνω των 65 έχουμε ότι:

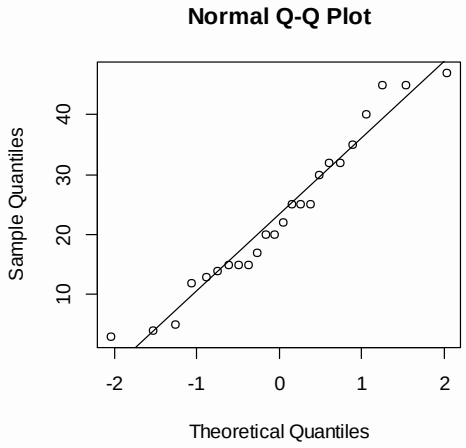
Διάγραμμα 5.4 (qqplot) Πίνακας 5.11 (Ελεγχος Shapiro-Wilk)



Data	>65
W	0.896
p-value	0.01504

Εδώ βλέπουμε ότι η κανονικότητα απορρίπτεται και για αυτό ευθύνεται η ακραία τιμή 80% για την Κυκλίνη Ε. Αν την απομακρύνουμε έχουμε ότι τα δεδομένα ακολουθούν την κανονική κατανομή (p-value>0.05)

Διάγραμμα 5.5 (qqplot)



Πίνακας 5.12 (Έλεγχος Shapiro-Wilk)

Data	>65
W	0.952
p-value	0.2987

Λόγω της αναντιστοιχίας που φαίνεται να υπάρχει μεταξύ του qqplot και του ελέγχου Shapiro-Wilk σχετικά με την κανονικότητα των ασθενών ηλικίας κάτω των 65 ετών και λόγω του μικρού πλήθους σε κάθε ομάδα είναι προτιμότερο να στραφούμε πάλι στο μη παραμετρικό τεστ για τον έλεγχο της ισότητας των γκρουπ. Οπότε μέσω του ελέγχου του Wilcoxon έχουμε ότι δεν απορρίπτεται η υπόθεση της ισότητας των δύο ομάδων, οπότε ούτε η ηλικία επηρεάζει την Κυκλίνη Ε.

Πίνακας 5.13 (Wilcoxon rank sum test)

W	189.5
p-value	0.9665

5.1.3 Η Κυκλίνη Ε ως κατηγορική μεταβλητή

Σχετικά με την Κυκλίνη θα μπορούσαμε να ακολουθήσουμε και άλλη οδό. Πολλές φορές η συγκεκριμένη μεταβλητή δεν θεωρείται συνεχής αλλά ως δίτιμη (0,1) ανάλογα με το αν η τιμή της είναι πάνω ή κάτω από μια τιμή κατώφλι η οποία συνήθως είναι είτε η τιμή 20 είτε η τιμή 30 (Khan,2003; Ioachim 2003) . Οπότε αν κάνουμε την συγκεκριμένη κωδικοποίηση μπορούμε να εφαρμόσουμε την θεωρία των διακριτών δεδομένων και παίρνουμε τους παρακάτω πίνακες συνάφειας όπως και τους αντίστοιχους X^2 ελέγχους.

Αρχικά θα δούμε τι γίνεται θεωρώντας ως κατώφλι την τιμή 20% και στην συνέχεια την τιμή 30

Πίνακας 5.14 (Exact Fisher Test)

Ηλικία	Κυκλίνη		
		<20	≥20
	<65	8	7
	≥65	12	13
Fisher's Exact Test			
p-value	1		
95% Conf.int.	0.29, 5.4		
sample estimates	0.53 0.48		

Βλέπουμε καθαρά (p-value>0.05) ότι δεν υπάρχει διαφορά μεταξύ της πιθανότητας η κυκλίνη να είναι κάτω από 20% στα άτομα κάτω των 65 και άνω των 65. Εδώ 0.53 είναι η πιθανότητα άτομο ηλικίας κάτω των 65 να έχει κυκλίνη μικρότερη από 20% και 0.48 η αντίστοιχη πιθανότητα για άτομο ηλικίας άνω των 65. Ενώ μέσω της τιμής του odds ratio έχουμε ότι:

$$odds\ ratio = 1.231$$

Αυτό σημαίνει ότι η σχετική πιθανότητα (odds) του να είναι η Κυκλίνη μικρότερη του 20% για τα άτομα κάτω των 65 είναι 1.231 φορές μεγαλύτερη από την αντίστοιχη σχετική πιθανότητα (odds) του να είναι η Κυκλίνη μικρότερη του 20% για τα άτομα άνω των 65 αλλά η διαφορά αυτή δεν είναι στατιστικά σημαντική καθώς το διάστημα εμπιστοσύνης περιλαμβάνει την μονάδα (0.29, 5.4)

Πίνακας 5.15 (Exact Fisher Test)

Φύλο	Κυκλίνη		
		<20	≥20
	άντρας	18	16
	γυναίκα	4	4
Fisher's Exact Test			
p-value	1		
95% Conf.int.	0.17, 7.11		
sample estimates	0.53 0.5		

Πάλι δεν μπορούμε να απορρίψουμε την υπόθεση ότι δεν υπάρχει διαφορά μεταξύ της πιθανότητας η κυκλίνη να είναι κάτω από 20% στους άντρες έναντι των γυναικών (p-value>0.05). Οι πιθανότητες έχουν παρόμοια ερμηνεία με πριν. Ενώ μέσω της τιμής του odds ratio έχω ότι:

$$odds\ ratio = 1.121$$

Αυτό σημαίνει ότι η σχετική πιθανότητα (odds) του να είναι η Κυκλίνη μικρότερη του 20 για τους άντρες είναι 1.121 φορές μεγαλύτερη από την αντίστοιχη σχετική πιθανότητα (odds) του να είναι η Κυκλίνη μικρότερη του 20% για τις γυναίκες αλλά η διαφορά αυτή δεν είναι στατιστικά σημαντική καθώς το διάστημα εμπιστοσύνης περιλαμβάνει την μονάδα (0.24, 5.25).

Θέτοντας την τιμή 30% παρατηρούμε ότι τα αποτελέσματα δεν διαφέρουν σε σχέση με πριν και πιο συγκεκριμένα:

Πίνακας 5.16 (Exact Fisher Test)

Ηλικία	Κυκλίνη		
		<30	≥30
	<65	11	4
	≥65	17	8
Fisher's Exact Test			
p-value	1		
95% Conf.int.	0.26, 7.3		
sample estimates	0.73 0.68		

Βλέπουμε καθαρά (p-value>0.05) ότι δεν υπάρχει διαφορά μεταξύ της πιθανότητας η κυκλίνη να είναι κάτω από 30% στα άτομα κάτω των 65 και άνω των 65. Ενώ μέσω της τιμής του odds ratio έχω ότι:

$$odds\ ratio = 1.285,$$

με διάστημα εμπιστοσύνης το (0.26, 7.3)

Πίνακας 5.17 (Exact Fisher Test)

Φύλο	Κυκλίνη		
		<30	≥30
	άντρας	24	10
	γυναίκα	6	2
Fisher's Exact Test			
p-value	1		
95% Conf.int.	0.06, 5.59		
sample estimates	0.70 0.75		

Πάλι δεν μπορούμε να απορρίψουμε την υπόθεση ότι δεν υπάρχει διαφορά μεταξύ της πιθανότητας η κυκλίνη να είναι κάτω από 30% στους άντρες έναντι των γυναικών (p-value>0.05). Ενώ μέσω της τιμής του odds ratio έχω ότι:

$$odds\ ratio = 0.8,$$

το οποίο πάλι δεν είναι στατιστικά σημαντικό αφού το 95% δ.ε περιλαμβάνει την μονάδα (0.06, 5.59).

5.2 Σχέσεις ανάμεσα στους κυτταρικούς δείκτες

Στην παρούσα ενότητα θα ελέγξουμε το αν υπάρχει κάποια σχέση μεταξύ των κυτταρικών δεικτών και πιο συγκεκριμένα ενδιαφερόμαστε για το αν υπάρχει σχέση ανάμεσα στα E2F1 και E2F4 με την κυκλίνη E και μετά εάν η Κυκλίνη E επηρεάζει τον δείκτη κυτταρικού πολλαπλασιασμού Ki67.

5.2.1 Σχέση ανάμεσα στις μεταβλητές E2F1 και E2F4 με την Κυκλίνη E

Η ανάλυση θα γίνει μέσω του εργαλείου της παλινδρόμησης με εξαρτημένη μεταβλητή την Κυκλίνη E και ανεξάρτητες τα E2F1 και E2F4. Το μοντέλο που θα χρησιμοποιήσουμε είναι το

$$Y = \beta_0 + \beta_1 * E2F1 + \beta_2 * E2F4$$

Πριν προχωρήσουμε στην παρουσίαση των συντελεστών παλινδρόμησης και την περαιτέρω ανάλυση θα πρέπει να δούμε αν ισχύουν οι προϋποθέσεις της παλινδρόμησης και αυτό θα το ελέγξουμε από τα κατάλοιπα της παλινδρόμησης.

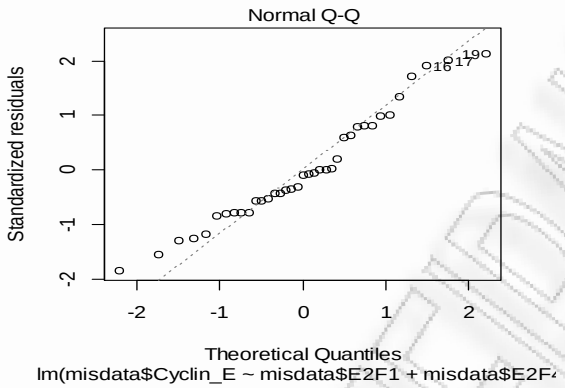
1. Κανονικότητα

   Η πρώτη προϋπόθεση που θα πρέπει να ισχύει είναι αυτή της κανονικότητας, δηλαδή ότι τα κατάλοιπα της παλινδρόμησης προέρχονται από την κανονική κατανομή. Σύμφωνα με το qqplot παρατηρούμε ότι η υπόθεση πρέπει λογικά να ισχύει

Πίνακας 5.18 (Ελεγχος Shapiro-Wilk)

Data	κατάλοιπα
W	0.963
p-value	0.2564

Διάγραμμα 5.6 (qqplot των καταλοίπων)

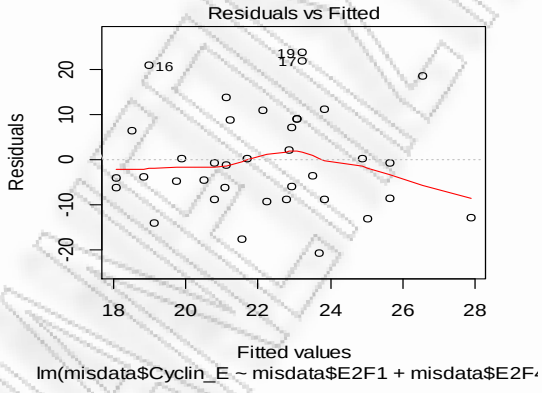


Μέσω του S-W τεστ η υπόθεση αυτή επιβεβαιώνεται.

2. Ανεξαρτησία των καταλοίπων

Η δεύτερη προϋπόθεση που πρέπει να ισχύει είναι αυτή της ανεξαρτησίας των καταλοίπων. Διαγραμματικά μπορούμε να πούμε ότι ισχύει η ανεξαρτησία καθώς το σχήμα έχει την προβλεπόμενη μορφή που δεν δηλώνει κάποιο συγκεκριμένο μοτίβο (pattern).

Διάγραμμα 5.7 (Ανεξαρτησία των καταλοίπων)



Πίνακας 5.19 (Τεστ ροών)

Runs Test - Two sided	
Data	κατάλοιπα
W	0.8388
p-value	0.4016

Το παραπάνω αποτέλεσμα επιβεβαιώνεται και από το τεστ ροών (runs test).

Από την στιγμή που ισχύουν τα παραπάνω μπορούμε να σχολιάσουμε τα αποτελέσματα της παλινδρόμησης.

Πίνακας 5.20 (Συντελεστές παραμέτρων)

Coefficients				
	Estimate	Std.	Error t value	Pr(>|t|)
(Intercept)	28.65955	5.55159	5.162	1.06e-05
E2F1	-0.06017	0.09029	-0.666	0.510
E2F4	-0.09095	0.07995	-1.138	0.263

Πίνακας 5.21 (Πίνακας ANOVA)

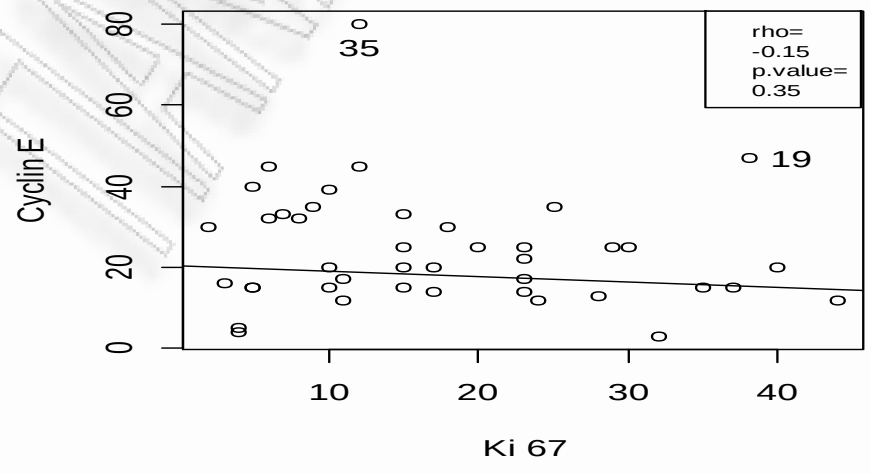
Analysis of Variance Table					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
E2F1	1	37.6	37.587	0.2871	0.5956
E2F4	1	169.4	169.441	1.2940	0.2633
Residuals	34	4452	130.941		

Από τους παραπάνω πίνακες βλέπουμε ότι οι συντελεστές των δεικτών E2F1 και E2F4 δεν είναι στατιστικά σημαντικοί (p-value>0.05) κάτι το οποίο επιβεβαιώνεται και από τον πίνακα ANOVA. Επίσης από τον πίνακα ANOVA και την στήλη Sum Sq παρατηρώ ότι η μεταβλητότητα που ερμηνεύει η κάθε μεταβλητή είναι πάρα πολύ μικρή σε σχέση με την συνολική. Άρα με βάση τα δεδομένα μας δεν μπορούμε να ισχυριστούμε ότι οι δείκτες E2F1 και E2F4 επηρεάζουν την κυκλίνη Ε.

5.2.2 Σχέση Ki 67 με Κυκλίνη Ε

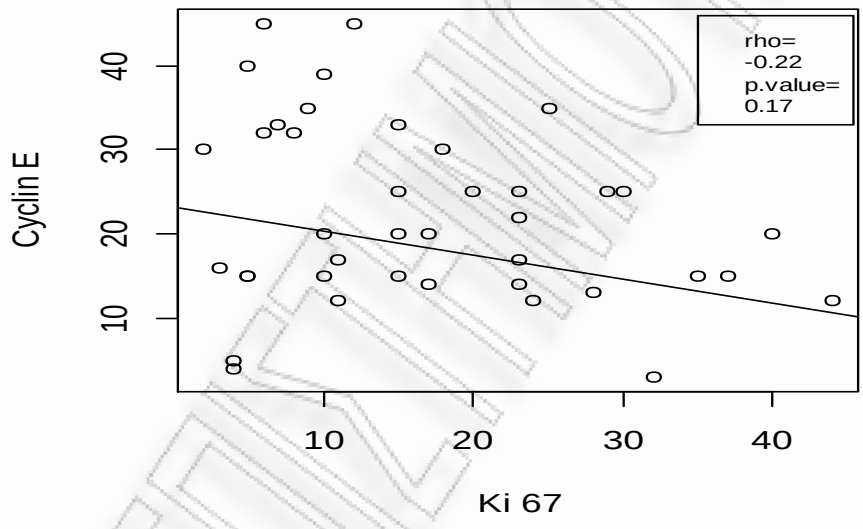
Επόμενο στοιχείο που μας ενδιαφέρει να ελέγξουμε είναι κατά πόσο και αν ο δείκτης κυτταρικού πολλαπλασιασμού Ki 67 σχετίζεται με την Κυκλίνη Ε.

Διάγραμμα 5.8 (Διάγραμμα διασποράς Ki 67 με Κυκλίνη Ε)



Σύμφωνα με το διάγραμμα αλλά και τον έλεγχο του Spearman βλέπουμε ότι δεν υπάρχει κάποια στατιστικά σημαντική συσχέτιση μεταξύ των 2 μεταβλητών (p-value>0.05). Παρόλα αυτά αν παρατηρήσουμε το διάγραμμα βλέπουμε ότι υπάρχουν δύο ακραίες παρατηρήσεις, η 19^{η} και η 35^{η}. Ειδικά η 19^{η} ξεφεύγει από το γενικότερο μοτίβο του σχήματος καθώς είναι η μόνη παρατήρηση όπου και οι δύο μεταβλητές παίρνουνε ταυτόχρονα υψηλές τιμές ενώ ο γενικός κανόνας είναι ότι για μικρές τιμές του Ki 67 έχω υψηλές τιμές για την Κυκλίνη Ε και το αντίστροφο. Οπότε αν απομακρύνουμε τις δύο αυτές παρατηρήσεις το διάγραμμα είναι το ακόλουθο.

Διάγραμμα 5.9 (Νέο διάγραμμα διασποράς Ki 67 με Κυκλίνη Ε)

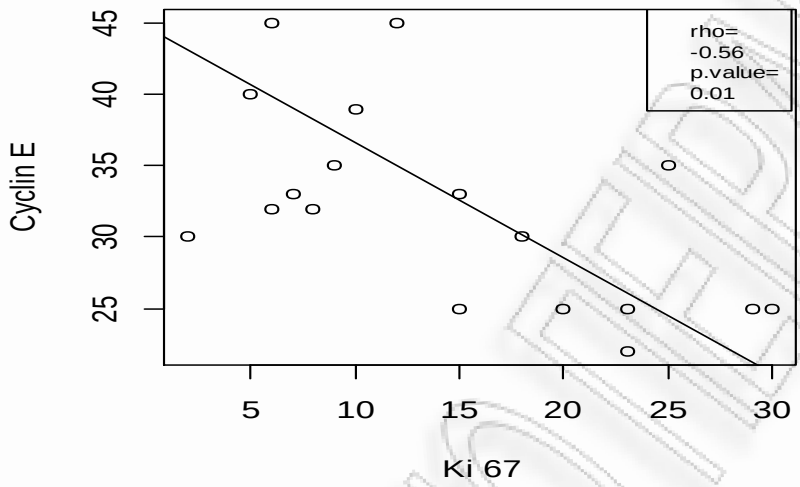


Παρατηρούμε ξανά ότι αν και η αρνητική συσχέτιση μεταξύ των μεταβλητών έχει αυξηθεί εντούτοις πάλι δεν μπορεί να θεωρηθεί στατιστικά σημαντική (p-value>0.05).

Όπως και πριν έτσι και εδώ αντί της συνεχούς περίπτωσης σχετικά με την Κυκλίνη έχει ενδιαφέρον και η συσχέτιση της με το Ki67 όταν η Κυκλίνη είναι πάνω ή κάτω από ένα όριο και πιο συγκεκριμένα τις τιμές 20% και 30% που είδαμε και πριν οπότε αρχικά θέτοντας ως όριο την τιμή 20% έχουμε.

1. Κυκλίνη μεγαλύτερη του 20%

Διάγραμμα 5.10 (Διάγραμμα διασποράς Ki 67 με Κυκλίνη Ε, περ.1)



Εδώ φαίνεται καθαρά ότι υπάρχει μια στατιστική σημαντική (p-value<0.05) αρνητική συσχέτιση μεταξύ του δείκτη Ki67 και της Κυκλίνης Ε όταν η δεύτερη παίρνει τιμές μεγαλύτερες του 20%. Στατιστικά σημαντικό αποτέλεσμα θα έχουμε αν τρέξουμε και μια παλινδρόμηση μεταξύ των δύο μεταβλητών.

Πίνακας 5.22 (συντελεστών των παραμέτρων)

Coefficients				
	Estimate	Std.	Error t value	Pr(>|t|)
(Intercept)	38.7276	8.1881	4.73	0.000227
Cyclin E	-0.7536	0.2503	-3.01	0.008297

Πίνακας 5.23 (Πίνακας ANOVA)

Analysis of Variance Table					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cyclin E	1	468.01	468.01	9.0627	0.008297
Residuals	16	826.27	51.64		

Ενώ σύμφωνα με το μοντέλο ερμηνεύεται το 32.17% της συνολικής μεταβλητότητας, επίσης οι υποθέσεις της παλινδρόμησης ισχύουν.

Πίνακας 5.24 (Έλεγχος υποθέσεων)

Runs Test - Two sided		Shapiro-Wilk normality test
Data	Residuals	
statistic	0.4859	0.976
p-value	0.627	0.8998

Στο σημείο αυτό θα δώσουμε και ένα παράδειγμα πως θα μπορούσαμε να χρησιμοποιήσουμε την stepwise μέθοδο παλινδρόμησης για να επιλέξουμε ποιες μεταβλητές θα βάλουμε στο μοντέλο.

Πίνακας 5.25 (Τα βήματα της stepwise μεθόδου)

Start: AIC=65.37

data1$Ki67 ~ data1$Cyclin_E + data1$E2F1 + data1$E2F4

	Df	Sum of Sq	RSS	AIC
- data1$E2F1	1	28.91	716.23	63.989
- data1$E2F4	1	59.46	746.78	64.616
<none>			687.32	65.371
- data1$Cyclin_E	1	515.55	1202.87	71.766

Step: AIC=63.99

data1$Ki67 ~ data1$Cyclin_E + data1$E2F4

	Df	Sum of Sq	RSS	AIC
<none>			716.23	63.989
- data1$E2F4	1	108.00	824.23	64.096
+ data1$E2F1	1	28.91	687.32	65.371
- data1$Cyclin_E	1	497.79	1214.02	69.905

Ο παραπάνω πίνακας μας λέει ότι το αρχικό μας μοντέλο το οποίο είναι το:

$$Ki67 \sim Cyclin\ E + E2F1 + E2F4 + \varepsilon\ ,$$

για το οποίο η τιμή του κριτηρίου AIC, βάση του οποίου θα λάβουμε την απόφαση μας είναι AIC=65.37. Σε πρώτη φάση αν από αυτό το μοντέλο αφαιρέσουμε κάποια από τις E2F1,4 τότε η τιμή του κριτηρίου πέφτει οπότε μπορούμε να τις διώξουμε από το μοντέλο ενώ αν διώξουμε την Κυκλίνη τότε η τιμή αυξάνεται οπότε πρέπει να μείνει. Σε δεύτερη φάση το μοντέλο μας είναι το

$$Ki67 \sim Cyclin\ E + E2F4 + \varepsilon$$

Εδώ παρατηρούμε πλέον ότι όποια κίνηση και αν κάνουμε η τιμή του AIC αυξάνεται οπότε αυτό είναι και το τελικό μας μοντέλο και αν κάνουμε την παλινδρόμηση έχω τον παρακάτω πίνακα.

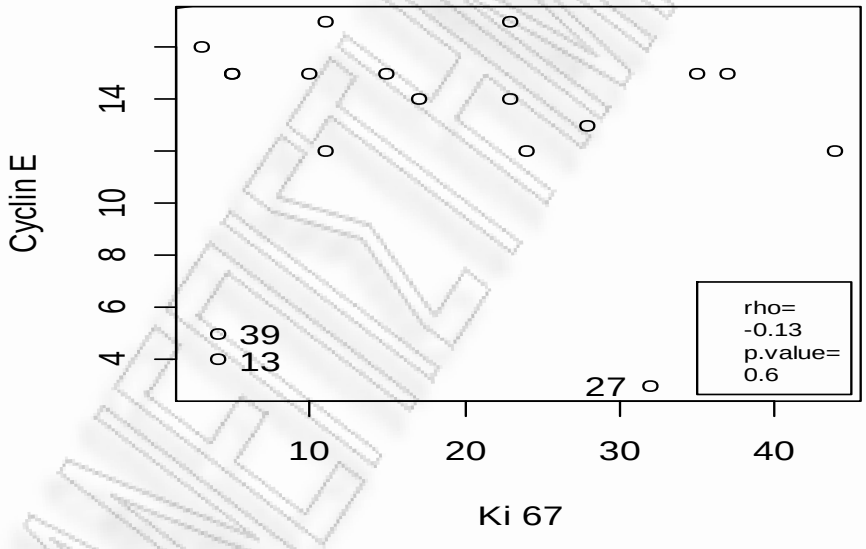
Πίνακας 5.26 (Συντελεστές των παραμέτρων)

Coefficients				
	Estimate	Std.	Error t value	Pr(>|t|)
(Intercept)	46.83469	11.10808	4.216	0.0012
Cyclin E	-0.85106	0.29469	-2.888	0.0136
E2F4	-0.11156	0.08293	-1.345	0.2035

Βλέπουμε ότι μόνο η Κυκλίνη είναι στατιστικά σημαντική (p-value<0.05) ενώ ερμηνεύεται το 36.47% της συνολικής μεταβλητότητας. Μια υποσημείωση που πρέπει να κάνουμε είναι ότι δεν έχουν οι συμπεριληφθεί οι δύο ακραίες παρατηρήσεις που παρατηρήσαμε στην αρχή.

2. Κυκλίνη μικρότερη του 20%

Διάγραμμα 5.11 (Διάγραμμα διασποράς Ki 67 με Κυκλίνη Ε, περ.2)

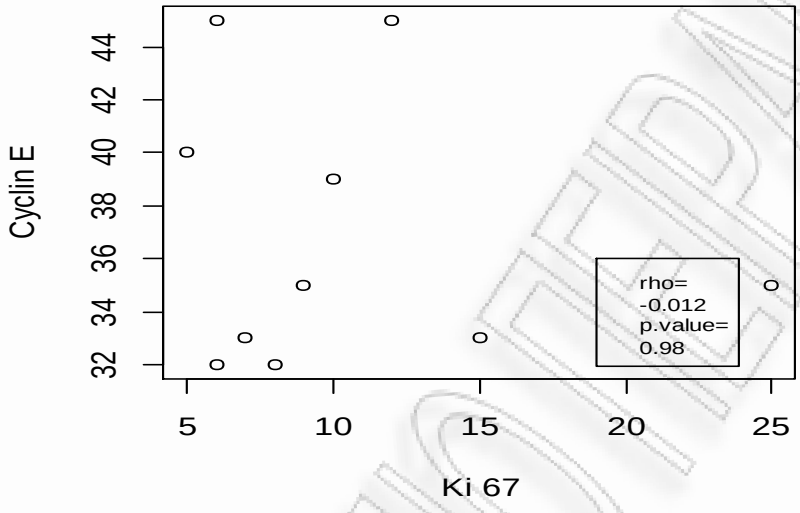


Παρατηρούμε ότι δεν υπάρχει στατιστικά σημαντική συσχέτιση μεταξύ των δύο μεταβλητών, ενώ επίσης παρατηρούμε και τρεις αρκετά ακραίες παρατηρήσεις οι οποίες όμως δεν επηρεάζουν το τελικό αποτέλεσμα.

Η δεύτερη κατηγορία με την οποία θα ασχοληθούμε είναι θέτοντας ως όριο την τιμή 30% για την Κυκλίνη Ε οπότε δουλεύοντας με τον ίδιο τρόπο όπως πριν έχουμε.

3. Κυκλίνη μεγαλύτερη του 30%

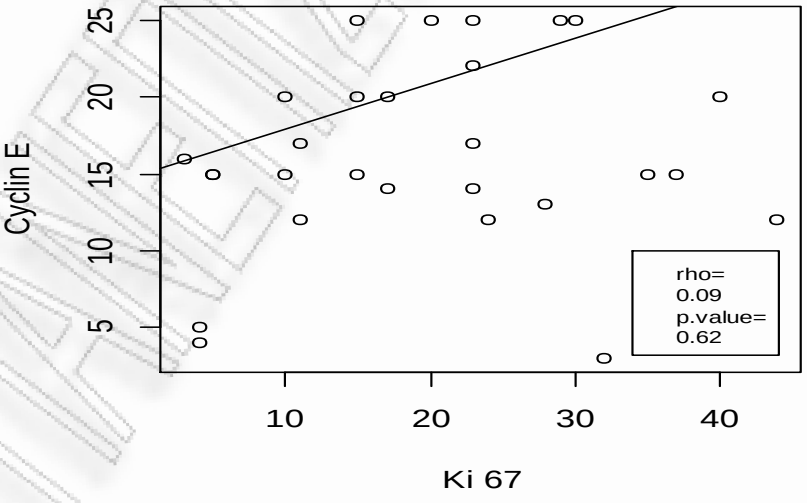
Διάγραμμα 5.12 (Διάγραμμα διασποράς Ki 67 με Κυκλίνη Ε, περ.3)



Ούτε σε αυτήν την περίπτωση παρατηρούμε κάποια στατιστικά σημαντική συσχέτιση μεταξύ των δύο μεταβλητών (p-value>0.05).

4. Κυκλίνη μικρότερη του 30%

Διάγραμμα 5.13 (Διάγραμμα διασποράς Ki 67 με Κυκλίνη Ε, περ.4)



Ούτε εδώ παρατηρούμε κάποια στατιστικά σημαντική συσχέτιση μεταξύ των δύο μεταβλητών (p-value>0.05).

Τα παραπάνω αποτελέσματα παρουσιάζονται συνοπτικά στον παρακάτω πίνακα.

Πίνακας 5.27 (Συγκεντρωτικά οι συντελεστές συσχέτισης)

Cyclin E	>20	<20	>30	<30
Spearman's rho	-0.56	-0.13	-0.012	0.09
p-value	0.01	0.6	0.98	0.62

Τέλος σχετικά με την μελέτη της σχέσης μεταξύ των δύο μεταβλητών σε ορισμένες μελέτες έχουν θέσει και στον δείκτη Ki67 μια τιμή κατώφλι και συγκεκριμένα την τιμή 10% (Ioachim et al). Οπότε υιοθετώντας το παραπάνω έχουμε ότι.

Πίνακας 5.28 (Exact Fisher test)

Ki67		Κυκλίνη	
		<20	≥20
	<10	7	8
	>10	15	12
p-value	0.7488		
95% Conf.int.	0.16, 2.97		
sample estimates	0.46 0.55		

Πάλι δεν μπορούμε να απορρίψουμε την υπόθεση ότι δεν υπάρχει διαφορά μεταξύ της πιθανότητας η κυκλίνη να είναι κάτω από 20% για τους ασθενείς των οποίων ο δείκτης Ki67 είναι κάτω από 10% (p-value>0.05). Ενώ μέσω της τιμής του odds ratio έχω ότι:

$$odds\ ratio = 0.7,$$

το οποίο πάλι δεν είναι στατιστικά σημαντικό αφού το 95% δ.ε περιλαμβάνει την μονάδα (0.16, 2.97).

Επαναλαμβάνοντας την διαδικασία θέτοντας ως όριο την τιμή 30% για την Κυκλίνη Ε έχουμε.

Πίνακας 5.29 (Exact Fisher test)

Ki67		Κυκλίνη	
		<30	≥30
	<10	8	7
	>10	22	5
Fisher's Exact Test			
p-value	0.07		
95% Conf.int.	0.05, 1.30		
sample estimates	0.53 0.81		

Πάλι, αν και οριακά, δεν μπορούμε να απορρίψουμε την υπόθεση ότι δεν υπάρχει διαφορά μεταξύ της πιθανότητας η κυκλίνη να είναι κάτω από 30% για τους ασθενείς των οποίων ο δείκτης Ki67 είναι κάτω από 10% (p-value>0.05). Ενώ μέσω της τιμής του odds ratio έχω ότι:

$$odds\ ratio = 0.27,$$

το οποίο πάλι δεν είναι στατιστικά σημαντικό αφού το 95% δ.ε περιλαμβάνει την μονάδα (0.05, 1.30). Πάντως στην συγκεκριμένη περίπτωση καλό είναι να έχουμε στο μυαλό μας ότι ίσως να υπάρχει κάποια σχέση η οποία όμως να μην είναι εμφανείς λόγω του μικρού δείγματος.

Κεφάλαιο 6

Ανάλυση σε επίπεδο ατόμου

Στο τρίτο και τελευταίο κομμάτι της ανάλυσης μας φεύγουμε από το επίπεδο του κυττάρου και κοιτάμε πλέον τι γίνεται σε επίπεδο ατόμου (ή οργάνου). Θα μας απασχολήσουν οι σχέσεις που αναπτύσσονται μεταξύ των Ki67, Κυκλίνης Ε, E2F1 και E2F4 με τους κλινικό-παθολογικούς δείκτες grade, stage και risk.

6.1 Ανάλυση σχετικά με τον βαθμό κακοήθειας (grade) του καρκίνου

Σε αυτήν την παράγραφο θα ερευνήσουμε την σχέση του βαθμού του καρκίνου με τους δύο δείκτες μέσω παραμετρικών ή μη ελέγχων, ανάλογα με το αν ισχύει η κανονικότητα των δεδομένων για κάθε κατηγορία της μεταβλητής grade και της λογιστικής παλινδρόμησης.

6.1.1 Σχέση grade με δείκτη Ki 67

Στον παρακάτω πίνακα παίρνουμε μια αρχική ιδέα σχετικά με τις τιμές του δείκτη Ki67 ανάλογα με το grade του καρκίνου.

Πίνακας 6.1 (Περιγραφικά μέτρα)

Grade	Μέση τιμή	Τυπική απόκλιση	Αριθμός ασθενών
G1	10.64	7.39	11
G2	14.38	10.14	16
G3	23.28	11.21	18

Από τον παραπάνω πίνακα βλέπουμε ότι υπάρχει κάποια διαφορά μεταξύ των τριών επιπέδων του grade και ειδικότερα παρατηρούμε ότι οι τιμές του δείκτη Ki67 για το επίπεδο G3 είναι αρκετά μεγαλύτερες σε σχέση με τα άλλα δύο επίπεδα. Επειδή το πλήθος των παρατηρήσεων ανά κατηγορία είναι αρκετά μικρό δεν έχει μεγάλο νόημα να κάνουμε έλεγχο κανονικότητας και να αποφασίσουμε για το αν θα χρησιμοποιήσουμε έναν παραμετρικό ή μη τεστ για να ελέγξουμε την ισότητα των τριών γκρουπ αλλά θα στραφούμε κατευθείαν στον μη παραμετρικό έλεγχο. Οπότε τον έλεγχο για το αν οι τιμές του δείκτη διαφέρουν ανάμεσα στα γκρουπ θα τον κάνουμε μέσω του τεστ Kruskal-Wallis και έχουμε ότι πράγματι υπάρχει διαφορά μεταξύ των ομάδων ($p\text{-value}<0.05$)

Πίνακας 6.2 (Έλεγχος Kruskal-Wallis)

Kruskal-Wallis chi-squared	10.5054
p-value	0.005233

Επόμενο στάδιο είναι να δούμε πια επίπεδα διαφέρουν μεταξύ τους. Σύμφωνα με τον παρακάτω πίνακα βλέπουμε ότι το επίπεδο G3 διαφέρει από τα άλλα δύο (p-value <0.05) ενώ για τα G1, G2 δεν μπορούμε να υποθέσουμε ότι διαφέρουν.

Πίνακας 6.3 (Έλεγχος των γκρουπ ανά 2)

Pairwise comparisons using Wilcoxon rank sum test		
	G1	G2
G2	0.322	-
G3	0.016	0.026

6.1.2 Σχέση grade με Κυκλίνη Ε

Δουλεύοντας όπως και πριν από τον παρακάτω πίνακα παίρνουμε μια αρχική ιδέα σχετικά με τις τιμές της Κυκλίνης Ε ανάλογα με το grade του καρκίνου.

Πίνακας 6.4 (Περιγραφικά μέτρα)

Grade	Μέση τιμή	Τυπική απόκλιση	Αριθμός ασθενών
G1	27.73	20.62	11
G2	22.53	9.4	16
G3	23.06	13.21	18

Από τον πίνακα βλέπουμε ότι με εξαίρεση ίσως τις τιμές της Κυκλίνης Ε για την πρώτη ομάδα δεν φαίνεται να υπάρχει κάποια διαφορά. Οπότε μέσω του τεστ Kruskal-Wallis έχουμε ότι πράγματι δεν υπάρχει διαφορά μεταξύ των ομάδων.

Πίνακας 6.5 (Έλεγχος Kruskal-Wallis)

Kruskal-Wallis chi-squared	0.1765
p-value	0.9155

6.1.3 Προσέγγιση με λογιστική παλινδρόμηση

Έχουμε την εξαρτημένη μας μεταβλητή που είναι το grade και τις 2 ανεξάρτητες που είναι οι δείκτες Ki 67 και Κυκλίνη Ε. Η μεταβλητή grade είναι μια διατακτική μεταβλητή με 3 επίπεδα ενώ οι δείκτες είναι συνεχείς μεταβλητές. Εφαρμόζοντας την διατακτική λογιστική παλινδρόμηση (βλ. παράρτημα Α) έχω τα εξής:

Πίνακας 6.6 (Έλεγχος σημαντικότητας του μοντέλου)

Model Likelihood Ratio Test	
LR X^2	11.38
d.f	2
p-value	0.0034

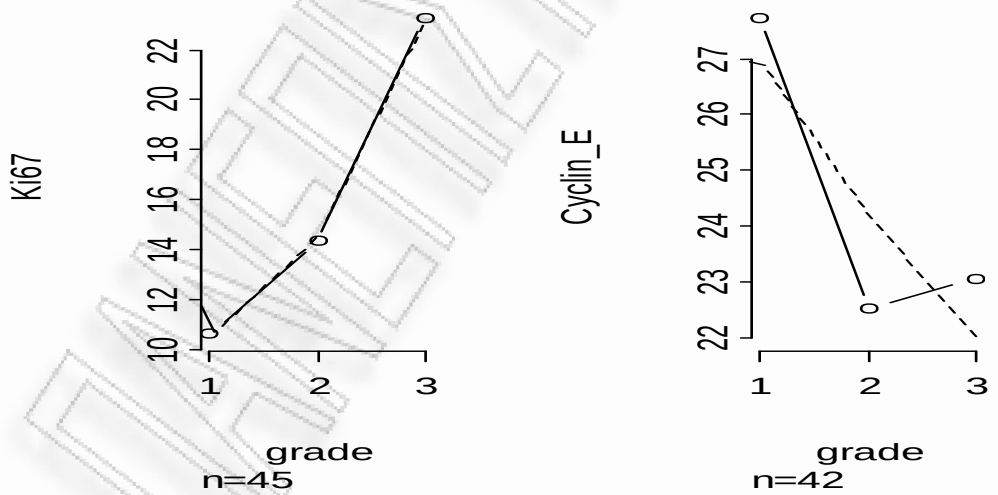
Από τον παραπάνω πίνακα καταλήγουμε στο αποτέλεσμα ότι το μοντέλο μας είναι στατιστικά σημαντικό, δηλαδή διαφέρει από το μοντέλο μόνο με τον σταθερό όρο.

Πίνακας 6.7 (Συντελεστές παραμέτρων)

	Coef.	S.E	Wald Z	P=value
grade≥2	-0.1895	0.8427	-0.22	0.8221
grade≥3	-2.0690	0.9109	-2.27	0.0231
Ki 67	0.0971	0.0330	2.94	0.0033
Κυκλίνη Ε	-0.0076	0.0227	-0.33	0.7387

Πριν περάσουμε στην παρουσίαση του μοντέλου και διάφορων στοιχείων σχετικά με αυτό θα πρέπει να ελέγξουμε αν ισχύει η υπόθεση των αναλογικών odds (βλ. 4.4.8.2) το οποίο θα γίνει με γραφικό τρόπο αλλά και στατιστικά.

Διάγραμμα 6.1 (Έλεγχος υπόθεσης των αναλογικών odds, 1ος τρόπος)

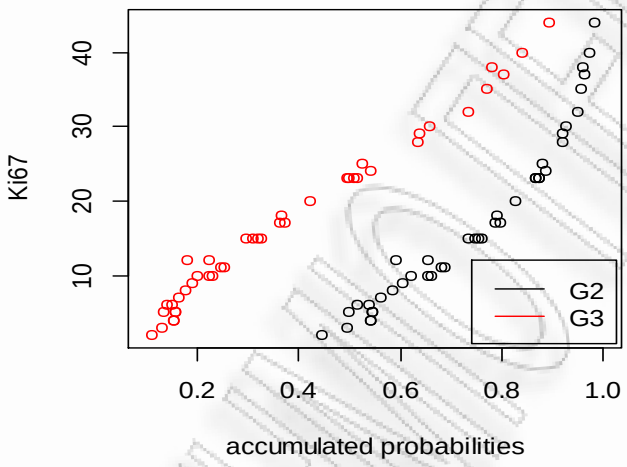


Στο παραπάνω διάγραμμα βλέπουμε το εξής για κάθε μια ανεξάρτητη μεταβλητή αναπαριστάται ο μέσος της μεταβλητής αυτής έναντι τον επιπέδων του παράγοντα grade (η έντονη γραμμή) όπως επίσης και η αναμενόμενη τιμή της μεταβλητής για κάθε επίπεδο του παράγοντα κάτω από την υπόθεση των αναλογικών odds (η διακεκομμένη γραμμή). Για να ισχύει η υπόθεση των αναλογικών odds πρέπει οι δύο γραμμές να μην διαφέρουν . Στην δικιά μας περίπτωση για το Ki67 βλέπουμε πως

αυτό ισχύει, ενώ δεν μας ενδιαφέρει τι γίνεται για την Κυκλίνη Ε αφού δεν είναι στατιστικά σημαντική (πάντως εδώ δεν γίνεται να την δεχτούμε καθώς οι δύο γραμμές δεν συμπίπτουν ούτε είναι σε γενικές γραμμές παράλληλες).

Εναλλακτικό σχήμα είναι αυτό που ισχύει στην περίπτωση των αναλογικών odds δηλαδή η παραλληλία των σωρευτικών πιθανοτήτων.

Διάγραμμα 6.2 (Έλεγχος υπόθεσης των αναλογικών odds, 2ος τρόπος)



Παρατηρούμε πως η υπόθεση μας πρέπει να ισχύει καθώς και οι δυο γραμμές έχουν παρόμοιο σχήμα και υπάρχει απλώς η μετατόπιση ως προς τον οριζόντιο άξονα.

Στατιστικά στο παραπάνω αποτέλεσμα μπορούμε να καταλήξουμε με δύο τρόπους. Πρώτον, να ελέγξουμε ένα μοντέλο υπό την υπόθεση των αναλογικών odds έναντι ενός που δεν ισχύει η συγκεκριμένη υπόθεση. Εναλλακτικά, εάν κάναμε έναν έλεγχο πιθανοφανειών συγκρίνοντας τις αποκλίσεις (deviances) μεταξύ ενός μοντέλου όπου ισχύει η υπόθεση των αναλογικών odds και ενός μοντέλου πολλαπλής λογιστικής παλινδρόμησης, όπου δηλαδή δεν υποθέτουμε ότι η μεταβλητή απόκρισης είναι διατάξιμη.

Σύμφωνα με τον πρώτο τρόπο έχουμε:

Πίνακας 6.8 (Έλεγχος υπόθεσης των αναλογικών odds, 1^{ος} τρόπος)

Μοντέλο	απόκλιση	Βαθμοί ελευθερίας
Αναλογικά odds	79.86801	80
Όχι αναλογικά odds	78.93318	78

Με p-value= 0.63, οπότε όπως περιμέναμε δεν απορρίπτεται η αρχική μας υπόθεση.

Με τον δεύτερο τρόπο έχω:

Πίνακας 6.9 (Έλεγχος υπόθεσης των αναλογικών odds, $2^{ος}$ τρόπος)

Μοντέλο	απόκλιση	Βαθμοί ελευθερίας
Αναλογικά odds	79.86801	2
Πολυωνυμικό	79.40693	6

Με p-value= 0.98, οπότε πάλι δεν απορρίπτεται η αρχική μας υπόθεση και άρα ισχύει η υπόθεση των αναλογικών odds. Οπότε μπορούμε να συνεχίσουμε την ανάλυση σύμφωνα με τα αποτελέσματα του πίνακα 6.1.3.2. Σύμφωνα με αυτόν το μοντέλο μου είναι το:

$$logit[P(grade \leq 1)] = 0.1895 - 0.0971 * Ki67 \qquad (6.1)$$

και

$$logit[P(grade \leq 2)] = 2.069 - 0.0971 * Ki67 \qquad (6.2)$$

Επίσης δεν υπάρχει πρόβλημα πολυσυγγραμικότητας των 2 ανεξάρτητων μεταβλητών καθώς οι τιμές του συντελεστή VIF και για τις δύο μεταβλητές είναι πολύ κοντά στην μονάδα και πιο συγκεκριμένα

$$VIF(Ki67) = 1.01 \; και \; VIF(Κυκλίνης \; E) = 1.01$$

Τέλος η area under curve (AUC) του μοντέλου που δείχνει το ποσοστό σωστής ταξινόμησης είναι 72.8%

Πριν συνεχίσουμε να κάνουμε μια παρατήρηση σχετικά με τις τιμές που παίρνουμε από τον πίνακα και τις τιμές του μοντέλου. Ο πίνακας μας δίνει τους συντελεστές των $logit[P(grade \geq i)] \; για \; i = 2,3$ το οποίο δεν είναι η συνηθισμένη μορφή που παρουσιάσαμε στο κεφάλαιο 4. Για να το επαναφέρουμε στην μορφή που ξέρουμε πρέπει απλώς να αλλάξουμε τα πρόσημα των συντελεστών και παίρνουμε τις σχέσεις (6.1) και (6.2).

Από την στιγμή που καταλήξαμε στο μοντέλο θα ήταν χρήσιμο να παρουσιάσουμε κάποια χρήσιμα στοιχεία σχετικά με τον δείκτη Ki67 και τις πιθανότητες που έχουμε για τις τρεις κατηγορίες για διάφορες τιμές του δείκτη. Αρχικά από την στιγμή που ο συντελεστής είναι αρνητικός αυτό σημαίνει ότι ο βαθμός κακοήθειας του καρκίνου είναι στοχαστικά αύξουσα συνάρτηση του Ki67, δηλαδή όσο αυξάνεται η τιμή του δείκτη τόσο πιο πιθανό είναι ο ασθενής να έχει έναν φτωχά διαφοροποιούμενο καρκίνο (δηλ. καρκίνος υψηλού βαθμού). Επιπλέον η τιμή του δείκτη είναι:

$$\hat{\beta} = -0.0971,$$

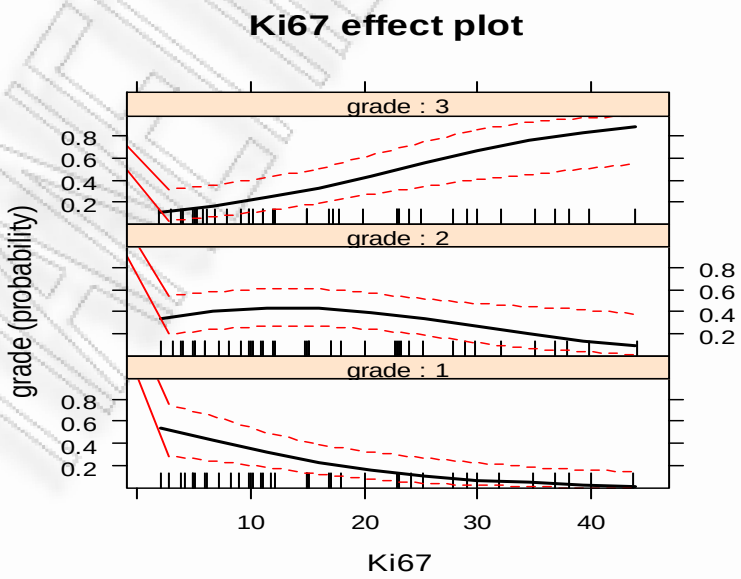
που σημαίνει ότι αν η τιμή του δείκτη Ki67 αυξηθεί κατά μία μονάδα τότε το odds ο ασθενής να πάσχει από καρκίνο με grade=1 έναντι όλων των άλλων επιπέδων είναι το $e^{\widehat{\beta}} = 0.9$ του αντίστοιχου odds για την αρχική τιμή του Ki67. Πιο καθαρά είναι τα συμπεράσματα αν δούμε τους πίνακες πιθανοτήτων για διάφορες τιμές του ki67 και θέτοντας ως τιμή για την Κυκλίνη την μέση της τιμή. Ο πρώτος πίνακας παρουσιάζει τις πιθανότητες $P(grade = i),\ i = 1,2,3$ και ο δεύτερος τις πιθανότητες $P(grade \geq i),\ i = 2,3$

Πίνακας 6.10 (Πιθανότητες ταξινόμησης για διάφορες τιμές του Ki67)

Ki67	grade=1	grade=2	grade=3
2	0.54441581	0.3423076	0.1132766
7	0.42376081	0.4043382	0.1719010
12	0.31155917	0.4362058	0.2522350
17	0.21783522	0.4281049	0.3540598
22	0.14631344	0.3825929	0.4710936
27	0.09540992	0.3131959	0.5913942
32	0.06095166	0.2373877	0.7016607
37	0.03840993	0.1689843	0.7926058
42	0.02399180	0.1147008	0.8613074

Βλέπουμε ότι όσο αυξάνεται η τιμή του δείκτη Ki67 τόσο αυξάνεται και η πιθανότητα ο ασθενής να πάσχει από καρκίνο υψηλού βαθμού. Διαφωτιστικό είναι και το παρακάτω διάγραμμα πιθανοτήτων για τα τρία επίπεδα.

Διάγραμμα 6.3 (Πιθανότητες ταξινόμησης για διάφορες τιμές του Ki67)



Σε παρόμοια αποτελέσματα καταλήγουμε παρατηρώντας τον πίνακα με τις σωρευτικές πιθανότητες.

Πίνακας 6.11 (Σωρευτικές πιθανότητες για διάφορες τιμές του Ki67)

Ki67	grade≥2	grade≥3
2	0.4555842	0.1132766
7	0.5762392	0.1719010
12	0.6884408	0.2522350
17	0.7821648	0.3540598
22	0.8536866	0.4710936
27	0.9045901	0.5913942
32	0.9390483	0.7016607
37	0.9615901	0.7926058
42	0.9760082	0.8613074

6.2 Ανάλυση σχετικά με το στάδιο (stage) του καρκίνου

Την ίδια διαδικασία που ακολουθήθηκε πριν θα ακολουθηθεί και εδώ σχετικά με το στάδιο του καρκίνου με στόχο να δούμε πως επηρεάζεται από τις δύο ανεξάρτητες μεταβλητές μας .

6.2.1 Σχέση stage με δείκτη Ki 67

Στον παρακάτω πίνακα παίρνουμε μια αρχική ιδέα σχετικά με τις τιμές του δείκτη Ki67 ανάλογα με το στάδιο του καρκίνου.

Πίνακας 6.12 (Περιγραφικά μέτρα)

Stage	Μέση τιμή	Τυπική απόκλιση	Αριθμός ασθενών
pTa	11.95	7.5	21
pT1	19	11.24	6
pT2	26.25	11.23	8

Από τον παραπάνω πίνακα βλέπουμε ότι υπάρχει διαφορά μεταξύ των τριών επιπέδων του stage. Τον έλεγχο για το αν οι τιμές του δείκτη διαφέρουν ανάμεσα στα γκρουπ θα τον κάνουμε μέσω του τεστ Kruskal-Wallis και έχουμε ότι πράγματι υπάρχει διαφορά μεταξύ των ομάδων (p-value<0.05).

Πίνακας 6.13 (Έλεγχος Kruskal-Wallis)

Kruskal-Wallis chi-squared	9.9017
p-value	0.007077

Μέσω των Wilcoxon test έχω ότι τα μόνα επίπεδα που διαφέρουν μεταξύ τους είναι τα Pta και Pt2 (p-value<0.05).

Πίνακας 6.14 (Έλεγχος των γκρουπ ανά δύο)

Pairwise comparisons using Wilcoxon rank sum test		
	Pta	Pt1
Pt1	0.191	-
Pt2	0.012	0.219

6.2.2 Σχέση stage με Κυκλίνη Ε

Δουλεύοντας όπως και πριν από τον παρακάτω πίνακα παίρνουμε μια αρχική ιδέα σχετικά με τις τιμές της Κυκλίνης Ε ανάλογα με το στάδιο του καρκίνου.

Πίνακας 6.15 (Περιγραφικά μέτρα)

stage	Μέση τιμή	Τυπική απόκλιση	Αριθμός ασθενών
pTa	23.90	16.82	20
pT1	29.8	15.27	5
pT2	20.62	9.92	8

Από τον πίνακα βλέπουμε ότι με εξαίρεση ίσως τις τιμές της Κυκλίνης Ε για την δεύτερη ομάδα δεν φαίνεται να υπάρχει κάποια διαφορά. Τον έλεγχο για το αν οι τιμές του δείκτη διαφέρουν ανάμεσα στα γκρουπ θα τον κάνουμε μέσω του τεστ Kruskal-Wallis και έχουμε ότι πράγματι δεν υπάρχει διαφορά μεταξύ των ομάδων.

Πίνακας 6.16 (Έλεγχος Kruskal-Wallis)

Kruskal-Wallis chi-squared	1.2382
p-value	0.5384

6.2.3 Προσέγγιση με λογιστική παλινδρόμηση

Όπως και η μεταβλητή grade έτσι και η μεταβλητή stage είναι διατάξιμη με τρία επίπεδα. Οπότε θα εφαρμόσουμε και εδώ την διατακτική λογιστική παλινδρόμηση

Πίνακας 6.17 (Έλεγχος σημαντικότητας του μοντέλου)

Model Likelihood Ratio Test	
LR X^2	10.75
d.f	2
p-value	0.0046

Από τον παραπάνω πίνακα καταλήγουμε στο συμπέρασμα ότι το μοντέλο μας είναι στατιστικά σημαντικό (p-value<0.05). επίσης η προβλεπτική ικανότητα του μοντέλου είναι 76.8%.

Οι τιμές των συντελεστών παρουσιάζονται στον παρακάτω πίνακα

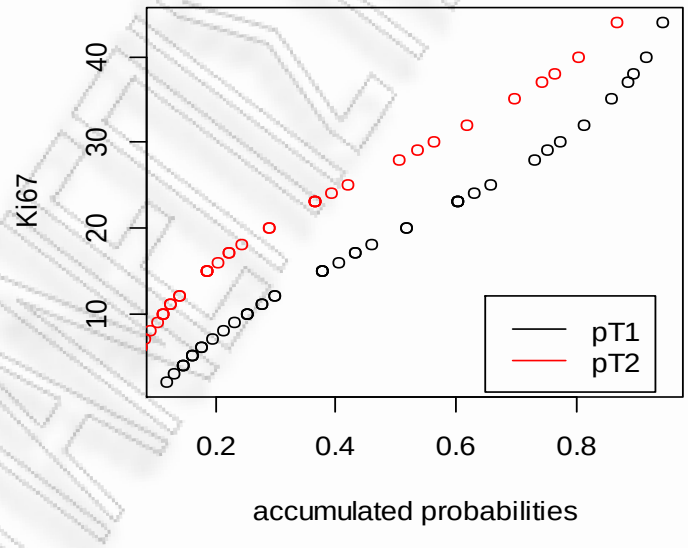
Πίνακας 6.18 (Συντελεστές παραμέτρων)

	Coef.	S.E	Wald Z	p-value
stage≥pT1	-2.2415	1.0643	-2.11	0.0352
stage≥pT2	-3.2164	1.1635	-2.76	0.0057
Ki 67	0.1157	0.0395	2.93	0.0034
Κυκλίνη Ε	-0.0085	0.0278	-0.31	0.7593

Όπου βλέπουμε πάλι ότι ο δείκτης Ki67 είναι στατιστικά σημαντικός ενώ η Κυκλίνη Ε όχι.

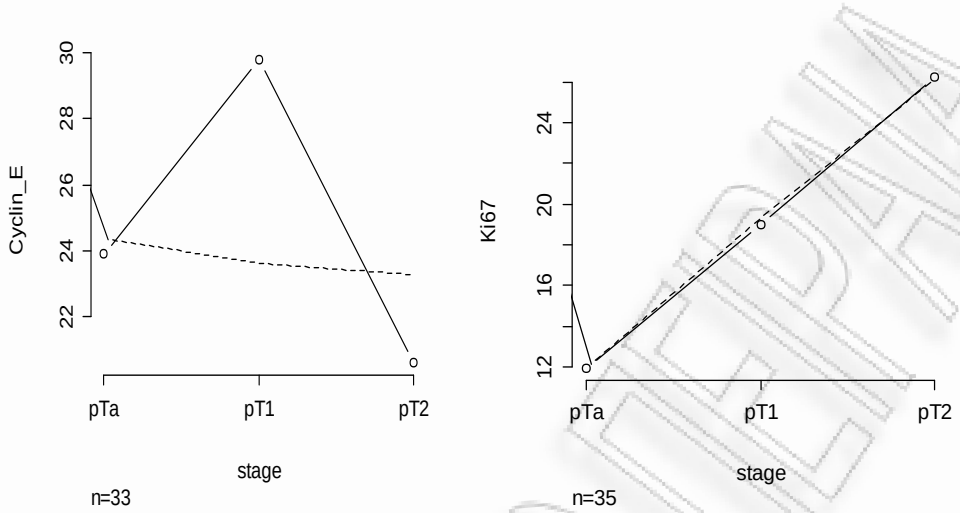
Όπως και πριν πρέπει πρώτα να ελέγξουμε αν ισχύει η υπόθεση των αναλογικών odds και στην συνέχεια να παρουσιάσουμε το μοντέλο. Τον έλεγχο θα τον κάνουμε όπως και προηγουμένως πρώτα σχηματικά και μετά μέσω στατιστικών ελέγχων.

Διάγραμμα 6.4 (Έλεγχος υπόθεσης των αναλογικών odds, 1ος τρόπος)



Φαίνεται ξεκάθαρα από το διάγραμμα των σωρευτικών πιθανοτήτων ότι οι 2 γραμμές έχουν παρόμοιο σχήμα κάτι το οποίο θέλουμε.

Διάγραμμα 6.5 (Έλεγχος υπόθεσης των αναλογικών odds, 2ος τρόπος)



Εδώ βλέπουμε ότι για το Ki67 οι δυο ευθείες είναι ουσιαστικά η μια πάνω στην άλλη οπότε πάλι δεν γίνεται να απορρίψουμε λογικά την υπόθεση των αναλογικών odds.

Σχετικά με τον στατιστικό έλεγχο στο συγκεκριμένο μοντέλο μπορεί να εφαρμοστεί μόνο ο δεύτερος τρόπος οπότε και έχουμε

Πίνακας 6.19 (Έλεγχος υπόθεσης των αναλογικών odds)

Μοντέλο	απόκλιση	Βαθμοί ελευθερίας
Αναλογικά odds	50.82946	2
Πολυωνυμικό	49.53182	6

Με p-value=0.86 πάλι δεν αποκλείουμε την υπόθεση των αναλογικών odds συνεπώς μπορούμε να συνεχίσουμε την ανάλυση σύμφωνα με τα αποτελέσματα του πίνακα 6.2.3.2. Σύμφωνα με αυτόν το μοντέλο μου είναι το:

$$logit[P(stage \leq pTa)] = 2.2415 - 0.1157 * Ki67$$

και

$$logit[P(stage \leq pT1)] = 3.2164 - 0.1157 * Ki67$$

Επίσης δεν υπάρχει πρόβλημα πολυσυγγραμικότητας των 2 ανεξάρτητων μεταβλητών καθώς οι τιμές του συντελεστή VIF και για τις δύο μεταβλητές ισούται με την μονάδα, ενώ τέλος η μέγιστη προβλεπτική ακρίβεια του μοντέλου (concordance) είναι 76.8%.

Όπως και για την μεταβλητή grade έτσι και εδώ από την στιγμή που ο συντελεστής είναι αρνητικός αυτό σημαίνει ότι το στάδιο του καρκίνου είναι στοχαστικά αύξουσα

συνάρτηση του Ki67, δηλαδή όσο αυξάνεται η τιμή του δείκτη τόσο πιο πιθανό είναι ο ασθενής να πάσχει από έναν καρκίνο σε ανεπτυγμένο στάδιο. Επιπλέον η τιμή του δείκτη είναι:

$$\hat{\beta} = -0.1157,$$

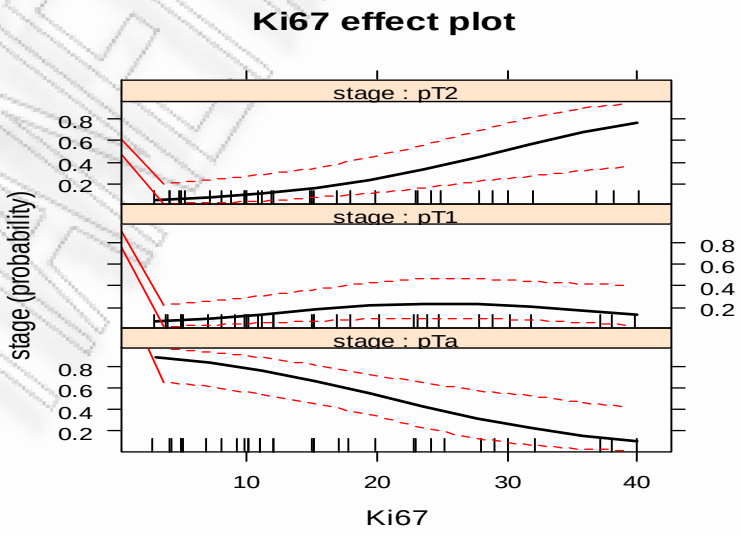
που σημαίνει ότι αν η τιμή του δείκτη Ki67 αυξηθεί κατά μία μονάδα τότε το odds ο ασθενής να πάσχει από καρκίνο με stage=pTa έναντι όλων των άλλων επιπέδων είναι το $e^{\hat{\beta}} = 0.89$ του αντίστοιχου odds για την αρχική τιμή του Ki67.

Πίνακας 6.20 (Πιθανότητες ταξινόμησης για διάφορες τιμές του Ki67)

Ki67	stage=pTa	stage =pT1	stage =pT2
2	0.9016298	0.05884232	0.03952787
7	0.8371032	0.09451287	0.06838389
12	0.7423455	0.14188820	0.11576628
17	0.6176439	0.19304802	0.18930803
22	0.4752514	0.23071580	0.29403284
27	0.3367724	0.23699574	0.42623187
32	0.2216035	0.20850867	0.56988784
37	0.1376459	0.15968742	0.70266671
42	0.0821402	0.10961206	0.80824774

Βλέπουμε ότι η πιθανότητες για τα στάδια pTa και pT2 μεταβάλλονται έντονα με την μεταβολή των τιμών του Ki67 ενώ σχετικά με το στάδιο pT1 βλέπουμε ότι αρχικά αυξάνεται η πιθανότητα ενώ στην συνέχεια ξαναπέφτει. Τα παραπάνω φαίνονται καθαρά και στο παρακάτω διάγραμμα.

Διάγραμμα 6.6 (ταξινόμησης για διάφορες τιμές του Ki67)



Από τον παρακάτω πίνακα βλέπουμε πως διαμορφώνονται οι σωρευτικές πιθανότητες και φαίνεται και από εδώ ότι το stage είναι αύξουσα συνάρτηση του

Ki67 καθώς βλέπουμε ότι καθώς αυξάνεται η τιμή του δείκτη Ki67 αυξάνεται και η πιθανότητα το stage να είναι άνω ενός επιπέδου έναντι κάτω από αυτό.

Πίνακας 6.21 (Σωρευτικές πιθανότητες για διάφορες τιμές του Ki67)

Ki67	Stage ≥pT1	Stage ≥pT2
2	0.09837019	0.03952787
7	0.16289676	0.06838389
12	0.25765448	0.11576628
17	0.38235605	0.18930803
22	0.52474864	0.29403284
27	0.66322761	0.42623187
32	0.77839651	0.56988784
37	0.86235413	0.70266671
42	0.91785980	0.80824774

6.3 Ανάλυση σχετικά με τον κίνδυνο (risk) που διατρέχει ο ασθενής

Στις δύο προηγούμενες παραγράφους είδαμε ότι ο βαθμός κακοήθειας (grade) όπως και το στάδιο (stage) του καρκίνου εξαρτάται από τον δείκτη Ki67 οπότε και ο κίνδυνος του ασθενή θα εξαρτάται από τον δείκτη αυτό. Στην συγκεκριμένη παράγραφο θα δούμε πως ο δείκτης κυτταρικού πολλαπλασιασμού Ki67 επηρεάζει την μεταβλητή risk.

6.3.1 Έλεγχος μέσω μη παραμετρικού ελέγχου

Αρχικά θα δούμε τι ισχύει με τις τιμές του Ki67 σε κάθε γκρουπ και στην συνέχεια θα ελέγξουμε την κανονικότητα μεταξύ των δύο γκρουπ ώστε να αποφασίσουμε πως θα ελέγξουμε τις διαφορές.

Πίνακας 6.22 (Περιγραφικά μέτρα)

Risk	Μέση τιμή	Τυπική απόκλιση	Αριθμός ασθενών
Low	11.4	6.86	25
high	24.05	11.66	20

Παρατηρώντας τις μέσες τιμές κάθε γκρουπ βλέπουμε ότι η τιμή του δείκτη για τους ασθενείς που ανήκουν στην κατηγορία ασθενών υψηλού ρίσκου έχουν αρκετά μεγαλύτερο δείκτη από ότι οι ασθενείς με χαμηλό ρίσκο.

Πίνακας 6.23 (Wilcoxon rank sum test)

W	88
p-value	0.0002

Βλέπουμε ξεκάθαρα ότι οι τιμές που παίρνει ο δείκτης Ki67 διαφέρει ανάλογα με το σε ποιο γκρουπ κινδύνου ανήκει ο ασθενής και πιο συγκεκριμένα οι ασθενείς που ανήκουν στην κατηγορία χαμηλού ρίσκου έχουν μικρότερη τιμή για τον δείκτη Ki67.

6.3.2 Έλεγχος μέσω λογιστικής παλινδρόμησης

Είδαμε ότι πράγματι η κατηγορία κινδύνου στην οποία ανήκει ο ασθενής επηρεάζεται από τον δείκτη Ki 67 ενδιαφέρον είναι πως όμως επιδρά ο δείκτης την μεταβλητή risk. Αυτό θα γίνει με την χρήση της λογιστικής παλινδρόμησης και το μοντέλο μας θα είναι της μορφής

$$logit[\pi(risk)] = a + \beta * Ki67$$

Πριν περάσουμε όμως στην παρουσίαση των συντελεστών θα ελέγξουμε εάν το μοντέλο μας είναι επαρκές κάτι το οποίο θα γίνει μέσω του ελέγχου λόγου πιθανοφανειών. Κάνοντας το έχουμε:

Πίνακας 6.24 (Έλεγχος σημαντικότητας του μοντέλου)

Log-rank Test	
Null deviance	61.827 για 44 β.ε
Residual deviance	45.156 για 43 β.ε
p-value	4.446274e-05

Βλέπουμε δηλαδή ότι επιβεβαιώνεται ότι ο δείκτης Ki67 είναι πράγματι σημαντικός. Αν θέλουμε να δούμε την προσαρμογή του μοντέλου αυτή ισούται με

$$\frac{deviance}{\beta\alpha\theta\mu o\acute{\upsilon}\varsigma\ \varepsilon\lambda\varepsilon\upsilon\theta\varepsilon\rho\acute{\iota}\alpha\varsigma} = \frac{45.156}{43} = 1.05,$$

δηλαδή η προσαρμογή του μοντέλου είναι αρκετά καλή.

Εάν θέλουμε να επιβεβαιώσουμε ότι η Κυκλίνη Ε δεν είναι στατιστικά σημαντική θα μπορούσαμε να τρέξουμε ένα μοντέλο λογιστικής παλινδρόμησης που θα περιλαμβάνει το όρο αυτό και να το συγκρίνουμε με το αρχικό μας μοντέλο. Κάνοντας το έχουμε.

Πίνακας 6.25 (Έλεγχος σημαντικότητας για το εναλλακτικό μοντέλο)

Log-rank Test	
Model1 deviance	45.156 για 43 β.ε
Model2 deviance	41.763 για 39 β.ε
p-value	0.4942906

Βλέπουμε δηλαδή ότι τα δύο μοντέλα δεν διαφέρουν μεταξύ τους οπότε δεν είναι ανάγκη να συμπεριλάβουμε την Κυκλίνη Ε στο μοντέλο.

Από την στιγμή που καταλήξαμε στο μοντέλο που θα χρησιμοποιήσουμε θα πρέπει να παρουσιάσουμε τους συντελεστές του, οι οποίοι φαίνονται στον παρακάτω πίνακα.

Πίνακας 6.26 (Συντελεστές παραμέτρων)

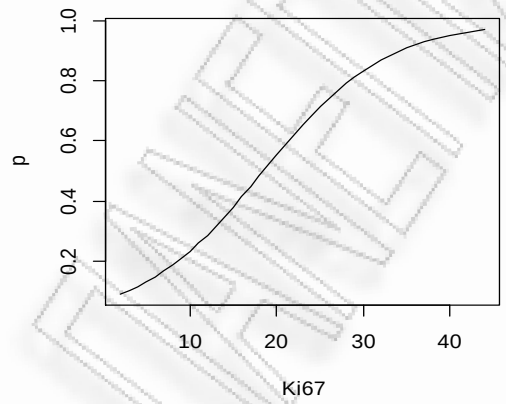
	Estimate	Std. Error	z value	p-value	95% δε
Intercept	-2.60802	0.80432	-3.243	0.00118	-4.40925 -1.19129
Ki67	0.14095	0.04364	3.230	0.00124	0.06582 0.24017

Άρα το μοντέλο μας είναι το παρακάτω

$$logit[\pi(risk)] = -2.60802 + 0.14095 * Ki67$$

Καταρχήν από την στιγμή που β>0 η $\pi(risk)$ είναι γνησίως αύξουσα ως προς τις τιμές του δείκτη. Επιπλέον από την στιγμή που $\beta = 0.14095$ αυτό σημαίνει ότι το odds ratio ισούται με exp(0.14095)=1.15 με 95% δ.ε (1.068 1.271). Δηλαδή αν ο δείκτης Ki67 αυξηθεί κατά μία μονάδα τότε το odds ο ασθενής να ανήκει στην ομάδα υψηλού κινδύνου έναντι του χαμηλού είναι 1.15 φορές μεγαλύτερο από ότι για την αρχική περίπτωση.

Διάγραμμα 6.7 (Διάγραμμα πιθανότητας του risk με Ki67)



Διαγ. 6.7 Μεταβολή της πιθανότητας να ανήκει κάποιος στην ομάδα υψηλού κινδύνου καθώς μεταβάλλεται ο δείκτης Ki67.

Στον παρακάτω πίνακα παρουσιάζεται η πιθανότητα ο ασθενής να ανήκει στην ομάδα υψηλού κινδύνου όταν ο Ki67 παίρνει τιμές ίσες με το 1^{ο}, 2^{ο},3^{ο} τεταρτημόριο μαζί με ένα 95% δ.ε αλλά και η τοπική κλίση μιας ευθείας στα σημεία αυτά.

Πίνακας 6.27 (Μεταβολή των πιθανοτήτων για διάφορες τιμές του Ki67)

Ki67	Probability	95% δ.ε	κλίση
Q1 (8)	0.185	0.0318 , 0.3388	0.021284133
Q2 (15)	0.379	0.2074 , 0.5505	0.033174156
Q3 (23)	0.653	0.4463 , 0.8603	0.031922546

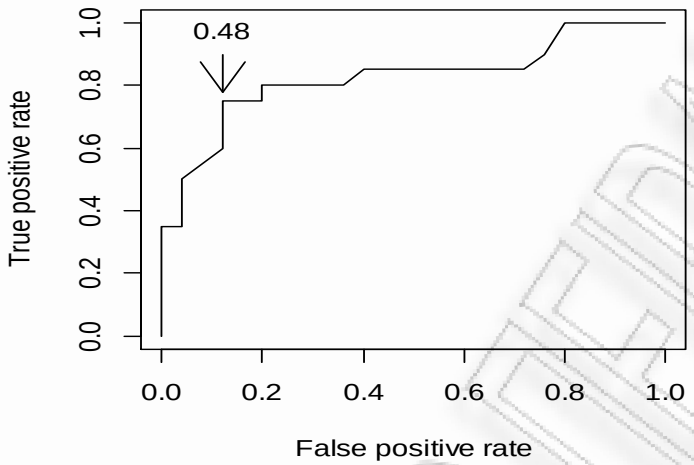
Από τον πίνακα μπορούμε να δούμε ότι όταν η τιμή του δείκτη μεταβάλλεται από το $1^{ο}$ στο $2^{ο}$ τεταρτημόριο τότε η πιθανότητα αυξάνεται κατά 19% ενώ η διαφορά μεταξύ $1^{ου}$ και $3^{ου}$ τεταρτημορίου είναι 46.7%. επίσης κοντά στο $1^{ο}$ τεταρτημόριο η πιθανότητα αυξάνεται με ρυθμό περίπου 2% ανά μονάδα αύξησης στην τιμή του δείκτη Ki67 ενώ για τα άλλα δύο τεταρτημόρια η αύξηση είναι περίπου 3% ανά μονάδα. Τέλος το διάμεσο επίπεδο αποτελεσματικότητας (median effective level) ισούται με

$$EL_{50} = 18.5,$$

το οποίο μας λέει ότι η «τοπική» ευθεία για Ki67 ίσο με 18.5 παρουσιάζει την μεγαλύτερη κλίση που ισούται με 0.035, δηλαδή ο ρυθμός αύξησης στο σημείο αυτό είναι 3.5% ανά μονάδα του Ki67.

Τέλος πολλές φορές είναι χρήσιμη η κατασκευή ενός πίνακα ταξινόμησης καθώς μας βοηθά στο να αντιληφθούμε οπτικά την προβλεπτική ικανότητα του μοντέλου λογιστικής παλινδρόμησης. Για την κατασκευή του πίνακα θα χρειαστούμε μια τιμή κατώφλι (cut-off point) με βάση την οποία θα επιλέγεται σε πιο από τα δύο γκρουπ θα ανήκει μια παρατήρηση. Την απόφαση για το πιο θα είναι το σημείο αυτό θα την πάρουμε μέσα από μια ROC καμπύλη η οποία είναι μια καμπύλη που ορίζεται από τα σημεία (1-ειδικότητα, ευαισθησία) και το εμβαδό που δημιουργείται κάτω από αυτήν καλείται δείκτης συμφωνίας (concordance index) και εκτιμά την πιθανότητα να είναι ίσες οι προβλέψεις με τις πραγματικές παρατηρήσεις. Όσο μεγαλύτερο είναι το νούμερο αυτό τόσο μεγαλύτερη θεωρείται η προβλεπτική ικανότητα του μοντέλου

Διάγραμμα 6.8 (Καμπύλη ROC)



Το σημείο 0.48 είναι αυτό που μεγιστοποιεί το άθροισμα των ποσοτήτων ευαισθησία (true positive rate) και ειδικότητας (1- false positive rate) και είναι η τιμή που θα χρησιμοποιήσουμε ως τιμή κατώφλι. Επίσης έχουμε ότι AUC= 0.824 που είναι αρκετά υψηλό.

Αφού καταλήξαμε στην τιμή κατώφλι μπορούμε να δημιουργήσουμε τον πίνακα μας ο οποίος είναι ο παρακάτω.

Πίνακας 6.28 (Πίνακας Ταξινόμησης)

	Πρόβλεψη για risk		
Risk	Low	High	Sum
Low	22	3	25
High	5	15	20
Sum	27	18	45

Σύμφωνα με τον παραπάνω πίνακα η πιθανότητα σωστής πρόβλεψης ισούται με:

$$P = P(\widehat{risk} = 1 \; και \; risk = 1) + P(\widehat{risk} = 0 \; και \; risk = 0) = \frac{22 + 15}{45} = \frac{37}{45}$$

$$= 0.82$$

Τέλος, σχετικά με τον πίνακα ταξινόμησης θα μπορούσαμε να κατασκευάσουμε πολλούς διαφορετικούς ανάλογα με την τιμή του κατωφλιού αλλά ο συγκεκριμένος με τον τρόπο που κατασκευάστηκε επιτυγχάνει τον καλύτερο διαχωρισμό από όλους τους άλλους.

6.4 Έλεγχος σχετικά με τους παράγοντες E2F1 και E2F4

Σε αυτήν την παράγραφο θα δούμε αν οι παράγοντες E2F1 και E2F4 επηρεάζουν τον βαθμό κακοήθειας ή το στάδιο του καρκίνου και κατά συνέπεια τον κίνδυνο που διατρέχει ο ασθενής.

6.4.1 Επίδραση στον βαθμό κακοήθειας του καρκίνου

Πρώτη κίνηση μας είναι να πάρουμε μια εικόνα μέσω των περιγραφικών μέτρων αρχικά για τον παράγοντα E2F1 και έπειτα για τον E2F4.

Πίνακας 6.29 (Περιγραφικά μέτρα σχετικά με τον παράγοντα E2F1)

Grade	Μέση τιμή	Τυπική απόκλιση	Αριθμός ασθενών
G1	35.5	21.79	10
G2	33	24.11	14
G3	28.81	18.82	16

Από τον πρώτο πίνακα βλέπουμε ότι οι τιμές που παίρνει ο παράγοντα και για τους τρεις βαθμούς κακοήθειας του καρκίνου είναι παραπλήσιες οπότε λογικά δεν θα υπάρχει στατιστικά σημαντική διαφορά μεταξύ των γκρουπ. Μέσω του ελέγχου Kruskal-Wallis έχουμε ότι δεν υπάρχει διαφορά μεταξύ των γκρουπ (p-value>0.05).

Πίνακας 6.30 (Έλεγχος Kruskal-Wallis)

Kruskal-Wallis chi-squared	0.5082
p-value	0.7756

. Εν συνεχεία, θα εφαρμόσουμε την ίδια μεθοδολογία και για τον παράγοντα E2F4. Κάνοντας το έχουμε τα εξής.

Πίνακας 6.31 (Περιγραφικά μέτρα σχετικά με τον παράγοντα E2F4)

Grade	Μέση τιμή	Τυπική απόκλιση	Αριθμός ασθενών
G1	50.45	23.71	11
G2	46.85	24.36	13
G3	53.06	25.32	18

Πίνακας 6.32 (Έλεγχος Kruskal-Wallis)

Kruskal-Wallis chi-squared	0.5195
p-value	0.7713

Από τον πρώτο πίνακα βλέπουμε ότι οι τιμές που παίρνει ο δείκτης και για τους τρεις βαθμούς κακοήθειας του καρκίνου είναι παραπλήσιες. Από τον έλεγχο Kruskal-Wallis έχουμε πάλι ότι δεν υπάρχει διαφορά μεταξύ των γκρουπ (p-value>0.05).

6.4.2 Επίδραση στο στάδιο του καρκίνου

Για να ελέγξουμε το αν οι δύο παράγοντες επιδρούν στο στάδιο του καρκίνου θα ακολουθήσουμε τα ίδια βήματα όπως στην προηγούμενη παράγραφο, οπότε για τον παράγοντα E2F1 έχουμε.

Πίνακας 6.33 (Περιγραφικά μέτρα σχετικά με τον παράγοντα E2F1)

Stage	Μέση τιμή	Τυπική απόκλιση	Αριθμός ασθενών
pTa	31.83	24.81	18
pT1	26	19.49	5
pT2	32.14	13.49	7

Από τον πρώτο πίνακα βλέπουμε ότι οι τιμές είναι σχετικά κοντά οπότε λογικά δεν αναμένουμε να υπάρχει στατιστικά σημαντική διαφορά. Ο έλεγχος Kruskal-Wallis επιβεβαιώνει την παραπάνω διαίσθηση (p-value>0.05).

Πίνακας 6.34 (Έλεγχος Kruskal-Wallis)

Kruskal-Wallis chi-squared	0.5195
p-value	0.7713

Κάνοντας τα ίδια και για τον παράγοντα E2F4 παίρνουμε τα εξής αποτελέσματα.

Πίνακας 6.35 (Περιγραφικά μέτρα σχετικά με τον δείκτη E2F4)

Stage	Μέση τιμή	Τυπική απόκλιση	Αριθμός ασθενών
pTa	51.84	24.33	19
pT1	44.17	28.35	6
pT2	57.50	23.45	8

Πίνακας 6.36 (Έλεγχος Kruskal-Wallis)

Kruskal-Wallis chi-squared	0.9808
p-value	0.6124

Από τον πρώτο πίνακα βλέπουμε ότι με εξαίρεση το δεύτερο στάδιο του καρκίνου όπου φαίνεται ότι οι τιμές του δείκτη E2F4 είναι μικρότερες δεν υπάρχει κάποια διαφορά. Σύμφωνα με τον έλεγχο Kruskal-Wallis βλέπουμε ότι δεν υπάρχει διαφορά μεταξύ των γκρουπ.

Κεφάλαιο 7

Ελλιπείς Τιμές

7.1 Είδη ελλιπών τιμών (missing data mechanisms)

Για να αποφασίσουμε το πως θα χειριστούμε τις ελλιπείς τιμές είναι χρήσιμο να ξέρουμε τον λόγο για τον οποίο λείπουν. Υπάρχουν γενικά τρία είδη ελλιπών τιμών:

1. Ελλείπουσες τιμές εντελώς στην τύχη (missingness completely at random, MCAR). Μια μεταβλητή είναι MCAR εάν η πιθανότητα να είναι ελλιπής είναι η ίδια για όλες τις μονάδες. Με απλά λόγια όταν οι τιμές που λείπουν αποτελούν τυχαίο δείγμα από το σύνολο των τιμών.
2. Ελλιπείς τιμές στην τύχη (missingness at random, MAR). Μια μεταβλητή είναι MAR εάν η πιθανότητα να είναι ελλιπής εξαρτάται μόνο από την διαθέσιμη πληροφόρηση που έχουμε. Για παράδειγμα έστω ότι έχουμε μια ερώτηση σχετικά με το εισόδημα όπου οι γυναίκες γενικά αποφεύγουν να την απαντούν σε σχέση με τους άντρες. Σε αυτή την περίπτωση δεν έχουμε MCAR καθώς οι ελλείπουσες τιμές στην ερώτηση δεν είναι στην τύχη αλλά από την στιγμή που έχουμε πλήρη καταγραφή για το φύλο των ατόμων τότε για δεδομένο το φύλο και του γεγονότος ότι οι ελλείπουσες οφείλονται μόνο στο φύλο η μεταβλητή μας είναι MAR.
3. Ελλιπείς τιμές όχι στην τύχη (missingness not at random, MCAR). Εάν μια μεταβλητή δεν ανήκει σε κάποια από τις δύο παραπάνω κατηγορίες τότε αναγκαστικά είναι MNAR, δηλαδή οι ελλείπουσες τιμές δεν είναι στην τύχη και οφείλεται είτε σε παράγοντες που δεν έχουμε λάβει υπόψη και οι οποίοι επηρεάζουν την μεταβλητή μας είτε στην μεταβλητή την ίδια. Για παράδειγμα έστω ότι το επίπεδο μόρφωσης επηρεάζει το αν κάποιος δηλώσει το εισόδημα του αλλά δεν το έχουμε καταγράψει ή γενικότερα όσοι έχουνε υψηλά εισοδήματα αποφεύγουν να τα δηλώσουν.

Δυστυχώς γενικά δεν μπορούμε να είμαστε σίγουροι εάν πράγματι τα δεδομένα μας είναι MAR ή αν αυτό εξαρτάται και από παράγοντες που δεν είναι καταγεγραμμένοι ή από τις ίδιες τις ελλείπουσες τιμές. Η ουσιαστική δυσκολία έγκειται στο ότι αυτοί οι παράγοντες μπορεί να επηρεάζουν δεν έχουν καταγραφεί

οπότε δεν μπορούμε να τους αποκλείσουμε. Γενικά πρέπει να κάνουμε υποθέσεις ή να ελέγξουμε για αναφορές σε άλλες μελέτες. Πρακτικά προσπαθούμε να εισάγουμε όσους περισσότερους παράγοντες σε ένα μοντέλο ώστε η MAR υπόθεση να είναι όσο το δυνατόν πιο πιθανή. Έτσι στο παράδειγμα μας η υπόθεση ότι η μη απάντηση στο ερώτημα για το εισόδημα εξαρτάται από το φύλο και την εκπαίδευση μπορεί να είναι αρκετά ισχυρή αλλά είναι πιο πιθανή από το να υποθέσουμε ότι είναι σταθερή ή ότι εξαρτάται μόνο από έναν από τους δύο παράγοντες.

7.2 Τρόποι αντιμετώπισης ελλιπών τιμών

Υπάρχουν διάφοροι τρόποι για την αντιμετώπιση του προβλήματος που παρουσιάζεται από την έλλειψη πλήρων δεδομένων. Εδώ θα παρουσιάσουμε συνοπτικά τους κυριότερους από αυτούς ξεκινώντας από τους πιο απλούς.

1. Ανάλυση με βάση τις πλήρεις παρατηρήσεις (complete cases analysis): Μια άμεση αντιμετώπιση του προβλήματος των ελλιπών δεδομένων είναι η διαγραφή των μονάδων για τις οποίες δεν υπάρχουν τα πλήρη δεδομένα. Αυτή η μέθοδος παρουσιάζει δύο βασικά μειονεκτήματα. Πρώτον γενικά οι εκτιμήσεις που προκύπτουν δεν είναι αμερόληπτες και μπορεί να είναι ανεπαρκείς λόγω του μικρότερου δείγματος. Εξαίρεση αποτελεί η περίπτωση MCAR όπου τότε οι εκτιμήσεις είναι αμερόληπτες. Δεύτερον σε περίπτωση που υπάρχουν αρκετές μεταβλητές με ελλιπή δεδομένα τότε θα χρειαστεί να διώξουμε ένα πολύ μεγάλο μέρος του δείγματος μας.
2. Ανάλυση με βάση τα διαθέσιμα δεδομένα (Available-case analysis): Στην περίπτωση αυτή διαφορετικές πτυχές ενός προβλήματος μελετούνται με διαφορετικά υποσύνολα δεδομένων. Το πρόβλημα που παρουσιάζει η συγκεκριμένη μέθοδος είναι ότι οι διάφορες συγκρίσεις που γίνονται δεν έχουν κάποια συνέπεια μεταξύ τους καθώς βασίζονται σε διαφορετικά μεταξύ τους υποσύνολα. Επίσης σε περίπτωση που οι ελλιπείς τιμές διαφέρουν συστηματικά από τις παρατηρηθείσες τότε τα αποτελέσματα που θα πάρουμε δεν θα είναι αμερόληπτα.
3. Αντικατάσταση με την μέση τιμή (mean imputation): Εδώ κάθε ελλιπής τιμή αντικαθίσταται από την μέση τιμή που προκύπτει από τις παρατηρηθείσες για την μεταβλητή αυτή. Τα μειονεκτήματα της μεθόδου αυτής είναι ότι αρχικά δεν γίνεται να χρησιμοποιηθεί για κατηγορικές μεταβλητές. Επιπλέον μπορεί παραποιήσει την κατανομή της μεταβλητής

οδηγώντας σε επιπλοκές κυρίως σε ότι αφορά την τυπική απόκλιση της μεταβλητής την οποία υποεκτιμά. Επίσης παραποιεί την σχέση μεταξύ μεταβλητών καθώς «οδηγεί» τις συσχετίσεις στο μηδέν.

4. Αντικατάσταση μέσω παλινδρόμησης (regression mean imputation): Εδώ για να αντικαταστήσουμε την ελλιπή τιμή μιας μεταβλητής δουλεύουμε ως εξής. Τρέχουμε μια παλινδρόμηση μεταξύ δύο μεταβλητών εκ των οποίων η μια έχει πλήρη δεδομένα και η άλλη όχι και με βάση την συνάρτηση που προκύπτει από τις πλήρεις περιπτώσεις υπολογίζουμε και παίρνουμε αμερόληπτες εκτιμήσεις για τις περιπτώσεις που έχουμε τα κενά. Το πρόβλημα πάλι είναι ότι υποεκτιμάται η διακύμανση.

Οι παραπάνω μεθοδολογίες είναι οι πιο απλές στην εφαρμογή τους αλλά όπως είδαμε αυτό δεν σημαίνει ότι είναι και ικανοποιητικές ως προς τα αποτελέσματα τους. (Gelman and Hill, 2007; Carpenter, 2009).

7.3 Πολλαπλή εισαγωγή (multiple imputation)

Όπως είδαμε οι προηγούμενες μέθοδοι δεν είναι και οι πιο κατάλληλοι για την αντιμετώπιση του προβλήματος των ελλιπών τιμών. Στην παράγραφο αυτή θα κάνουμε μια συνοπτική παρουσίαση της multiple imputation (MI) που χρησιμοποιείται αρκετά για την αντιμετώπιση αυτού του προβλήματος. Για μια πιο ενδελεχή ανάλυση μπορεί κάποιος να ανατρέξει στις εργασίες των Rubin (1987,1996) και Schafer (1997).

7.3.1 Περιγραφή της μεθόδου

Τα τελευταία χρόνια και με την ανάπτυξη των στατιστικών πακέτων έχει γίνει πιο ευρεία η χρήση της MI μεθόδου για την αντιμετώπιση των ελλιπών τιμών. Η κεντρική ιδέα της μεθόδου αυτής είναι ότι αντί να συμπληρώσουμε ένα κενό στα δεδομένα μας με μία μόνο τιμή, θα ήταν προτιμότερο να την αντικαταστήσουμε με περισσότερες έτσι ώστε να αναδειχτεί η αβεβαιότητα μας σχετικά με το μοντέλο που ακολουθήθηκε ώστε να καλυφθούν οι τιμές. Η μέθοδος MI το επιτυγχάνει αυτό δημιουργώντας περισσότερες της μίας τιμές (συνήθως 5) για κάθε κενό στα δεδομένα μας, καθεμία εκ των οποίων έχει προκύψει από ένα λίγο διαφορετικό μοντέλο και οι οποίες αντικατοπτρίζουν την μεταβλητότητα μεταξύ των δειγμάτων (sampling variability). Με βάση αυτές δημιουργεί πολλαπλά πλήρη σετ δεδομένων στα οποία πλέον μπορούμε να κάνουμε την ανάλυση που μας ενδιαφέρει ενώ και τα αποτελέσματα

μπορούν να συνδυαστούν. Για παράδειγμα στην περίπτωση που εφαρμόζουμε γραμμική παλινδρόμηση και ενδιαφερόμαστε για τις τιμή του συντελεστή β εκτιμούμε για κάθε ένα από τα M σετ τους συντελεστές $\hat{\beta}_i$ όπως και τα αντίστοιχα τυπικά σφάλματα s_i. Μια συνολική εκτίμηση του συντελεστή β είναι τότε ο $\hat{\beta} = \frac{1}{M}\sum_{i=1}^{M}\hat{\beta}_i$ ενώ μια εκτίμηση της διασποράς που όμως αντικατοπτρίζει τόσο την διακύμανση μέσα σε κάθε σετ όσο και ανάμεσα στα σετ είναι η $V_\beta = W + \left(1+\frac{1}{M}\right) * B$

Όπου

$$W = \frac{1}{M}\sum_{i=1}^{M}\hat{s}_i^{\,2}$$

και

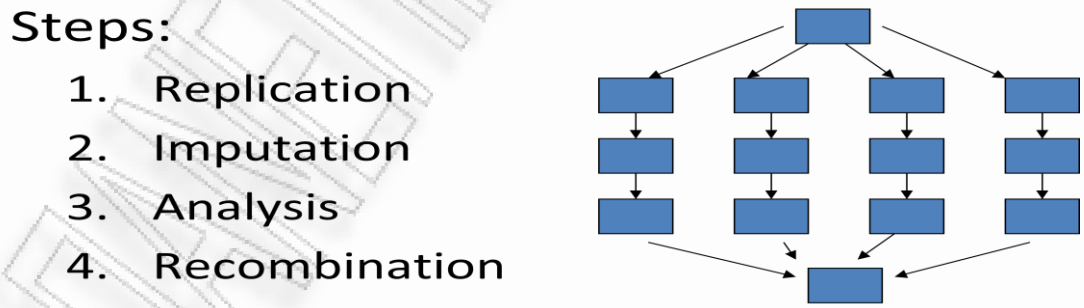
$$B = \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\beta}_i - \hat{\beta}\right)^2$$

Εάν θέλουμε να ελέγξουμε την υπόθεση β=β_0 αυτό θα γίνει μέσω της στατιστικής συνάρτησης

$$T = \frac{\hat{\beta} - \beta_0}{\sqrt{V_\beta}} \sim t_\nu ,$$

όπου $\nu = (M-1) * \left[1 + \frac{W}{\left(1+\frac{1}{M}\right)*B}\right]^2$.

Η όλη διαδικασία περιγράφεται σχηματικά από το παρακάτω σχεδιάγραμμα



Διάγραμμα 7.3.1 απεικόνιση της διαδικασίας MI

(πηγή: Paul T. von Hippel 2010)

Μια άλλη ποσότητα που υπάρχει είναι το ποσοστό χαμένης πληροφορίας για μια μεταβλητή (rate of missing information). Όταν ενεργούμε ανάλυση που είναι βασισμένη στην MI η διακύμανση μεταξύ των σετ δεδομένων που έχουν δημιουργηθεί δείχνει την στατιστική αβεβαιότητα που υπάρχει λόγω των ελλιπών δεδομένων. Ο Rubin (1987) παρέχει διαγνωστικά μέτρα που δείχνουν πόσο ισχυρά

επηρεάζονται οι ποσότητες που έχουν εκτιμηθεί από τα ελλιπή δεδομένα. Το εκτιμώμενο ποσοστό χαμένης πληροφορίας για μια μεταβλητή δίνεται από τον τύπο

$$\gamma = \frac{r + \frac{2}{\nu + 3}}{r + 1},$$

όπου $r = \frac{\left(1+\frac{1}{M}\right)*B}{W}$,δηλαδή η σχετική αύξηση στην διακύμανση λόγω των ελλειπουσών τιμών.

Ο δείκτης αυτός καθώς και το πλήθος Μ των εκτιμήσεων προσδιορίζουν την αποτελεσματικότητα της MI. Σύμφωνα με τον Rubin (1987) 3 έως 10 επαναλήψεις είναι αρκετές για να έχουμε ικανοποιητικά αποτελέσματα. Εξαίρεση αποτελεί η περίπτωση που ο δείκτης γ είναι αρκετά μεγάλος για κάποια μεταβλητή οπότε ίσως χρειαστούν περισσότερες επαναλήψεις. Αυτό όμως δεν είναι πάντα εφικτό και ειδικότερα σε περιπτώσεις που υπάρχουν πολλές μεταβλητές στα δεδομένα μας. Παρακάτω παρουσιάζεται ένας πίνακας σχετικά με την αποτελεσματικότητα της μεθόδου για διάφορες τιμές των γ, Μ.

	γ				
m	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Πίνακας 7.3.1 (πηγή: http://sites.stat.psu.edu/~jls/mifaq.html#few)

Τα αποτελέσματα που προκύπτουν από την MI είναι ισχυρά υπό την υπόθεση ότι οι ελλιπείς τιμές είναι MAR (η περίπτωση MCAR είναι υποπερίπτωση της MAR) και για αυτό τον λόγο προσπαθούμε στα μοντέλα που θα χρησιμοποιήσουμε ώστε να καλύψουμε τις ελλείπουσες τιμές να βάλουμε τις περισσότερες αν όχι όλες τις μεταβλητές που έχουμε στην διάθεση μας έτσι ώστε η υπόθεση μας να είναι όσο το δυνατόν πιο ισχυρή (Carpenter 2009).

7.3.2 Μοντέλα εισαγωγής (imputation models)

Ο τελικός σκοπός της μεθόδου αυτής είναι να παράγει έγκυρα συμπεράσματα για τους εκτιμητές που μας ενδιαφέρουν και προκύπτουν από τα ολοκληρωμένα δεδομένα. Για να επιτευχθεί αυτό οι τιμές που προκύπτουν θα πρέπει να διατηρούν τη δομή των δεδομένων όπως και την αβεβαιότητα για αυτή τη δομή και να λαμβάνουν υπόψη τυχόν γνώση για την διαδικασία που δημιούργησε τα ελλιπή δεδομένα. Δύο βασικές προσεγγίσεις έχουν προταθεί η joint modeling (jm) και η Sequential regression multiple imputation (SRMI). Η (jm) μένει πιο κοντά στην θεωρία ενώ η (SRMI) δίνει περισσότερη έμφαση στα δεδομένα (Stef van Buuren, 2007).

1. Joint modeling: Η προσέγγιση αυτή χωρίζει τις παρατηρήσεις σε γκρουπ με παρόμοια μοτίβα σχετικά με τις ελλείπουσες τιμές και γεμίζει τις κενές τιμές για κάθε μοτίβο σύμφωνα με ένα μοντέλο για όλες τις μεταβλητές που είναι κοινό για όλες τις παρατηρήσεις. Κάποια κλασικά μοντέλα είναι αυτά της πολυμεταβλητής κανονικής κατανομής για συνεχείς μεταβλητές και τα log-linear για κατηγορικές μεταβλητές. Θεωρητικά η συγκεκριμένη μέθοδος είναι ισχυρή αλλά στην πράξη δεν έχει την ελαστικότητα που απαιτείται ώστε να αντιπροσωπεύσει την πολυπλοκότητα των δομών που έχουν τα δεδομένα σε πολλές μελέτες. Σε μια τέτοια περίπτωση η συγκεκριμένη μέθοδος είναι δύσκολο να εφαρμοστεί γιατί οι τυπικές προδιαγραφές των πολυμεταβλητών κατανομών δεν έχουν την ελαστικότητα ώστε να συμπεριλάβουν αυτά τα χαρακτηριστικά.
2. Sequential regression multiple imputation (SRMI): Σε αυτήν την περίπτωση το μοντέλο για την εκτίμηση των ελλειπουσών τιμών ορίζεται ξεχωριστά για κάθε μεταβλητή, με τις υπόλοιπες να χρησιμοποιούνται ως προβλεπτικοί παράγοντες αυτής. Σε κάθε βήμα του αλγορίθμου δημιουργούνται εκτιμήσεις για τις ελλείπουσες τιμές μιας μεταβλητής, οι οποίες στην συνέχεια χρησιμοποιούνται για την εκτίμηση των τιμών μιας άλλης μεταβλητής. Η συγκεκριμένη διαδικασία συνεχίζεται έως ότου να επιτευχθεί η σύγκλιση των τιμών. Συγκρινόμενη με την προηγούμενη μέθοδο η συγκεκριμένη έχει το πλεονέκτημα ότι είναι σχετικά εύκολο να συμπεριλάβει χαρακτηριστικά πολύπλοκων δεδομένων σε μοντέλα μονοπαραγοντικής παλινδρόμησης. Για παράδειγμα στην περίπτωση μιας συνεχούς μεταβλητής μπορεί να χρησιμοποιηθεί η γραμμική παλινδρόμηση και σε μια δίτιμη κατηγορική η λογιστική παλινδρόμηση (Yulei He, 2010).

Πλέον πολλά στατιστικά πακέτα (STATA,SPSS κτλ.) διαθέτουν τις επιλογές ώστε να πραγματοποιήσουν την MI. Στην R κάποια πακέτα που υπάρχουν για αυτό τον σκοπό είναι τα mi, MICE και AMELIA.

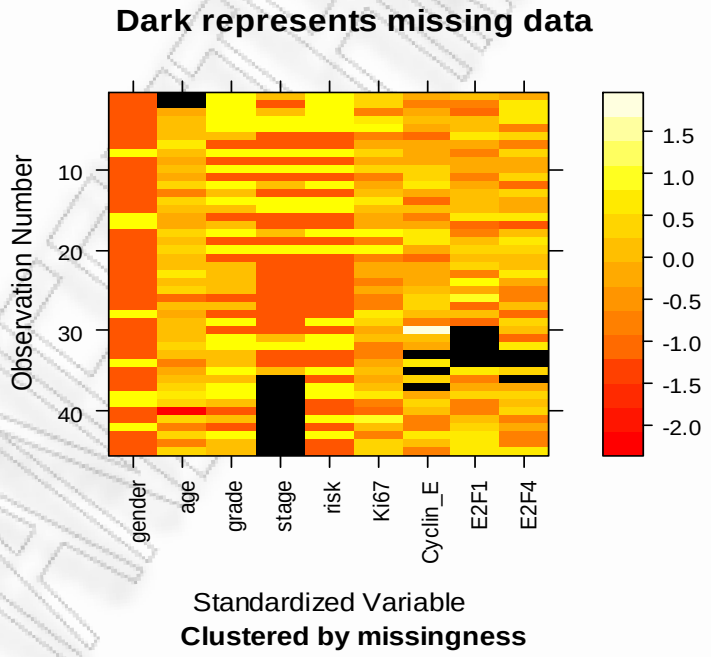
7.4 Εφαρμογή της μεθόδου στα δεδομένα μας

Όπως είδαμε τα αρχικά μας δεδομένα δεν είναι ολοκληρωμένα αλλά ελλειπή και αυτό μπορεί να επηρεάσει τα αποτελέσματα της ανάλυσης μας. Για αυτό το λόγο μέσω των κατάλληλων εργαλείων θα καλύψουμε τα κενά στα δεδομένα μας και θα δούμε αν υπάρχουν τυχών διαφοροποιήσεις στα δεδομένα μας.

7.4.1 Εκτίμηση των ελλιπών τιμών και έλεγχος

Αρχικά μέσω του παρακάτω διαγράμματος μπορούμε να δούμε τι κατάσταση επικρατεί σχετικά με το ποσοστό των ελλειπουσών τιμών. Παρατηρούμε ότι τα περισσότερα κενά υπάρχουν στην μεταβλητή stage και τα λιγότερα στην age όπως φανερώνουν τα μαύρα κομμάτια στο παρακάτω σχήμα.

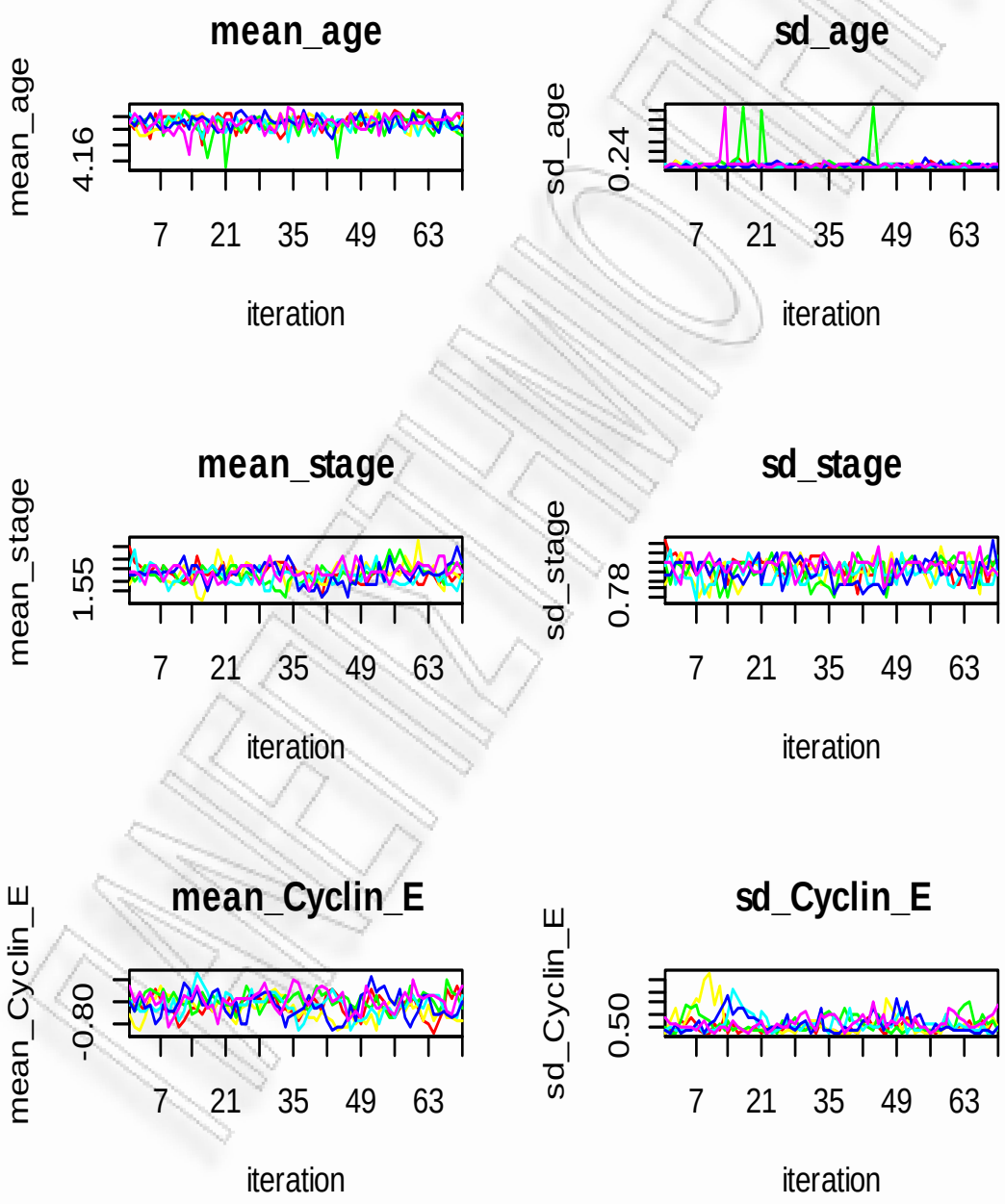
Διάγραμμα 7. 1 (Πλήθος ελλιπών τιμών ανά μεταβλητή)

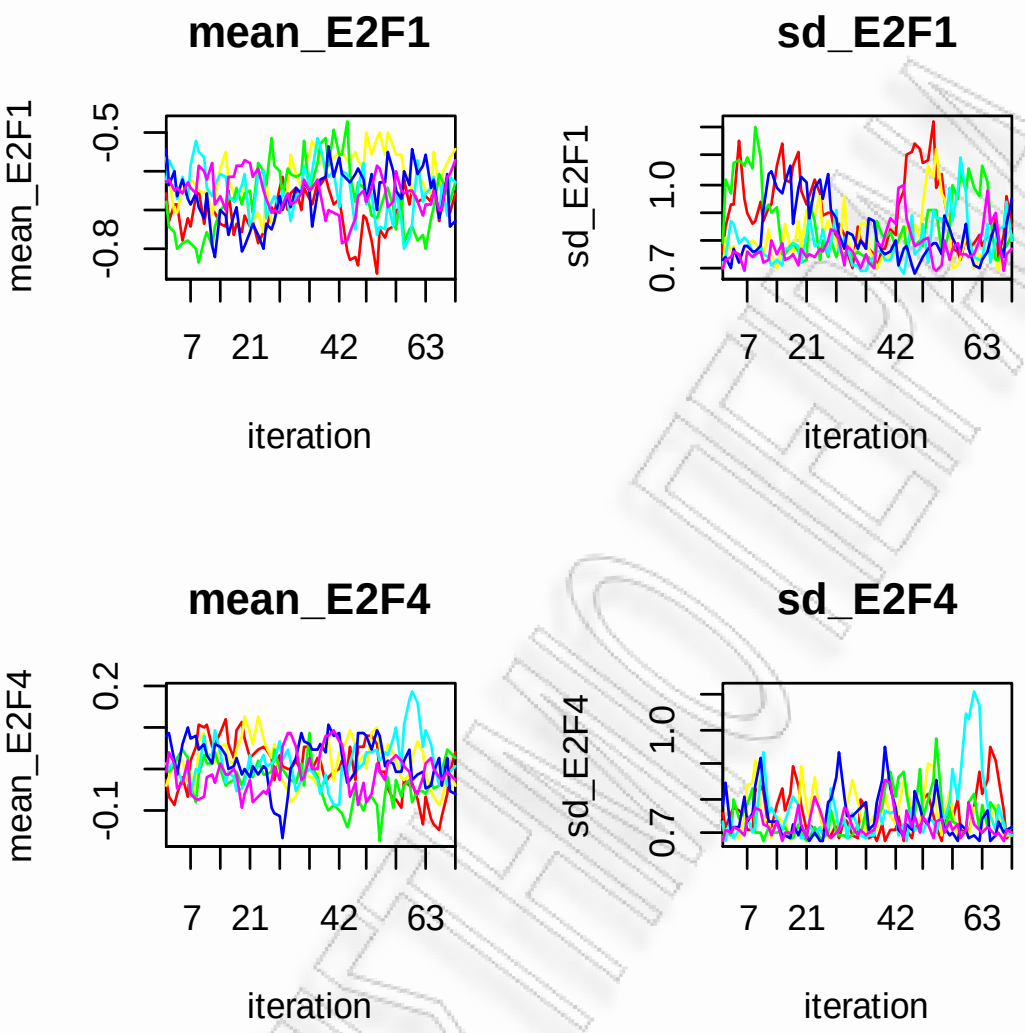


Πριν παρουσιάσουμε τις εκτιμηθείσες τιμές θα πρέπει να δούμε αν ο αλγόριθμος που χρησιμοποιήσαμε έχει συγκλίνει και επίσης αν οι τιμές που παίρνουμε είναι πιθανόν να προέρχονται από το συγκεκριμένο δείγμα.

Α) Σύγκλιση: Δεν υπάρχει κάποιος ξεκάθαρος τρόπος για να ελέγξουμε την σύγκλιση του αλγόριθμου . Αυτό που γίνεται συχνά είναι να δημιουργήσουμε διαγράμματα των παραμέτρων έναντι του αριθμού των επαναλήψεων. Αν το κάνουμε αυτό έχουμε τα παρακάτω διαγράμματα όπου αυτό που θέλουμε είναι οι γραμμές που δημιουργούνται να «μπλέκονται» μεταξύ τους και να μην υπάρχει κάποια που να διαφέρει από τις υπόλοιπες.

Διάγραμμα 7. 2 (Μέση τιμή και διακύμανση για κάθε πλήρη μεταβλητή)

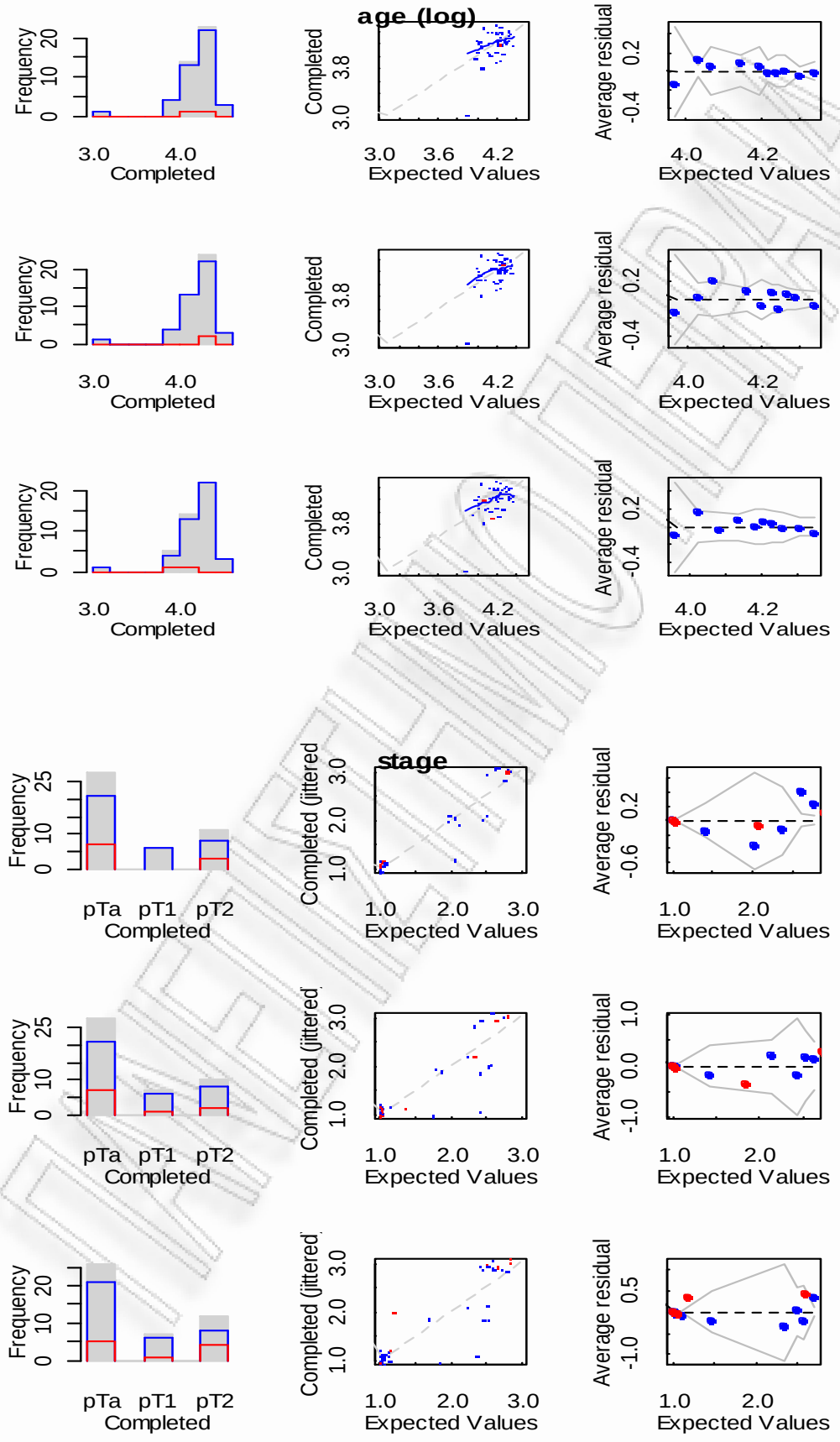


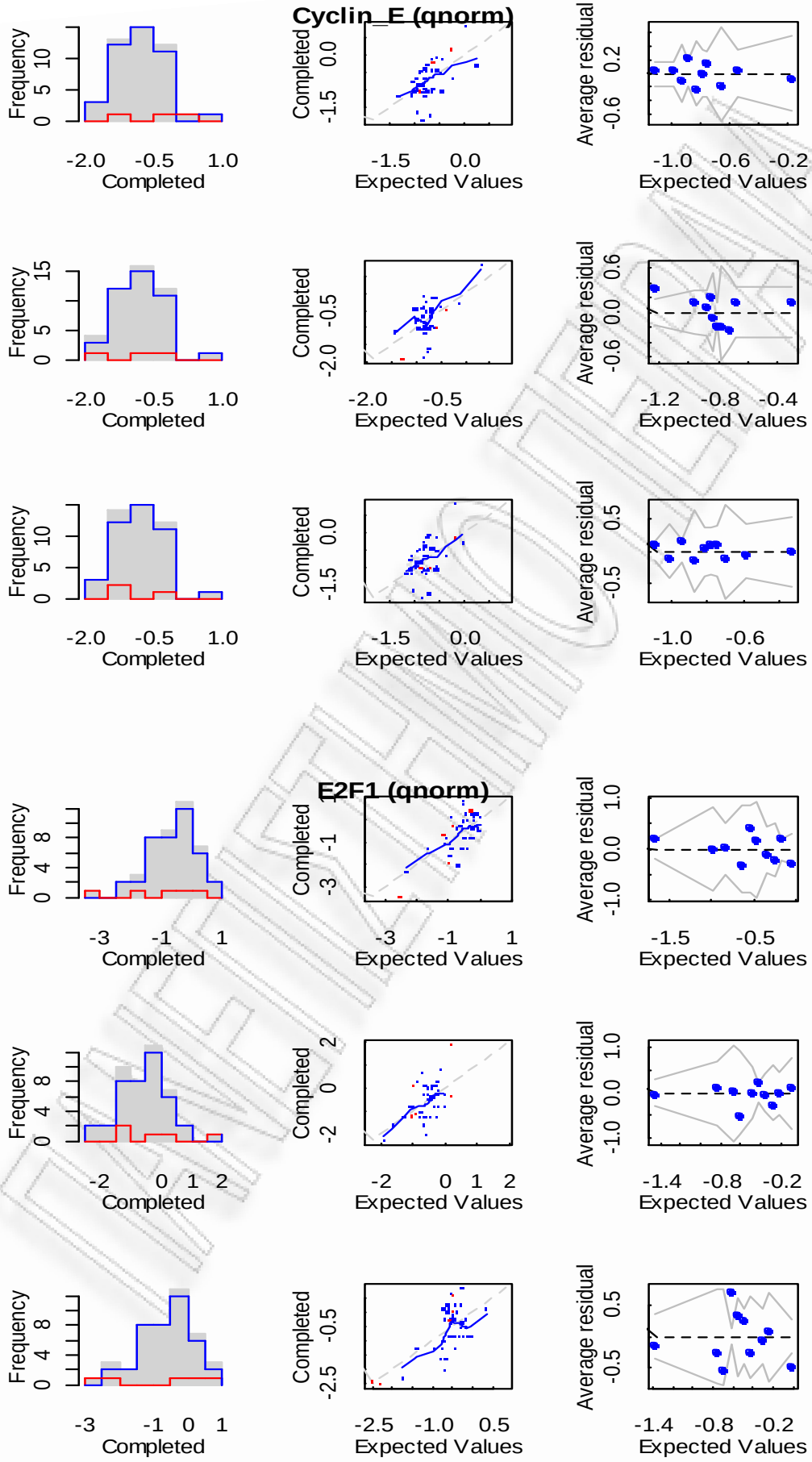


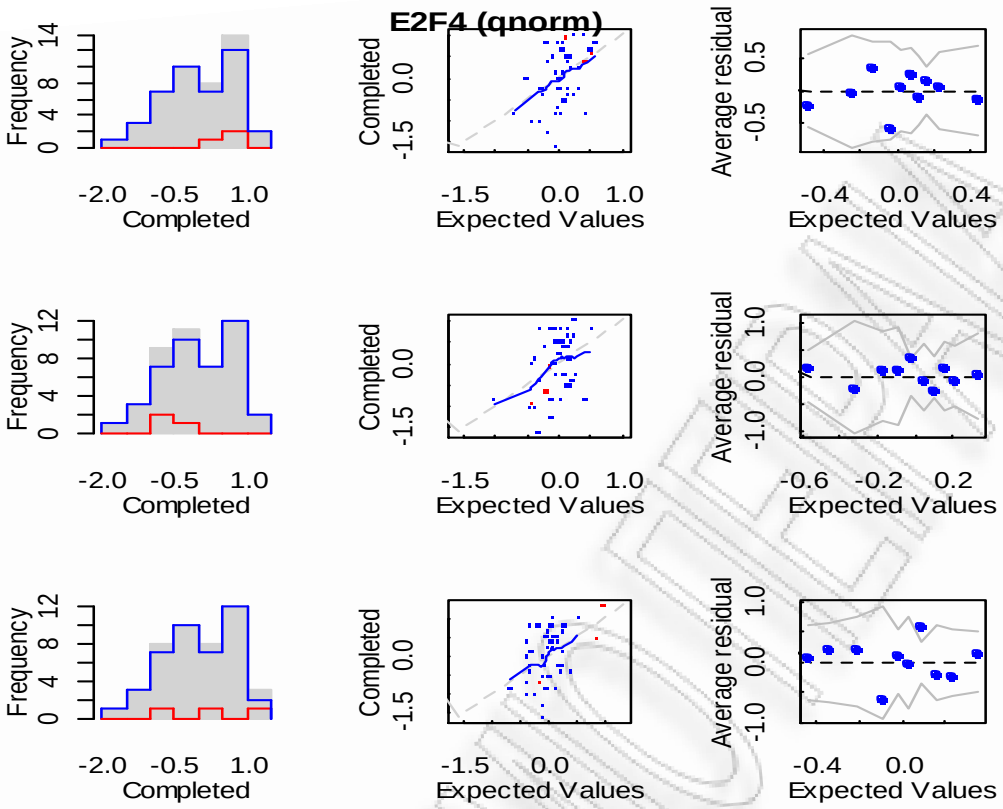
Από τα παραπάνω βλέπουμε ότι ισχύει αυτό που θέλαμε, δηλαδή όλες οι γραμμές είναι μαζί και δεν υπάρχει κάποια που να ακολουθά διαφορετική πορεία αν και σε μερικές περιπτώσεις υπάρχουν σύντομες αποκλίσεις. Παρόλα αυτά βλέπουμε ότι όλες οι γραμμές συγκλίνουν στο ίδιο και άρα μπορούμε να πούμε ότι ο αλγόριθμος έχει συγκλίνει.

Β) από την στιγμή που ο αλγόριθμος έχει συγκλίνει είναι σημαντικό να δούμε τι ισχύει με τις τιμές που παίρνουμε. Αυτό θα το κάνουμε μέσω διαγραμμάτων, δυστυχώς όμως για τα τρία πρώτα από τα έξι πλήρη σετ δεδομένων, για κάθε μεταβλητή.

Διάγραμμα 7.3 (διάφορα χρήσιμα διαγράμματα για κάθε μεταβλητή)




Cyclin_E (qnorm)
Frequency
Completed
Expected Values
Average residual
E2F1 (qnorm)



Αυτά που παίρνουμε από το παραπάνω διάγραμμα είναι αρχικά ένα ιστόγραμμα συχνοτήτων για κάθε μεταβλητή όπου το μπλε χρώμα αντιστοιχεί στις διαθέσιμες τιμές που είχαμε και το κόκκινο σε αυτές που υπολογίσαμε. Από αυτά βλέπουμε ότι οι τιμές που έχουν υπολογιστεί βρίσκονται μέσα σε λογικά πλαίσια και δεν διαφέρουν πολύ από τις παρατηρηθείσες τιμές. Στην συνέχεια, στο διάγραμμα διασποράς σχεδιάζουμε την παρατηρηθείσα ή υπολογισμένη τιμή κάθε παρατήρησης έναντι της αντίστοιχης εκτιμηθείσας τιμής. Εδώ παρατηρούμε ότι για κάθε μεταβλητή δεν υπάρχει στατιστική διαφορά μεταξύ των παρατηρηθεισών και εκτιμώμενων τιμών. Τέλος το τρίτο σχήμα είναι ένα διάγραμμα καταλοίπων όπου μας ενδιαφέρει τα κατάλοιπα αυτά να βρίσκονται κατά κανόνα μέσα στα όρια που ορίζονται από τις δύο γραμμές καθώς αυτό δείχνει ότι τα μοντέλα που χρησιμοποιήθηκαν για τον υπολογισμό των τιμών είναι ικανοποιητικά, κάτι το οποίο συμβαίνει εδώ.

7.5 Ανάλυση με βάση τα νέα πλήρη δεδομένα

Επόμενο στάδιο της ανάλυσης μας είναι να επαναλάβουμε τις αναλύσεις που πραγματοποιήσαμε στα κεφάλαια 5 και 6 χρησιμοποιώντας τα πλήρη δεδομένα που έχουμε στην διάθεση μας και παίρνουμε τις αποφάσεις μας ελέγχοντας τους συγκεντρωτικούς συντελεστές.

7.5.1 Σχέση Κυκλίνης Ε με τους παράγοντες E2F1 και E2F4.

Το πρώτο πράγμα που ελέγξαμε στο 5ο κεφάλαιο είναι η σχέση μεταξύ της Κυκλίνης Ε με τους παράγοντες E2F1 και E2F4. Πραγματοποιώντας την παλινδρόμηση σε όλα τα δεδομένα που έχουμε τα εξής αποτελέσματα.

Πίνακας 7.1 (Συντελεστές παραμέτρων)

Coefficients				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.29112	0.07181	4.054	0.00382
E2F1	-0.03005	0.15616	-0.192	0.855
E2F4	-0.07027	0.11094	-0.633	0.547

Παρατηρούμε ξανά ότι δεν μπορούμε να υποθέσουμε κάποια σχέση μεταξύ των παραγόντων E2F1 και E2F4 με την Κυκλίνη Ε καθώς οι συντελεστές δεν είναι στατιστικά σημαντικοί (p-value>0.05).

7.5.2 Σχέση μεταξύ του δείκτη Ki 67 και της Κυκλίνης Ε

Στο 5ο κεφάλαιο είχαμε δει ότι λαμβάνοντας υπόψη όλα τα δεδομένα δεν υπήρχε κάποια σχέση μεταξύ των δύο δεικτών. Το ίδιο αποτέλεσμα έχουμε και τώρα όπως δείχνει και ο παρακάτω πίνακας.

Πίνακας 7.2 (Συντελεστές παραμέτρων)

Coefficients				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.20373	0.03229	6.309	9.76e-07
Cyclin E	-0.13628	0.11302	-1.206	0.24

Παρατηρούμε ότι ο συντελεστής της Κυκλίνης Ε δεν είναι στατιστικά σημαντικός οπότε ξανά καταλήγουμε στο αρχικό μας συμπέρασμα. Στην συνέχεια παρουσιάζουμε για κάθε ένα από τα έξι νέα σετ δεδομένων τον συντελεστή συσχέτισης του Spearman και τα αντίστοιχα p-values όπως κάναμε και στο 5ο κεφάλαιο. Από τον παρακάτω πίνακα βλέπουμε ότι στατιστικά σημαντικά αποτελέσματα έχουμε μόνο στην περίπτωση που πάρουμε ως κατώφλι για την Κυκλίνη Ε την τιμή 20%, όπου παρατηρούμε ότι υπάρχει αρνητική συσχέτιση μεταξύ της Κυκλίνης Ε με τον δείκτη Ki 67. Σε όλες τις άλλες περιπτώσεις δεν έχουμε κάποιο στατιστικά σημαντικό αποτέλεσμα.

Πίνακας 7.3 (συντελεστές του Spearman)

Spearman's rho						
	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6
CyclinE >20	-0.568	-0.505	-0.593	-0.604	-0.442	-0.522
	0.008	**0.027**	**0.007**	**0.003**	**0.05**	**0.018**
CyclinE <20	-0.061	-0.081	-0.053	-0.03	0.06	-0.027
	0.78	0.704	0.8023	0.872	0.784	0.899
CyclinE >30	0.102	0.45	-0.165	-0.151	0.169	-0.084
	0.724	0.122	0.6267	0.637	0.598	0.793
CyclinE <30	-0.02	-0.019	-0.009	-0.009	0.056	0.011
	0.913	0.916	0.957	0.961	0.761	0.95

7.5.3 Σχέση μεταξύ του grade και των δεικτών Ki 67 και Κυκλίνης Ε

Σε επίπεδο ατόμου αρχικά ελέγξαμε εάν και κατά πόσο οι δείκτες Ki 67 και Κυκλίνη Ε επηρεάζουν τον βαθμό διαφοροποίησης του καρκίνου. Στον παρακάτω πίνακα εμφανίζονται τα συγκεντρωτικά αποτελέσματα.

Πίνακας 7.4 (Συντελεστές παραμέτρων)

Coefficients				
	Estimate	Std. Error	t-value	Pr(>|t|)
grade≥2	0.1028	0.7120	-0.1444	0.886
grade≥3	-1.6555	0.7864	-2.1052	0.0422
Ki67	7.3042	3.381	2.1602	0.03694
Cyclin E	-0.1858	1.402	-0.1325	0.8957

Παρατηρούμε ξανά ότι μόνο ο δείκτης κυτταρικού πολλαπλασιασμού Ki67 επηρεάζει τον βαθμό κακοήθειας του καρκίνου.

7.5.4 Σχέση μεταξύ του grade και των παραγόντων E2F1 και E2F4

Εδώ θα δείξουμε μέσω της χρήσης ενός γενικευμένου γραμμικού μοντέλου ότι ξανά οι παράγοντες E2F1 και E2F4 δεν επηρεάζουν τον δείκτη κακοήθειας του καρκίνου.

Πίνακας 7.5 (Συντελεστές παραμέτρων)

Coefficients				
	Estimate	Std. Error	t-value	Pr(>|t|)
grade≥2	1.1546	0.7106	1.6247	0.1185
grade≥3	-0.3914	0.6839	-0.5723	0.5724
E2F1	-0.5972	1.1973	-0.4988	0.6302
E2F4	0.3426	0.9798	0.3496	0.7286

7.5.5 Σχέση μεταξύ του stage και των δεικτών Ki 67 και Κυκλίνης Ε

Επόμενο βήμα μας ήταν ο έλεγχος του κατά πόσο το στάδιο του καρκίνου επηρεάζεται από τους δύο παραπάνω δείκτες. Ακολουθώντας την ίδια διαδικασία όπως και πριν έχουμε τον παρακάτω πίνακα.

Πίνακας 7.6 (Συντελεστές παραμέτρων)

Coefficients				
	Estimate	Std. Error	t-value	Pr(>|t|)
stage≥pT1	-1.9162	0.8233	-2.3275	0.0274
stage≥pT2	-2.8925	0.9107	-3.1760	0.0032
Ki67	8.90977	3.565	2.49903	0.02198
Cyclin E	0.05728	1.407	0.1750	0.96779

Βλέπουμε ξανά ότι τα αποτελέσματα είναι παρόμοια με τα αρχικά μας στο κεφάλαιο 6. Από τις 2 ανεξάρτητες μεταβλητές μόνο ο δείκτης Ki 67 επηρεάζει το στάδιο του καρκίνου.

7.5.6 Σχέση μεταξύ του stage και των παραγόντων E2F1 και E2F4

Τα αποτελέσματα που παίρνουμε σχετικά με το στάδιο του καρκίνου και την σχέση του με τους παράγοντες E2F1 και E2F4 δεν διαφέρουν από την αρχική μας ανάλυση. Δηλαδή δεν παρατηρείται κάποια στατιστικά σημαντική σχέση μεταξύ των δεικτών και του σταδίου του καρκίνου.

Πίνακας 7.7 (Συντελεστές παραμέτρων)

Coefficients				
	Estimate	Std. Error	t-value	Pr(>|t|)
stage≥pT1	-0.4369	0.7183	-0.6082	0.5499
stage≥pT2	-1.2145	0.7145	-1.6998	0.0973
E2F1	0.3929	1.336	0.2941	0.7768
E2F4	-0.1059	1.036	-0.1021	0.9195

7.5.7 Σχέση μεταξύ του risk και των δεικτών Ki 67 και Κυκλίνης Ε

Το τελευταίο πράγμα που ελέγξαμε ήταν η σχέση μεταξύ του κινδύνου που διατρέχει ο ασθενής και των δύο δεικτών. Οι τελικοί συντελεστές που παίρνουμε είναι παρόμοιοι με αυτούς που είχαμε πάρει με την αρχική ανάλυση.

Πίνακας 7.8 (Συντελεστές παραμέτρων)

Coefficients				
	Estimate	Std.	z value	Pr(>|z|)
(Intercept)	-2.4658	1.0447	-2.36	0.01826
Ki67	12.7305	3.9450	3.227	0.00125
Cyclin E	0.3376	2.4456	0.138	0.89019

Όπως περιμέναμε από την στιγμή που η Κυκλίνη Ε δεν είναι στατιστικά σημαντική όσον αφορά την σχέση της με τον βαθμό κακοήθειας και το στάδιο του καρκίνου θα ήταν απίθανο να βρούμε ότι σχετίζεται με τον κίνδυνο (risk) του ασθενή.

Τέλος, δύο παρατηρήσεις σχετικά με την διαδικασία και τα αποτελέσματα που βγάλαμε.

Πρώτον, βλέποντας τους πίνακες 7.4, 7.5 και 7.6 και συγκρίνοντας τους με τους αντίστοιχους στο $6^{ο}$ κεφάλαιο (βλ. πιν. 6.9, 6.22 και 6.32) παρατηρούμε ότι οι συντελεστές του Ki67 δεν είναι οι ίδιοι και πιο συγκεκριμένα οι συντελεστές των πινάκων του $7^{ου}$ κεφαλαίου (όπως και τα τυπικά σφάλματα) είναι πολλαπλασιασμένα με το 100. Πρόκειται για ένα καθαρά τεχνικό θέμα και οφείλεται στο γεγονός ότι για να εφαρμόσουμε την τεχνική της πολλαπλής εισαγωγής έπρεπε να δηλώσουμε στο πρόγραμμα ότι οι τιμές των δεικτών είναι ποσοστά ώστε να αποφευχθεί το ενδεχόμενο να πάρουμε τιμή που δεν είναι εφικτή (π.χ αρνητική ή μεγαλύτερη του 100). Για αυτό τον λόγο έπρεπε να διαιρέσουμε όλες τις τιμές με το 100.

Δεύτερον, θα πρέπει να κρατήσουμε στο μυαλό μας ότι η πολλαπλή εισαγωγή είναι ευαίσθητη στις υποθέσεις που κάνουμε κατά την διαδικασία της δημιουργίας των νέων τιμών και ειδικότερα στην υπόθεση ότι οι ελλιπής τιμές είναι MAR. Η συγκεκριμένη θεωρία όμως είναι γενικότερη και μπορεί να εφαρμοστεί και όταν οι ελλιπείς τιμές είναι MNAR αλλά σε αυτήν την περίπτωση το μοντέλο που εφαρμόζεται στα πλήρη δεδομένα είναι λάθος για τις ελλιπείς τιμές και άρα είναι ακατάλληλο για την εισαγωγή των τιμών. Εκτός και αν έχουμε επιπλέον εξωτερικά δεδομένα, δεν είναι δυνατόν να προβλέψουμε το μέγεθος του σφάλματος. Πάντως

στην δική μας περίπτωση λόγω της φύσης του προβλήματος αλλά και ύστερα από τις διαβεβαιώσεις του κ. Κοτσίνα, κατόπιν προσωπικής επικοινωνίας, δεν έχουμε κάποιον ισχυρό λόγο για να αμφιβάλουμε για την υπόθεση ότι τα ελλιπή δεδομένα είναι MAR.

Κεφάλαιο 8

Μη παραμετρική παλινδρόμηση

Στα κεφάλαια που προηγήθηκαν ασχοληθήκαμε με την παραμετρική προσέγγιση στο πρόβλημα μας. Σε αυτό το κεφάλαιο θα ασχοληθούμε με την μη παραμετρική παλινδρόμηση (nonparametric regression), κύριο χαρακτηριστικό της οποίας είναι η χαλάρωση της υπόθεσης της γραμμικότητας και η αντικατάσταση της με την υπόθεση της ομαλότητας της πληθυσμιακής συνάρτησης παλινδρόμησης. Το κόστος αυτής της χαλάρωσης της υπόθεσης της γραμμικότητας είναι κυρίως υπολογιστικό και σε μερικές περιπτώσεις ένα πιο δυσνόητο αποτέλεσμα. Το πλεονέκτημα είναι η πιο ακριβής εκτίμηση της συνάρτησης παλινδρόμησης. Η παρακάτω παρουσίαση έχει στηριχτεί κυρίως στα δύο βιβλία του John Fox "Nonparametric simple regression" και "Multiple and generalized nonparametric regression". Στο τελευταίο μέρος του κεφαλαίου μας θα προσεγγίσουμε κάποια σημεία του προβλήματος μας μέσω της μη παραμετρικής παλινδρόμησης..

8.1 Σύγκριση παραμετρικής με μη παραμετρική παλινδρόμηση

Η παραμετρική προσέγγιση έχει το πλεονέκτημα ότι είναι πιο αποτελεσματική εάν το μοντέλο είναι σωστό. Εάν έχουμε καλή πληροφόρηση σχετικά με την κατάλληλη οικογένεια μοντέλων τότε θα πρέπει να προτιμήσουμε ένα παραμετρικό μοντέλο καθώς και οι παράμετροι έχουν επίσης κάποια διαισθητική ερμηνεία. Τα μη παραμετρικά μοντέλα δεν γίνεται να περιγράψουν την σχέση μεταξύ των ερμηνευτικών μεταβλητών και της μεταβλητής απόκρισης με την μορφή ενός τύπου, αλλά αυτό συχνά γίνεται γραφικά. Από την άλλη πλευρά η μη παραμετρική προσέγγιση είναι πιο ευέλικτη, καθώς όταν μοντελοποιούμε καινούργια δεδομένα συχνά δεν ξέρουμε ποια είναι η κατάλληλη μορφή του μοντέλου, οπότε αυτή η έρευνα μπορεί να γίνει μέσω της μη παραμετρικής παλινδρόμησης και για αυτό η μη παραμετρική προσέγγιση είναι ιδιαίτερα χρήσιμη όταν έχουμε μικρή εμπειρία σχετικά με τα δεδομένα. Τέλος με την παραμετρική προσέγγιση μπορεί εύκολα να γίνει λάθος επιλογή μοντέλου, ενώ η μη παραμετρική προσέγγιση έχει λιγότερες υποθέσεις και κατά συνέπεια είναι πιο δύσκολο να γίνουν σοβαρά λάθη (J.Faraway 2006).

8.2 Απλή μη παραμετρική παλινδρόμηση (simple nonparametric regression)

Όπως και με τις προηγούμενες μεθόδους έτσι και εδώ ξεκινάμε με την περίπτωση που έχουμε μια μεταβλητή απόκρισης και μία ανεξάρτητη μεταβλητή. Η απλή μη παραμετρική παλινδρόμηση έχει συχνά τον χαρακτηρισμό «εξομαλυντής του διαγράμματος διασποράς» καθώς η κύρια χρησιμότητα της είναι η χάραξη μιας ομαλής καμπύλης σε ένα διάγραμμα διασποράς δύο μεταβλητών.

8.2.1 Τοπικός μέσος και εκτίμηση μέσω πυρήνων (Kernel estimation)

Η βασική ιδέα πίσω από τον τοπικό μέσο είναι ότι, δεδομένου ότι η συνάρτηση παλινδρόμησης είναι ομαλή (smooth) οι παρατηρήσεις με τιμή x κοντά στην κεντρική τιμή x_0 μας δίνουν πληροφορίες σχετικά με την $f(x_0)=\mu|x_0$. Ο τρόπος που λειτουργεί η παραπάνω διαδικασία είναι η εξής: περνάμε ένα «πλαίσιο» πάνω από τα δεδομένα παίρνοντας τον μέσο των παρατηρήσεων που πέφτουν μέσα σε αυτό το πλαίσιο. Γενικά δεν γίνεται να εκτιμήσουμε την συνάρτηση παλινδρόμησης για ένα άπειρο αριθμό από x αλλά μπορούμε να εκτιμήσουμε το $\hat{f}(x)$ για ένα μεγάλο αριθμό από κεντρικά x. Σχετικά με το μέγεθος του πλαισίου μπορούμε να επιλέξουμε να έχει σταθερό πλάτος (bandwidth) ή να μεταβάλλεται ώστε σε κάθε ένα να περιλαμβάνονται ένα συγκεκριμένο αριθμό παρατηρήσεων m. αυτοί είναι οι m κοντινότεροι γείτονες της κεντρικής τιμής x.

Η εκτίμηση μέσω πυρήνων (ή ο τοπικά σταθμισμένος μέσος) είναι η επέκταση των τοπικών μέσων, η βασική ιδέα εδώ είναι ότι κατά την εκτίμηση της $f(x_0)$ θέλουμε να δίνεται μεγαλύτερη βαρύτητα στις παρατηρήσεις που είναι κοντά στην κεντρική x_0 και λιγότερη σε όσες είναι πιο μακριά. Έστω $z_i = (x_i - x_0)/h$ όπου h το εύρος του εκτιμητή Kernel και έχει παρόμοιο ρόλο με το πλάτος του πλαισίου στον τοπικό μέσο. Χρειαζόμαστε μια συνάρτηση πυρήνα (Kernel function) K(z) η οποία δίνει μεγαλύτερο βάρος στις παρατηρήσεις κοντά στο κεντρικό x_0 και μειώνεται ομαλά και συμμετρικά καθώς το |z| μεγαλώνει. Έχοντας υπολογίσει τα βάρη $w_i = K[(x_i - x_0)/h]$ συνεχίζουμε με τον υπολογισμό της προσαρμοσμένης τιμής στο x_0 μέσω της σχέσης

$$\hat{f}(x_0) = \hat{y}|x_0 = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}.$$

Οι δύο πιο συνηθισμένες επιλογές για τις συναρτήσεις των πυρήνων είναι ο Γκαουσιανός ή κανονικός (Gaussian ή normal) και ο τρικυβικός (tricube).

- Ο κανονικός πυρήνας είναι απλά η συνάρτηση πυκνότητας της τυπικής κανονικής:

$$K_N(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

όπου εδώ h είναι η τυπική απόκλιση μιας κανονικής κατανομής κεντραρισμένη στο x_0. Παρατηρήσεις σε αποστάσεις μεγαλύτερες από 2h από την κεντρική τιμή έχουν βάρος σχεδόν 0, επειδή η πυκνότητα της κανονικής κατανομής είναι μικρή πέρα από 2 τυπικές αποκλίσεις από τον μέσο.

- Ο τρικυβικός πυρήνας δίνεται από την σχέση:

$$K_T(z) = \begin{cases} (1-|z|^3)^3 & για\ |z| < 1 \\ 0 & για\ |z| \geq 1 \end{cases}$$

Εδώ το h είναι το μισό πλάτος ενός πλαισίου κεντραρισμένο στην κεντρική τιμή x_0. Παρατηρήσεις που πέφτουν έξω από το παράθυρο έχουν βάρος 0.

Στους παραπάνω τύπους υποθέσαμε ότι το πλάτος των πλαισίων h είναι σταθερό, αλλά οι εκτιμητές μέσω πυρήνων μπορούν εύκολα να προσαρμοστούν στην περίπτωση που δουλέψουμε στην περίπτωση με τους m κοντινότερους γείτονες. Η προσαρμογή αυτή είναι πιο απλή για εκτιμητές όπως ο τρικυβικός ο οποίος πέφτει στο 0, απλώς ορίζουμε το h(x) ώστε ένας συγκεκριμένος αριθμός m παρατηρήσεων να περιλαμβάνονται στο πλαίσιο. Το κλάσμα m/n ονομάζεται εύρος (span) της ομαλότητας του πυρήνα. Τους εκτιμητές μέσω πυρήνων μπορούμε να τους εκτιμήσουμε είτε για ένα σύνολο τιμών ισότιμα κατανεμημένες κατά μήκος των x είτε για τις διατεταγμένες παρατηρήσεις $x_{(i)}$.

8.2.2 Τοπική πολυωνυμική παλινδρόμηση (local polynomial regression)

Η τοπική πολυωνυμική παλινδρόμηση παρέχει μια γενικά επαρκή μέθοδο μη παραμετρικής παλινδρόμησης η οποία επεκτείνεται άμεσα και στην πολλαπλή παλινδρόμηση (multiple regression), στην προσθετική παλινδρόμηση (additive regression) και στην γενικευμένη μη παραμετρική παλινδρόμηση (generalized nonparametric regression). Μια εφαρμογή της τοπικής πολυωνυμικής παλινδρόμησης

ονομάζεται lowess (ή loess) και είναι η πιο συνηθισμένη μέθοδος μη παραμετρικής παλινδρόμησης.

Έστω ένα πολυώνυμο p βαθμού για την μεταβλητή πρόβλεψης x.

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon$$

Η τοπική πολυωνυμική παλινδρόμηση επεκτείνει την εκτίμηση Kernel στο κεντρικό σημείο x_0 σε πολυωνυμική μορφή χρησιμοποιώντας τοπικά βάρη $w_i = K[(x_i - x_0)/h]$. Η σταθμισμένη παλινδρόμηση ελαχίστων τετραγώνων (weighted least squares regression) δίνεται από την σχέση:

$$y_i = a_i + b_{1i}(x_i - x_0) + b_{2i}(x_i - x_0)^2 + \cdots + b_{pi}(x_i - x_0)^p + e_i,$$

η οποία ελαχιστοποιεί το σταθμισμένο άθροισμα τετραγώνων των υπολοίπων (weighted residual sum of squares) $\sum_{i=1}^{n} {w_i}^2 {e_i}^2$. Μέσω αυτής της διαδικασίας η προσαρμοσμένη τιμή για το x_0 είναι $\hat{y}|x_0 = a$. Η διαδικασία αυτή επαναλαμβάνεται και για τις άλλες παρατηρήσεις των x. όπως και πριν έτσι και εδώ το πλάτος h μπορεί να είναι σταθερό ή να μεταβάλλεται με βάση τις κοντινότερες παρατηρήσεις.

Η πιο συνηθισμένη επιλογή είναι p=1 δηλαδή η γραμμική μορφή, η οποία μειώνει την μεροληψία σε σχέση με τους εκτιμητές μέσω πυρήνων που είδαμε προηγουμένως και αντιστοιχούν στην περίπτωση p=0. Αυτό το πλεονέκτημα είναι πιο ορατό στα άκρα όπου οι kernel εκτιμητές φαρδαίνουν. Οι τιμές p=2 και p=3 παράγουν πιο ευέλικτες παλινδρομήσεις, οι οποίες μπορεί να μειώνουν την μεροληψία αλλά αυτό συνεπάγεται μεγαλύτερη διακύμανση. Επίσης προκύπτει ένα πλεονέκτημα των πολυωνύμων μονού βαθμού έναντι αυτών με άρτιο βαθμό, δηλαδή το πολυώνυμο $1^{ο}$ βαθμού γενικά προτιμάται από του μηδενικού βαθμού και όμοια αυτό ισχύει για το $3^{ο}$ έναντι του $2^{ου}$.

Η απόδειξη για το παραπάνω οφείλεται στο γεγονός ότι η μεροληψία της εκτίμησης Kernel εξαρτάται από την κατανομή των X ενώ της τοπικής γραμμικής (locally linear) όχι οπότε στα άκρα των δεδομένων όπου οι παρατηρήσεις είναι άνισα κατανεμημένες έχει μικρότερη μεροληψία και από την στιγμή που και οι δύο εκτιμήσεις ασυμπτωτικά έχουν την ίδια διακύμανση οι τοπικά γραμμικές εκτιμήσεις έχουν μικρότερο MSE. Το παραπάνω αποτέλεσμα γενικεύεται και στην περίπτωση που έχουμε ένα πολυώνυμο άρτιας τάξης p και ένα περιττής p+1, οπότε το πολυώνυμο περιττού έχει ασυμπτωτικά μικρότερο MSE.

8.2.3 Θέματα γύρω από το πλάτος του πλαισίου

Σχετικά με το μέγεθος των πλαισίων αντιμετωπίζουμε ένα πρόβλημα, καθώς το πλάτος πλαισίου μειώνεται ταυτόχρονα μειώνεται και η μεροληψία της εκτίμησης αλλά αυξάνεται η διακύμανση της. Εμείς θέλουμε η μεροληψία (bias) να είναι όσο το δυνατόν μικρότερη οπότε θέλουμε τα πλαίσια να είναι μικρά, αυτό όμως τις περισσότερες φορές έχει ως αποτέλεσμα να αυξάνεται η δειγματική διακύμανση της εκτίμησης του δειγματικού μέσου $\bar{y}_i$. Το μέσο τετραγωνικό σφάλμα της εκτίμησης δίνεται από τον τύπο

$$\mathrm{MSE}\left(\hat{\mathrm{f}}(\mathrm{x}_0)\right) = \mathrm{E}[(\hat{\mathrm{y}}|\mathrm{x}_0 - \mu|\mathrm{x}_0)^2] = \mathrm{bias}^2\left[\hat{\mathrm{f}}(\mathrm{x}_0)\right] + \mathrm{V}\left[\hat{\mathrm{f}}(x_0)\right],$$

όπου

$$\mathrm{bias}\left[\hat{\mathrm{f}}(\mathrm{x}_0)\right] = \mathrm{E}(\hat{\mathrm{y}}|\mathrm{x}_0) - \mathrm{f}(\mathrm{x}_0) \cong \frac{\mathrm{h}^2}{2}\mathrm{s_k}^2\mathrm{f}''(\mathrm{x}_0) \text{ και } \mathrm{V}\left[\hat{\mathrm{f}}(x_0)\right] \cong \frac{\mathrm{s}^2\mathrm{a_k}^2}{\mathrm{nhp}(\mathrm{x}_0)}$$

ενώ s^2 η εκτιμώμενη διακύμανση των σφαλμάτων, h το πλάτος, n το μέγεθος του δείγματος, $\mathrm{f}''(\mathrm{x}_0)$ η 2^{η} παράγωγος της συνάρτησης παλινδρόμησης στο x_0, $\mathrm{p}(\mathrm{x}_0)$ η συνάρτηση πιθανότητας για την κατανομή των x στο σημείο x_0 (αν πχ η $\mathrm{p}(\mathrm{x}_0)$ είναι μικρή αυτό σημαίνει ότι στο σημείο αυτό τα δεδομένα είναι αραιά) και τέλος $\mathrm{s_k}^2$, $\mathrm{a_k}^2$ θετικές σταθερές που εξαρτώνται από την συνάρτηση Kernel.

Σκοπός δικός μας είναι για κάθε τιμή των x όπου πρέπει να εκτιμήσουμε την f(x)=μ|x να βρούμε το κατάλληλο πλάτος h* το οποίο ελαχιστοποιεί το MSE. Σύμφωνα με τα παραπάνω βλέπουμε ότι το να μειώσουμε και την μεροληψία και την διακύμανση ταυτόχρονα δεν γίνεται. Τα μικρά παράθυρα έχουν μικρή μεροληψία και μεγάλη διακύμανση ενώ το αντίθετο ισχύει όταν έχουμε μεγάλα πλαίσια. Αυτό δεν ισχύει όταν μόνο όταν έχουμε μεγάλο δείγμα. Ο τύπος που δίνει το κατάλληλο h* στο σημείο x_0 είναι ο:

$$h^*(x_0) = \left[\frac{{a_k}^2}{{s_k}^4} * \frac{s^2}{np(x_0)[f''(x_0)]^2}\right]^{1/5}$$

(John Fox 2000)

Παρατηρούμε ότι όταν $\mathrm{f}''(\mathrm{x}_0) = 0$ το βέλτιστο πλάτος απειρίζεται που σημαίνει ότι στα δεδομένα μας προσαρμόζεται το γραμμικό μοντέλο. Αυτά ισχύουν όταν επιλέγουμε εμείς το πλάτος, αν δουλέψουμε με την μέθοδο των κοντινότερων γειτόνων το πλάτος για κάθε τιμή του x_0 εξαρτάται από τον σταθερό αριθμό s και δεν λαμβάνει υπόψη του την καμπυλότητα της συνάρτησης παλινδρόμησης.

8.2.4 Επιλογή του κατάλληλου εύρους (span)

Ας υποθέσουμε ότι είμαστε στην περίπτωση που τα πλαίσια εξαρτώνται από τις κοντινότερες παρατηρήσεις και δεν είναι σταθερά και ας υποθέσουμε για λόγους ευκολίας ότι είμαστε στην περίπτωση p=1. Το πρόβλημα της επιλογής του κατάλληλου πλάτους πλαισίου ανάγεται στο πρόβλημα επιλογής του κατάλληλου δείκτη span. Μια γενικά αποτελεσματική μέθοδος είναι μέσω διαδοχικών δοκιμών αν και γενικά η τιμή s=0.5 είναι ικανοποιητική. Αν η προσαρμοσμένη καμπύλη παλινδρόμησης δεν φαίνεται ομαλή (smooth) μπορούμε να αυξήσουμε το νούμερο, ενώ αν φαίνεται ομαλή δοκιμάζουμε να το χαμηλώσουμε και ελέγχουμε αν εξακολουθεί να είναι ομαλή. Αυτό που θέλουμε είναι την μικρότερη τιμή s που επιτυγχάνει αυτό το αποτέλεσμα.

Μια επιπλέον οπτική προσέγγιση έχει να κάνει με τα κατάλοιπα της παλινδρόμησης $e_i = y_i - \hat{y}_i$ όπου προσπαθούμε να τα καταστήσουμε ασυσχέτιστα με την ερμηνευτική μεταβλητή x. Εάν η καμπύλη παλινδρόμησης υπερεξομαλύνει (oversmooth) τα δεδομένα τότε παρουσιάζεται μια συστηματική σχέση μεταξύ των καταλοίπων και της x. εάν δεν το κάνει αυτό τότε τα κατάλοιπα έχουν μέση τιμή 0 ανεξάρτητα από την τιμή της x. Οπότε ψάχνουμε το μεγαλύτερο s για το οποίο τα κατάλοιπα είναι ασυσχέτιστα με το x.

Πέρα της οπτικής μεθόδου για τον υπολογισμό του κατάλληλου εύρους η οποία δίνει καλά αποτελέσματα, υπάρχει και η καθαρά υπολογιστική που την χρησιμοποιούμε είτε για να βρούμε το βέλτιστο σταθερό (fixed) ή μεταβαλλόμενο πλάτος πλαισίου. Η μέθοδος με την οποία γίνεται ονομάζεται διασταυρωμένη επικύρωση ή cross-validation (CV), κύρια ιδέα της οποίας είναι να παραλείψουμε την i παρατήρηση από την τοπική παλινδρόμηση για την τιμή x_i. Δηλώνουμε ως $E(y|x_i) = \hat{y}_{-i}|x_i$, όπου παραλείποντας την i παρατήρηση η προσαρμοσμένη τιμή $\hat{y}_{-i}|x_i$ είναι ανεξάρτητη από την παρατηρηθείσα τιμή y_i. Η CV συνάρτηση είναι η εξής:

$$CV(s) = \frac{\sum_{i=1}^{n}[\hat{y}_{-i}(s) - y_i]^2}{n},$$

όπου $\hat{y}_{-i}(s)$ είναι το $\hat{y}_{-i}|x_i$ για εύρος s. Σκοπός μας είναι να βρούμε ποια τιμή του s ελαχιστοποιεί το CV.

Η συνάρτηση CV εκτιμά το μέσο αναμενόμενο τετραγωνικό σφάλμα (mean average squared error) ή MASE για τα παρατηρημένα x.

$$\mathrm{MASE(s)} = \mathrm{E}\left\{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}[\hat{\mathrm{y}}_{\mathrm{i}}(\mathrm{s}) - \mu_{\mathrm{i}}]^2}{\mathrm{n}}\right\}$$

Λόγω της ανεξαρτησίας των $\hat{\mathrm{y}}_{-\mathrm{i}}$ και y_i η αναμενόμενη τιμή της CV(s) είναι

$$\mathrm{E[CV(s)]} = \frac{\sum_{i=1}^{n} \mathrm{E}[\hat{y}_{-i}(s) - y_i]^2}{n} \cong \mathrm{MASE(s)} + \sigma^2$$

Η αντικατάσταση των μ_{i} με $\mathrm{y_i}$ αυξάνει την $\mathrm{E[CV(s)]}$ κατά σ^2 αλλά επειδή το σ είναι σταθερό η τιμή για την οποία ελαχιστοποιείται η $\mathrm{E[CV(s)]}$ είναι η ίδια (σχεδόν) που ελαχιστοποιεί την ποσότητα MASE.

Αν και η μέθοδος της διασταυρωμένης επικύρωσης δίνει αρκετά ικανοποιητικά αποτελέσματα ως προς την εύρεση του κατάλληλου δείκτη s θα πρέπει να λάβουμε υπόψη ότι η τιμή s που παίρνουμε είναι απλώς μια εκτίμηση και για ως εκ τούτου παρουσιάζει μεταβλητότητα. Ειδικά σε μικρά δείγματα η μεταβλητότητα μπορεί να είναι σημαντική, επίσης για μικρά δείγματα η παραπάνω μέθοδος τείνει να δίνει τιμές για τον δείκτη s που είναι πολύ μικρές. Για αυτό τον λόγο καλό θα είναι να μην αποδεχόμαστε αμέσως την τιμή που παίρνουμε με την παραπάνω μέθοδο αλλά να την χρησιμοποιήσουμε ως τιμή οδηγό και με βάση αυτή να ψάξουμε να βρούμε την κατάλληλη τιμή.

8.2.5 Πώς να κάνουμε την τοπική παλινδρόμηση ανθεκτική στις ακραίες τιμές

Όπως και στην γραμμική παλινδρόμηση έτσι και εδώ οι ακραίες τιμές έχουν ως αποτέλεσμα ο εκτιμητής ελαχίστων τετραγώνων να είναι λάθος. Μια αντιμετώπιση του συγκεκριμένου προβλήματος είναι με το να υποβαθμίσουμε την επίδραση των ακραίων τιμών.

Ας υποθέσουμε ότι προσαρμόζουμε την τοπική παλινδρόμηση στα δεδομένα μας και παίρνουμε τις εκτιμήσεις $\hat{\mathrm{y}}_{\mathrm{i}}$ και τα αντίστοιχα κατάλοιπα $\mathrm{e_i} = \mathrm{y_i} - \hat{\mathrm{y}}_{\mathrm{i}}$. Μεγάλα κατάλοιπα αντιπροσωπεύουν ακραίες παρατηρήσεις σε σχέση με τις προσαρμοσμένες. Ορίζουμε τα βάρη $\mathrm{W_i} = \mathrm{W(e_i)}$ όπου η συνάρτηση $W(\cdot)$ θέτει μέγιστο βάρος στα μηδενικά κατάλοιπα και μειώνει το βάρος όσο τα κατάλοιπα μεγαλώνουν κατά απόλυτη τιμή. Δύο είναι οι πιο συνηθισμένες συναρτήσεις ορισμού των βαρών:

- Η πιο δημοφιλής επιλογή είναι η συνάρτηση bisquare ή biweight

$$W_i = W_B(e_i) = \begin{cases} \left[1 - \left(\frac{e_i}{cS}\right)^2\right]^2 & |e_i| < cS \\ 0 & |e_i| > cS \end{cases} \quad \text{(John Fox 2000)}$$

(John Fox 2000)

Όπου S ένα μέτρο διασποράς των καταλοίπων, όπως η διάμεσος των απόλυτων καταλοίπων δηλ. S=median$|e_i|$ και c είναι μια σταθερά που εξισορροπεί την αντίσταση στις ακραίες παρατηρήσεις με την αποδοτικότητα όταν έχουμε κανονικά σφάλματα. Μικρές τιμές του c έχουν ως αποτέλεσμα μεγαλύτερη αντίσταση στις ακραίες τιμές αλλά έχει μικρότερη απόδοση αν τα σφάλματα ακολουθούν την κανονική κατανομή. Διαλέγοντας c=7 (και χρησιμοποιώντας ως μέτρο διασποράς την διάμεσο των απόλυτων αποκλίσεων) πετυχαίνουμε το 95% της αποτελεσματικότητας συγκρινόμενη με την μέθοδο των ελαχίστων τετραγώνων όταν τα σφάλματα είναι κανονικά, συνήθως προτιμάται η τιμή c=6.

- Η δεύτερη επιλογή είναι η συνάρτηση του Huber:

$$W_i = W_H(e_i) = \begin{cases} 1 & |e_i| < cS \\ cS/|e_i| & |e_i| > cS \end{cases}$$

Σε αντίθεση με την προηγούμενη περίπτωση η συνάρτηση Huber δεν γίνεται να πάρει την τιμή 0 ενώ θέτοντας c=2 έχουμε πάλι το 95 % της αποτελεσματικότητας για σφάλματα που ακολουθούν την κανονική κατανομή

Ο τρόπος που δουλεύουμε είναι ο εξής αφού υπολογίσουμε τα αρχικά βάρη (είτε τα bisquare είτε τα Huber) ξανατρέχουμε την παλινδρόμηση στα διάφορα x_i και θέλουμε να ελαχιστοποιήσουμε του σταθμισμένο άθροισμα τετραγώνων των υπολοίπων (weighted residual sum of squares) $\sum_{i=1}^{n} {w_i}^2 {W_i}^2 {e_i}^2$. Επειδή οι ακραίες τιμές θα επηρεάσουν τις αρχικές προσαρμοσμένες τιμές και κατά συνέπεια και τα αρχικά σφάλματα και βάρη θα πρέπει να επαναλάβουμε την διαδικασία κάμποσες φορές μέχρι πλέον οι προσαρμοσμένες τιμές $\hat{y}_i$ να είναι σταθερές. Συνήθως δύο με τέσσερις επαναλήψεις αρκούν.

8.2.6 Έλεγχοι υποθέσεων

Όπως και στο απλό γραμμικό μοντέλο έτσι και εδώ ο έλεγχος για το αν η ανεξάρτητη μεταβλητή μας είναι στατιστικά σημαντική γίνεται μέσω ενός F test με στατιστική συνάρτηση ελέγχου:

$$\mathrm{F} = \frac{(\mathrm{TSS} - \mathrm{RSS})/(\mathrm{df}_{\mathrm{mod}} - 1)}{\mathrm{RSS}/\mathrm{df}_{\mathrm{res}}} ,$$

όπου TSS είναι το ολικό άθροισμα τετραγώνων (total sum of squares), RSS είναι το άθροισμα τετραγώνων των καταλοίπων του μη παραμετρικού μοντέλου (residual sum of squares), ενώ $\mathrm{df}_{\mathrm{mod}} = \mathrm{trace}(\mathbf{S}), \mathrm{trace}(\mathbf{SS}')$ ή $\mathrm{trace}(\mathbf{2S} - \mathbf{SS}')$ είναι οι βαθμοί ελευθερίας του μοντέλου και $\mathrm{df}_{\mathrm{res}} = \mathrm{n} - \mathrm{df}_{\mathrm{mod}}$ είναι οι βαθμοί ελευθερίας των καταλοίπων ενώ η διακύμανση των σφαλμάτων εκτιμάται από την σχέση $S^2 = \frac{\sum \mathrm{e_i}^2}{\mathrm{df}_{\mathrm{res}}}$. Όμοια με το απλό γραμμικό μοντέλο έτσι και εδώ μεγάλες τιμές του F οδηγούν σε απόρριψη της υπόθεσης της ανεξαρτησίας.

Ένας άλλος έλεγχος που υπάρχει είναι αυτός της γραμμικότητας μεταξύ της μεταβλητής απόκρισης και της ανεξάρτητης, ο οποίος γίνεται συγκρίνοντας το μη παραμετρικό μοντέλο με το απλό γραμμικό και αυτό γίνεται επειδή η γραμμική σχέση θεωρείται υποπερίπτωση μιας γενικότερης μη γραμμικής. Η στατιστική του ελέγχου δίνεται από την σχέση:

$$\mathrm{F} = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(\mathrm{df}_{\mathrm{mod}} - 2)}{\mathrm{RSS}_1/\mathrm{df}_{\mathrm{res}}} ,$$

όπου RSS_0 είναι το άθροισμα τετραγώνων του γραμμικού μοντέλου και RSS_1 το αντίστοιχο άθροισμα για το μη παραμετρικό μοντέλο.

Πριν προχωρήσουμε παρακάτω να ξεκαθαρίσουμε ποιος είναι ο πίνακας **S** που χρησιμοποιείται για τον υπολογισμό των βαθμών ελευθερίας του μοντέλου. Όπως είδαμε παραπάνω στην τοπική πολυωνυμική παλινδρόμηση οι προσαρμοσμένες τιμές $\hat{\mathrm{y}}_\mathrm{i}$ είναι σταθμισμένα αθροίσματα των παρατηρήσεων $\mathrm{y_i}$: $\hat{y}_i = \sum_{j=1}^{n} s_{ij} y_j$. Όπου τα βάρη s_{ij} είναι συναρτήσεις των x (η κατάσταση περιπλέκεται κάπως όταν υπάρχουν επαναλήψεις (βλέπε 8.2.5) όπου τα βάρη εξαρτώνται και από τα $\mathrm{y_i}$. Όλα αυτά τα s_{ij} βάρη μπορούν να συγκεντρωθούν σε έναν πίνακα που ονομάζεται πίνακας ομαλότητας **S** (smoother matrix), ο οποίος έχει την μορφή

$$\mathbf{S}_{(\mathrm{nxn})} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & & s_{2n} \\ \vdots & & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{pmatrix}$$

Οπότε $\hat{\mathbf{y}}_{(\mathrm{nx1})} = \mathbf{S}\mathbf{y}_{(\mathrm{nx1})}$, όπου $\hat{\boldsymbol{y}} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n]'$ είναι το διάνυσμα στήλη με τις προσαρμοσμένες τιμές και $\boldsymbol{y} = [y_1, y_2, \dots, y_n]'$ το διάνυσμα στήλη με τις παρατηρηθείσες τιμές.

8.3 splines

Οι splines είναι πολυωνυμικές συναρτήσεις που περιορίζονται στο να ενώνονται ομαλά σε διάφορα σημεία που ονομάζονται κόμβοι (knots). Αν και χρησιμοποιούνται κυρίως για παρεμβολή μπορούν εντούτοις να χρησιμοποιηθούν στην παραμετρική και μη παραμετρική παλινδρόμηση. Στις περισσότερες εφαρμογές χρησιμοποιούνται οι κυβικές (cubic) splines.

8.3.1 Splines παλινδρόμησης

Μια προσέγγιση στην απλή περίπτωση της παλινδρόμησης είναι να προσαρμόσουμε στο x ένα πολυώνυμο μεγάλου βαθμού,

$$y_i = a + \beta_1 x_i + \beta_2 {x_i}^2 + \cdots + \beta_p {x_i}^p + \varepsilon_i$$

το οποίο είναι ικανό να περιγράψει συσχετίσεις που ποικίλουν σε μορφή. Το πρόβλημα είναι ότι οι προσαρμοσμένες τιμές δεν εκτιμώνται τοπικά αλλά επηρεάζονται από μακρινές παρατηρήσεις και επίσης οι εκτιμήσεις υπόκεινται σε μεγάλη δειγματική διακύμανση.

Ως μια εναλλακτική λύση μπορούμε να διαχωρίσουμε τα δεδομένα σε κομμάτια και να εφαρμόσουμε διαφορετικές πολυωνυμικές παλινδρομήσεις σε κάθε ένα από αυτά. Ένα μειονέκτημα αυτής της προσέγγισης είναι ότι οι καμπύλες που προσαρμόζονται σε κάθε κομμάτι είναι σχεδόν βέβαιο ότι δεν θα είναι συνεχής.

Οι κυβικές splines παλινδρόμησης (cubic regression splines) προσαρμόζουν ένα πολυώνυμο τρίτου βαθμού σε κάθε κομμάτι με τους επιπλέον περιορισμούς ότι οι καμπύλες ενώνονται στα όρια του κάθε κομματιού (κόμβοι) και οι δύο πρώτες παράγωγοι είναι συνεχείς στους κόμπους έτσι ώστε να επιτυγχάνεται μια ομαλή καμπύλη.

Οι φυσικές κυβικές splines παλινδρόμησης (natural cubic regression splines) προσθέτουν κόμπους στα όρια των δεδομένων και υποθέτει ότι ισχύει η γραμμική προσέγγιση πέρα από τους τελικούς κόμβους. Με αυτόν τον τρόπο αποφεύγεται η ακραία συμπεριφορά στα άκρα των δεδομένων. Ας υποθέσουμε ότι έχουμε K εσωτερικούς κόμπους και δύο στα άκρα που χωρίζουν τα δεδομένα σε K+1 κομμάτια.

Κάθε κυβική παλινδρόμηση χρησιμοποιεί τέσσερις παραμέτρους αλλά υπάρχουν τρεις περιορισμοί σε κάθε εσωτερικό κόμπο και επιπλέον οι δύο περιορισμοί της γραμμικότητας πέρα από τους ακραίους κόμπους που συνεπάγεται με $4(k+1) - 3k - 2 = k + 2$ ανεξάρτητες παράμετροι. Με τις τιμές των κόμβων δεδομένες η spine παλινδρόμηση είναι στην ουσία ένα γραμμικό μοντέλο. Η πρακτική δυσκολία είναι ότι είναι δύσκολο να αποφασίσουμε πόσοι κόμποι χρειάζονται και που πρέπει να τοποθετηθούν.

8.3.2 Εξομαλυντές splines

Σε αντίθεση με τις splines παλινδρόμησης, οι εξομαλυντές splines (smoothing splines) αποτελούν την λύση στο παρακάτω πρόβλημα μη παραμετρικής παλινδρόμησης: Να βρούμε μία συνάρτηση $\hat{f}(x)$ με δύο συνεχής παραγώγους, η οποία ελαχιστοποιεί το άθροισμα τετραγώνων ποινής (penalized sum of squares),

$$ss^*(h) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + h \int_{x_{min}}^{x_{max}} [f''(x)]^2 dx \qquad (8.1),$$

όπου h είναι μια σταθερά εξομάλυνσης, ανάλογη με το πλάτος πλαισίου των τοπικά πολυωνυμικών εκτιμητών. Ο πρώτος όρος της παραπάνω εξίσωσης είναι το άθροισμα τετραγώνων των σφαλμάτων, ενώ ο δεύτερος είναι η ποινή τραχύτητας (roughness penalty), η οποία είναι μεγάλη όταν η ολοκληρωτική δεύτερη παράγωγος της συνάρτησης παλινδρόμησης $f''(x)$ είναι μεγάλη, δηλαδή όταν η $f(x)$ αλλάζει γρήγορα κλίση. Όταν η σταθερά εξομάλυνσης h είναι ίση με 0, τότε η $\hat{f}(x)$ απλά παρεμβάλει τα δεδομένα, αν όμως το h είναι αρκετά μεγάλο, τότε η $\hat{f}(x)$ θα επιλεγεί έτσι ώστε η $\hat{f}''(x)$ να είναι παντού 0, η οποία είναι ισοδύναμη με μια γενική γραμμική εφαρμογή ελαχίστων τετραγώνων στα δεδομένα.

Αποδεικνύεται ότι η συνάρτηση $\hat{f}(x)$ που ελαχιστοποιεί την συνάρτηση (8.1) είναι μια φυσική κυβική spline με κόμπους στις διακριτές παρατηρηθείσες τιμές των x. Αν και από το αποτέλεσμα φαίνεται ότι χρειάζονται n παράμετροι (όταν όλες οι τιμές των x είναι διαφορετικές), η ποινή τραχύτητας θέτει επιπλέον περιορισμούς μειώνοντας έτσι σημαντικά τον αριθμό των παραμέτρων για τους εξομαλυντές splines, αποτρέποντας έτσι την $\hat{f}(x)$ από το να παρεμβάλει τα δεδομένα. Είναι συχνή πρακτική να επιλέγεται η σταθερά εξομάλυνσης h έμμεσα θέτοντας τον κατάλληλο αριθμό παραμέτρων για την εξομάλυνση.

8.4 Τοπική πολυωνυμική πολλαπλή παλινδρόμηση

Η μέθοδος της τοπικής πολυωνυμικής παλινδρόμησης επεκτείνεται άμεσα από την απλή περίπτωση στην πολλαπλή και είναι σχετικά απλή στην εφαρμογή της. Επιπλέον η τοπική πολυωνυμική πολλαπλή παλινδρόμηση γενικεύεται εύκολα στην περίπτωση που τα δεδομένα μας είναι δίτιμα ή γενικά δεν ακολουθούν την κανονική κατανομή. Τέλος η εφαρμογή της τοπικής πολυωνυμικής πολλαπλής παλινδρόμησης που ονομάζεται lowess ή loess είναι η πιο διαδεδομένη μέθοδος μη παραμετρικής παλινδρόμησης.

8.4.1 Τα βάρη μέσω πυρήνων στην πολλαπλή παλινδρόμηση

Για να αποκτήσουμε την προσαρμοσμένη τιμή $\hat{y}|x_0$ στο σημείο $x_0=(x_{01},x_{02},\ldots,x_{0k})$' πραγματοποιούμε μια σταθμισμένη πολυωνυμική παλινδρόμηση ελαχίστων τετραγώνων του y σε σχέση με τα x δίνοντας έμφαση στις παρατηρήσεις κοντά στο κεντρικό σημείο. Υπάρχουν δύο μέθοδοι υπολογισμού των βαρών στην πολλαπλή παλινδρόμηση.

1. Υπολογίζουμε τα περιθώρια βάρη (marginal weights) ξεχωριστά για κάθε παράγοντα και στην συνέχεια παίρνουμε το γινόμενο των οριακών βαρών, το οποίο σημαίνει ότι για τον j παράγοντα και την i παρατήρηση υπολογίζουμε το περιθώριο βάρος

$$w_{ij} = K\left(\frac{x_{ij} - x_{0j}}{h_j}\right),$$

όπου x_{0j} είναι η κεντρική τιμή του παράγοντα j και h_j το οριακό πλάτος για αυτόν τον παράγοντα. Το πλάτος, όπως και στην απλή περίπτωση, έτσι και εδώ μπορεί να είναι σταθερό ή να προσαρμόζεται ώστε να συμπεριλαμβάνει ένα σταθερό αριθμό από τις κοντινότερες τιμές του x_j. Έχοντας βρει τα οριακά βάρη για τους κ παράγοντες το τελικό βάρος που αντιστοιχεί στην i παρατήρηση της τοπικής παλινδρόμησης είναι το γινόμενο τους:

$$w_i = w_{i1}w_{i2} \ldots w_{ik}$$

2. Υπολογίζουμε την απόσταση $D(\boldsymbol{x}_i, \boldsymbol{x}_0)$ μεταξύ των τιμών των x των παραγόντων για την i παρατήρηση και το κεντρικό $\mathbf{x}_0$. Τα βάρη μπορούν να υπολογιστούν κατευθείαν μέσω αυτών των αποστάσεων.

$$w_i = K\left(\frac{D(\boldsymbol{x}_i, \boldsymbol{x}_0)}{h}\right)$$

Ξανά το πλάτος h μπορεί είτε να είναι σταθερό ή να προσαρμόζεται σύμφωνα με τις κοντινότερες παρατηρήσεις στην κεντρική τιμή. Όμως υπάρχουν διάφοροι τρόποι για να υπολογίσουμε τις αποστάσεις

- Απλή Ευκλείδεια απόσταση (Simple Euclidean distance)

  $$D_E(\mathbf{x}_i, \mathbf{x}_0) = \sqrt{\sum_{j=1}^{k} (x_{ij} - x_{0j})^2}$$

  Η συγκεκριμένη απόσταση έχει νόημα μόνο όταν τα x έχουν ίδιες μονάδες μέτρησης.

- Scaled Ευκλείδεια απόσταση (Scaled Euclidean distance)

  Είναι η πιο συνηθισμένη προσέγγιση για τον ορισμό αποστάσεων καθώς επιτρέπει τον υπολογισμό αποστάσεων μεταξύ μεταβλητών με διαφορετικές μονάδες μέτρησης. Στην περίπτωση αυτή δεν υπολογίζουμε την απόσταση μεταξύ των x αλλά μεταξύ των τυποποιημένων τους τιμών z,

  $$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j},$$

  όπου $\bar{x}_j$ και s_j είναι ο μέσος και η τυπική απόκλιση του x_j. Η scaled Ευκλείδεια απόσταση μεταξύ μιας παρατήρησης $\mathbf{x}_i$ και του κεντρικού σημείου $\mathbf{x}_0$ είναι:

  $$D_s(\mathbf{x}_i, \mathbf{x}_0) = \sqrt{\sum_{j=1}^{k} (z_{ij} - z_{0j})^2}$$

- Γενικευμένη απόσταση (Generalized distance)

  Με τον τρόπο αυτό η απόσταση προσαρμόζεται τόσο σύμφωνα με την διασπορά των x όσο και με την τυχόν συσχέτιση που έχουν μεταξύ τους.

  $$D_G(\mathbf{x}_i, \mathbf{x}_0) = \sqrt{(x_i - x_0)' V^{-1} (x_i - x_0)},$$

  όπου V είναι ο πίνακας διακυμάνσεων των x.

Πέρα από την απλή Ευκλείδεια απόσταση, που για να χρησιμοποιηθεί πρέπει οι παράγοντες να έχουν ίδια κλίμακα μέτρησης, το ποια από τις άλλες θα επιλέξουμε δεν παίζει πολύ σημαντικό ρόλο.

8.4.2 Επιλογή εύρους, συμπερασματολογία και τάξη πολυωνύμου

Σχετικά με την επιλογή του εύρους για την πολλαπλή τοπική πολυωνυμική παλινδρόμηση οι μέθοδοι που υπάρχουν είναι οι ίδιες με την απλή περίπτωση, δηλαδή μέσω της οπτικής δοκιμής και μέσω της διασταυρωμένης επικύρωσης.

Σχετικά με την στατιστική συμπερασματολογία μέσω των F τεστ μπορούμε να συγκρίνουμε διάφορα εναλλακτικά μοντέλα μεταξύ τους και να λάβουμε τις ανάλογες αποφάσεις σχετικά με την σημαντικότητα των μεταβλητών. Έστω ότι θέλουμε να ελέγξουμε κατά πόσο η μεταβλητή X_j επηρεάζει την μεταβλητή απόκρισης Y Η στατιστική συνάρτηση που χρησιμοποιούμε είναι η

$$F = \frac{(RSS_0 - RSS_1)/(df_1 - df_0)}{RSS_1/df_{res}},$$

όπου τα RSS είναι τα αθροίσματα τετραγώνων των καταλοίπων. το RSS_1 αντιστοιχεί στο πλήρες μοντέλο και το RSS_0 στο μοντέλο που έχουμε παραλείψει τον όρο που θέλουμε να ελέγξουμε την σημαντικότητα του. Κάτω από την μηδενική υπόθεση η στατιστική υπόθεση ακολουθεί F κατανομή με $df_1 - df_0$ και $df_{res} = n - df_1$ βαθμούς ελευθερίας.

Όσον αφορά την τάξη του πολυωνύμου πρέπει να πούμε ότι στην πολλαπλή παλινδρόμηση όσο αυξάνεται η τάξη του πολυωνύμου αυξάνονται και οι όροι του και μάλιστα σημαντικά, οπότε στην πράξη θεωρούμε πολυώνυμα πρώτου και δευτέρου βαθμού. Μια τετραγωνική προσαρμογή συνιστάται αν η καμπυλότητα της επιφάνειας παλινδρόμησης αλλάζει πολύ γρήγορα για να περιγραφεί ικανοποιητικά από έναν γραμμικό εκτιμητή. Σε κάποιο βαθμό αυτό μπορεί να διορθωθεί μειώνοντας την τιμή της σταθεράς s καθώς αυτό καθιστά την γραμμική παλινδρόμηση πιο ελαστική. Παρόλα αυτά την επιλογή μεταξύ γραμμικής ή τετραγωνικής μορφής θα την πάρουμε είτε οπτικά είτε μέσω ενός F τεστ ώστε να δούμε αν οι επιπλέον όροι της τετραγωνικής μορφής είναι σημαντικοί.

8.4.3 Εμπόδια στην πολλαπλή μη παραμετρική παλινδρόμηση

Αν και φαίνεται απλή η γενίκευση από την απλή στην πολλαπλή περίπτωση, εντούτοις υπάρχουν δύο σοβαρά ζητήματα

1. Η «κατάρα» των διαστάσεων» (curse of dimensionality): καθώς το πλήθος των παραγόντων αυξάνεται, ο αριθμός των σημείων που είναι κοντά στο κεντρικό σημείο μειώνεται απότομα οπότε για να συμπεριλάβουμε ένα συγκεκριμένο αριθμό παρατηρήσεων στις τοπικές προσαρμογές θα πρέπει να

αυξήσουμε πολύ τα όρια των πλαισίων και κατά συνέπεια η εκτίμηση της $f(x_0)$ δεν θα είναι καλή λόγω της μεγάλης μεροληψίας που θα υπάρχει.

2. Δυσκολίες στην ερμηνεία: η μη παραμετρική παλινδρόμηση δεν δίνει κάποια εξίσωση σχετικά με την μέση απόκριση και τις ανεξάρτητες μεταβλητές, οπότε πρέπει να παρουσιάσουμε την επιφάνεια απόκρισης γραφικά το οποίο σε περίπτωση που έχουμε πολλούς παράγοντες (συνήθως περισσότερους από τρεις) παρουσιάζει δυσκολία στο να γίνει κατανοητό.

Για την αντιμετώπιση των παραπάνω προβλημάτων έχουν δημιουργηθεί μέθοδοι όπως τα αθροιστικά μοντέλα παλινδρόμησης (additive regression models) που θα παρουσιαστούν στην συνέχεια.

8.5 Αθροιστικά μοντέλα παλινδρόμησης (additive regression models)

Στην πολλαπλή μη παραμετρική παλινδρόμηση μοντελοποιούμε την μέση απόκριση των y ως μια γενική εξομαλυμένη συνάρτηση των x.

$$E(y|x_1, x_2, \dots, x_k) = f(x_1, x_2, \dots, x_k)$$

Στην γραμμική παλινδρόμηση αντίθετα η μέση απόκριση της μεταβλητής απόκρισης μοντελοποιείται ως μια γραμμική συνάρτηση των ανεξάρτητων μεταβλητών.

$$E(y|x_1, x_2, \dots, x_k) = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

Ένα αθροιστικό μοντέλο παλινδρόμησης υποθέτει ότι η μέση απόκριση του y είναι το άθροισμα ξεχωριστών όρων για κάθε ανεξάρτητη μεταβλητή, αλλά αυτή οι όροι υποθέτουμε ότι είναι εξομαλυμένες συναρτήσεις των x δηλαδή,

$$E(y|x_1, x_2, \dots, x_k) = a + f_1(x_1) + f_2(x_2) + \cdots + f_k(x_k)$$

Επειδή τα συγκεκριμένα μοντέλα αποκλείουν τις αλληλεπιδράσεις μεταξύ των x είναι πιο περιοριστικά από τα γενικά μοντέλα μη παραμετρικής παλινδρόμησης αλλά πιο ευέλικτα από τα γραμμικά μοντέλα. Ένα πλεονέκτημα τους είναι το γεγονός ότι μετατρέπουν το αρχικό πρόβλημα σε μια σειρά από μερικά (partial) προβλήματα παλινδρόμησης δύο διαστάσεων κάνοντας έτσι πιο εύκολο τόσο τους υπολογισμούς όσο και την ερμηνεία. Επειδή κάθε μερική παλινδρόμηση είναι δισδιάστατη μπορούμε να εκτιμήσουμε την μερική σχέση μεταξύ του y και ενός x_j για παράδειγμα μέσω της τοπικής πολυωνυμικής παλινδρόμησης. Αυτό που πρέπει όμως να κάνουμε είναι με κάποιο τρόπο να απομακρύνουμε τις επιδράσεις των άλλων παραγόντων και όχι απλά να αδιαφορήσουμε για αυτούς.

8.6 Ημιπαραμετρικά μοντέλα και μοντέλα με αλληλεπιδράσεις

Στην παράγραφο αυτή θα περιγράψουμε δύο παραλλαγές των προσθετικών μοντέλων:

1. Τα ημιπαραμετρικά μοντέλα (semiparametric models), τα οποία είναι προσθετικά μοντέλα στα οποία μερικοί όροι εισέρχονται μη παραμετρικά ενώ άλλοι εισέρχονται γραμμικά.
2. Μοντέλα στα οποία κάποιοι από τους παράγοντες επιτρέπεται να αλληλεπιδρούν.

Επίσης είναι πιθανόν αυτά τα δύο να συνδυαστούν. Δηλαδή ορισμένοι όροι να εισέρχονται γραμμικά στο μοντέλο και ταυτόχρονα να υπάρχουν και αλληλεπιδράσεις.

Ένα ημιπαραμετρικό μοντέλο παλινδρόμησης γράφεται στην παρακάτω μορφή

$$y_i = a + b_1 x_{i1} + \cdots + b_r x_{ir} + f_{r+1}(x_{i,r+1}) + \cdots + f_k(x_{ik}) + \varepsilon_i \,,$$

όπου τα σφάλματα ε_i υποθέτουμε πάλι ότι είναι ανεξάρτητα και είναι κανονικά κατανεμημένα με σταθερή διακύμανση. Οι πρώτοι r παράγοντες εισέρχονται στο μοντέλο γραμμικά ενώ για τους υπόλοιπους k-r υποθέτουμε απλώς ότι η σχέση τους με την y ακολουθούν κάποια ομαλή κατανομή. Το συγκεκριμένο μοντέλο μπορεί να εκτιμηθεί μέσω της μεθόδου backfitting. Σε κάθε επανάληψη, όλοι οι γραμμικοί όροι εκτιμούνται σε ένα στάδιο: δημιουργούμε τα μερικά κατάλοιπα που απομακρύνουν τις εκτιμήσεις των μη παραμετρικών όρων και στην συνέχεια κάνουμε παλινδρόμηση των καταλοίπων αυτών με τους γραμμικούς όρους ώστε να πάρουμε τις ανανεωμένες εκτιμήσεις των συντελεστών b. τα ημιπαραμετρικά μοντέλα εφαρμόζονται όταν για κάποιο λόγο πιστεύουμε ότι κάποιος παράγοντας εισέρχεται στην παλινδρόμηση γραμμικά. Αρκετά συχνά όταν κάποια από τα x είναι βουβές μεταβλητές , δηλαδή εκφράζουν την επίδραση ενός ή περισσότερων κατηγορικών μεταβλητών, τότε τους συγκεκριμένους όρους τους εισάγουμε στο μοντέλο με την γραμμική τους μορφή.

Ένα μοντέλο που επιτρέπει αλληλεπιδράσεις (εδώ συγκεκριμένα μεταξύ των x_1 και x_2) έχει την παρακάτω μορφή

$$y_i = a + f_{12}(x_{i1}, x_{i3}) + f_3(x_{i3}) \ldots + f_k(x_{ik}) + \varepsilon_i$$

Ξανά το μοντέλο αυτό υπολογίζεται μέσω της μεθόδου backfitting, όπου για να υπολογίσουμε το f_{12} εφαρμόζουμε έναν εξομαλυντή πολλαπλής παλινδρόμησης όπως την τοπική πολυωνυμική πολλαπλή παλινδρόμηση.

Σχετικά με τους ελέγχους που μπορεί να μας ενδιαφέρουν εδώ είναι στην περίπτωση των ημιπαραμετρικών να δούμε αν πράγματι κάποιοι όροι εισέρχονται

γραμμικά στο μοντέλο. Ενώ στην $2^{η}$ περίπτωση ελέγχουμε αν πράγματι υπάρχει η αλληλεπίδραση. Ο τρόπος για να το ελέγξουμε αυτό και στις δύο περιπτώσεις είναι ξανά μέσω του F test όπως έχει παρουσιαστεί και πιο πάνω.

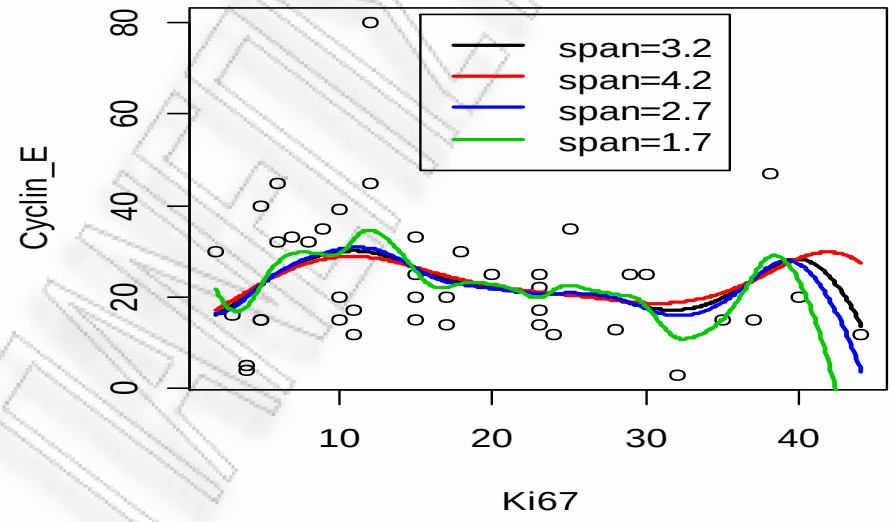
8.7 Εφαρμογή

Δεδομένου του μικρού σχετικά δείγματος η συγκεκριμένη μέθοδος δεν ενδείκνυται ειδικά στην περίπτωση που έχουμε περισσότερες της μιας ανεξάρτητες μεταβλητές. Για το λόγο αυτό ο σκοπός του συγκεκριμένου κεφαλαίου είναι κυρίως να δείξει πως μπορούμε να χρησιμοποιήσουμε τα εργαλεία της μη παραμετρικής παλινδρόμησης σε περίπτωση που τα δεδομένα μας είναι αρκετά μεγάλα ώστε να το επιτρέψουν

8.7.1 Σχέση Ki 67 με Κυκλίνη Ε

Έστω ότι θέλουμε να πάρουμε μια αρχική εικόνα σχετικά με τις δύο μεταβλητές χωρίς να κάνουμε κάποια υπόθεση για τις κατανομές τους. Αυτή θα την πάρουμε μέσω του διαγράμματος διασποράς όπου έχουμε εφαρμόσει και την καμπύλη lowess (βλ. 8.2.2).

Διάγραμμα 8.1 (προσαρμογή καμπύλη lowess στα δεδομένα)



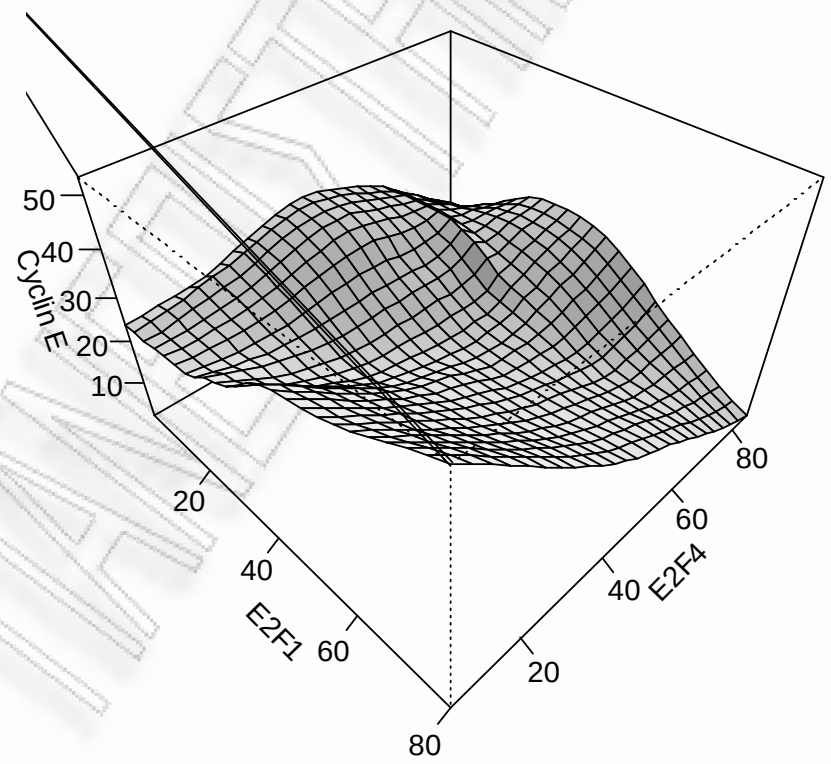
Από το παραπάνω σχήμα παρατηρούμε ότι οι δύο μεταβλητές δεν φαίνεται να έχουν κάποια σχέση κάτι το οποίο επιβεβαιώνεται όπως είδαμε πριν όταν κάναμε την απλή παλινδρόμηση. Σημείωση η αρχική γραμμή που παίρνουμε αν υπολογίσουμε το κατάλληλο εύρος είναι η μαύρη βλέπουμε όμως ότι αν το μειώσουμε λίγο πάλι η

καμπύλη πάλι είναι ομαλή (η μπλε). Αν μειώσουμε το εύρος ακόμα περισσότερο (πράσινη γραμμή) τότε βλέπουμε ότι η καμπύλη δεν μπορεί να χαρακτηριστεί ομαλή ενώ αν το αυξήσουμε πέρα από το αρχικό εύρος (κόκκινη γραμμή) βλέπουμε ότι η καμπύλη υπερεξομαλύνεται οπότε δεν είναι κατάλληλη.

8.7.2 Σχέση Κυκλίνης Ε με τους παράγοντες E2F1 και E2F4

Μια άλλη υπόθεση που θέλαμε να ελέγξουμε είναι κατά πόσο οι δείκτες E2F1 και E2F4 επηρεάζουν την Κυκλίνη Ε. Δουλεύοντας όπως πριν θα μπορούσαμε να πάρουμε ένα αντίστοιχο σχήμα όπως με πριν αλλά και για τις τρεις μεταβλητές μαζί και επιλέγοντας ως εύρος s=0.5. Κοιτώντας τους άξονες μπορούμε να πούμε ότι όσο αυξάνεται το E2F1 αυξάνεται η κυκλίνη και όταν αυξάνεται το E2F4 μειώνεται η Κυκλίνη.

Διάγραμμα 8.2 (Τοπική πολυωνυμική πολλαπλή παλινδρόμηση της Κυκλίνης Ε με τους δείκτες και E2F2 E2F4)



Μπορούμε ακόμα να ελέγξουμε την σημαντικότητα καθεμίας από τις ανεξάρτητες μεταβλητές, διώχνοντας την από το μοντέλο, μέσω ενός F test για την αλλαγή του αθροίσματος τετραγώνων (RSS) μεταξύ των μοντέλων με και δίχως την μεταβλητή. Η διαδικασία αυτή παρουσιάζεται στους δύο παρακάτω πίνακες.

Πίνακας 8.1 (Πίνακας ανάλυσης διακύμανσης)

	Analysis of Variance			
	ENP	RSS	F-value	p-value
Full model	6.86	2710.6	2.3202	0.06251
Model (E2F1)	2.95	4375.2		

Σύμφωνα με τον παραπάνω πίνακα δεν μπορούμε να απορρίψουμε την υπόθεση ότι το μοντέλο μόνο με το E2F1 δεν διαφέρει από το πλήρες μοντέλο (p-value>0.05), δηλαδή το E2F4 δεν είναι στατιστικά σημαντικό.

Πίνακας 8.2 (Πίνακας ανάλυσης διακύμανσης)

	Analysis of Variance			
	ENP	RSS	F-value	p-value
Full model	6.86	2710.6	2.3203	0.05904
Model (E2F4)	2.95	4485.3		

Εδώ δεν μπορούμε να απορρίψουμε ότι το μοντέλο μόνο με το E2F4 δεν διαφέρει από το πλήρες μοντέλο (p-value>0.05), δηλαδή το E2F1 δεν είναι στατιστικά σημαντικό.

Δηλαδή εξαρχής μπορούμε να πάρουμε μια εικόνα σχετικά με το τι συμβαίνει σχετικά με τις μεταβλητές μας και πράγματι όπως και μέσω την συνηθισμένης πολλαπλής παλινδρόμησης βλέπουμε ότι με βάση τα δεδομένα δεν μπορούμε να πούμε ότι η Κυκλίνη Ε επηρεάζεται από τους 2 δείκτες.

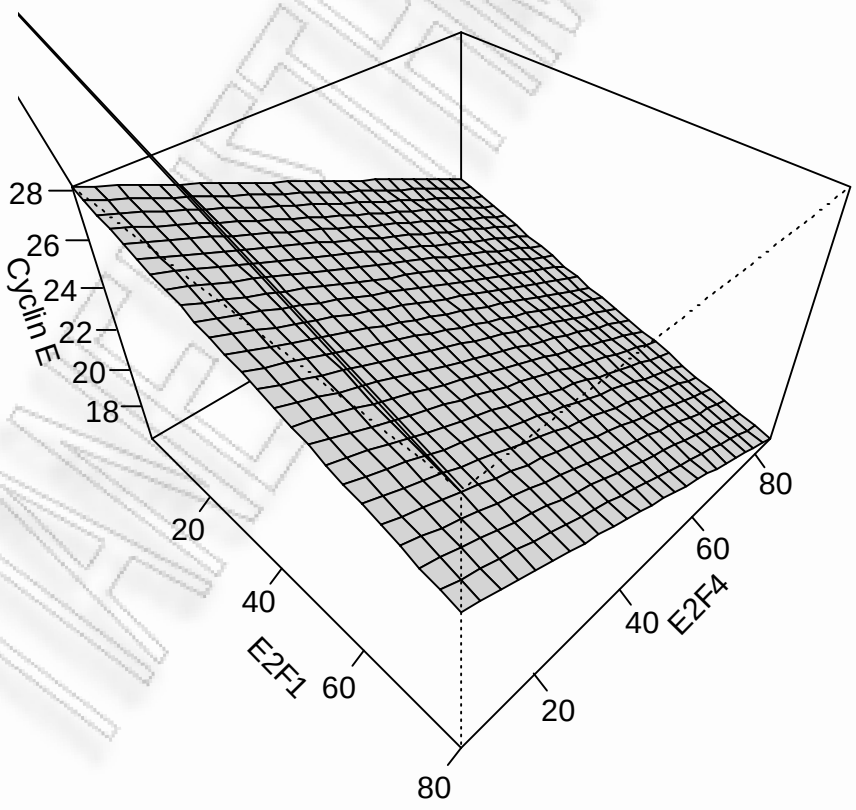
Όπως έχουμε αναφέρει η προηγούμενη μέθοδος έχει το πρόβλημα των διαστάσεων, όπου όσο αυξάνονται οι ανεξάρτητες μεταβλητές τόσο αυξάνεται ο αριθμός των παρατηρήσεων που απαιτούνται αλλά και η δυσκολία της ερμηνείας των σχημάτων. Οπότε μια εναλλακτική μέθοδος είναι το προσθετικό μοντέλο σύμφωνα με το οποίο έχω παρόμοια αποτελέσματα με πριν.

Πίνακας 8.3 (Συντελεστές παραμέτρων)

Parametric coefficients				
	Estimate	Std. Error	t value	p-value
intercept	22.162	1.881	11.78	1.5e-13
Approximate significance of smooth terms				
	edf	Ref.df	F	p-value
s(E2F1)	1	1	0.444	0.510
s(E2F4)	1	1	1.294	0.263

Ενώ το διάγραμμα που παρουσιάζει την σχέση μεταξύ των μεταβλητών είναι το παρακάτω όπου το οπτικό αποτέλεσμα μας ξεγελά καθώς μας δίνει την αίσθηση ότι πράγματι πρέπει να υπάρχει κάποια σχέση μεταξύ των μεταβλητών μας ενώ όπως είδαμε αυτό δεν ισχύει..

Διάγραμμα 8.3 (προσθετική μη παραμετρική παλινδρόμηση της Κυκλίνης Ε με τους δείκτες και E2F2 E2F4)



Μια παρατήρηση, το σύμβολο s(…) σημαίνει ότι σε κάθε όρο έχει προσαρμοστεί ο εξομαλυντής splines. Αν θέλουμε να γενικεύσουμε τα παραπάνω μπορούμε να χρησιμοποιήσουμε το ημιπαραμετρικό μοντέλο που μπορεί να περιέχει και αλληλεπίδραση των μεταβλητών κλπ όπως παρουσιάσαμε προηγουμένως (βλ. 8.6).

ΚΕΦΑΛΑΙΟ 9

Σύγκριση αποτελεσμάτων

Το κεφάλαιο αυτό αποτελείται από τρία μέρη, στο πρώτο μέρος γίνεται μια σύνοψη των αποτελεσμάτων της ανάλυσης μας, στο δεύτερο γίνεται μια επισκόπηση της βιβλιογραφίας που υπάρχει πάνω σε αυτό το θέμα και στο τελευταίο μέρος θα γίνει μια σύγκριση των αποτελεσμάτων της δικής μας ανάλυσης και αυτών που έχουμε παραθέσει.

9.1 Σύνοψη

Στην παράγραφο αυτή θα παρουσιάσουμε συνοπτικά τα αποτελέσματα της ανάλυσης που προηγήθηκε. Το πρώτο κομμάτι αφορούσε την ανάλυση σε επίπεδο κυττάρου, αρχικά ελέγξαμε κατά πόσο τα χαρακτηριστικά των ασθενών (φύλο, ηλικία) επηρεάζουν τους δείκτες Ki67 και Κυκλίνη Ε. Από την ανάλυση μας είδαμε ότι δεν προκύπτει κάποιο στατιστικά σημαντικό αποτέλεσμα. Στην συνέχεια ερευνήσαμε εάν οι πρωτεΐνες E2F1 και E2F4 επηρεάζουν την εμφάνιση της κυκλίνης Ε αλλά ούτε εδώ καταλήξαμε σε κάποιο στατιστικά σημαντικό αποτέλεσμα. Τρίτο κομμάτι στην ανάλυση μας ήταν η σχέση που υπάρχει μεταξύ του δείκτη Ki67 και της κυκλίνης Ε. Εδώ ακολουθήσαμε δύο οδούς, πρώτα ελέγξαμε την σχέση τους σε ολόκληρο το δείγμα όπου δεν βρήκαμε κάποιο στατιστικά σημαντικό αποτέλεσμα. Αντίθετα όταν οριοθετήσαμε τις τιμές κατωφλιού για την Κυκλίνη Ε είδαμε ότι για τιμές της Κυκλίνης Ε μεγαλύτερες του 20% υπάρχει αρνητική σχέση με τον δείκτη Ki67. Το δεύτερο κομμάτι αφορούσε την ανάλυση σε επίπεδο οργάνου (ή ατόμου). Εκεί είδαμε ότι ο δείκτης κυτταρικού πολλαπλασιασμού Ki 67 έχει θετική συσχέτιση τόσο με τον βαθμό κακοήθειας όσο και με το στάδιο του καρκίνου και κατά συνέπεια έχει θετική συσχέτιση με τον κίνδυνο που διατρέχει ο ασθενής. Σχετικά με την Κυκλίνη Ε είδαμε ότι σχετικά με τον βαθμό του καρκίνου όταν αυξάνεται ο βαθμός κακοήθειας του καρκίνου η κυκλίνη Ε μειώνεται παρόλα αυτά το συγκεκριμένο αποτέλεσμα δεν ήταν στατιστικά σημαντικό. Σε σχέση με το στάδιο του καρκίνου είδαμε ότι καθώς το στάδιο μεταβάλλεται από pTa σε pT1 τότε αυξάνεται η Κυκλίνη Ε και στην συνέχεια ξαναπέφτει, ούτε αυτό όμως το αποτέλεσμα ήταν στατιστικά σημαντικό. Σχετικά με την πρωτεΐνη E2F1 είδαμε ότι μειώνεται καθώς αυξάνεται ο

βαθμός του καρκίνου, ενώ μειώνονται καθώς το στάδιο μεταβάλλεται από pTa σε pT1 και στην συνέχεια αυξάνεται πάλι, όμως ξανά τα αποτελέσματα δεν είναι στατιστικά σημαντικά. Σχετικά με την πρωτεΐνη E2F4 είδαμε ότι οι τιμές του είναι σταθερές και στα τρία επίπεδα του βαθμού του καρκίνου αν και παρουσιάζει το μέγιστο για το τρίτο επίπεδο του βαθμού, ενώ όσον αφορά το στάδιο του καρκίνου παρουσιάζει παρόμοια εικόνα με την πρωτεΐνη E2F1.

9.2 Διεθνής βιβλιογραφία

Σχετικά με την διεθνή βιβλιογραφία μπορούμε να πούμε ότι τα αποτελέσματα είναι σε κάποιο βαθμό αντικρουόμενα μεταξύ τους. Σχετικά με τον ρόλο του E2F1 έρευνες έχουν δείξει ότι η πρωτεΐνη E2F1 έχει ογκοκατασταλτική δράση καθώς μείωση του παράγοντα είχε σαν αποτέλεσμα μια πιο επιθετική ασθένεια (Hui-zi Chen et al. 2009, Rebecca Kinkade 2008) ενώ σύμφωνα με άλλες η πρωτεΐνη E2F1 έχει ογκοκατασταλτική δράση στους διηθητικούς καρκίνους ενώ στους επιφανειακούς ισχύει το αντίστροφο δηλαδή έχει ογκογενή δράση αν και τονίζεται ότι αυτό πιθανόν να οφείλεται σε διάφορες βιολογικές και κλινικές διαφορές μεταξύ των δύο αυτών ομάδων (Yoshida et al,2008). Για την πρωτεΐνη E2F4 δεν υπάρχουν πολλές αναφορές σχετικά με την επίδραση του στον καρκίνο πάντως σύμφωνα με ορισμένους ερευνητές (Hui-zi Chen et al) έχει ογκογενή δράση.

Όσον αφορά την Κυκλίνη Ε, και εδώ δεν υπάρχουν ξεκάθαρα αποτελέσματα σχετικά με την σχέση της με άλλους παράγοντες. Πιο συγκεκριμένα σύμφωνα με τα αποτελέσματα διάφορων μελετών (T.Kamai et al 2001) η Κυκλίνη Ε έχει αρνητική συσχέτιση με το στάδιο του καρκίνου και τον βαθμό κακοήθειας του και πιο συγκεκριμένα έχει παρατηρηθεί ότι όσο μεγαλώνει το στάδιο του καρκίνου και ο βαθμός κακοήθειας τόσο μειώνεται η ποσότητα της Κυκλίνης Ε. σύμφωνα με άλλη μελέτη (E. Ioachim et al 2004) η Κυκλίνη αυξάνεται στα αρχικά στάδια, κατά την μετάβαση από το στάδιο Ta στο T1, και στην συνέχεια μειώνεται πάλι επιβεβαιώνοντας την ιδέα ότι παίζει ρόλο στα αρχικά στάδια της καρκινογένεσης. Επίσης δεν παρουσιάζεται κάποια συσχέτιση της με τον βαθμό κακοήθειας του καρκίνου. Αλλού (Kazuhide Makiyama et al 2000) η Κυκλίνη Ε έχει θετική σχέση με τον βαθμό κακοήθειας ενώ δεν υπάρχει κάποια σχέση με το στάδιο του καρκίνου. Τέλος, σε άλλη μελέτη (Joseph J. del Pizzo et al 1999) η έκφραση της Κυκλίνης Ε ήταν μικρότερη για καρκίνους με βαθμό κακοήθειας G3 έναντι των άλλων δύο, παρόλα αυτά το αποτέλεσμα της δεν ήταν στατιστικά σημαντικό

Τέλος σχετικά με τον δείκτη κυτταρικού πολλαπλασιασμού Ki 67 όλες οι μελέτες συμφωνούν ότι έχει θετική συσχέτιση τόσο με το στάδιο του καρκίνου όσο και με τον βαθμό κακοήθειας ενώ επιπλέον παρουσιάζεται αρνητική συσχέτιση μεταξύ του δείκτη Ki67 και της Κυκλίνης Ε για τιμές της Κυκλίνης Ε μεγαλύτερες του 30% και θετική για τιμές μικρότερες του 30% (A A Khan et al 2003)

9.3 Σύγκριση αποτελεσμάτων

Συγκρίνοντας τα αποτελέσματα μας με τα αντίστοιχα των άλλων μελετών παρατηρούμε ότι αρχικά συμφωνούν σχετικά με την θετική συσχέτιση που παρουσιάζει ο δείκτης Ki67 με το στάδιο και τον βαθμό κακοήθειας του καρκίνου. Επιπλέον, βλέπουμε ότι σύμφωνα με την δική μας ανάλυση η αρνητική συσχέτιση μεταξύ του δείκτη Ki67 και της Κυκλίνης Ε ισχύει για τιμές της Κυκλίνης Ε μεγαλύτερες το 20% και όχι του 30% όπως παρουσιάζει η προηγούμενη έρευνα ενώ δεν βρήκαμε κάποια θετική συσχέτιση μεταξύ τους ανεξάρτητα της τιμής κατωφλιού. Σχετικά με την πρωτεΐνη E2F1 βλέπουμε ότι αν και δεν ήταν στατιστικά σημαντικό το αποτέλεσμα εντούτοις καθώς αυξάνεται ο βαθμός κακοήθειας του καρκίνου τόσο μειώνονται οι τιμές του οπότε ίσως να επιβεβαιώνεται ο ογκοκατασταλτικός χαρακτήρας του. Σε σχέση με την πρωτεΐνη E2F4 δεν μπορούμε να πούμε τίποτα καθώς δεν έχουμε έστω κάποια ένδειξη που να συντελεί στο ότι πιθανόν την πρωτεΐνη έχει ογκογενή δράση. Τέλος, για την Κυκλίνη Ε φαίνεται να επιβεβαιώνεται ότι είναι κάπως σύνθετο το ζήτημα καθώς σε επίπεδο ενδείξεων βλέπουμε ότι τα αποτελέσματα μας συμφωνούν με τα αντίστοιχα που λένε ότι παίζει ρόλο στα αρχικά στάδια της καρκινογένεσης.

9.4 Επίλογος

Είδαμε ότι τα αποτελέσματα μας συμφωνούν σε μεγάλο βαθμό με άλλες έρευνες που έχουν γίνει πάνω στον συγκεκριμένο τομέα. Τα περισσότερα όμως δεν βρέθηκαν στατιστικά σημαντικά ώστε να είμαστε σίγουροι για αυτά και αποτελούν κυρίως ενδείξεις. Κατά την άποψη μας τα συγκεκριμένα αποτελέσματα μπορούν να χρησιμοποιηθούν ως βάση για μια μελλοντική έρευνα μεγαλύτερης κλίμακας η οποία θα μπορέσει να επιβεβαιώσει τα αποτελέσματα της συγκεκριμένης αλλά και να ελέγξει εάν αυτές οι ενδείξεις που είχαμε αλλά δεν μπορέσαμε να αποδείξουμε στατιστικά ότι ισχύουν έχουν πράγματι βάση.

Παραρτήματα

Παράρτημα Α: Ανάλυση στην R

```
library(foreign)
misdata<-
read.spss("C:/Users/user/Desktop/missingdata.sav",to.data.frame=TRUE)
summary(misdata)
```

Αρχικά θα φορτώσουμε το πακέτο foreign καθώς μέσω της εντολής του read.spss θα φορτώσουμε το σετ των δεδομένων μας. στην συνέχεια μέσω της εντολής **summary(…)** παίρνουμε όλα τα περιγραφικά μέτρα που παρουσιάζουμε στο κεφάλαιο 4.

```
hist(misdata$Cyclin_E,xlab="CyclinE",labels=TRUE,xlim=c(0,90),ylim=c(
0,20),col=c(2:6),main="Histogram of Cyclin E")
```

Με την εντολή **hist(…)** παίρνουμε το ιστόγραμμα συχνοτήτων για τις συνεχείς μας μεταβλητές και στην συγκεκριμένη περίπτωση για την Κυκλίνη Ε. Όμοια γίνεται και για τις άλλες συνεχείς μας μεταβλητές.

misdata$Cyclin_E: Παράμετρος που δηλώνει ότι το ιστόγραμμα αφορά την μεταβλητή Κυκλίνη Ε που βρίσκεται στο σετ δεδομένων misdata.

xlab="CyclinE": Παράμετρος που δηλώνει ότι ο άξονας Χ θα έχει την ονομασία CyclinE

labels=TRUE: Παράμετρος που δηλώνει ότι στο σχήμα θα εμφανίζεται ο αριθμός των παρατηρήσεων που ανήκει σε κάθε κατηγορία.

xlim=c(0,90),ylim=c(0,20): Παράμετροι που δηλώνουν τα όρια των αξόνων.

main="Histogram of Cyclin E": Παράμετρος που δηλώνει το όνομα του σχήματος.

col=c(2:6): Παράμετρος που δηλώνει τα χρώματα που θα έχουν οι ράβδοι που δημιουργούνται. Εδώ έχουμε τέσσερα χρώματα που αντιστοιχούν στους αριθμούς 4-6

```
boxplot(misdata$Cyclin_E,main="Cyclin E",horizontal=T,col=2)
text(locator(1),"outlier")
```

Με την εντολή **boxplot(…)** δημιουργούμε το θηκόγραμμα για την Κυκλίνη Ε και όμοια για τις άλλες μεταβλητές.

horizontal=T,col=2: Παράμετρος που δηλώνει ότι το σχήμα μας θα είναι οριζόντιο αντί για κάθετο και θα έχει χρώμα κόκκινο (αυτό αντιστοιχεί στο νούμερο 2).

text(locator(1),"outlier"): Παράμετρος που δηλώνει ότι στο σημείο που θα επιλέξουμε εμείς (μέσω του **locator(1)**) θα εμφανίσει το κείμενο «outlier».

```
table1<-table(misdata$grade,misdata$risk);fisher.test(table1)
table2<-table(misdate$stage,misdate$grade);
library(vcdExtra);GKgamma(table2)
```

Μέσω της εντολής **table(…)** δημιουργείται ο πίνακας συνάφειας μεταξύ των μεταβλητών grade, stage, ενώ με την εντολή **fisher.test(…)** ελέγχουμε κατά πόσο οι κατηγορικές μας μεταβλητές είναι ανεξάρτητες. Όμοια δουλεύουμε και για τις άλλες 2 περιπτώσεις. Τέλος, αφού φορτώσουμε το πακέτο vcdExtra με την εντολή GKgamma(…) υπολογίζουμε το μέτρο συνάφειας γ των Goodman – Kruskal.

```
risk.low <- table1[,"low"];risk.low
risk.total <- margin.table(table1,1);risk.total
prop.test<-prop.trend.test(risk.low,risk.total);prop.test
r1<-sqrt(prop.test$statistic/(sum(table1)-1));r1
```

Με την μεταβλητή risk.low παίρνουμε την στήλη του πίνακα **table1** που αντιστοιχεί σε **risk=low**. Με την μεταβλητή risk.total παίρνουμε τον περιθώριο πίνακα (**margin.table**) του αρχικού μας πίνακα (**table1**) αθροίζοντας ανά γραμμή (**1**). Με την εντολή **prop.trend.test** κάνουμε τον έλεγχο γραμμικής τάσης παίρνοντας το παρακάτω output.

```
Chi-squared Test for Trend in Proportions
data:  risk.low out of risk.total ,
 using scores: 1 2 3
X-squared = 32.1664, df = 1, p-value = 1.415e-08
```

Η μεταβλητή **r1** είναι ο συντελεστής συσχέτισης των 2 κατηγορικών μεταβλητών

Sqrt(…): Συμβολίζει την τετραγωνική ρίζα της ποσότητας μέσα στην παρένθεση

prop.test$statistic: Είναι η τιμή του στατιστικού τεστ **X-squared** = 32.1664

sum(table1): Είναι ουσιαστικά το σύνολο των παρατηρήσεων μας.

Την ίδια διαδικασία ακολουθούμε και στην περίπτωση του ελέγχου stage-risk.

```
qqnorm(misdata$Ki67);qqline(misdata$Ki67)
shapiro.test(misdata$Ki67)
```

Με την εντολή **qqnorm(…)** παίρνουμε το **qqplot** για τον δείκτη Ki 67 ενώ με την εντολή **qqline(…)** προσθέτουμε την διαγώνια γραμμή. Με την εντολή **Shapiro.test(…)** κάνουμε τον έλεγχο κανονικότητας για τον δείκτη Ki 67. Όμοια δουλεύουμε και για τις άλλες περιπτώσεις.

```
data_male<-subset(misdata,gender=="male",select="Cyclin_E")
data_male<-na.omit(data_male)
std.m<-sd(data_male$Cyclin_E)
data_female<-subset(misdata,gender=="female",select="Cyclin_E")
std.f<-sd(data_female$Cyclin_E)
wilcox.test(data_male$Cyclin_E,data_female$Cyclin_E)
```

Με την μεταβλητή **data_male** ορίζουμε ένα νέο σετ δεδομένων μέσω της εντολής **subset(…)** που για να δημιουργηθεί επιλέξαμε από το αρχικό μας σετ **misdata** τις περιπτώσεις των αντρών (**gender=="male"**) και μας ενδιαφέρει η μεταβλητή της Κυκλίνης Ε (**select="Cyclin_E"**). Μέσω της εντολής **na.omit (…)** διώχνουμε τις παρατηρήσεις με ελλιπείς τιμές, είναι απαραίτητο για να κάνουμε τον έλεγχο μεταξύ των δύο γκρουπ. Η εντολή **sd(…)** μας δίνει την τυπική απόκλιση για την Κυκλίνη Ε για τους άντρες, ενώ όμοια δουλεύουμε και για το **data_female**. Μέσω του **Wilcox.test(…)** κάνουμε τον μη παραμετρικό έλεγχο του wilcoxon για την ισότητα των δύο γκρουπ. Δουλεύοντας με τον ίδιο τρόπο παίρνουμε τα 2 σετ για τον δείκτη Ki 67, αλλάζουμε μόνο το όρισμα select σε **select="Ki67"**. Για να πάρουμε τα σετ των δεδομένων ανάλογα με τις ηλικίες των ασθενών αντί για **gender=="…"** γράφουμε **age>65** ή **age≤65**.

```
Cyclin_E_over20<-ifelse(misdata$Cyclin_E>20,1,0)
Cyclin_E_over30<-ifelse(misdata$Cyclin_E>30,1,0)
age_over65<-ifelse(misdata$age>=65,1,0)
Ki67_over10<-ifelse(misdata$Ki67>10,1,0)
data.new<-
cbind(misdata,Cyclin_E_over20,Cyclin_E_over30,age_over65,Ki67_over10)
```

Με τις 4 πρώτες εντολές ορίζουμε 4 νέες μεταβλητές, μέσω της εντολής **ifelse(…)**, όπου εάν οι αρχικές τιμές είναι πάνω από την τιμή που δηλώνεται στην **ifelse(…)** τότε

η νέα τιμή είναι η μονάδα αλλιώς μηδέν. data.new είναι το νέο σετ δεδομένων εάν στο αρχικό μας σετ (misdata) προσθέσουμε με μορφή στήλης τις νέες μας μεταβλητές μέσω της εντολής **cbind**. Στην συνέχεια πάλι μέσω των εντολών **table(…)** και **fisher.test(…)** κάνουμε τους ελέγχους που μας ενδιαφέρουν μεταξύ των πινάκων συνάφειας στο κεφάλαιο 5.

```
regres.1<-lm(datanew$Cyclin_E~datanew$E2F1+datanew$E2F4)
summary(regres.1);anova(regres.1)
qqnorm(regres.1$residuals);qqline(regres.1$residuals)
shapiro.test(regres.1$residuals)
plot(regres.1,which=1)
library(lawstat);runs.test(regres.1$residuals)
```

Για να τρέξουμε την γραμμική παλινδρόμηση στην R χρησιμοποιούμε την εντολή **lm(…)**. Για να δούμε τους συντελεστές του μοντέλου και αν είναι στατιστικά σημαντικοί χρησιμοποιούμε την εντολή **summary(…)** ενώ τον πίνακα ANOVA τον παίρνουμε μέσω της εντολής **anova(…).** Τρέχοντας την εντολή **summary (…)** παίρνω το παρακάτω output

```
Call:
lm(formula = datanew$Cyclin_E ~ datanew$E2F1 + datanew$E2F4)
Residuals:
   Min     1Q  Median     3Q    Max
-20.671  -8.637  -1.103   8.763  23.790

Coefficients:
                  Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)       28.65955    5.55159   5.162 1.06e-05 ***
datanew$E2F1 -0.06017    0.09029  -0.666    0.510
datanew$E2F4 -0.09095    0.07995  -1.138    0.263
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.44 on 34 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared: 0.04444,    Adjusted R-squared: -0.01177
```

```
F-statistic: 0.7905 on 2 and 34 DF,  p-value: 0.4618
```

Ενώ με την εντολή **anova(…)** έχω

```
Analysis of Variance Table
Response: datanew$Cyclin_E
                Df  Sum Sq  Mean Sq  F value  Pr(>F)
datanew$E2F1    1    37.6    37.587   0.2871  0.5956
datanew$E2F4    1   169.4   169.441   1.2940  0.2633
Residuals      34  4452.0   130.941
```

Ξανά μέσω των εντολών **qqnorm(…)** και **shapiro.test(…)** ελέγχω την κανονικότητα των καταλοίπων. Για να ελέγξουμε την ανεξαρτησία των καταλοίπων γραφικά θα το κάνουμε μέσω της εντολής **plot(regres.1,which=1)** που μας δίνει το διάγραμμα 5.7, το όρισμα which=1 μπαίνει επειδή αν δεν το βάζαμε τότε μέσω της εντολής **plot(regres.1)** θα παίρναμε 4 διαφορετικά σχεδιαγράμματα τα οποία όμως δεν τα χρειαζόμαστε. Στατιστικά την ανεξαρτησία την ελέγχουμε μέσω της εντολής **runs.test(…)** του πακέτου **lawstat**. Με παρόμοιο τρόπο δουλεύουμε σε όλες τις περιπτώσεις που χρησιμοποιούμε την γραμμική παλινδρόμηση.

```
cor.test(misdata$Ki67,misdata$Cyclin_E,method="spearman",alternative
="two.sided")
plot(misdata$Ki67, misdata$Cyclin_E,xlab="Ki 67",ylab="Cyclin E")
fit<-lm(misdata$Ki67~ misdata$Cyclin_E)
abline(fit)
legend(35,46,cex=0.7,c("rho=",-0.15,"p.value=",0.35))
identify(misdata$Ki67,misdata$Cyclin_E)
misdata<-misdata[-c(19,35),]
```

Επόμενο κομμάτι της ανάλυσης μας ήταν η σχέση μεταξύ του δείκτη Ki 67 με την Κυκλίνη Ε. με την εντολή **cor.test(…)** υπολογίζουμε τον συντελεστή συσχέτισης του Spearman μεταξύ του δείκτη Ki 67 και Κυκλίνης Ε.

method="spearman": Παράμετρος που δηλώνει ποιον συντελεστή συσχέτισης θέλουμε να υπολογίσουμε, εμείς επιλέξαμε την μέθοδο Spearman ενώ οι άλλες δύο διαθέσιμες είναι οι **"pearson"** και **"kendall"**.

Με την εντολή **plot(…)** κατασκευάζουμε το διάγραμμα διασποράς των δύο μεταβλητών. Με την εντολή **lm(…)** τρέχουμε την γραμμική παλινδρόμηση μεταξύ των δύο μεταβλητών με σκοπό μέσω της εντολής **abline(…)** να προσαρμόσουμε την ευθεία παλινδρόμησης πάνω στο διάγραμμα διασποράς. Μέσω της εντολής

legend(…) δημιουργείται ένα πλαίσιο κειμένου πάνω στο διάγραμμα όπου τοποθετείται το περιεχόμενο του **c(…)**. Οι τιμές 35,46 δηλώνουν το που θα τοποθετηθεί το πλαίσιο ενώ το **cex=0.7** έχει να κάνει με το μέγεθος του πλαισίου.

Η εντολή **identify(…)** αλλάζει το σύμβολο του ποντικιού από το «βελάκι» σε έναν σταυρό που άμα τον σύρουμε πάνω από τις κουκίδες στο διάγραμμα διασποράς μας και «κλικάρουμε» σε κάποια από αυτές δείχνει σε ποια παρατήρηση αντιστοιχεί και με αυτό τον τρόπο πήραμε το διάγραμμα 5.8.

Τέλος με την εντολή **misdata[-c(19,35),]** διαγράφουμε από το σετ των δεδομένων μας τις δύο αυτές παρατηρήσεις.

Μέσω της εντολής **subset(…)** και ακολουθώντας την ίδια διαδικασία με παραπάνω παίρνουμε τα διαγράμματα διασποράς και τους συντελεστές του Spearman για όλες τις περιπτώσεις της κυκλίνης Ε.

```
data1<-subset(misdata,Cyclin_E>20,select=c(Ki67,Cyclin_E,E2F4,E2F1))
data1<-na.omit(data1)
fit<-lm(data1$Ki67~data1$Cyclin_E+data1$E2F1+data1$E2F4)
step<-stepAIC(fit,direction="both",k=2);summary(step);anova(step)
```

Στο κομμάτι που συγκρίναμε την σχέση μεταξύ του δείκτη Ki 67 και της Κυκλίνης Ε δείξαμε πως μπορεί κάποιος να τρέξει την stepwise παλινδρόμηση ώστε αν έχει πολλές μεταβλητές να μπορέσει να επιλέξει ποιες θα βάλει στο μοντέλο της παλινδρόμησης. Το βασικό κομμάτι των παραπάνω εντολών είναι ότι δεν πρέπει να υπάρχουν ελλιπείς παρατηρήσεις στα δεδομένα μας, οπότε πρέπει να διώξουμε τις παρατηρήσεις αυτές μέσω της εντολής na.omit(…). Στην συνέχεια τρέχουμε το μοντέλο της παλινδρόμησης με όλες τις μεταβλητές μέσα μέσω της εντολής lm(…). Τελευταίο κομμάτι είναι να τρέξουμε την stepwise παλινδρόμηση μέσω της εντολής stepAIC(…) που μας δίνει τον πίνακα 5.25 και με τις εντολές summary(step) και anova(step) παίρνουμε τις αποφάσεις μας σχετικά με την σημαντικότητα του μοντέλου.

direction="both": παράμετρος που δηλώνει ότι θα εκτελεστεί η stepwise μέθοδος, άλλες επιλογές **backward** και **forward**.

k=2: Παράμετρος που δηλώνει ότι η απόφαση για τις μεταβλητές που θα μείνουν στο μοντέλο θα γίνεται με βάση το AIC κριτήριο.

```
x1<-subset(misdata,grade==1,select="Ki67");summary(x1)
sapply(x1,sd);dim(x1)
y1<-subset(misdata,grade==2,select="Ki67");summary(y1)
sapply(y1,sd);dim(y1)
z1<-subset(misdata,grade==3,select="Ki67");summary(z1)
sapply(z1,sd);dim(z1)
a<-c(x1,y1,z1)
g <- factor(rep(1:3, c(11,16,18)),labels = c("G1","G2","G3"))
kruskal.test(a,g)
pairwise.wilcox.test(misdata$Ki67,misdata$grade)
```

Στο 6^{ο} κεφάλαιο κάναμε σύγκριση των παραγόντων Ki 67, Κυκλίνη Ε, E2F1, E2F4 για κάθε επίπεδο των μεταβλητών grade, stage και risk. Στην περίπτωση του

ελέγχου για την ισότητα μεταξύ των διαφόρων επιπέδων του grade ως προς το Ki67 αυτό γίνεται με τις παραπάνω εντολές, ενώ παρόμοιες είναι οι εντολές και στις άλλες περιπτώσεις. Δημιουργήσαμε 3 σετ δεδομένων (x1, y1, z1) για κάθε επίπεδο του grade και μέσω της εντολής **sapply(…)** για κάθε ένα υπολογίσαμε την τυπική απόκλιση, ενώ την εντολή **dim(…)** που δίνει τις διαστάσεις του σετ την χρησιμοποιήσαμε για να πάρουμε το πλήθος των ασθενών ανά κατηγορία. Η μεταβλητή **a** είναι η ένωση των τριών αυτών σετ ενώ η μεταβλητή **g** είναι ένας παράγοντας (factor) με 3 επίπεδα σύμφωνα με τα σετ (x1, y1, z1).

factor(…): Εντολή που ορίζει ότι η μεταβλητή στην οποία αντιστοιχεί είναι παράγοντας.

rep(1:3, c(11,16,18)),labels = c("G1","G2","G3"): Εντολή που σημαίνει επανέλαβε τους αριθμούς 1 μέχρι 3 σύμφωνα με τους αριθμούς στο c(…) και στον 1^{ο} αριθμό αντιστοιχεί η ταμπέλα G1 κ.ο.κ.

Ο Kruskal-Wallis έλεγχος για την ισότητα των γκρουπ γίνεται με την εντολή **kruskal.test(…)** ενώ για να δούμε ποια γκρουπ διαφέρουν μεταξύ τους γίνεται μέσω του **pairwise.wilcox.test(…)**.

```
##ordinal logistic regression##
attach(misdata)
library(rms)
grade<-as.ordered(grade)


ddist <- datadist(Ki67,Cyclin_E)
options(datadist='ddist')
fit <- lrm(grade ~Ki67+Cyclin_E);fit
vif(fit)
summary(fit,Ki67=c(2,3))
deviance(fit)

par(mfrow=c(2,1))
plot.xmean.ordinaly(grade ~Ki67+Cyclin_E, cr=FALSE, topcats=2)
par(mfrow=c(1,1))
```

Σε δεύτερο επίπεδο πραγματοποιήσαμε την διατακτική λογιστική παλινδρόμηση μέσω του πακέτου **rms**, αφού πρώτα θέσαμε την μεταβλητή grade ως διατάξιμη μεταβλητή μέσω της εντολής **as.ordered(…)**. Σημείωση με τις ακριβώς ίδιες εντολές που παρουσιάζονται παρακάτω δουλεύουμε και για την περίπτωση του stage.

Οι επόμενες 2 εντολές (**datadist(…),options(…)**) αποθηκεύουν πληροφορίες σχετικά με τις μεταβλητές που περιλαμβάνουν και τις κατανομές τους, οι

πληροφορίες αυτές μπορούν αργότερα να χρησιμοποιηθούν από το πακέτο ώστε να δείξει την επίδραση των ανεξάρτητων μεταβλητών στο μοντέλο.

Μέσω της εντολής **lrm(…)** τρέχουμε την διατακτική λογιστική παλινδρόμηση και παίρνουμε το παρακάτω output

Logistic Regression Model

lrm(formula = grade ~ Ki67 + Cyclin_E)

Frequencies of Missing Values Due to Each Variable

grade	Ki67	Cyclin_E
0	0	3

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	42	LR chi2	11.38	R2	0.268	C	0.728
1	11	d.f.	2	g	1.279	Dxy	0.456
2	15	Pr(> chi2)	0.0034	gr	3.592	gamma	0.458
3	16			gp	0.192	tau-a	0.308
max |deriv|	1e-07			Brier	0.166		

	Coef	S.E.	Wald Z	Pr(>|Z|)
y>=2	-0.1895	0.8427	-0.22	0.8221
y>=3	-2.0690	0.9109	-2.27	0.0231
Ki67	0.0971	0.0330	2.94	0.0033
Cyclin_E	-0.0076	0.0227	-0.33	0.7387

Στο κάτω μέρος του output έχουμε τους συντελεστές του μοντέλου, ενώ στο άνω μέρος αυτά που αξίζει να κρατήσουμε είναι το **Pr(> chi2)=0.0034** που είναι ο έλεγχος πιθανοφάνειας για την σημαντικότητα του μοντέλου, ο δείκτης **c= 0.728** που εκφράζει την area under curve (AUC) που είναι ένα μέτρο που εκφράζει την προβλεπτική ικανότητα του μοντέλου και ίσως το R2=0.268 που είναι το pseudo R^2.

Με την εντολή **vif(…)** ελέγχουμε εάν υπάρχει πρόβλημα πολυσυγγραμικότητας μεταξύ των μεταβλητών και με την εντολή **deviance(…)** παίρνουμε την απόκλιση του μοντέλου μας όπως και την απόκλιση του μοντέλου χωρίς μεταβλητές.

Η εντολή summary(…) μας δίνει τα odds ratio για τις ανεξάρτητες μεταβλητές μας, το όρισμα **Ki67=c(2,3)** μας δίνει το odds ratio για το Ki 67 κατά μία μονάδα και είναι ουσιαστικά το exp(0.0971). ενδεικτικό είναι το output που παίρνουμε.

Effects	Response : grade						
Factor	Low	High	Diff.	Effect	S.E.	Lower 0.95	Upper 0.95
Ki67	2	3	1	0.10	0.03	0.03	0.16
Odds Ratio	2	3	1	1.10	NA	1.03	1.18
Cyclin_E	15	32	17	-0.13	0.39	-0.89	0.63
Odds Ratio	15	32	17	0.88	NA	0.41	1.88

Παρατηρούμε ότι για τον δείκτη Ki 67 που θέσαμε ως όριο το (2,3) η στήλη Diff.=1 ενώ για την κυκλίνη Ε παίρνει την διαφορά της ελάχιστης έναντι της μέγιστης τιμής.

```
par(mfrow=c(1,2))
plot.xmean.ordinaly(grade ~Ki67+Cyclin_E)
par(mfrow=c(1,1))
```

Μια υπόθεση που πρέπει να ελέγξουμε είναι αυτή των αναλογικών odds την οποία διαγραμματικά την ελέγχουμε μέσω της εντολής **plot.xmean.ordinaly(…)** που μας δίνει το διάγραμμα 6.1. Οι εντολές par(mfrow=c(1,2)) και par(mfrow=c(1,1)) αυτό που κάνουν είναι η μεν $1^{η}$ να χωρίζει κάθετα την οθόνη που εμφανίζονται τα σχήματα ώστε να παίρνουμε 2 το ένα δίπλα στο άλλο και η $2^{η}$ το επαναφέρει στην αρχική κατάσταση ώστε να εμφανίζει ένα σχήμα.

```
prob1<-predict(fit, type="fitted")[,1]
prob2<-predict(fit, type="fitted")[,2]
plot(prob1,Ki67,type="p",xlab="accumulated probabilities"
,xlim=c(0.1,1));points(prob2,Ki67,col=2)
legend(0.74,12,c("G2","G3"),lty=c(1,1),col=c(1,2))
```

Για να πάρουμε το διάγραμμα 6.2 ακολουθούμε τις παραπάνω εντολές όπου prob1 και prob2 είναι οι αθροιστικές πιθανότητες και plot το διάγραμμα. Με την εντολή **points(…)** προσθέτουμε κουκκίδες στο υπάρχον διάγραμμα.

predict(fit, type="fitted")[,1]: Εντολή που προβλέπει τις αθροιστικές πιθανότητες, βάζοντας το όρισμα [,1] σημαίνει ότι παίρνουμε την $1^{η}$ στήλη που αντιστοιχεί στην πιθανότητα P(grade≥G2) ενώ το όρισμα [,2] σημαίνει ότι παίρνουμε την $2^{η}$ στήλη που αντιστοιχεί στην πιθανότητα P(grade≥G3).

```
##1° τρόπος##
library(VGAM)
fit1 <- vglm(grade ~Ki67+Cyclin_E,family=cumulative(parallel=T))
fit2 <- vglm(grade
~Ki67+Cyclin_E,family=cumulative(parallel=F),maxit=50)

pchisq(deviance(fit1)-deviance(fit2),df=df.residual(fit1)-
df.residual(fit2),lower.tail=FALSE)

##2° τρόπος##
Library(nnet)
unordered <- multinom(grade ~Ki67+Cyclin_E)
summary(unordered)

pchisq(fit$deviance[2]-unordered$deviance,df=unordered$edf-
fit$stats[4],lower.tail=FALSE)
```

Για να ελέγξουμε στατιστικά αν ισχύει η υπόθεση των αναλογικών odds χρησιμοποιήσαμε δύο τρόπους. Σχετικά με τον πρώτο, τρέξαμε δύο φορές την διατακτική παλινδρόμηση μέσω της εντολής **vglm(…)** που βρίσκεται στο πακέτο **VGAM**, όπου την πρώτη φορά υποθέσαμε ότι ισχύει η υπόθεση των αναλογικών odds και την άλλη όχι μέσω του ορίσματος **family=cumulative(parallel=T ή F)**. Στην συνέχεια κάναμε έλεγχο λόγου πιθανοφανειών μέσω της εντολής **pchisq(…)** όπου μας δίνει την πιθανότητα η τιμή που παίρνουμε από τις αποκλίσεις των δύο μοντέλων **(deviance(fit1)-deviance(fit2))** να προέρχονται από την κατανομή X^2 με βαθμούς ελευθερίας ίσους με την διαφορά των βαθμών ελευθερίας των μοντέλων **(df=df.residual(fit1)-df.residual(fit2))**. Όμοια είναι η λογική και στον δεύτερο τρόπο μόνο που τώρα ελέγχουμε ένα μοντέλο που ισχύει η υπόθεση των αναλογικών odds με ένα μοντέλο πολλαπλής λογιστικής παλινδρόμησης (όχι διατακτικής) μέσω της εντολής **multinom(…)** του πακέτου **nnet**.

```
d1<-data.frame(Ki67=seq(2,44,5),Cyclin_E=mean(Cyclin_E,na.rm = TRUE))
predict(fit,d1,type="fitted.ind"); predict(fit, d1, type="fitted")
library(effects);library(MASS)
fit3 <- polr(grade~Ki67+Cyclin_E)
plot(effect("Ki67", fit3))
```

Για να πάρουμε τους πίνακες 6.10, 6.11 αλλά και το διάγραμμα 6.3 δουλεύουμε ως εξής. Δημιουργούμε ένα νέο σετ δεδομένων μέσω της εντολής data.frame(…) που περιλαμβάνει τα παρακάτω.

Ki67=seq(2,44,5): Παράμετρος που δηλώνει ότι ο δείκτης Ki 67 θα παίρνει τιμές από το 2 έως το 44 με βήμα ανά 5, δηλαδή 2-7-12 κ.ο.κ.

Cyclin_E=mean(Cyclin_E,na.rm = TRUE): Η κυκλίνη είναι σταθερή και ίση με την μέση της τιμή που προκύπτει αφού αφαιρέσουμε τις ελλιπείς τιμές ώστε να μπορέσει να υπολογιστεί.

Αφού γίνει αυτό μέσω των εντολής **predict(…)** και αλλάζοντας το όρισμα type παίρνω τους δύο πίνακες. Σημείωση, το όρισμα **type="fitted.ind"** μου δίνει τις πιθανότητες P(grade=i), i=1,2,3 ενώ το όρισμα **type="fitted"** τις πιθανότητες P(grade≥i), i=2,3

Τέλος, για να πάρουμε το διάγραμμα 6.3 αρχικά θα φορτώσουμε τα πακέτα effects και MASS. Στην συνέχεια, θα τρέξουμε μια παλινδρόμηση μέσω της εντολής polr(…) και μέσω της εντολής plot(…) παίρνουμε το διάγραμμα.

effect("Ki67", fit3): Παράμετρος που δηλώνει ότι μας ενδιαφέρει η επίδραση του Ki 67 σύμφωνα με το μοντέλο fit3.

```
risk<-as.ordered(risk)
##logistic regression##
mylogit<-glm(risk~Ki67,family=binomial(link="logit"),
na.action=na.pass)
summary(mylogit);anova(mylogit)

dev.<-mylogit$null.deviance - mylogit$deviance
df<-mylogit$df.null - mylogit$df.residual
pchisq(dev.,df,lower.tail=FALSE)

mylogit2<-glm(risk~Cyclin_E+Ki67,family=binomial(link="logit"),
na.action=na.omit)

dev2.<-mylogit$deviance - mylogit2$deviance
df2<-mylogit$df.residual - mylogit2$df.residual
pchisq(dev2.,df2,lower.tail=FALSE)

exp(mylogit$coefficients[2])
confint(mylogit)
```

```
exp(confint(mylogit))
```

Το τελευταίο κομμάτι το οποίο θα μας απασχολήσει σχετικά με το 6^{ο} κεφάλαιο είναι η λογιστική παλινδρόμηση που εφαρμόσαμε σχετικά με την μεταβλητή risk. Η εντολή που πραγματοποιεί την λογιστική παλινδρόμηση είναι η **glm(…)** θέτοντας ως οικογένεια (family) το όρισμα **binomial(link="logit")**. Ξανά τους συντελεστές του μοντέλου κτλ. θα τα πάρουμε μέσω των εντολών **summary(…)** και **anova(…)**.

na.action=na.pass: Παράμετρος που δηλώνει ότι οι ελλιπής παρατηρήσεις θα παραλείπονται από την ανάλυση.

Επόμενο στάδιο είναι ο έλεγχος σημαντικότητας του μοντέλου αλλά και το αν θα πρέπει να συμπεριληφθεί η Κυκλίνη Ε στο μοντέλο σύμφωνα με το **mylogit2**. Οι 2 αυτοί έλεγχοι θα γίνουν σύμφωνα με τον έλεγχο λόγου πιθανοφανειών όπως και πριν.

Σχετικά με τις τρεις επόμενες εντολές, η πρώτη μου δίνει το e^{β} που είναι ουσιαστικά το odds ratio για τον συντελεστή του Ki 67. Οι άλλες δύο δίνουν το 95% δ.ε για τον συντελεστή και το odds ratio του Ki 67 αντίστοιχα.

```
pi<-predict.glm(mylogit, type="response",
newdata=data.frame(Ki67=quantile(Ki67)))
##Κλίση ευθείας##
incline<-mylogit$coefficients[2]*pi*(1-pi)
EL50<--mylogit$coefficients[1]/mylogit$coefficients[2]
mylogit$coefficients[2]/4
```

Η μεταβλητή pi περιλαμβάνει τις πιθανότητες ο ασθενής να ανήκει στην ομάδα υψηλού κινδύνου για τιμές του Ki 67 ίσες με τα ποσοστιαία σημεία μέσω της εντολής **predict.glm(…)** σύμφωνα με το μοντέλο **mylogit**.

type="response": Παράμετρος που δηλώνει ότι θα υπολογιστούν οι εκτιμηθείσες πιθανότητες.

newdata=data.frame(Ki67=quantile(Ki67)): Παράμετρος που δηλώνει ότι θα υπολογιστούν οι πιθανότητες για τιμές του Ki 67 που ισούνται με τα ποσοστημόρια (quantile) και βρίσκονται σε ένα σετ δεδομένων (dataframe).

Οι επόμενες εντολές μας βοηθούν να εκτιμήσουμε κάποια στοιχεία σχετικά με το μοντέλο της λογιστικής παλινδρόμησης όπως αναφέρονται στην ενότητα 3.4.1. η μεταβλητή incline είναι η κλίση της καμπύλης παλινδρόμησης για τις τιμές pi που υπολογίσαμε προηγουμένως. Η μεταβλητή EL50 είναι το διάμεσο επίπεδο

αποτελεσματικότητας (median effective level), ενώ η 3η εντολή μας δίνει την κλίση στο σημείο αυτό. Συγκεντρωτικά όλα τα παραπάνω βρίσκονται στον πίνακα 6.27.

mylogit$coefficients: Με την συγκεκριμένη εντολή παίρνουμε τους συντελεστές του μοντέλου της λογιστικής παλινδρόμησης. Βάζοντας τον δείκτη [1] παίρνουμε τον συντελεστή του σταθερού όρου ενώ ο δείκτης [2] αντιστοιχεί στον συντελεστή του Ki 67.

```
p <- predict.glm(mylogit, type="response")
data<-cbind(Ki67,p);data<-data[order(Ki67),]
plot(data,type="l")
```

Με τις παραπάνω εντολές παίρνουμε το διάγραμμα 6.7, αυτό που πρέπει να τονίσουμε είναι ότι τώρα στην εντολή **predict.glm(…)** δεν περιέχει το όρισμα **newdata** που σημαίνει ότι θα υπολογιστούν οι πιθανότητες για τα δεδομένα που έχουμε στο αρχικό σετ. Στην συνέχεια θα δημιουργήσουμε ένα νέο σετ δεδομένων που θα περιλαμβάνει την ένωση κατά στήλες (**cbind**) του Ki 67 με τις αντίστοιχες πιθανότητες και το οποίο θα το διατάξουμε ως προς το Ki 67 μέσω της εντολής **order(Ki 67)** ώστε να πάρουμε το σωστό σχήμα με την εντολή **plot(…)**.

type="l": Παράμετρος που δηλώνει ότι το σχήμα μας θα έχει την μορφή γραμμής.

```
a<-mylogit$coefficients[1]
b<-mylogit$coefficients[2]
vec2<-
c(exp(a+b*quantile(Ki67)[3])/(1+exp(a+b*quantile(Ki67)[3]))^2,quantil
e(Ki67)[3]*exp(a+b*quantile(Ki67)[3])/(1+exp(a+b*quantile(Ki67)[3]))^
2)
as.var<-t(vec2)%*%summary(mylogit)$cov.unscaled %*%vec2
SE2<-sqrt(as.var)
ci2<-c(pi[3]-1.96*SE2,pi[3]+1.96*SE2)
```

Οι παραπάνω εντολές υπολογίζουν τα διαστήματα εμπιστοσύνης για την πιθανότητα επιτυχίας p(x). Εδώ παρουσιάζεται το 95% δ.ε για την διάμεσο του Ki 67 **(quantile(Ki67)[3])** αλλά όμοια γίνεται και για άλλες τιμές. Σχετικά με την μεθοδολογία δεν θα επεκταθούμε καθώς είναι η εφαρμογή της μεθόδου Δέλτα που ξεφεύγει από τα όρια της εργασίας μας αναλυτική περιγραφή της υπάρχει όμως στις σημειώσεις του μαθήματος «Ανάλυση διακριτών δεδομένων» (Ηλιόπουλος 2011). Συνοπτικά η μεταβλητή **vec2** είναι ένα διάνυσμα 2 στοιχείων, η **as.var** είναι η ασυμπτωτική διακύμανση και προκύπτει ως το εσωτερικό γινόμενο (συμβ. %*%)

μεταξύ του ανάστροφου διανύσματος **vec2** (**t(vec2)**) με τον πίνακα διακυμάνσης συνδυακύμανσης των συντελεστών του μοντέλου και με το διάνυσμα **vec2**. Η μεταβλητή **SE2** είναι το τυπικό σφάλμα και το **ci2** είναι το διάνυσμα με το 95% δ.ε.

```
library(ROCR)
data<-data.frame(p,risk)
pred <- prediction(data$p,data$risk)
perf1 <- performance(pred, "auc")
perf2 <- performance(pred, "sens", "spec")
sum<-perf2@y.values[[1]]+perf2@x.values[[1]]
perf2@alpha.values[[1]][which.max(sum)]
perf3 <- performance(pred,"tpr","fpr")
plot(perf3);text(0.10,0.95,label=0.48);arrows(0.12,0.90,0.12,0.77)
```

Στα πλαίσια της λογιστικής παλινδρόμησης κατασκευάσαμε την καμπύλη ROC ώστε να βρούμε το κατάλληλο σημείο αποκοπής. Αρχικά πρέπει να φορτώσουμε το πακέτο ROCR και στην συνέχεια για το σετ δεδομένων data εκτελέσαμε την εντολή prediction η οποία μας δίνει το παρακάτω output. Εκεί παίρνουμε αρκετά στοιχεία από τα οποία αυτά που μας ενδιαφέρουν είναι τα: "cutoffs" που μας δίνει τα διάφορα σημεία αποκοπής. "fp" που μας δείχνει το πλήθος των εσφαλμένα θετικών ταξινομήσεων, δηλαδή το μοντέλο ταξινομεί τον ασθενή στην κατηγορία risk=high ενώ κανονικά ανήκει στην risk=low. "tp" που μας δείχνει το πλήθος των σωστά θετικών ταξινομήσεων, δηλαδή το μοντέλο ταξινομεί σωστά τον ασθενή στην κατηγορία risk=high. "tn" και "fn" που είναι τα αντίστοιχα με τα προηγούμενα και αφορούν την ταξινόμηση του ασθενή στην κατηγορία risk=low και όλα αυτά για κάθε σημείο αποκοπής ξεχωριστά

```
.....
Slot "cutoffs":
[[1]]
 [1]       Inf 0.97324184 0.95391083 0.93980548 0.93131977 0.91094601
 [7] 0.87016180 0.83486265 0.81450300 0.79225783 0.71417299 0.68455576
[13] 0.65335919 0.55255200 0.48227910 0.44723135 0.41270025 0.37900726
[19] 0.28564816 0.25777506 0.23173945 0.20759797 0.18536463 0.16501657
[25] 0.14650049 0.12973922 0.11463806 0.10109043 0.08898292
```

```
Slot "fp":
[[1]]
 [1]  0  0  0  0  0  0  0  0  1  1  1  1  3  3  3  5  5  9 10 12 15 16 17 18 19
[26] 20 23 24 25
Slot "tp":
[[1]]
 [1]  0  1  2  3  4  5  6  7  7  8  9 10 12 14 15 15 16 16 17 17 17 17 17 17 18
[26] 20 20 20 20
Slot "tn":
[[1]]
 [1] 25 25 25 25 25 25 25 25 24 24 24 24 22 22 22 20 20 16 15 13 10  9  8  7  6
[26]  5  2  1  0
Slot "fn":
[[1]]
 [1] 20 19 18 17 16 15 14 13 13 12 11 10  8  6  5  5  4  4  3  3  3  3  3  3  2
[26]  0  0  0  0
```

Η εντολή **performance(pred, "auc")** μας δίνει την τιμή για την area under curve, ενώ η εντολή **performance(pred, "sens", "spec")** μας δίνει τις διάφορες τιμές για την ευαισθησία (sens) και την ειδικότητα (spec) για διάφορες τιμές του σημείου αποκοπής.

```
An object of class "performance"
Slot "x.name":
[1] "Specificity"
Slot "y.name":
[1] "Sensitivity"
Slot "alpha.name":
[1] "Cutoff"
Slot "x.values":
[[1]]
 [1] 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 0.96 0.96 0.96 0.96 0.88 0.88 0.88
[16] 0.80 0.80 0.64 0.60 0.52 0.40 0.36 0.32 0.28 0.24 0.20 0.08 0.04 0.00
Slot "y.values":
[[1]]
```

```
 [1] 0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.35 0.40 0.45 0.50 0.60 0.70 0.75
[16] 0.75 0.80 0.80 0.85 0.85 0.85 0.85 0.85 0.85 0.90 1.00 1.00 1.00 1.00
Slot "alpha.values":
[[1]]
 [1]       Inf 0.97324184 0.95391083 0.93980548 0.93131977 0.91094601
 [7] 0.87016180 0.83486265 0.81450300 0.79225783 0.71417299 0.68455576
[13] 0.65335919 0.55255200 0.48227910 0.44723135 0.41270025 0.37900726
[19] 0.28564816 0.25777506 0.23173945 0.20759797 0.18536463 0.16501657
[25] 0.14650049 0.12973922 0.11463806 0.10109043 0.08898292
```

Επόμενο βήμα είναι να αθροίσουμε τις τιμές της ευαισθησίας και της ειδικότητας και να βρούμε εκείνο το σημείο που μεγιστοποιεί το άθροισμα τους. Στην μεταβλητή **sum** είναι αποθηκευμένα όλα αυτά τα αθροίσματα και η επόμενη εντολή μας δίνει την τιμή alpha που αντιστοιχεί στο μέγιστο στοιχείο της **sum**.

Το **performance(pred,"tpr","fpr")** δίνει ένα όμοιο με τα προηγούμενα output αυτή την φορά σχετικά με την ευαισθησία και το (1-ειδικότητα) και θα χρησιμοποιηθεί ώστε να πάρουμε μέσω της **plot(…)** την καμπύλη roc.

```
thresh  <- 0.48
pFac <- cut(p, breaks=c(-Inf, thresh, Inf), labels=c(0, 1))
cTab <- table(risk,pFac);addmargins(cTab)
sum(diag(cTab)) / sum(cTab)
```

Το τελευταίο κομμάτι της ανάλυσης μας ήταν η κατασκευή του πίνακα ταξινόμησης. Η μεταβλητή thresh είναι το σημείο αποκοπής που βρήκαμε προηγουμένως. με την εντολή cut(…) παίρνουμε τις εκτιμηθείσες πιθανότητες που είχαμε βρει προηγουμένως (p) και δημιουργούμε μια νέα μεταβλητή την pFac με τον εξής τρόπο, εάν η αρχική πιθανότητα είναι στο διάστημα (-∞, thresh) τότε η νέα τιμή είναι το μηδέν αλλιώς η μονάδα. Μετά δημιουργούμε τον πίνακα ταξινόμησης στον οποίο με την εντολή addmargins(…) προσθέτουμε τα περιθώρια αθροίσματα. Η τελευταία εντολή μας δίνει την πιθανότητα σωστής πρόβλεψης του μοντέλου.

sum(diag(cTab)): Άθροισμα των διαγώνιων στοιχείων του πίνακα

sum(cTab): Άθροισμα όλων των στοιχείων του πίνακα

Παράρτημα Β: Μη παραμετρική παλινδρόμηση στην R

```
misdata1<-subset(misdata,select=c("Ki67","Cyclin_E"))
misdata1<-na.omit(misdata1)
attach(misdata1)
##simple##
library(KernSmooth)
plot(Ki67,Cyclin_E);
h<-dpill(Ki67,Cyclin_E)
fit1 <- locpoly(Ki67,Cyclin_E, bandwidth = h)
fit2 <- locpoly(Ki67,Cyclin_E, bandwidth = h+1)
fit3 <- locpoly(Ki67,Cyclin_E, bandwidth = h-0.5)
fit4 <- locpoly(Ki67,Cyclin_E, bandwidth = h-1.5)
lines(fit1,lwd=2);lines(fit2,col=2,lwd=2);lines(fit3,col=4,lwd=2);lin
es(fit4,col=3,lwd=2)
legend(locator(1),c("span=3.2","span=4.2","span=2.7","span=1.7"),lwd=
c(2,2,2,2),col=c(1,2,4,3))
detach(misdata1)
```

Με την πρώτη εντολή παίρνουμε ένα υποσύνολο από του αρχικού σετ δεδομένων με τις μεταβλητές Ki67 και Cyclin_E από το οποίο αφαιρούμε τις ελλιπείς παρατηρήσεις μέσω της εντολής **na.omit(…)** ενώ με την εντολή **attach(…)** το ορίζουμε και ως σετ αναφοράς, δηλαδή οι επόμενες εντολές αφορούν τις μεταβλητές στο συγκεκριμένο σετ. Στην συνέχεια δημιουργούμε το διάγραμμα διασποράς των μεταβλητών Ki67 και Cyclin_E. Για να τρέξουμε την απλή μη παραμετρική παλινδρόμηση πρέπει να φορτώσουμε το πακέτο **KernSmooth**. Αφού γίνει αυτό στην μεταβλητή **h** ορίζουμε το πλάτος του πλαισίου που θα χρησιμοποιηθεί ώστε να γίνει η παλινδρόμηση μέσω της εντολής **dpill(…)**. Οι μεταβλητές **fit(i)** είναι η εφαρμογή της απλής μη παραμετρικής παλινδρόμησης πάνω στα δεδομένα που μας ενδιαφέρουν μέσω της εντολής **locpoly(…)** αλλάζοντας σε κάθε περίπτωση το πλάτος πλαισίου (**bandwith**) κατά λίγο. Στην συνέχεια τις τέσσερις αυτές καμπύλες τις προσαρμόζουμε πάνω στο διάγραμμα διασποράς που έχουμε μέσω της εντολής **lines(…)** και κάνουμε την ανάλυση μας. Με την εντολή **legend(…)** διαλέγουμε ένα σημείο πάνω στο γράφημα όπου εμφανίζεται το κείμενο που περιγράφεται στο **c(…)** και με αυτόν τον τρόπο παίρνουμε το διάγραμμα 8.1. Με την τελευταία εντολή το σετ misdata1 παύει να είναι το σετ αναφοράς.

```
## multinomial##
misdata2<-subset(misdata,select=c("Cyclin_E","E2F1","E2F4"))
misdata2<-na.omit(misdata2)
attach(misdata2)
mn <- loess(Cyclin_E ~ E2F1+E2F4, span=.5, degree=1)
F1<-seq(min(E2F1), max(E2F1), len=25)
F4<-seq(min(E2F4), max(E2F4), len=25)
newdata <- expand.grid(E2F1=F1,E2F4=F4)
fit.Cyclin_E <- matrix(predict(mn, newdata), 25, 25)
persp(F1,F4, fit.Cyclin_E, theta=45, phi=40,
ticktype="detailed",xlab="E2F1", ylab="E2F4", zlab="Cyclin E",
expand=2/3,shade=0.3)
mn.F1 <- loess(Cyclin_E ~ E2F1, span=.7, degree=1)
mn.F4 <- loess(Cyclin_E ~ E2F4, span=.7, degree=1)
anova(mn.F1, mn);anova(mn.F4, mn)
```

Στο δεύτερο κομμάτι της μη παραμετρικής παλινδρόμησης ελέγξαμε τι γίνεται μεταξύ των E2F1 και E2F4 με την Κυκλίνη Ε. αυτό γίνεται μέσω της εντολής **loess(…)** στην οποία θέσαμε το όνομα **mn**. Οι μεταβλητές **F1, F4** είναι μια σειρά 25 στοιχείων από την ελάχιστη έως την μέγιστη τιμή των E2F1 και E2F4 αντίστοιχα και για τις οποίες δημιουργούμε ένα dataframe με όλους τους συνδυασμούς τους, ενώ η μεταβλητή **fit.Cyclin_E** περιέχει έναν πίνακα με τις εκτιμώμενες τιμές για την Κυκλίνη Ε σύμφωνα με το μοντέλο mn για το σετ δεδομένων **newdata**. Όλα τα παραπάνω θα μας βοηθήσουν ώστε μέσω της εντολής **persp(…)** να πάρουμε το διάγραμμα 8.2.

theta=45, phi=40, expand=2/3,shade=0.3: Παράμετροι που αλλάζουν την εικόνα του σχήματος.

Τέλος οι δύο εντολές **anova(…)** μας βοηθούν να δούμε εάν οι μεταβλητές E2F1 και E2F4 είναι στατιστικά σημαντικές (βλ. πίνακες 8.1 και 8.2).

```
##additive##
library(mgcv)
mna <- gam(Cyclin_E ~ s(E2F1) + s(E2F4))
summary(mna)
fit.Cyclin_E <- matrix(predict(mna, newdata), 25, 25)
persp(F1,F4,fit.Cyclin_E,theta=45,phi=40,ticktype="detailed",xlab="E2
F1", ylab="E2F4", zlab="Cyclin E", expand=2/3,shade=0.3)
detach(misdata2)
```

Μια εναλλακτική προσέγγιση της πολυωνυμικής παλινδρόμησης ώστε να αντιμετωπίσουμε το πρόβλημα των διαστάσεων είναι η προσθετική (additive) παλινδρόμηση. Αυτό γίνεται μέσω της εντολής **gam(…)** από το πακέτο **mgcv**. Μέσω της **summary(…)** παίρνουμε την απόφαση για τους συντελεστές μας (βλ. πίνακα 8.3). Οι υπόλοιπες εντολές είναι όμοιες με πριν και μας βοηθούν στο να πάρουμε το διάγραμμα 8.3.

Παράρτημα Γ: Πολλαπλή εισαγωγή (multiple imputation) στην R

```
misdata[,6:9]<-misdata[,6:9]/100

library(mi)
mdf <- missing_data.frame(misdata, favor_positive = TRUE)
show(mdf)
mdf <- change(mdf, y = "age", what = "imputation_method", to = "pmm")
mdf <-change(mdf, y ="stage", what="class", to="ordered-categorical")
mdf <-change(mdf, y = "proportion", what = "family", to = gaussian())
image(mdf)
imp <- mi(mdf, n.chains = 6, n.iter = 60)
BA <- mi2BUGS(imp)
traceplot(BA,mfrow =
c(3,2),varname=c("mean_age","sd_age","mean_stage","sd_stage","mean_Cy
clin_E","sd_Cyclin_E"))
traceplot(BA,mfrow =
c(2,2),varname=c("mean_E2F1","sd_E2F1","mean_E2F4","sd_E2F4"))
plot(imp,"E2F4")
analysis <- pool(Cyclin_E~E2F1+E2F4 ,data = imp)
display(analysis);summary(analysis)
```

Πρώτο βήμα μας είναι οι στήλες που έχουν τις τιμές των δεικτών τις διαιρούμε με το 100 ώστε το πακέτο mi που θα χρησιμοποιήσουμε να καταλάβει ότι πρόκειται για ποσοστά. Η μεταβλητή **mdf** παρουσιάζει έναν πίνακα με όλες τις μεταβλητές μας όπου βλέπουμε ποιες έχουν ελλιπείς τιμές, τον τύπο των μεταβλητών μας καθώς και με ποιον τρόπο θα υπολογίσει τις ελλιπείς τιμές. Με την εντολή **change(…)** μπορούμε να κάνουμε διάφορες τροποποιήσεις στον πίνακα **mdf**, για παράδειγμα διαβάζοντας τα δεδομένα μας το πρόγραμμα δεν αντιλαμβάνεται ότι η μεταβλητή stage είναι διατάξιμη οπότε πρέπει να το διορθώσουμε ή για την μεταβλητή ηλικία θέλουμε να παίρνει ακέραιες τιμές ενώ για τα ποσοστά που είναι συνεχείς μεταβλητές θέλουμε να ορίσουμε ότι προέρχονται από την κανονική κατανομή που είναι η συνηθέστερη για τέτοιου είδους μεταβλητές. Το τελικό output είναι το παρακάτω

Object of class missing_data.frame with 45 observations on 9 variables

There are 9 missing data patterns

Append '@patterns' to this missing_data.frame to access the corresponding pattern for every observation or perhaps use table()

	type	missing	method	model
gender	binary	0	<NA>	<NA>
age	positive-continuous	2	pmm	linear
grade	ordered-categorical	0	<NA>	<NA>
stage	ordered-categorical	10	ppd	ologit
risk	binary	0	<NA>	<NA>
Ki67	proportion	0	<NA>	<NA>
Cyclin_E	proportion	3	ppd	linear
E2F1	proportion	5	ppd	linear
E2F4	proportion	3	ppd	linear

	family	link	transformation
gender	<NA>	<NA>	<NA>
age	gaussian	identity	log
grade	<NA>	<NA>	<NA>
stage	multinomial	logit	<NA>
risk	<NA>	<NA>	<NA>
Ki67	<NA>	<NA>	qnorm
Cyclin_E	gaussian	identity	qnorm
E2F1	gaussian	identity	qnorm
E2F4	gaussian	identity	qnorm

Με την εντολή **image(…)** παίρνουμε το Διάγραμμα 7.1 που βλέπουμε το μέγεθος των ελλιπών τιμών ανά μεταβλητή. Με την εντολή **mi(…)** δημιουργούμε τα πλήρη σετ δεδομένων και με την **mi2BUGS(…)** ελέγχουμε εάν η διαδικασία έχει συγκλίνει. Την σύγκληση της μεθόδου την βλέπουμε και διαγραμματικά μέσω της εντολής **traceplot(…)** από όπου παίρνουμε το Διάγραμμα 7.2. Μέσω της εντολής **plot(…)** παίρνουμε το διάγραμμα 7.3 όπου έχουμε τα έξι σχήματα για κάθε μεταβλητή που μας επιτρέπουν να δούμε κατά πόσο οι τιμές που πήραμε είναι λογικές. Τέλος με την

εντολή **pool(…)** κάνουμε την ανάλυση που μας ενδιαφέρει σύμφωνα με τα πλήρη δεδομένα, τα αποτελέσματα της οποίας τα βλέπουμε μέσω των εντολών **display(…)** ή **summary(…)**. Στο παραπάνω πλαίσιο οι εντολές αφορούν τον έλεγχο για την σχέση μεταξύ της Κυκλίνης Ε και των παραγόντων E2F1 και E2F4 (βλ. πίνακας 7.1) ενώ όμοια πραγματοποιούμε και τα υπόλοιπα είδη των παλινδρομήσεων.

Βιβλιογραφία

Ηλεκτρονική (τελευταία ενημέρωση (6/6/2012)

1) http://en.wikipedia.org/wiki/Bladder_cancer

2) http://www.emedicinehealth.com/bladder_cancer/article_em.htm

3) http://www.cancer.gov/cancertopics/types/bladder

4) http://info.cancerresearchuk.org/cancerstats/keyfacts/bladder-cancer/

5) http://info.cancerresearchuk.org/cancerstats/types/bladder/

6) http://en.wikipedia.org/wiki/Ki-67_(protein)

7)http://www.andrologia.gr/Templates/ArticlesContinuous.asp?C=OurodoxouKystews&OpenArticle=84

8)http://el.wikipedia.org/wiki/%CE%9F%CF%85%CF%81%CE%BF%CF%80%CE%BF%CE%B9%CE%B7%CF%84%CE%B9%CE%BA%CF%8C_%CF%83%CF%8D%CF%83%CF%84%CE%B7%CE%BC%CE%B1

9)http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1334&context=etd&sei-redir=1&referer=http%3A%2F%2Fwww.google.gr%2Furl%3Fsa%3Dt%26rct%3Dj%26q%3Drebecca%2Bkinkade%2Bbladder%2Bcancer%26source%3Dweb%26cd%3D2%26ved%3D0CCAQFjAB%26url%3Dhttp%253A%252F%252Fscholarcommons.usf.edu%252Fcgi%252Fviewcontent.cgi%253Farticle%253D1334%2526context%253Detd%26ei%3DumbwTtSoEoOE8gOnluWkAQ%26usg%3DAFQjCNEM3MGhcNYO5MP4NYYUYz-l-xBFBg#search=%22rebecca%20kinkade%20bladder%20cancer%22

10) http://www.nature.com/nrc/journal/v9/n11/fig_tab/nrc2696_T3.html

11) http://www.onco.gr/documents/RigasAthanasiou.pdf

12)http://www.powershow.com/view/3e500-MjkzZ/Filling_Holes_in_Your_Data_Multiple_Imputation_in_Education_Research_flash_ppt_presentation

13) http://carcin.oxfordjournals.org/content/25/3/375.short

14) www.e-urology.gr

15)http://web.campbell.edu/faculty/garrett/PHAR%20408/cell%20cycle%20checkpoints.jpg

16) http://www.andrologia.gr

17) http://www.iatronet.gr/article.asp?art_id=247

Ελληνική

1) Μ. Κούτρας (2010), *Ανάλυση παλινδρόμησης και ανάλυση διακύμανσης,* Σημειώσεις ΠΜΣ «εφαρμοσμένη στατιστική».

2) Μ. Κούτρας, Χ. Ευαγγελάρας (2010), *Πολλαπλή Παλινδρόμηση*, Σημειώσεις ΠΜΣ «εφαρμοσμένη στατιστική».

3) Μ. Κούτρας, Χ. Ευαγγελάρας (2010), *ειδικά θέματα στην πολλαπλή παλινδρόμηση*, Σημειώσεις ΠΜΣ «εφαρμοσμένη στατιστική».

4) Γ. Ηλιόπουλος (2011), *Ανάλυση διακριτών δεδομένων*, Σημειώσεις ΠΜΣ «εφαρμοσμένη στατιστική».

5) Κ. Πολίτης (2011), *Γενικευμένα γραμμικά μοντέλα,* Σημειώσεις ΠΜΣ «εφαρμοσμένη στατιστική».

6) Μ. Κατέρη (2011), *Βιοστατιστική και στατιστικές μέθοδοι στην επιδημιολογία*, Σημειώσεις ΠΜΣ «εφαρμοσμένη στατιστική».

Ξένη

1) Alan Agresti (2007), *An introduction to categorical data analysis 2nd edition*, Wiley

2) Andrew Gelman, Jennifer Hill (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press

3) James R. Carpenter (2009), *Statistical modeling with missing data using multiple imputation* (Lecture notes), London School of Hygiene & Tropical Medicine

4) Yulei He, *Missing Data Analysis Using Multiple Imputation : Getting to the Heart of the Matter*, Circ Cardiovasc Qual Outcomes 2010,3,98-105

5) Stef van Buuren, Multiple imputation of discrete and continuous data by fully conditional specification, *Statistical Methods in Medical Research* 2007, **16**: 219–242

6) Rubin DB, (1987), *Multiple Imputation for Nonresponse in Surveys,*Wiley

7) Rubin DB. *Multiple imputation after 18_ years (with discussion)*. J Am Stat Assoc. 1996;91:473– 489

8) E. Ioachim, M. Michael, N.E. Stavropoulos, E. Kitsiou, K. Hastazeris, M. Salmas, S. Stefanaki, N.J. Agnantis, *Expression Patterns of Cyclins D1, E and Cyclin-Dependent Kinase Inhibitors p21(Waf1/Cip1) and p27(Kip1) in Urothelial Carcinoma: Correlation with Other Cell-Cycle-Related Proteins (Rb, p53, Ki-67 and PCNA) and Clinicopathological Features*, Urol Int 2004;73:65–73

9) Kenichi Yoshida, (2008), *Control of Cellular Physiology by E2F Transcription Factors*, Research Signpost

10) A A Khan, P D Abel, K S Chaudhary, Z Gulzar, G W H Stamp, E-N Lalani, *Inverse correlation between high level expression of cyclin E and proliferation index in transitional cell carcinoma of the bladder*, J Clin Pathol: Mol Pathol 2003;56:353–361

11) T Kamai, K Takagi, H Asami, Y Ito, H Oshima and K-I Yoshida, *Decreasing of p27Kip1 and cyclin E protein levels is associated with progression from superficial into invasive bladder cancer*, British Journal of Cancer (2001) 84(9), 1242–1251

12) Joseph J. Del Pizzo, Andrew Borkowski, Stephen C. Jacobs and Natasha Kyprianou, *Loss of Cell Cycle Regulators p27Kip1 and Cyclin E in Transitional Cell Carcinoma of the Bladder Correlates with Tumor Grade and Patient Survival*, *American Journal of Pathology,* Vol. 155, No. 4, October 1999

13) Kazuhide Makiyama, Mitsunobu Masuda, Yasuo Takano, Masayuki Iki, Tomoyuki Asakura, Yutaka Suwa, Sumio Noguchi, Masahiko Hosaka, *Cyclin E overexpression in transitional cell carcinoma of the bladder*, Cancer Letters 151 (2000) ,193-198

14) John Fox, (2000), *Nonparametric simple regression smoothing scatterplots*, Sage publications

15) John Fox, (2000), *Multiple and generalized nonparametric regression*, Sage publications

16) Julian J. Faraway, (2006), *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Hall/CRC