



Στατιστική Ανάλυση Δεδομένων των προπτυχιακών
φοιτητών του τμήματος Μαθηματικών Αθήνας στο διάστημα
των ετών 2000 έως 2015

ΙΩΑΝΝΗΣ ΓΙΑΓΚΙΝΗΣ



Διπλωματική Εργασία στο ΠΜΣ
στην Στατιστική και Επιχειρησιακή Έρευνα

ΑΘΗΝΑ 2017

Πανεπιστήμιο Αθηνών

Τμήμα Μαθηματικών

Αθήνα 2017

Η παρούσα Διπλωματική Εργασία εκπονήθηκε
στα πλαίσια των σπουδών για την απόκτηση του
Μεταπτυχιακού Διπλώματος Ειδίκευσης στην
Στατιστική και Επιχειρησιακή Έρευνα
που απονέμει το τμήμα Μαθηματικών του
Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών

Εγκρίθηκε στις από Εξεταστική Επιτροπή αποτελούμενη από τους:

Ονοματεπώνυμο	Βαθμίδα	Υπογραφή
Λουκία Μελικοτσιδου	Επίκουρη καθηγήτρια	
Απόστολος Μπουρνέτας	Καθηγητής	
Φώτης Σιάννης(επιβλέπων)	Επίκουρος Καθηγητής	

Πίνακας Περιεχομένων

1.Εισαγωγή.....	6
2.Θεωρία	7
2.1 Πολλαπλό Γραμμικό Μοντέλο.....	7
2.2 Το κλασσικό Πολλαπλό Γραμμικό Μοντέλο.....	7
2.2.1 Βασικές Υποθέσεις του Γραμμικού Μοντέλου.....	8
2.3 Εκτίμηση των Παραμέτρων του Γραμμικού Μοντέλου	9
2.3.1 Εκτίμηση του β	9
2.4 Γενικευμένα Γραμμικά Μοντέλα -GLM.....	13
2.4.1. Εισαγωγή	13
2.4.2 Παραδείγματα-Ειδικές περιπτώσεις	14
Παράδειγμα 1.....	14
Παράδειγμα 2.....	14
Παράδειγμα 3.....	14
2.4.3 Κλασσική Εκτίμηση για Γενικευμένα Γραμμικά Μοντέλα	15
2.4.5 GLM, Λογιστική και Poisson παλινδρόμηση.	17
Γ.Γ.Μ. για Διωνυμικά Δεδομένα.....	17
Το Μοντέλο της Λογιστικής Παλινδρόμησης	17
Poisson Παλινδρόμηση	18
2.4.6.GLMs	19
GLMM (Ορισμός).....	19
Η παράμετρος που καθορίζει τη διασπορά.....	20
2.4.7 Η πιθανοφάνεια και η ψευδοπιθανοφάνεια	21
Η συνάρτηση πιθανοφάνειας για τυχαίες επιδράσεις και για randomcoefficientmodels	21
Η ψευδο-πιθανοφάνεια για τα covariance pattern models	22
Η γενική μορφή συνάρτησης της ψευδο-πιθανοφάνειας για όλα τα GLMMs.....	23
2.5 Ανάλυση Επιβίωσης	24
2.5.1 Εισαγωγή	24
Βασικές έννοιες	24
2.5.2 Μοντελοποίηση με χρήση του αναλογικού μοντέλου Cox.....	25
2.6 Χρονοσειρές	27

3. Δεδομένα Μελέτης.....	35
3.1 Εισαγωγή	35
3.2 Δεδομένα εργασίας.....	36
3.2.1 Περιγραφή πρωτογενών –ακατέργαστων δεδομένων-Περιγραφή Δεδομένων.....	36
Περιγραφή πρώτου πακέτου με δεδομένα	36
Περιγραφή δεύτερου πακέτου με δεδομένα.....	42
3.3 Μια πρώτη ανάλυση (περιγραφική) πάνω στην σταδιοδρομία των φοιτητών	43
3.3.1 Μέσος όρος απόκτησης πτυχίου ανά έτος εισαγωγής και πλήθος πτυχιούχων και επι πτυχίω	43
3.3.2 Ποσοστά Πτυχιούχων επί το σύνολο Πτυχιούχοι- Επί Πτυχίω	45
3.3.3 Μέσο εξάμηνο αποφοίτησης φοιτητών ανά έτος εισαγωγής.	46
3.4 Πίνακες επίδοσης στις εξεταστικές.....	47
3.5 Καθυστέρηση απόκτησης πτυχίου λόγω αργοπορίας σε συγκεκριμένα μαθήματα... ..	51
Προσπάθειες μέχρι την επιτυχία στο εκάστοτε μάθημα σε εξάμηνα	52
4. Μοντελοποίηση μέσω περιγραφικών μέτρων χρόνου απόκτησης πτυχίου - βαθμολογίας.....	54
4.1 Πρώτο Ερώτημα Έρευνας: Παράγοντες καθυστέρησης χρόνου απόκτησης πτυχίου ..	54
Πρώτο ερώτημα έρευνας	54
4.1.1 Περιγραφή μεταβλητών.....	55
Μέτρα θέσης χρόνου καθυστέρησης πτυχίου	55
Μέτρα θέσης χρόνου καθυστέρησης πτυχίου όσον αφορά το Βαθμό	55
Μέτρα θέσης χρόνου καθυστέρησης πτυχίου όσον αφορά την Κατεύθυνση	56
Μέτρα θέσης χρόνου καθυστέρησης πτυχίου όσον αφορά το Sex.....	56
Μέτρα θέσης χρόνου καθυστέρησης πτυχίου όσον αφορά το Έτος Εισαγωγής.....	57
Μέτρα θέσης χρόνου καθυστέρησης πτυχίου όσον αφορά την Έκταξη.....	57
4.1.2 Περιγραφική Στατιστική	58
Ιστόγραμμα χρόνου καθυστέρησης πτυχίου όσον αφορά το Βαθμό.....	59
Ιστόγραμμα χρόνου καθυστέρησης πτυχίου όσον αφορά την Κατεύθυνση	60
Ιστόγραμμα χρόνου καθυστέρησης πτυχίου όσον αφορά το Φύλο	61
Ιστόγραμμα χρόνου καθυστέρησης πτυχίου όσον αφορά την Χρονιά Εισαγωγής.....	62
Ιστόγραμμα χρόνου καθυστέρησης πτυχίου όσον αφορά την Έκταξη	64
4.1.3 Μοντελοποίηση.....	65
Διαδικασία επιλογής μεταβλητών στο μοντέλο	65
Επεξήγηση Αποτελεσμάτων προηγούμενου πίνακα	67

4.2 Δεύτερο Ερώτημα Έρευνας : Παράγοντες που επηρεάζονται τον τελικό βαθμό πτυχίου	71
Δεύτερο Ερώτημα Έρευνας	71
Μέτρα Θέσης - Περιγραφική Στατιστική.....	72
4.2.1 Μοντελοποίηση.....	78
Επεξήγηση Αποτελεσμάτων	80
4.3 Τρίτο Μοντέλο Μοντελοποίηση με χρήση Ανάλυση Επιβίωσης με καταληκτικό γεγονός τον χρόνο απόκτησης πτυχίου	82
4.3.1 Περιγραφή Μοντέλου	82
4.3.2 Μοντελοποίηση με το μοντέλο του Cox	86
Συμπεράσματα:	87
Πρόβλεψη προσέλευσης αναμενόμενου αριθμού φοιτητών σε μια εξέταση με χρήση ανάλυσης χρονοσειρών.....	89
5. Παράρτημα	97
5.1 Παράρτημα Α-Υπολογισμός κόστους φοιτητή που έγραψε από 2 και κάτω	97
5.2 Παράρτημα Β – Κώδικας R	100
5.2.1 Κώδικας για έτος επιτυχίας.....	100
5.2.2 Προσπάθειες μέχρι την πρώτη επιτυχία και Προσπάθειες μέχρι την πρώτη επιτυχία αλλά εξαιρώ τις προσπάθειες που πήρανε 0 ή 1 ή 2.....	101
5.2.3. Πόσοι δήλωσαν-πέρασαν μάθημα ανά εξάμηνο και Σεπτέμβρη	102
6. Βιβλιογραφία	140

1.Εισαγωγή

Σκοπός της παρούσας διπλωματικής είναι η Στατιστική Ανάλυση των δεδομένων των προπτυχιακών φοιτητών του τμήματος Μαθηματικών Αθήνας των ετών 2000 έως 2015. Αρχικά παραθέτουμε κάποιες εισαγωγικές στατιστικές μεθόδους μοντελοποίησης: Γραμμικά μοντέλα, Γενικευμένα γραμμικά μοντέλα (Glm) και τελικά τα Γενικευμένα γραμμικά μοντέλα τυχαίων επιδράσεων. Στην πορεία γίνεται αναλυτική παρουσίαση των δεδομένων των φοιτητών που χρησιμοποιήθηκαν. Ολοκληρώνουμε με τα μοντέλα για την στατιστική ανάλυση των δεδομένων και τα όποια συμπεράσματα.

Θα ήθελα ιδιαίτερα να ευχαριστήσω τον επιβλέποντα καθηγητή κ.Σιάννη Φώτη για την καθοδήγηση του, το υλικό που μου έδωσε για μελέτη και τις χρήσιμες συμβουλές του σε όλα τα στάδια της εκπόνησης αυτής της διπλωματικής εργασίας αλλά και τα όσα έμαθα από αυτόν κατά τη διάρκεια των σπουδών μου στο μεταπτυχιακό αυτό πρόγραμμα. Ευχαριστώ επίσης την κ.Ειρήνη Νινέττα Χαμάου για την βοήθεια της στην επεξήγηση των δεδομένων και ειδικότερα στην αποσαφήνιση των συντομογραφιών. Τα δεδομένα παραλήφθησαν την Τρίτη 19 Απριλίου 2016 από το τμήμα μηχανογράφησης της διεύθυνσης Εκπαίδευσης και Έρευνας του ΕΚΠΑ που βρίσκεται στην Πανεπιστημιούπολη στο υπόγειο του κτηρίου του Μαθηματικού κατόπιν αίτησης μέσω της γραμματείας του τμήματος μαθηματικού Αθήνας.

Τα δεδομένα βρισκότουσαν σε μορφή excel ανεπεξέργαστα (raw) με μερικές ελλείψεις σε μεταβλητές στοιχεία κάποιων φοιτητών. Το πρώτο excel περιείχε στοιχεία από όλους τους προπτυχιακούς φοιτητές όπως τον αριθμό μητρώου, το φύλο, τον τρόπο εισαγωγής στο πανεπιστήμιο, την σπουδαστική κατάσταση, την εθνικότητα, την υπηκοότητα, κλπ, λεπτομερή παρουσίαση τους δίνεται αργότερα. Στο δεύτερο excel οι πληροφορίες αφορούν τα μαθήματα που δηλώνουν οι φοιτητές: πόσες φορές το δηλώνουν, τότε το περνάνε κάθε μάθημα (σε ποιο εξάμηνο) και με τι βαθμό. Ακόμα μας χορηγήθηκαν αρχεία excel με επεξήγηση όλων των κωδικών-συντομογραφιών (κωδικοί εισαγωγής, κωδικοί σχολείων κτλ) που υπάρχουν στα προηγούμενα αρχεία.

Αρχικά θα γίνει μια περιγραφή του μαθηματικού υπόβαθρου (θεωρία) που θα χρησιμοποιήσουμε για την ανάλυση των δεδομένων. Στην συνέχεια θα ακολουθήσουν κάποια πρώτα συμπεράσματα από τα δεδομένα (περιγραφική στατιστική), και τελικά μέσω στατιστικών μοντελοποιήσεων θα απαντήσουμε στα διάφορα ερωτήματα που έχουμε θέσει.

2.Θεωρία

2.1 Πολλαπλό Γραμμικό Μοντέλο

Σε αυτό το κεφάλαιο κάνουμε μία αναφορά στο κλασικό πολλαπλό γραμμικό μοντέλο. Στο πολλαπλό γραμμικό μοντέλο μας ενδιαφέρει να εξετάσουμε την σχέση μεταξύ μίας εξαρτημένης μεταβλητής και ενός συνόλου ανεξάρτητων μεταβλητών ξαθώς και να μπορέσουμε να πραγματοποιήσουμε προβλέψεις για την εξαρτημένη μεταβλητή. Η εξαρτημένη μεταβλητή συμβολίζεται με Y και το σύνολο των ανεξάρτητων συμβολίζονται με X_i . Η Y είναι τυχαία μεταβλητή ενώ οι οι ανεξάρτητες μεταβλητές θεωρούνται σταθερές με γνωστές τιμές και όχι τυχαίες.

2.2 Το κλασικό Πολλαπλό Γραμμικό Μοντέλο

Το γραμμικό μοντέλο με μία μόνο ερμηνευτική μεταβλητή έχει την μορφή

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_r X_{ri} + \varepsilon_i,$$

όπου $X_1, X_2, \dots, X_r$ οι r ανεξάρτητες μεταβλητές, $\beta = (\beta_0, \beta_1, \dots, \beta_r)$ είναι το διάνυσμα που περιέχει τις άγνωστες παραμέτρους του γραμμικού μοντέλου και ε το τυχαίο σφάλμα, που αποτελείται από κάποιους τυχαίους παράγοντες που δεν προσμετρούνται κατά την κατασκευή του μοντέλου. Το διάνυσμα ε όπως και το διάνυσμα Y θεωρείται τυχαία μεταβλητή.

Όταν έχουμε n ανεξάρτητες παρατηρήσεις για το Y το μοντέλο παίρνει την παρακάτω μορφή,

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \dots + \beta_r X_{1r} + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \dots + \beta_r X_{2r} + \varepsilon_2$$

$$\dots \dots \dots$$

$$Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \dots + \beta_r X_{nr} + \varepsilon_n$$

Σε σημειογραφία πινάκων έχουμε,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1r} \\ 1 & X_{21} & X_{22} & \dots & X_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Σε πίνακική μορφή $\underline{Y} = X \cdot \underline{\beta} + \underline{\varepsilon}$

όπου :

$$\underline{Y} = (Y_1, Y_2, \dots, Y_n)',$$

$$\underline{\beta}_0 = (\beta_0, \beta_1, \dots, \beta_r)',$$

$$\underline{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$$

και

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix},$$

2.2.1 Βασικές Υποθέσεις του Γραμμικού Μοντέλου

Το γραμμικό μοντέλο είναι δομημένο κάτω από μερικές υποθέσεις για το διάνυσμα $\underline{\varepsilon}$. Για το σφάλμα της j δοκιμής έχουμε,

1. $E(\varepsilon_j) = 0, j = 1, 2, \dots, n$
2. $Var(\varepsilon_j) = \sigma^2, j = 1, 2, \dots, n$
3. $Cov(\varepsilon_j, \varepsilon_k) = 0, j \neq k$

Ή με μορφή πινάκων για το τυχαίο σφάλμα,

1. $E(\underline{\varepsilon}) = \mathbf{0}$,
2. $Cov(\underline{\varepsilon}) = E(\underline{\varepsilon}\underline{\varepsilon}') = \sigma^2 I_n$

Η πρώτη υπόθεση μας εξηγεί ότι δεν απαιτείται κανένας άλλος παράγοντας για να προβλέψουμε την Y και οποιαδήποτε μεταβλητότητα της Y οφείλεται σε τυχαίους παράγοντες. Η δεύτερη υπόθεση αναφέρει ότι όλοι οι τυχαίοι παράγοντες έχουν ίδια διασπορά, ενώ η τρίτη ότι οι όροι του σφάλματος είναι ασυσχέτιστοι.

Επίσης υποθέτουμε ότι ,

- $\varepsilon_j \sim N(0, \sigma^2)$ (υπόθεση κανονικότητας) ή
- $\underline{\varepsilon} \sim N_n(0, \sigma^2 I_n)$.

για να μας βοηθήσει στην πραγματοποίηση όλων εκείνων των στατιστικών τεστ που απαιτούνται για το γραμμικό μοντέλο. Από αυτές τις υποθέσεις συνεπάγονται οι παρακάτω ιδιότητες για την εξαρτημένη μεταβλητή Y .

1. $E(Y) = X\beta$
2. $Cov(Y) = \sigma^2 I_n$
3. $Y \sim N_n(X\beta, \sigma^2 I_n)$

2.3 Εκτίμηση των Παραμέτρων του Γραμμικού Μοντέλου

2.3.1 Εκτίμηση του β

Επειδή η ευθεία παλινδρόμησης (2.1) είναι άγνωστη, είναι ανάγκη να εκτιμήσουμε το διάνυσμα β με τέτοιο τρόπο έτσι ώστε να βρούμε την δειγματική εξίσωση εκείνη που εμμενεί καλύτερα την γραμμική συσχέτιση μεταξύ της εξαρτημένης και της ανεξάρτητης μεταβλητής και που για νέες τιμές των ανεξάρτητων μεταβλητών θα δίνει τη καλύτερη πρόβλεψη για την εξαρτημένη μεταβλητή Y . Αν λοιπόν $\hat{\beta}$ μία οποιαδήποτε εκτίμηση του β τότε βασικός στόχος είναι η διαφορά $Y_j - b_0 - b_1X_{j1} - b_2X_{j2} - \dots - b_rX_{jr}$ να είναι μικρή. Για να βρούμε τον κατάλληλο εκτιμητή, ο οποίος κάνει ελάχιστη αυτήν την διαφορά θα εκτιμήσουμε τα β_i με την μέθοδο των ελαχίστων τετραγώνων. Σκοπός μας είναι να κάνουμε ελάχιστο το,

$$\sum_{i=1}^n (\varepsilon_i)^2 = \sum_{i=1}^n (Y_j - \beta_0 - \beta_1X_{j1} - \beta_2X_{j2} - \dots - \beta_rX_{jr})^2$$

Οι εκτιμητές που προκύπτουν από αυτήν την μέθοδο ονομάζονται εκτιμητές ελαχίστων τετραγώνων και συμβολίζονται ως $\hat{\beta}$ ενώ οι διαφορές $Y_j - \hat{\beta}_0 - \hat{\beta}_1X_{j1} - \hat{\beta}_2X_{j2} - \dots - \hat{\beta}_rX_{jr}$, $j=1,2, \dots, n$ καλούνται κατάλοιπα.

Επίσης οι εκτιμώμενες τιμές του Y είναι $\hat{Y} = Z\hat{\beta} = E(Y)$

Αποτέλεσμα

Έστω ότι ο πίνακας X έχει πλήρη τάξη $r + 1 \leq n$, τότε η τιμή του $\hat{\beta}$ που εκτιμά με την μέθοδο των ελαχίστων τετραγώνων τα β είναι,

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Ιδιότητες

- Για το $\hat{\beta}$ συνεπάγονται οι εξής ιδιότητες,
 - $E(\hat{\beta}) = \beta$ (αμερόληπτος)
 - Ο εκτιμητής αυτός είναι γραμμικός ως προς Y
 - Είναι άριστος δηλαδή από όλους τους γραμμικούς εκτιμητές του β αυτός έχει την ελάχιστη διασπορά όπου $Con(\hat{\beta}) = \sigma^2(X'X)^{-1}X'Y$.
 - Ο εκτιμητής αυτός είναι επίσης εκτιμητής μεγίστης πιθανοφάνειας για τον β .

Σχόλια για τις Ιδιότητες

Αν υπάρχει ο αντίστροφος του πίνακα $(X' \cdot X)$ και βλέποντας ότι ο πίνακας

$$\frac{\partial^2 SSE(b)}{\partial(b^2)} = 2 \cdot X' \cdot X$$

είναι πάντα θετικά ημιορισμένος παίρνουμε τις παρακάτω εκτιμήτριες ελαχίστων τετραγώνων του $\underline{\beta}$:

$$\underline{\hat{\beta}} = (X' \cdot X)^{-1} \cdot X' \cdot \underline{Y}$$

Η παραπάνω εκτιμήτρια αποδεικνύεται ότι είναι αμερόληπτη, δηλαδή $E(\hat{\beta}) = \beta$ και αποτελεσματική (δηλαδή έχει την ελάχιστη διακύμανση από όλες τις γραμμικές αμερόληπτες εκτιμήτριες του β). Αποδεικνύεται επίσης ότι είναι και η εκτιμήτρια μέγιστης πιθανοφάνειας (ε.μ.π.). Η διασπορά του $\hat{\beta}$ δίνεται από την σχέση

$$V(\hat{\beta}) = (X' \cdot X)^{-1} \sigma^2$$

και σαν γραμμικός συνδυασμός κανονικής ακολουθεί κανονική κατανομή.

Δηλαδή, $\hat{\beta} \sim N_{p+1}(\beta, (X' \cdot X)^{-1} \sigma^2)$.

Η αμερόληπτη εκτιμήτρια του σ^2 δίνεται από την σχέση $\hat{\sigma}^2 = \frac{e'e}{n-k-1}$, η οποία είναι αποτελεσματική, ενώ η εκτιμήτρια μέγιστης πιθανοφάνειας του σ^2 είναι η $\hat{\sigma}^2 = \frac{e'e}{n}$ η οποία δεν είναι αμερόληπτη, ούτε αποτελεσματική. Οι ε.μ.π. των β , σ^2 είναι συνεπείς, ασυμπτωτικά αποτελεσματικές και ασυμπτωτικά κανονικές. Ορίζοντας ως $S(\hat{\beta}) = \hat{\sigma}^2 (X' \cdot X)^{-1}$ που τυπικά είναι η εκτιμήτρια της διασποράς της εκτιμήτριας των β , όταν η διασπορά των καταλοίπων είναι άγνωστη, αποδεικνύεται ότι ισχύει η σχέση

$$\frac{\hat{\beta}_i - \beta_i}{S(\hat{\beta}_i)} \sim t(n - k - 1)$$

όπου $S(\hat{\beta}_i)$ είναι το (i+1)-οστό διαγώνιο στοιχείο του πίνακα $S(\hat{\beta}_i)$ και από εδώ μπορούμε να σχηματίσουμε τα διαστήματα εμπιστοσύνης

$$\hat{\beta}_i \pm t_{\alpha/2}(n - k - 1) S(\hat{\beta}_i)$$

Εδώ να αναφέρουμε ότι για πιο πολύπλοκους ελέγχους της μορφής, πχ

$$H_0: \begin{cases} \beta_1 = 0 \\ \beta_2 - \beta_3 = 0 \end{cases}$$

ή σε μορφή πινάκων $H_0: R \cdot \beta = r$, όπου $R = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$, $\beta = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, χρησιμοποιούμε για ελεγχουσυνάρτηση την σχέση

$$F = \frac{(R\hat{\beta} - r)' [RS(\hat{\beta})R']^{-1} (R\hat{\beta} - r)}{j}$$

η οποία κάτω από την H_0 ακολουθεί την κατανομή $F(j, n - k - 1)$, που j είναι ο αριθμός των γραμμών του πίνακα r .

ΠΑΡΑΤΗΡΗΣΕΙΣ:

1. Όταν οι μεταβλητές X_i δεν είναι γραμμικώς ανεξάρτητες (υπάρχει δηλαδή κάποιο $c \neq 0$ και $i \neq j : x_i = cx_j$) τότε λέμε ότι παρουσιάζεται το πρόβλημα της πολυσυγγραμμικότητας, όπου δύο ή και παραπάνω μεταβλητές έχουν γραμμική σχέση η μία με την άλλη, με μαθηματική σχέση αυτό σημαίνει πως ο συντελεστής συσχέτισης των δύο μεταβλητών είναι 1. Γενικά η πολυσυγγραμμικότητα είναι ένα πρόβλημα το οποίο παρουσιάζεται όταν δύο μεταβλητές έχουν άμεση σχέση η μία με την άλλη και αυτό μπορεί να οδηγήσει σε λάθος συμπεράσματα αφού η μία μεταβλητή θα απορροφάει σημαντικότητα από την άλλη και θέλει ιδιαίτερη προσοχή στην αντιμετώπιση της. Πολλές φορές δεν μπορούμε να ξεχωρίσουμε την συμβολή κάθε μεταβλητής X στην ερμηνεία της Y .

2. Η σχέση ότι ο πίνακας

$$\frac{\partial^2 SSE(b)}{\partial (b^2)} = 2 \cdot X' \cdot X$$

είναι θετικά ημιορισμένος μας εξασφαλίζει ότι οι εκτιμήτριες

$$\underline{\hat{\beta}} = (X' \cdot X)^{-1} \cdot X' \cdot \underline{Y}$$

αντιστοιχούν πράγματι σε σημείο ελαχίστου και όχι μεγίστου ή σημείου καμπής.

3. Για $p = 1$ έχουμε το απλό γραμμικό μοντέλο, στο οποίο υπάρχει μόνο μία ανεξάρτητη μεταβλητή X και μόνο μία εξαρτημένη μεταβλητή Y , η οποία μπορεί να προσεγγιστεί ικανοποιητικά από μία γραμμική συνάρτηση του X , δηλαδή $Y = \beta_0 + \beta_1 X + \varepsilon$, όπου ε παριστάνει ένα τυχαίο σφάλμα μη παρατηρήσιμο (σφάλμα μετρήσεων, επιδράσεις από άλλες εξαρτημένες μεταβλητές που δεν συμπεριλήφθησαν στο μοντέλο).

Στην συνέχεια ορίζουμε τις ποσότητες

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2,$$

$$SSE = \sum_{i=1}^n [y_i - (\hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip})]^2,$$

$$SSR = \sum_{i=1}^n (\hat{b}_1 x_{i1} + \dots + \hat{b}_p x_{ip} - \bar{y})^2,$$

για τις οποίες ισχύει η σχέση $SST = SSR + SSE$, με την προϋπόθεση βέβαια ότι τα $\hat{\beta}$ είναι εκτιμήτριες ελαχίστων τετραγώνων. Έτσι ένα μοντέλο θα το θεωρούμε καλό όταν ένα μεγάλο κομμάτι του SST , το αποδίδει στα SSR και ένα μικρό στο SSE . Ορίζοντας $R^2 = \frac{SSR}{SST}$ (Coefficient of determination) με $0 \leq R^2 \leq 1$, έχουμε ότι όσο πιο μεγάλο το R^2 τόσο καλύτερο το μοντέλο (όπου αν $R^2 = 1 \Leftrightarrow SSE = 0 \Leftrightarrow y_i =$

$\hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}$) ενώ όταν το R^2 παίρνει την τιμή μηδέν, συμπεραίνουμε ότι η μεταβλητή Y δεν επηρεάζεται από τις ανεξάρτητες μεταβλητές $x_1, x_2, \dots, x_p$ και στην περίπτωση αυτή θα πρέπει να αναζητήσουμε διαφορετικό σύνολο ανεξάρτητων μεταβλητών. Το πρόβλημα με το R^2 είναι ότι όσο προσθέτουμε όρους, αυτό μεγαλώνει και μπορεί να πάρει τιμές κοντά στο 1 χωρίς αναγκαστικά να υπάρχει συσχέτιση μεταξύ της Y και της X . Για αυτό δεν μπορούμε πάντα να χρησιμοποιήσουμε το R^2 για να δούμε αν το μοντέλο μας είναι ισχυρό.

Για αυτό συνήθως χρησιμοποιούμε το $R_{adj}^2 = 1 - \frac{n-1}{n-p}(1 - R^2)$ (όπου p η διάσταση του διανύσματος $\underline{\beta}$) για το οποίο ισχύει πάντα $R_{adj}^2 \leq R^2$.

Τέλος θα αναφερθούμε στο p -value για τον έλεγχο μιας στατιστικής υπόθεσης. Ορίζουμε το p -value ως την πιθανότητα να πάρει η στατιστική μας συνάρτηση ελέγχου τιμή τόσο ακραία ή περισσότερο ακραία από αυτή που παρατηρήσαμε. Δηλαδή το p -value είναι το ελάχιστο επίπεδο σημαντικότητας για να απορρίψουμε την μηδενική υπόθεση. Οπότε απορρίπτω την μηδενική υπόθεση H_0 σε επίπεδο στατιστικής σημαντικότητας όταν $p\text{-value} < \alpha$. Συνήθως συγκρίνω σε επίπεδο στατιστικής σημαντικότητας $\alpha = 5\%$ ή 1% (δηλαδή $\alpha = 0,01$ ή $\alpha = 0,05$).

2.4 Γενικευμένα Γραμμικά Μοντέλα -GLM

2.4.1. Εισαγωγή

Η πρόοδος στη στατιστική θεωρία μαζί με την ανάπτυξη των υπολογιστών μας επέτρεψαν να δημιουργήσουμε μεθόδους ανάλογους με αυτές που έχουν αναπτυχθεί για τα γραμμικά μοντέλα σε περιπτώσεις που οι μεταβλητές απόκρισης ακολουθούν κατανομή διαφορετική από την κανονική, δεν είναι απαραίτητα συνεχείς (μπορεί να είναι κατηγορικές μεταβλητές) και δεν χρειάζεται να είναι στην απλή μορφή: $Y = X \cdot \beta + \varepsilon$. Μια σημαντική παρατήρηση είναι ότι πολλές από τις χρήσιμες ιδιότητες της κανονικής κατανομής κατέχει η ομάδα κατανομών που ανήκει στην « εκθετική οικογένεια ».

Η εκτίμηση των παραμέτρων του γραμμικού μοντέλου $Y = X \cdot \beta + e$ επεκτάθηκε στην εκτίμηση παραμέτρων συναρτήσεων της μορφής $Y = g(X \cdot \beta) + \varepsilon$, όπου g πραγματική συνάρτηση. Θεωρητικά οι διαδικασίες εκτίμησης είναι απλές. Στη πράξη απαιτούν ένα μεγάλο όγκο υπολογισμών οι οποίοι έγιναν εφικτοί μόνο μέσω υπολογιστών με τη βοήθεια αριθμητικών προσεγγίσεων μη γραμμικών συναρτήσεων. Θα δούμε στη συνέχεια πως επεκτείνονται τα κλασσικά γραμμικά μοντέλα σε γενικευμένα γραμμικά μοντέλα.

Το γενικευμένο γραμμικό μοντέλο είναι ένα 'εργαλείο' για την μοντελοποίηση μιας μεταβλητής απόκρισης Y στη μορφή $g(X \cdot \beta) + \varepsilon$, όπου η συνάρτηση g λέγεται συνδετική συνάρτηση. Τα μοντέλα αυτά επιτρέπουν κάποιο βαθμό μη γραμμικότητας στην σχέση μεταξύ των μεταβλητών X και Y , διατηρώντας παράλληλα τον κεντρικό ρόλο για τον γραμμικό εκτιμητή $X \cdot \beta$ όπως και στο απλό γραμμικό μοντέλο.

Κάθε γενικευμένο γραμμικό μοντέλο αποτελείται από τρεις συνιστώσες:

1. Την κατανομή της μεταβλητής απόκρισης
2. Μια γραμμική παράμετρο πρόβλεψης που περιέχει τις μεταβλητές παλινδρόμησης x_i
3. Την συνάρτηση σύνδεσης η οποία ενώνει τη γραμμική παράμετρο πρόβλεψης με τη μέση τιμή της απόκρισης

Εξετάζοντας το μοντέλο μας πιο αναλυτικά παρατηρούμε ότι ορίζεται από ένα σύνολο τυχαίων μεταβλητών $Y = (Y_1, Y_2, \dots, Y_n)$, οι οποίες ονομάζονται μεταβλητές απόκρισης, κάθε μία από τις οποίες ακολουθεί μία κατανομή $Y_i / (X_1, X_2, \dots, X_p) \sim \text{κατανομή}(\theta)$ από την εκθετική οικογένεια κατανομών με θ : διάνυσμα παραμέτρων και X_j : επεξηγηματικές μεταβλητές. Υπάρχει συνάρτηση g , η οποία είναι μονότονη και διαφορίσιμη και επιλέγεται με τέτοιο τρόπο ώστε να ικανοποιούνται τυχόντες περιορισμοί για το θ . Η γραμμική παράμετρος πρόβλεψης του μοντέλου, μέσω της οποίας η κατανομή των Y_i εξαρτάται από τα x_i , είναι της μορφής: $n = \beta_0 + \beta_1 \cdot X_1 + \dots + \beta_p \cdot X_p$, ώστε ο τρόπος σύνδεσης των παραμέτρων ή κάποιας εκ των παραμέτρων της τυχαίας συνιστώσας με τη γραμμική παράμετρο πρόβλεψης του

μοντέλου να είναι μέσω μιας συνάρτησης σύνδεσης (linkfunction) g έτσι ώστε: $g(\theta) = \eta = \beta_0 + \beta_1 \cdot X_1 + \dots + \beta_p \cdot X_p$.

Συνήθως έχουμε ένα τυχαίο δείγμα n παρατηρήσεων $(Y_i, X_{1i}, \dots, X_{pi}, i = 1, \dots, n)$ με την βοήθεια του οποίου εκτιμούμε τα $(\beta_0, \beta_1, \dots, \beta_p)$.

2.4.2 Παραδείγματα-Ειδικές περιπτώσεις

Παράδειγμα 1

Έστω τυχαίες συνιστώσες $Y_1, \dots, Y_n$ κανονικά κατανεμημένες. Η κανονική κατανομή ανήκει στην εκθετική οικογένεια κατανομών. Επιπλέον οι επεξηγηματικές μεταβλητές εισάγουν το γραμμικό μοντέλο διαμέσου ενός γραμμικού εκτιμητή: $\eta_i = x_i^T \cdot \beta$, οπότε η σύνδεση μεταξύ των $E(Y) = \mu$ και του γραμμικού εκτιμητή η γίνεται διαμέσου της συνδετικής συνάρτησης: $\mu_i = \eta_i, i = 1, 2, \dots, n$.

Παράδειγμα 2

Έστω τυχαία συνιστώσα $Y_i \sim \text{Binomial}(n_i, p_i)$ όπου Y_i εκφράζει τον αριθμό επιτυχιών σε n_i επαναλήψεις και οι επεξηγηματικές μεταβλητές $X_j, j = 1, \dots, p$ μπορεί να είναι συνεχείς ή διακριτές. Ξέρουμε ότι η διωνυμική κατανομή ανήκει στην εκθετική οικογένεια κατανομών με $A(\eta) = \eta \cdot \log(1 + e^\eta)$.

Γνωρίζουμε ότι: $\eta \cdot p = E(Y) = A'(\eta) \Rightarrow \eta = (A')^{-1}(\eta \cdot p) \Rightarrow \eta = \log\left(\frac{p}{1-p}\right)$, άρα η συνδετική συνάρτηση που προκύπτει είναι: $\log(p/(1-p)) = \eta_i = \beta_0 + \beta_1 \cdot X_{1i} + \dots + \beta_p \cdot X_{pi}$ οπότε $g(p) = \text{logit}(p/(1-p))$, με $g: \mathbb{R} \rightarrow (0,1)$.

Παράδειγμα 3

Έστω τυχαία συνιστώσα $Y_i \sim \text{Poisson}(\lambda_i)$, όπου Y_i εκφράζει αριθμό εμφανίσεων γεγονότων σε ένα χρονικό/χωρικό διάστημα. Οι επεξηγηματικές μεταβλητές X_j μπορεί να είναι συνεχείς ή διακριτές, $j = 1, \dots, p$. Γνωρίζουμε ότι η κατανομή *Poisson* ανήκει στην εκθετική οικογένεια κατανομών με $A(\eta) = e^\eta$.

Έχουμε: $\lambda = E(Y) = A'(\eta) \Rightarrow \eta = (A')^{-1}(\lambda) \Rightarrow \eta = \log(\lambda)$.

Άρα η συνδετική συνάρτηση (linkfunction) που προκύπτει είναι:

$$\log(\lambda_i) = \eta_i = \beta_0 + \beta_1 \cdot X_{1i} + \dots + \beta_p \cdot X_{pi}.$$

Άρα $g(\eta) = \log(\eta)$.

Στο *Poisson* μοντέλο παρατηρούμε ότι $\mu = E(Y) > 0$ οπότε δεν μπορούμε να έχουμε $\mu = X^T \cdot \beta$, αφού για το δεξί μέλος της ισότητας δεν ισχύει ο απαραίτητος περιορισμός. Γιαυτό είναι αναγκαίο στην περίπτωση αυτή να ορίσουμε την συνάρτηση $g: \mathbb{R}^+ \rightarrow \mathbb{R}$.

2.4.3 Κλασσική Εκτίμηση για Γενικευμένα Γραμμικά Μοντέλα

Το Γ.Γ.Μ. ορίζεται σε σχέση με ένα σύνολο από ανεξάρτητες τυχαίες αποκρίσεις $y_1, y_2, \dots, y_n$ με μέσους $\mu_1, \mu_2, \dots, \mu_n$, που ακολουθούν κατανομές από την εκθετική οικογένεια κατανομών, που έχουν την ίδια συναρτησιακή μορφή $f(y_i; n_i) = \exp[n_i T_i(y_i) - A(n_i) + H(y_i)]$, με $n(\cdot), T(\cdot), A(\cdot), H(\cdot)$ πραγματικές συναρτήσεις.

Όπως είπαμε το μοντέλο αυτό περιέχει τις μεταβλητές παλινδρόμησης τις $x_1, x_2, \dots, x_p$ και κατασκευάζεται βάση της γραμμικής παραμέτρου πρόβλεψης: $\mathbf{n} = \mathbf{x}^T \cdot \boldsymbol{\beta} = \beta_0 + \sum_{i=1}^p \beta_i \cdot x_i$. Η σύνδεση ανάμεσα στην κατανομή των Y και στην γραμμική παράμετρο πρόβλεψης $\mathbf{n}$ γίνεται μέσω της συνάρτησης σύνδεσης: $\mathbf{n}_i = g(\mu_i), i = 1, 2, \dots, n$ όπου $\mu_i \equiv E(Y_i), i = 1, 2, \dots, n$ οπότε $g(\mu_i) = g(E(Y_i)) = \mathbf{n}_i = \mathbf{x}_i^T \cdot \boldsymbol{\beta}, i = 1, 2, \dots, n$.

Στις κατανομές των Y που ανήκουν στην εκθετική οικογένεια κατανομών ισχύει $E(Y) = A'(n)$ οπότε $n = (A')^{-1}(\mu)$. Από τις προηγούμενες σχέσεις έχουμε: $\mathbf{n} = (A')^{-1}(g^{-1}(\mathbf{x}_i^T \cdot \boldsymbol{\beta}))$. Ορίζουμε ως κανονική συνάρτηση σύνδεσης την $g(\mu) = (A')^{-1}(\mu)$, η οποία δηλώνει ότι υπάρχει σύνδεση μεταξύ του μέσου και της γραμμικής παραμέτρου πρόβλεψης με μέση τιμή απόκρισης: $E(y_i) = g^{-1}(\mathbf{n}_i) = g^{-1}(\mathbf{x}_i^T \cdot \boldsymbol{\beta})$.

Παρακάτω δίνεται ένας πίνακας των κανονικών συνδετικών συναρτήσεων που χρησιμοποιούνται στις πιο συχνές κατανομές από την εκθετική οικογένεια κατανομών στα Γ.Γ.Μ.

Κατανομή	Κανονική Συνδετική Συνάρτηση
Κανονική	$\mathbf{n}_i = \mu_i$
Διωνυμική	$\mathbf{n}_i = \log(p/(1-p))$
Poisson	$\mathbf{n}_i = \log(\mu_i)$
Εκθετική	$\mathbf{n}_i = \frac{1}{\mu_i}$
Γάμμα	$\mathbf{n}_i = \frac{1}{\mu_i}$
Αντίστροφη Κανονική	$\mathbf{n}_i = \frac{1}{\mu_i^2}$

Πίνακας: Συνδετικές συναρτήσεις διαφόρων κατανομών.

Η εκτίμηση των παραμέτρων των Γ.Γ.Μ. δεν μπορεί να γίνει με αναλυτικό τρόπο, λόγω της πολυπλοκότητας που παρουσιάζει. Γιαυτό καταφεύγουμε σε κάποια μέθοδο αριθμητικής μεγιστοποίησης.

Μία τέτοια μέθοδος είναι και ο αλγόριθμος *Newton-Raphson*. Σκοπός της μεθόδου είναι η μεγιστοποίηση της συνάρτησης πιθανοφάνειας $\log L(\theta) = l(\theta)$. Τα βήματα της είναι τα παρακάτω:

1. Δίνουμε αρχικές τιμές στο $\theta = (\theta_1, \theta_2, \dots, \theta_p)$.
2. Υπολογίζουμε $\theta_{t+1} = \theta_t - \frac{l'(\theta_t)}{l''(\theta_t)}$.
3. Επαναλαμβάνουμε το 2^ο βήμα μέχρι $\theta_{t+1} \cong \theta_t$, δηλαδή μέχρι να επιτευχθεί σύγκλιση.

Πολλές φορές είναι αναγκαίο να χρησιμοποιήσουμε ένα διαφορετικό αλγόριθμο. Μια μορφή της μεθόδου Newton-Raphson όταν έχει χρησιμοποιηθεί η κανονική συνδετική συνάρτηση είναι ο αλγόριθμος Fisher-scoring ο οποίος είναι μία ειδική περίπτωση του αλγορίθμου επαναλαμβανόμενων σταθμισμένων ελαχίστων τετραγώνων.

Στα Γ.Γ.Μ. ο αλγόριθμος Fisher-scoring εφαρμόζεται με τα παρακάτω βήματα:

1. Δίνουμε αρχικές τιμές στα θ .
2. Παίρνουμε $\theta_{t+1} = \theta_t + \frac{l'(\theta_t)}{E(l''(\theta_t))}$, με $l'(\theta_t) = x^T \cdot W \cdot z$, όπου z διάνυσμα με συντεταγμένες $z_i = (Y_i - \mu_i)g'(\mu_i)$, W : διαγώνιος πίνακας με $W_{ii} = (g'(\mu_i)^2 \cdot A''(n_i))^{-1}$, $i = 1, 2, \dots, n$ και $E(l''(\theta_t)) = x^T \cdot W_t \cdot x$.
3. Επαναλαμβάνουμε το 2^ο βήμα μέχρι να επιτευχθεί σύγκλιση.

Οι δύο παραπάνω αλγόριθμοι έχουν διαφορά στο ότι ο αλγόριθμος *Newton – Raphson* χρησιμοποιεί την $l''(\theta)$, την παρατηρούμενη πληροφορία του *Fisher*, ενώ ο αλγόριθμος *Fisher scoring* χρησιμοποιεί την $E(l''(\theta))$ δηλαδή την αναμενόμενη πληροφορία του *Fisher*.

2.4.5 GLM, Λογιστική και Poisson παλινδρόμηση.

Τα πιο σημαντικά μοντέλα που συναντάμε στα γενικευμένα γραμμικά μοντέλα είναι αυτά της Λογιστικής και της Poisson παλινδρόμησης. Πρόκειται στην ουσία για Γ.Γ. Μ. για διωνυμικά και Poisson δεδομένα αντίστοιχα.

Γ.Γ.Μ. για Διωνυμικά Δεδομένα

Η λογιστική παλινδρόμηση είναι μέθοδος πολυπαραγοντικής στατιστικής ανάλυσης που χρησιμοποιεί ένα σύνολο ανεξάρτητων μεταβλητών για τη διερεύνηση της κίνησης μιας κατηγορικής εξαρτημένης μεταβλητής.

Στις περιπτώσεις αυτές χρησιμοποιούμε την τιμή 1 για την επιτυχία και την τιμή 0 για την αποτυχία. Στην συνέχεια ο υπολογισμός της μέσης τιμής της εξαρτημένης δίτιμης μεταβλητής ορίζει την αναλογία p των επιτυχιών στο σύνολο των δυνατών τιμών της.

Το κατάλληλο μοντέλο εδώ είναι η λογιστική παλινδρόμηση. Με την οποία εκτιμάμε την πιθανότητα επιτυχίας p μιας δίτιμης μεταβλητής για ένα πλήθος τιμών μίας ή περισσότερων ανεξάρτητων μεταβλητών. Ακόμα το μοντέλο λογιστικής παλινδρόμησης εκτός από την δυνατότητα εκτίμησης δίνει και την δυνατότητα να προβλέπει την επίδραση κάθε ανεξάρτητης μεταβλητής στη διαμόρφωση των τιμών της εξαρτημένης μεταβλητής. Στην λογιστική παλινδρόμηση, αντίθετα από την πολλαπλή παλινδρόμηση είναι δυνατό να χρησιμοποιηθούν ως εξαρτημένες μεταβλητές εκτός από αναλογικές αριθμητικές μεταβλητές (ratioscale) και κατηγορικές μεταβλητές.

Εάν $y_i \sim \text{Binomial}(n_i, p_i)$ η λογιστική παλινδρόμηση δίνεται από :

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = n_i = x_i^T \cdot \beta$$
 , όπου έχουμε *logit* συνδετική συνάρτηση. Παρόμοια με το λογιστικό μοντέλο έχουμε και το κανονικό μοντέλο πιθανότητας. Η διαφορά τους είναι πως για τον υπολογισμό της πιθανότητας p χρησιμοποιούμε την αθροιστική συνάρτηση πιθανότητας της κανονικής κατανομής. Δηλαδή αν $y_i \sim \text{Bernoulli}(\Phi(n_i))$ τότε $\text{probit}(p_i) = \Phi^{-1}(p_i)$, όπου εδώ έχουμε συνδετική συνάρτηση της μορφής *probit*.

Το Μοντέλο της Λογιστικής Παλινδρόμησης

Εκφράζουμε την πιθανότητα επιτυχίας p , μιας δίτιμης μεταβλητής Y , η οποία ακολουθεί Διωνυμική κατανομή μέσω ενός γραμμικού μοντέλου: $p = x_i^T \cdot \beta$, όπου x : τιμές μιας ανεξάρτητης μεταβλητής X . Το μείζον πρόβλημα που συναντάμε εδώ είναι πως αν και οι τιμές της p θεωρητικά δεν μπορούν να βρίσκονται εκτός του διαστήματος $[0,1]$, από την άλλη όμως οι τιμές της ποσότητας $x_i^T \cdot \beta$ κυμαίνονται σε όλο το εύρος του διαστήματος των πραγματικών αριθμών. Γιαυτό το λόγο αναπαραμετροποιούμε το μοντέλο. Θεωρούμε το φυσικό λογάριθμο του λόγου της

πιθανότητας επιτυχίας προς την πιθανότητα αποτυχίας. Τον λόγο αυτό τον ονομάζουμε odds. Έτσι έχω:

$odds = \left(\frac{p}{1-p}\right)$ και $\log(odds) = \log\left(\frac{p}{1-p}\right)$. Ο λόγος αυτός παίρνει τιμές σε όλο το διάστημα των πραγματικών αριθμών.

Επομένως $\log\left(\frac{p}{1-p}\right) = x_i^T \cdot \beta$. Θέτοντας $Z = x_i^T \cdot \beta$ στην προηγούμενη σχέση, έχουμε: $p = \frac{e^Z}{1+e^Z}$.

Η τελευταία σχέση μας δίνει την εκτίμηση της πιθανότητας επιτυχίας της δίτιμης μεταβλητής Y , για δεδομένες τιμές των επεξηγηματικών μεταβλητών. Ακόμα παρατηρούμε ότι η αύξηση κατά μία μονάδα της επεξηγηματικής μεταβλητής αυξάνει τα odds αυξάνονται πολλαπλασιαστικά κατά $\exp$ του αντίστοιχου β ή αλλιώς τα $\log(odds)$ αυξάνονται κατά αντίστοιχο β .

Poisson Παλινδρόμηση

Η κατανομή Poisson είναι γνωστό πως είναι η κατανομή των σπάνιων γεγονότων και χρησιμοποιείται όταν θέλουμε να μετρήσουμε το πλήθος των γεγονότων-συμβάντων που συμβαίνουν στην μονάδα του χρόνου. Η τυχαία μεταβλητή Y που εκφράζει την κατανομή Poisson εκφράζει το πλήθος των συμβάντων στην μονάδα μέτρησης και έχει συνάρτηση πιθανότητας :

$$f(y_i) = \frac{e^{-\mu_i} \cdot \mu_i^{y_i}}{y_i!}, i = 1, 2, \dots, n.$$

Εδώ χρησιμοποιούμε τη λογαριθμική συνδετική συνάρτηση που έχει μορφή: $g = \log(\mu_i) = \eta_i = x_i^T \cdot \beta, i = 1, 2, \dots, n$.

Πολλές φορές στην Poisson παλινδρόμηση αντί να έχουμε αριθμό εμφανίσεων γεγονότων σε κάποια μονάδα (χρόνου, χώρου και άλλων) έχουμε το ρυθμό εμφανίσεων των γεγονότων, π.χ. ο αριθμός των ασθενών που αρρώστησαν διαφέρει ανάλογα με την περιοχή που μελετάμε. Αυτό συμβαίνει όταν χρησιμοποιούμε κάποιους όρους σαν αντιστάθμισμα. Αυτούς τους όρους τους ορίζουμε σαν offset.

Αντί να έχουμε $\log \mu_i = \beta_0 + \beta_1 \cdot x_i$, έχουμε $\log \mu_i = \log e_i + \beta'_0 + \beta'_1 \cdot x_i$ (όπου $\log e_i$ έχει το ρόλο του offset για την i -παρατήρηση). Παρατηρούμε ότι για αύξηση της μεταβλητής x κατά μία μονάδα, έχουμε αύξηση του λογαρίθμου της αναμενόμενης τιμής της μεταβλητής y , $\log \mu_i = \log(E(y_i))$, κατά αντίστοιχο β .

2.4.6. GLMMs

Τα GLMMs επεκτείνουν τα GLM ώστε να περιλαμβάνουν τις τυχαίες επιδράσεις (random effects), που οφείλονται στο τυχαίο συντελεστή (random coefficients) και σε πρότυπα (patterns) που ακολουθεί η διακύμανση .

GLMM (Ορισμός)

Τα GLMM μπορούν να ορισθούν σαν :

$$y = \mu + e.$$

Όπως και στο GLM, μ είναι το διάνυσμα των αναμενόμενων μέσων όρων των παρατηρήσεων και e τα σφάλματα. Το διάνυσμα αυτό συνδέεται με τις παραμέτρους του μοντέλου μέσω της link συνάρτησης που συμβολίζεται και εδώ με g :

$$g(\mu) = X\alpha + Z\beta,$$

όπου X και Z είναι ο σταθερός και ο τυχαίος πίνακας σχεδιασμού, και α και β είναι τα διανύσματα των σταθερών και τυχαίων παραμέτρων όπως και στο κανονικό συνδυασμένο μοντέλο. Η τυχαία επίδραση β , υποθέτουμε ότι ακολουθεί την κανονική κατανομή:

$$\beta \sim N(0, G)$$

όπου G πίνακας διαστάσεων $q \times q$, όπου q είναι ο αριθμός των παραμέτρων που εκφράζουν τις τυχαίες επιδράσεις. Στα μοντέλα τυχαίων επιδράσεων, ο πίνακας G είναι πάντα διαγώνιος και οι τυχαίες επιδράσεις θεωρούνται μεταξύ τους ασυσχέτιστες. Εάν οι κύριες επιδράσεις θεωρηθούν ως τυχαίες στο απλό πολυκεντρικό παράδειγμα με τρεις κεντρικές, τότε ο πίνακας G έχει την μορφή:

$$G = \begin{pmatrix} \sigma_c^2 & 0 & 0 \\ 0 & \sigma_c^2 & 0 \\ 0 & 0 & \sigma_c^2 \end{pmatrix}, \text{ όπου } \sigma_c^2: \text{ είναι η κεντρική συνδιασμένη διακύμανση.}$$

Ο πίνακας διακυμάνσεων είναι ο $var(y) = V = var(\mu) + R$, όπου R είναι ο πίνακας διακυμάνσεων των κατάλοιπων, δηλαδή $var(e)$. Επομένως ο πίνακας V παρατηρούμε ότι δεν υπολογίζεται εύκολα όπως συμβαίνει στην περίπτωση κανονικών δεδομένων όπου $V = ZGZ' + R$. Αυτό συμβαίνει επειδή μ δεν είναι γραμμική συνάρτηση του β . Μία πρώτη εκτίμηση μπορεί να γίνει από το παρακάτω μοντέλο προσέγγισης: $V \approx BZGZ'B + R$, όπου B είναι ο διαγώνιος πίνακας των διακυμάνσεων που καθορίζεται από τις γνωστές κατανομές (π.χ. $B = diag\{\mu_i(1 - \mu_i)\}$, για διωνυμικά δεδομένα). Στις τυχαίες επιδράσεις και στα μοντέλα με τυχαίους συντελεστές, ο πίνακας υπολοίπων, R είναι διαγώνιος όταν τα κατάλοιπα θεωρηθούν ασυσχέτιστα. Οι διαγώνιες διακυμάνσεις είναι ίσες με τις αναμενόμενες διακυμάνσεις για τις γνωστές κατανομές και γι'αυτό $R = AB$ όπως και στα GLMs.

Στις τυχαίες επιδράσεις και στα μοντέλα με τυχαίους συντελεστές, ο πίνακας V μπορεί να γραφτεί ως $V \approx \mathbf{BZGZ}'\mathbf{B} + \mathbf{AB}$. Στα μοντέλα οι διακυμάνσεις που προκύπτουν από πρότυπα (covariancepatternmodel), τα συσχετισμένα κατάλοιπα επιτρέπονται και ο πίνακας R μπορεί να εκφραστεί σαν γινόμενο του πίνακα συσχετίσεων P γραμμένο σε γραμμική μορφή και του AB με την μορφή $R = \mathbf{A}^{1/2}\mathbf{B}^{1/2}\mathbf{P}\mathbf{B}^{1/2}\mathbf{A}^{1/2}$.

Ο λόγος που μετατρέπουμε το πίνακα P σε γραμμική μορφή είναι γιατί μπορεί έτσι να παραμετροποιηθεί ώστε να έχει συμπεριφορά στις διακυμάνσεις με τον ίδιο τρόπο όπως όταν έχω κανονικά δεδομένα. Αυτό είναι κάτι που δεν θα μας απασχολήσει προς το παρόν.

Η παράμετρος που καθορίζει τη διασπορά

Όπως και στα GLMs, η διακύμανση σε επίπεδο καταλοίπων μπορεί να μειωθεί (ή να αυξηθεί) χρησιμοποιώντας μία παράμετρο που καθορίζει την διασπορά. Η διακύμανση των καταλοίπων πολλαπλασιάζεται από την παράμετρο της διασποράς που συμβολίζεται με φ , έτσι ώστε, $R = \varphi \mathbf{A}^{1/2}\mathbf{B}^{1/2}\mathbf{P}\mathbf{B}^{1/2}\mathbf{A}^{1/2}$.

Εάν οι παρατηρούμενες διακυμάνσεις των καταλοίπων είναι ακριβώς ίδιες με τις αναμενόμενες ($\mathbf{A}^{1/2}\mathbf{B}^{1/2}\mathbf{P}\mathbf{B}^{1/2}\mathbf{A}^{1/2}$) τότε παράμετρος της διασποράς φ ισούται με 1. Πολλές φορές αυτό δεν συμβαίνει επειδή η τιμή που παίρνει η παράμετρος φ επηρεάζεται και από άλλους παράγοντες.

2.4.7 Η πιθανοφάνεια και η ψευδοπιθανοφάνεια

Ένας συνήθης τρόπος εφαρμογής του GLMM βασίζεται στην μεγιστοποίηση της πιθανοφάνειας για τις παραμέτρους του μοντέλου. Όμως μία δυσκολία που υπάρχει είναι ότι η συνάρτηση πιθανοφάνειας μπορεί να καθοριστεί μόνο για τις τυχαίες επιδράσεις και στα μοντέλα με τυχαίους συντελεστές. Δεν υπάρχει πιθανοφάνεια για μοντέλα `covariancepatternmodel` όπως δεν υπάρχει μία γενική πολυμεταβλητή κατανομή για μη-κανονικά δεδομένα (για κανονικά δεδομένα χρησιμοποιείται η πολυμεταβλητή κανονική κατανομή). Θα δούμε πως θα ξεπεράσουμε την δυσκολία αυτή ορίζοντας μία εναλλακτική συνάρτηση γνωστή με το όνομα ψευδο-πιθανοφάνεια, που έχει πολλές ομοιότητες με την γνωστή μας πιθανοφάνεια.

Στην παρακάτω ενότητα αρχικά θα ορίσουμε την πιθανοφάνεια για τις τυχαίες επιδράσεις και τα μοντέλα τυχαίων συντελεστών, στην συνέχεια θα ορίσουμε την ψευδο-πιθανότητα για `covariancepatternmodel` και τέλος θα δώσουμε μία γενική μορφή η οποία είναι κατάλληλη για όλους τους τύπους μικτών μοντέλων.

Η συνάρτηση πιθανοφάνειας για τυχαίες επιδράσεις και για `randomcoefficientmodels`

Για αυτά τα μοντέλα μπορούμε να υπολογίσουμε τη συνάρτηση πιθανοφάνειας από το γινόμενο των πιθανοφανειών του $\mathbf{y}/\boldsymbol{\beta}$ και $\boldsymbol{\beta}$ (όπου $\mathbf{y}/\boldsymbol{\beta}$ εννοείται $\mathbf{y}$ δεδομένου του $\boldsymbol{\beta}$, όπου θεωρούμε $\boldsymbol{\beta}$ το σαν σταθερά). Η πιθανοφάνεια για τις σταθερές επιδράσεις (fixedeffects) $\mathbf{a}$, και για τις παραμέτρους της διακύμανσης του πίνακα $\mathbf{G}, \boldsymbol{\gamma}_G$, μπορεί να βρεθεί με το παρακάτω τύπο:

$$L(\mathbf{a}, \boldsymbol{\gamma}_G; \mathbf{y}) = L(\mathbf{a}; \mathbf{y}/\boldsymbol{\beta})L(\boldsymbol{\gamma}_G; \boldsymbol{\beta}). \quad (\text{B})$$

Ας θεωρήσουμε ότι $\boldsymbol{\beta}$ ακολουθεί μία πολυμεταβλητή κανονική κατανομή, $\boldsymbol{\beta} \sim N(\mathbf{O}, \mathbf{G})$, τότε αντικαθιστώντας όπου $L(\boldsymbol{\gamma}_G; \boldsymbol{\beta})$ την πολυμεταβλητή κανονική κατανομή έχουμε:

$$L(\mathbf{a}, \boldsymbol{\gamma}_G; \mathbf{y}) \propto L(\mathbf{a}; \mathbf{y}/\boldsymbol{\beta})|\mathbf{G}|^{-1/2}\exp(-1/2\boldsymbol{\beta}'\mathbf{G}^{-1}\boldsymbol{\beta}).$$

Επειδή έχουμε υποθέσει ασυσχέτιστα κατάλοιπα εύκολα καθορίζεται χρησιμοποιώντας την κατανομή της $\mathbf{y}/\boldsymbol{\beta}$ (π.χ. διωνυμική, Poisson), θα έχουν την παρακάτω γνωστή μορφή:

$$\begin{aligned} L(\mathbf{a}; \mathbf{y}/\boldsymbol{\beta}) &= \exp\left\{\mathbf{y}'\mathbf{A}^{-1}\boldsymbol{\theta} - b(\boldsymbol{\theta})^{-1/2}'\mathbf{A}^{-1}b(\boldsymbol{\theta})^{-1/2} + K\right\} \\ &\propto \exp\left\{\mathbf{y}'\mathbf{A}^{-1}\boldsymbol{\theta} - b(\boldsymbol{\theta})^{-1/2}'\mathbf{A}^{-1}b(\boldsymbol{\theta})^{-1/2}\right\} \end{aligned}$$

Όπου:

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta}.$$

$\mathbf{A} = \text{diag}\{a_i\}$, όπου a_i είναι σταθεροί όροι.

$b(\boldsymbol{\theta}) = (b(\theta_1), b(\theta_2), \dots, b(\theta_n))'$ όπου b είναι η συνάρτηση που χρησιμοποιείται στην γενική μορφή της κατανομής και $K = \text{σταθερά}$.

Η συνολική πιθανοφάνεια για $\boldsymbol{\alpha}$ και $\boldsymbol{\gamma}_G$ μπορούν να εκφραστούν ως:

$$L(\boldsymbol{\alpha}, \boldsymbol{\gamma}_G; \mathbf{y}) \propto \exp\left\{\mathbf{y}'\mathbf{A}^{-1}\boldsymbol{\theta} - b(\boldsymbol{\theta})^{-1/2'}\mathbf{A}^{-1}b(\boldsymbol{\theta})^{-1/2}\right\} |\mathbf{G}|^{-1/2} \exp\left(-\frac{1}{2\boldsymbol{\beta}'\mathbf{G}^{-1}\boldsymbol{\beta}}\right).$$

με αντίστοιχη log-πιθανοφάνεια:

$$\log\{L(\boldsymbol{\alpha}, \boldsymbol{\gamma}_G; \mathbf{y})\} = \mathbf{y}'\mathbf{A}^{-1}\boldsymbol{\theta} - b(\boldsymbol{\theta})^{-1/2'}\mathbf{A}^{-1}b(\boldsymbol{\theta})^{-1/2} - 1/2 \log|\mathbf{G}| - 1/2 \boldsymbol{\beta}'\mathbf{G}^{-1}\boldsymbol{\beta} + K. (\Gamma)$$

Η ψευδο-πιθανοφάνεια για τα covariance pattern models

Σε αυτά τα μοντέλα οι παρατηρήσεις συσχετίζονται και το μοντέλο μπορεί να παραμετροποιηθεί από την σταθερή επίδραση $\boldsymbol{\alpha}$, και τις διακυμάνσεις των παραμέτρων που χρησιμοποιούνται στον πίνακα $\mathbf{R}$, που τις έχουμε συμβολίσει ως $\boldsymbol{\gamma}_R$.

Αφού μία γενική πολυμεταβλητή κατανομή δεν είναι διαθέσιμη για δεδομένα που δεν ακολουθούν κανονική κατανομή, δεν μπορούμε να ορίσουμε μία συνήθη πιθανοφάνεια. Αυτή η δυσκολία ξεπερνιέται ορίζοντας μία ψευδο-πιθανοφάνεια, $QL(\boldsymbol{\alpha}, \boldsymbol{\gamma}_R; \mathbf{y})$, που έχει παρόμοιες ιδιότητες με την συνήθη πιθανοφάνεια. Το διαφορικό της $\log$ ως προς $\boldsymbol{\alpha}$ έχει την ίδια μορφή με την μορφή του διαφορικού ως προς $\boldsymbol{\alpha}$ της συνήθους πιθανοφάνειας. Το διαφορικό της συνήθους πιθανοφάνειας ως προς $\boldsymbol{\alpha}$ έχει την μορφή:

$$\delta \log\{L(\boldsymbol{\alpha}, \boldsymbol{\gamma}_G; \mathbf{y})\} / \delta \boldsymbol{\alpha} = \mathbf{X}'\mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu})$$

Τώρα αφού $\mathbf{R} = \mathbf{A}\mathbf{B}$, έχουμε ότι $\mathbf{A}^{-1} = \mathbf{B}\mathbf{R}^{-1}$, και άρα

$$\delta \log\{L(\boldsymbol{\alpha}, \boldsymbol{\gamma}_G; \mathbf{y})\} / \delta \boldsymbol{\alpha} = \mathbf{X}'\mathbf{B}\mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\mu}),$$

όπου $\mathbf{R}$ ο πίνακας διακυμάνσεων των καταλοίπων.

Η γενική μορφή συνάρτησης της ψευδο-πιθανοφάνειας για όλα τα GLMMs

Είναι χρήσιμο να ορίσουμε μία μορφή για την ψευδο-πιθανοφάνεια η οποία θα είναι κατάλληλη για όλους τους τύπους που μπορεί να συναντήσουμε στα GLMMs και θα περιέχει τις τυχαίες επιδράσεις, τους συντελεστές και τα covariance patterns. Θα το κατορθώσουμε αυτό αντικαθιστώντας $L(\mathbf{a}; \mathbf{y}/\boldsymbol{\beta})$ στην σχέση (B) με $QL(\mathbf{a}, \boldsymbol{\gamma}_R; \mathbf{y}/\boldsymbol{\beta})$. Κάνοντας αυτό έχουμε : $QL(\mathbf{a}, \boldsymbol{\gamma}; \mathbf{y}) = QL(\mathbf{a}, \boldsymbol{\gamma}_R; \mathbf{y}/\boldsymbol{\beta})L(\boldsymbol{\gamma}_G; \boldsymbol{\beta})$,

και

$$\log\{QL(\mathbf{a}, \boldsymbol{\gamma}; \mathbf{y})\} = \log\left\{QL\left(\mathbf{a}, \boldsymbol{\gamma}_R; \frac{\mathbf{y}}{\boldsymbol{\beta}}\right)\right\} - 1/2 \log|\mathbf{G}| - 1/2 \boldsymbol{\beta}'\mathbf{G}^{-1}\boldsymbol{\beta} + K, (D)$$

όπου $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_G, \boldsymbol{\gamma}_R)$.

Αυτή η συνάρτηση ανταποκρίνεται σε μία συνήθης πιθανοφάνεια ανεξαιρέτως αν τα κατάλοιπα είναι ασυσχέτιστα .

2.5 Ανάλυση Επιβίωσης

2.5.1 Εισαγωγή

Στο παρόν κεφάλαιο θα αναφερθούμε στη θεωρία της ανάλυσης επιβίωσης. Ανάλυση Επιβίωσης είναι ο κλάδος της στατιστικής ο οποίος ασχολείται με την περιγραφή και την ανάλυση των δεδομένων σε σχέση με τον χρόνο από μια καλώς ορισμένη αρχική χρονική στιγμή έως την πραγματοποίηση κάποιου συγκεκριμένου γεγονότος ή το τέλος ενός ορισμένου χρόνου. Το γεγονός καλείται καταληκτικό γεγονός. Στη βιοστατιστική συνήθως είναι ο θάνατος.

Τα δεδομένα επιβίωσης γενικά δεν είναι συμμετρικά κατανεμημένα. Συνήθως ένα ιστόγραμμα από χρόνους επιβίωσης κάποιας ομάδας ατόμων με ίδια χαρακτηριστικά τείνει να έχει θετικό συντελεστής λοξότητας, δηλαδή είναι λοξή προς τα δεξιά και περιέχει το μεγαλύτερο κομμάτι των παρατηρήσεων. Οπότε θα ήταν παράλογο να υποθέσουμε κανονική κατανομή. Αυτός είναι και ένας λόγος που δεν χρησιμοποιούμε τα συνηθισμένα μοντέλα.

Βασικές έννοιες

Ένα ακόμα χαρακτηριστικό που συναντάμε στα δεδομένα επιβίωσης και κάνουν τις συνηθισμένες στατιστικές μεθόδους μη εφαρμόσιμες είναι ότι τα δεδομένα είναι συνήθως censored (λογοκριμένα). Ο χρόνος που συμβαίνει το καταληκτικό γεγονός ενός ατόμου λέγεται λογοκριμένος όταν το καταληκτικό γεγονός δεν παρατηρηθεί μέχρι την λήξη της έρευνας. Αυτό μπορεί να οφείλεται στο γεγονός ότι δεν του έχει συμβεί το καταληκτικό γεγονός της έρευνας όταν ολοκληρώνεται η έρευνα, είτε γιατί το άτομο σταμάτησε να συμμετέχει στην έρευνα.

Ας υποθέσουμε ότι το T συμβολίζει τον χρόνο επιβίωσης ενός ατόμου, με συνάρτηση κατανομής $F(T)$ και πυκνότητας $f(t)$, προφανώς αφού αναφερόμαστε σε χρόνο η τυχαία μεταβλητή δεν μπορεί να είναι αρνητική. Τότε ορίζεται η συνάρτηση επιβίωσης η οποία υποδηλώνει την πιθανότητα κάποιο άτομο να έχει επιβιώσει μετά την χρονική στιγμή t , δηλαδή έχουμε την σχέση

$$S(t) = P(T > t) = 1 - F(t).$$

Μια άλλη πολύ χρήσιμη σχέση που μας ενδιαφέρει στην ανάλυση επιβίωσης είναι η συνάρτηση κινδύνου (*hazard function*), η πιθανότητα να συμβεί καταληκτικό γεγονός σε κάποιον δεδομένου ότι έχει επιβιώσει μέχρι στιγμής. Αυτό μαθηματικά σημαίνει

$$h(t) = \lim_{dt \rightarrow 0} \frac{P(t \leq T < t + dt | T > t)}{dt}$$

Αναλογικότητα είναι η υπόθεση ότι η συνάρτηση κινδύνου μιας ομάδας ατόμων με κάποιο χαρακτηριστικό είναι ανάλογη με αυτή της ομάδας χωρίς αυτό το χαρακτηριστικό (και όμοια για περιπτώσεις με πάνω από μια δυνατές ομάδες ή και συνεχείς τιμές). Η υπόθεση της αναλογικότητας δεν πρέπει να γίνεται χωρίς ενδείξεις ύπαρξης αυτής. Το πρώτο που έχουμε να κάνουμε όταν θέλουμε να υποθέσουμε αναλογικότητα είναι να κοιτάζουμε στα γραφήματα κατά πόσο για τις διαφορετικές τιμές μια μεταβλητής οι γραφικές παραστάσεις των συναρτήσεων επιβίωσης είναι ανάλογες μεταξύ τους μέσα στον χρόνο. Αν απ'την άλλη τέμνονται ή είναι απλά παράλληλες δεν μπορούμε να υποθέσουμε αναλογικότητα.

2.5.2 Μοντελοποίηση με χρήση του αναλογικού μοντέλου Cox

Μοντελοποιώντας μια ανάλυση επιβίωσης μπορούμε παραδείγματος χάριν να ερευνήσουμε πως η επιβίωση μιας ομάδα από ασθενείς εξαρτάται από τις τιμές μιας ή και παραπάνω περιγραφικών μεταβλητών. Στην ανάλυση επιβίωσης το ενδιαφέρον μας επικεντρώνεται στην συνάρτηση κινδύνου στο να πάθει το άτομο το καταληκτικό γεγονός κάθε χρονική στιγμή και έχουμε την απευθείας μοντελοποίηση της συνάρτησης κινδύνου, σε σύγκριση με της συνήθεις μεθόδους όπου μοντελοποιούμε την γραμμική σχέση κάποιων μεταβλητών. Παρόλα αυτά πολλές αρχές και διαδικασίες των γραμμικών μοντέλων ισχύουν και εδώ.

Υπάρχουν δύο κύριοι λόγοι για να μοντελοποιήσουμε τα δεδομένα από μια ανάλυση επιβίωσης, ο ένας είναι να δούμε ποιοι παράγοντες επηρεάζουν την επιβίωση ενός ατόμου μέσω των περιγραφικών μεταβλητών και ο άλλος να μπορούμε να εκτιμήσουμε την επιβίωση κάποιου ατόμου.

Το μοντέλο που θα μιλήσουμε αρχικά είναι το μοντέλο αναλογικού κινδύνου του Cox, το οποίο υποθέτει την ύπαρξη αναλογικών κινδύνων, ενώ δεν υποθέτει κάποια κατανομή. Για να περιγράψουμε τι σημαίνει η υπόθεση αναλογικού κινδύνου, σαν παράδειγμα ας υποθέσουμε μια μελέτη για κάποια πάθηση όπου δοκιμάζονται δύο είδη θεραπείας, η καινούργια και η ήδη υπάρχουσα τότε με βάση την υπόθεση που αναφέραμε έχουμε για τις συναρτήσεις κινδύνων των ατόμων από την μια θεραπεία και από την άλλη, την σχέση

$$h_{\text{καινούρια}}(t) = \psi \cdot h_{\text{παλιά}}(t),$$

όπου ψ είναι ένας σταθερός θετικός αριθμός και είναι ο λόγος των δύο συναρτήσεων κινδύνου. Ανάλογα του αν το ψ είναι μικρότερο ή μεγαλύτερο του 1 μπορούμε να βγάλουμε συμπέρασμα για το ποια θεραπεία είναι καλύτερη.

Αφού το ψ είναι θετικό μπορούμε να το γράψουμε στην μορφή $\beta = \log \psi$. Ας θέσουμε τώρα την μεταβλητή X που παίρνει την τιμή 1 αν το άτομο βρίσκεται στην καινούρια θεραπεία, αλλιώς παίρνει την τιμή 0. Τότε η συνάρτηση επιβίωσης του ατόμου θα έχει την μορφή

$$h_j(t) = e^{\beta x_j} \cdot h_0(t),$$

όπου $j=1, \dots, n$ με n =πλήθος του δείγματος και $h_0(t)$ συνάρτηση κινδύνου της υπάρχουσας συνηθισμένης θεραπείας.

Γενικεύοντας το μοντέλο του Cox και θεωρώντας πως τα στοιχεία της έρευνάς μας δεν αποτελούνται μόνο από χρόνους του καταληκτικού γεγονότος ή χρόνους λογοκρισίας, αλλά και περιγραφικές μεταβλητές $x_1, \dots, x_p$, (ρφυσικός) οι οποίες μπορεί να είναι συνεχείς, διακριτές και κατηγορικές, έχουμε το γενικό αναλογικό μοντέλο:

$$h_j(t) = e^{\eta_j} \cdot h_0(t),$$

όπου

$$\eta_j = \beta_1 x_{1j} + \dots + \beta_p x_{pj}$$

Η $h_0(t)$ λέγεται αρχική ή μηδενική συνάρτηση κινδύνου και είναι η συνάρτηση κινδύνου του ατόμου που έχει όλες τις τιμές των μεταβλητών 0.

Η εκτίμηση των παραμέτρων $\beta_1, \dots, \beta_p$ γίνεται με την χρήση μερικής πιθανοφάνειας.

2.6 Χρονοσειρές

Στο παρακάτω κεφάλαιο θα ορίσουμε κάποιες γενικές έννοιες, μοντέλα και σχέσεις όσον αφορά την ανάλυση χρονοσειρών.

Ως χρονοσειρά ορίζουμε ένα δείγμα από παρατηρήσεις, για το ίδιο μέγεθος X σε κάποιες συνεχόμενες χρονικές στιγμές. Στην ουσία ενδιαφερόμαστε να μελετήσουμε την εξάρτηση αυτού του μεγέθους την στιγμή t σε σχέση με τις προηγούμενες χρονικές στιγμές. Έτσι συμβολίζουμε ως χρονοσειρά μεγέθους n

$$\{x_t\}_n = \{x_1, \dots, x_n\}$$

όπου x_i είναι η τιμή της παραμέτρου που ερευνάμε την χρονική στιγμή i .

Βασικά χαρακτηριστικά Χρονοσειρών

1.Στασιμότητα

Από μια στάσιμη χρονοσειρά αναμένεται η διακύμανση των τιμών να μην αλλάζει μέσα στο χρόνο. Σε μια μη στάσιμη όμως, αναμένουμε να υπάρχουν κάποιες **τάσεις** (πχ για κάποιο καιρό να υπάρχει ανοδική τάση και για κάποια άλλη πτωτική για την τιμή της βενζίνης λόγω διεθνών πολιτικών συνθηκών), ή κάποια **περιοδικότητα** (πχ ανάλογα την εποχή οι γεννήσεις είναι περισσότερες), το οποίο λέγεται αλλιώς και **εποχικότητα**, όπως και τις δύο παραπάνω περιπτώσεις.

2.Αιτιοκρατία και στοχαστικότητα

Ένα χαρακτηριστικό που θέλουμε να ελέγξουμε είναι αν υπάρχει κάτι 'κρυμμένο' πίσω από την εναλλαγή των τιμών, δηλαδή αν οι τιμές αλλάζουν λόγω κάποιας τυχαίας διαδικασίας η οποία πρέπει να βρεθεί και να αναλυθεί.

3.Γραμμικότητα και μη γραμμικότητα

Όπως στο γραμμικό μοντέλο ελέγξαμε την ύπαρξη γραμμικής σχέσης, έτσι κι εδώ θέλουμε να δούμε κατά πόσο ισχύει η γραμμική σχέση της τιμής x_t με τις προηγούμενες χρονικά παρατηρήσεις του ίδιου μεγέθους X .

Αθροιστική Μοντελοποίηση Χρονοσειράς

Με βάση τα προηγούμενα μια χρονοσειρά γίνεται θεωρηθεί ότι αποτελείται από τρεις συνιστώσες :

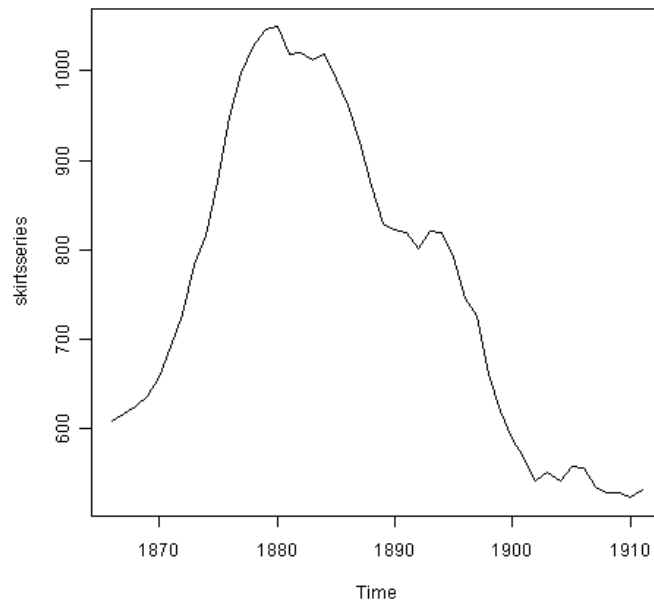
1. Την περιοδικότητα s_t
2. Την τάση μ_t
3. Την τυχειότητα y_t (υπόλοιπα)

Αν υποθέσουμε ότι οι τρεις παραπάνω χρονοσειρές παράγουν την πραγματική τιμή της χρονοσειράς του μεγέθους X , προστιθέμενη η μία στην άλλη τότε έχουμε την σχέση

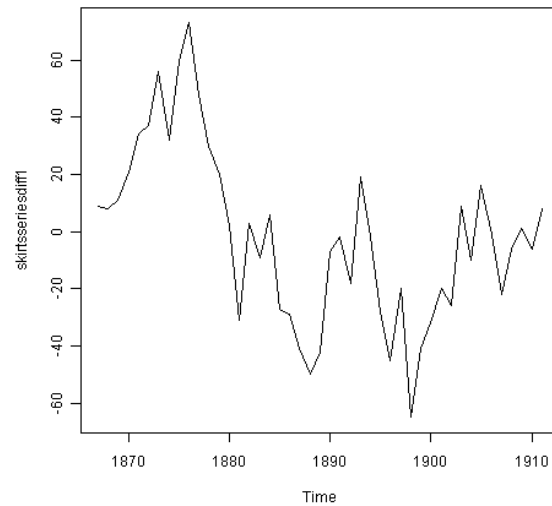
$$x_t = s_t + \mu_t + y_t$$

Στόχος μας είναι να 'ξεχωρίσουμε' τις τρεις αυτές συνιστώσες (αν υπάρχουν και οι τρεις, αλλιώς όσες υπάρχουν) ώστε να μπορούμε να τις μελετήσουμε ξεχωριστά και να βγάλουμε τα όποια συμπεράσματα.

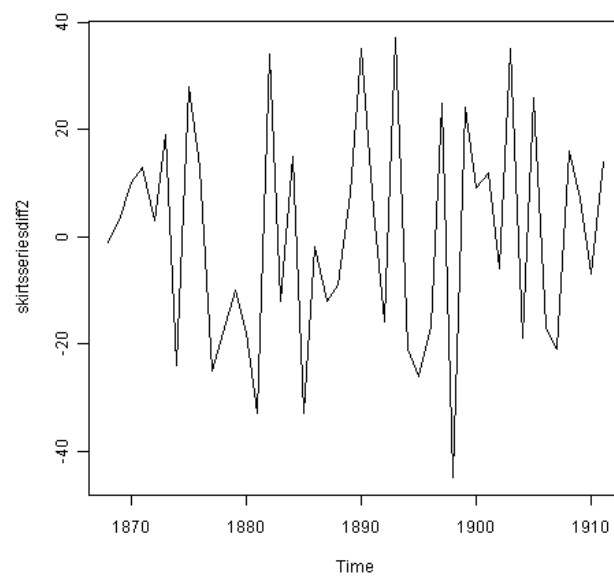
Ενδεικτικά ένας τρόπος απαλοιφή της τάσης είναι οι πρώτες διαφορές όπου αφαιρούμε από κάθε τιμή της χρονοσειράς την προηγούμενή της, $x_t - x_{t-1}$ και παρατηρούμε κατά πόσο έχει επαλειφθεί η όποια τάση υπάρχει στο γράφημα της χρονοσειράς, επαναλαμβάνουμε μέχρι να μην φαίνεται κάποια τάση. Παραδείγματος χάρη στο παρακάτω παράδειγμα, μια χρονοσειράς



εφαρμόζουμε την πρώτη διαφορά



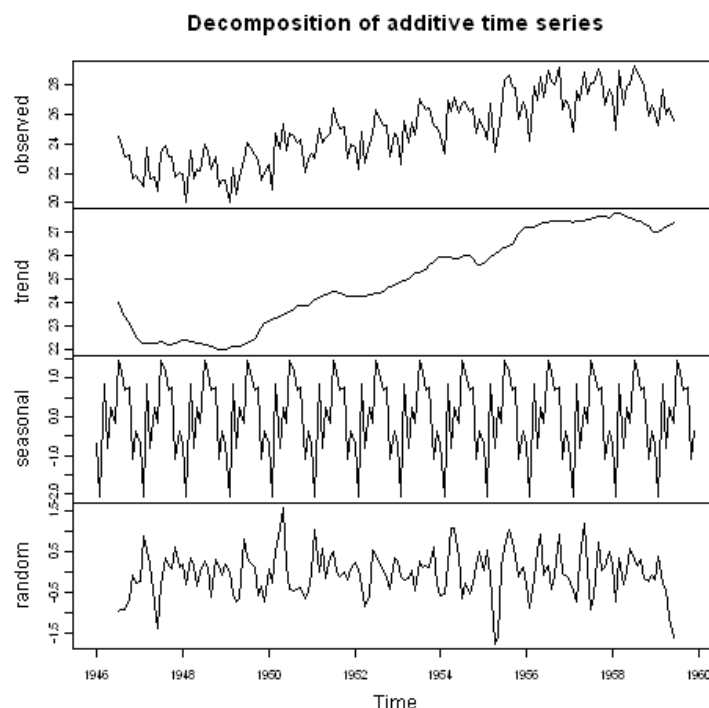
και εφαρμόζοντας άλλη μια



παρατηρούμε ότι οι τάσεις που υπήρχαν στο πρώτο Σχήμα πλέον δεν υπάρχουν και έχουμε μια χρονοσειρά χωρίς κάποια τάση.

Για την απαλοιφή της εποχικότητας ένας τρόπος είναι το να αφαιρούμε τον μέσο της εποχής από κάθε τιμή που ανήκει στην συγκεκριμένη εποχή.

Αφαιρώντας την χρονοσειρά των εποχικών μέσων από αυτή των πρώτων διαφορών έχουμε την χρονοσειρά των τυχαίων επιδράσεων, σχηματικά για ένα γενικό παράδειγμα θα είχαμε το παρακάτω



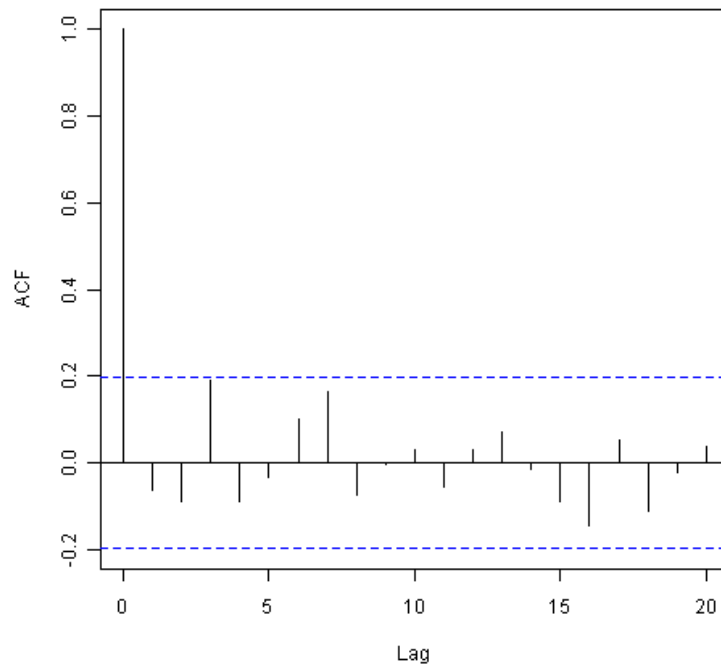
Συσχέτιση σε χρονοσειρά

Είναι λογικό να θεωρήσουμε ότι ενδέχεται μεταξύ των τιμών μια χρονοσειράς λόγω του ότι αναφέρεται στο ίδιο χαρακτηριστικό να υπάρχει ένας βαθμός συσχέτισης. Στην ανάλυση χρονοσειρών αυτό ονομάζεται αυτοσυσχέτιση μεταξύ των χρονικών στιγμών t και $t-r$ (δηλαδή χρονικές στιγμές απόστασης r) και εκτιμάται από την σχέση

$$\widehat{\rho}_r = \text{Corr}(x_t, x_{t-r}) = \frac{\sum_{t=r+1}^n (x_{t-r} - \bar{x})(x_t - \bar{x})}{\sum_{t=r+1}^n (x_t - \bar{x})^2}$$

Το παραπάνω έχει νόημα όταν η χρονοσειρά είναι στάσιμη (δηλαδή χωρίς τάση και εποχικότητα), οπότε για να ελεγχθεί η ύπαρξη αυτής πρέπει να ελέγξουμε την απαλλαγμένη από αυτά τα δύο χαρακτηριστικά χρονοσειρά.

Αφού γίνει αυτό χρησιμοποιούμε τα διαγράμματα αυτοσυσχέτισης (acf) και της μερικής αυτοσυσχέτισης (pacf) για να ελέγξουμε την ύπαρξη συσχέτισης. Η μερική αυτοσυσχέτιση είναι ένα διάγραμμα σταθμισμένο το οποίο ελέγχει καθαρά την μεταξύ συσχέτιση μεταξύ δύο χρονικών στιγμών απαλλάσσοντας τα από συσχέτιση λόγω των λοιπών χρονικών στιγμών.



Η διακεκομμένη γραμμή είναι το όριο όπου από εκεί και ύστερα θεωρούμε ότι υπάρχει αυτοσυσχέτιση.

Μοντέλα αυτοσυσχέτισης AR(Autoregressive models)

Σε αυτά τα μοντέλα θεωρούμε ότι η τιμή της χρονοσειράς στη στιγμή t εξαρτάται γραμμικά από τις προηγούμενες p τιμές. Έτσι η μορφή του μοντέλου **AR(p)** είναι της μορφής

$$x_t = c + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \dots + \varphi_p x_{t-p} + e_t$$

Με c συμβολίζεται ο σταθερός όρος και e_t ο λευκός θόρυβος (η τυχαιότητα)

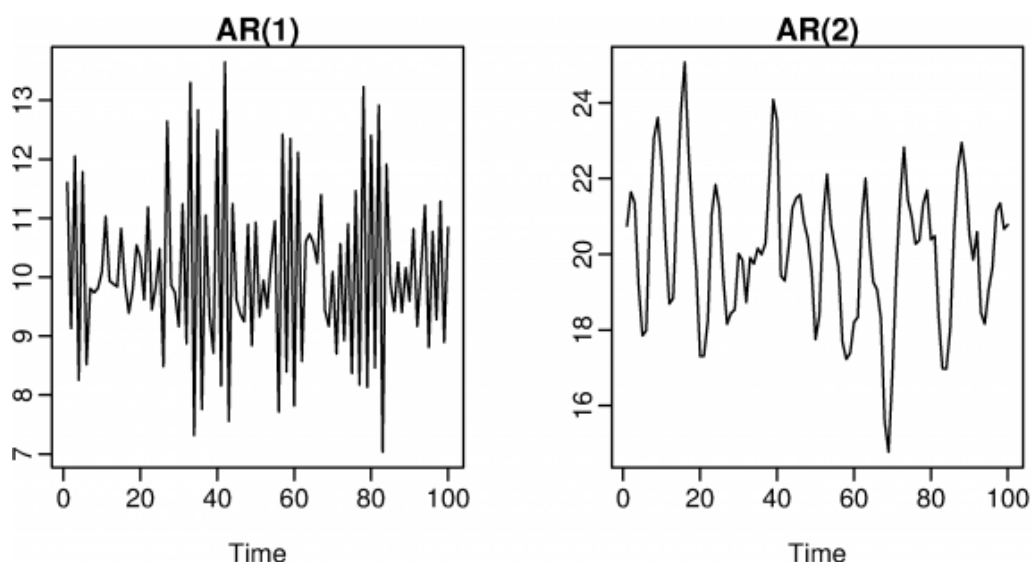
Σαν παράδειγμα για ένα AR(1)

- Όταν $\phi_1=0$, το x_t είναι λευκός θόρυβος
- όταν $\phi_1=1$ και $c=0$, η χρονοσειρά είναι τυχαίος περίπατος
- όταν $\phi_1<0$, η χρονοσειρά αλλάζει τιμές μεταξύ θετικών κι αρνητικών

Για να χρησιμοποιήσουμε τα Μοντέλα αυτοσυσχέτισης AR, πρώτα ‘φέρνουμε’ τις τιμές μας σε στάσιμη μορφή.

Υπάρχουν και κάποιο άλλοι περιορισμοί, ενδεικτικά αναφέρουμε

- για AR(1): $-1 < \phi_1 < 1$.
- για AR(2) : $-1 < \phi_2 < 1$, $\phi_1 + \phi_2 < 1$, $\phi_2 - \phi_1 < 1$.



Μοντέλα κινούμενου μέσου MA [Moving average models]

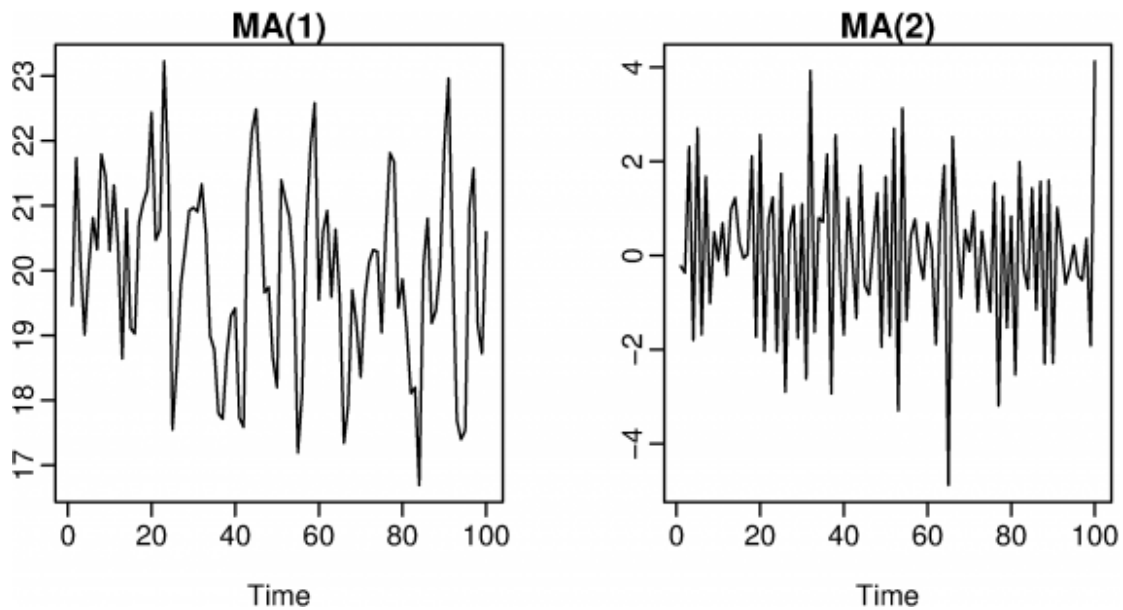
Τα Moving average μοντέλα θεωρούν ότι υπάρχει ένας μέσος ο οποίος κινείται μέσα στο χρόνο και αλλάζει από κάποιες τυχαίες επιδράσεις κάθε φορά , αυτός επηρεάζει γραμμικά το μοντέλο άρα έχει τη μορφή

$$x_t = c + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \dots + \theta_q e_{t-q}$$

Όπου e_t είναι ο λευκός θόρυβος την χρονική στιγμή t . Το παραπάνω μοντέλο είναι ένα MA(q) μοντέλο στο οποίο θεωρήσαμε ότι η χρονοσειρά επηρεάζεται από τις q προηγούμενες χρονικές στιγμές .

Γίνεται να φέρουμε κάθε AR(p) στη μορφή ενός MA(∞). Σαν παράδειγμα για ένα μοντέλο AR(1):

$$\begin{aligned} x_t &= \phi_1 y_{t-1} + e_t = \phi_1 (\phi_1 y_{t-2} + e_{t-2}) + e_t = \dots \\ &= e_t + \phi_2^2 e_{t-2} + \phi_1 e_{t-1} + \phi_3^3 e_{t-3} + \phi_4^4 e_{t-4} + \dots \end{aligned}$$



Μοντέλα αυτοσυσχέτισης κινούμενου μέσου- ARIMA

Αν συνδυαστούν τα δύο παραπάνω τότε μας δίνεται ένα μοντέλο της μορφής

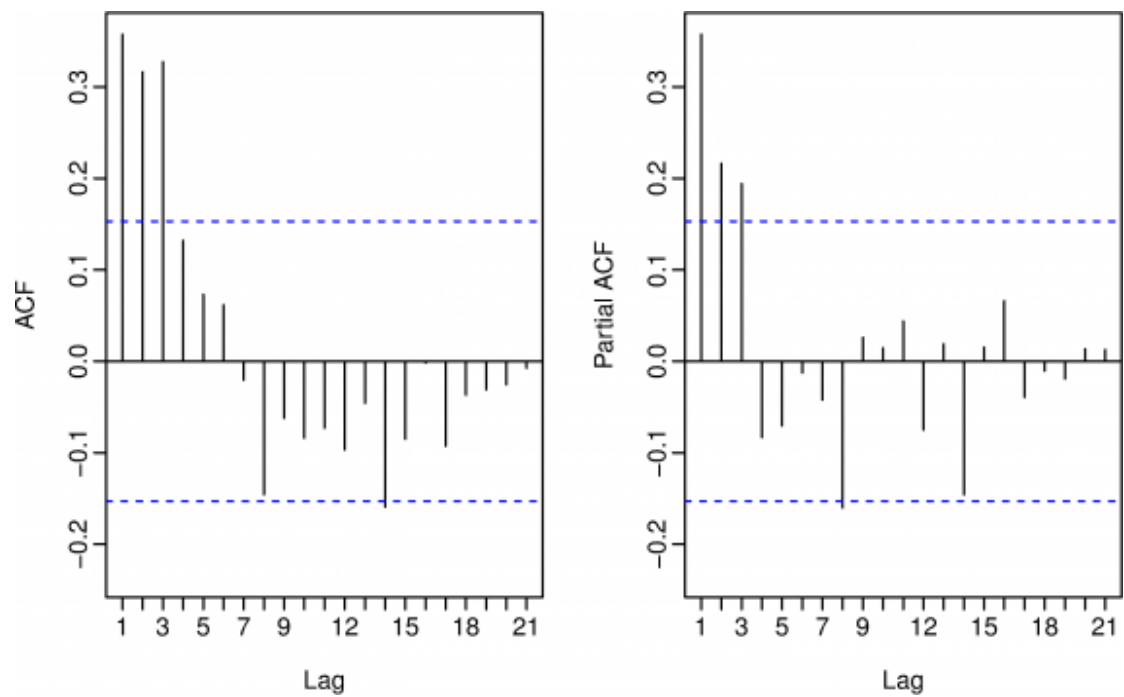
$$x'_t = c + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \dots + \theta_q e_{t-q}$$

$$\varphi_1 x'_{t-1} + \varphi_2 x'_{t-2} + \dots + \varphi_p x'_{t-p} + e_t$$

Όπου $x'_t = x_t - x_{t-1}$ (όσες φορές χρειαστεί μέχρι να έχουμε στασιμότητα).

Το μοντέλο γράφεται ως ARIMA(p,d,q) , με p,q να έχουν την σημασία που δόθηκε στα προηγούμενα δύο μοντέλα AR(p), MA(q) και το d συμβολίζει πόσες φορές έχουμε 'πάρει' την διαφορά $x_t - x_{t-1}$ μέχρι πετύχουμε στασιμότητα.

Η εύρεση των p,q είναι μια διαδικασία που θέλει εμπειρία από τον ερευνητή και κατεύθυνση σε αυτό δίνουν τα διαγράμματα acf, pacf για το q και το p αντίστοιχα.



Στην περίπτωση που έχουμε εποχιακή χρονοσειρά χρησιμοποιούμε τα εποχιακά ARIMA μοντέλα $ARIMA(p,d,q)(P,D,Q)_m$ που είναι η γενίκευση του παραπάνω. Τα P,D,Q είναι το τι συμβαίνει μέσα στους χρόνους της κάθε περιόδου εσωτερικά, με τον χρόνο κάθε περιόδου να κρατάει m χρονικές τιμές.

3. Δεδομένα Μελέτης

3.1 Εισαγωγή

Τα δεδομένα παρελήφθησαν από το τμήμα μηχανογράφησης της διεύθυνσης Εκπαίδευσης και Έρευνας του ΕΚΠΑ κατόπιν αίτησης στην γραμματεία του τμήματος Μαθηματικού ΕΚΠΑ.

Αποτελούνται κυρίως από δύο μεγάλα πακέτα σε μορφή λογιστικών φύλλων (excels). Συνοπτικά το πρώτο περιέχει πληροφορίες για όλα τα άτομα που υπήρξαν φοιτητές κατά τα έτη 2000 έως και 2015, όσον αφορά τα προσωπικά τους δεδομένα, όπως φύλο, καταγωγή, λύκειο κλπ, αλλά και τον τρόπο εγγραφής: Πανελλήνιες, Erasmusκλπ, όπως και αναγράφεται το αν κάποιος εγκατέλειψε ή είναι ανενεργός, ενώ για τους πτυχιούχους έχουμε επιπλέον πληροφορίες: Βαθμό, Κατεύθυνση κλπ.

Το δεύτερο λογιστικό φύλλο περιέχει αναλυτικά τι έκανε ο κάθε φοιτητής όσον αφορά τις δηλώσεις μαθημάτων, το αν συμμετείχε στις εξετάσεις των μαθημάτων αυτών και τη βαθμολογία του.

Όπως ήδη αναφέραμε στην εισαγωγή τα δεδομένα ήταν μέχρι ενός σημείου, μη επεξεργασμένα (raw) και για να φτάσουν σε μια αναγνωρίσιμη μορφή από τα προγραμματιστικά πακέτα ήθελε αρκετές τροποποιήσεις, καθότι και αρκετές από τις μεταβλητές των επομένων μοντέλων που θα περιγραφθούν στα επόμενα κεφάλαια έγιναν από συμπτύξεις και συμπεράσματα από τα επάνω δεδομένα, π.χ. για να βρεθεί ο τόπος καταγωγής έπρεπε να γίνει σύγκριση των παραμέτρων: Πόλη, Λύκειο, Εθνικότητα, Υπηκοότητα.

Τέλος να αναφερθεί ότι παντού δίνεται μόνο ο αριθμός μητρώου του φοιτητή, πουθενά δεν υπάρχουν ονοματεπώνυμα των φοιτητών, για την διασφάλιση των προσωπικών δεδομένων.

3.2 Δεδομένα εργασίας

3.2.1 Περιγραφή πρωτογενών –ακατέργαστων δεδομένων-Περιγραφή Δεδομένων

Περιγραφή πρώτου πακέτου με δεδομένα

Το

πρώτο πακέτο με δεδομένα (excel) για τους φοιτητές του τμήματος κατά τα έτη 2000-2015 περιείχε τα παρακάτω στοιχεία σε στήλες :

1. Αρ.Μητρώου
2. Φύλο
3. Ημ/νια Γέννησης
4. Βαθμός απολυτηρίου
5. Έτος απολυτηρίου
6. Κωδικός Πανελλαδικών
7. Μονάδες εισαγωγής
8. Κωδικός Λυκείου
9. Λύκειο
10. Σημειώσεις
11. Εθνικότητα
12. Υπηκοότητα
13. Χώρα Γέννησης
14. Μόνιμη Διεύθυνση
15. Πόλη
16. Κωδικός Εισαγωγής
17. Έτος εγγραφής
18. Εξάμηνο Εγγραφής
19. Σπουδαστική κατάσταση
20. Ημ/νια τελευταίας σπουδαστικής μεταβολής
21. STU_ACTIVE
22. Εξάμηνο φοιτητή
23. Ημ/νια ανακήρυξης
24. Βαθμός πτυχίου
25. Ακαδημαϊκό έτος Πτυχίου
26. Κατεύθυνση
27. Σύνολο προβιβάσιμων μαθημάτων
28. Έτος Τελευταίας Κίνησης (Δήλωση ή εξέταση)

Πιο αναλυτικά

1.Αριθμός μητρώου:Ο Α.Μ. κάθε φοιτητή. Συνολικά έχουμε**6624 φοιτητές**.

2. Φύλο: Το φύλο κάθε φοιτητή, εκ των οποίων το 44% είναι κορίτσια (2877)

3.Ημερομηνία Γέννησης: Ο γηραιότερος γεννήθηκε το 1952 και ο νεότερος το 1998.

4.Βαθμός Απολυτηρίου: Ο βαθμός με τον οποίο οι φοιτητές τελείωσαν το λύκειο, Δυστυχώς περιέχει πάρα πολλές ελλείψεις καθώς άλλες φορές συμπληρώνεται αλλά άλλες φορές όχι.

5. Έτος απολυτηρίου: Η χρονιά που ο φοιτητής τελείωσε το λύκειο, μικρότερη τιμή είναι το 1980.

6.Κωδικός Πανελλαδικών:Κωδικοποίηση που αφορά τον τρόπο εισαγωγής στο Τμήμα του Μαθηματικού.

7.Μονάδες εισαγωγής: Τα μόρια των πανελλαδικών, περιέχει πάρα πολλές ελλείψεις

8.Κωδικός Λυκείου:Κάποια κωδικοποίηση των λυκείων

9.Λύκειο: Το λύκειο που αποφοίτησε ο φοιτητής

10.Σημειώσεις: Ειδικές σημειώσεις για κάθε φοιτητή πχ Εράσμους, (συνήθως δεν συμπληρώνεται)

11.Εθνικότητα(με τον όρο εθνικότητα δηλώνουμε σε ποιο έθνος (και όχι κράτος) ανήκουμε.)

Ελλάδα	6113
Κενό	257
Κύπρος	172
Λοιπές Χώρες	27
Αλβανία	54

12.Υπηκοότητα (Η υπηκοότητα, σύμφωνα με την ελληνική νομοθεσία, είναι η νομική σχέση ή δεσμός που έχει ένα άτομο με τη χώρα στην οποία ανήκει, στη χώρα δηλαδή τς οποίας είναι υπήκοος.)

Ελλάδα	6034
Κενό	226

Κύπρος	226
Λοιπές Χώρες	27
Αλβανία	61

13.Χώρα Γέννησης

Ελλάδα	6113
Κενό	155
Κύπρος	242
Λοιπές Χώρες	162
Αλβανία	72

Στις λοιπές χώρες έχουμε τις: Αίγυπτος, Αγγλία, Γερμανία, Βουλγαρία, Ιταλία, Καζαχστάν, Κίνα και Σουηδία.

11.Μόνιμη Διεύθυνση: Η διεύθυνση μόνιμης κατοικίας. Προσοχή, σημαντικό είναι ότι πέρα από τις ελλείψεις κάποιοι φοιτητές ενδέχεται να δηλώνουν όχι το 'πατρικό' τους σπίτι αλλά το σπίτι που ενοικίασαν στην Αθήνα.

15. Πόλη: Σε συνέχεια του προηγούμενου

16.Κωδικός Εισαγωγής:

	Τι σημαίνει	Συχνότητα
ΑΓΝ	ΑΓΝΩΣΤΟΣ ΚΩΔΙΚΟΣ	0
ΑΔΣ	ΑΠΟΦΑΣΗ Δ.Σ. ΤΟΥ ΤΜΗΜΑΤΟΣ	0
ΑΘΑ	ΩΣ ΑΘΛΗΤΗΣ	15
ΑΘΓ	ΩΣ ΑΘΛΗΤΡΙΑ	7
ΑΘΛ	ΑΘΛΗΤΕΣ	0
ΑΚΑ	ΑΛΛΟΓΕΝΕΙΣ ΑΛΛΟΔΑΠΟΙ ΚΟΙΝΟΤ. ΥΠΟΤΡΟΦΟΙ	0
ΑΛΑ	ΩΣ ΑΛΛΟΔΑΠΟΣ	2
ΑΛΓ	ΩΣ ΑΛΛΟΔΑΠΗ	6
ΑΛΕ	ΑΠΟΦΟΙΤΟΙ ΛΥΚΕΙΩΝ ΕΞΩΤΕΡΙΚΟΥ	5
ΑΛΚ	ΑΛΛΑΓΗ ΚΑΤΕΥΘΥΝΣΗΣ	0
ΑΛΣ	ΑΛΛΟΔΑΠΟΣ(-Η) Σ.Α.Ν.	0
ΑΜΕ	ΑΝΑΔΡ.ΕΓΚΡ. ΜΕΤΕΓΓΡΑΦΗΣ ΑΠΟ Γ.Σ.	1
ΑΥΑ	ΩΣ ΑΛΛΟΔΑΠΟΣ ΥΠΟΤΡΟΦΟΣ	1
ΑΥΓ	ΩΣ ΑΛΛΟΔΑΠΗ ΥΠΟΤΡΟΦΟΣ	14
ΔΘΑ	ΔΙΕΘΝΕΙΣ ΑΝΤΑΛΛΑΓΕΣ	0
ΔΙΚ	ΚΑΤΑΤΑΞΗ ΔΙΚΑΤΣΑ - ΔΟΑΤΑΠ	0
ΔΙΚ2	ΚΑΤΑΤΑΞΗ ΔΙΚΑΤΣΑ	0
ΕΑΦ	ΕΠΑΝΕΓΓΡΑΦΗ ΜΕΤΑ ΑΠΟ ΑΠΟΦΟΙΤΗΡΙΟ	2
ΕΓΕ	ΕΝΤΑΞΗ ΣΤΟ Γ ΕΞΑΜΗΝΟ (ΑΠΟΦ.ΣΥΓΚΛΗΤΟΥ)	0
ΕΓΣ	ΕΠΑΝΕΓΓΡΑΦΗ (ΑΠΟΦ. ΓΕΝ. ΣΥΝΕΛΕΥΣΗΣ)	0
ΕΕΠ	ΕΝΤΑΞΗ ΑΠΟ ΕΙΔΙΚΑ ΠΡΟΓΡΑΜΜΑΤΑ	0
ΕΙΔ	ΕΙΔΙΚΕΣ ΑΝΑΓΚΕΣ	0
ΕΙΠ	ΕΙΔΙΚΕΣ ΠΕΡΙΠΤΩΣΕΙΣ	0

ΕΚΛ	ΑΝΩΤΕΡΕΣ ΕΚΚΛΗΣΙΑΣΤΙΚΕΣ ΣΧΟΛΕΣ	0
ΕΞΑ	ΩΣ ΕΛΛΗΝΑΣ ΕΞΩΤΕΡΙΚΟΥ	0
ΕΞΓ	ΩΣ ΕΛΛΗΝΙΔΑ ΕΞΩΤΕΡΙΚΟΥ	3
ΕΞΟ	ΕΞΟΜΟΙΩΣΗ ΠΤΥΧΙΟΥ ΠΑΙΔΑΓΩΓΙΚΩΝ ΑΚΑΔΗΜΙΩΝ	0
ΕΠΑ	ΚΑΘ' ΥΠΕΡ/Φ.152/Β6/198/00	0
ΕΠΒ	ΚΑΘ' ΥΠΕΡΒ.Φ243/Β6/755/00	0
ΕΠΓ	ΚΑΘ' ΥΠΕΡΒ.Φ151/20049/Β6/272/07	0
ΕΠΔ	ΚΑΘ' ΥΠΕΡΒ.Φ151/17104/Β6/259/06	0
ΕΠΕ	ΥΠΟΥΡΓΙΚΗ ΑΠΟΦΑΣΗ Φ.253/Β6/755/30-5-2000	0
ΕΠΖ	ΥΠΟΥΡΓΙΚΗ ΑΠΟΦΑΣΗ ΕΣΩΤ.Φ.25225/Β6/200/6-4-00	0
ΕΠΗ	ΚΑΘ' ΥΠΕΡΒ.Φ151/20049/Β6 ΦΕΚ 156-Α/04.09.09	8
ΕΠΘ	ΥΠΟΥΡΓΙΚΗ ΑΠΟΦΑΣΗ Φ.152/Β6/198/2000Ν.3794/4.9.2009	2
ΕΠΞ	ΕΠΑΝΕΝΤΑΞΗ ΣΕ ΝΕΟ ΠΡΟΓΡΑΜΜΑ (ΕΞΑΜΗΝΑ)	3
ΕΠΥ	ΚΑΘ' ΥΠΕΡΒΑΣΗ 5%	3
ΚΑΕ	ΚΑΤΑΤΑΞΗ ΠΤΥΧΙΟΥΧΟΥ ΑΕΙ	46
ΚΑΤ	ΚΑΤΑΤΑΚΤΗΡΙΕΣ	0
ΚΒΑ	ΚΑΤΑΤΑΞΗ ΜΕ BACHELOR	0
ΚΔΕ	ΚΑΤΑΤΑΞΗ ΔΙΕΤΟΥΣ	4
ΚΔΙ	ΚΑΤΑΤΑΞΗ ΔΙ.ΚΑ.ΤΣΑ.	1
ΚΜΑ	ΚΑΤΑΤΑΞΗ ΜΕ MASTER	0
ΚΞΠ	ΚΑΤΟΧΟΣ ΤΙΤΛΩΝ ΣΠΟΥΔΩΝ ΞΕΝΩΝ ΠΑΝΕΠΙΣΤΗΜΙΩΝ	0
ΚΟΑ	ΚΟΙΝΟΤΙΚΟΙ ΑΛΛΟΔΑΠΟΙ	0
ΚΠΤ	ΚΑΤΑΤΑΞΗ ΠΤΥΧΙΟΥΧΟΥ ΑΕΙ-ΤΕΙ	23
ΚΤΔ	ΚΑΤΑΤΑΞΗ ΠΤΥΧΙΟΥΧΟΥ ΔΡΑΜΑΤΙΚΗΣ ΣΧΟΛΗΣ	0
ΚΤΕ	ΚΑΤΑΤΑΞΗ ΠΤΥΧΙΟΥΧΟΥ ΤΕΙ	5
ΚΥΑ	ΩΣ ΚΥΠΡΙΟΣ	89
ΚΥΓ	ΩΣ ΚΥΠΡΙΑ	127
ΚΥΠ	ΚΥΠΡΙΟΙ	14
ΚΩΦ	ΚΩΦΑΛΛΟΙ	0
ΛΥΓ	ΛΟΓΟΙ ΥΓΕΙΑΣ - ΑΝΕΥ ΕΞΕΤΑΣΕΩΝ	0
ΜΑΒ	ΔΙΑΚΡΙΣΗ ΣΕ ΕΠΙΣΤΗΜΟΝΙΚΕΣ ΟΛΥΜΠΙΑΔΕΣ	1
ΜΑΓ	ΜΕΤΕΓΓΡΑΦΗ ΛΟΓΩ ΑΠΩΛΕΙΑΣ ΓΟΝΕΑ	0
ΜΑΘ	ΜΑΘΗΜΑΤΙΚΗ ΟΛΥΜΠΙΑΔΑ	0
ΜΑΝ	ΜΕΣΟΓΕΙΑΚΗ ΑΝΑΙΜΙΑ	0
ΜΕΑ	ΜΕΤΕΓΓΡΑΦΗ ΑΘΛΗΤΗ	3
ΜΕΞ	ΜΕΤΕΓΓΡΑΦΗ ΕΞΩΤΕΡΙΚΟΥ	1
ΜΕΣ	ΜΕΤΕΓΓΡΑΦΗ ΕΞΩΤΕΡΙΚΟΥ	336
ΜΕΤ	ΑΠΟ ΜΕΤΕΓΓΡΑΦΗ	1
ΜΘΓ	ΜΕΤΕΓΓΡΑΦΗ ΛΟΓΩ ΑΣΘΕΝΕΙΑΣ ΓΟΝΕΑ	11
ΜΘΡ	ΜΟΥΣΟΥΛΜΑΝΟΙ ΘΡΑΚΗΣ	11
ΜΙΣ	ΜΕΤΕΓΓΡΑΦΗ ΙΣΟΒΑΘΜΙΣΑΝΤΩΝ	0
ΜΚΥ	ΜΕΤΕΓΓΡΑΦΗ ΑΠΟ ΚΥΠΡΟ	0
ΜΛΑ	ΜΕΤΕΓΓΡΑΦΗ ΛΟΓΩ ΦΟΙΤΟΥΝΤΟΣ ΑΔΕΛΦΟΥ	0
ΜΛΓ	ΜΕΤΕΓΓΡΑΦΗ ΛΟΓΩ ΦΟΙΤΟΥΝΤΟΣ ΓΟΝΕΑ	0
ΜΛΘ	ΜΕΤΕΓΓΡΑΦΗ ΛΟΓΩ ΑΣΘΕΝΕΙΑΣ	13
ΜΛΜ	ΜΕΤΕΓΓΡΑΦΗ ΩΣ ΜΗΤΕΡΑ ΑΝΗΛΙΚΩΝ	0
ΜΛΣ	ΜΕΤΕΓΓΡΑΦΗ ΛΟΓΩ ΦΟΙΤΟΥΝΤΟΣ ΣΥΖΥΓΟΥ	0

ΜΟΡ	ΜΕΤΕΓΓΡΑΦΗ ΒΑΣΕΙ Ν3282/04 (ΩΣ ΟΡΦΑΝΟ ΤΕΚΝΟ)	68
ΜΤΘ	ΜΕΤΑΦΟΡΑ ΘΕΣΗΣ	276
ΜΤΛ	ΜΕΤΕΓΓΡΑΦΗ ΩΣ ΤΕΚΝΟ ΠΟΛΥΜΕΛΟΥΣ ΟΙΚΟΓΕΝΕΙΑΣ	118
ΜΤΠ	ΜΕΤΕΓΓΡΑΦΗ ΩΣ ΤΕΚΝΟ ΠΟΛΥΤΕΚΝΩΝ	130
Ν2640	Ν.2640/98	16
Ν3763	Ν.3763/2009	0
Ν3794	Ν.3794/09	0
ΟΜΟ	ΩΣ ΟΜΟΓΕΝΗΣ	4
ΟΜΥ	ΩΣ ΟΜΟΓΕΝΗΣ ΥΠΟΤΡΟΦΟΣ	1
ΠΑΔΕ	ΠΑΝΕΛΛΗΝΙΕΣ ΕΞΕΤΑΣΕΙΣ (ΑΠΟΦΑΣΗ ΔΙΟΙΚΗΤΙΚΟΥ ΕΦΕΤΕΙΟΥ ΑΘΗΝΩΝ)	0
ΠΑΚ	ΠΑΙΔΑΓΩΓΙΚΕΣ ΑΚΑΔΗΜΙΕΣ	0
ΠΑΝ	ΠΑΝΕΛΛΗΝΙΕΣ ΕΞΕΤΑΣΕΙΣ	4699
ΠΑΣ	ΠΑΝΕΛΛΗΝΙΕΣ ΕΞΕΤΑΣΕΙΣ (ΑΠΟΦΑΣΗ ΣΥΓΚΛΗΤΟΥ)	4
ΠΚΚ	ΠΑΝΕΛΛΗΝΙΕΣ ΕΙΔ.ΠΕΡΙΠΤ.-ΚΟΙΝΩΝ. ΚΡΙΤΗΡΙΑ	35
ΠΠΟ	ΠΑΝΕΛΛΗΝΙΕΣ ΕΙΔΙΚΕΣ ΠΕΡΙΠΤΩΣΕΙΣ-ΠΟΛΥΤΕΚΝΟΙ	66
ΠΣΕ	ΕΓΓΡΑΦΗ ΩΣ ΠΡΩΗΝ Π.Σ.Ε.	0
ΠΤΡ	ΠΑΝΕΛΛΗΝΙΕΣ ΕΙΔΙΚΕΣ ΠΕΡΙΠΤΩΣΕΙΣ-ΤΡΙΤΕΚΝΟΙ	56
ΣΑΝ	ΑΠΟ ΣΧΟΛΗ ΑΞΙΩΜΑΤΙΚΩΝ ΝΟΣΗΛΕΥΤΙΚΗΣ	0
ΣΕΙ	ΣΕΙΣΜΟΠΛΗΚΤΟΣ	0
ΣΤΡ	ΣΤΡΑΤΙΩΤΙΚΕΣ ΣΧΟΛΕΣ	3
ΤΕΑ	ΤΕΚΝΑ ΑΝΑΠΗΡΩΝ	0
ΤΕΞ	ΤΕΚΝΑ ΕΛΛΗΝΩΝ ΕΞΩΤΕΡΙΚΟΥ	20
ΤΕΠ	ΤΕΚΝΑ ΠΟΛΥΤΕΚΝΩΝ	157
ΤΠΛ	ΤΕΚΝΑ ΠΟΛΥΜΕΛΩΝ ΟΙΚΟΓΕΝΕΙΩΝ	213
ΤΥΞ	ΤΕΚΝΑ ΕΛΛΗΝΩΝ ΥΠΑΛΛΗΛΩΝ ΤΟΥ ΕΞΩΤΕΡΙΚΟΥ	0
ΤΥΦ	ΤΥΦΛΟΙ	2
ΥΑΑ	ΥΠΟΤΡΟΦΟΣ ΑΚΑΔΗΜΙΑΣ ΑΘΗΝΩΝ	0
ΥΙΚ	ΩΣ ΥΠΟΤΡΟΦΟΣ Ι.Κ.Υ.	0
ΥΥΕ	ΩΣ ΥΠΟΤΡΟΦΟΣ ΥΠ.ΕΞ.	0
ΥΥΠ	ΩΣ ΥΠΟΤΡΟΦΟΣ Υ.Π.Ε.Π.Θ	0
ΦΕΠ	ΦΟΙΤΗΤΕΣ ΕΙΔΙΚΩΝ ΠΡΟΓΡΑΜΜΑΤΩΝ	0
ΦΠΨ	ΕΝΤΑΞΗ ΑΠΟ Φ.Π.Ψ.	0
ΦΣΓ	ΕΚ ΤΟΥ ΚΑΤΑΡΓΗΘΕΝΤΟΣ ΦΥΣΙΟΓΝΩΣΤΙΚΟΥ	0

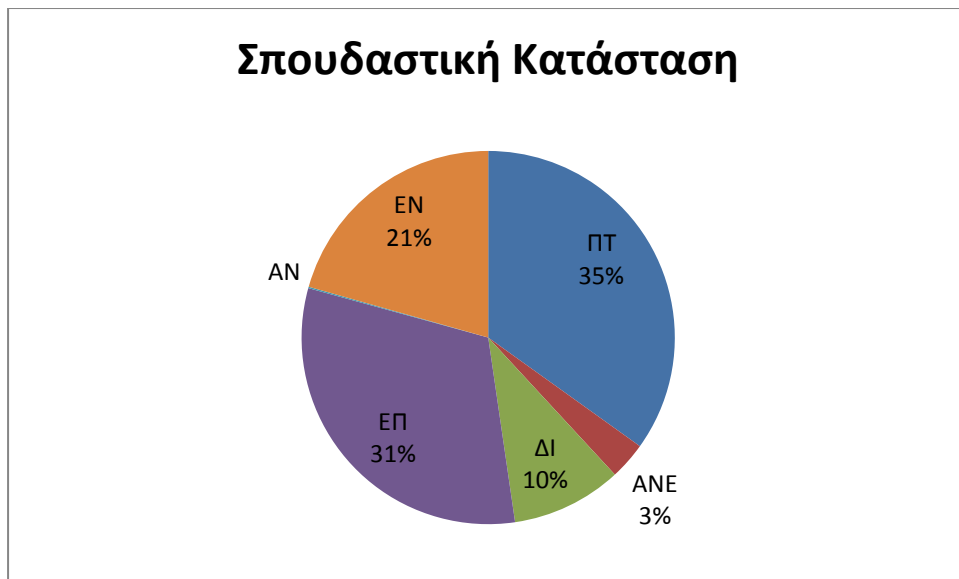
17.Έτος εγγραφής: από το 2000 έως το 2015

18.Εξάμηνο Εγγραφής: Από το 1 (6402φοιτητές) έως το 9

19.Σπουδαστική κατάσταση:

- ΑΝ:Σε αναστολή 8
- ΑΝΕ:Ανενεργός 214
- ΔΙ:Διακοπή 634
- ΕΝ:Ενεργός 1365

- ΕΠ: Επί πτυχίο 2091
- ΠΤ: πτυχιούχος 2311



20.Ημερομηνία τελευταίας σπουδαστικής μεταβολής: Πότε ήταν η τελευταία φορά που ο φοιτητής έκανε οποιαδήποτε μεταβολή στην γραμματεία (δήλωσε μαθήματα, έδωσε μάθημα ή κάποια άλλη κίνηση).

21.STU_ACTIVE: Κωδικοποίηση αν ο φοιτητής είναι ενεργός ή όχι, 0-1.

22.Εξάμηνο φοιτητή: Σε περίπτωση που είναι πτυχιούχος είναι το εξάμηνο που ολοκλήρωσε τις σπουδές του

23. Ημερομηνία ανακήρυξης: Πότε ανακηρύχτηκε πτυχιούχος ο φοιτητής

24.Βαθμός πτυχίου: Μόνο για τους έχοντες πτυχίο

25.Ακαδημαϊκό έτος Πτυχίου

26.Κατεύθυνση: Μόνο για τους πτυχιούχους

27.Σύνολο προβιβάσιμων μαθημάτων: Υπάρχει άτομο με περασμένα 60 μαθήματα!

28.Έτος Τελευταίας Κίνησης (Δήλωση ή εξέταση)

Περιγραφή δευτέρου πακέτου με δεδομένα

Το δεύτερο πακέτο με δεδομένα (δευτερο excel) , το οποίο συνολικά έχει 858.681 εγγραφές, περιέχει το τι έκανε ο κάθε φοιτητής ξεχωριστά στην έως την ημερομηνία που πήραμε τα δεδομένα από το σύστημα. Ο χρόνος που αποτελούν αυτά τα δεδομένα είναι από το ΔΗΛ(Χ) 2000-2001 έως το ΕΞ(Χ) 2015-2016, δηλαδή δήλωση Χειμερινού του 2000-01 έως Χειμερινές Εξετάσεις του 2015-2016. Δηλαδή μπορείς να εντοπίσεις τι δήλωσε ο φοιτητής σε κάθε εξάμηνο, αν έδωσε ένα μάθημα και τι βαθμό πήρε.

Συγκεκριμένα οι τιμές είναι οι παρακάτω:

1. **Ακαδημαϊκό Έτος:** Σε ποιο έτος έκανε τι
2. **Κωδικός περιόδουδήλωσης/εξεταστικής**
3. **KAR_RECORDTYPE:** Κάποιο είδος κωδικοποίησης για χρήση
4. **AM:** Αριθμός Μητρώου Φοιτητή
5. **Εξάμηνο μαθήματος:** Το εξάμηνο που ανήκει το κάθε μάθημα με βάση τον οδηγό σπουδών
6. **Μάθημα:** κωδικός μαθήματος
7. **Βαθμός:** Υπάρχει μόνο αν έδωσε το μάθημα

3.3 Μια πρώτη ανάλυση (περιγραφική) πάνω στην σταδιοδρομία των φοιτητών

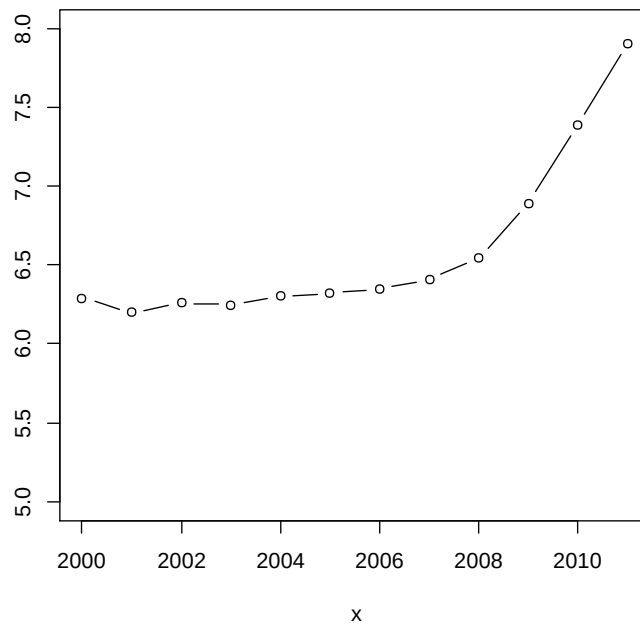
Αρχικά θα προσπαθήσουμε να βγάλουμε κάποια πρώτα συμπεράσματα, όπως και ενδιαφέρουσες παρατηρήσεις μέσα από τα δύο αυτά πακέτα δεδομένων με χρήση Περιγραφικής Στατιστικής. Αργότερα στην συνέχεια της εργασίας θα γίνει χρήση συγκεκριμένων μοντέλων.

3.3.1 Μέσος όρος απόκτησης πτυχίου ανά έτος εισαγωγής και πλήθος πτυχιούχων και επι πτυχίω

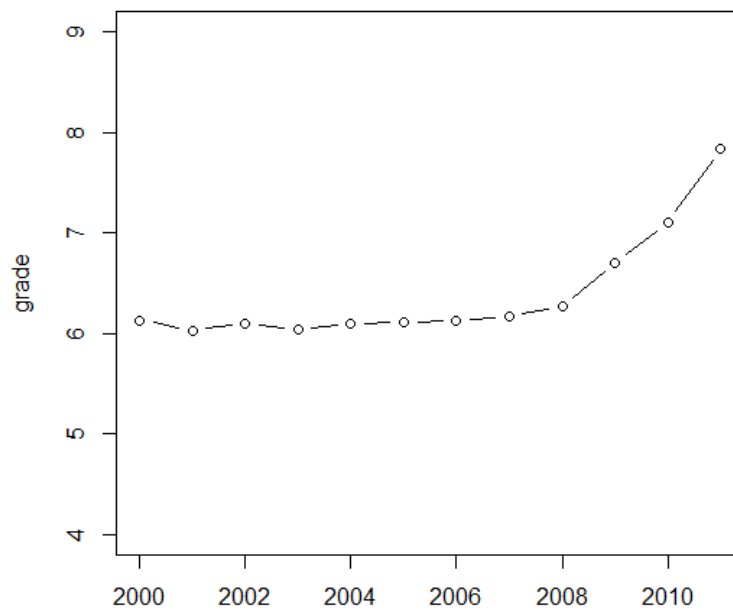
Ακ. Έτος	Μέσος Όρος πτυχίου	Διάμεσος	Σύνολο Πτυχιούχων	Επί πτυχίω έως και το 2015
2000	6.290373	6,13	295	86
2001	6.199746	6.03	315	101
2002	6.258731	6.10	268	159
2003	6.246782	6.04	261	138
2004	6.305093	6.09	216	144
2005	6.319524	6.11	231	163
2006	6.348358	6.12	201	140
2007	6.407069	6.17	174	171
2008	6.544783	6.27	138	204
2009	6.894831	6.70	89	264
2010	7.388427	7.10	89	257
2011	7.907429	7.84	35	257

Στον παραπάνω πίνακα έχουμε τον μέσο όρο πτυχίου για κάθε έτος εισαγωγής, το σύνολο πτυχιούχων και το σύνολο των ατόμων που δεν έχουν ολοκληρώσει ακόμα τις σπουδές τους. Όσον αφορά το μέσο όρο υπάρχει μια μικρή βελτίωση χρόνο με το χρόνο. Η παρατήρηση αυτή επιβεβαιώνεται και από τη διάμεσο. Η αισθητή βελτίωση κατά τα έτη 2009-2011 είναι λόγω του ότι το σύνολο των Πτυχιούχων περιέχει κυρίως τα άτομα που ολοκληρώνουν τις σπουδές τους 'γρήγορα' και συνήθως αυτοί πηγαίνουν και καλά βαθμολογικά. Από την πρώτη και την δεύτερη στήλη έχουμε τα επόμενο γραφήματα:

Mean grade



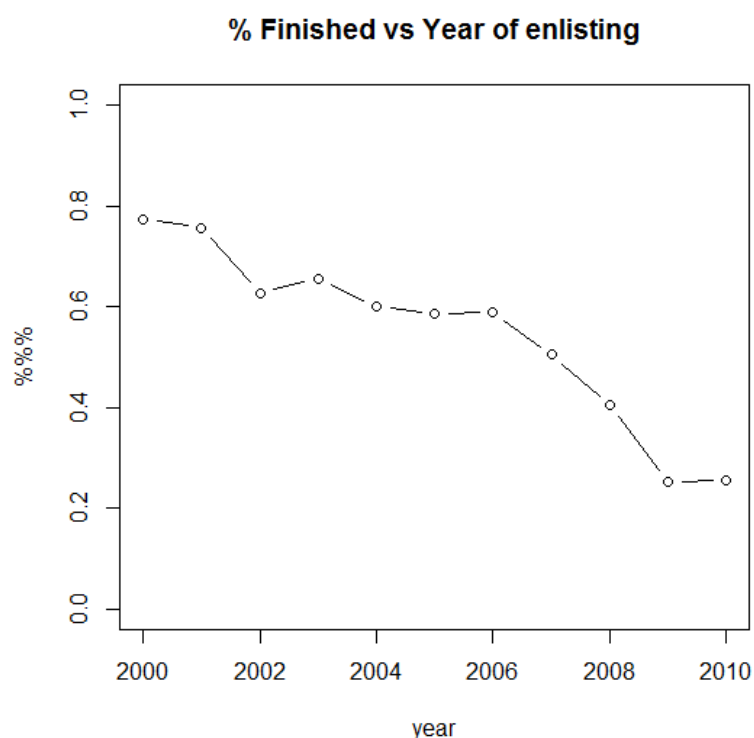
Median Grade



3.3.2 Ποσοστά Πτυχιούχων επί το σύνολο Πτυχιούχοι- Επί Πτυχίω

Όσον αφορά τα ποσοστά των πτυχιούχων στο σύνολο Πτυχιούχοι – Επί Πτυχίω έχουμε τα παρακάτω, επί τις 100 (δεν περιλαμβάνει όσους έχουν διαγραφεί ή είναι Ανενεργοί).

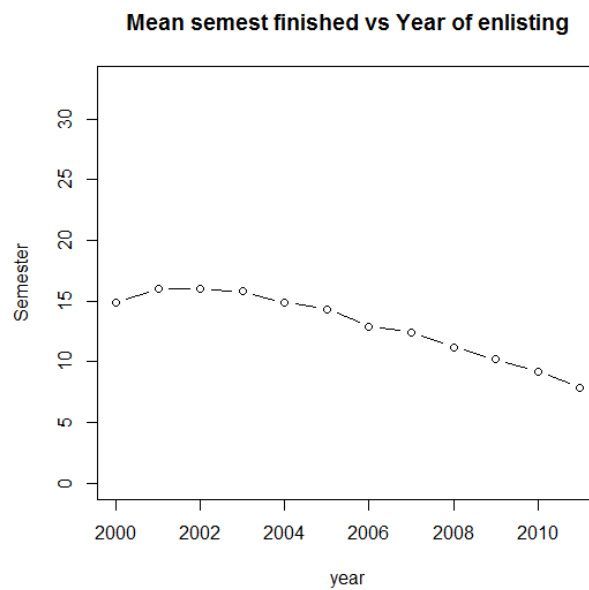
2000	77.4%
2001	75.7%
2002	62.7%
2003	65.4%
2004	60,0%
2005	58.6%
2006	58.9%
2007	50.4%
2008	40.6%
2009	25.2%
2010	25.7%
2011	12.3%



Το παραπάνω διάγραμμα αναφέρει τα ποσοστά πτυχιούχων έναντι του συνόλου πτυχιούχων και επί πτυχίω, σε σχέση με το πότε εισήχθησαν στην σχολή. Είναι λογικό τα τελευταία χρόνια να είναι λιγότεροι αφού ένα μεγάλο ποσοστό δεν έχει προλάβει να ολοκληρώσει τις σπουδές τους σε σχέση με τους λοιπούς.

3.3.3 Μέσο εξάμηνο αποφοίτησης φοιτητών ανά έτος εισαγωγής.

Τέλος ας δούμε και το μέσο εξάμηνο που αποφοίτησαν οι φοιτητές κάθε έτους εισαγωγής:



Στο οποίο και παρατηρούμε μια βελτίωση όσον αφορά τον χρόνο που ολοκληρώνει ο φοιτητής και όχι μόνο τα τελευταία χρόνια, που έτσι κι αλλιώς μόνο λίγοι έχουν ολοκληρώσει.

3.4 Πίνακες επίδοσης στις εξεταστικές

Στους επόμενους πίνακες θα αναφερθούμε ενδεικτικά στο τι συμβαίνει ενδεικτικά στα τέσσερα Υποχρεωτικά μαθήματα Απειροστικός Λογισμός (101), Πιθανότητες (241), Στατιστική (541) και Πραγματική Ανάλυση (401), ενώ για όλα τα υπόλοιπα μαθήματα μπορούμε να βγάλουμε παρόμοια αποτελέσματα αλλά δεν θα τα αναφέρουμε ως πλεονασμό για το τι μελετά η παρούσα διπλωματική.

Όπου υπάρχει το σύμβολο ‘–’ δεν είχαμε προσφερόμενο το μάθημα το συγκεκριμένο χρονικό διάστημα, ενώ με ‘*’ υπάρχει έλλειψη στοιχείων.

Όπου λείπουν ολόκληρα φοιτητικά έτη σημαίνει πως το μάθημα είχε άλλη κωδικοποίηση και ενδεχομένως διαφορετική διδαχθήσα ύλη.

Απειροστικός Λογισμός 1

	Δήλωσε Χειμερ.	Έδωσε	Πέρασε	Δήλωσε Εαρινό	Έδωσε	Πέρασε	Σεπτέμβρης εδωσε	Πέρασε	Σύνολο ≥ 5
2000	433	247	14	-	-	-	112	23	37
2001	766	504	177	-	-	-	190	22	199
2002	954	504	69	-	-	-	313	67	136
2003	1169	724	128	-	-	-	384	88	216
2004	1317	701	120	-	-	-	412	119	239
2005	1424	750	183	-	-	-	388	80	263
2006	1424	691	184	-	-	0	428	103	287
2007	1434	824	227	903	497	138	341	54	419
2008	1214	497	174	820	401	53	288	94	321
2009	1243	686	140	825	369	65	309	50	255
2010	1281	574	113	759	264	61	204	25	199
2011	1244	633	198	658	220	32	222	38	268
2012	1183	619	197	*	192	18	329	86	301

2013	1095	517	147	*	190	75	301	53	275
2014	1200	626	238	*	256	67	256	44	349
2015	1041	432	98	*	*	*	*	*	98

Πιθανότητες 1

	Δήλωσε	Έδωσε	Πέρασε	Δήλωσε	Έδωσε	Πέρασε	Σεπτέμβρης έδωσε	Πέρασε	Σύνολο >5
2001	363	180	41	-	-	-	136	40	81
2002	667	302	83	-	-	-	210	99	182
2003	810	319	62	-	-	-	253	85	147
2004	993	401	111	-	-	-	293	162	273
2005	1039	427	81	-	-	-	331	43	124
2006	1305	450	139	-	-	-	351	95	234
2007	1288	592	126	-	-	-	345	106	232
2008	1273	517	144	780	287	91	286	114	349
2009	1237	455	86	832	363	159	228	64	309
2010	1124	320	125	658	219	38	177	100	263
2011	1068	382	143	604	179	72	163	33	248
2012	864	269	117	*	288	108	231	49	274
2013	771	248	83	*	288	41	247	152	276
2014	896	325	91	*	144	28	176	61	180
2015	920	254	96	*	*	*	*	*	96

Στατιστική 1

	Δήλωσε	Έδω σε	Πέρασ ε	Δήλωσε	Έδωσε	Πέρασε	Σεπτέμβρης Έδωσε	Πέρασε	Σύνολ ο >5
2002	161	45	22	-	-	-	35	16	38
2003	366	133	47	4	1	0	89	38	85
2004	528	185	59	133	57	19	144	50	128
2005	763	250	102	272	110	52	125	42	196
2006	900	221	115	178	0	0	145	49	164
2007	1019	299	86	453	257	88	238	87	261
2008	922	315	82	680	196	47	244	76	205
2009	1063	327	124	664	223	114	202	59	297
2010	974	297	92	746	332	149	141	42	283
2011	802	221	67	810	409	138	212	65	270
2012	812	249	73	*	263	101	216	62	236
2013	611	178	96	*	442	166	270	104	366
2014	713	239	57	*	329	106	233	99	262
2015	645	208	108	*	*	*	*	*	108

Πραγματική Ανάλυση

	Δήλωσε	Έδωσε	Πέρασε	Δήλωσε	Έδωσε	Πέρασε	Σεπτέμβρης Έδωσε	Πέρασε	Σύνολο >5
2002	-	-	-	591	153	41	53	27	68
2003	-	-	-	663	218	44	118	31	75
2004	176	94	34	775	302	107	157	27	168
2005	306	129	42	972	231	26	178	39	107
2006	495	271	114	953	148	24	187	30	168
2007	836	298	68	995	305	36	256	40	144
2008	862	249	58	1199	467	146	279	55	259
2009	920	233	35	1124	455	164	274	66	265
2010	1148	493	188	708	198	57	190	66	311
2011	1078	523	338	537	131	30	153+1	39	407
2012	975	414	144	*	129	18	255	63	225
2013	884	455	261	*	102	16	195	56	333
2014	959	439	145	*	128	27	235	48	220
2015	923	421	260	*	*	*	*	*	260

3.5 Καθυστέρηση απόκτησης πτυχίου λόγω αργοπορίας σε συγκεκριμένα μαθήματα

Ολοκληρώνουμε τον έλεγχο σε συγκεκριμένα μαθήματα δείχνοντας πρώτα πόσο απόκλιση υπάρχει σε κάθε μάθημα ξεχωριστά, από το προσφερόμενο χρόνο που θεωρείται ως σωστό (το εξάμηνο που σύμφωνα με τον οδηγό σπουδών πρέπει να δίνει ένας φοιτητής το μάθημα).. Για να δούμε το ακριβές χρόνο αφαιρούμε τις παρακάτω τιμές από το έτος που προσφέρεται το κάθε μάθημα. Έπειτα δείχνουμε τα στατιστικά του κάθε μαθήματος, δηλαδή πόσες προσπάθειες χρειάζονται μέχρι να έρθει η επιτυχής εξέταση.

Απειροστικός Λογισμός Ι

Min	1 τετ/ριο	Median	Mean	3	max	sd
1 έτος	1 έτος	2 έτος	3,108 έτος	4 έτος	15 έτος	2,26

Μάθημα 1 έτους: Μέση διαφορά από το κανονικό έτος: (περίπου) 2 έτη

Στατιστική 1

Min	1	Median	Mean	3	max	sd
1 έτος	4 έτος	5 έτος	5,655 έτος	7 έτος	15 έτος	2,18

Μάθημα 3 έτους: Μέση διαφορά από το κανονικό έτος: 2 και μισό έτη

Πραγματική

Min	1	Median	Mean	3	max	sd
1 έτος	4 έτος	5 έτος	5.58 έτος	7 έτος	16 έτος	2.663858

Μάθημα 3 έτους: Μέση διαφορά από το κανονικό έτος: 2 και μισό έτη

Πιθανότητες 1

Min	1	Median	Mean	3	max	sd
1 έτος	3 έτος	4 έτος	4.182 έτος	5 έτος	15 έτος	2.13

Μάθημα 2 έτους: Μέση απόκλιση περίπου 2 έτη

Προσπάθειες μέχρι την επιτυχία στο εκάστοτε μάθημα σε εξάμηνα

Απειροστικός Λογισμός I

Min.	1st Qu.	Median	Mean	3	Max.	sd
1	1	3	3.353	4	22	2.739246

Δίχως να μετράμε τα 0-2 (δηλαδή περιπτώσεις που ο φοιτητής συνήθως δεν θεωρείται καλά προετοιμασμένος)

Min.	1st Qu.	Median	Mean	3	Max.	sd
1	1	1	1.805	2	12	

Στατιστική

Min.	1st Qu.	Median	Mean	3	Max.	sd
1.000	1.000	2.000	2.459	3.000	22	1.94553

Δίχως να μετράμε τα 0-2

Min.	1st Qu.	Median	Mean	3	Max.	sd
1.000	1.000	1	1.482	2.000	6.000	

Πιθανότητες I

Min.	1st Qu.	Median	Mean	3	Max.	sd
1	1	2	2,739	3	22	2.152582

Δίχως να μετράμε τα 0-2

Min.	1st Qu.	Median	Mean	3	Max.	sd
1	1	1	1,506	2	8	0,836

Πραγματική Ανάλυση

Min.	1st Qu.	Median	Mean	3	Max.	sd
1.000	1.000	2.000	2.728	3.000	18.000	2.380756

Δίχως να μετράμε τα 0-2

Min.	1st Qu.	Median	Mean	3	Max.	sd
1.000	1.000	1.000	1.514	2.000	8.000	0,877

4. Μοντελοποίηση μέσω περιγραφικών μέτρων χρόνου απόκτησης πτυχίου - βαθμολογίας

4.1 Πρώτο Ερώτημα Έρευνας: Παράγοντες καθυστέρησης χρόνου απόκτησης πτυχίου

Πρώτο ερώτημα έρευνας

Το πρώτο ερώτημα που μας δίνεται να απαντήσουμε είναι : **Για ποιους παράγοντες κάποιος καθυστερεί να πάρει πτυχίο;** Στο πρώτο μας αυτό μοντέλο θα θεωρήσουμε ως εξαρτημένη τυχαία μεταβλητή το εξάμηνο το οποίο ο πτυχιούχος φοιτητής ολοκλήρωσε τις σπουδές του, από το οποίο θα αφαιρέσουμε το 7 δηλαδή το κατώτερο όριο απόκτησης πτυχίου, ούτως ώστε η εξαρτημένη μας μεταβλητή να είναι πλέον ο χρόνος καθυστέρησης εκφρασμένος σε εξάμηνα από το κατώτερο όριο απόκτησης πτυχίου. Σαν περιγραφικές (ανεξάρτητες) μεταβλητές θα χρησιμοποιήσουμε τις εξής:

- 1)Το φύλο του φοιτητή
 - 2)Το βαθμό πτυχίου
 - 3)Την κατεύθυνση που επέλεξε (Θεωρητική ή Εφαρμοσμένη)
 - 4)Την εποχή που έγινε δεκτός στην σχολή, χωρισμένη σε τρεις κατηγορίες 2000-04, 2005-2007, και 2008-2015
- και τέλος :
- 5)Τον τόπο μόνιμης κατοικίας του, πριν γίνει δεκτός στην σχολή, σε τρεις κατηγορίες Αττική, Εκτός Αττικής και Κύπρος. (τόπος καταγωγής)

Για την μοντελοποίηση θα χρησιμοποιήσουμε τα μοντέλα GLMM με random effects.

Συγκεντρωτικά έχουμε τον παρακάτω πίνακα

Εξαρτημένη μεταβλητή	Χρόνος καθυστέρησης απόκτησης πτυχίου
Περιγραφικές Μεταβλητές	Βαθμός
	Κατεύθυνση
	Φύλο
	Εποχή
	Καταγωγή

4.1.1 Περιγραφή μεταβλητών

Όπως είπαμε ως εξαρτημένη μεταβλητή θα θεωρήσουμε τον χρόνο καθυστέρησης του φοιτητή για να αποκτήσει πτυχίο, δηλαδή ο σκοπός μας είναι να βρούμε τι τελικά φαίνεται με βάση το μοντέλο που θα περιγράψουμε να επηρεάζει και κατά πόσο τον χρόνο που τελειώνει ο εκάστοτε φοιτητής. Να επισημάνουμε εδώ ότι το πλήθος των πτυχιούχων είναι 2312 το οποίο θα είναι και το μέγεθος δείγματος. Παρακάτω δίνονται κάποια μέτρα θέσης για να κάνουμε κάποιες πρώτες παρατηρήσεις.

Μέτρα θέσης χρόνου καθυστέρησης πτυχίου

Min	1 Τετ/ριο	Διάμεσος	Μέσος	3 Τετ/ριο	Max
0 εξάμηνα	4 εξάμηνα	6 εξάμηνα	7.143 εξάμηνα	10 εξάμηνα	25 εξάμηνα

Αρχικά παρατηρούμε ότι υπάρχουν φοιτητές που ολοκληρώνουν την σχολή στα 3.5 ακριβώς χρόνια (7 εξάμηνα) που είναι και το ελάχιστο, πιο συγκεκριμένα αυτοί είναι 21 άτομα. Το 25% καθυστερεί έως και 4 εξάμηνα δηλαδή δύο χρόνια. Η διάμεσος είναι στα 6 παραπάνω εξάμηνα, δηλαδή οι μισοί πτυχιούχοι έχουν ολοκληρώσει μέχρι και τα 6.5 συνολικά χρόνια σπουδών. Το 3^ο τεταρτημόριο είναι στα 10 παραπάνω εξάμηνα, ενώ υπάρχουν και οι ακραίες τιμές με μέγιστη τα 25 παραπάνω εξάμηνα. Προφανώς και δεν υπολογίζεται κατά πόσο οι συγκεκριμένοι φοιτητές ήταν ενεργοί ή απλά επανήλθαν στις σπουδές τους έπειτα από καιρό (περιορισμός έρευνας).

Ως πρώτη περιγραφική (ανεξάρτητη) μεταβλητή θα ασχοληθούμε με τον **βαθμό απόκτησης πτυχίου** και θα το συγκρίνουμε σε σχέση με την εξαρτημένη μας μεταβλητή τον χρόνο καθυστέρησης πτυχίου. Παρακάτω θα κατηγοριοποιήσουμε λίγο την μεταβλητή για να φαίνονται κάποια πιο ξεκάθαρα αποτελέσματα, στο μοντέλο θα είναι όμως ως συνεχής.

Μέτρα θέσης χρόνου καθυστέρησης πτυχίου όσον αφορά το Βαθμό

	Min	1 Τετ/ριο	Διάμεσος	Μέσος	3 Τετ/ριο	Max
5.0 έως 6.5	1 εξ/να	6 εξ/να	8 εξ/να	8,554 εξ/να	11 εξ/να	22 εξ/να
6.5 έως 7.5	0 εξ/να	2 εξ/να	4 εξ/να	4,915 εξ/να	6,25 εξ/να	25 εξ/να
7.5 έως 8.5	0 εξ/να	1,5 εξ/να	2 εξ/να	2,867 εξ/να	4,1 εξ/να	11 εξ/να
8.5 έως 10	0 εξ/να	1 εξ/να	2 εξ/να	1,988 εξ/να	3 εξ/να	6 εξ/να

Το +1 εξάμηνο είναι τέσσερα χρόνια ακριβώς

Από τον παραπάνω πίνακα είναι εμφανές ότι όσο καλύτερος είναι ο φοιτητής βαθμολογικά, τόσο ανάλογα γρήγορος είναι και στην ολοκλήρωση των σπουδών

του. Αυτό γίνεται πιο εμφανές με απλή σύγκριση των μέσων όπου πχ οι Αριστούχοι μαθητές του 8.5 και πάνω κατά μέσο όρο ολοκληρώνουν στα σχεδόν 4.5 χρόνια, ενώ οι μαθητές του 6.5 έως 7.5 στα 6,5 χρόνια.

Σαν δεύτερη περιγραφική μεταβλητή έχουμε την κατεύθυνση που επέλεξε ο πτυχιούχος, Θεωρητικά ή Εφαρμοσμένα Μαθηματικά. Εδώ να αναφέρουμε ότι από τους πτυχιούχους μόλις το 11.5% επέλεξε την Θεωρητική κατεύθυνση, ενώ να αναφερθεί ότι τα δεδομένα δεν μας παρέχουν την πληροφορία για το αν τελικά επιλέχθηκαν και οι δύο από κάποιον φοιτητή.

Μέτρα θέσης χρόνου καθυστέρησης πτυχίου όσον αφορά την Κατεύθυνση

	Min	1 Τετ/ριο	Διάμεσος	Μέσος	3 Τετ/ριο	Max
Θεωρητικά	0	2	4	5,2	7,25	21
Εφαρμοσμένα	0	4	7	7,39	10	25

Οι αριθμοί είναι σε επιπλέον εξάμηνα

Από το παραπάνω παρατηρούμε ότι φαίνεται μια καλύτερη χρονική επίδοση στα άτομα που επέλεξαν την Θεωρητική Κατεύθυνση, με το μέσο στα Θεωρητικά να κυμαίνεται στα 5.2 επιπλέον εξάμηνα (δηλαδή +δυόμιση χρόνια), ενώ στα Εφαρμοσμένα στο 7.39 (παραπάνω από τρία και μισό χρόνια επιπλέον).

Η επόμενη ανεξάρτητη μεταβλητή μας θα είναι το Φύλο του φοιτητή. Στο δείγμα των πτυχιούχων έχουμε 1191 αγόρια και 1121 κορίτσια. Εδώ να παρατηρήσουμε ότι στην πράξη έχουμε μοιρασμένους πτυχιούχους (πάνω κάτω μοιρασμένη, 48% κορίτσια – 52% αγόρια).

Μέτρα θέσης χρόνου καθυστέρησης πτυχίου όσον αφορά το Sex

	Min	1 Τετ/ριο	Διάμεσος	Μέσος	3 Τετ/ριο	Max
Γυναίκες	0	4	6	6.697	9	22
Άντρες	0	4	7	7.563	10	25

Οι αριθμοί είναι σε επιπλέον εξάμηνα

Από το παραπάνω πίνακα παρατηρούμε μια ελαφρώς καλύτερη τιμή (εδώ καλύτερα όσο λιγότερα) στις γυναίκες σχεδόν σε όλα τα μέτρα θέσης, κατά ένα περίπου εξάμηνο ή και κατά δύο εξάμηνα. Αυτό παρατηρούμε από την σύγκριση των μέσων όρων καθυστέρησης σε εξάμηνα.

Ως επόμενη περιγραφική μεταβλητή έχουμε την Εποχή (**Έτος Εισαγωγής**) που έγινε δεκτός στο τμήμα η οποία θα είναι κατηγορική χωρισμένη σε τρεις κατηγορίες 2000-2004, 2005-2007 και 2008-2015. Εδώ να σημειωθεί ότι η κατηγοριοποίηση έχει γίνει με γνώμονα ότι το 2008 ξέσπασε η παγκόσμια οικονομική κρίση, όπως επιπλέον ότι σε αυτά τα χρόνια 2008-2015 η διαθέσιμη πληροφορία (η βοήθεια που δέχονται οι φοιτητές από ποικίλες μορφές) αυξήθηκε πάρα πολύ (σημειώσεις φοιτητών στην eclass, πρόσβαση σε σημειώσεις, υλικό μαθημάτων και βιντεοδιαλέξεις), αλλά και οκίνδυνος διαγραφής με το γνωστό νόμο Διαμαντοπούλου (νόμος 4009/2011). Τα έτη 2005- 2007 είναι ας πούμε τα χρόνια πριν την κρίση, στα οποία υπήρξε σαφώς μια καλύτερευση στην πληροφορία και στην πρόσβαση σε σχέση με την πρώτη κατηγορία 2000-2004. Έχουμε τα παρακάτω χρήσιμα μέτρα θέσης.

Μέτρα θέσης χρόνου καθυστέρησης πτυχίου όσον αφορά το Έτος Εισαγωγής

	Min	1 Τετ/ριο	Διάμεσος	Μέσος	3 Τετ/ριο	Max
2000-04	0	5	8	8.551	12	25
2005-07	1	4.75	7	6.683	8	13
2008-15	0	2	3	3.12	4	8

Οι αριθμοί είναι σε επιπλέον εξάμηνα

Από τα παραπάνω φαίνεται μια καλύτερευση γενικά όσο περνάνε τα χρόνια. Εδώ όμως πρέπει να δώσουμε προσοχή καθώς στα έτη 2008 και έπειτα οι << ας πούμε όχι τόσο καλοί φοιτητές>> δεν έχουν προλάβει να ολοκληρώσουν τις σπουδές τους μέχρι στιγμής άρα το δείγμα μας όσον αφορά αυτούς είναι κυρίως από άτομα που το τελειώνουν μέσα στο πολύ 7 έτη πάνω-κάτω.

Τελευταία περιγραφική μεταβλητή θα είναι ο τόπος μόνιμης κατοικίας του φοιτητή, για να μπορέσουμε να ελένουμε αν η πίεση λόγω των εξόδων παραμονής στην Αθήνα είναι κίνητρο γρήγορης ολοκλήρωσης σπουδών. Οι κατηγορίες είναι: Αττική, Εκτός Αττικής και Κύπρος, με αντίστοιχα ποσοστά 60%, 24% και 16% (ποσοστά στο περίπου).

Μέτρα θέσης χρόνου καθυστέρησης πτυχίου όσον αφορά την 'Καταγωγή'

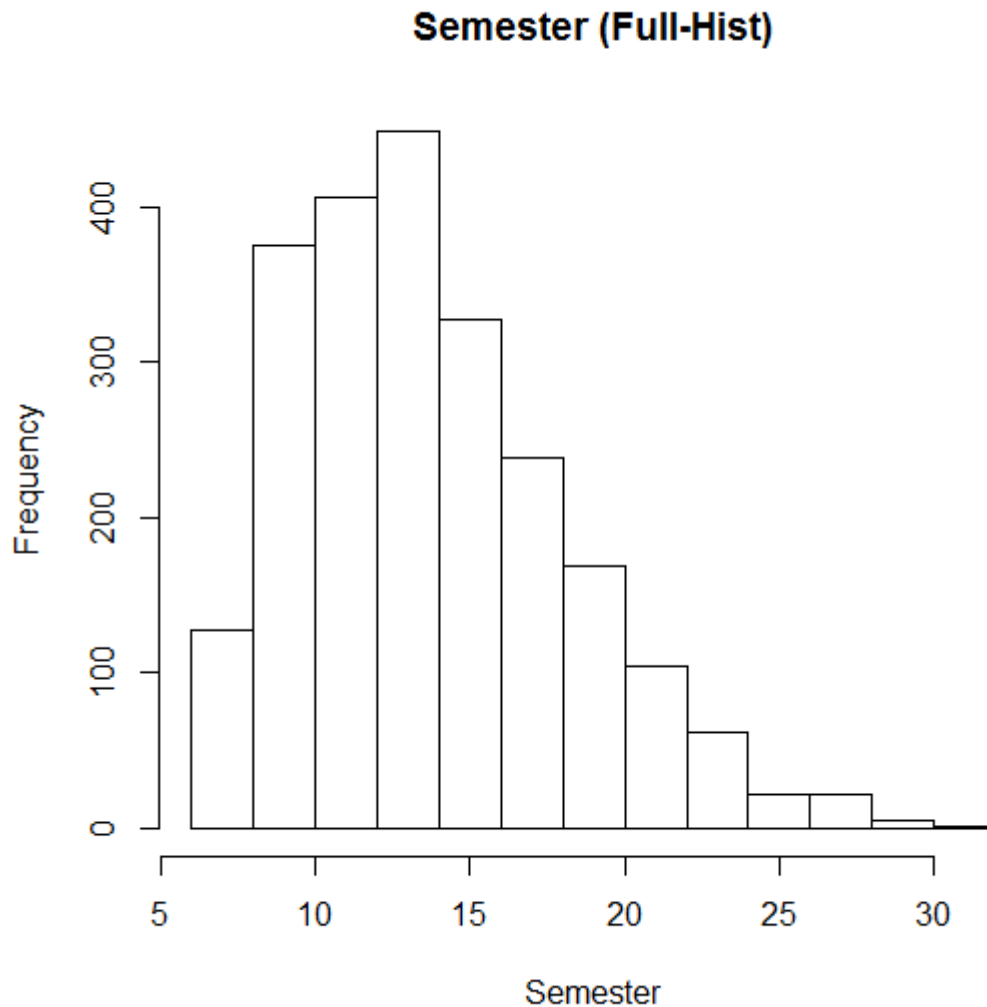
	Min	1 Τετ/ριο	Διάμεσος	Μέσος	3 Τετ/ριο	Max
Αττική	0	4	7	7.5	10	25
Εκτός Αττικής	0	4	6	6.86	9	22
Κύπρος	0	2	3	3.72	4	16

Είναι εμφανές ότι οι φοιτητές από την Κύπρο, βιάζονται πιο πολύ να ολοκληρώσουν τις σπουδές του σε σχέση με τις άλλες δύο κατηγορίες, όπου οι τιμές κυμαίνονται σχετικά κοντά, λίγο καλύτερα για τους εκτός Αττικής.

4.1.2 Περιγραφική Στατιστική

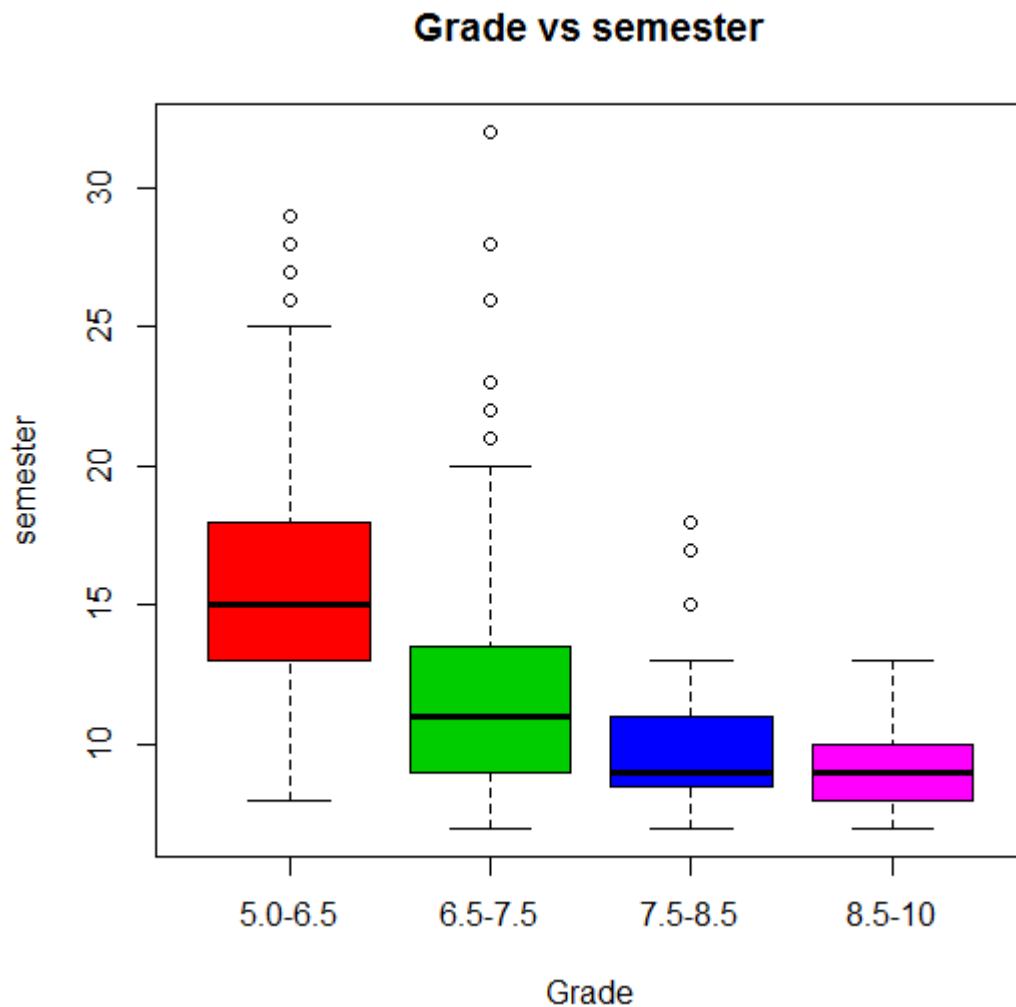
Στο υποκεφάλαιο εδώ θα συνεχίσουμε με Περιγραφική Στατιστική όπου θα προσπαθήσουμε να εκμειεύσουμε και από εκεί κάποια συμπεράσματα, πριν το τελικό μοντέλο.

Αρχικά θα αναφερθούμε στο ιστόγραμμα του χρόνου απόκτησης του πτυχίου, στο παρακάτω σχήμα, στο οποίο και προφανώς βλέπουμε αυτά που είχαμε παρατηρήσει και από τα μέτρα θέσης προηγουμένως.



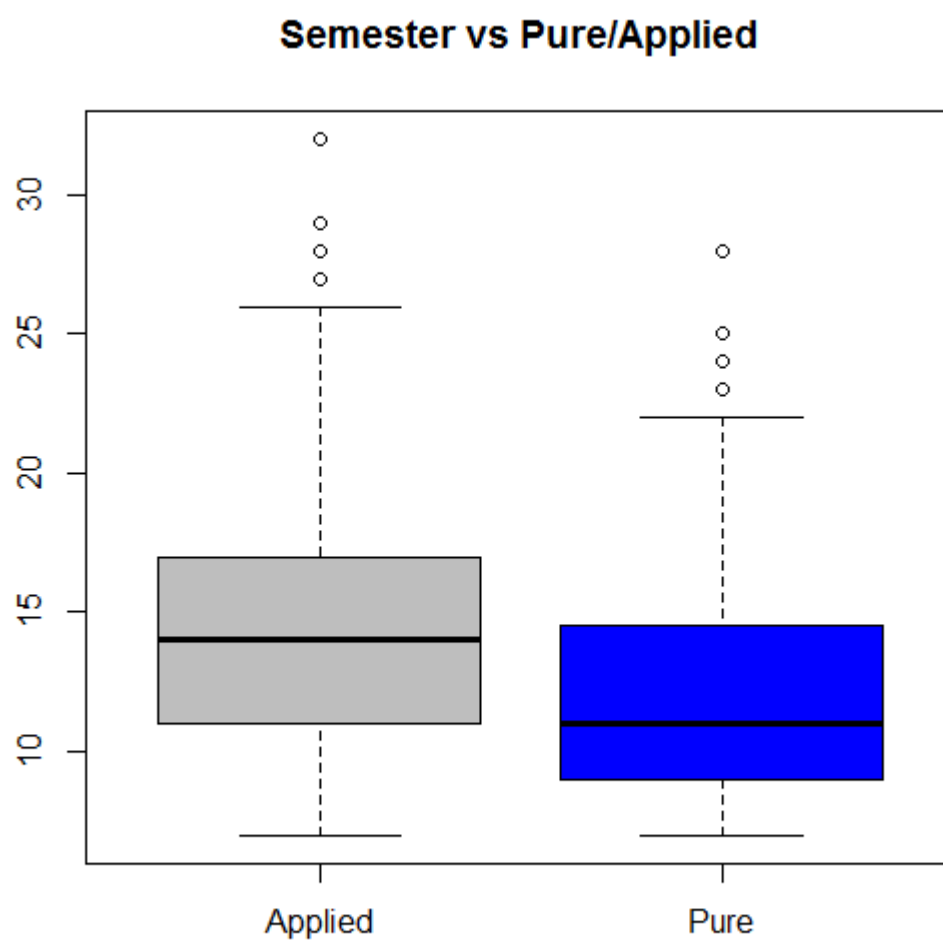
Συνεχίζουμε με τα θηκογράμματα για το χρόνο καθυστέρησης πτυχίου για κάθε επεξηγηματική μεταβλητή ξεχωριστά. Να σημειωθεί ότι εδώ δεν κρίνεται αναγκαίο να αφαιρέσουμε το κατώτερο όριο, για να γίνει κατανοητή η όποια ύπαρξη διαφοράς, αφού το σχήμα θα ήταν το ίδιο με μια τοποθέτηση προς τα κάτω. Να αναφερθεί ότι τα ποσοστά στα εξάμηνα είναι 14%, 18%, 26%, 25% και 17%, αντίστοιχα για τις κατηγορίες "λιγότερα από 10", "10-11", "12-14", "15-18" και 19 και άνω.

Ιστόγραμμα χρόνου καθυστέρησης πτυχίου όσον αφορά το Βαθμό



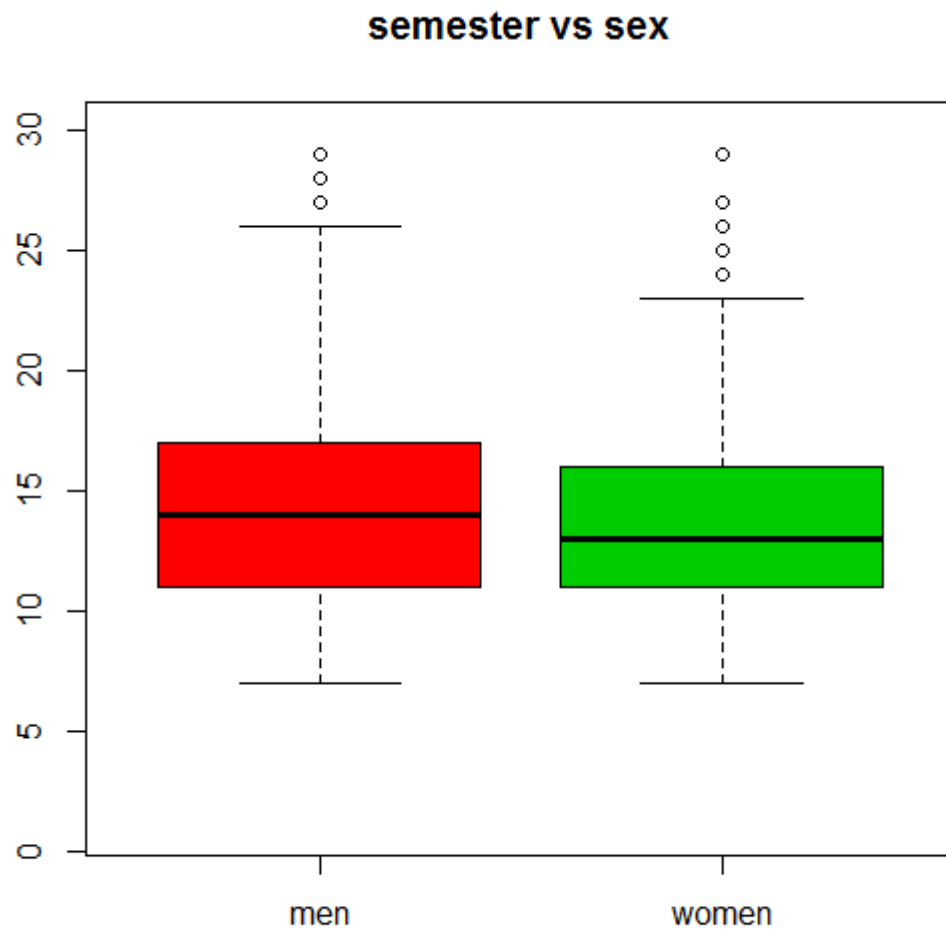
Από το παραπάνω γράφημα βλέπουμε το πόσο μειώνονται τα εξάμηνα σε όλα τα μέτρα θέσης (1^ο τεταρτημόριο, διάμεσος, minimum, maximum) όσο μεγαλύτερη βαθμολογία έχει πετύχει ο πτυχιούχος. Είναι ξεκάθαρο το πόσο όσο μεγαλώνει η βαθμολογία τόσο και ο φοιτητής τελειώνει πιο σύντομα τις σπουδές του, η διαφορά αυτή είναι εμφανής μεταξύ των δύο πρώτων κατηγοριών.

Ιστόγραμμα χρόνου καθυστέρησης πτυχίου όσον αφορά την Κατεύθυνση



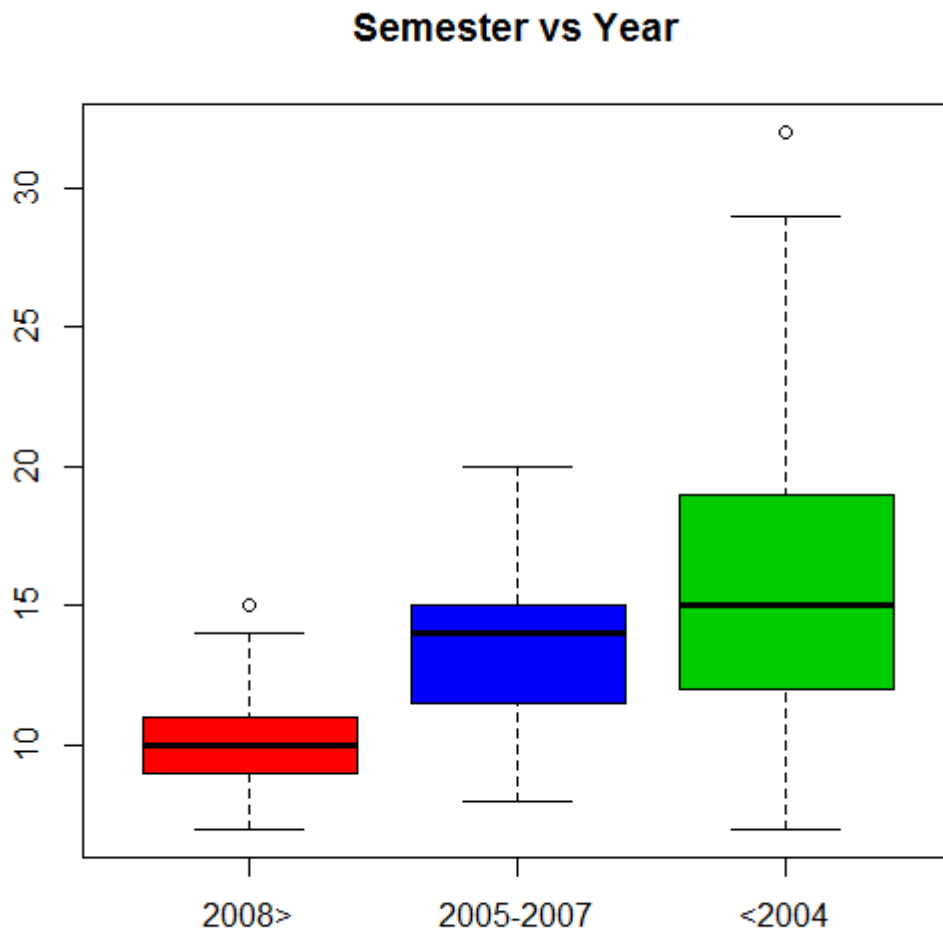
Οι φοιτητές που επέλεξαν Θεωρητική έχουν μια καλύτερη επίδοση στο χρόνο απόκτησης πτυχίου σε όλα τα μέτρα θέσης.

Ιστόγραμμα χρόνου καθυστέρησης πτυχίου όσον αφορά το Φύλο



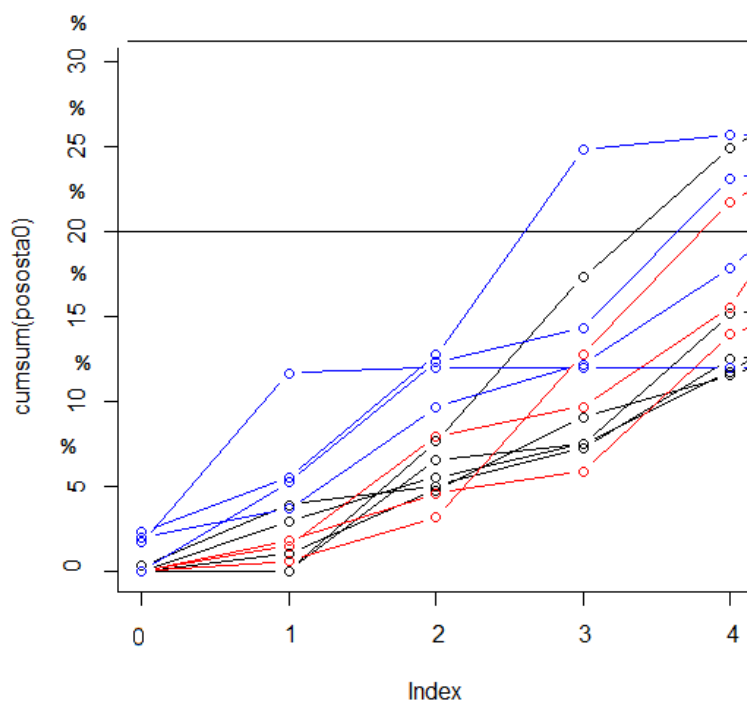
Οι διαφορές με βάση τα φύλα δεν φαίνεται να είναι σημαντικές. Ωστόσο η διάμεσος είναι σχετικά μεγαλύτερη στους άντρες, όπως και όλα τα άλλα μέτρα θέσης.

Ιστόγραμμα χρόνου καθυστέρησης πτυχίου όσον αφορά την Χρονιά Εισαγωγής



Από το παραπάνω είναι εμφανές ότι τα τελευταία χρόνια οι φοιτητές ολοκληρώνουν πιο σύντομα τις σπουδές τους. Εξαιρώντας ακόμα και την κατηγορία αυτών που μπήκαν μετά το 2008 (που μπορεί να μεροληπτεί) στις άλλες δύο κατηγορίες βλέπουμε πάλι το ίδιο.

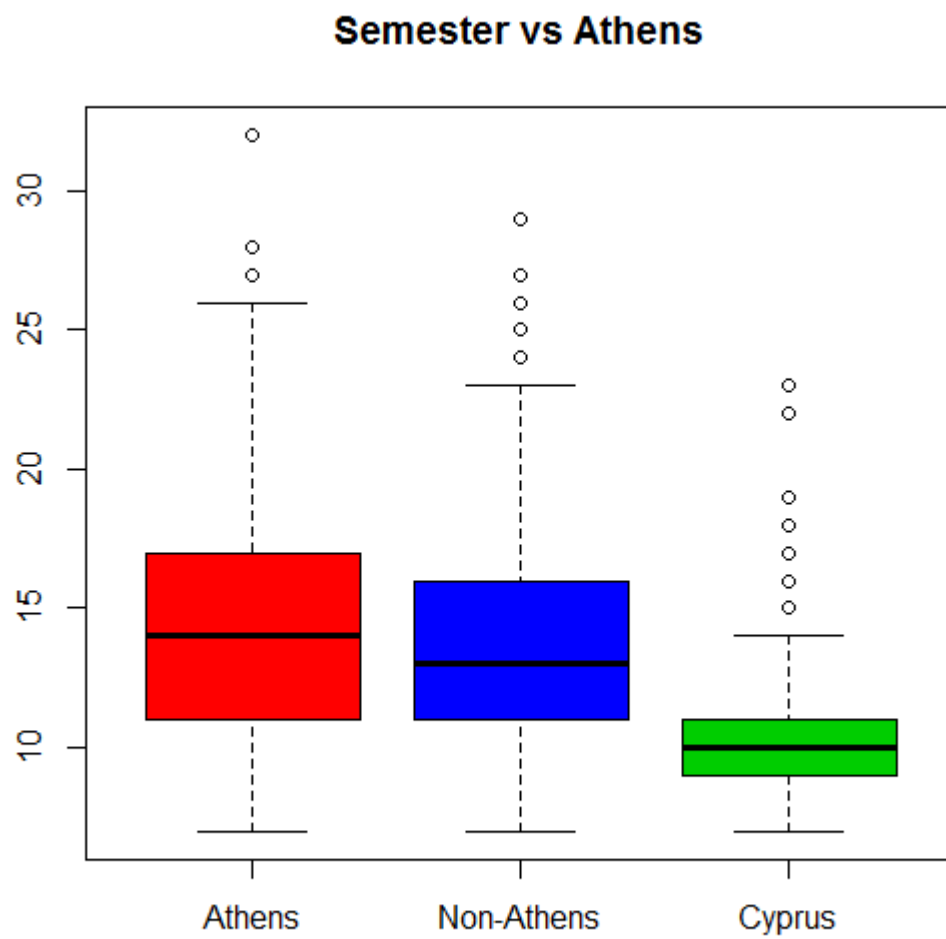
Παρακάτω θα ερευνήσουμε κατά πόσο μπορούμε να επιτρέψουμε αυτή τη μεροληψία να υπάρχει στο μοντέλο, συγκρίνοντας τι συμβαίνει για κάθε επιπλέον εξάμηνο όσον αφορά τις αντίστοιχες χρονικές περιόδους ξεχωριστά.



Με μπλέ χρώμα είναι το τι έκανε το αντίστοιχο ποσοστό του συνόλου των φοιτητών (όχι μόνο των πτυχιούχων) που ανήκουν στην κατηγορία 2008 και έπειτα, με τα επιπλέον εξάμηνα, παρατηρούμε ότι τα πηγαίνουν αρκετά καλύτερα από τις άλλες δύο κατηγορίες, κόκκινο (2005-07) και μαύρο (≤ 2004). Αυτό μαζί με το προηγούμενο θηκόγραμμα μας κάνει να πιστεύουμε ότι ναι μεν θα υπάρχει μια κάποια μεροληψία, αλλά όντως φαίνεται να υπάρχει καλυτέρευση τα τελευταία χρόνια.

Γενικά άλλωστε άμα κοιτάξουμε το συνολικό γράφημα, μέχρι εκεί που έχουν προλάβει να φτάσουν χρονικά οι τελευταίες χρονιές (μέχρι εκεί που γίνονται ευθεία) φαίνεται ότι τα πάνε καλύτερα από τις περισσότερες χρονιές των άλλω δύο κατηγοριών.

Ιστόγραμμα χρόνου καθυστέρησης πτυχίου όσον αφορά την Καταγωγή



Από το παραπάνω βλέπουμε ότι οι Κύπριοι φοιτητές ολοκληρώνουν πιο σύντομα σε σχέση με τις άλλες δύο κατηγορίες οι οποίες φαίνονται να κινούνται στα ίδια.

4.1.3 Μοντελοποίηση

Όπως ήδη αναφέραμε το μοντέλο που θα χρησιμοποιήσουμε θα είναι ένα γενικευμένο γραμμικό μοντέλο με τυχαίες επιδράσεις. Οι τυχαίες επιδράσεις θα εισαχθούν στο κάθε άτομο ξεχωριστά, δηλαδή υποθέτουμε ότι για κάποιο αδιευκρίνιστο λόγω επηρεάζεται η διάρκεια σπουδών του κάθε φοιτητή (το οποίο έχει και λογική εξήγηση), (άρα τυχαίος όρος είναι ο φοιτητής). Παραδείγματος χάριν ένας φοιτητής με κάποια προβλήματα υγείας σε κάποια στιγμή της ζωής του, αυτό μπορεί να θεωρηθεί μια τυχαία επίδραση η οποία δεν μπορεί να ελεγχθεί.

Όπως ήδη είπαμε εξαρτημένη μας μεταβλητή αυτού του μοντέλου θα είναι ο χρόνος καθυστέρησης απόκτησης πτυχίου, για τον οποίο θα υποθέσουμε πως ακολουθεί Poisson(μ) κατανομή. Οπότε με τα συνηθισμένα γενικευμένα γραμμικά μοντέλα, με βάση και το σχετικό link-function θα ήταν της μορφής

$$\log(\mu) = X' \cdot \beta$$

όμως λόγω των randomeffectτελικά το μοντέλο είναι το

$$\log(\mu) = X' \cdot \beta + Z' \cdot r$$

Όπου β , ο πίνακας των συντελεστών των ανεξαρτήτων μεταβλητών, ενώ r είναι τα randomeffects.

Στην πράξη το μοντέλο που θα υποθέσουμε θα είναι της μορφής

$$\log(\mu) = \beta_0 + r_i + \beta_1 \cdot X_1 + \dots + \beta_p \cdot X_p$$

με τις τυχαίες επιδράσεις να ακολουθούν Κανονική $N(0, u^2)$, όπου $u^2 = \text{var}(r)$.

Διαδικασία επιλογής μεταβλητών στο μοντέλο

Η διαδικασία επιλογής μοντέλου θα είναι **αλγοριθμικά** η παρακάτω:

-Έλεγε μία προς μία τις περιγραφικές μεταβλητές αν είναι στατιστικά σημαντικές, στο μοντέλο με μόνη περιγραφική την κάθε φορά μία και εξαρτημένη αυτή του χρόνου καθυστέρησης.

-Όσες είναι στατιστικά σημαντικές θα τοποθετηθούν σε ένα νέο μοντέλο.

-Από αυτό το μοντέλο θα εξαιρείται μια προς μία κάθε φορά όποια βρεθεί περισσότερο στατιστικά μη σημαντική. Στη πράξη αυτή με το μεγαλύτερο p-valueμεγαλύτερο του 0,05. (backwardelimination)

-Όταν τελικά δεν υπάρχει στατιστικά μη σημαντική μεταβλητή σταματάω και αυτό είναι το τελικό μοντέλο.

Με βάση την παραπάνω μεθοδολογία έχουμε τον παρακάτω πίνακα.

	Εκτίμηση β	Τυπική απόκλιση	P value
Φύλο	-0.1170600 (γυναίκα)	0.0004075	σχεδόν 0
Βαθμός	-0.55415	0.01634	σχεδόν 0
Εποχή	-0.28720 (2005-008)	0.02635	
	-1.00086 (>2008)	0.03906	σχεδόν 0
Καταγωγή	-0.08810 (Εκτός Αθήνας)	0.02865	
	-0.72364 Κύπρος	0.06068	σχεδόν 0
Κατεύθυνση	-0.38017 (θεωρητική)	0.04242	σχεδόν 0

Παρατηρούμε ότι όλες οι μεταβλητές είναι στατιστικά σημαντικές στο απλό γενικευμένο γραμμικό μοντέλο με τυχαίες επιδρασεις, με μια περιγραφική μεταβλητή της μορφής

$$\log(\mu) = \beta_0 + r_i + \beta_1 \cdot X_1$$

Οπότε θα προχωρήσουμε στο επόμενο βήμα, όπως έχουμε αναφέρει με όλες τις στατιστικά σημαντικές μεταβλητές από το προηγούμενο βήμα.

Θα είναι της μορφής

$$\log(\mu) = \beta_0 + r_i + \beta_1 \cdot \Phi\acute{\upsilon}\lambda\omicron + \beta_2 \cdot \text{Εποχή}1 + \beta_3 \cdot \text{Εποχή}2 + \beta_4 \cdot \text{Κατευθυνση}1 + \beta_6 \cdot \text{Βαθμός} + \beta_5 \cdot \text{Καταγωγή}2 + \beta_7 \cdot \text{Καταγωγή}1$$

με $r_i \sim N(0, u^2)$.

Από το οποίο έχουμε τα παρακάτω αποτελέσματα:

Μεταβλητή	Εκτίμηση β	Τυπική απόκλιση	Pvalue
Σταθ. Όρος	5.09573	0.10041	σχεδόν 0
Κατεύθυνση	0.08946 (θεωρητική)	0.03437	0.00925
Βαθμός	-0.47031	0.01636	σχεδόν 0
Εποχή	-0.24647 (2005-08)	0.02181	σχεδόν 0
	-0.71920 (>'08)	0.03557	
Καταγωγή	-0.0442 (Εκτός Αθήνας)	0.0209	σχεδόν 0
	-0.46964 (Κύπρος)	0.04931	
Φύλο	-0.12108 (γυναίκα)	0.01861	σχεδόν 0

με $r_i \sim N(0, 0.04724)$.

Παρατηρούμε ότι όλες οι επεξηγηματικές μεταβλητές είναι στατιστικά σημαντικές, σε επίπεδο στατιστικής σημαντικότητας $\alpha=0,05$, άρα δεν είναι ανάγκη να συνεχίσουμε σε backwardelimination, και αυτό είναι το τελικό μας μοντέλο.

Συμπέρασμα: Δηλαδή και οι πέντε επεξηγηματικές μεταβλητές επηρεάζουν τελικά τον χρόνο που ο φοιτητής ολοκληρώνει τις σπουδές του.

Επεξήγηση Αποτελεσμάτων προηγούμενου πίνακα

Φύλο $\beta=-0.12108$

Υπάρχει αρνητική σχέση μεταξύ της εξαρτημένης μεταβλητής του επιπλέον χρόνου απόκτησης πτυχίου με το αν ο φοιτητής είναι γυναίκα (αυτό γιατί στην τιμή όπου η κατηγορική μεταβλητή Φύλο παίρνει την τιμή 1, είναι όταν έχουμε γυναίκα). Δηλαδή οι κοπέλες κάνουν λιγότερα επιπλέον εξάμηνα να ολοκληρώσουν τις σπουδές τους. Πιο συγκεκριμένα αντιστρέφοντας την link-function, δηλαδή

$$\exp(-0.12108) = 0.8859631$$

έχουμε ότι ο λόγος των επιπλέον εξαμήνων

$$\frac{\mu_1}{\mu_2} = 0.8859631$$

όπου μ_1 είναι ο μέσος των γυναικών, δηλαδή μια γυναίκα η οποία έχει όλα τα άλλα χαρακτηριστικά ίδια με έναν άντρα ενδέχεται να ολοκληρώσει τις σπουδές της στο περίπου 88% των επιπλέον εξαμήνων που θα χρειαστεί ένας άντρας. Πχ αν ένας

άντρας (με ίδια όλα τα άλλα χαρακτηριστικά) χρειάζεται 10 επιπλέον εξάμηνα, η αντίστοιχη γυναίκα θα χρειαστεί 9 επιπλέον εξάμηνα.

Παρατηρούμε ότι στην πράξη αυτό δεν είναι και τόσο σημαντικό, αλλά αφού στο μοντέλο υπάρχει επιρροή και είναι στατιστικά σημαντική εμείς την κρατάμε.

Βαθμός $\beta = -0.47031$

Με παρόμοιο τρόπο έχουμε ότι υπάρχει αρνητική σχέση μεταξύ, της συνεχούς τυχαίας μεταβλητής του βαθμού με την εξαρτημένη (χρόνος καθυστέρησης απόκτησης πτυχίου). Δηλαδή οι λεγόμενοι 'καλοί' φοιτητές με τους υψηλούς βαθμούς ολοκληρώνουν γρηγορότερα την προπτυχιακή φοιτητική τους ζωή. Πιο συγκεκριμένα αντιστρέφοντας

$$\exp(-0.473) = 0.6231301$$

έχουμε ότι ο λόγος των επιπλέον εξαμήνων

$$\frac{\mu_1}{\mu_2} = 0.6231301$$

όπου μ_1 είναι ο μέσος των ατόμων όπου έχουν πάρει μια μονάδα παραπάνω βαθμολογικά σε σχέση με αυτούς του μ_2 , δηλαδή έστω ότι ένας φοιτητής έχει πάρει 5.3 και ο άλλος 6.3 τότε ο μ_1 είναι αυτός με το 6.3 και ο μ_2 αυτός με το 5.3 (και έστω ότι έχουν όλα τα άλλα χαρακτηριστικά ίδια) τότε ο μ_1 ενδέχεται να ολοκληρώσει τις σπουδές του στο 62% των επιπλέον εξαμήνων του μ_2 , όπως και πριν αν ο μ_2 μείνει στη σχολή 10 επιπλέον εξάμηνα ο μ_1 θα χρειάζεται γύρω στα 6.

Κατεύθυνση $\beta = 0.08946$

Υπάρχει θετική σχέση μεταξύ της εξαρτημένης μεταβλητής του επιπλέον χρόνου απόκτησης πτυχίου με το αν ο φοιτητής έχει διαλέξει την θεωρητική Κατεύθυνση. Αυτό στην πράξη σημαίνει ότι το άτομο που τελικά επιλέγει θεωρητική θα δυσκολευτεί να τελειώσει χρονικά σε σχέση με αυτό που θα επιλέξει εφαρμοσμένη κατεύθυνση. Προσοχή πρέπει να δοθεί ότι μέχρι στιγμής είχαμε παρατηρήσει το ανάποδο (περιγραφικά αλλά και από το μοντέλο με μόνο την Κατεύθυνση ως μεταβλητή), αυτό γιατί τώρα απορροφώντας οι άλλες στατιστικά σημαντικές μεταβλητές το γεγονός ότι μάλλον φαίνεται οι πιο 'καλοί' βαθμολογικά φοιτητές επιλέγουν θεωρητική κατ/ση, τελικά φαίνεται ότι η θεωρητική δυσκολεύει τους φοιτητές πιο πολύ στο να τελειώσουν γρήγορα τις σπουδές τους. Συγκεκριμένα

$$\exp(0.08946) = 1.093584$$

Από εδώ φαίνεται ότι η διαφορά είναι ελάχιστη της τάξης του 10% επιπλέον εξαμήνα. Αυτό σημαίνει ότι τελικά αυτοί που θα επιλέξουν Θεωρητικά μαθηματικά, αλλά δεν θα είναι και τόσο επιμελείς θα καθυστερήσουν οριακά σε σχέση με αυτούς των Εφαρμοσμένων.

Εποχή, $\beta = -0.71920$ (2008-2015) και $\beta = -0.24647$ (2005-07)

Με βάση τους παραπάνω συντελεστές βλέπουμε σαφή βελτίωση όσο περνάνε τα χρόνια. Εδώ μπορούμε να δώσουμε πολλές εξηγήσεις όπως την βελτίωση της πληροφορίας και την βελτίωση της πρόσβασης σε αυτήν (eclass- σημειώσεις φοιτητών- παλιά θέματα- βιντεοδιαλέξεις), όπως ακόμα και τον κίνδυνο διαγραφής ή τις πτυχιακές εξεταστικές. Ακόμα μια εξήγηση είναι ότι την περίοδο της κρίσης ο φοιτητής δεν έχει άλλη διέξοδο εκτός των σπουδών του. Έχουμε ότι

$$\exp(-0.71920) = 0.487141 \text{ και } \exp(-0.24647) = 0.7815548,$$

δηλαδή ένα άτομο που πέρασε στη σχολή μετά το 2008 ενδέχεται να χρειαστεί σε σχέση με κάποιο άλλο με ίδια χαρακτηριστικά το 48% των επιπλέον εξαμήνων, ενώ ένα άτομο που μπήκε μεταξύ του 2005 και 2007 σχεδόν το 80% των εξαμήνων, ενώ ένα άτομο που μπήκε μετά το 2008 σε σχέση με κάποιο που μπήκε το 2005 και 2007 έχουμε ότι ο λόγος τους θα είναι 0.6232983. δηλαδή θα χρειαστεί το 62% των επιπλέον εξαμήνων. Εδώ να δοθεί προσοχή ότι οι όχι τόσο 'επιμελείς' φοιτητές μετά το 2008 δεν έχουν προλάβει ακόμα να ολοκληρώσουν τις σπουδές τους αυτό δημιουργεί μια ας πούμε μεροληψία, παρόλα αυτά βλέποντας και το 2005 και 2007 ότι υπάρχει πτώση σε σχέση με παλιότερα φαίνεται ότι η τάση καλυτέρευσης υπάρχει.

Καταγωγή, $\beta = -0.0442$ (Εκτός Αθήνας), $\beta = -0.46964$ (Κύπρος)

Παρόμοια παρατηρούμε ότι ουσιαστικά οι φοιτητές από την Κύπρο βιάζονται πολύ περισσότερο να ολοκληρώσουν τις σπουδές τους και σε σχέση με τους Εντός Αττικής φοιτητές, αλλά και με τους Εκτός. Συγκεκριμένα οι εντός Αττικής ολοκληρώνουν στο (0.6252273) 62% των επιπλέον εξαμήνων, ενώ με τους εκτός στο (0.6534822) 65% των επιπλέον εξαμήνων. Η διαφορά των άλλων δύο Εκτός Αττικής και Εντός Αττικής είναι (0.9567626) είναι στο 96% υπέρ των εκτός Αττικής (αμελητέα, αν κάποιος εντός Αττικής θέλει 10 επιπλέον εξάμηνα, ο εκτός θέλει μισό εξάμηνο λιγότερο). Εδώ φαίνεται ότι τελικά οι φοιτητές που προέρχονται εκτός Αττικής τελικά το γεγονός ότι έχουν περισσότερα έξοδα λόγω ενοικίου δεν είναι κίνητρο για γρήγορη ολοκλήρωση των σπουδών τους, ενώ για τους κύπριους φοιτητές είναι. Αυτό μπορεί να οφείλεται στο γεγονός ότι ενδεχομένως θέλουν να

επιστρέψουν στην Κύπρο, ενώ αντιθέτως οι εκτός Αττικής πολλές φορές ακόμα και μετά το τέλος παραμένουν στην Αθήνα για εύρεση εργασίας.

Συγκεντρωτικά:

- Οι καλοί βαθμολογικά τελειώνουν γρήγορα
- Οι Κύπριοι φοιτητές βιάζονται να τελειώσουν
- Σε σχέση με πιο παλιά, έχουμε καλύτερους χρόνους απόκτησης πτυχίου

Σύγκριση αποτελεσμάτων GLMMr.e. με σύνηθες GLM

Μεταβλητή	Εκτίμηση β GLMMr.e.	Εκτίμηση β με σύνηθες GLM
Σταθ. Όρος	5.09573	5.13799
Κατεύθυνση	0.08946 (θεωρητική)	0.08790
Βαθμός	-0.47031	-0.47343
Εποχή	-0.24647 (2005-08)	-0.25633
	-0.71920 (>'08)	-0.72431
Καταγωγή	-0.0442 (Εκτός Αθήνας)	-0.04022
	-0.46964 (Κύπρος)	-0.46124
Φύλο	-0.12108 (γυναίκα)	-0.11593

Εδώ να γίνει σαφές ότι η κύρια αλλαγή είναι στα διαστήματα εμπιστοσύνης, για την παρούσα διπλωματική η ανάλυση παραλείπεται ως πλεονασμός .

4.2 Δεύτερο Ερώτημα Έρευνας : Παράγοντες που επηρεάζονται τον τελικό βαθμό πτυχίου

Δεύτερο Ερώτημα Έρευνας

Το επόμενο ερώτημα προς απάντηση είναι : Τι επηρεάζει τον βαθμό πτυχίου ;Δηλαδή τώρα θα θεωρήσουμε ως εξαρτημένη τυχαία μεταβλητή την βαθμολογία με την οποία ο πτυχιούχος φοιτητής ολοκλήρωσε τις σπουδές του, κοινώς το βαθμό πτυχίου. Σαν περιγραφικές μεταβλητές θα χρησιμοποιήσουμε ότι και στο πρώτο μοντέλο.

1)Το φύλο του φοιτητή

2)Επιπλέον εξάμηνα για την απόκτηση πτυχίου, από το ελάχιστο όριο απόκτησης

3)Την κατεύθυνση που επέλεξε (θεωρητική ή Εφαρμοσμένη)

4)Την εποχή που έγινε δεκτός στην σχολή, χωρισμένη σε τρεις υποκατηγορίες

2000-2004, 2005-2007, και 2008-2015

και τέλος:

5)Τον τόπο μόνιμης κατοικίας του, πριν γίνει δεκτός στην σχολή, σε τρεις κατηγορίες Αττική, Εκτός Αττικής και Κύπρος. (τόπος καταγωγής)

Για την μοντελοποίηση θα χρησιμοποιήσουμε τα μοντέλα GLMMμε randomeffects, εδώ προφανώς αφού η βαθμολογία είναι συνεχής τιμή δεν γίνεται να υποθέσουμε Poisson.

Συγκεντρωτικά έχουμε τον παρακάτω πίνακα:

Εξαρτημένη μεταβλητή	Βαθμός
Περιγραφικές Μεταβλητές	Εξάμηνα
	Κατεύθυνση
	Φύλο
	Εποχή
	Καταγωγή

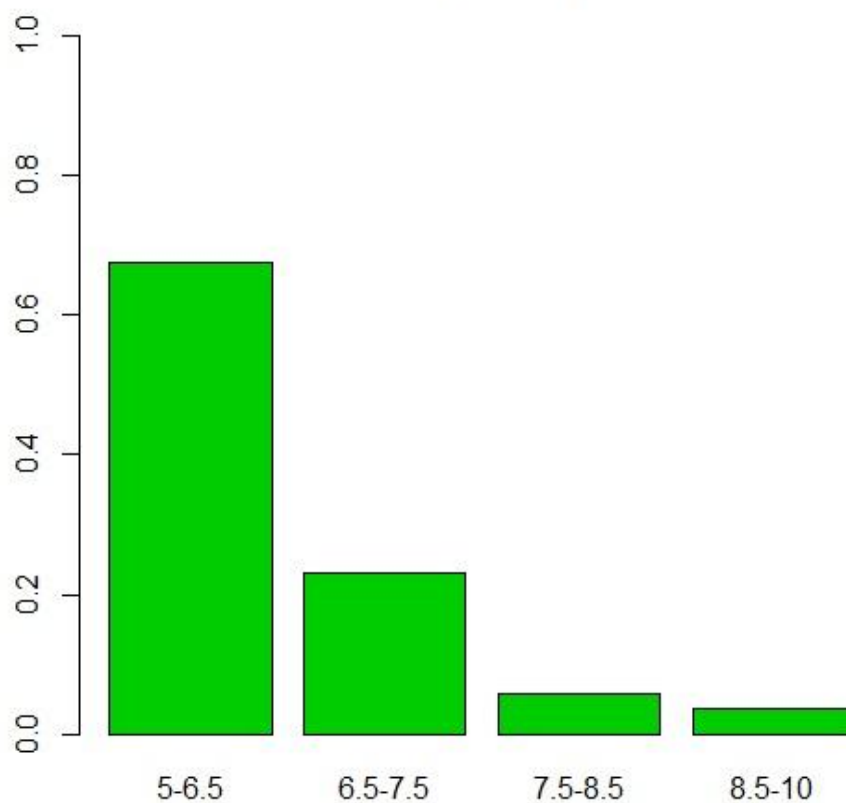
Στο παρών μοντέλο θα είμαστε πιο σύντομοι, καθότι την όλη λογική αντιμετώπισης την έχουμε προαναφέρει.

Μέτρα Θέσης - Περιγραφική Στατιστική

Min	1 Τετ/ριο	Διάμεσος	Μέσος	3 Τετ/ριο	Max
5.220	5.840	6.16	6.393	6.700	10.000

Από το παραπάνω πίνακα παρατηρούμε ότι ο μέσος βαθμός απόκτησης πτυχίου είναι 6.393 το οποίο εντάσσεται στην κλίμακα ΚΑΛΩΣ. Αξιοσημείωτο είναι ότι το 75% των φοιτητών βρίσκεται κάτω από το 6.7. Γύρω στο 70% των πτυχιούχων δηλαδή ολοκληρώνουν τις σπουδές τους με την κατώτερη ονομαστική βαθμολογία ΚΑΛΩΣ. Αυτό φαίνεται και από το παρακάτω:

Percentages of grades

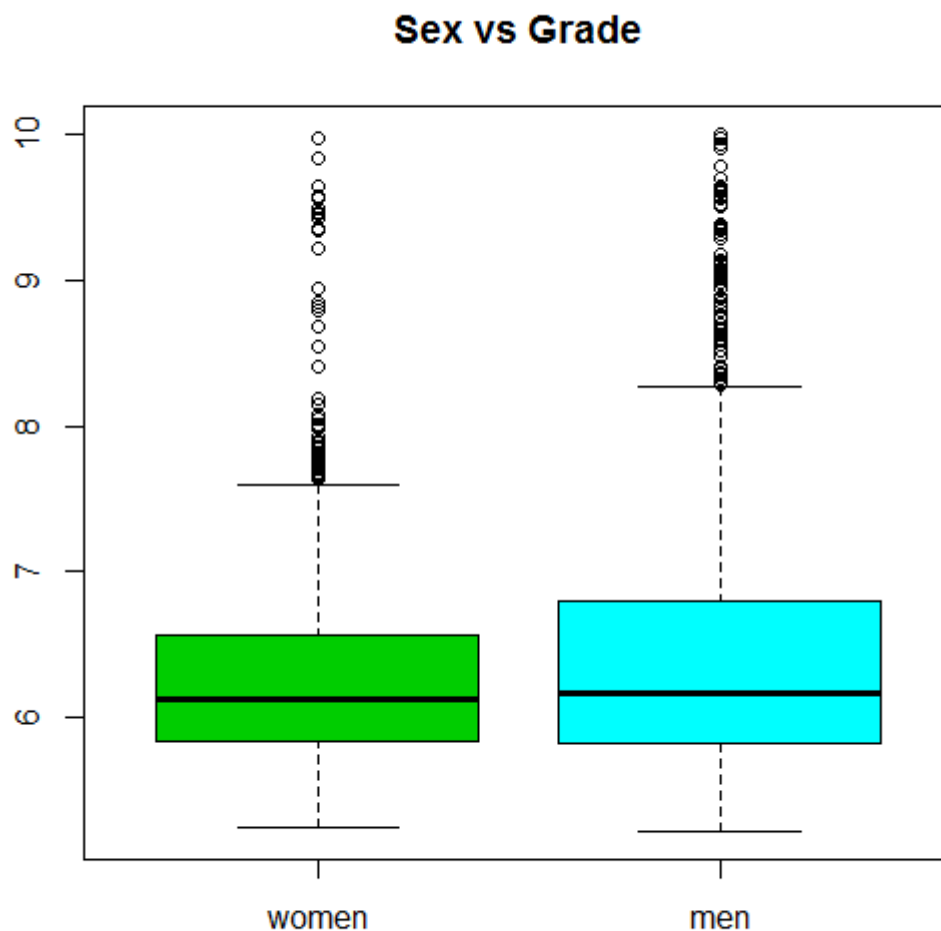


67%	23%	6%	4%
-----	-----	----	----

Παρατηρούμε ότι μόνο ένα 10% είναι πάνω από το 7.5, το οποίο ενδεχομένως σε άλλες σχολές να μην είναι κάτι συνηθισμένο.

Φύλο του φοιτητή σε σχέση με την βαθμολογία

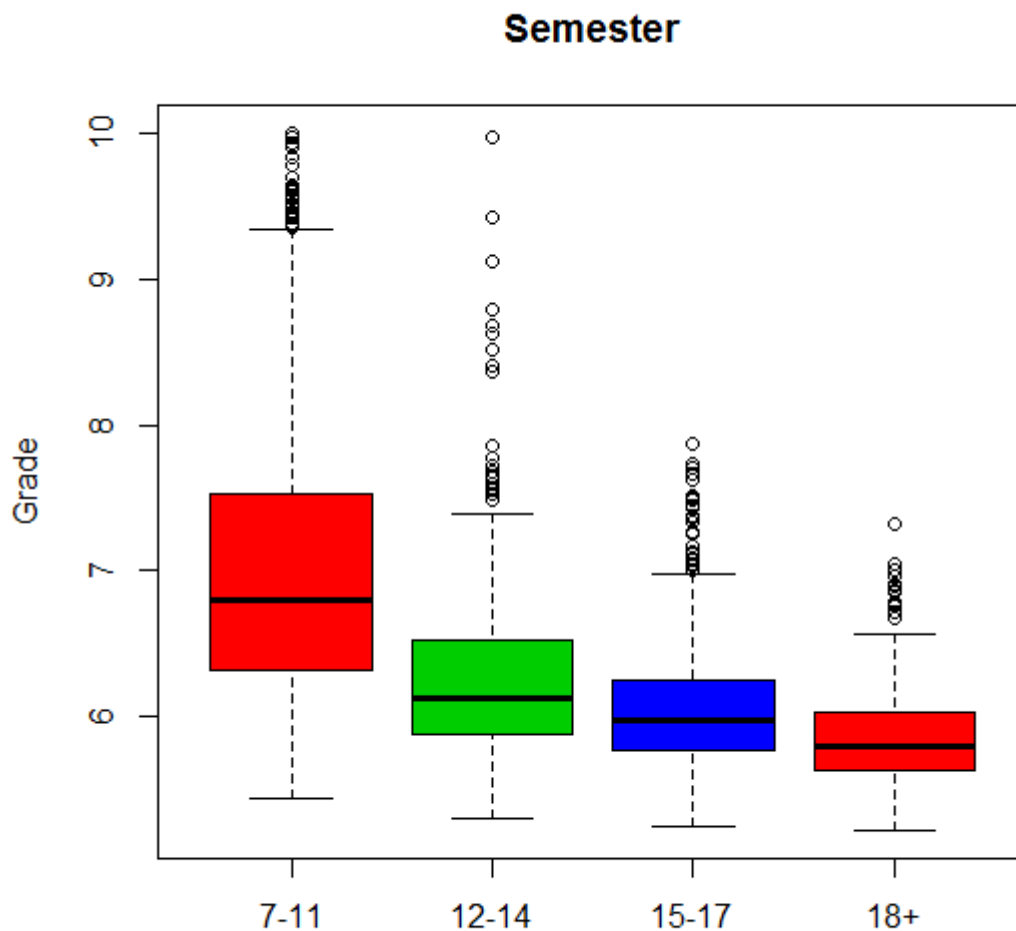
	min	1 τετ/ριο	Διαμεσος	Μέσος	3 τετ/ριο	Μέγιστος
Άντρας	5.220	5.815	6.170	6.646	6.8	10
Γυναίκα	5.240	5.840	6.130	6.311	6.560	9.980



Από τα παραπάνω παρατηρούμε μια ελάχιστη καλύτερη επίδοση στους άντρες φοιτητές. Αυτή βέβαια δεν φαίνεται να είναι και πολύ μεγάλη π.χ. συγκρίνοντας τους μέσους έχουμε διαφορά 0.34 μονάδων.

Εξάμηνα σε σχέση με την βαθμολογία

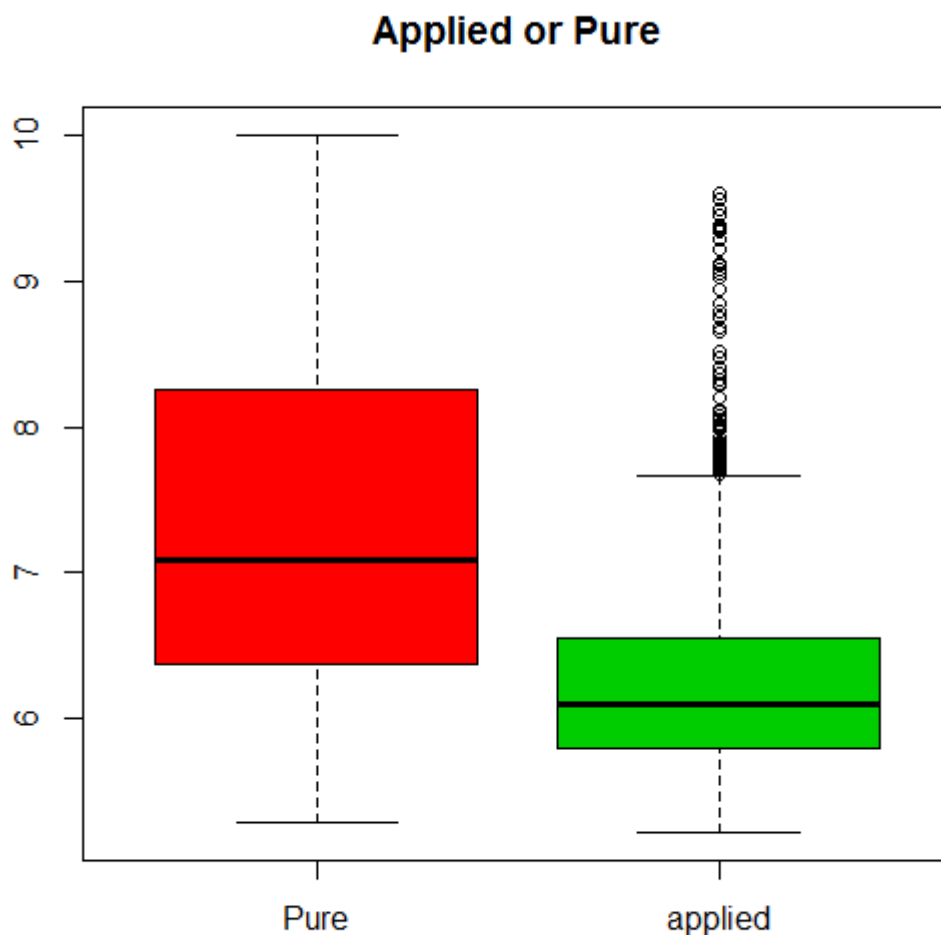
	Min	1 Τετ/ριο	Διάμεσος	Μέσος	3 Τετ/ριο	Max
7-11	5.440	6.315	6.800	7.042	7.522	10.000
12-14	5.300	5.9	6.125	6.245	6.520	9.980
15-17	5.240	5.760	5.970	6.065	6.250	7.870
18+	5.220	5.630	5.800	5.866	6.03	7.320



Η σχέση μεταξύ βαθμολογίας και στο κατά πόσο χρόνο χρειάζεται ο φοιτητής να ολοκληρώσει τις σπουδές του έχει ήδη ειπωθεί στο προηγούμενο μοντέλο, από τα παραπάνω σχήματα είναι προφανές ότι όσο πιο γρήγορα τελειώνει τόσο καλύτερα τα πηγαίνει στον βαθμό του, δηλαδή είναι επιμελής και στα δύο.

Κατεύθυνση σε σχέση με την βαθμολογία

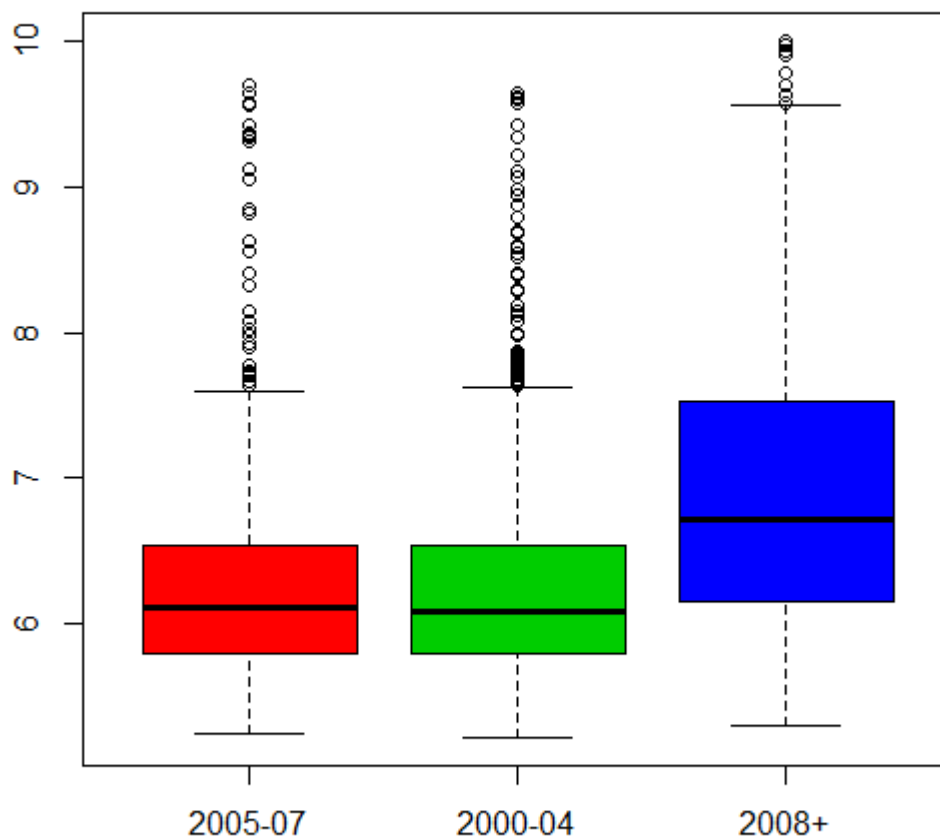
	Min	1 Τετ/ριο	Διάμεσος	Μέσος	3 Τετ/ριο	Max
Εφαρ/να	5.220	5.800	6.100	6.265	6.550	9.600
Θεωρ/κα	5.290	6.375	7.080	7.372 8	8.247	10.000



Από τα παραπάνω είναι φανερό ότι οι φοιτητές που επέλεξαν θεωρητική έχουν καλύτερες βαθμολογικά επιδόσεις.

Χρόνια εισαγωγής σε σχέση με την βαθμολογία

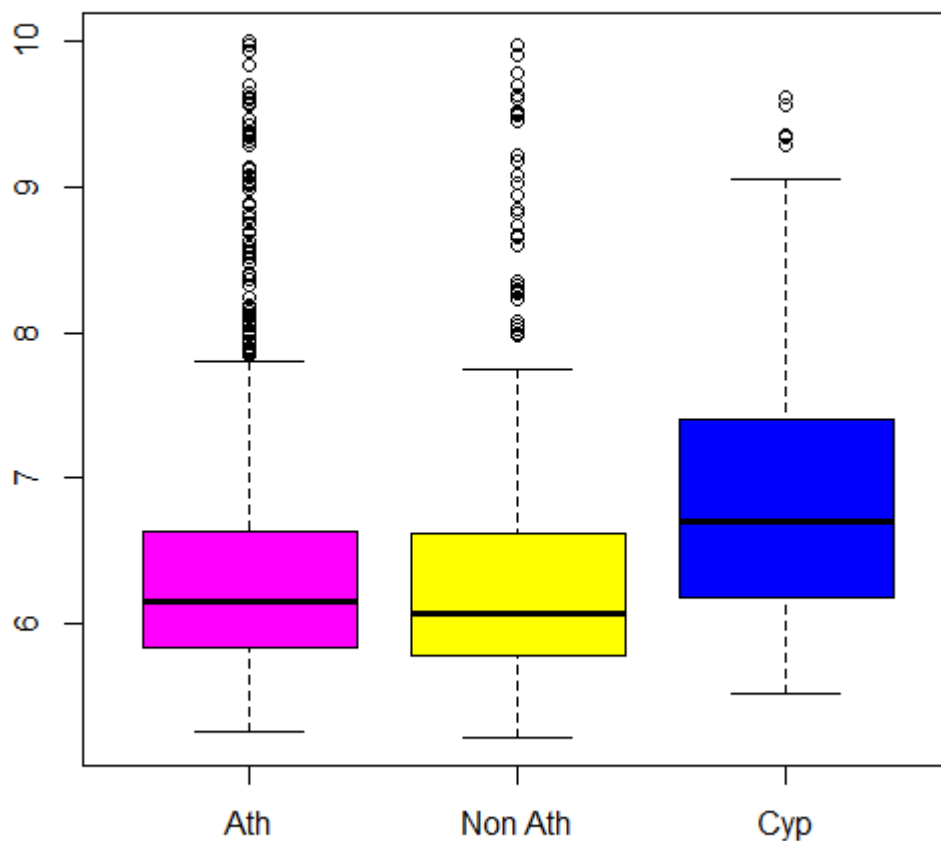
	Min	1 Τετ/ριο	Διάμεσος	Μέσος	3 Τετ/ριο	Max
2008-15	5.300	6.155	6.720	6.983	7.520	10.000
2005-07	5.240	5.800	6.110	6.333	6.532	9.700
2000-04	5.220	5.800	6.080	6.257	6.530	9.640



Σε σχέση με το έτος εισαγωγής και το βαθμό πτυχίου παρατηρούμε ότι στα έτη 2000-2007 (οι περίοδοι 2000-2004 & 2005-2007 μαζί) δεν έχουμε μεγάλες διαφορές , από εκεί και ύστερα όμως υπάρχει καλυτέρευση, εδώ ενδέχεται να συνδράμουν τα αίτια που αναφέραμε και σε ότι αφορά την σχέση επιπλέον εξάμηνο – χρονιά εισαγωγής, δηλαδή αύξηση πληροφορίας, πτυχιακή εξεταστική, διέξοδος από κρίση. **Σημαντικός ανασταλτικός παράγοντας** όμως είναι ότι ενδεχομένως οι όχι τόσο δυνατοί βαθμολογικά φοιτητές να μην έχουν ολοκληρώσει τις σπουδές τους, άλλωστε στο μέσο όρο η απόκλιση είναι λίγο περισσότερο από μισή μονάδα.

Καταγωγή σε σύγκριση με την βαθμολογία

	Min	1 Τετ/ριο	Διάμεσος	Μέσος	3 Τετ/ριο	Max
Αθήνα	5.250	5.830	6.150	6.377	6.630	10
Εκτός Αθήνας	5.220	5.780	6.070	6.343	6.620	9.970
Κύπρος	5.520	6.180	6.700	6.899	7.400	9.610



Όπως βλέπουμε οι Κύπριοι φοιτητές μοιάζει να έχουν μια καλύτερη βαθμολογική επίδοση σε σχέση με τους υπόλοιπους φοιτητές, ενώ οι άλλες δυο κατηγορίες δεν φαίνεται να έχουν ουσιαστική διαφορά.

4.2.1 Μοντελοποίηση

Το μοντέλο και αυτή τη φορά θα επιλεχθεί με την διαδικασία που ορίσαμε προηγουμένως, δηλαδή θα κρατήσουμε το τελικό μοντέλο όπου όλες οι περιγραφικές μεταβλητές θα είναι στατιστικά σημαντικές σε επίπεδο στατιστικής σημαντικότητας $\alpha=0.05$. Εδώ βέβαια η εξαρτημένη είναι συνεχής και θα υποθέσουμε κανονική κατανομή, όπου η linkfunction είναι πιο άμεση γνωστή μας από το απλό γραμμικό μοντέλο

$$Y = X' \cdot \beta$$

οπότε λόγω των random effect τελικά το μοντέλο είναι το

$$Y = X' \cdot \beta + Z' \cdot r$$

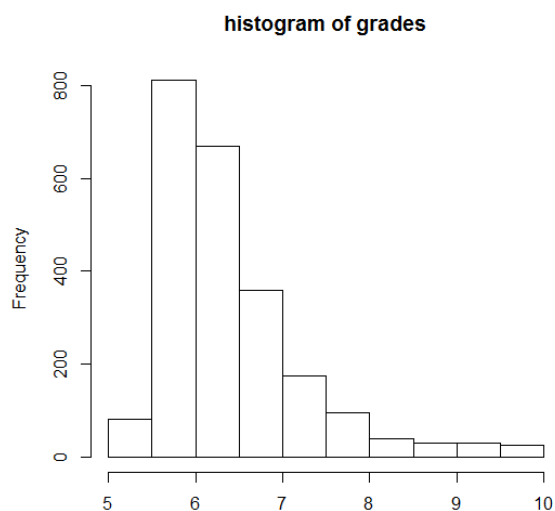
Όπου Y η εξαρτημένη τ.μ. του βαθμού πτυχίου.

Όπως και πριν β , ο πίνακας των συντελεστών των ανεξαρτήτων μεταβλητών, ενώ είναι τα random effects.

Και εδώ παρόμοια το μοντέλο που θα υποθέσουμε θα είναι της μορφής

$$Y = \beta_0 + r_i + \beta_1 \cdot X_1 + \dots + \beta_p \cdot X_p$$

με τις τυχαίες επιδράσεις να ακολουθούν Κανονική $N(0, u^2)$.



Με βάση τη μεθοδολογία όπου αρχικά εξετάζουμε ατομικά τις περιγραφικές μεταβλητές έχουμε τον παρακάτω πίνακα:

	Εκτίμηση β	Τυπική απόκλιση	Μπαίνει στο μοντέλο
Φύλο	-0.15806 (γυναίκα)	0.03432	Ναι
Επιπλέον Χρόνος	-0.102221	0.003364	Ναι
Εποχή	0.09723 (2005-008)	0.03856	
	0.72634 (>2008)	0.04725	Ναι
Καταγωγή	-0.03715 (Εκτός Αθήνας)	0.03891	
	0.52948 Κύπρος	0.07105	Ναι
Κατεύθυνση	1.10282 (θεωρητική)	0.04850	Ναι

Παρατηρούμε και εδώ ότι όλες οι μεταβλητές είναι στατιστικά σημαντικές στο μοντέλο

$$Y = \beta_0 + r_i + \beta_1 \cdot X_1$$

Οπότε στο επόμενο βήμα, θα είναι της μορφής

$$Y = \beta_0 + r_i + \beta_1 \cdot \text{Φύλο} + \beta_2 \cdot \text{Εποχή1} + \beta_3 \cdot \text{Εποχή2} + \beta_4 \cdot \text{Κατευθυνση1} + \beta_6 \cdot \text{ΕπιπλέονΧρόνος} + \beta_5 \cdot \text{Καταγωγή2} + \beta_7 \cdot \text{Καταγωγή1}$$

με $r_i \sim N(0, u^2)$. Στο οποίο θα συνεχίσουμε με backwardelim/tion.

Παρατηρούμε τα παρακάτω αποτελέσματα:

Μεταβλητή	Εκτίμηση β	Τυπική απόκλιση	Pvalue
Σταθ. Όρος	6.975123	0.040028	σχεδόν 0
Επιπλέον Χρόνος	-0.086245	0.003521	σχεδόν 0
Φύλο	-0.152536 (γυναίκα)	0.026663	σχεδόν 0
Εποχή	-0.099763 (2005-07)	0.031696	σχεδόν 0
	0.154432 (2008+)	0.042329	
Καταγωγή	-0.035340	0.030099	σχεδόν 0
	0.294768 (Κύπρος)	0,04	
Κατεύθυνση	0.867641 (θεωρητικά)	0.042002	σχεδόν 0

με $r_i \sim N(0,0.1614)$.

Επεξήγηση Αποτελεσμάτων

Επιπλέον εξάμηνα για την απόκτηση πτυχίου $\beta=-0,086246$

Υπάρχει αρνητική σχέση, δηλαδή όσο αργείς να τελειώσεις τόσο δεν τα πας καλά και με το βαθμό σου. Πράγμα που το είχαμε παρατηρήσει και στο μοντέλο 1. Δείχνει ουσιαστικά μια συνέπεια όσον αφορά τους ας πούμε επιμελείς φοιτητές. Πιο συγκεκριμένα για κάθε χρόνο (2 εξάμηνα) επιπλέον που χρειάζεται κάποιος για την απόκτηση πτυχίου έχουμε -0.17249 αφαίρεση από το βαθμό που θα παίρνει ένας άλλος φοιτητής με ίδια όλα τα άλλα χαρακτηριστικά. Πχ ο δεύτερος αναμένεται να πάρει πτυχίο με 7 ο άλλος αναμένεται με 6.82.

Φύλο $\beta=-0.152536$ (γυναίκα)

Υπάρχει αρνητική σχέση και εδώ δηλαδή φαίνεται ότι οι φοιτήτριες, πιθανόν να έχουν χαμηλότερες βαθμολογικές επιδόσεις. Συγκεκριμένα αν μια κοπέλα που έχει βαθμό 8 και ίδια χαρακτηριστικά με ένα αγόρι σε όλα τα άλλα, ενδέχεται να πάρει τελικό βαθμό 7.152536.

Χρόνος Εισαγωγής, $\beta(2005-07)=-0.099763$ και $\beta=0.154432$ (2008-2015)

Παρατηρούμε με βάση το αν κάποιος έχει μπει στην σχολή από το 2008 και έπειτα υπάρχει μια καλυτέρευση στις βαθμολογίες. Ενώ το 2005-07 ήταν χειρότερο από το 2000-2004 όσον αφορά αυτό τον τομέα, σε ελάχιστο βέβαια βαθμό αφού το -0.099763 είναι αμελητέα ποσότητα. Με ίδια λογική με τα προηγούμενα παραδείγματα αν έχουμε ένα φοιτητή που αναμένεται να πάρει 7 και μπήκε πριν το 2005, ο φοιτητής με τα ίδια χαρακτηριστικά του 2008 αναμένεται να πάρει 7.15. Εδώ υπάρχει ο κίνδυνος μεροληψίας αφού από ένα έτος και μετά οι ενδεχομένως όχι καλοί βαθμολογικά δεν έχουν ολοκληρώσει ακόμα τις σπουδές τους.

Καταγωγή, $\beta= -0.035340$ (Εκτός Αθηνών) και 0.294768 (Κύπρος)

Παρατηρούμε καλύτερη επίδοση από τους Κύπριους φοιτητές με +0,294 στην τελική βαθμολογία σε σύγκριση με κάποιον που έχει ίδια όλα τα άλλα χαρακτηριστικά. Για την σύγκριση Αττική –Εκτός Αττικής οι εκ Αττικής έχουν ελάχιστα καλύτερες επιδόσεις, οι οποίες όμως είναι αμελητέες βαθμολογικά.(δέκατα)

Κατεύθυνση $\beta=0.867641$ (Θεωρητικά)

Υπάρχει θετική σχέση, δηλαδή τα άτομα που επιλέγουν θεωρητική κατεύθυνση ενδέχεται να έχουν βαθμολογική επιτυχία σε σχέση με τα άτομα των εφαρμοσμένων και μάλιστα κατά +0.867641.

Συγκεντρωτικά

- Όσοι τελειώνουν γρήγορα παίρνουν και καλό βαθμό
- Οι Κύπριοι φοιτητές έχουν καλύτερη βαθμολογική επίδοση (σαν σύνολο)
- Τα άτομα της Θεωρητικής τελειώνουν με καλύτερο βαθμό
- Οι άντρες φοιτητές έχουν ελάχιστα καλύτερες επιδόσεις

4.3 Τρίτο Μοντέλο Μοντελοποίηση με χρήση Ανάλυση Επιβίωσης με καταληκτικό γεγονός τον χρόνο απόκτησης πτυχίου

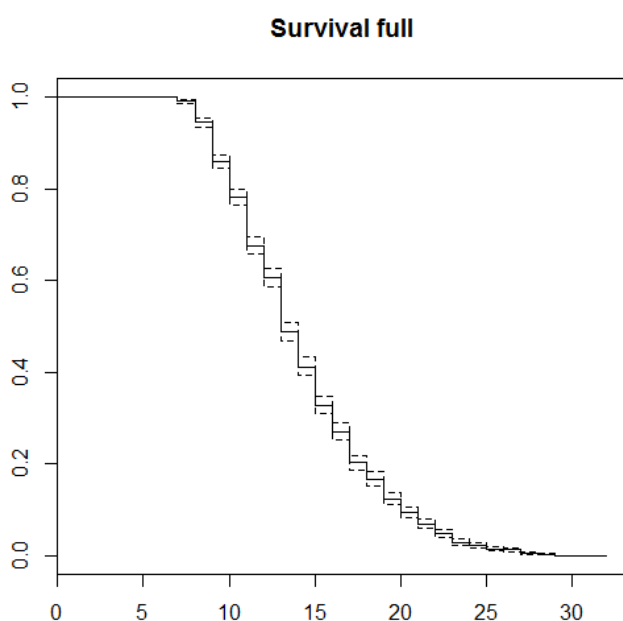
4.3.1 Περιγραφή Μοντέλου

Σε αυτήν την Τρίτη μοντελοποίηση θα χρησιμοποιήσουμε την θεωρία της ανάλυσης επιβίωσης, θεωρώντας ως τελικό γεγονός το συμβάν του πτυχίου. Στην πράξη είναι ένας εναλλακτικός τρόπος ανάλυσης του πρώτου μοντέλου όπου εξαρτημένη ορίσαμε τον χρόνο απόκτησης πτυχίου. Το μοντέλο με το οποίο θα δουλέψουμε είναι το *Coxproportionalmodel*, (μοντέλο αναλογικής επιβίωσης του Cox) το οποίο θεωρεί ότι η *hazardfunction* δηλαδή η συνάρτηση ρίσκου του να συμβεί η τελική πράξη απόκτησης του πτυχίου στον άμεσο χρόνο είναι ανάλογη της *hazardfunction* του μοντέλου με τις αρχικές τιμές των μεταβλητών, δηλαδή

$$h_i(t) = \exp(X' \cdot \beta) \cdot h_0(t)$$

Η μεθοδολογία και η ανάλυση δεν είναι τόσο αναλυτική όσο τα προηγούμενα μοντέλα, σκοπός είναι να γίνει μια σύγκριση με το πρώτο μοντέλο και αν κατά πόσο παρατηρούμε όμοια πράγματα.

Στόχος μας ήταν αρχικά οι μη πτυχιούχοι να χρησιμοποιηθούν ως *censored* (λογοκριμένα) στοιχεία, όμως λόγω μεγάλης έλλειψης δύο κύριων στοιχείων όπως κατεύθυνσης, έως τώρα βαθμολογίας μας κάνει να μην τους περιλάβουμε στο μοντέλο. Άρα το μοντέλο μας θα περιέχει μόνο *uncensored* μεταβλητές δηλαδή μεταβλητές που ξέρουμε πότε συνέβη το τελικό γεγονός.

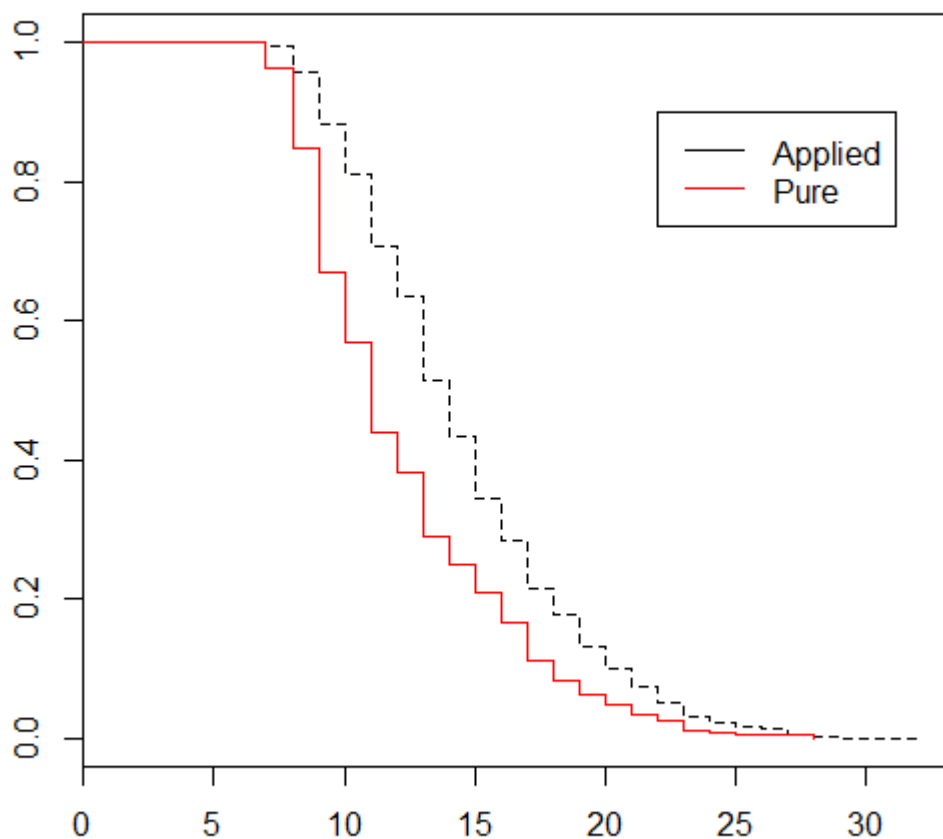


Το παραπάνω γράφημα ουσιαστικά το έχουμε αναλύσει όταν περιγράψαμε τα ποσοστημόρια του χρόνου απόκτησης πτυχίου, εδώ είναι ας πούμε από την 'ανάποδη' καθώς είναι το γράφημα επιβίωσης $S(y)=P(T>y)$. Η διάμεσος είναι γύρω στο 13, ενώ πχ παρατηρούμε ότι μόνο το 20% δεν έχει πάρει πτυχίο από 16.5 εξάμηνα και έπειτα.

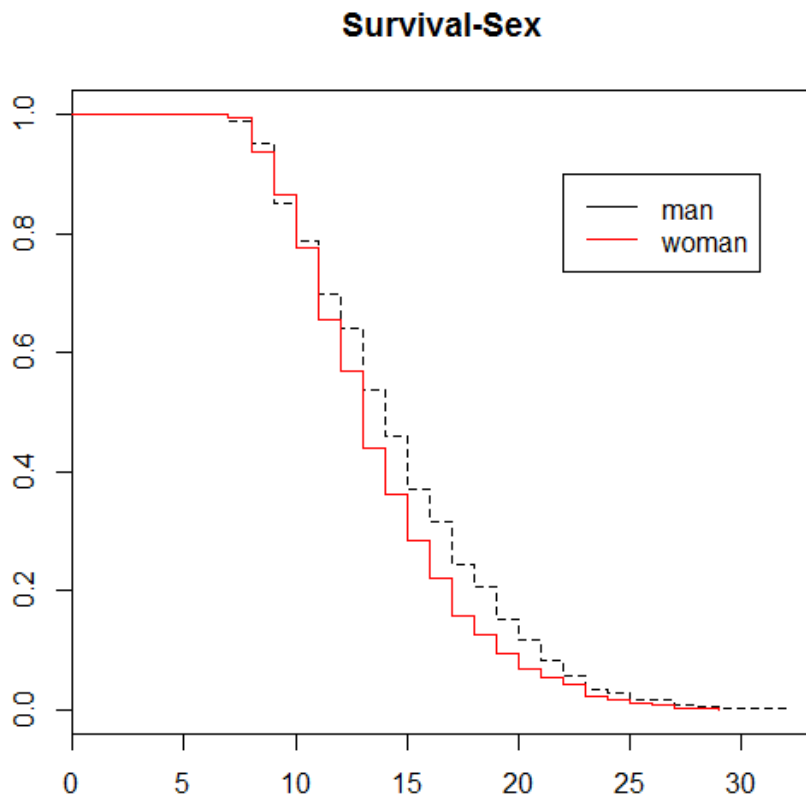
Θα ελέγξουμε γραφικά την υπόθεση αναλογικότητας της επιβίωσης για κάθε περιγραφική μεταβλητή, ώστε να δούμε αν το αναλογικό μοντέλο δουλεύει καλώς.

Περιγραφικές Μεταβλητές	Βαθμός
	Κατεύθυνση
	Φύλο
	Εποχή
	Καταγωγή

Survival-Pure or appl



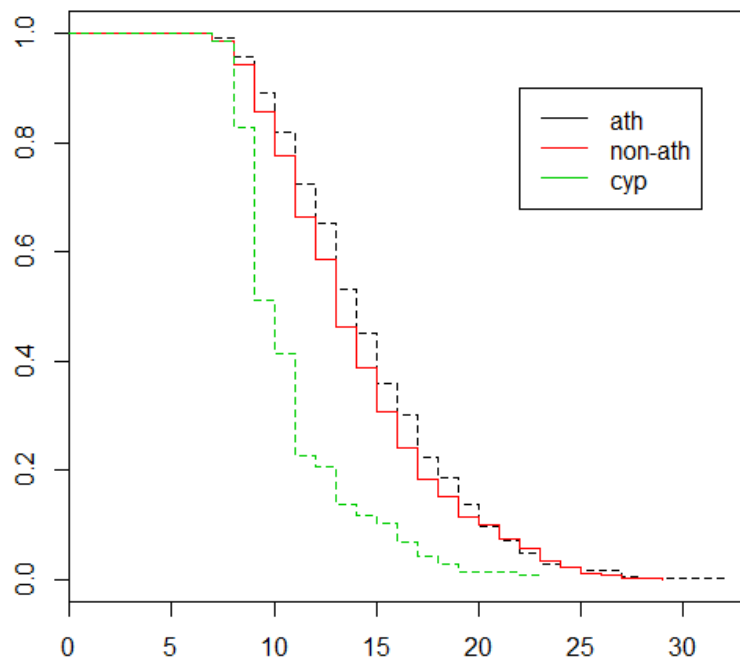
Όσον αφορά την Κατεύθυνση παρατηρούμε ότι οι φοιτητές των θεωρητικών τελειώνουν πιο γρήγορα, αφού η γραμμή επιβίωσης τελειώνει πιο γρήγορα και παίρνει τιμές και πιο γρήγορα, ενώ βλέπουμε μια αναλογικότητα στις δυο γραμμές.



Για το αν οι φοιτήτριες ή οι φοιτητές έχουν διαφορά βλέπουμε ότι η γραμμή επιβίωσης τελειώνει πιο γρήγορα για τα κορίτσια όπως και μια αναλογικότητα από το 13 και έπειτα.

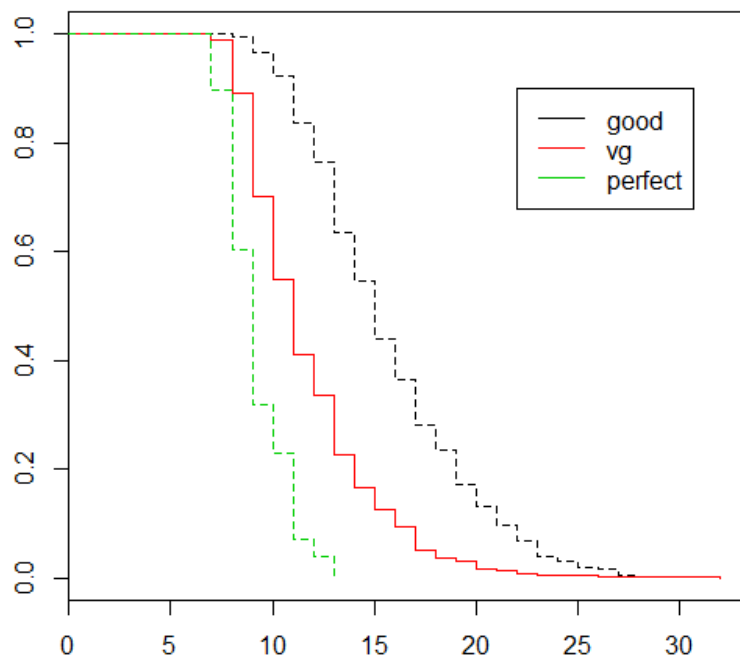
Επειδή τα συμπεράσματα για το αν υπάρχει διαφορά στα εξάμηνα έχουν γίνει στα προηγούμενα μοντέλα από εδώ και κάτω στα σχήματα θα αναφέρουμε μόνο την αναλογικότητα.

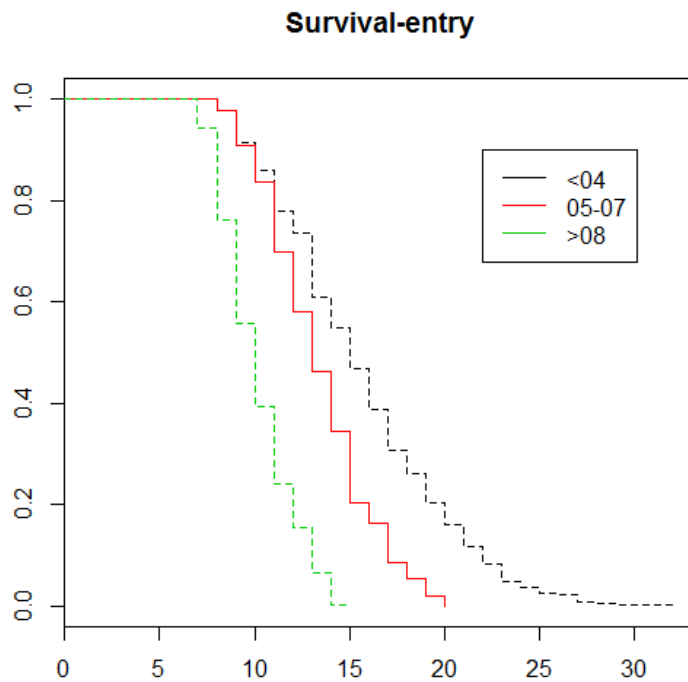
Survival-Cyprus-Athens-Non-Athens



Για τον τόπο μόνιμη κατοικίας παρατηρούμε μια αναλογικότητα κυρίως όσον αφορά Κύπριος και μη Κύπριος, οι άλλες δυο κατηγορίες είναι στα ίδια επίπεδα.

Survival-Grade





Η αναλογικότητα είναι προφανής και στις τρεις κατηγορίες και στα δύο σχήματα, αντίστοιχα. Οπότε και μπορούμε να χρησιμοποιήσουμε τα αναλογικά μοντέλα του Cox.

4.3.2 Μοντελοποίηση με το μοντέλο του Cox

Με την ίδια τακτική με τα προηγούμενα μοντέλα, θα ερευνήσουμε πρώτα το τι κάνει η κάθε μια μεταβλητή μόνη της. Έχουμε ξεχωριστά για την κάθε μεταβλητή:

	Εκτίμηση β	Εκτίμηση εκθετική τιμή β	Μπαίνει στο μοντέλο
Φύλο	0.21365 (γυναίκα)	1.23819	Ναι
Βαθμός	0.89214	2.44035	Ναι
Εποχή	0.6977 (2005-008)	2.0092	
	1.9465 (>2008)	7.0039	Ναι
Καταγωγή	0.10314 (Εκτός Αθήνας)	1.10864	
	1.07497 Κύπρος	2.92989	Ναι

Κατεύθυνση	0.48611 (θεωρητική)	1.62599	Ναι
------------	------------------------	---------	-----

Οπότε όπως και προηγουμένως όλες οι μεταβλητές μπαίνουν στο μοντέλο.

Για το πλήρες αυτό μοντέλο βρίσκουμε τις παρακάτω εκτιμήσεις:

	Εκτίμηση β	Εκτίμηση εκθετική τιμή β	p-value
Φύλο	0.29186 (γυναίκα)	1.33892	σχεδόν 0
Βαθμός	0.81615	2.26177	σχεδόν 0
Εποχή	0.64161 (2005-008)	1.89954	
	1.63019 (>2008)	5.10486	σχεδόν 0
Καταγωγή	0.10524 (Εκτός Αθήνας)	1.11098	
	1.04339 Κύπρος	2.83882	σχεδόν 0
Κατεύθυνση	-0.19871 (θεωρητική)	0.81979	

Τελικά όλες οι μεταβλητές παραμένουν στο μοντέλο, αφού είναι στατιστικά σημαντικές. Το μοντέλο μας δηλαδή θα είναι της μορφής

$$h_i(t) = \exp(\beta_1 \cdot \text{Φύλο} + \beta_2 \cdot \text{Εποχή1} + \beta_3 \cdot \text{Εποχή2} + \beta_4 \cdot \text{Κατευθυνση1} + \beta_5 \cdot \text{Βαθμός} + \beta_6 \cdot \text{Καταγωγή2} + \beta_7 \cdot \text{Καταγωγή1}) \cdot h_0(t)$$

Συμπεράσματα:

Φύλλο $\exp(\beta) = 1.33892$ (γυναίκα)

Οι φοιτήτριες έχουν μεγαλύτερη πιθανότητα να πάρουν πιο γρήγορα πτυχίο. Συγκεκριμένα με βάση το μοντέλο η πιθανότητα να πάρουν αυτή τη στιγμή πτυχίο σε σχέση με κάποιον φοιτητή είναι αυξημένη κατά περίπου 34%.

Βαθμός $\exp(\beta) = 2.26177$

Για κάθε ολόκληρη βαθμολογική μονάδα η πιθανότητα να πάρει κάποιος πτυχίο αυτή τη στιγμή, όπως ορίζεται από το μοντέλο είναι η διπλάσια +26%.

Χρονιά εισαγωγής $\exp(\beta)=1.89954$ (05-07) και 5.10486 (2008>)

Όπως είδαμε και στο μοντέλο 1 η σειρά πάει από το καλύτερο στο χειρότερο 2008-2015, 2005-2007 και πριν το 2004. Εδώ παρατηρούμε σχεδόν 5πλάσια πιθανότητα για τους νέους φοιτητές σε σύγκριση με αυτούς του 2004 και πριν. Ενώ στο άλλο ζευγάρι έχουμε αύξηση κατά 90% στην πιθανότητα.

Κατεύθυνση $\exp(\beta)= 0.81979$ (θεωρητική)

Εδώ υπάρχει αρνητική σχέση (φαίνεται από το αρνητικό πρόσημο στην εκτίμηση του β), όπως και στο Μοντέλο 1 η πιθανότητα να πάρει κάποιος αυτή τη στιγμή πτυχίο τελικά μειώνεται κατά $(1-0.81979 \text{ περίπου})$ 20% αν έχει επιλέξει θεωρητική κατεύθυνση. Με βάση όλα τα προηγούμενα φαίνεται ότι η κατεύθυνση σε απόλυτο βαθμό επηρεάζει ως προς την δυσκολία και όχι ότι γενικά οι φοιτητές της θεωρητικής αργούνε να την τελειώσουν.

Τόπος μόνιμης κατοικίας, $\exp(\beta)= 1.11098$ (εκτός Αθηνών) 2.83882 Κύπρος

Και εδώ παρατηρούμε ότι και στο μοντέλο 1. Για τους Κύπριους έχουμε $\times 2.83882$ στον κίνδυνο να πάρει κάποιος πτυχίο σε σχέση με τους εκ Αττικής, ενώ στους εκ Αττικής με τους εκτός είναι μόλις +11% πιθανότητα για τους εκτός Αθηνών.

Πρόβλεψη προσέλευσης αναμενόμενου αριθμού φοιτητών σε μια εξέταση με χρήση ανάλυσης χρονοσειρών

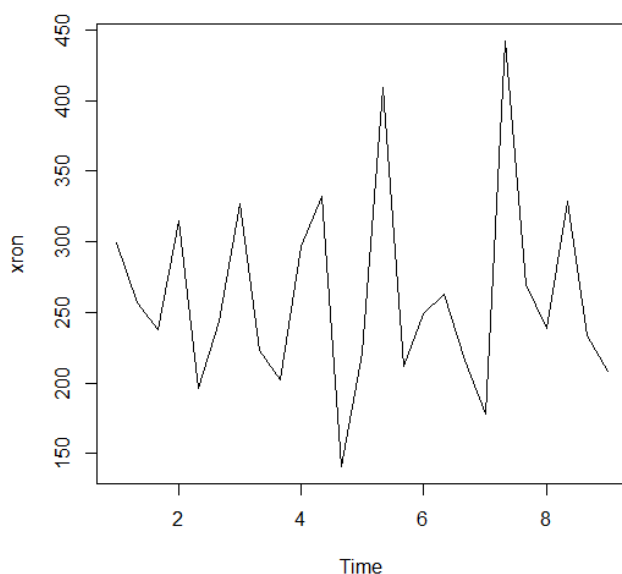
Το να προβλέψουμε το πόσοι φοιτητές αναμένονται να έρθουν σε κάποιο μάθημα την επόμενη εξέταση είναι πολύ χρήσιμο, ώστε να μπορέσουμε να είμαστε προετοιμασμένοι για το πλήθος των επιτηρητών, το πλήθος των αιθουσών, τα φύλλα Α4 που πρέπει να εκτυπωθούν τα θέματα, όπως και τις κόλες αναφοράς που θα χρειαστούν για να γράψουν τα θέματα οι φοιτητές, μέχρι και ένα αναμενόμενο κόστος που θα έχει η εξέταση.

Θα χρησιμοποιήσουμε ως δεδομένα για κάθε μάθημα ξεχωριστά το πόσοι φοιτητές προσήλθαν στην κάθε εξέταση, ανάλογα και το πότε ήταν αυτή Χειμερινό Εξάμηνο/Εαρινό Εξάμηνο /Επαναληπτική Εξέταση Σεπτεμβρίου και με χρήση της ανάλυσης χρονοσειρών θα δώσουμε μια πρόβλεψη, για το επόμενο διάστημα. Ενδεικτικά θα αναφέρουμε το μάθημα της Στατιστικής.

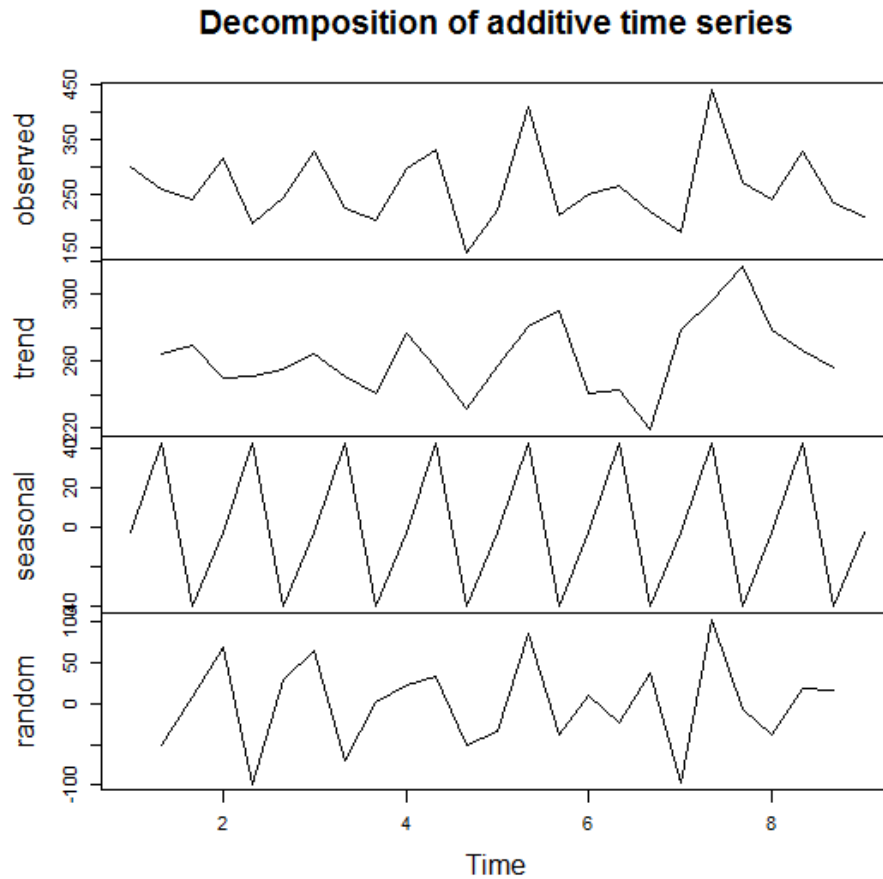
Στατιστική Ι (541)

Θα ελέγχουμε αρχικά τον αριθμό προσέλευσης σε κάθε εξέταση στο υποχρεωτικό μάθημα της Στατιστικής. Το δείγμα προέρχεται από το χρονικό διάστημα Χειμερινό εξάμηνο 2007 έως και το Χειμερινό εξάμηνο 2015. Να σημειώσουμε ότι στο διάστημα αυτό έχουμε εξετάσεις σε όλα τα εξάμηνα και το Σεπτέμβρη. Το δείγμα μας (σε μορφή χρονοσειράς) είναι το παρακάτω

$$\{x_t\}_n = \{ 299, 257, 238, 315, 196, 244, 327, 223, \\ 202, 297, 332, 141, 221, 409, 212, 249, 263, 216, 178, \\ 442, 270, 239, 329, 233, 208 \}$$

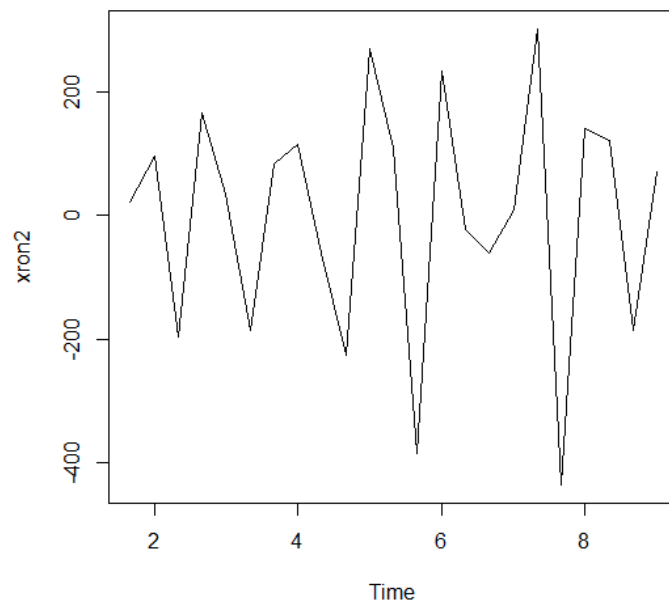


Σχηματικά ξεχωρίζοντας την τάση, την εποχικότητα και το τυχαίο, έχουμε το παρακάτω

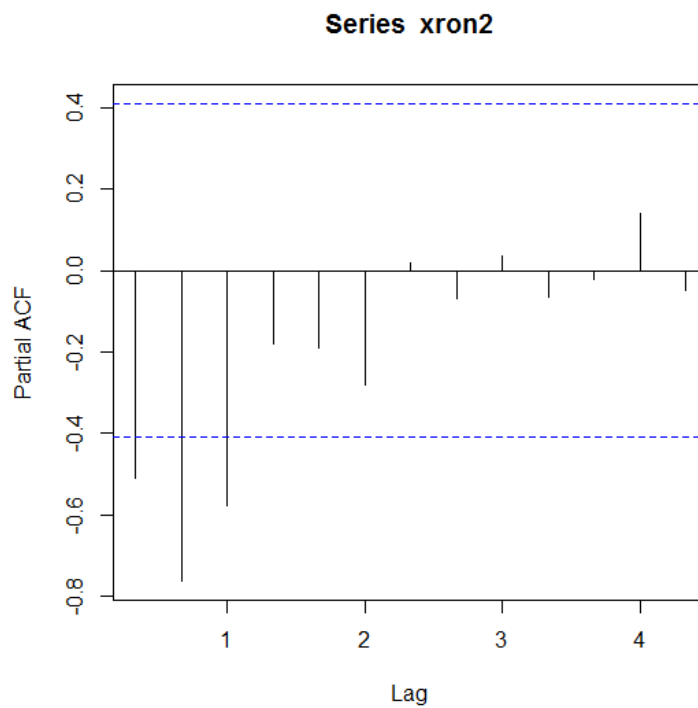


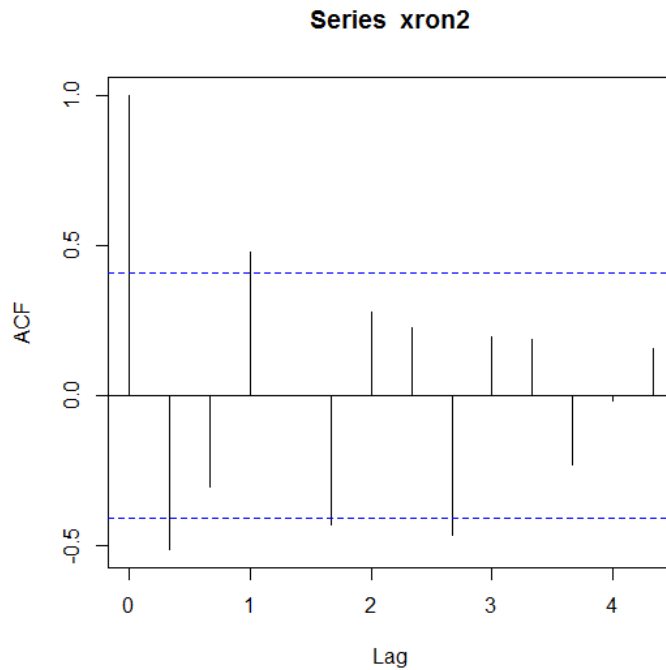
Όπως γνωρίζουμε οι τρεις εξετάσεις μιας χρονιάς παρουσιάζουν μεγάλη σχέση μεταξύ τους, η εποχικότητα εδώ θα είναι η κάθε χρονιά ξεχωριστά. Από το παραπάνω παρατηρούμε τάσεις, οι τάσεις είναι λογικές αφού οι καθηγητές δεν είναι σταθεροί στο συγκεκριμένο μάθημα, ο φοιτητής ενδέχεται να επιλέγει πότε θα προσφέρεται από τον καθηγητή που προτιμά.

Θα χρησιμοποιήσουμε το εποχιακό ARIMA μοντέλο. Πρώτη δουλειά μας να απαλλαγούμε από την τάση. Στη συνέχεια δίνεται το γράφημα των δεύτερων διαφορών.



Θα ελέγξουμε με τα γραφήματα acf και pacf τις συσχετίσεις και όσον αφορά τον κινούμενο μέσο.





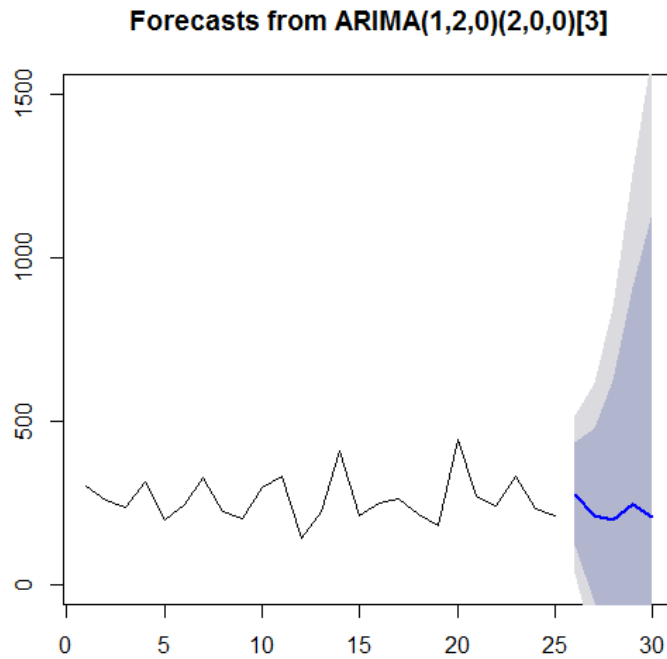
Παρατηρούμε από το διάγραμμα racf ότι υπάρχει σχέση μέσα στην ίδια την εποχή, αλλά και του χρόνου με τον επόμενο. Αυτό είναι κάτι που το ξέρουμε έτσι κι αλλιώς, πέρα του ότι άμα δεν το περάσεις το Χειμερινό, με μεγάλη πιθανότητα θα το δώσεις την επόμενη εξέταση και όμοια για το εαρινό, κάποιος που απέτυχε το Σεπτέμβρη ενδέχεται να τολμήσει την επόμενη χρονιά.

Με βάση αυτά θα υποθέσουμε ένα εποχικό μοντέλο

$$arima(1,2,0) (2, 0, 0)_3$$

Το οποίο δίνει τις παρακάτω σημειακές προβλέψεις για τις επόμενες πέντε εξετάσεις

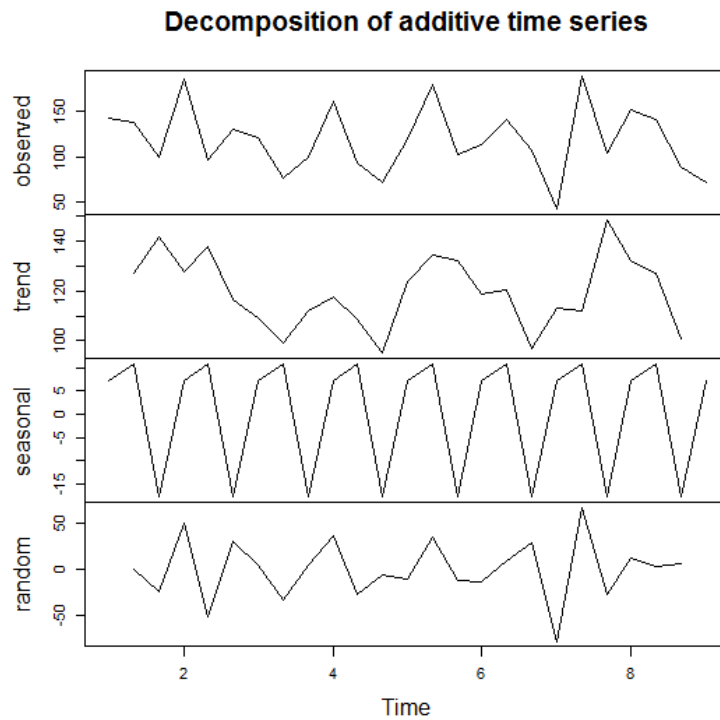
Εαρινό 2015	Σεπτέμβριος	Χειμερινό 2016	Εαρινό 2016	Σεπτέμβριος
275	211	199	243	205



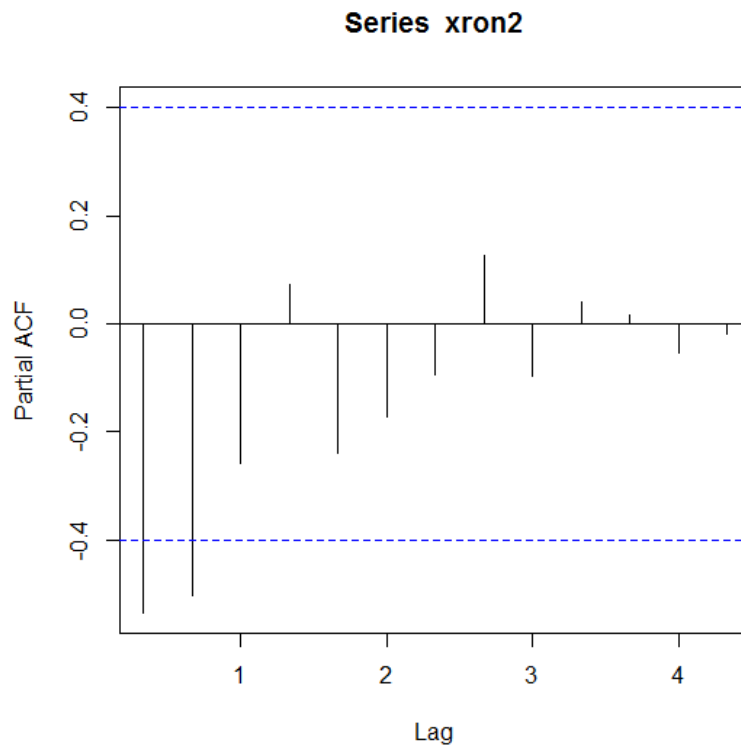
Ένα άλλο ερώτημα είναι πόσοι φοιτητές έρχονται στην εξέταση χωρίς να είναι καλά προετοιμασμένοι, θα ορίσουμε την ιδιότητα αυτή στα άτομα που έγραψαν από 2 και κάτω. Για την Στατιστική Ι έχουμε λοιπόν την παρακάτω χρονοσειρά στον ίδιο χρόνο με πριν.

$\{y_t\}_n = \{143, 139, 100, 186, 97, 130, 122, 76, 99, 161, 93, 72, 120, 180, 103, 113, 141, 108, 42, 190, 104, 152, 141, 89, 72\}$

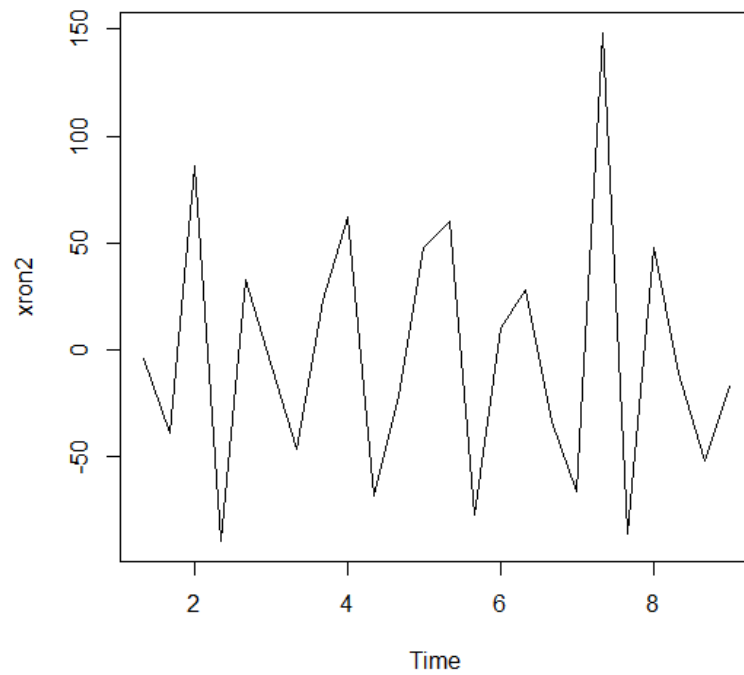
Παρακάτω δίνεται το διάγραμμα με όλα τα χαρακτηριστικά της χρονοσειράς



Είναι λογικό υπάρχει μια λογική σχέση με την προηγούμενη χρονοσειρά του συνολικού πλήθους που ήρθε στην εξέταση και στην τάση και εποχικότητα έχουμε παρόμοια συμπεράσματα. Όπως φαίνεται στο διάγραμμα `pacf` η συσχέτιση μεταξύ των τιμών μέσα στο έτος υπάρχει, αλλά μεταξύ των ετών όχι απαραίτητα.



Το διάγραμμα πρώτων διαφορών μας δίνει το παρακάτω, χωρίς τάση



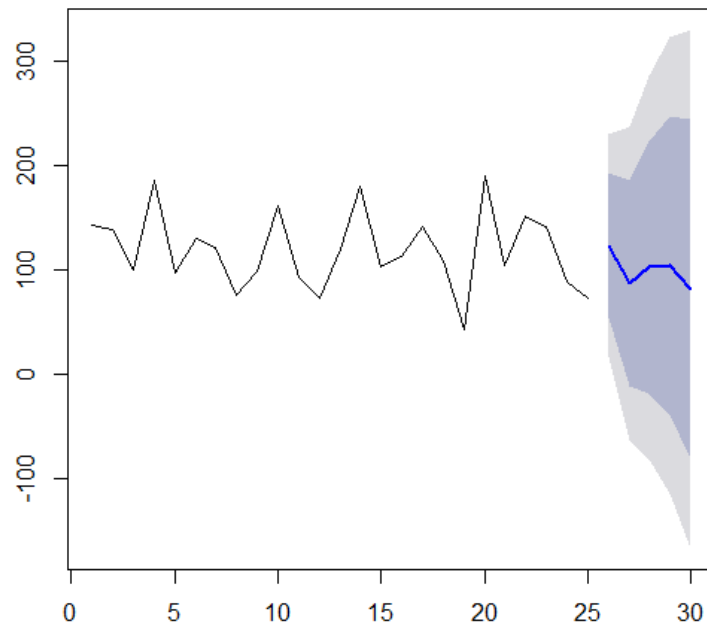
Κατάλληλο μοντέλο φαίνεται να είναι το

$$arima(0,1,0) (2,0,0)_3$$

Έχουμε τις παρακάτω προβλέψεις

Εαρινό 2015	Σεπτέμβριος	Χειμερινό 2016	Εαρινό 2016	Σεπτέμβριος
123	87	102	104	81

Forecasts from ARIMA(0,1,0)(2,0,0)[3]



5. Παράρτημα

5.1 Παράρτημα Α-Υπολογισμός κόστους φοιτητή που έγραψε από 2 και κάτω

Για να υπολογίσω έναν μέσο κόστος (λειτουργικά έξοδα) από τα δεδομένα μου για τα άτομα που μπορεί να θεωρηθούν από αδιάβαστα έως όχι τόσο καλά προετοιμασμένα (εκτός περιπτώσεων όπου η αποτυχία μπορεί να προέρχεται από άλλους παράγοντες πχ όχι τόσο καλή δουλειά από το μέρος του διδάσκοντα) έχω τα παρακάτω βήματα:

-Από τον πίνακα κάθε μαθήματος παίρνω τον μέσο όρο του πόσοι φοιτητές προσέρχονται στην κάθε εξέταση (χειμερινό/εαρινό/επαναληπτική).

-Από τον πίνακα του ίδιου μαθήματος με τα ποσοστά των φοιτητών που πήραν βαθμό κάτω από 3 (0-2) (δίνονται παρακάτω) παίρνω τον αντίστοιχο μέσο όρο

-Πολλαπλασιάζω τους δύο πάνω μέσο όρους με το κόστος εξέτασης(π.χ. σελίδες Α4, ρεύμα, χρόνος βαθμολόγησης, εργατώρες για την επιτήρηση κλπ)

Ποσοστά φοιτητών που γράψανε από 3 και κάτω ανά μάθημα ανά έτος

Απειροστικός Λογισμός Ι

	Χειμερινό	Εαρινό	Σεπτέμβριος
2000	12,15%	-	35,71%
2001	55,56%	-	34,74%
2002	37,70%	-	56,55%
2003	40,19%	-	57,03%
2004	32,81%	-	48,06%
2005	47,33%	-	43,04%
2006	48,48%	-	60,28%
2007	54,00%	56,94%	38,12%
2008	64,79%	34,66%	56,94%
2009	41,84%	27,10%	38,83%
2010	46,52%	46,21%	30,88%
2011	54,34%	22,73%	29,28%
2012	50,73%	26,56%	58,36%
2013	54,35%	54,74%	29,90%
2014	66,13%	39,84%	36,33%
2015	39,81%	*	*

Πιθανότητες Ι

	Χειμερινό	Εαρινό	Σεπτέμβρης
2000	-	-	-
2001	43,33%	-	47,06%
2002	55,30%	-	63,33%
2003	36,99%	-	53,36%
2004	40,90%	-	73,72%
2005	33,96%	-	22,66%
2006	48,67%	-	39,89%
2007	36,82%	-	42,61%
2008	39,65%	53,66%	50,35%
2009	43,08%	68,87%	60,53%
2010	55,31%	21,92%	72,88%
2011	61,78%	60,34%	52,15%
2012	70,26%	60,76%	36,36%
2013	45,16%	23,26%	71,66%
2014	46,46%	39,58%	47,73%
2015	55,51%	-	-

Στατιστική Ι

	Χειμερινό	Εαρινό	Σεπτέμβριος
2002	80,00%	-	80,00%
2003	60,15%	-	52,81%
2004	49,73%	57,89%	50,00%
2005	19,70%	62,73%	40,00%
2006	67,87%	-	44,14%
2007	52,17%	45,91%	57,98%
2008	40,95%	50,51%	46,72%
2009	62,69%	65,92%	50,99%
2010	45,79%	71,99%	48,94%
2011	45,70%	55,99%	51,42%
2012	54,62%	46,39%	50,00%
2013	76,40%	57,01%	61,48%
2014	36,40%	57,14%	61,80%
2015	65,38%	*	*

Πραγματική Ανάλυση Ι

	Χειμερινό	Εαρινό	Σεπτέμβριος
2001	-	50,00%	50,00%
2002	-	35,29%	66,04%
2003	33,33%	31,19%	38,98%
2004	39,36%	30,58%	27,39%
2005	44,96%	15,58%	28,09%
2006	63,47%	33,11%	22,99%
2007	31,21%	19,34%	21,09%
2008	35,34%	55,25%	39,43%
2009	31,33%	61,32%	47,45%
2010	63,89%	48,48%	53,68%
2011	86,04%	40,46%	50,65%
2012	59,18%	27,91%	46,67%
2013	82,86%	25,49%	42,56%
2014	62,64%	39,06%	42,98%
2015	86,94%	*	*

5.2 Παράρτημα Β – Κώδικας R

5.2.1 Κώδικας για έτος επιτυχίας

```
dedomena2=read.csv2("g:\\oliko_mam.csv",header=F)

dedo=data.frame(etos=dedomena2$V1,kodikos_period=dedomena2$V2,kodikos=dedomena2$V3,foit=dedomena2$V5,foitExaminoMath=dedomena2$V6,Mathima=dedomena2$V7,Perase=dedomena2$V8)

summary(dedo)

#####edo gia opoioidipote mathima allazo MONO ton kodiko 12101 ston antistoixo#####

##### Statistiki I as poume einai to 12541, Grammiki I 12121#####

thesi2=0

d2=0

for(l in 1:length(dedo$Mathima))

{ if((dedo$Mathima[l]=='12101')&(dedo$Perase[l]>4))

d2=c(d2,l)

}

thesi2=d2

pin=0

length(thesi2)

for(j in 2:length(thesi2))

{pin[j]=dedo$etos[(dedo$foit==dedo$foit[thesi2[j]])&(dedo$Mathima=='12101')&(dedo$Perase>4)]-min(dedo$etos[dedo$foit==dedo$foit[thesi2[j]]],na.rm =T)

}

###o pinakas me to etos pou perase to mathima -1###

pin

summary(pin)

###prostheto +1 gia na exo to akrives etos##
```

5.2.2 Προσπάθειες μέχρι την πρώτη επιτυχία και Προσπάθειες μέχρι την πρώτη επιτυχία αλλά εξαιρώ τις προσπάθειες που πήρανε 0 ή 1 ή 2

```
dedomena2=read.csv2("g:\\oliko_mam.csv",header=F)

dedo=data.frame(etos=dedomena2$V1,kodikos_period=dedomena2$V2,kodikos=dedomena2$V3,foit=dedomena2$V5,foitExaminoMath=dedomena2$V6,Mathima=dedomena2$V7,Perase=dedomena2$V8)

summary(dedo)

d=duplicated(dedo$foit)

length(d)

thesi=0

for(l in 1:858681){ if (d[l]=='FALSE') thesi=c(thesi,l)}

##6016=length(thesi)##

synolo=matrix(1000,6016,3)

#####edo gia opoioidipote mathima allazo MONO ton kodiko 12101 ston antistoixo#####

##### Statistiki I as poume einai to 12541, Grammiki I 12121#####

for(j in 2:6016){

synolo[j,1]=length(dedo$Mathima[(dedo$foit==dedo$foit[thesi[j]])&(dedo$Mathima=='12101')&(dedo$Perase>=0)])

synolo[j,2]=length(dedo$Mathima[(dedo$foit==dedo$foit[thesi[j]])&(dedo$Mathima=='12101')&(dedo$Perase>2)])

synolo[j,3]=length(dedo$Mathima[(dedo$foit==dedo$foit[thesi[j]])&(dedo$Mathima=='12101')&(dedo$Perase>=5)])

}

library(xlsx)

synolo2=synolo

###an thelo ta stoixeia se excel

write.xlsx(synolo2, "g:/apeiro22.xlsx")

synolo2=synolo[-1,]

synolo2=synolo2[-length(synolo2[,1]),]
```

```
###prospatheies mexri tin proti epityxia
```

```
summary(synolo2[,1][synolo2[,3]==1])
```

```
###prospatheies mexri tin proti epityxia alla exairo otan pirane 0-1-2
```

```
summary(synolo2[,2][synolo2[,3]==1])
```

5.2.3. Πόσοι δήλωσαν-πέρασαν μάθημα ανά εξάμηνο και Σεπτέμβρη

```
dedomena2=read.csv2("G:\\oliko.csv",header=F)
```

```
dedomena2$V7[is.na(dedomena2$V7)==TRUE]=-2
```

```
dedo=data.frame(etos=dedomena2$V1,kodikos_period=dedomena2$V2,kodikos=dedomena2$V3,foiti=dedomena2$V4,ExaminoMath=dedomena2$V5,Mathima=dedomena2$V6,Perase=dedomena2$V7)
```

```
summary(dedo)
```

```
##example simplirosis
```

```
mathima=12101
```

```
year=2000
```

```
#an
```

```
func=function(year,mathima){
```

```
a=c(0,0,0,0,0,0,0)
```

```
a[1]=sum(dedo$kodikos[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_period==(year*1000+101))]==1)
```

```
a[2]=sum(dedo$Perase[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_period==year*1000+1021)]!=-2)
```

```
a[3]=sum(dedo$Perase[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_period==year*1000+1021)]>4)
```

```
a[4]=sum(dedo$kodikos[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_period==(year+1)*1000+31)]==1)
```

```
a[5]=sum(dedo$Perase[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_period==year*1000+1061)]>=0)
```

```
a[6]=sum(dedo$Perase[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_p  
eriod==(year*1000+1061)]>4)
```

```
#septemvris
```

```
a[7]=sum(dedo$Perase[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_p  
eriod==(year+1)*1000+91)]>=5)
```

```
a
```

```
}
```

```
#dilose xeim/edose/perase - ear/edose/perase/ septemvris
```

```
#to 2013-2014 ti kanan sto 12101(apeirostikos)
```

```
func(2013,12101)
```

```
#to 2000 ti kanan sto 12151(syndyastyki)
```

```
func(2000,12151)
```

5.2.4 Όλοσκόδικας

```
dedomena=read.csv2('e:/foititesFina_l.csv',header=T)
```

```
#gia na exo ptyxio vathmos_ptixiou oxi -1000
```

```
#sinolo ptixiouxon
```

```
#arithmos ptyxiouxon
```

```
length(dedomena$vathmos_ptixiou[dedomena$vathmos_ptixiou!=-1000])
```

```
#arthmos foithvtn poy grafthkan esto kai mia fora
```

```
length(dedomena$vathmos_ptixiou)
```

```
#agoria ptyxiouxoi
```

```
agoria=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$gender==1)]
```

```
length(agoria)
```

```
#koritsia ptyxiouxoi
```



```
koritsia=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=-1000)&(dedomena$gender==2)]
```

```
length(koritsia)
```

```
#pe;ilipsi gia ta examhna ana filo kai genika
```

```
summary(dedomena$examino_foit[dedomena$vathmos_ptixiou!=-1000])
```

```
summary(agoria)
```

```
summary(koritsia)
```

```
#gia pararthma ergsias
```

```
#gia kataskeyh barplot
```

```
pososta=matrix(0,max(max(agoria),max(koritsia)),1)
```

```
pososta2=matrix(0,max(max(agoria),max(koritsia)),1)
```

```
for(j in 1:max(max(agoria),max(koritsia)))
```

```
{pososta[j]=length(agoria[agoria==j]) }
```

```
for(j in 1:max(max(agoria),max(koritsia)))
```

```
{pososta2[j]=length(koritsia[koritsia==j]) }
```

```
pososta=pososta/length(agoria)
```

```
pososta2=pososta2/length(koritsia)
```

```
pinakas=cbind(pososta*100,100*pososta2)
```

```
#barplot examhna apokthshs ptyxiou sigkrisi agoria-koritsia gia pararthma
```

```
barplot(as.matrix(t(pinakas)),beside=T,col=c("darkblue","red"),names.arg=c(seq(1,32)))
```

```
text(80,10,c("agoria"),col=1)
```

```
text(80,12,c("koritsia"),col=2)
```

```
#plot examhna apokthshs ptyxiou sigkrisi agoria-koritsia gia pararthma
```

```
plot(pososta,xlim=c(7,35),ylim=c(0,0.3),type='l',xlab="examino ptixiou",main="Fillo vs  
examino ptixiou")
```

```
lines(pososta2,type='l',col=2)
```

```
text(30,0.20,c("agoria"),col=1)
```

```
text(30,0.22,c("koritsia"),col=2)
```

```
#gia kataskeyh boxplot
```

```
boxplot(agoria,koritsia, names=c("agoria","koritsia"),col=c(2,3),ylim=c(1,30),main="examhno  
apofitishs vs sex")
```

```
#####
```

```
#gia bathmo ptyxiou
```

```
summary(dedomena$vathmos_ptixiou[dedomena$vathmos_ptixiou!=-1000])
```

```
#posoi foithtes exoun parei 5 eos 6.5 , 6.5 me 7.5 , 7.5 me 8.5 .. 8.5 me 10 ktl
```

```
pente=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$vathmos_ptixiou<=6.5)]
```

```
eximisi=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$vathmos_ptixiou>6.5)&(dedomena$vathmos_ptixiou<=7.5)]
```

```
eftams=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$vathmos_ptixiou>7.5)&(dedomena$vathmos_ptixiou<=8.5)]
```

```
oktomisi=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$vathmos_ptixiou>8.5)]
```

```
temp=c(length(pente),length(eximisi),length(eftams),length(dedomena$vathmos_ptixiou[de  
domena$vathmos_ptixiou!=-1000])-length(pente)-length(eximisi)-  
length(eftams))/length(dedomena$vathmos_ptixiou[dedomena$vathmos_ptixiou!=-1000])
```

```
barplot(as.matrix(t(temp)),col="3",names.arg=c("5-6.5","6.5-7.5","7.5-8.5","8.5-10"),ylim=c(0,1),main="Pososta Ptixiouxon ana vathmologia")
```

```
#istogramma
```

```
hist(dedomena$vathmos_ptixiou[dedomena$vathmos_ptixiou!=-1000],breaks=10,main="vathmoi",xlab="")
```

```
####Sigkrisi examhno apogoitisis me bathmos
```

```
pen1=matrix(0,max(dedomena$examinofit[dedomena$vathmos_ptixiou!=-1000]),1)
```

```
exi1=matrix(0,max(dedomena$examinofit[dedomena$vathmos_ptixiou!=-1000]),1)
```

```
xoro=matrix(0,max(dedomena$examinofit[dedomena$vathmos_ptixiou!=-1000]),1)
```

```
eftam1=matrix(0,max(dedomena$examinofit[dedomena$vathmos_ptixiou!=-1000]),1)
```

```
oktomisi1=matrix(0,max(dedomena$examinofit[dedomena$vathmos_ptixiou!=-1000]),1)
```

```
#apo edo kai kato einai gia barplot-gia pararthma
```

```
for(j in 1:32)
```

```
{pen1[j]=length(pen1[pen1==j]) }
```

```
for(j in 1:32)
```

```
{exi1[j]=length(eximisi[eximisi==j]) }
```

```
for(j in 1:32)
```

```
{xoro[j]=length(eftams[eftams==j]) }
```

```
for(j in 1:32)
```

```
{oktomisi1[j]=length(oktomisi[oktomisi==j]) }
```

```
par(mfrow=c(2,2))
```

```
barplot(as.matrix(t(pente1/sum(pente1))),ylim=c(0,0.5),main="5-  
6.5",names.arg=c(seq(1,32)))
```

```
barplot(as.matrix(t(exi1/sum(exi1))),ylim=c(0,0.5),main="6.5",names.arg=c(seq(1,32)))
```

```
barplot(as.matrix(t(xoro/sum(xoro))),ylim=c(0,0.5),main="7.5-8.5",names.arg=c(seq(1,32)))
```

```
barplot(as.matrix(t(oktomisi1/sum(oktomisi1))),ylim=c(0,0.5),main="8.5-  
10",names.arg=c(seq(1,32)))
```

```
title(sub="Vathmos vs Examino Apofitisis",col="5")
```

```
#8 gia pararthma plot
```

```
plot(pente1/sum(pente1),xlim=c(7,35),ylim=c(0,0.6),type='l',xlab="examino  
ptixiou",main="Vatmos vs examino ptixiou",ylab="Pososto")
```

```
lines(exi1/sum(exi1),type='l',col=2)
```

```
lines(xoro/sum(xoro),type='l',col=3)
```

```
lines(oktomisi1/sum(oktomisi1),type='l',col=4)
```

```
text(30,0.50,c("5.0-6.5"),col=1)
```

```
text(30,0.47,c("6.5-7.5"),col=2)
```

```
text(30,0.44,c("7.5-8.5"),col=3)
```

```
text(30,0.41,c("8.5>"),col=4)
```

```
#boxplot vathmos vs examhno ptyxiou
```

```
boxplot(pente,eximisi,eftams,oktomisi,names=c("5.0-6.5","6.5-7.5","7.5-8.5","8.5-  
10"),col=c(2,3,4,6),main=" vathmos vs examhno ptyxiou")
```

```
#####
```

```
#epomenh grammh oxi xrhsimh
```

```
summary(dedomena$etos_eggrafis)
```

#10) Krish apo 2008 kai epeita, nomos gia diagrafh foithton, eclass kalh ..shmeioseis foitoitvn

#Prin thn krish apo to 2005 eos 2007

#to poly prin thn krish apo 2000 eos 2004 ME PASOK

```
krisi=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=  
1000)&(dedomena$etos_eggrafis>2007)]
```

```
prin_krisi=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=  
1000)&(dedomena$etos_eggrafis<2007)&(dedomena$etos_eggrafis>2004)]
```

```
poli_prin=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=  
1000)&(dedomena$etos_eggrafis<2005)]
```

```
boxplot(krisi,prin_krisi,poli_prin,col=c(2,4,3),names=c("Κρίση","Πριν την Κρίση", "ΜΕ  
ΠΑΣΟΚ"))
```

#11)gia pararthma bathmoi me krish kai prin thn krish kai me pasok

```
pasok=matrix(0,32,1)
```

```
for(j in 1:32)
```

```
{pasok[j]=length(poli_prin[poli_prin==j]) }
```

```
nd=matrix(0,32,1)
```

```
for(j in 1:32)
```

```
{nd[j]=length(prin_krisi[prin_krisi==j]) }
```

```
syriza=matrix(0,32,1)
```

```
for(j in 1:32)
```

```
{syriza[j]=length(krisi[krisi==j]) }
```

```
par(mfrow=c(3,1))
```

```
barplot(as.matrix(t(syriza/sum(syriza))),ylim=c(0,0.5),main="Meta tin  
Krisi",names.arg=c(seq(1,32)))
```

```
barplot(as.matrix(t(pasok/sum(pasok))),ylim=c(0,0.5),main="Poli prin tin  
Krisi",names.arg=c(seq(1,32)))
```

```
barplot(as.matrix(t(nd/sum(nd))),ylim=c(0,0.5),main="Prin tin Krisi",names.arg=c(seq(1,32)))  
title(sub="Krisi vs Examino Apofitisis",col="5")
```

```
###einai kapos meroliptiko giati oi kaloi mathites telionoun kai pio grigora, ara den ehoun  
prolavei na teliosoun oi kakoi
```

```
###13
```

```
summary(dedomena$katethinsi[(dedomena$vathmos_ptixiou!=-1000)])
```

```
efta=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$katethinsi==557)]
```

```
exi=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$katethinsi==556)]
```

```
#boxplot
```

```
boxplot(efta,exi,names=c("Efarmosme","Pure"),col=c(16,12),main="examhno apofithshs vs  
kateuthnisi")
```

```
###14 gia parartima
```

```
theos=efas=matrix(0,32,1)
```

```
for(j in 1:32)
```

```
{theos[j]=length(exi[exi==j]) }
```

```
for(j in 1:32)
```

```
{efas[j]=length(efta[efta==j]) }
```

```
par(mfrow=c(2,1))
```

```
barplot(as.matrix(t(theos/sum(theos))),ylim=c(0,0.5),main="Theor",names.arg=c(seq(1,32)))
```

```
barplot(as.matrix(t(efas/sum(efas))),ylim=c(0,0.5),main="Efar",names.arg=c(seq(1,32)))
```

```
title(sub="Efarmosena-Theor vs Examino Apofitisis",col="5")
```

```
plot(theos/sum(theos),xlim=c(7,35),ylim=c(0,0.3),type='l',xlab="examino  
ptixiou",main="Vatmos vs examino ptixiou",ylab="Pososto")
```

```
lines(efas/sum(efas),type='l',col=2)
```

```
text(30,0.24,c("To allo"),col=1)
```

```
text(30,0.26,c("Efar"),col=2)
```

```
##15 tha antikatastathei
```

```
ellinas=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$ypikootita=='ΕΛΛ')]
```

```
kp=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$ypikootita=='ΚΥΠ')]
```

```
allo=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$ypikootita!='ΚΥΠ')&(dedomena$ypikootita!='ΕΛΛ')&(dedomena$ypikoot  
ita=='')]
```

```
boxplot(ellinas,kp,names=c("Ellinas","Kiprios"))
```

```
greek=cyp=allos=matrix(0,32,1)
```

```

for(j in 1:32)
{cyp[j]=length(kp[kp==j]) }

for(j in 1:32)
{allos[j]=length(allo[allo==j]) }

for(j in 1:32)
{greek[j]=length(ellinas[ellinas==j]) }


par(mfrow=c(3,1))

barplot(as.matrix(t(greek/sum(greek))),ylim=c(0,0.4),main="greek",names.arg=c(seq(1,32)))
barplot(as.matrix(t(allos/sum(allos))),ylim=c(0,0.4),main="allos",names.arg=c(seq(1,32)))
barplot(as.matrix(t(cyp/sum(cyp))),ylim=c(0,0.4),main="cyp",names.arg=c(seq(1,32)))

title(sub="Katagogi ton eidon vs Examino Apofitisis",col="5")

####

```



```
##pame pio kato itane test.dokimh gia ta epomena.den xreiazetai sthn ergasia
```

```
library(lme4)
```

```
Y=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=-1000)]-7
```

```
X=dedomena$gender[(dedomena$vathmos_ptixiou!=-1000)]
```

```
X2=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-1000)]
```

```
id=dedomena$ arithmos_mhtroou[(dedomena$vathmos_ptixiou!=-1000)]
```

```
summary(glmer(formula = Y~ factor(X)+as.numeric(X2)+(1|id ), family = poisson()))
```

```
####
```

```
for(j in 1:max(max(agoria),max(koritsia)))
```

```
{pososta[j]=length(agoria[agoria==j]) }
```

```
for(j in 1:max(max(agoria),max(koritsia)))
```

```
{pososta2[j]=length(koritsia[koritsia==j]) }
```

```
temp=c(sum(pososta[1:12]),sum(pososta[13:17]),sum(pososta[18:22]),sum(pososta[23:27]),  
sum(pososta[28:32]))
```

```
temp2=c(sum(pososta2[1:12]),sum(pososta2[13:17]),sum(pososta2[18:22]),sum(pososta2[2  
3:27]),sum(pososta2[28:32]))
```

```
temp3=temp+temp2
```

```

expected=temp3*sum(temp)/sum(temp3)

expected=c(expected,temp3*sum(temp2)/sum(temp3))

sum((c(temp,temp2)-expected)^2/expected)

qchisq(0.95, (5-1)*(2-1), ncp = 0, lower.tail = TRUE, log.p = FALSE)

#If the test statistic is improbably large according to that chi-squared distribution, then one
rejects the null hypothesis of independence.#

#####

#####

###16

library(lme4)

#metatroph dedomenon

Y=dedomena$examino_foit[(dedomena$vathmos_ptixiou!=-1000)]-7

fyllo=dedomena$gender[(dedomena$vathmos_ptixiou!=-1000)]

vathmos=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-1000)]

#id=arithmos mhtroou. Edo mpainei to random effect.kathe atomo exei to diko tou random
effect

id=dedomena$ arithmos_mhtroou[(dedomena$vathmos_ptixiou!=-1000)]

###17 krisi###

#metatroph dedomenon. Epoxi-krish-pro krish

Epoxi=matrix(0,length(dedomena[,1]),1)

for(i in 1:length(dedomena[,1])){

```

```

if((dedomena$vathmos_ptixiou[i]!=-1000)&(dedomena$etos_eggrafis[i]>2007)) {Epoxi[i]=3}

else if ((dedomena$vathmos_ptixiou[i]!=-
1000)&(dedomena$etos_eggrafis[i]<=2007)&(dedomena$etos_eggrafis[i]>2004)){Epoxi[i]=2
}

else if ((dedomena$vathmos_ptixiou[i]!=-
1000)&(dedomena$etos_eggrafis[i]<2005)){Epoxi[i]=1}

}

```

```

Epoxi=Epoxi[Epoxi!=0]

```

```

#poupame=kateuthnisi

```

```

poupame=matrix(0,length(dedomena[,1]),1)

```

```

for(i in 1:length(dedomena[,1])){

```

```

if((dedomena$katethinsi[i]==557)&(dedomena$vathmos_ptixiou[i]!=-1000)){poupame[i]=1}

```

```

else if((dedomena$katethinsi[i]==556)&(dedomena$vathmos_ptixiou[i]!=-1000)){
poupame[i]=2}

```

```

else if((dedomena$katethinsi[i]==0)&(dedomena$vathmos_ptixiou[i]!=-1000)){
poupame[i]=3}

```

```

}

```

```

poupame=poupame[poupame!=0]

```

```

pureornot=poupame

```

```

p_exi=1-length(pureornot[pureornot==1])/length(pureornot[pureornot!=3])

```

```

pureornot

```

```

for(i in 1:2312){

```

```

if(pureornot[i]==3){

```

```

u=runif(1,0,1)

```

```

if(u>p_exi)

```

```
pureornot[i]=1  
else  
pureornot[i]=2  
}}
```

```
poli=read.table("C:\\Users\\apostolas\\Documents\\POLI.txt",header=F)  
poli=as.matrix(poli)  
length(poli[poli==0])  
length(poli[poli==1])  
length(poli[poli==2])
```

```
pith_ath=length(poli[poli==0])/1938  
length(poli)
```

```
for(i in 1:2312){  
  if(poli[i]==4){  
    u=runif(1,0,1)  
    if(u>pith_ath)  
      poli[i]=1  
    else  
      poli[i]=0  
  }  
}
```

```
#18 elegxos ;oti exoun idio megethow
```

```
length(Epoxi)
```

```
length(Y)
```

```
length(pureornot)
```

```
#19 Montela GLMM me random effect to 1 id
```

```
#exartimenh y vs bathmos
```

```
summary(glmer(formula = Y~ as.numeric(vathmos)+(1|id ), family = poisson()))
```

```
#exartimenh y vs fyllo
```

```
summary(glmer(formula = Y~ factor(fyllo)+(1|id ), family = poisson()))
```

```
#exartimenh y vs epoxi
```

```
summary(glmer(formula = Y~ +factor(Epoxi)+(1|id ), family = poisson()))
```

```
#exartimenh y vs kateuthnisi
```

```
summary(glmer(formula = Y~ factor(pureornot)+(1|id ), family = poisson()))
```

```
summary(glmer(formula = Y~ factor(poli)+(1|id ), family = poisson()))
```

```
#exarth me ola. Full Montelo.Edw meinane ola logo tou p-value-statistikhs simantikohtas
```

```
summary(glmer(formula = Y~  
factor(fyllo)+as.numeric(vathmos)+factor(Epoxi)+factor(pureornot)+factor(poli)+(1|id ),  
family = poisson()))
```

```
####
```

```
####
```

```
####
```

```
####
```

```
ag=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=  
1000)&(dedomena$gender==1)]
```

```

kor=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$gender==2)]

summary(ag)

summary(kor)

agoria3=round(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$gender==1)],0)

koritsia3=round(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$gender==2)],0)

boxplot(kor,ag,names=c("ginaikes","agori"),col=c(11,13))

pososta3=matrix(0,max(max(agoria3),max(koritsia3)),1)
pososta23=matrix(0,max(max(agoria3),max(koritsia3)),1)
for(j in 1:10)
{pososta3[j]=length(agoria3[agoria3==j]) }
for(j in 1:10)
{pososta23[j]=length(koritsia3[koritsia3==j]) }
pososta3=pososta3/length(agoria3)
pososta23=pososta23/length(koritsia3)

par(mfrow=c(2,1))

barplot(as.matrix(t(pososta3)),ylim=c(0,0.6),main="Agoria",names.arg=c(seq(1,10)))
barplot(as.matrix(t(pososta23)),ylim=c(0,0.6),main="Koukles",names.arg=c(seq(1,10)))

title(sub="Vathmos vs Fylo",col="5")

```

```

plot(pososta3,xlim=c(5,10),ylim=c(0,1),type='l',xlab="vathmos",main="Fillo vs examino
ptixiou")

lines(pososta23,type='l',col=2)

text(9,0.8,c("agoria"),col=1)

text(9,0.92,c("koritsia"),col=2)

#####

summary(dedomena$etos_eggrafis)

kr=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis>2007)]

pr_b=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis<2007)&(dedomena$etos_eggrafis>2004)]

p_prin=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis<2005)]

boxplot(pr_b,p_prin,krisi_b,names=c("2005-07","2000-04","2008+"),col=c(2,3,4))

krisi_b=round(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis>2007)],0)

prin_krisi_b=round(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis<2007)&(dedomena$etos_eggrafis>2004)],0)

poli_prin_b=round(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis<2005)],0)

pasokb=ndb=syrizab=matrix(0,10,1)

for(j in 1:10)

{pasokb[j]=length(poli_prin_b[poli_prin_b==j]) }

for(j in 1:10)

{ndb[j]=length(prin_krisi_b[prin_krisi_b==j]) }

syrizab=matrix(0,10,1)

for(j in 1:10)

{syrizab[j]=length(krisi_b[krisi_b==j]) }

```

```

par(mfrow=c(3,1))

barplot(as.matrix(t(pasokb)/sum(pasokb)),ylim=c(0,1),main="Poli Prin tin
Krisi",names.arg=c(seq(1,10)))

barplot(as.matrix(t(ndb)/sum(ndb)),ylim=c(0,1),main="Prin tin Krisi",names.arg=c(seq(1,10)))

barplot(as.matrix(t(syrizab)/sum(syrizab)),ylim=c(0,1),main="Me
Syriza",names.arg=c(seq(1,10)))

title(sub="Vathmos vs Krisi",col="5")

plot(pasokb/sum(pasokb),xlim=c(5,10),ylim=c(0,1),type='l',xlab="vathmos",main="Fillo vs
examino ptixiou",ylab="pososta")

lines(ndb/sum(ndb),type='l',col=3)

lines(syrizab/sum(syrizab),type='l',col=2)

text(9,0.8,c("Poli prin"),col=1)

text(9,0.92,c("Krisi"),col=2)

text(9,0.85,c("Prin tin Krisi"),col=3)

#####

summary(dedomena$skatefthinsi[(dedomena$vathmos_ptixiou!=-1000)])

eftac=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-
1000)&(dedomena$skatefthinsi==557)]

exic=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-
1000)&(dedomena$skatefthinsi==556)]

boxplot(exic,eftac,names=c("Pure","appli"),col=c(2,3))

```



```
eftab=round(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-
1000)&(dedomena$skatefthinsi==557)],0)
```

```
exib=round(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-
1000)&(dedomena$skatefthinsi==556)],0)
```

```
theoritiki=statistiki=matrix(0,10,1)
```

```
for(j in 1:10)
```

```
{theoritiki[j]=length(exib[exib==j]) }
```

```
for(j in 1:10)
```

```
{statistiki[j]=length(eftab[eftab==j]) }
```

```
par(mfrow=c(2,1))
```

```
barplot(as.matrix(t(theoritiki)/sum(theoritiki)),ylim=c(0,0.6),main="Theoria",names.arg=c(se
q(1,10)))
```

```
barplot(as.matrix(t(statistiki)/sum(statistiki)),ylim=c(0,0.6),main="Vlakes",names.arg=c(seq(1
,10)))
```

```
title(sub="Θεωρητική και άλλο",col="5")
```

```
###
```

```
Y2=as.numeric(vathmos)
```

```
hist(Y2)
```

```
examino=Y
```

```
length(Epoxi)
```

```
length(Y)
```

```
length(pureornot)
```

```
summary(lmer(formula = Y2~ factor(Epoxi)+(1|id),control=lmerControl(check.nobs.vs.nlev =
"ignore", check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore")))
```

```
summary(lmer(formula = Y2~
as.numeric(examino)+(1|id),control=lmerControl(check.nobs.vs.nlev = "ignore",
check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore")))
```

```
summary(lmer(formula = Y2~
factor(pureornot)+(1|id),control=lmerControl(check.nobs.vs.nlev = "ignore",
check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore")))
```

```
summary(lmer(formula = Y2~ factor(fyllo)+(1|id),control=lmerControl(check.nobs.vs.nlev =
"ignore", check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore")))
```

```
summary(lmer(formula = Y2~ factor(poli)+(1|id),control=lmerControl(check.nobs.vs.nlev =
"ignore", check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore")))
```

```
temp=lmer(formula = Y2~ factor(poli)+(1|id),control=lmerControl(check.nobs.vs.nlev =
"ignore", check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore"))
```

```
temp2=lmer(formula = Y2~ (1|id),control=lmerControl(check.nobs.vs.nlev = "ignore",
check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore"))
```

```
anova(temp,temp2)
```

```
m1=lmer(formula = Y2~ factor(fyllo)+factor(Epoxi)+as.numeric(examino)+
factor(pureornot)+(1|id),control=lmerControl(check.nobs.vs.nlev = "ignore",
check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore"))
```

```
m2=lmer(formula = Y2~ factor(fyllo)+factor(Epoxi)+as.numeric(examino)+factor(poli)+
factor(pureornot)+(1|id),control=lmerControl(check.nobs.vs.nlev = "ignore",
check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore"))
```

```
anova(m1,m2)
```

```
summary(m2)
```

```
####
```

```
summary(dedomena$vathmos[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$examino_foit<12)])
```

```

summary(dedomena$vathmos[(dedomena$vathmos_ptixiou!=-
1000)&(dedomena$examinio_foit>11)&(dedomena$examinio_foit<15)])

summary(dedomena$vathmos[(dedomena$vathmos_ptixiou!=-
1000)&(dedomena$examinio_foit>14)&(dedomena$examinio_foit<18)])

summary(dedomena$vathmos[(dedomena$vathmos_ptixiou!=-
1000)&(dedomena$examinio_foit>18)])

w=dedomena$vathmos[(dedomena$vathmos_ptixiou!=-
1000)&(dedomena$examinio_foit<12)]

ww=dedomena$vathmos[(dedomena$vathmos_ptixiou!=-
1000)&(dedomena$examinio_foit>11)&(dedomena$examinio_foit<15)]

wwe=dedomena$vathmos[(dedomena$vathmos_ptixiou!=-
1000)&(dedomena$examinio_foit>14)&(dedomena$examinio_foit<18)]

qq=dedomena$vathmos[(dedomena$vathmos_ptixiou!=-
1000)&(dedomena$examinio_foit>18)]

boxplot(w,ww,wwe,qq,names=c("7-11","12-14","15-
17","18+"),col=c(2,3,4,66),main="Semester",ylab='Grade')

#####

#####

dedomena=read.csv2('c:/foititesFina_l2.csv',header=T)

#gia na exo ptyxio vathmos_ptixiou oxi -1000

#sinolo ptixiouxon

#arithmos ptyxiouxon

length(dedomena$vathmos_ptixiou[dedomena$vathmos_ptixiou!=-1000])

dedomena[1,]

```

```

mo=matrix(0,12,1)

mo[1]=mean(dedomena$svathmos_ptixiou[(dedomena$svathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2000)])

mo[2]=mean(dedomena$svathmos_ptixiou[(dedomena$svathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2001)])

mo[3]=mean(dedomena$svathmos_ptixiou[(dedomena$svathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2002)])

mo[4]=mean(dedomena$svathmos_ptixiou[(dedomena$svathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2003)])

mo[5]=mean(dedomena$svathmos_ptixiou[(dedomena$svathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2004)])

mo[6]=mean(dedomena$svathmos_ptixiou[(dedomena$svathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2005)])

mo[7]=mean(dedomena$svathmos_ptixiou[(dedomena$svathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2006)])

mo[8]=mean(dedomena$svathmos_ptixiou[(dedomena$svathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2007)])

mo[9]=mean(dedomena$svathmos_ptixiou[(dedomena$svathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2008)])

mo[10]=mean(dedomena$svathmos_ptixiou[(dedomena$svathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2009)])

mo[11]=mean(dedomena$svathmos_ptixiou[(dedomena$svathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2010)])

mo[12]=mean(dedomena$svathmos_ptixiou[(dedomena$svathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2011)])

```

```

mo1=matrix(0,12,1)

mo1[1]=length(dedomena$svathmos_ptixiou[(dedomena$svathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2000)])

mo1[2]=length(dedomena$svathmos_ptixiou[(dedomena$svathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2001)])

```

```

mo1[3]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2002)])

mo1[4]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2003)])

mo1[5]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2004)])

mo1[6]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2005)])

mo1[7]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2006)])

mo1[8]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2007)])

mo1[9]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2008)])

mo1[10]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2009)])

mo1[11]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2010)])

mo1[12]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=
1000)&(dedomena$etos_eggrafis==2011)])

```

mo1

Spoud_katastasi

```
lele=matrix(0,12,1)
```

```

lele[1]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou==
1000)&(dedomena$Spoud_katastasi==3)&(dedomena$etos_eggrafis==2000)])

lele[3]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou==
1000)&(dedomena$Spoud_katastasi==3)&(dedomena$etos_eggrafis==2002)])

lele[2]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou==
1000)&(dedomena$Spoud_katastasi==3)&(dedomena$etos_eggrafis==2001)])

lele[4]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou==
1000)&(dedomena$Spoud_katastasi==3)&(dedomena$etos_eggrafis==2003)])

lele[5]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou==
1000)&(dedomena$Spoud_katastasi==3)&(dedomena$etos_eggrafis==2004)])

```

```

lele[6]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou==1000)&(dedomena$Spoud_katastasi==3)&(dedomena$etos_eggrafis==2005)])

lele[7]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou==1000)&(dedomena$Spoud_katastasi==3)&(dedomena$etos_eggrafis==2006)])

lele[8]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou==1000)&(dedomena$Spoud_katastasi==3)&(dedomena$etos_eggrafis==2007)])

lele[7+1+1]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou==1000)&(dedomena$Spoud_katastasi==3)&(dedomena$etos_eggrafis==2008)])

lele[10]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou==1000)&(dedomena$Spoud_katastasi==3)&(dedomena$etos_eggrafis==2009)])

lele[12]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou==1000)&(dedomena$Spoud_katastasi==3)&(dedomena$etos_eggrafis==2011)])

lele[11]=length(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou==1000)&(dedomena$Spoud_katastasi==3)&(dedomena$etos_eggrafis==2010)])

lele

o=mo1/(mo1+lele)*100

x=c(2000:2011)

plot(x,mo,ylim=c(0,10),type='b',main="Mean grade",ylab="")

plot(x,o,ylim=c(0,100),type='b',main="Degree percentage",ylab="",xlab='year')

barplot(as.matrix(t(mo)),beside=T,ylim=c(0,7),col=c("darkblue"),names.arg=c(seq(2000,2011)),space=c(0,0.5))

####

###

dedomena2=read.csv2("G:\\oliko.csv",header=F)

dedomena2$V7[is.na(dedomena2$V7)==TRUE]=-2

```

```
dedo=data.frame(etos=dedomena2$V1,kodikos_period=dedomena2$V2,kodikos=dedomena2$V3,foiti=dedomena2$V4,ExaminoMath=dedomena2$V5,Mathima=dedomena2$V6,Perase=dedomena2$V7)
```

```
summary(dedo)
```

```
#posoi dilosan to 12121 to 2000
```

```
sum(dedo$kodikos[(dedo$etos==2000)&(dedo$Mathima==12121)]==1)
```

```
#to xeimerino 2001021 posoi to edosan
```

```
sum(dedo$Perase[(dedo$etos==2000)&(dedo$Mathima==12121)&(dedo$kodikos_period==2001021)]>=0)
```

```
#to xeimerino 2001091 posoi to edosan septem,vris
```

```
sum(dedo$Perase[(dedo$etos==2000)&(dedo$Mathima==12121)&(dedo$kodikos_period==2001091)]>=0)
```

```
#posoi to perasan to 12121 to 2000
```

```
sum(dedo$Perase[(dedo$etos==2000)&(dedo$Mathima==12121)]>=5)
```

```
#posoi dilosan to 12121 to 2000
```

```
sum(dedo$kodikos[(dedo$etos==2001)&(dedo$Mathima==12121)]==1)
```

```
#to xeimerino 2001021 posoi to edosan
```

```
sum(dedo$Perase[(dedo$etos==2001)&(dedo$Mathima==12121)&(dedo$kodikos_period==
2002021)]>=0)
```

```
#to xeimerino 2001091 posoi to edosan septem,vris
```

```
sum(dedo$Perase[(dedo$etos==2001)&(dedo$Mathima==12121)&(dedo$kodikos_period==
2002091)]>=0)
```

```
#posoi to perasan to 12121 to 2001
```

```
sum(dedo$Perase[(dedo$etos==2001)&(dedo$Mathima==12121)]>=5)
```

```
mathima=12121
```

```
year=2000
```

```
#posoi dilosan to mathima to year to xeimerino kai posoi to perasan
```

```
sum(dedo$kodikos[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_perio
d==(year)*1000+101)]==1)
```

```
sum(dedo$Perase[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_period
==year*1000+1021)]>4)
```

```
#posoi dilosan to 12121 to 2000 to allo
```

```
sum(dedo$kodikos[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_perio
d==(year+1)*1000+31)]==1)
```

```
sum(dedo$Perase[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_period
==year*1000+1061)]>4)
```

```
#septemvris
```

```
sum(dedo$Perase[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_period
==(year+1)*1000+91)]>=5)
```

```
mathima=12101
```

```
year=2000
```



```

#posoi dilosan to mathima to year to xeimerino kai posoi to perasan

func=function(year,mathima){

a=c(0,0,0,0,0,0,0)

a[1]=sum(dedo$kodikos[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_p
eriod==(year*1000+101))]==1)

a[2]=sum(dedo$Perase[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_p
eriod==year*1000+1021)]!=-2)

a[3]=sum(dedo$Perase[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_p
eriod==year*1000+1021)]>4)


#posoi dilosan to 12121 to 2000 to allo

a[4]=sum(dedo$kodikos[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_p
eriod==(year+1)*1000+31)]==1)

a[5]=sum(dedo$Perase[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_p
eriod==year*1000+1061)]>=0)

a[6]=sum(dedo$Perase[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_p
eriod==year*1000+1061)]>4)


#septemvris

a[7]=sum(dedo$Perase[(dedo$etos==year)&(dedo$Mathima==mathima)&(dedo$kodikos_p
eriod==(year+1)*1000+91)]>=5)

a

}

#dilose xeim/edose/perase - ear/edose/perase/ septemvris

func(2013,12101)

####

ag=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=1000)&(dedomena$gender==1)]

```

```
kor=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=  
1000)&(dedomena$gender==2)]
```

```
summary(ag)
```

```
summary(kor)
```

```
agoria3=round(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=  
1000)&(dedomena$gender==1)],0)
```

```
koritsia3=round(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=  
1000)&(dedomena$gender==2)],0)
```

```
boxplot(kor,ag,names=c("ginaikes","agori"),col=c(11,13))
```

```
pososta3=matrix(0,max(max(agoria3),max(koritsia3)),1)
```

```
pososta23=matrix(0,max(max(agoria3),max(koritsia3)),1)
```

```
for(j in 1:10)
```

```
{pososta3[j]=length(agoria3[agoria3==j]) }
```

```
for(j in 1:10)
```

```
{pososta23[j]=length(koritsia3[koritsia3==j]) }
```

```
pososta3=pososta3/length(agoria3)
```

```
pososta23=pososta23/length(koritsia3)
```

```
par(mfrow=c(2,1))
```

```
barplot(as.matrix(t(pososta3)),ylim=c(0,0.6),main="Agoria",names.arg=c(seq(1,10)))
```

```
barplot(as.matrix(t(pososta23)),ylim=c(0,0.6),main="Koukles",names.arg=c(seq(1,10)))
```

```
title(sub="Vathmos vs Fylo",col="5")
```

```
plot(pososta3,xlim=c(5,10),ylim=c(0,1),type='l',xlab="vathmos",main="Fillo vs examino  
ptixiou")
```

```
lines(pososta23,type='l',col=2)
```

```
text(9,0.8,c("agoria"),col=1)
```

```
text(9,0.92,c("koritsia"),col=2)
```

```
#####
```

```
summary(dedomena$etos_eggrafis)
```

```
kr=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$etos_eggrafis>2007)]
```

```
pr_b=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$etos_eggrafis<2007)&(dedomena$etos_eggrafis>2004)]
```

```
p_prin=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$etos_eggrafis<2005)]
```

```
boxplot(pr_b,p_prin,krisi_b, names=c("2005-07","2000-04","2008+"),col=c(2,3,4))
```

```
krisi_b=round(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$etos_eggrafis>2007)],0)
```

```
prin_krisi_b=round(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$etos_eggrafis<2007)&(dedomena$etos_eggrafis>2004)],0)
```

```
poli_prin_b=round(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-  
1000)&(dedomena$etos_eggrafis<2005)],0)
```

```
pasokb=ndb=syrizab=matrix(0,10,1)
```

```
for(j in 1:10)
```

```
{pasokb[j]=length(poli_prin_b[poli_prin_b==j]) }
```

```

for(j in 1:10)

{ndb[j]=length(prin_krisi_b[prin_krisi_b==j]) }


syrizab=matrix(0,10,1)

for(j in 1:10)

{syrizab[j]=length(krisi_b[krisi_b==j]) }


par(mfrow=c(3,1))


barplot(as.matrix(t(pasokb)/sum(pasokb)),ylim=c(0,1),main="Poli Prin tin
Krisi",names.arg=c(seq(1,10)))

barplot(as.matrix(t(ndb)/sum(ndb)),ylim=c(0,1),main="Prin tin Krisi",names.arg=c(seq(1,10)))

barplot(as.matrix(t(syrizab)/sum(syrizab)),ylim=c(0,1),main="Me
Syriza",names.arg=c(seq(1,10)))


title(sub="Vathmos vs Krisi",col="5")


plot(pasokb/sum(pasokb),xlim=c(5,10),ylim=c(0,1),type='l',xlab="vathmos",main="Fillo vs
examino ptixiou",ylab="pososta")


lines(ndb/sum(ndb),type='l',col=3)

lines(syrizab/sum(syrizab),type='l',col=2)

text(9,0.8,c("Poli prin"),col=1)

text(9,0.92,c("Krisi"),col=2)

text(9,0.85,c("Prin tin Krisi"),col=3)


#####

```

```

summary(dedomena$katethinsi[(dedomena$vathmos_ptixiou!=-1000)])

eftac=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-
1000)&(dedomena$katethinsi==557)]

exic=dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-
1000)&(dedomena$katethinsi==556)]

boxplot(exic,eftac,names=c("Pure","appli"),col=c(2,3))

eftab=round(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-
1000)&(dedomena$katethinsi==557)],0)

exib=round(dedomena$vathmos_ptixiou[(dedomena$vathmos_ptixiou!=-
1000)&(dedomena$katethinsi==556)],0)

theoritiki=statistiki=matrix(0,10,1)

for(j in 1:10)

{theoritiki[j]=length(exib[exib==j]) }

for(j in 1:10)

{statistiki[j]=length(eftab[eftab==j]) }

par(mfrow=c(2,1))

barplot(as.matrix(t(theoritiki)/sum(theoritiki)),ylim=c(0,0.6),main="Theoria",names.arg=c(se
q(1,10)))

barplot(as.matrix(t(statistiki)/sum(statistiki)),ylim=c(0,0.6),main="Vlakes",names.arg=c(seq(1
,10)))

title(sub="Θεωρητική και άλλο",col="5")

###

```

```

Y2=as.numeric(vathmos)

hist(Y2)

examino=Y

length(Epoxi)

length(Y)

length(pureornot)


summary(lmer(formula = Y2~ factor(Epoxi)+(1|id),control=lmerControl(check.nobs.vs.nlev =
"ignore", check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore")))

summary(lmer(formula = Y2~
as.numeric(examino)+(1|id),control=lmerControl(check.nobs.vs.nlev = "ignore",
check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore")))

summary(lmer(formula = Y2~
factor(pureornot)+(1|id),control=lmerControl(check.nobs.vs.nlev = "ignore",
check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore")))

summary(lmer(formula = Y2~ factor(fyllo)+(1|id),control=lmerControl(check.nobs.vs.nlev =
"ignore", check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore")))

summary(lmer(formula = Y2~ factor(poli)+(1|id),control=lmerControl(check.nobs.vs.nlev =
"ignore", check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore")))

temp=lmer(formula = Y2~ factor(poli)+(1|id),control=lmerControl(check.nobs.vs.nlev =
"ignore", check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore"))

temp2=lmer(formula = Y2~ (1|id),control=lmerControl(check.nobs.vs.nlev = "ignore",
check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore"))

anova(temp,temp2)

m1=lmer(formula = Y2~ factor(fyllo)+factor(Epoxi)+as.numeric(examino)+
factor(pureornot)+(1|id),control=lmerControl(check.nobs.vs.nlev = "ignore",
check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore"))

m2=lmer(formula = Y2~ factor(fyllo)+factor(Epoxi)+as.numeric(examino)+factor(poli)+
factor(pureornot)+(1|id),control=lmerControl(check.nobs.vs.nlev = "ignore",
check.nobs.vs.rankZ = "ignore", check.nobs.vs.nRE="ignore"))

anova(m1,m2)

summary(m2)

#####

```

```

poli=read.table("C:\\Users\\apostolas\\Documents\\POLI.txt",header=F)

poli=as.matrix(poli)

length(poli[poli==0])

length(poli[poli==1])

length(poli[poli==2])

pith_ath=length(poli[poli==0])/1938

length(poli)


for(i in 1:2312){

if(poli[i]==4){

u=runif(1,0,1)

if(u>pith_ath)

poli[i]=1

else

poli[i]=0

}}

summary(glmmer(formula = Y~ factor(poli)+(1|id ), family = poisson()))

summary(glmmer(formula = Y~
factor(fyllo)+as.numeric(vathmos)+factor(Epoxi)+factor(pureornot)+(1|id )+factor(poli),
family = poisson()))

#

dedomena2=read.csv2("g:\\oliko_mam.csv",header=F)


dedo=data.frame(etos=dedomena2$V1,kodikos_period=dedomena2$V2,kodikos=dedomena2$V3,foit=dedomena2$V5,foitExaminoMath=dedomena2$V6,Mathima=dedomena2$V7,Perase=dedomena2$V8)

summary(dedo)

```

```
#####edo gia opoioidipote mathima allazo MONO ton kodiko 12101 ston  
antistoixo#####
```

```
##### Statistiki I as poume einai to 12541, Grammiki I 12121#####
```

```
thesi2=0
```

```
d2=0
```

```
for(l in 1:length(dedo$Mathima))
```

```
{ if((dedo$Mathima[l]=='12101')&(dedo$Perase[l]>4))
```

```
d2=c(d2,l)
```

```
}
```

```
thesi2=d2
```

```
pin=0
```

```
length(thesi2)
```

```
for(j in 2:length(thesi2))
```

```
{pin[j]=dedo$etos[(dedo$foit==dedo$foit[thesi2[j]])&(dedo$Mathima=='12101')&(dedo$Per  
ase>4)]-min(dedo$etos[dedo$foit==dedo$foit[thesi2[j]]],na.rm =T)
```

```
}
```

```
###o pinakas me to etos pou perase to mathima -1###
```

```
pin
```

```
summary(pin)
```

```
###prostheto +1 gia na exo to akrives etos##
```


####

```
dedomena2=read.csv2("g:\\oliko_mam.csv",header=F)
```

```
dedo=data.frame(etos=dedomena2$V1,kodikos_period=dedomena2$V2,kodikos=dedomena2$V3,foit=dedomena2$V5,foitExaminoMath=dedomena2$V6,Mathima=dedomena2$V7,Perase=dedomena2$V8)
```

```
summary(dedo)
```

```
d=duplicated(dedo$foit)
```

```
length(d)
```

```
thesi=0
```

```
for(l in 1:858681){ if (d[l]=='FALSE') thesi=c(thesi,l)}
```

```
length(dedo$Mathima[(dedo$foit==dedo$foit[thesi[2]])&(dedo$Mathima=='12101')&(dedo$Perase>=0)])
```

```
length(dedo$Mathima[(dedo$foit==dedo$foit[thesi[2]])&(dedo$Mathima=='12101')&(dedo$Perase>2)])
```

```
length(dedo$Mathima[(dedo$foit==dedo$foit[thesi[2]])&(dedo$Mathima=='12101')&(dedo$Perase>4)])
```

```
synolo=matrix(1000,6016,3)
```

```
for(j in 2:6016){
```

```
synolo[j,1]=length(dedo$Mathima[(dedo$foit==dedo$foit[thesi[j]])&(dedo$Mathima=='12101')&(dedo$Perase>=0)])
```

```
synolo[j,2]=length(dedo$Mathima[(dedo$foit==dedo$foit[thesi[j]])&(dedo$Mathima=='12101')&(dedo$Perase>2)])
```

```
synolo[j,3]=length(dedo$Mathima[(dedo$foit==dedo$foit[thesi[j]])&(dedo$Mathima=='12101')&(dedo$Perase>=5)])
```

```
}
```

```
library(xlsx)
```

```
synolo2=synolo
```

```
write.xlsx(synolo2, "g:/apeiro22.xlsx")
```

```

synolo2=synolo[-1,]
synolo2=synolo2[-length(synolo2[,1]),]
oliko=synolo2[,1][synolo2[,3]==1]
meriko=synolo2[,2][synolo2[,3]==1]
summary(synolo2[,2][synolo2[,3]==1])
summary(oliko)

write.xlsx(synolo2, "g:/apeiro22.xlsx")

dedomena2=read.csv2("g:\\oliko_mam.csv",header=F)

dedo=data.frame(etos=dedomena2$V1,kodikos_period=dedomena2$V2,kodikos=dedomena2$V3,foit=dedomena2$V5,foitExaminoMath=dedomena2$V6,Mathima=dedomena2$V7,Perase=dedomena2$V8)

summary(dedo)

thesi2=0
d2=0
for(l in 1:length(dedo$Mathima))
{ if((dedo$Mathima[l]=='12101')&(dedo$Perase[l]>4))
d2=c(d2,l)
}
thesi2=d2

#pote to perna
dedo$etos[(dedo$foit==dedo$foit[thesi2[2]])&(dedo$Mathima=='12101')&(dedo$Perase>4)]

#pote mpike
min(dedo$etos[dedo$foit==dedo$foit[thesi2[3]]],na.rm =T)

pin=0

```

```

length(thesi2)

for(j in 2:length(thesi2))

{pin[j]=dedo$etos[(dedo$foit==dedo$foit[thesi2[j]])&(dedo$Mathima=='12101')&(dedo$Per
ase>4)]-min(dedo$etos[dedo$foit==dedo$foit[thesi2[j]]],na.rm =T)

}

library(xlsx)

write.xlsx(pin, "g:/apeiro223.xlsx")

dedomena2=read.csv2("g:\\oliko_mam.csv",header=F)

dedo=data.frame(etos=dedomena2$V1,kodikos_period=dedomena2$V2,kodikos=dedomena2$V3,foit=dedomena2$V5,foitExaminoMath=dedomena2$V6,Mathima=dedomena2$V7,Perase=dedomena2$V8)

summary(dedo)

d=duplicated(dedo$foit)

length(d)

thesi=0

for(l in 1:858681){ if (d[l]=='FALSE') thesi=c(thesi,l)}

##6016=length(thesi)##

synolo=matrix(1000,6016,3)

#####edo gia opoioidipote mathima allazo MONO ton kodiko 12101 ston antistoixo#####

##### Statistiki I as poume einai to 12541, Grammiki I 12121#####

```

```

for(j in 2:6016){

synolo[j,1]=length(dedo$Mathima[(dedo$foit==dedo$foit[thesi[j]])&(dedo$Mathima=='121
01')&(dedo$Perase>=0)])

synolo[j,2]=length(dedo$Mathima[(dedo$foit==dedo$foit[thesi[j]])&(dedo$Mathima=='121
01')&(dedo$Perase>2)])

synolo[j,3]=length(dedo$Mathima[(dedo$foit==dedo$foit[thesi[j]])&(dedo$Mathima=='121
01')&(dedo$Perase>=5)])

}

library(xlsx)

synolo2=synolo

###an thelo ta stoixeia se excel

write.xlsx(synolo2, "g:/apeiro22.xlsx")

synolo2=synolo[-1,]

synolo2=synolo2[-length(synolo2[,1]),]

###prospatheies mexri tin proti epityxia

summary(synolo2[,1][synolo2[,3]==1])

###prospatheies mexri tin proti epityxia alla exairo otan pirane 0-1-2

summary(synolo2[,2][synolo2[,3]==1])

```

6. Βιβλιογραφία

- [1] Brown Helen , Prescott Robin. Applied Mixed Models in Medicine Wiley (1999)
- [2] Σιάννης Φώτιος. Introduction to Survival Analysis (2008)
- [3] Παπαδάτος Νικόλαος Θεωρία Πιθανοτήτων, (Αθήνα 2006)
- [4] Collet David, Modelling Survival data in Medical Research(2004)
- [5] Δαμιανού Χαράλαμπος, Κούτρας Μάρκος Εισαγωγή στη Στατιστική I, II(ΑΘΗΝΑ, 1998)
- [6] ΕΚΠΑ, Τμήμα Μαθηματικών, Έκθεση εσωτερικής αξιολόγησης, modip.uoa.gr/fileadmin/modip.uoa.gr/...Reports/IER-Ma8hmatiko_2010.pdf (Μάιος 2010)
- [7] Χαραλαμπίδης Χαράλαμπος Θ. Πιθανοτήτων και Εφαρμογές I και II
- [8] Μελιγκοτσίδου Λ. ,Γενικευμένα γραμμικά μοντέλα -Σημειώσεις παραδόσεων (eclass)
- [9]] Σιάννης Φώτιος -Introduction to Longitudinal Data Analysis (April 27, 2012)