ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΩΝ ΘΕΩΡΙΑ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΜΕ ΤΗ ΧΡΗΣΗ ΤΟΥ ΣΤΑΤΙΣΤΙΚΟΥ ΠΑΚΕΤΟΥ R

ΣΤΑΥΡΟΥΛΑ ΠΑΝΤΖΑΒΕΛΗ

ΕΠΙΒΛΕΠΟΥΣΑ: ΒΟΝΤΑ ΦΙΛΙΑ

ΑΝΑΠΛΗΡΩΤΡΙΑ ΚΑΘΗΓΗΤΡΙΑ Ε.Μ.Π.

Επιτροπή Καθηγητών: Βόντα Φιλία, Καρώνη Χρυσηίς,

Κουκουβίνος Χρήστος

Αθήνα, Μάρτιος 2017

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ

ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΩΝ ΘΕΩΡΙΑ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΜΕ ΤΗ ΧΡΗΣΗ ΤΟΥ ΣΤΑΤΙΣΤΙΚΟΥ ΠΑΚΕΤΟΥ R

ΣΤΑΥΡΟΥΛΑ ΠΑΝΤΖΑΒΕΛΗ

ΕΠΙΒΛΕΠΟΥΣΑ: ΒΟΝΤΑ ΦΙΛΙΑ

ΑΝΑΠΛΗΡΩΤΡΙΑ ΚΑΘΗΓΗΤΡΙΑ Ε.Μ.Π.

Επιτροπή Καθηγητών: Βόντα Φιλία, Καρώνη Χρυσηίς,

Κουκουβίνος Χρήστος

Αθήνα, Μάρτιος 2017

Ευχαριστίες

Θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες σε όσους συνέβαλαν στο να φέρω εις πέρας την παρούσα προπτυχιακή Διπλωματική Εργασία. Ιδιαίτερα θα ήθελα να ευχαριστήσω θερμά την Καθηγήτρια Εθνικού Μετσόβιου Πολυτεχνείου κα. Φιλία Βόντα. Η υπομονή της, η άμεση καθοδήγησή της και το συνεχές ενδιαφέρον της ήταν απαραίτητα συστατικά προκειμένου να ολοκληρωθεί με επιτυχία η εργασία μου. Επίσης αξίζει να ευχαριστήσω την οικογένεια μου και τους φίλους μου για την κατανόησή τους και την υποστήριξη που μου παρέχουν σε όλη τη διάρκεια των σπουδών μου.

Περίληψη

Η στατιστική μοντελοποίηση, η επιλογή δηλαδή του κατάλληλου μοντέλου ανάμεσα από ένα πλήθος υποψήφιων προσεγγιστικών μοντέλων, θεωρείται μία παραδοσιακή προσέγγιση στη στατιστική συμπερασματολογία. Η επιλογή του βέλτιστου μοντέλου επιτυγχάνεται μέσα από την ανάπτυξη κριτηρίων πληροφορίας όπως είναι το AIC (κριτήριο πληροφορίας του Akaike), το BIC (μπεϋζιανό κριτήριο πληροφορίας) και τις επεκτάσεις τους (AICc, BICc). Επιπλέον παρατίθενται κάποιες μέθοδοι επιλογής μεταβλητών οι οποίες αποσκοπούν στην επιλογή των μεταβλητών που προσεγγίζουν ικανοποιητικά το μοντέλο όπως είναι η μέθοδος προσθήκης και αφαίρεσης μεταβλητών. Με τη χρήση τους επιδιώκεται η επιλογή του κατάλληλου υποσυνόλου ανεξάρτητων μεταβλητών που περιγράφουν με τον καλύτερο δυνατό τρόπο τη μεταβλητή απόκρισης ή εξαρτημένη μεταβλητή.

Στόχος της εργασίας είναι η σύγκριση των δύο αυτών κριτηρίων πληροφορίας, AIC και BIC, προκειμένου να γίνει η βέλτιστη μοντελοποίηση. Και τα δύο κριτήρια εξετάζονται και συγκρίνονται ως προς τη συνέπεια και την αποδοτικότητά τους μέσα από κατάλληλα θεωρήματα. Τέλος, η σύγκριση τους ολοκληρώνεται μέσα από προσομοιώσεις μοντέλου πολλαπλής παλινδρόμησης με τη βοήθεια του στατιστικού πακέτου R καθώς επίσης μέσα από προσαρμογή σε πραγματικά δεδομένα. Όσον αφορά τις προσομοιώσεις, αφού υπολογιστεί το ποσοστό που επιλέγεται το σωστό μοντέλο σε όλες τις επαναλήψεις και τα σφάλματα τύπου I και II, εξάγονται συμπεράσματα για το ποιο από τα δύο κριτήρια πληροφορίας συμπεριφέρεται καλύτερα.

Abstract

The statistical modeling, the choice that is, of the appropriate model among a plurality of candidate approximate models, is considered as a traditional approach to statistical inference. In order to identify the optimal model, several information criteria have been developed such as AIC (Akaike information criterion), the BIC (Bayesian information criterion) and their extensions (AICc, BICc). Furthermore, several variable selection methods are used in order to select the variables that best fit our data such as the Forward Procedure and the Backward Elimination Procedure. Their aim is to identify the most appropriate subset of the independent variables that significantly describe the response or dependent variable.

The purpose of this thesis is to compare the two information criteria, AIC and BIC, in order to achieve best model fitting. Both criteria are examined and compared for consistency and efficiency using appropriate theorems. Finally, simulations from a multiple regression model contribute to that comparison, using the statistical package R as well as with an application of the general linear model to real data. As far as the simulations are concerned, in order to find which of the two model selection criteria behaves better, the percentage of times that a correct model is selected is being computed as well as the errors of type I.

Περιεχόμενα

Εισαγωγή

Αντικείμενο Εργασίας

Η επιλογή μοντέλων, γνωστή ευρέως ως Model Selection, αποτελεί σημαντικό κομμάτι της Στατιστικής. Ενώ σε ένα σύνολο δεδομένων μπορεί κανείς να προσαρμόσει πολλά μοντέλα, είναι δύσκολο να βρεθεί εκείνο που προσαρμόζεται καλύτερα στα δεδομένα. Από ένα σύνολο υποψήφιων μεταβλητών πρέπει να επιλεχθεί εκείνο το υποσύνολο που επιδρά σημαντικά στην εξαρτημένη μεταβλητή ή μεταβλητή απόκρισης. Για την εύρεση του βέλτιστου μοντέλου και την επιλογή των κατάλληλων ανεξάρτητων μεταβλητών έχουν αναπτυχθεί διάφορα κριτήρια πληροφορίας που έχουν ως στόχο να επιλέξουν εκείνο το μοντέλο που προσεγγίζει καλύτερα το πραγματικό μοντέλο.

Σκοπός της παρούσας διπλωματικής εργασίας λοιπόν είναι η ανάπτυξη και η σύγκριση δύο κριτηρίων επιλογής μοντέλων, του AIC (κριτήριο πληροφορίας του Akaike) και του BIC (το μπεϋζιανό κριτήριο πληροφορίας). Η μελέτη καλείται να αποδείξει ποιο από τα δύο κριτήρια είναι πιο αξιόπιστο και να γίνει η μεταξύ τους σύγκριση ως προς τη συνέπεια και αποδοτικότητά τους.

Μέσα Έρευνας

Το πρόγραμμα που χρησιμοποιήθηκε για τη σύγκριση των δύο κριτηρίων είναι το στατιστικό πακέτο R (έκδοση Rx64 3.3.0), πρόγραμμα στατιστικής ανάλυσης. Έγινε εγκατάσταση μέσα από την εργαλειοθήκη της R, το πακέτο MASS, το οποίο ήταν απαραίτητο για να αναγνωριστούν οι δοθέντες εντολές του κώδικα. Τέλος για την εξαγωγή των αποτελεσμάτων και τη δημιουργία των πινάκων στους οποίους παρουσιάζονται οι τιμές των δύο κριτηρίων και τα σφάλματα τύπου I και II χρησιμοποιήθηκε το Microsoft Excel.

Διάρθρωση

Στο 1ο κεφάλαιο της διπλωματικής εργασίας γίνεται αναλυτική περιγραφή της ανάλυσης παλινδρόμησης για το απλό και το γραμμικό μοντέλο. Αναλύονται μέθοδοι εκτίμησης των παραμέτρων ενός μαθηματικού μοντέλου όπως είναι η μέθοδος ελαχίστων τετραγώνων, η μέθοδος ροπών, η μέθοδος εκτίμησης μέγιστης πιθανοφάνειας, στατιστικοί έλεγχοι υποθέσεων και παρατίθενται τα μέτρα καταλληλόλητας, R^2 και $C_p - Mallows$.

Στο 2ο κεφάλαιο αναλύεται το κριτήριο πληροφορίας AIC (Akaike information criterion) συνοδευόμενο με τη μαθηματική του απόδειξη. Δίνεται η βασική έννοια της απόστασης

Kullback-Leibler καθώς έπαιξε σημαντικό ρόλο στη συλλογική πορεία του Akaike για την ανάπτυξη του κριτηρίου. Τέλος γίνεται μία μικρή αναφορά και σε άλλα κριτήρια παρόμοια του AIC καθώς και στις επεκτάσεις τους (AICc, TIC, QAIC, QAICc).

Στο 3^{ο} κεφάλαιο γίνεται μία εισαγωγή στη Μπεϋζιανή Στατιστική και αναλύεται το Μπεϋζιανό Κριτήριο Πληροφορίας, BIC (Bayesian Criterion Information). Το κριτήριο συνοδεύεται από τη μαθηματική του απόδειξη καθώς και από τα θεωρήματα συνέπειας και αποδοτικότητας προκειμένου να εξεταστεί η αξιοπιστία του σε σύγκριση με το AIC.

Στο 4^{ο} κεφάλαιο αρχικά αναλύονται οι μέθοδοι προσθήκης και αφαίρεσης μεταβλητών. Έπειτα γίνονται προσομοιώσεις ενός δοθέντος πολλαπλού μοντέλου παλινδρόμησης μέσα από τις οποίες γίνεται σύγκριση του AIC με το BIC και εξάγονται συμπεράσματα για το ποιο από τα δύο κριτήρια είναι πιο αξιόπιστο. Υπολογίζονται τα σφάλματα τύπου I και II και πόσες φορές επιλέγεται το σωστό μοντέλο. Οι προσομοιώσεις έγιναν στο στατιστικό πακέτο R (έκδοσης 3.3.0).

Σε όλη τη διπλωματική εργασία χρησιμοποιήθηκαν αρκετά παραδείγματα προκειμένου να γίνουν κατανοητές βασικές έννοιες και μέθοδοι που χρησιμοποιήθηκαν καθώς και για να γίνει αντιληπτή η χρήση των κριτηρίων πληροφορίας προκειμένου να επιλεχθεί το «καλύτερο» μοντέλο που προσαρμόζεται κατάλληλα στα δεδομένα.

Κεφάλαιο 1ο

Βασικές έννοιες

1.1 Εισαγωγή στη Στατιστική

Διανύουμε μία εποχή τεχνολογικής ανάπτυξης και πειραματικής έκρηξης στην οποία η **Στατιστική** αποτελεί σημαντικό παράγοντα. Όλοι μας έχουμε βρεθεί στη θέση να προσπαθούμε να βρούμε τη βέλτιστη λύση προκειμένου να λύσουμε ή πιθανόν να προσεγγίσουμε ένα πρόβλημα. Η Στατιστική λοιπόν είναι μία επιστήμη η οποία αφορά τη συλλογή, την ταξινόμηση, την παρουσίαση και την ερμηνεία παρατηρήσεων με κύριο σκοπό την εξαγωγή συμπερασμάτων για τη λήψη βέλτιστων αποφάσεων. Είναι η επιστήμη της αβεβαιότητας, την οποία εμείς καλούμαστε να προσδιορίσουμε. Διακρίνεται στην περιγραφική στατιστική (descriptive statistics) η οποία χρησιμοποιεί αριθμητικές και γραφικές μεθόδους για την εξαγωγή συμπερασμάτων και στην επαγωγική στατιστική (inferential statistics) η οποία μελετά ολόκληρους τους πληθυσμούς μέσα από τα δεδομένα του δείγματος και εξάγει συμπεράσματα.

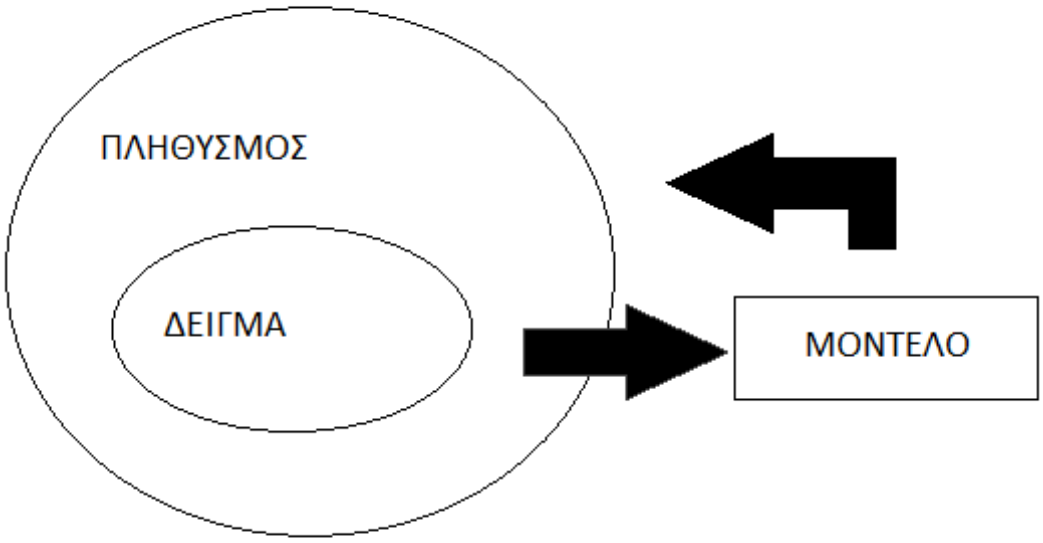
Η στατιστική είναι ο κλάδος των εφαρμοσμένων επιστημών μέσα από τον οποίο χρησιμοποιούνται μαθηματικά μοντέλα για την επίλυση προβλημάτων σε διάφορους τομείς όπως ιατρική, χρηματοοικονομικά, μάρκετινγκ, αστρονομία, αρχαιολογία, ψυχολογία κ.α. Τα άτομα που ασχολούνται με τη στατιστική ονομάζονται στατιστικοί αναλυτές των οποίων αρμοδιότητα είναι η ανάπτυξη κατάλληλων μοντέλων και η συμπερασματολογία, η οποία επιτυγχάνεται με τη χρήση στατιστικών πακέτων όπως είναι η R,το Minitab,SPSS κ.α.

Η λέξη στατιστική (statistics) προέρχεται από τη λατινική λέξη "status" που χρησιμοποιήθηκε για να αποδώσει την έννοια της συγκροτημένης πολιτείας. Τον όρο στατιστική αναφέρει ο Σωκράτης (Ξενοφώντας "Απομνημονεύματα'') καθώς και ο Αριστοτέλης στο έργο του ''Πολιτεία'' και στις αρχές του δέκατου ένατου αιώνα καθιερώθηκε ως η έννοια της συλλογής και ταξινόμησης δεδομένων.

1.2 Ανάλυση Παλινδρόμησης

Τα μαθηματικά μοντέλα σήμερα αποτελούν την πιο διαδεδομένη μέθοδο μελέτης φυσικών, κοινωνικών, οικονομικών και ιατρικών φαινομένων. Κύριο συστατικό για την επιτυχημένη κατασκευή ενός μοντέλου είναι η παρατήρηση και η εμπειρία προκειμένου να

διατυπωθούν οι κατάλληλες πληροφορίες και να γίνει η σύγκριση του προτεινόμενου μοντέλου με τα ήδη υπάρχοντα αφού ανατροφοδοτηθεί με νέα στοιχεία.



Διάγραμμα 1.1: Διαδικασία παραγωγής μοντέλου

Πηγή: Στατιστικά Πακέτα 1, Στέλιος Ζήμερας, Πανεπιστήμιο Αιγαίου, 2003

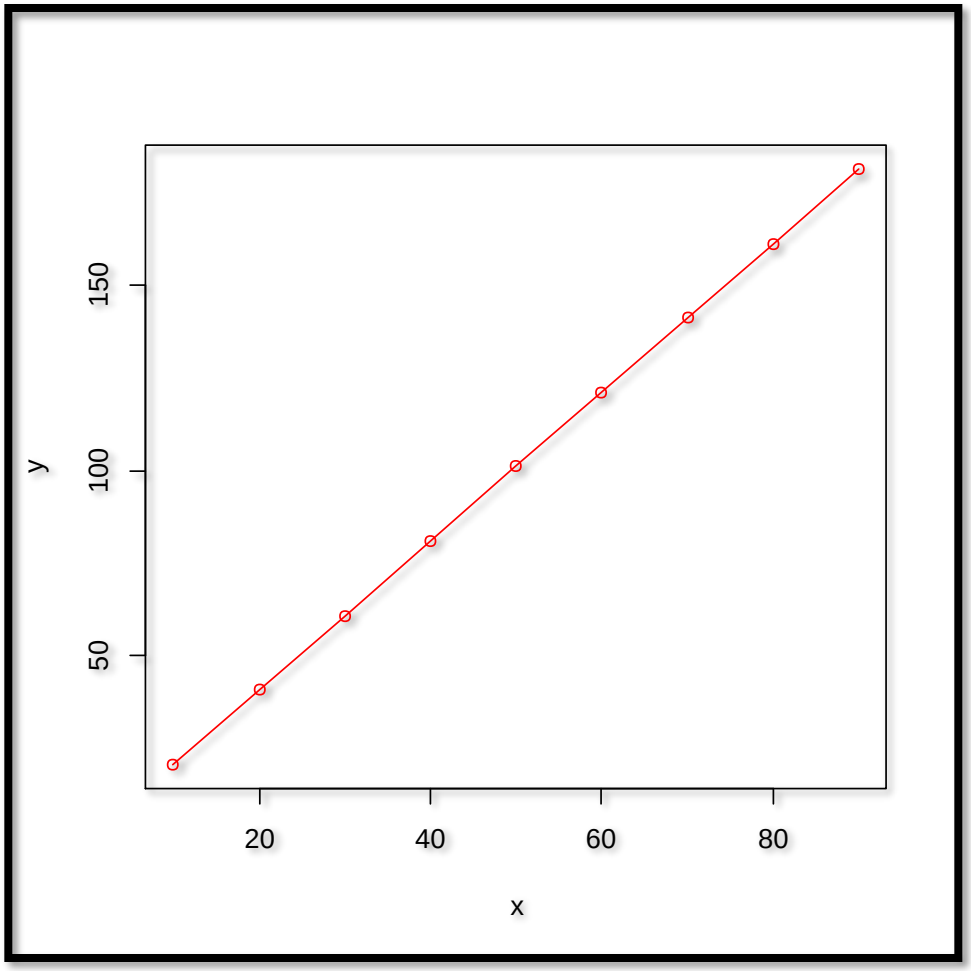
Η παραγωγή μοντέλου που αναφέραμε παραπάνω περιγράφει τη στοχαστική δομή μεταξύ κάποιων στοχαστικών μεταβλητών. Με μια πρώτη σκέψη θα μπορούσε εύκολα να αναρωτηθεί κάποιος τι είναι ακριβώς οι στοχαστικές μεταβλητές και σε τι κατηγορίες διακρίνονται. Εμείς καλούμαστε να εξετάσουμε τη σχέση μεταξύ δύο ή περισσότερων τέτοιων μεταβλητών. Αυτό επιτυγχάνεται με τη βοήθεια της ανάλυσης παλινδρόμησης (regression analysis). Η ανάλυση παλινδρόμησης είναι μία στατιστική τεχνική μέσα από την οποία μπορούμε να μελετήσουμε και να εκτιμήσουμε την τιμή μιας εξαρτημένης μεταβλητής ή μεταβλητή απόκρισης (dependent, response variable) σε σχέση με τις τιμές των ανεξάρτητων ή ελεγχόμενων ή επεξηγηματικών μεταβλητών (independent, predictor, explanatory variables). Ανεξάρτητη μεταβλητή X θεωρείται εκείνη την οποία μπορούμε να ελέγξουμε ή να

καθορίσουμε την τιμή της (π.χ. το ύψος της διαφημιστικής δαπάνης ενός προϊόντος) ενώ εξαρτημένη μεταβλητή Y είναι εκείνη της οποίας η τιμή εξαρτάται από τη μεταβολή των τιμών των ανεξάρτητων μεταβλητών (π.χ. η αντοχή ενός υλικού) (Γ. Παπαδόπουλος-Ανάλυση Παλινδρόμησης ,pdf)

Η συναρτησιακή λοιπόν σχέση που συνδέει τις μεταβλητές X,Y είναι Y=f(X) και έχει καθιερωθεί ως μία ντετερμινιστική σχέση (deterministic relationship) όπου η τιμή της μεταβλητής Y καθορίζεται απόλυτα από την τιμή της X. Στον τομέα της στατιστικής όμως η σχέση αυτή μετατρέπεται σε στοχαστική (stochastic relationship) και παίρνει τη μορφή Y=f(X)+ε όπου ε είναι το τυχαίο σφάλμα το οποίο θεωρείται ότι ακολουθεί γνωστή κατανομή και συνήθως θεωρείται ότι $\varepsilon \sim N(0, \sigma^2)$ δηλαδή το σφάλμα έχει μέση τιμή 0 και σ^2 άγνωστη. Επιδιώκεται λοιπόν να δημιουργηθεί ένα μοντέλο,Y=f(X), έτσι ώστε να προσδιοριστεί η τιμή της μεταβλητής Y με βάση τις τιμές της μεταβλητής X.

1.2.1 Απλό Γραμμικό Μοντέλο

Στη στατιστική μοντελοποίηση προσπαθούμε να μελετήσουμε την εξάρτηση της μεταβλητής Y από μία ή περισσότερες μεταβλητές $X_1, X_2, \ldots X_n$. Εάν έχουμε μία μόνο ανεξάρτητη μεταβλητή X και η συνάρτηση που χρησιμοποιούμε είναι γραμμική τότε πρόκειται για το απλό γραμμικό μοντέλο (linear model) το οποίο είναι της μορφής $Y=\beta_0+\beta_1 X+\varepsilon \Leftrightarrow E(Y|X)=\beta_0 + \beta_1 X$ όπου $\varepsilon \sim N(0, \sigma 2)$ και β_0, β_1 σταθερές. Η ευθεία αυτή $E(Y|X)=\beta_0 +\beta_1 X$ ονομάζεται ευθεία παλινδρόμησης. Στο παρακάτω σχήμα απεικονίζεται η ευθεία παλινδρόμησης (απλή γραμμική παλινδρόμηση, simple-linear-regression, όπου υπάρχει τέλεια συσχέτιση μεταξύ της μεταβλητής απόκρισης Y και της επεξηγηματικής μεταβλητής X.



Διάγραμμα 1.2: Διάγραμμα Ευθείας Παλινδρόμησης

Η μέση τιμή της Y για ορισμένη τιμή της X βρίσκεται πάνω σε μία ευθεία με σταθερό όρο β_0 και κλίση β_1. Οι β_0,β_1,σ^2 είναι άγνωστες παράμετροι τις οποίες θέλουμε να εκτιμήσουμε είτε με τη μέθοδο ελαχίστων τετραγώνων είτε με τη μέθοδο μέγιστης πιθανοφάνειας όπως θα δούμε αναλυτικά παρακάτω. Η παράμετρος β_0 εκφράζει την μέση τιμή της μεταβλητής Y όταν η μεταβλητή X πάρει την τιμή 0 ενώ η παράμετρος β_1 εκφράζει το πόσο θα μεταβληθεί η αναμενόμενη τιμή της Y εάν η μεταβλητή X αυξηθεί κατά μία μονάδα. Αξιοσημείωτο είναι ότι ο όρος γραμμικό αναφέρεται στις παραμέτρους και όχι στις μεταβλητές.

Στο απλό γραμμικό μοντέλο ισχύουν οι εξής παραδοχές:

- Γραμμικότητα (linearity): Οι μέσες τιμές των Y είναι γραμμικές συναρτήσεις των X,βρίσκονται δηλαδή σε ευθεία γραμμή, κάτι που μπορεί να ελεγχθεί μέσα από το διάγραμμα διασποράς καθώς επίσης και από τα διαγράμματα υπολοίπων (residual

plots) όπου αναπαριστάνονται γραφικά τα (x_i,$\widehat{\varepsilon_i}$) ή τα ($\widehat{y_i}$, $\widehat{\varepsilon_i}$) όπου τα σημεία θα πρέπει να είναι κατανεμημένα κατά συστηματικό τρόπο κοντά στην ευθεία παλινδρόμησης και όχι τυχαία προκειμένου να επιτευχθεί η γραμμικότητα.

- Ομοσκεδαστικότητα (Homoscedasticity): Η διασπορά των κατανομών των Y είναι σταθερή για κάθε τιμή της X δηλαδή $Var(Y/X=x_i)=\sigma^2$.Η ομοσκεδαστικότητα μπορεί πάλι να ελεγχθεί μέσα από διαγράμματα διασποράς ή διαγράμματα υπολοίπων σε σχέση με τις τιμές των ανεξάρτητων μεταβλητών X όπου θα πρέπει τα ζεύγη των τιμών αυτών να κατανέμονται τυχαία.
- Ανεξαρτησία σφαλμάτων (Independence): Οι τιμές των Y είναι ανεξάρτητες μεταξύ τους για κάθε επίπεδο των X. Κάνοντας το διάγραμμα υπολοίπων σε σχέση με μία σειρά δεδομένων ελέγχεται εάν τα υπόλοιπα κατανέμονται με τυχαίο τρόπο προκειμένου να υπάρχει ανεξαρτησία.
- Κανονικότητα (Normality) σφαλμάτων: Η Y ακολουθεί κανονική κατανομή σε όλα τα επίπεδα της X. Μέσα από την κατασκευή ιστογραμμάτων μπορεί να ελεγχθεί αν η υπόθεση κανονικότητας σφαλμάτων είναι ορθή.

1.2.2 Γενικό Γραμμικό Μοντέλο

Πολλές φορές είναι πρακτικό να χρησιμοποιήσουμε περισσότερες από μία ανεξάρτητες μεταβλητές προκειμένου να μελετήσουμε με ακρίβεια κάποιο πρόβλημα ή φαινόμενο. Μοντέλα παλινδρόμησης που περιέχουν δύο ή περισσότερες μεταβλητές ονομάζονται γενικά ή πολλαπλά γραμμικά μοντέλα (multiple linear models) και έχουν τη μορφή

$$Yi = \beta 0 \ + \beta 1 Xi1 + \ \dots + \beta k Xik \ + \varepsilon i \ (\mu \varepsilon \ i = 1, \dots \ , n) \Leftrightarrow$$

$$\Leftrightarrow E(Y|Xi) = \ \beta 0 \ + \beta 1 Xi1 + \ \dots + + \beta k Xik, \acute{o}\pi ou \ \varepsilon i \sim N(0, \sigma 2),$$

ε_i ανεξάρτητα και β_0 καλείται ο σταθερός όρος της συνάρτησης παλινδρόμησης δηλαδή είναι η τιμή που παίρνει η μεταβλητή Y όταν $Xi1 = Xi2 = \cdots = Xik = 0$ ενώ ο συντελεστής β_j για j=1,..,k δείχνει την μεταβολή της τιμής της μεταβλητής Y όταν η μεταβλητή X_j αυξηθεί κατά μία μονάδα και οι υπόλοιπες μεταβλητές παραμείνουν σταθερές. Υπό τη μορφή πινάκων το γενικό γραμμικό μοντέλο γράφεται ως εξής:

$$Y = \tilde{X}\beta + \varepsilon$$

Όπου Y=$\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix}$, $\tilde{X}$=$\begin{pmatrix} 1 & X_{11} & X_{1k} \\ \vdots & \ddots & \vdots \\ 1 & \cdots X_{n1} & X_{nk} \end{pmatrix}$, β=$\begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{pmatrix}$, ε=$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix}$

Όπως και στο απλό γραμμικό μοντέλο έτσι και στο γενικό ισχύει η γραμμικότητα, η ομοσκεδαστικότητα, η ανεξαρτησία σφαλμάτων και η κανονικότητα σφαλμάτων και μπορεί να ελεγχθεί όπως περιγράψαμε παραπάνω με τη διαφορά ότι στο απλό γραμμικό μοντέλο τη γραμμικότητα την ελέγχουμε με το διάγραμμα διασποράς των (xi, yi) ενώ στο γενικό γραμμικό μοντέλο θα κατασκευάσουμε το διάγραμμα διασποράς για κάθε επεξηγηματική μεταβλητή ξεχωριστά. Κάποια παραδείγματα μοντέλων πολλαπλής γραμμικής παλινδρόμησης είναι τα εξής:

- $Y=\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \varepsilon$ (μοντέλο πρώτου πολυωνυμικού βαθμού)
- $Y=\beta_0 + \beta_1X_1 + \beta_2X_2^2 + \beta_3X_3^2 + \varepsilon$ (μοντέλο δευτέρου πολυωνυμικού βαθμού)

Παρατηρούμε δηλαδή όπως αναφέραμε και παραπάνω ότι η γραμμικότητα αναφέρεται στη γραμμική σχέση των συντελεστών και όχι των μεταβλητών. Για κάθε τιμή $Xi1, \dots, Xik$ μπορούμε να υπολογίσουμε την προβλεπόμενες τιμές $\widehat{Y_\iota} = \widehat{\beta_0} + \widehat{\beta_1}X_{i1} + \cdots + +\widehat{\beta_k}X_{ik}$ όπου $\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_k}$ είναι οι εκτιμήσεις των συντελεστών οι οποίες υπολογίζονται από τη μέθοδο ελαχίστων τετραγώνων όπως θα δούμε αναλυτικά παρακάτω.

1.3 Μέθοδοι Εκτίμησης Παραμέτρων

1.3.1 Μέθοδος Ελαχίστων Τετραγώνων

Η μέθοδος ελαχίστων τετραγώνων η οποία χρησιμοποιήθηκε για πρώτη φορά από τον Γάλλο μαθηματικό Legendre και μετέπειτα από τον Γερμανό μαθηματικό Gauss,περιγράφει τη στοχαστική εξάρτηση μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών. Στην περίπτωση του απλού γραμμικού μοντέλου σκοπός μας είναι η εκτίμηση των συντελεστών β_0, β_1 προκειμένου να προσδιοριστεί μία εκτίμηση $\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1 X_1}$ της ευθείας $E(Y|X_1)= \beta_0 + \beta_1 x_1$, η οποία ονομάζεται ευθεία ελαχίστων τετραγώνων. Έστω ότι δίνονται οι τιμές (x_1 ,y_1) του τυχαίου δείγματος (x_1,Y_1). Έτσι η στοχαστική προσέγγιση του μοντέλου γίνεται

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X_1$$

Η απόκλιση της πραγματικής τιμής Y από την παραπάνω προσεγγιστική τιμή αποτελεί το σφάλμα της παλινδρόμησης και συμβολίζεται ως:

$$\varepsilon = Y - (\beta_0 - \beta_1 X_1).$$

Προκειμένου λοιπόν να εκτιμήσουμε τις παραμέτρους β_0,β_1 θα ελαχιστοποιήσουμε τις ποσότητες ε_i και συγκεκριμένα θα βρούμε τις τιμές των β_0 και β_1 για τις οποίες ελαχιστοποιείται το άθροισμα των τετραγώνων των ε_i. Δηλαδή:

- $\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$ (1)

Παραγωγίζουμε την (1) ως προς β_0 και β_1 και εξισώνουμε τις παραστάσεις με την τιμή μηδέν δημιουργώντας έτσι τις κανονικές εξισώσεις.

Συγκεκριμένα για το απλό γραμμικό μοντέλο όπου έχουμε να εκτιμήσουμε δύο παραμέτρους, τις β_0 και β_1,οι κανονικές εξισώσεις είναι οι εξής:

- $\frac{\partial \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_0} = 0 \Leftrightarrow -\sum_{i=1}^{n} y_i + n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = 0$ (2)

- $\frac{\partial \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1} = 0 \Leftrightarrow -\sum_{i=1}^{n} y_i x_i + \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = 0$ (3)

Λύνοντας τις (2), (3) ως προς β_0,β_1:

- $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$
- $\widehat{\beta_1} = \frac{-n\bar{x}\bar{y} + \sum_{i=1}^{n} y_i x_i}{-n\bar{x}^2 + \sum_{i=1}^{n} x_i^2}$ ή $\widehat{\beta_1} = \frac{S_{xy}}{S_x^2}$

Χρησιμοποιώντας λοιπόν τις παραπάνω σχέσεις προκύπτει η ευθεία ελαχίστων τετραγώνων της μορφής $\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X$ η οποία είναι η εκτίμηση της ευθείας $E(Y|X)=\beta_0 + \beta_1 X$.Γνωρίζοντας ότι:

$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$

$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$

$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

η εκτιμήτρια ευθεία παλινδρόμησης γίνεται $\hat{Y} = \bar{y} - \widehat{\beta_1}\bar{x} + \widehat{\beta_1} X = \bar{y} + \widehat{\beta_1}(X - \bar{x}) = \bar{y} + \frac{S_{xy}}{S_{xx}}(X - \bar{x})$ όπου $\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$. Είναι εύκολο να παρατηρήσουμε ότι η ευθεία ελαχίστων τετραγώνων διέρχεται από το σημείο $(\bar{x}, \bar{y})$. Αξιοσημείωτο είναι ότι οι τιμές $\widehat{\beta_0}, \widehat{\beta_1}$ είναι οι

εκτιμήσεις των β_0,β_1 με βάση το δείγμα $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Οι εκτιμήσεις των παραμέτρων διαφέρουν προφανώς από δείγμα σε δείγμα.

1.3.2 Μέθοδος Μέγιστης Πιθανοφάνειας

Το κριτήριο μέγιστης πιθανοφάνειας είναι ένα από τα πιο διαδεδομένα κριτήρια εκτίμησης. Επιλέγοντας το μοντέλο με τη μεγαλύτερη τιμή της λογαριθμικής πιθανοφάνειας ανάμεσα σε ένα πλήθος προσαρμοσμένων μοντέλων μπορούμε να πάρουμε το καταλληλότερο μοντέλο το οποίο είναι πιο κοντά στην πραγματική κατανομή. Δηλαδή εάν επιλέξουμε τις τιμές των παραμέτρων οι οποίες μεγιστοποιούν τη λογαριθμική πιθανοφάνεια τότε θα οδηγηθούμε σε αρκετά καλό μοντέλο. Η μέθοδος αυτή λοιπόν ονομάζεται μέθοδος μέγιστης πιθανοφάνειας και είναι η πιο γνωστή μέθοδος εύρεσης μιας εκτιμήτριας για τις παραμέτρους θ μιας κατανομής F. Θεωρείται αρκετά αξιόπιστη διότι με μία απλή διαδικασία οδηγούμαστε σε εκτιμήτριες με πολύ καλές ιδιότητες.

Έστω ένα τυχαίο δείγμα μεγέθους n, δηλαδή ένα σύνολο ανεξάρτητων τυχαίων μεταβλητών $x1, x2, \ldots, xn$ με σ.π.π. $f(x;\theta)$ που εξαρτάται από ένα διάνυσμα αγνώστων παραμέτρων $\theta = (\theta 1, \theta 2, \ldots \theta\kappa)$.Η εκτίμηση της παραμέτρου θ μπορεί να γίνει με τη βοήθεια της εκτιμήτριας συνάρτησης της θ η οποία ουσιαστικά είναι μια στατιστική συνάρτηση T των $x1, x2, .., xn$ δηλαδή $T = t(X) = t(x1, x2, \ldots, xn)$. Η από κοινού σ.π.π. των $x1, x2, \ldots, xn$ ονομάζεται συνάρτηση πιθανοφάνειας (likelihood) και ορίζεται ως:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)\ ,\theta\ \epsilon\ \Theta$$

Είναι δηλαδή μία συνάρτηση του θ και ορίζει την πιθανότητα που έχει το γεγονός το δείγμα να προέρχεται από την υποτιθέμενη κατανομή με παράμετρο θ. Η εκτιμήτρια του $\theta = (\theta_1, \theta_2, \ldots, \theta_\kappa)$ καλείται εκτιμήτρια μέγιστης πιθανοφάνειας(Ε.Μ.Π) της παραμέτρου θ αν μεγιστοποιεί την πιθανοφάνεια L(θ),δηλαδή $\hat{\theta} = max_{\theta\epsilon\Theta} L(\theta)$.Μεγιστοποιώντας δηλαδή την λογαριθμημένη συνάρτηση πιθανοφάνειας *L*(θ) ως προς θ θα προκύψουν οι εκτιμήτριες μέγιστης πιθανοφάνειας της θ δηλαδή οι $\widehat{\theta_1}, \widehat{\theta_2}, \ldots, \widehat{\theta_\kappa}$.

Θεωρείται πιο εύκολο να αναζητηθεί το σημείο μεγίστου της $l(\theta) = logL(\theta)$ αντί της L(θ) δεδομένου ότι και οι δύο έχουν το ίδιο σημείο μεγίστου καθώς η λογαριθμική συνάρτηση είναι αύξουσα. Η Ε.Μ.Π. του θ λαμβάνεται από τη λύση του συστήματος n-εξισώσεων:

$$\frac{\partial l(\theta)}{\partial \theta_j} = 0\ \gamma\iota\alpha\ \kappa\acute{\alpha}\theta\varepsilon\ j = 1, \ldots, \kappa$$

Είναι ασυμπτωτικά αμερόληπτη για το θ δηλαδή ακολουθεί ασυμπτωτικά πολυδιάστατη κανονική κατανομή με διάνυσμα μέσων τιμών θ. (Κούτρας-Μπούτσικας,2012)

Παρακάτω ακολουθούν δύο παραδείγματα σχετικά με την εύρεση την εκτιμήτριας της μέγιστης πιθανοφάνειας.

Παράδειγμα1:

Έστω ένα τυχαίο δείγμα μεγέθους n όπου οι $t_1, t_2, \dots, t_n$ ακολουθούν την εκθετική κατανομή με σ.π.π $f(t) = \lambda e^{-\lambda t}\ t > 0, \lambda > 0$

Έτσι η συνάρτηση πιθανοφάνειας ορίζεται ως:

$$L(\lambda|t_1, t_2, \dots, t_n) = \prod_{i=1}^{n} f(t_i) = \prod_{i=1}^{n} (\lambda e^{-\lambda t_i}) = \lambda^n e^{-\lambda \sum_{i=1}^{n} t_i}$$

και αφού λογαριθμήσουμε την παραπάνω σχέση θα έχουμε:

$$l = lnL = nln\lambda - \lambda \sum_{i=1}^{n} t_i$$

Για να εκτιμήσουμε την παράμετρο λ θα μεγιστοποιήσουμε τη συνάρτηση $l(\lambda)$ως προς λ,δηλαδή θέτουμε: $\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i = 0$

Επομένως $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i} = \frac{1}{\bar{t}}$ και παρατηρούμε ότι $\frac{\partial^2 l}{\partial \lambda^2} = -\frac{n}{\lambda^2} < 0$ το οποίο είναι αναμενόμενο καθώς η εκτιμήτρια $\hat{\lambda}$ αντιστοιχεί στη μεγιστοποίηση της lnL και η αρνητική τιμή της δευτέρου παραγώγου $-\frac{\partial^2 l}{\partial \lambda^2} = \frac{n}{\hat{\lambda}^2} = n\bar{t}^2$καλείται παρατηρούμενη πληροφορία (observed information). (Χ.Καρώνη,2009)

Παράδειγμα2:

Έστω τυχαίο δείγμα μεγέθους n με τ.μ.$x_1, x_2, \dots, x_n$ οι οποίες ακολουθούν διωνυμική κατανομή, η οποία εκφράζει το πλήθος των επιτυχιών στις n επαναλήψεις Bernoulli με τιμές

$x = (0,1, \dots, N)$ και συνάρτηση μάζας πιθανότητας

$$f(x; \theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x}\ x = 0,1, \dots, n, 0 < \theta < 1$$

Η συνάρτηση πιθανοφάνειας γίνεται

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} \binom{N}{x_i} \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{nN - \sum_{i=1}^{n} x_i}$$

Και λογαριθμώντας την παραπάνω σχέση έχουμε:

$$l(\theta) = lnL(\theta) = ln\prod_{i=1}^{n}\binom{N}{x_i} + \sum_{i=1}^{n} x_i ln\theta + (nN - \sum_{i=1}^{n} x_i) ln(1-\theta)$$

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{nN - \sum_{i=1}^{n} x_i}{1-\theta} = 0 \Leftrightarrow \frac{1-\theta}{\theta} = \frac{nN - \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i} \Leftrightarrow \theta = \frac{\bar{x}}{N}$$

και καθώς η εκτιμήτρια $\hat{\theta}$ αντιστοιχεί στη μεγιστοποίηση της lnL θα πρέπει η δεύτερη παράγωγος να είναι αρνητική.

$$\frac{\partial^2 l(\theta)}{\partial \theta} = -\frac{\sum_{i=1}^{n} x_i}{\theta^2} - \frac{nN - \sum_{i=1}^{n} x_i}{(1-\theta)^2} = -\frac{nN}{\bar{x}} - \frac{n(N-\bar{x})}{\left(1-\frac{\bar{x}}{N}\right)^2} < 0$$

Άρα η Ε.Μ.Π της παραμέτρου θ είναι $\widehat{\theta} = \frac{\bar{X}}{N}$.(Φουσκάκης, pdf)

1.3.3 Μέθοδος Ροπών

Η λεγόμενη μέθοδος των ροπών (method of moments) είναι από τις παλιότερες μεθόδους εκτίμησης και δίκαια έχει πάρει τη συγκεκριμένη ονομασία καθώς χρησιμοποιεί τις ροπές προκειμένου να γίνει εκτίμηση των παραμέτρων. Στη στατιστική λοιπόν η ποσότητα $E(X^k)$ ονομάζεται απλή ροπή k-τάξης της X όπου k ένας θετικός αριθμός και X μία τυχαία μεταβλητή και συμβολίζεται ως μ'_k.Έτσι έχουμε

$$\mu'_k = E(X^k) = \begin{cases} \sum_x x^k P(X = x) \ , \alpha\nu \ X \ \delta\iota\alpha\kappa\rho\iota\tau\acute{\eta} \\ \int_{-\infty}^{+\infty} x^k f(x)dx \ , \alpha\nu \ X \ \sigma\upsilon\nu\varepsilon\chi\acute{\eta}\varsigma \end{cases}$$

Ενώ για ένα τυχαίο δείγμα $X_1, X_2, .., ., X_n$ορίζουμε την δειγματική ροπή τάξης-k ως:

$$m'_k = \frac{\sum_{i=1}^{n} x_i^k}{n} \quad k = 1,2 \ldots$$

Για να εφαρμόσουμε λοιπόν τη μέθοδο των ροπών σε ένα τυχαίο δείγμα$(X_1, X_2, \ldots, X_n)$ από ένα πληθυσμό X με συνάρτηση πυκνότητας πιθανότητας ή μάζας πιθανότητας$f(x;\theta)$ $\kappa\alpha\iota$ $\theta = (\theta_1, \theta_2, \ldots, \theta_m) \in \Theta$ θα εξισώσουμε τις πρώτες δειγματικές ροπές με τις αντίστοιχες ροπές του πληθυσμού και θα πάρουμε ένα σύστημα m εξισώσεων όπου m είναι ο αριθμός των παραμέτρων που επιθυμούμε να εκτιμήσουμε. Έτσι :

$$m'_k = \mu'_k$$

και θέτοντας π.χ. στη συνεχή περίπτωση ως

$$\mu'_k = E(X^k) = \int x^k f(x;\theta) = g_k(\theta) \ k = 1, \ldots, m$$

προκύπτει το σύστημα:

$$g_k(\theta) = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Λύνοντας το σύστημα υπολογίζουμε τις εκτιμήτριες $\widehat{\theta_1}, \widehat{\theta_2}, \dots, \widehat{\theta_m}$. (Ι. Παναρέτου & Ε.Ξεκαλάκη,2000)

1.3.4 Στατιστικός έλεγχος υποθέσεων

Μέχρι τώρα αναφερθήκαμε στον τρόπο με τον οποίο μπορούμε να εκτιμήσουμε την τιμή μιας παραμέτρου. Πολλές φορές όμως καλούμαστε να συγκρίνουμε την τιμή της παραμέτρου σε σχέση με κάποια άλλη δεδομένη τιμή η οποία έχει φυσική σημασία για το πρόβλημα που εξετάζουμε. Η Στατιστική συμπερασματολογία μας προσφέρει μία συμπερασματική μέθοδο, το λεγόμενο στατιστικό έλεγχο υποθέσεων (hypothesis testing) ο οποίος μας επιτρέπει να ελέγξουμε εάν η τιμή της παραμέτρου που εξετάζουμε είναι μικρότερη ή μεγαλύτερη από μία δεδομένη τιμή.

Έστω θ λοιπόν η παράμετρος που θέλουμε να εκτιμήσουμε τότε μπορούμε να έχουμε τους εξής ελέγχους: $H_0: \theta = \theta_0$ $(\mu\eta\delta\varepsilon\nu\iota\kappa\acute{\eta}\ \upsilon\pi\acute{o}\theta\varepsilon\sigma\eta)$

$$H_1: \theta \neq \theta_0\ (\varepsilon\nu\alpha\lambda\lambda\alpha\kappa\tau\iota\kappa\acute{\eta}\ \upsilon\pi\acute{o}\theta\varepsilon\sigma\eta)$$

Ο παραπάνω έλεγχος ονομάζεται αμφίπλευρος ενώ οι παρακάτω έλεγχοι ονομάζονται μονόπλευροι και ορίζονται ως:

$$H_0: \theta = \theta_0 \qquad H_0: \theta = \theta_0$$

$$H_1: \theta > \theta_0 \quad \text{ή} \quad H_1: \theta < \theta_0$$

οι οποίοι ονομάζονται μονόπλευροι έλεγχοι. Η επιλογή του μονόπλευρου ή του αμφίπλευρου ελέγχου εξαρτάται από το πρόβλημα το οποίο θέλουμε να μελετήσουμε. Αφού λοιπόν έχουμε ορίσει τη μηδενική και την εναλλακτική υπόθεση, υπολογίζουμε τώρα την ελεγχοσυνάρτηση για τα δεδομένα μας το οποίο είναι μία δειγματοσυνάρτηση που ορίζεται ως:

$$\frac{[(\varepsilon\kappa\tau\iota\mu\acute{\eta}\tau\rho\iota\alpha\ \tau\sigma\upsilon\ \theta) - (\eta\ \tau\iota\mu\acute{\eta}\ \tau\sigma\upsilon\ \theta\ \pi\sigma\upsilon\ \pi\rho\sigma\acute{\varepsilon}\kappa\upsilon\psi\varepsilon\ \alpha\pi\acute{o}\ \tau\sigma\nu\ \acute{\varepsilon}\lambda\varepsilon\gamma\chi\sigma\ H_0)]}{se(\widehat{\theta})}$$

Δηλαδή κατασκευάζουμε μία κατάλληλη στατιστική συνάρτηση $T = T(X_1, X_2, \dots, X_n)$ ώστε να υπολογίσουμε αυτό που συμβαίνει στο δείγμα υπό τη μηδενική υπόθεση H_0. Έπειτα ορίζουμε την κρίσιμη περιοχή η οποία αποτελείται από τις τιμές της ελεγχοσυνάρτησης για τις οποίες απορρίπτουμε τον H_0 επηρεαζόμενη από ένα επίπεδο σημαντικότητας α (συνήθως 0.05 ή 0.01) το οποίο ταυτίζεται με την πιθανότητα σφάλματος τύπου Ι. Το επίπεδο σημαντικότητας ή αλλιώς

παρατηρούμενη στάθμη ορίζεται ως η πιθανότητα η τιμή της ελεγχοσυνάρτησης που παρατηρείται να είναι μεγαλύτερη (ή γενικότερα πιο extreme) από την τιμή που προέκυψε από το δείγμα (Φ.Κολυβά, Ε.Μπόρα, 1995). Γενικότερα ο παραμετρικός χώρος αποτελείται από την κρίσιμη περιοχή που αναφέραμε παραπάνω και από την περιοχή αποδοχής η οποία αποτελείται από τις τιμές της ελεγχοσυνάρτησης για τις οποίες δεν απορρίπτουμε τον H_0. Γενικότερα αν η τιμή της ελεγχοσυνάρτησης βρίσκεται μέσα στην κρίσιμη περιοχή τότε απορρίπτουμε την H_0 αλλιώς δεν απορρίπτουμε την H_0 σε επίπεδο σημαντικότητας α.

Συνοψίζοντας:

	H_0 σωστή	H_1 σωστή
Απορρίπτω H_0	$Σφάλμα\ τύπου\ I$ $πιθανότητα(\alpha)$	Ορθή απόφαση
Δεν απορρίπτω H_0	Ορθή απόφαση	$Σφάλμα\ τύπου\ II$ $πιθανότητα(\beta)$

Διάγραμμα 1.3: Πίνακας Έλεγχος Υποθέσεων

Πηγή: Φουσκάκης, ΣΕΜΦΕ, Στατιστική Συμπερασματολογία, pdf

Όπου :

- $\alpha = P(σφάλμα\ τύπου\ I) = P(απορρίπτω\ H_0 | H_0\ σωστή)$
- $\beta = P(σφάλμα\ τύπου\ II) = P(δεν\ απορρίπτω\ H_0 | H_1\ σωστή)$

Η πιθανότητα $P(απορρίπτω\ H_0 | H_1\ σωστή)$ ονομάζεται ισχύς του ελέγχου. Όταν λοιπόν απορρίπτεται η H_0, το δείγμα χαρακτηρίζεται στατιστικά σημαντικό (statistically significant) δηλαδή διαφέρει σημαντικά από αυτό που αναμενόταν υπό την H_0.(Φουσκάκης pdf σημειώσεις)

<u>Παράδειγμα:</u>

Έστω X τ.μ. η οποία υπολογίζει τη δύναμη θραύσης μιας ατσάλινης ράβδου. Εάν η ατσάλινη ράβδος παραχθεί με τη μέθοδο I τότε X ~ N(50,36). Μια καινούρια μέθοδος II που είναι υπό δοκιμή δίνει X~N(55,36). Εάν δοθούν 16 ατσάλινες ράβδοι που φτιάχτηκαν με τη μέθοδο II, πώς θα μπορούσε να ελεγχθεί εάν η αύξηση της δύναμης θραύσης είναι πραγματική;

Λύση:

Σύμφωνα με τις μεθόδους I,II ορίζω τις εξής υποθέσεις:

$H_0: \mu = 50 (μηδενική\ υπόθεση)$

ή

$H_1: \mu = 55 (εναλλακτική\ υπόθεση)$

Από τυχαίο δείγμα 16 ατσάλινων ράβδων, έστω ότι η δειγματική μέση τιμή προέκυψε 53.

Δεδομένου ότι $\bar{X} \sim N\left(50, \frac{36}{16}\right)$ $όταν\ \mu = 50$

Και $\bar{X} \sim N\left(55, \frac{36}{16}\right)$ $όταν\ \mu = 55$

Υπολογίζω την ελεγχοσυνάρτηση : $Z=\frac{\bar{X}-\mu}{SE(\bar{X})} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$

Τότε το σφάλμα τύπου I είναι:

$$\alpha = P(απορρίπτω\ H_0: H_0\ σωστή) = P(απορρίπτω\ H_0 | \mu = 50)$$

$$= P(\bar{X} > 53 | \mu = 50) = P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} > \frac{53-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(\frac{\bar{X}-50}{\frac{6}{4}} > \frac{53-50}{\frac{6}{4}}\right) = 0.0228$$

Και το σφάλμα τύπου II είναι :

$$\alpha = P(απορρίπτω\ H_1: H_1\ σωστή) = P(απορρίπτω\ H_1 | \mu = 55)$$

$$= P(\bar{X} \leq 53 | \mu = 55) = P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{53-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(\frac{\bar{X}-55}{\frac{6}{4}} \leq \frac{53-55}{\frac{6}{4}}\right) = 0.0913$$

Συνήθως όταν η P τιμή ελέγχου είναι μικρότερη από το επίπεδο σημαντικότητας α τότε απορρίπτουμε H_0.Για $\alpha = 0.01\ παρατηρώ\ ότι\ 0.0228 > 0.01\ επομένως\ δεν\ απορρίπτω\ H_0$

$ενώ\ για\ \alpha = 0.05\ παρατηρώ\ ότι\ 0.0228 < 0.05\ επομένως\ απορρίπτω\ H_0\ σε\ \alpha = 5\%$

1.4 Μέτρα Καταλληλόλητας

Ένα κοινό πρόβλημα σε εφαρμογές μοντέλων πολλαπλής γραμμικής παλινδρόμησης είναι η σωστή επιλογή ανεξάρτητων μεταβλητών ή επιλογή του βέλτιστου μοντέλου. Στόχος είναι η επιλογή ενός μικρότερου υποσυνόλου μεταβλητών από ένα πλήθος που είναι

διαθέσιμες ώστε να προκύψει ένα φειδωλό μοντέλο. Υπάρχουν διάφορες προσεγγίσεις για την επιλογή ενός τέτοιου υποσυνόλου όπως είναι η μέθοδος προσθήκης μεταβλητών, η μέθοδος αφαίρεσης και η βηματική παλινδρόμηση που αναλύονται στο 4ο κεφάλαιο (Lance,2005). Για την καλύτερη επιλογή ενός μοντέλου χρήσιμα εργαλεία αποτελούν τα μέτρα καταλληλόλητας όπως είναι ο συντελεστής προσδιορισμού R^2, το κριτήριο $C_p - Mallows$ και τα κριτήρια πληροφορίας AIC και BIC, τα οποία θα αναλυθούν στα επόμενα κεφάλαια.

1.4.1 Συντελεστής συσχέτισης – Συντελεστής Προσδιορισμού

Η ανάλυση παλινδρόμησης μελετά την εξάρτηση μεταξύ μεταβλητών. Η γραμμική συσχέτιση μεταξύ δύο μεταβλητών X,Y προσδιορίζεται από τον συντελεστή συσχέτισης(correlation coefficient) ο οποίος σ' ένα δείγμα δεδομένων ορίζεται ως η συνδιακύμανση των μεταβλητών διαιρεμένη με τις τυπικές αποκλίσεις των X,Y. Η συνδιακύμανση 2 μεταβλητών διακρίνεται σε:

- Συνδιακύμανση δείγματος (sample covariance) η οποία ορίζεται ως:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Συνδιακύμανση πληθυσμού(population covariance) η οποία ορίζεται ως:

  $\sigma_{xy} = E\,(x_i - \mu_x)\big(y_i - \mu_y\big)$ όπου μ_x, μ_y είναι οι πληθυσμιακές μέσες τιμές των μεταβλητών X,Y.

Η θετική συνδιακύμανση υποδεικνύει μία θετική γραμμική σχέση μεταξύ των μεταβλητών και η αρνητική συσχέτιση δείχνει αντίστοιχα το αντίθετο. Έτσι ο συντελεστής συσχέτισης διακρίνεται σε:

- Συντελεστή συσχέτισης πληθυσμού ρ (population correlation coefficient) και ορίζεται ως:

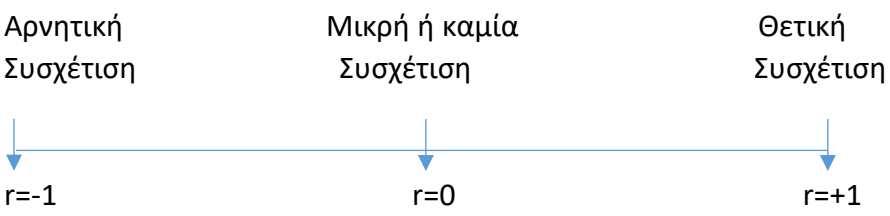
$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

➢ Συντελεστή συσχέτισης δείγματος r Pearson, sample correlation coefficient) ο οποίος είναι η σημειακή εκτίμηση του συντελεστή συσχέτισης και ορίζεται ως:

$\widehat{\rho_{xy}} = r_{xy} = \frac{s_{xy}}{s_x s_y}$, (R Tutorial book by Chi Yau,2014)

Ο συντελεστής συσχέτισης λοιπόν μπορεί να θεωρηθεί ως ένα μέτρο που ελέγχει πως μεταβάλλεται μια τ.μ. ως προς μία άλλη. Ενώ η τιμή της συνδιασποράς σ_{xy} παίρνει τιμές που εξαρτώνται από το πεδίο τιμών των X,Y, ο συντελεστής συσχέτισης ρ_{xy} παίρνει τιμές στο διάστημα [-1,1]. Οι τιμές του συντελεστή συσχέτισης ερμηνεύονται ως εξής:

- Αν ρ=1 τότε υπάρχει τέλεια θετική γραμμική σχέση ανάμεσα στις μεταβλητές X,Y.
- Αν ρ=0 τότε οι μεταβλητές X,Yδεν συνδέονται γραμμικά, δηλαδή είναι ασυσχέτιστες. Αξιοσημείωτο είναι ότι μπορεί οι X,Y να μην είναι ανεξάρτητες αλλά να συνδέονται μη-γραμμικά μεταξύ τους.
- Αν ρ= -1 τότε υπάρχει τέλεια αρνητική σχέση ανάμεσα στις μεταβλητές X,Y.



Διάγραμμα 1.4: Διάγραμμα για το συντελεστή συσχέτισης r

Όσο πιο κοντά στη μονάδα είναι η τιμή του συντελεστή συσχέτισης τόσο πιο ισχυρή είναι η σχέση μεταξύ των μεταβλητών. Εάν θέλουμε να εκφράσουμε τη σχέση μεταξύ των μεταβλητών X,Y σε ποσοστό τότε χρησιμοποιούμε το συντελεστή προσδιορισμού r^2 ή R^2(coefficient of determination) ο οποίος εκφράζει το ποσοστό μεταβολής της μεταβλητής Y σε σχέση με το ποσό μεταβολής της μεταβλητής X. Συγκεκριμένα ο συντελεστής προσδιορισμού R^2 είναι ο λόγος της διακύμανσης των εκτιμημένων τιμών της εξαρτημένης μεταβλητής προς τη διακύμανση των πραγματικών τιμών της εξαρτημένης μεταβλητής και υπολογίζεται ως εξής:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST},$$

Όπου: SSE: άθροισμα τετραγώνων των υπολοίπων
SST: συνολικό άθροισμα τετραγώνων
SSR: άθροισμα τετραγώνων των συνιστωσών παλινδρόμησης

Οι τιμές του συντελεστή προσδιορισμού R^2 κυμαίνονται στο διάστημα [0,1] και προφανώς όσο η τιμή πλησιάζει προς το 1 τόσο καλύτερη προσαρμογή έχει το μοντέλο. Η ερμηνεία των παραπάνω ορίων έχει ως εξής:

- $R^2 = 1$ οι επεξηγηματικές μεταβλητές εξηγούν το 100% της διακύμανσης της εξαρτημένης μεταβλητής και άρα έχουμε ένα τέλειο μοντέλο.
- $R^2 = 0$ δεν υπάρχει καμία συσχέτιση μεταξύ της εξαρτημένης μεταβλητής και των επεξηγηματικών μεταβλητών.

ο οποίος εκφράζει το ποσοστό μεταβολής της μεταβλητής Y σε σχέση με το ποσό μεταβολής της μεταβλητής X.
Ένας άλλος συντελεστής συσχέτισης ο οποίος είναι εύκολο να υπολογιστεί χωρίς τη βοήθεια υπολογιστή είναι ο συντελεστής Spearman's rho ο οποίος συμβολίζεται ως r_s,όπου το r δηλώνει το συντελεστή συσχέτισης και ο χαρακτήρας s προέρχεται από το στατιστικό Spearman. Στην πραγματικότητα ο συντελεστής συσχέτισης Spearman ορίζεται ως ρ_s ενώ r_s θεωρείται ο εκτιμημένος συντελεστής συσχέτισης βάσει του δείγματος και υπολογίζεται από τον τύπο:

$$r_s = 1 - \frac{6\sum D_i^2}{n(n^2-1)}$$

όπου D_i θεωρείται η διαφορά των τάξεων μεταξύ των τιμών των συμμεταβλητών X_i, Y_i και n το μέγεθος του δείγματος. Όπως και ο συντελεστής συσχέτισης Pearson έτσι και ο συντελεστής Spearman παίρνει τιμές στο διάστημα [-1,1].(Paul Gingrich, 2004)

1.4.2 Το κριτήριο C_p του Mallows

Ένα κοινό πρόβλημα σε εφαρμογές μοντέλων πολλαπλής γραμμικής παλινδρόμησης είναι η σωστή επιλογή ανεξάρτητων μεταβλητών ή επιλογή του βέλτιστου μοντέλου. Στόχος είναι η επιλογή ενός μικρότερου υποσυνόλου μεταβλητών από ένα πλήθος που είναι διαθέσιμες ώστε να προκύψει ένα φειδωλό μοντέλο. Υπάρχουν διάφορες προσεγγίσεις για την επιλογή ενός τέτοιου υποσυνόλου όπως είναι η μέθοδος προσθήκης μεταβλητών, η μέθοδος αφαίρεσης και η βηματική παλινδρόμηση που αναλύονται στο 4ο κεφάλαιο (Lance,2005).
Όσον αφορά την επιλογή του βέλτιστου μοντέλου χρησιμοποιούνται διάφορα κριτήρια πληροφορίας όπως είναι το AIC ή το BIC καθώς και το κριτήριο C_p του Mallows. Το κριτήριο αυτό

βοηθά στον καθορισμό ενός "optimal" υποσυνόλου ανεξαρτήτων μεταβλητών και είχε προταθεί από τον Mallows (1973). Συνδέεται άμεσα με τον συντελεστή προσδιορισμού και αποσκοπεί στο να ισορροπήσει το πλήθος των αριθμών μεταβλητών που περιλαμβάνονται στο μοντέλο.

Η στατιστική συνάρτηση $C_p - Mallows$ είναι άλλο ένα μέτρο αξιολόγησης της καταλληλόλητας του μοντέλου και ορίζεται ως εξής:

$$C_p = \frac{SSE(k)}{s^2} - n + 2p$$

Ορίζοντας ως p τον αριθμό των επεξηγηματικών μεταβλητών στο μοντέλο και n τον αριθμό των παρατηρήσεων. Ο όρος $s^2 = MSE$ είναι το μέσο τετραγωνικό σφάλμα για ολόκληρο το μοντέλο (δηλαδή για το μοντέλο που περιέχει όλες τις p επεξηγηματικές μεταβλητές και ο όρος $SSE(k)$ ορίζεται ως το άθροισμα τετραγώνων των υπολοίπων για το μοντέλο που περιλαμβάνει k μεταβλητές από το σύνολο p επεξηγηματικών μεταβλητών). Αφού υπολογιστεί η τιμή του C_p για όλα τα μοντέλα με τους δυνατούς συνδυασμούς των υποψήφιων μεταβλητών τότε ο προσδιορισμός του καταλληλότερου μοντέλου μέσω του κριτηρίου αυτού γίνεται μέσω της γραφικής παράστασης του συνόλου των επεξηγηματικών μεταβλητών ως προς την τιμή C_p. Αξιοσημείωτο είναι ότι όταν το σωστό μοντέλο έχει p παραμέτρους τότε $E[C_p] = p$, το οποίο είναι αναμενόμενο καθώς αν το p μοντέλο είναι σωστό τότε η τιμή του C_p είναι αρκετά κοντά στην τιμή p. Δηλαδή το σημείο στο οποίο αντιστοιχεί η μικρότερη τιμή του C_p θεωρείται και το καταλληλότερο μοντέλο.

Κεφάλαιο 2ο

Περιγραφή του Akaike κριτηρίου πληροφορίας (AIC)

2.1 Εισαγωγή

Τα τελευταία χρόνια η στατιστική δίνει όλο και περισσότερη έμφαση στα κριτήρια αξιολόγησης μοντέλου μέσα από τα οποία επιτυγχάνεται η επιλογή του καλύτερου μοντέλου ανάμεσα σε ένα πλήθος πολύπλοκων μοντέλων (Bozdogan-Akaike, 2000). Η επιλογή ενός μοντέλου με λίγες παραμέτρους μπορεί να οδηγήσει σε μη ρεαλιστικές εικασίες και σε μεγάλο συστηματικό σφάλμα με αποτέλεσμα η πρόβλεψή μας να γίνει ελλιπής. Τέτοιου είδος μοντέλα δεν είναι ικανά να περιγράψουν το δείγμα μας καθώς επίσης κι ολόκληρο τον πληθυσμό. Γι' αυτό το λόγο χρησιμοποιούνται τα κριτήρια επιλογής μοντέλου όπως είναι το λεγόμενο AIC, το οποίο αναπτύχθηκε πρώτη φορά από τον Akaike(1973) ως ένας τρόπος σύγκρισης διαφορετικών μοντέλων. Το μοντέλο είναι ένα εργαλείο το οποίο προσπαθεί να αποκαλύψει τη σχέση μεταξύ του δοθέντος αποτελέσματος και μιας συγκεκριμένης μεταβλητής**.** Παρ' όλα αυτά διαφορετικά κριτήρια είναι πιθανόν να οδηγήσουν σε διαφορετικά αποτελέσματα το οποίο μας εμποδίζει να διακρίνουμε την αξιοπιστία του κάθε μοντέλου. Η 'πραγματικότητα' δεν μπορεί να συμπεριληφθεί σε ένα μόνο μοντέλο. Αναζητούμε λοιπόν ένα καλό μοντέλο με σκοπό να προσεγγίσουμε τα εμπειρικά μας δεδομένα (Keneth, Anderson,1994). Απαραίτητο συστατικό για να προσεγγίσουμε θεωρητικά το κριτήριο πληροφορίας AIC θεωρείται η απόσταση Kullback-Leibler (1951) η οποία θεωρείται βασικό κριτήριο προκειμένου να εκτιμήσουμε το πόσο κατάλληλα προσεγγίζει ένα μοντέλο την πραγματική κατανομή δεδομένων που έχουμε. Το AIC λοιπόν προέρχεται από την ασυμπτωτική εκτίμηση της K-L απόκλισης και παρέχει ένα χρήσιμο εργαλείο για την επιλογή μοντέλων τα οποία εκτιμώνται μέσω της μεθόδου της μέγιστης πιθανοφάνειας. Με βάση το κριτήριο K-L περιγράφεται η απόδειξη του κριτηρίου πληροφορίας του Akaike που θα περιγράψουμε παρακάτω αφού ορίσουμε αναλυτικά το κριτήριο πληροφορίας Kullback-Leibler.

2.2 Απόσταση κατά Kullback–Leibler (K-L distance)

Όπως αναφέραμε παραπάνω με τη βοήθεια του κριτηρίου AIC προσπαθούμε να εντοπίσουμε ποιο από τα υποψήφια μοντέλα προσεγγίζει καλύτερα την πραγματική κατανομή ή αλλιώς ποιο μοντέλο ελαχιστοποιεί την απόσταση μεταξύ της προσαρμοσμένης και της

πραγματικής κατανομής. Ένα εργαλείο μέτρησης αυτής της απόστασης είναι η ποσότητα πληροφορίας κατά Kullback-Leibler. Το 1951 ο S.Kullback και ο R.A. Leibler δημοσίευσαν ένα πλέον αναγνωρίσιμο άρθρο (Kullback and Leibler 1951) το οποίο προσδιορίζει την έννοια της "πληροφορίας" σύμφωνα με την αντίληψη του R.A. Fisher. Το συμπέρασμα αυτής της έρευνας γνωστό ως K-L πληροφορία αποτελεί ένα θεμελιώδες κομμάτι της επιστήμης κι έχει τις ρίζες του στην ιδέα της εντροπίας του Boltzmann (Burnham and Anderson,2004).

2.2.1 Μοντέλο συνεχούς κατανομής

Υπάρχουν αρκετοί τρόποι για να μελετήσουμε την παραμετρική προσέγγιση της σωστής κατανομής f ενός μοντέλου από μία άλλη κατανομή g αλλά η απόσταση η οποία συνδέεται άμεσα με την μέθοδο μέγιστης πιθανοφάνειας είναι η λεγόμενη Kullback-Leibler distance και συμβολίζεται με 'I'. Η πληροφορία Kullback-Leibler (K-L) θεωρείται ουσιαστικά η πληροφορία που χάνεται όταν το μοντέλο προσπαθεί να προσεγγίσει την πραγματική κατανομή ή εναλλακτικά είναι η «απόσταση» του μοντέλου που έχουμε υποθέσει από το πραγματικό. Θέτουμε $x_n = \{x_1, x_2, \dots, x_n\}$ ένα σύνολο από n ανεξάρτητες παρατηρήσεις οι οποίες επιλέγονται τυχαία από μια άγνωστη συνάρτηση κατανομής πιθανότητας F(x). Η F(x) ουσιαστικά αποτελεί το πραγματικό μοντέλο ή την πραγματική κατανομή ενώ θεωρούμε ως G(x) ένα αυθαίρετα ορισμένο μοντέλο. Ας θεωρήσουμε ότι έχουμε ένα μοντέλο συνεχούς κατανομής όπου f είναι η συνάρτηση πυκνότητας πιθανότητας της F(x) που αντιστοιχεί στο πραγματικό μοντέλο η οποία δεν εξαρτάται από παραμέτρους και g είναι είναι η συνάρτηση πυκνότητας πιθανότητας της G(x) η οποία προσεγγίζει το μοντέλο που έχουμε υποθέσει ως πραγματικό. Οι Cover και ο Thomas(1991) υποστήριξαν ότι "η απόσταση Kullback-Leibler είναι ένα μέτρο ανεπάρκειας που υποστηρίζει ότι η κατανομή μας είναι η g όταν η πραγματική κατανομή είναι η f".

Η K-L λοιπόν ορίζεται από τον παρακάτω τύπο:

$$I(F;G) = E_F[\log\{\frac{F(x)}{G(x)}\}] \qquad (2.1)$$

όπου το E_F αναπαριστά την αναμενόμενη τιμή ως προς τη συνάρτηση κατανομής F. Στην περίπτωση που οι συναρτήσεις κατανομής πιθανότητας είναι συνεχείς τότε ο παραπάνω τύπος παίρνει την εξής μορφή:

$$I(f;g) = \int f(x) log \frac{f(x)}{g(x|\theta)} dx \qquad (2.2)$$

Η $I(f;g)$ έχει τις εξής ιδιότητες:

$(i)\ I(f;g) \geq 0$

$(ii)\ I(f;g) = 0 \Leftrightarrow f(x) = g(x)$

Το κριτήριο πληροφορίας $I(f;g)$ δεν μπορεί να χρησιμοποιηθεί ευθέως στην επιλογή μοντέλου καθώς απαιτείται γνώση για το πραγματικό μοντέλο και για τις παραμέτρους θ στο προσεγγιστικό μοντέλο. Στην ανάλυση δεδομένων οι παράμετροι των μοντέλων πρέπει να εκτιμώνται και υπάρχει συχνά αβεβαιότητα για αυτή την εκτίμηση. Τέλος καλό θα είναι να γίνεται ελαχιστοποίηση της αναμενόμενης K-L πληροφορίας παρά της πραγματικής ανάμεσα στο πλήθος R υποψήφιων μοντέλων.

Η K-L πληροφορία μπορεί να εκφραστεί ως:

$$I(f;g) = \int (f(x)logf(x))\,dx - \int (f(x)logg(x|\theta))dx$$

ή

$$I(f;g) = E_f\left[\log\left(f(x)\right)\right] - E_f[\log(g(x|\theta)] \qquad (2.3)$$

Θεωρώντας την $E_f\left[\log\left(f(x)\right)\right]$ ως σταθερα C, διότι δεν εξαρτάται από το θ, η παραπάνω σχέση γίνεται:

$$\boldsymbol{I(f;g) = C - E_f[\log(g(x|\theta))]} \qquad (2.4)$$

όπου το $I = \int (f(x)\log f(x))\,dx$ δεν εξαρτάται από τα δεδομένα μας ή το μοντέλο μας ενώ μόνο η $E_f[\log(g(x|\theta)]$ είναι απαραίτητο να εκτιμηθεί για κάθε μοντέλο. (Burnham-Anderson,2004)

Όπως αναφέραμε στην ανάλυση δεδομένων οι παράμετροι των μοντέλων πρέπει να εκτιμηθούν και φυσικά υπάρχει αβεβαιότητα σε αυτή τη εκτίμηση. Η διαφορά της χρήσης της εκτιμημένης παραμέτρου $\hat{\theta}$ έναντι της παραμέτρου θ είναι αρκετά σημαντική και οδηγεί στην ελαχιστοποίηση της προσδοκώμενης K-L απόστασης έναντι της πραγματικής. Δηλαδή έχουμε:

$$\hat{I}(f;g) = \int f(x)\log(\frac{f(x)}{g(x|\hat{\theta}(y))})dx$$

Αξιοσημείωτο είναι ότι η μεταβλητή x του ολοκληρώματος δεν αναπαριστά τα δεδομένα τα

οποία όμως προσδιορίζονται από την μεταβλητή y. Η μέση τιμή της K-L απόστασης ως προς την εκτιμημένη παράμετρο $\hat{\theta}$ αναμένεται να είναι:

$$E_{\hat{\theta}}[\hat{l}(f;g)] = \int f(y)[\int f(x)\log(\frac{f(x)}{g(x|\hat{\theta}(y))})dx]dy$$

Λόγω της δομής του μοντέλου g ισχύει:

$$E_{\hat{\theta}}[\hat{l}(f;g)] > I(f,g)$$

Από τον παραπάνω τύπο έχουμε:

$$\hat{I}(f;g) = \int f(x)\log f(x)\,dx - \int f(x)\log(\,g(x|\hat{\theta}(y)))dx =$$

$$= constant - E_x[\log(g\left(x\middle|\hat{\theta}(y)\right)]$$

Επιπλέον,

$$E_{\hat{\theta}}[\hat{I}(f;g)] = \int f(y)\left[\int f(x)\log f(x)\,dx\right]dy - \int f(y)\ [\int f(x)\log\left(g\left(x\middle|\hat{\theta}(y)\right)dx\right]dy =$$

$$= [\int f(y)dy][\int f(x)\log f(x)dx] - \int f(y)\,[\int \mathrm{f(x)}\log g\left(x\middle|\hat{\theta}(y)\right)dx]dy \Leftrightarrow$$

$$E_{\hat{\theta}}[\hat{I}(f;g)] = Constant - E_yE_x[\log(g\left(x\middle|\hat{\theta}(y)\right)] \qquad (2.5)$$

όπου $E_yE_x[\log(g\left(x\middle|\hat{\theta}(y)\right)]$ θεωρείται η σχετική προσδοκώμενη K-L απόσταση. Σκοπός μας είναι να καθορίσουμε μία μέθοδο προκειμένου να επιλέξουμε το κατάλληλο μοντέλο g_i η οποία να ελαχιστοποιεί μεταξύ του πλήθους των μοντέλων $g_1, \dots, g_R$ την K-L απόσταση. Το μοντέλο $g_i(x|\theta)$ το οποίο μεγιστοποιεί την $E_yE_x[\log(g\left(x\middle|\hat{\theta}(y)\right)]$ θεωρείται το καλύτερο K-L μοντέλο όταν η παράμετρος θ πρέπει να εκτιμηθεί και η θεωρία μέγιστης πιθανοφάνειας χρησιμοποιείται για να πραγματοποιηθεί αυτή η εκτίμηση.

2.2.2 Μοντέλο διακριτής κατανομής

Αν οι συναρτήσεις κατανομών πιθανότητας είναι διακριτά μοντέλα των οποίων οι συναρτήσεις μάζας πιθανότητας δίνονται από $\{g(x_i)\colon i = 1,2,\dots\}$ $και$ $\{f(x_i)\colon i = 1,2,\dots\}$,τότε η K-L πληροφορία μπορεί να εκφραστεί ως:

$$I(f;g) = \sum_{i=1}^{\infty} f(x_i)\,\{log\,\frac{f(x_i)}{g(x_i)}\} \qquad (2.5)$$

(Konishi-Kitagawa,2008)

Έστω το μοντέλο $p_i = (p_1, p_2, \dots, p_k)$ το οποίο χαρακτηρίζει την πραγματική κατανομή με p_i την πιθανότητα να συμβεί το ω_i γεγονός και $q_i = (q_1, q_2, \dots, q_k)$ το μοντέλο διακριτής κατανομής που υποθέτουμε τα οποία ικανοποιούν αντίστοιχα τις εξής συνθήκες:

i) $0 < p_i < 1,\ 0 < q_i < 1$

ii) $\sum_{i=1}^{k} p_i = p_1 + \cdots + p_k = 1,\ \ \sum_{i=1}^{k} q_i = q_1 + \cdots + q_k = 1$

Ο λογάριθμος $log\,\frac{p}{q}$ είναι μία τυχαία μεταβλητή και όταν συμβεί το γεγονός ω_i παίρνει τη μορφή $log\,\frac{p_i}{q_i}$. Ορίζουμε λοιπόν ως:

$$I(p;q) = Elog\left(\frac{p}{q}\right) = \sum_{i=1}^{k} p_i\left(\frac{p_i}{q_i}\right) \qquad (2.6)$$

την Kullback-Leibler ποσότητα πληροφορίας η οποία ικανοποιεί τα ακόλουθα:

i) $I(p;q) \geq 0$

ii) $I(p;q) = 0 \iff p_i = q_i\ \mu\varepsilon\ i = (1,2,\dots,k)$

Ισχυριζόμαστε λοιπόν ότι όσο πιο μικρή είναι η τιμή της $I(p;q)$ και επομένως πιο κοντά στο 0 τόσο πιο κοντά είναι το μοντέλο q στη σωστή κατανομή. Τα μοντέλα p, q αντιστοιχούν στα μοντέλα συνεχών κατανομών f , g που αναφέραμε παραπάνω.

Παράδειγμα 2.1

Δύο παρουσιαστές baseball, Α και Β, πρόβλεψαν ότι η πιθανότητα μίας ομάδας να νικήσει είναι 0.7 και 0.5 αντίστοιχα. (Αυτά είναι τα μοντέλα διωνυμικής κατανομής q_A=(0.7,0.3) και q_B=(0.5,O.5) αντίστοιχα. Αν η σωστή πιθανότητα νίκης είναι 0.4 είναι φανερό ποιος παρουσιαστής ήταν σωστός. Εάν όμως η πιθανότητα νίκης γίνει 0.6 ποια πρόβλεψη είναι σωστή;

Απάντηση

Συγκρίνοντας τα δύο μοντέλα έχουμε:

$$I(p\ ;\ qA\ ) = 0.6log(0.6/0.7) + 0.4log(0.4/0.3) = -0.0925 + 0.1151 = 0.0226$$

$$I(p\ ;\ qB) = 0.6log(0.6/0.5) + 0.4log(0.4/0.5) = 0.1094 - 0.0893 = 0.0201$$

Επομένως σύμφωνα με το κριτήριο πληροφορίας K-L ο παρουσιαστής B είναι πιο κοντά στην αλήθεια καθώς $l(p;qB) < I(p;qA)$.

(Y. Sakamoto, M. Ishigro and G. Kitagawa,1986)

2.2.3 Εντροπία

Σύμφωνα με τον Akaike(1977) η **"εντροπία αρχής μεγίστου"** βασίζεται στο γεγονός ότι η αρνητική ποσότητα πληροφορίας K-L είναι η εντροπία του Boltzmann. Στην πραγματικότητα η K-L πληροφορία ονομάζεται αρνητική εντροπία (negative entropy ή "negentropy"). Συνεπώς

$$Boltzmann's\ entropy = -\log(\frac{f(x)}{g(x)}) \qquad (2.7)$$

Τότε,

$$-Boltzmann's\ entropy = \log(\frac{f(x)}{g(x)})$$

Και,

$$\text{K-L} = E_f(-Boltzmann's\ entropy)$$

$$= E_f(\log\left(\frac{f(x)}{g(x)}\right))$$

$$= \int f(x)\log\left(\frac{f(x)}{g(x)}\right)dx \qquad (2.8)$$

Επομένως η ελαχιστοποίηση της απόστασης K-L ισούται με τη μεγιστοποίηση της εντροπίας. Έτσι λοιπόν δικαιολογείται και ο όρος **"εντροπία αρχής μεγίστου".** Παρόλα αυτά η μεγιστοποίηση της εντροπίας μπορεί να οδηγήσει σε ένα μοντέλο με μεγαλύτερη αβεβαιότητα ενώ η ελαχιστοποίηση της K-L θεωρείται πιο ευθύς προσέγγιση καθώς δίνει ως αποτέλεσμα ένα προσεγγιστικό μοντέλο το οποίο χάνει ελάχιστη ποσότητα πληροφορίας από τα δεδομένα μας (Kenneth, Anderson,1998).

Στην περίπτωση που έχουμε μοντέλο διακριτής κατανομής $f = \{f_1, \dots, f_k\}$ η εντροπία μπορεί να εκφραστεί ως μία ποσότητα η οποία μεταβάλλεται αναλογικά με το λογάριθμο της πιθανότητας W ότι η κατανομή σχετικών συχνοτήτων ενός δείγματος με n παρατηρήσεις από το μοντέλο που

έχουμε υποθέσει συμφωνεί με την πραγματική κατανομή. Συγκεκριμένα υποθέτουμε ότι έχουμε ένα μοντέλο που ακολουθεί f κατανομή και έστω τυχαίο δείγμα μεγέθους n για το οποίο οι συχνότητες των κατηγοριών είναι $\{n_1, \dots, n_k\}, (n_1 + n_2 + \cdots + n_k) = n$ είτε οι σχετικές συχνότητες $\{g_1, \dots, g_k\}, (g_i = \frac{n_i}{n})$. Η πιθανότητα με την οποία οι συχνότητες $\{n_1, \dots, n_k\}$ πραγματοποιούνται από αυτό το μοντέλο είναι

$$\mathrm{W} = \frac{\mathrm{n!}}{\mathrm{n_1!,..,n_k!}} \mathrm{f_1^{n_1}, \dots, f_k^{n_k}} \qquad (2.9)$$

Λογαριθμώντας λοιπόν την παραπάνω ποσότητα και χρησιμοποιώντας την προσέγγιση του Stirling ο οποίος υποστήριξε ότι

$$logn! \sim nlogn - n \text{ (Stirling's approximation)}$$

Έχουμε,

$$logW = logn! - \sum_{i=1}^{k} logn_i! + \sum_{i=1}^{k} n_i logf_i$$

$$\approx nlogn - n - \sum_{i=1}^{k} n_i logn_i + \sum_{i=1}^{k} n_i + \sum_{i=1}^{k} n_i logf_i$$

$$= -\sum_{i=1}^{k} n_i \log\{\frac{n_i}{n}\} + \sum_{i=1}^{k} n_i logf_i$$

$$= \sum_{i=1}^{k} n_i \log\{\frac{f_i}{g_i}\}$$

$$= n\sum_{i=1}^{k} g_i \log\{\frac{f_i}{g_i}\}$$

$$= nB(g; f)$$

Συνεπώς,

$$B(g; f) \sim n^{-1} logW \qquad (2.10)$$

Η εντροπία $B(g; f)$ είναι προσεγγιστικά ανάλογη με το λογάριθμο της πιθανότητας του γεγονότος ότι οι σχετικές συχνότητες του δείγματος που προέρχεται από το μοντέλο που έχουμε υποθέσει συμφωνεί με την πραγματική κατανομή. Αξιοσημείωτο είναι ότι η K-L πληροφορία δεν είναι η πιθανότητα η οποία περιέχει την κατανομή που προέρχεται από το μοντέλο με την πραγματική κατανομή αλλά θεωρείται ως η πιθανότητα που περιέχει τα παρατηρούμενα δεδομένα από το μοντέλο (Konishi, Kitagawa, 2008).

2.3 Κριτήρια Πληροφορίας (IC)

Η επιλογή ενός μοντέλου με πολύ λίγες παραμέτρους μπορεί να οδηγήσει σε μη ρεαλιστικές υποθέσεις, σε μεγάλη μεροληψία καθώς και σε φτωχή πρόβλεψη. Τέτοιου είδους μοντέλα δεν είναι αρκετά ικανά προκειμένου να περιγράψουμε το δείγμα μας καθώς επίσης και ολόκληρο τον πληθυσμό. Παρ' όλα αυτά ένα ακόμη συχνό πρόβλημα στη στατιστική μοντελοποίηση είναι η ύπαρξη πολλών παραμέτρων στο μοντέλο καθώς έτσι αυξάνεται η πολυπλοκότητα. Οι ερευνητές επιδιώκουν ένα μοντέλο το οποίο θα προσαρμόζεται στα δεδομένα ενώ το πραγματικό μοντέλο δεν είναι γνωστό και αυτό επιτυγχάνεται με τη χρήση διάφορων κριτηρίων πληροφορίας όπως είναι το AIC (Akaike, 1973), CAIC (Bozdogan, 1987), BIC (Schwarz, 1978) και το διορθωμένο κριτήριο BIC(Sclove, 1987). Σύμφωνα με την αρχή της φειδωλότητας επιλέγεται το μοντέλο εκείνο με τις λιγότερες παραμέτρους το οποίο όμως είναι ικανό να περιγράψει καλύτερα τα δεδομένα. Η αρχή της φειδωλότητας (principle of Parsimony) υποστηρίζει ότι ένα απλό (φειδωλό) μοντέλο θεωρείται καλύτερο σε σχέση με ένα πολύπλοκο (Snuhua Hu,2007).

Αρχή της φειδωλότητας-Principle of Parsimony

Στο 14ο αιώνα ο Γουλιμέος του Όκαμ (Άγγλος Φραγκισκανός μοναχός, φιλόσοφος και θεολόγος) διατύπωσε το λεγόμενο «Το Ξυράφι του Όκαμ ή Λεπίδα του Όκαμ», μία επιστημονική αρχή η οποία εκφράζεται ως: «Κανείς δεν θα πρέπει να προβαίνει σε περισσότερες εικασίες από αυτές που είναι απαραίτητες» ή στα μαθηματικά μπορεί να εξηγηθεί ως όταν δύο θεωρίες παρέχουν εξίσου ακριβείς προβλέψεις πάντα επιλέγουμε την απλούστερη (Wikipedia, Ξυράφι του Όκαμ) .

Το Ξυράφι του Όκαμ έχει μία μακρινή ιστορία τόσο στην επιστήμη όσο και στην τεχνολογία και ενσωματώνεται στην αρχή της φειδωλότητας. Αντίστοιχα σύμφωνα με τον Άλμπερτ Αϊνστάιν, «όλα θα πρέπει να γίνονται όσο πιο απλά είναι πιθανόν αλλά όχι απλούστερα». Η επιτυχία στην ανάλυση πραγματικών δεδομένων εξαρτάται κυρίως από την επιλογή του βέλτιστου προσεγγιστικού μοντέλου. Η ανάλυση δεδομένων στις βιολογικές επιστήμες πρέπει να βασίζεται σε ένα φειδωλό μοντέλο που παρέχει μια ακριβή προσέγγιση στα πραγματικά δεδομένα. Αυτό δεν θα πρέπει να θεωρηθεί ως η αναζήτηση για το «αληθινό μοντέλο» (Burnham, Anderson,2002).

Ότι πιο απλό οδηγεί στο πιο βέλτιστο. Αυτό υποστηρίζει η αρχή της φειδωλότητας καθώς και στη Στατιστική προτιμώνται κυρίως μοντέλα που είναι εύκολο να γίνουν κατανοητά και αντιληπτά από τον καθένα. Η πολυπλοκότητα του μοντέλου είναι αναγκαίο να εξισορροπήσει την καλή προσαρμογή του προκειμένου να αποφευχθεί το λεγόμενο

"overfitting". Η αρχή της φειδωλότητας αναγκάζει τους ερευνητές να εγκαταλείψουν τα πολύπλοκα μοντέλα που προκαλούν έστω και μικρές τροποποιήσεις στα παρατηρούμενα δεδομένα. Η χρήση πολλών μεταβλητών για τη μοντελοποίηση μιας σχέσης με μία μεταβλητή απόκρισης μπορεί να παραβεί λοιπόν τη συγκεκριμένη αρχή. Επομένως συνιστάται η χρήση κατάλληλου μεγέθους μεταβλητών που μπορεί εύκολα να εκτιμηθεί.

Σύμφωνα με τον Atkinson(1980) η γενική εξίσωση του κριτηρίου πληροφορίας (IC) είναι η ακόλουθη:

$$\boldsymbol{IC = -2l + A_n k} \qquad (2.11)$$

Όπου

l: λογαριθμική συνάρτηση πιθανοφάνειας

$\mathrm{A_n}$: σταθερά του μεγέθους n του δείγματος

k: ο αριθμός των παραμέτρων στο μοντέλο

Η παραπάνω έκφραση κριτηρίου μερικές φορές χρησιμοποιείται ως $G^2 + A_n k$ όπου G^2 είναι η απόκλιση (Collins and Lanza,2010). Καθένα από τα παραπάνω κριτήρια πληροφορίας αποσκοπούν στην επιλογή μοντέλων με τη μεγαλύτερη τιμή της παράστασης $\mathbf{l - A_n k}$. Για ιστορικούς όμως λόγους έχει καθιερωθεί να υπολογίζεται η μικρότερη τιμή της (2.11), δηλαδή της ποσότητας $\boldsymbol{-2l + A_n k}$. Η τιμή της σταθεράς $\boldsymbol{A_n}$ μεταβάλλεται ανάλογα με το κριτήριο πληροφορίας το οποίο χρησιμοποιείται. Στον παρακάτω πίνακα παρουσιάζονται συνοπτικά οι τιμές της για κάποια κριτήρια πληροφορίας:

ΚΡΙΤΗΡΙΑ (CRITERIA)	ΟΡΟΣ ΠΟΙΝΗΣ (PENALTY WEIGHT)
AIC	$\mathrm{A_n} = 2$
CAIC	$\mathrm{A_n} = log(n+1)$
BIC	$\mathrm{A_n} = \log(n)$
Adjusted BIC	$\mathrm{A_n} = \log(\frac{n+2}{24})$

Διάγραμμα 2.1: Οι όροι ποινής για τα κριτήρια AIC,CAIC,BIC, AdBIC

Πηγή: (Dziak, Coffman, Lanza, Runze, 2015)

2.3.1 Περιγραφή και απόδειξη του κριτηρίου πληροφορίας του Akaike AIC

Ένα σημαντικό βήμα στη στατιστική μοντελοποίηση είναι να υιοθετηθεί η φιλοσοφία σχετικά με τα μοντέλα και την ανάλυση δεδομένων και να επιλεχθεί το κατάλληλο κριτήριο πληροφορίας. Το κλειδί όμως για την κατανόηση αυτής της φιλοσοφίας είναι η ικανότητα αναγνώρισης του σωστού μοντέλου μέσα από το οποίο τα δεδομένα θα προσεγγίζουν με όσο δυνατόν μεγαλύτερη ακρίβεια την πραγματικότητα. Όπως αναφέρεται και στις δημοσιεύσεις του Bozdogan (1987) και Burnham και Anderson (2004) το κριτήριο πληροφορίας του Akaike(1973), το λεγόμενο AIC, θεωρήθηκε αξιοσημείωτο για την στατιστική έρευνα, μοντελοποίηση και αξιολόγηση μοντέλων. Ο Akaike λοιπόν αφού πρότεινε τη χρήση της απόστασης Kullback-Leibler για την επιλογή μοντέλων και διαμόρφωσε μία σχέση μεταξύ της μέγιστης πιθανοφάνειας και της απόστασης αυτής, ανάπτυξε το κριτήριο AIC προκειμένου να εκτιμήσει την απόσταση αυτή. Θέτοντας λοιπόν στην εξίσωση (2.11) $\boldsymbol{A_n = 2}$ προκύπτει το AIC,το οποίο έχει την εξής μορφή:

$$\mathbf{AIC} = -\mathbf{2}(\mathbf{maximum\ log\ likelihood}) + \mathbf{2}(\mathbf{number\ of\ parameters})$$

$$\mathbf{AIC} = -\mathbf{2}\ \mathbf{l}(\widehat{\boldsymbol{\theta}}) + \ \mathbf{2}\boldsymbol{k} \qquad (2.12)$$

όπου $l(\hat{\theta}) = \sum_{i=1}^{k} logf(x_i, \hat{\theta})$είναι η μέγιστη συνάρτηση πιθανοφάνειας (loglikelihood) και $\widehat{\theta}$ είναι η εκτιμήτρια μεγίστης πιθανοφάνειας (Sakamoto 1986). Η σταθερά 2 χρησιμοποιείται για ιστορικούς λόγους (Burnham and Anderson,2002). Στην περίπτωση της μεθόδου ελαχίστων τετραγώνων με κανονικά κατανεμημένα σφάλματα το AIC παίρνει την εξής έκφραση:

$$AIC = nlog(\hat{\sigma}^2) + 2k$$

όπου $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n}$ και ε_i είναι τα εκτιμημένα υπόλοιπα για ένα συγκεκριμένο υποψήφιο μοντέλο (Burnham&Anderson,1998).

Όπως αναφέρθηκε και στον ορισμό της απόστασης Kullback-Leibler, η απόσταση ενός προσεγγιστικού μοντέλου από το πραγματικό ορίζεται ως:

$$I(f; g) = E_f\left[\log\left(f(x)\right)\right] - E_f[\log(g(x|\theta)]$$

Θεωρώντας την $E_f\left[\log\left(f(x)\right)\right]$ ως σταθερα C η παραπάνω σχέση γίνεται:

$$I(f; g) = C - E_f[\log(g(x|\theta))]$$

όπου $x_n = \{x_1, x_2, \dots, x_n\}$ ένα σύνολο από n ανεξάρτητες παρατηρήσεις, f η συνάρτηση

πυκνότητας του πραγματικού μοντέλου και g η συνάρτηση πυκνότητας του προσεγγιστικού. Ο όρος $E_f\left[\log\left(f(x)\right)\right]$ δεν εξαρτάται από τα δεδομένα άρα αρκεί να εκτιμηθεί η ποσότητα $E_f[\log(g(x|\theta)]$. Σύμφωνα με τον τύπο (2.5) μία εκτίμηση της απόστασης αυτής είναι:

$$E_{\hat{\theta}}\left[\hat{I}(f;g)\right] = Constant - E_y E_x[\log(f\left(x\middle|\hat{\theta}(y)\right)]$$

Προκειμένου λοιπόν να ελαχιστοποιηθεί η απόσταση Kullback-Leibler, δηλαδή η $\hat{I}(f;g)$, αρκεί να μεγιστοποιηθεί η προσδοκούμενη απόσταση K-L, δηλαδή η ποσότητα $E_y E_x[\log(f\left(x\middle|\hat{\theta}(y)\right)]$. Σε συνδυασμό με τη συλλογική πορεία του Akaike θα πρέπει λοιπόν να καθοριστεί μία μέθοδος μέσα από την οποία θα επιλεγεί το κατάλληλο μοντέλο g_i που θα μεγιστοποιεί τον παραπάνω όρο.

Για να προσεγγιστεί το ζήτημα αυτό θεωρείται ένα συνεχές τυχαίο διάνυσμα $\vec{x}$ με συνάρτηση πυκνότητας πιθανότητας $f(\vec{x}|\vec{\theta})$ όπου $\vec{\theta} = (\theta_1, \theta_2, \dots, \theta_\kappa)$, διάνυσμα παραμέτρων κ-διάστασης. Έστω $\vec{\theta}^*$ το διάνυσμα παραμέτρων του πραγματικού μοντέλου το οποίο προσπαθεί να προσεγγίσει το $\vec{\theta}$. Τότε η απόσταση Kullback-Leibler που ορίστηκε στον τύπο (2.3)

$$I(f;g) = E_f\left[\log\left(f(x)\right)\right] - E_f[\log(g(x|\theta)]$$

παίρνει την εξής μορφή:

$$I(\vec{\theta}^*, \vec{\theta}) = E[\log f(x|\vec{\theta}^*)] - E\left[logf(x|\vec{\theta})\right] =$$

$$= \int f(\vec{x}|\vec{\theta}^*)\log f(\vec{x}|\vec{\theta}^*)\, d\vec{x} - \int f(\vec{x}|\vec{\theta}^*)\, logf(\vec{x}|\vec{\theta}) d\vec{x} =$$

$$= H(\vec{\theta}^*, \vec{\theta}^*) - H(\vec{\theta}^*, \vec{\theta}) \quad (2.13)$$

όπου $H(\vec{\theta}^*, \vec{\theta})$ η αναμενόμενη λογαριθμική πιθανοφάνεια που δείχνει πόσο καλά προσαρμόζεται το προσεγγιστικό μοντέλο στο πραγματικό και όπου $H(\vec{\theta}^*, \vec{\theta}^*)$ η αρνητική εντροπία Shannon, θεωρία της πληροφορίας, η οποία ισούται με:

$$H(\vec{\theta}^*, \vec{\theta}^*) = H(\vec{\theta}^*) = constant$$

Καθώς στόχος είναι η εκτίμηση της $\hat{H}(\vec{\theta}^*, \vec{\theta})$, η σχέση (2.13) γίνεται:

$$\hat{H}(\vec{\theta}^*, \vec{\theta}) = -\hat{I}(\vec{\theta}^*, \vec{\theta}) + \hat{H}(\vec{\theta}^*, \vec{\theta}^*) \qquad (2.14)$$

Η εκτίμηση της $\hat{H}(\vec{\theta}^*, \vec{\theta})$ είναι η μέση λογαριθμική πιθανοφάνεια $l_n(\vec{\theta})$. Προκειμένου να ελαχιστοποιηθεί η απόσταση Kullback-Leibler αρκεί να μεγιστοποιηθεί η αναμενόμενη λογαριθμική πιθανοφάνεια.

Ορισμός της μέσης λογαριθμικής πιθανοφάνειας:

Η συνάρτηση πιθανοφάνειας για το μοντέλο με σ.π.π. : $f(\vec{x}|\vec{\theta})$ ορίζεται όπως αναφέρεται και στο 1ο κεφάλαιο ως:

$$L(\vec{\theta}) = f(x_1, x_2, \dots, x_n|\vec{\theta}) = \prod_{i=1}^{n} logf(x_i|\vec{\theta})$$

Λογαριθμώντας την παραπάνω συνάρτηση έχω:

$$l(\vec{\theta}) = logL(\vec{\theta}) = \sum_{i=1}^{n} logf(x_i|\vec{\theta})$$

Έτσι η μέση λογαριθμική πιθανοφάνεια ορίζεται ως:

$$\frac{l(\vec{\theta})}{n} = \frac{\sum_{i=1}^{n} logf(x_i|\vec{\theta})}{n} = l_n(\vec{\theta})$$

Αντικαθιστώντας : $E\left(\frac{l(\vec{\theta})}{n}\right) = \hat{H}(\vec{\theta}^*, \vec{\theta})$ η σχέση (2.14) γίνεται:

$$I(\vec{\theta}^*, \vec{\theta}) = \hat{H}(\vec{\theta}^*) - \frac{l(\vec{\theta})}{n} = \hat{H}(\vec{\theta}^*) - \frac{\sum_{i=1}^{n} logf(x_i|\vec{\theta})}{n} = \hat{H}(\vec{\theta}^*) - l_n(\vec{\theta})$$

Άρα $$l_n(\vec{\theta}) = -I(\vec{\theta}^*, \vec{\theta}) + \hat{H}(\vec{\theta}^*)$$

Η μέση λογαριθμική πιθανοφάνεια $l_n(\vec{\theta})$ παίρνει λοιπόν τη μέγιστη τιμή όταν $I(\vec{\theta}^*, \vec{\theta}) = 0$, ή εναλλακτικά όταν η απόσταση του προσεγγιστικού μοντέλου από το πραγματικό μοντέλο είναι μηδέν. Δηλαδή τα δύο μοντέλα ταυτίζονται και $f(\vec{x}|\vec{\theta}) = f(\vec{x}|\vec{\theta}^*)$. Τότε η μέγιστη τιμή που θα πάρει θα είναι $l_n(\vec{\theta}) = \hat{H}(\vec{\theta}^*)$, θα ισούται δηλαδή μόνο με την αρνητική εντροπία Shannon. Άρα θα υπάρχει θ_0, το σημείο στο οποίο ελαχιστοποιείται η K-L απόσταση, τέτοιο ώστε:

$$\theta_0 = \theta = \theta^*$$

Ο Akaike βασιζόμενος στα παραπάνω επιδίωκε να εκτιμήσει την ποσότητα $E_y E_x[\log\left(f\left(x\middle|\hat{\theta}(y)\right)\right]$ προκειμένου να κατασκευάσει το κατάλληλο κριτήριο πληροφορίας. Έστω θ_0 η τιμή του θ που ελαχιστοποιεί την πληροφορία K-L, $I(f, g)$. Επειδή όμως η θ_0 είναι άγνωστη, ο Akaike, χρησιμοποίησε μία εκτίμηση της, τη $\hat{\theta}$ η οποία συντελεί στην ελαχιστοποίηση της αναμενόμενης τιμής της K-L. Για να εκτιμήσει λοιπόν

την $E_y E_x[\log\left(f\left(x\middle|\hat{\theta}(y)\right)\right]$, μεγιστοποίησε την ποσότητα $logL(\vec{\theta})$ για κάθε προσεγγιστικό μοντέλο g. Απέδειξε ότι η εκτιμήτρια αυτή δεν είναι αμερόληπτη και ότι η μεροληψία της υπολογίζεται προσεγγιστικά ως k, ο αριθμός δηλαδή των παραμέτρων στο προσεγγιστικό μοντέλο και ο όρος διόρθωσης της μεροληψίας. Έτσι μία ολοκληρωμένη εκτίμηση της αναμενόμενης τιμής της πληροφορίας K-L είναι $logL(\vec{\theta}) - k$. Για ιστορικούς λόγους ο Akaike πολλαπλασίασε την εκτίμηση αυτή με -2 και έτσι όρισε το κριτήριο πληροφορίας AIC(Akaike Information Criterion):

$$\boldsymbol{AIC = -2logL(\vec{\theta}) + 2k}$$

Αξιοσημείωτο είναι ότι όσο μεγαλύτερος είναι ο αριθμός των παραμέτρων που χρησιμοποιήθηκαν στο μοντέλο τόσο αυξάνεται η τιμή της μέγιστης λογαριθμικής πιθανοφάνειας. Ο όρος διόρθωσης του σφάλματος λοιπόν αποσκοπεί στο να αποτρέψει το μοντέλο να γίνει πολύπλοκο. Το κριτήριο πληροφορίας του Akaike έχει ως στόχο την επιλογή μοντέλων που πιθανόν να έχουν λίγες παραμέτρους αλλά παρ' όλα αυτά να προσαρμόζονται καλά στα δεδομένα. Παρακάτω παρατίθενται δύο παραδείγματα επιλογής του κατάλληλου μοντέλου με τη χρήση του κριτηρίου AIC.

<u>Παράδειγμα 2.1-Σύγκριση του AIC ενός μοντέλου εκθετικής κατανομής με ένα μοντέλο κατανομής Weibull</u>

Σύγκριση μεταξύ ενός μοντέλου εκθετικής κατανομής (exponential distribution) με ένα που ακολουθεί κατανομή Weibull. Η συνάρτηση πυκνότητας πιθανότητας της εκθετικής κατανομής είναι f(y) = $\theta e^{-\theta y}$, y>0, θ>0 με συνάρτηση πιθανοφάνειας:

L(θ |y_1, $y_{2, \dots,}$ y_n) = $\prod_{i=1}^{n} f(y_i, \theta)$ =$\prod_{i=1}^{n} \theta e^{-\theta y_i}$ = $\theta^n e^{-\sum_{i=1}^{n} \theta y_i}$

Επομένως η λογαριθμημένη συνάρτηση πιθανοφάνειας είναι η εξής:

l(θ) = logL(θ) = log ($\theta^n e^{-\sum_{i=1}^{n} \theta y_i}$) = nlogθ - $\sum_{i=1}^{n} \theta y_i$

Άρα για την εκτιμήτρια $\hat{\theta}$ έχω:

l($\hat{\theta}$) = nlog$\hat{\theta}$ - $\sum_{i=1}^{n} \hat{\theta} y_i$ = $\sum_{i=1}^{n} (log\hat{\theta} - \hat{\theta} y_i)$

Αντίστοιχα η συνάρτηση πυκνότητας πιθανότητας της κατανομής Weibull είναι:

f(y) =exp{-$(\theta y)^{\gamma}$}$\theta^{\gamma}\gamma y^{\gamma-1}$ όπου για γ=1 καταλήγω στην εκθετική κατανομή.

Έτσι η συνάρτηση πιθανοφάνειας γίνεται:

L(θ, γ |y_1, $y_{2, \dots,}$ y_n) = $\prod_{i=1}^{n} f(y_i)$ = $\theta^{n\gamma}$ $\gamma^n \Pi_{i=1}^{n} y_i^{\gamma-1}$exp{$\sum_{i=1}^{n} [-(\theta y_i)^{\gamma}]$}

και η λογαριθμημένη συνάρτηση πιθανοφάνειας για τις $\hat{\theta}$,$\hat{\gamma}$ είναι η εξής:

l($\widehat{\theta}$, $\hat{\gamma}$) =$\sum_{i=1}^{n}\{-\hat{\theta}y_i)^{\hat{\gamma}} + \hat{\gamma}log\hat{\theta} + log\hat{\gamma} + (\hat{\gamma}-1)logy_i\}$

για κάθε μοντέλο (εκθετική κατανομή και Weibull) έχουμε:

AIC(exp) = -2$\sum_{i=1}^{n}(log\hat{\theta} - \hat{\theta}y_i)$ – 2 (διότι k=1)

AIC(wei) = -2$\sum_{i=1}^{n}\{-\hat{\theta}y_i)^{\hat{\gamma}} + \hat{\gamma}log\hat{\theta} + log\hat{\gamma} + (\hat{\gamma}-1)logy_i\}$ – 4 (διότι k=2)

Το μοντέλο με τη μικρότερη τιμή του AIC θα θεωρηθεί ως το καταλληλότερο μοντέλο για τα δεδομένα μας. (Claeskens and Hjort, 2008)

<u>Παράδειγμα 2.2-Μοντέλο πολλαπλής γραμμικής παλινδρόμησης</u>

Ας θεωρήσουμε ότι έχουμε ένα μοντέλο γραμμικής παλινδρόμησης της μορφής

$Y_i=x_{i,1}\beta_1+ x_{i,2}\beta_2 + \dots + x_{i,p}\beta_p + \varepsilon_i = x_i^T\beta + \varepsilon_i$ για i=1, …, n

όπου $Y=(Y_1,Y_2,\dots,Y_n)^T$ το διάνυσμα των εξηρτημένων μεταβλητών,

$x_i = (x_{i,1}, \dots,x_{i,p})^T$ το σύνολο των ανεξάρτητων μεταβλητών για το i άτομο,

$\beta=(\beta_1, \dots, \beta_p)^T$το διάνυσμα των συντελεστών παλινδρόμησης και $\varepsilon=(\varepsilon_1, \dots, \varepsilon_n)$ όπου κάθε σφάλμα ακολουθεί την $N(0, \sigma^2)$.

Το μοντέλο αυτό γράφεται υπό τη μορφή πινάκων ως $Y=X\beta + \varepsilon$ όπου X είναι ένας nxp πίνακας.

Η συνάρτηση πιθανοφάνειας ορίζεται ως:

$L(\beta, \sigma)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y_i - x_i^T\beta)^2\right\}=(\frac{1}{\sqrt{2\pi}\sigma})^n\exp\{\Sigma_{i=1}^{n}\,[\![\left[-\frac{1}{2\sigma^2}(y_i - x_i^T\beta)^2\right]\!]\}$

Άρα η λογαριθμική συνάρτηση πιθανοφάνειας γίνεται:

$L(\beta,\sigma) =\sum_{i=1}^{n}\{-\frac{1}{2}\log(2\pi) - \frac{1}{2}[(y_i - x_i^T\beta)^2/\sigma^2] - log\sigma\}$ και δεδομένου ότι $\widehat{\sigma^2} = n^{-1}SSE\left(\hat{\beta}\right) = n^{-1}\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2$ η l $(\hat{\beta},\hat{\sigma})$ γίνεται:

$l\,(\hat{\beta},\hat{\sigma}) = -nlog\hat{\sigma} - \frac{n}{2} - \frac{n}{2}log2\pi$

και AIC = -2l $(\hat{\beta},\hat{\sigma})$ +2k = $2nlog\hat{\sigma} + n + nlog2\pi + 2(p+1)$

Άρα σύμφωνα με το κριτήριο AIC το καλύτερο μοντέλο θα θεωρηθεί εκείνο για το οποίο η τιμή του AIC θα είναι μικρότερη. (laeskens and Nils Lid Hjort, 2008)

2.4 Ανάλυση των κριτηρίων TIC, AICc, QAIC

Ο Akaike, το 1973, έθεσε το κριτήριο πληροφορίας Kullback-Leibler ως μία θεμελιώδη αρχή για την επιλογή μοντέλων, την απέδειξε και υποστήριξε ότι η ποσότητα αυτή μπορεί να εκτιμηθεί από τα εμπειρικά δεδομένα και από την εκτιμήτρια μέγιστης πιθανοφάνειας. Η εύρεση της σχέσης μεταξύ της πληροφορίας K-L και της μέγιστης πιθανοφάνειας,(2.4), έπαιξε καταλυτικό ρόλο στην ανάλυση δεδομένων. Τρία χρόνια μετά την ανακάλυψη του Akaike, το 1976, ο Takeuchi διαμορφώνει μία τροποποιημένη μορφή του κριτηρίου AIC, το λεγόμενο TIC (Takeuchi,K.1976) το οποίο θεωρείται αρκετά χρήσιμο όταν τα υποψήφια μοντέλα προσεγγίζουν ικανοποιητικά την κατανομή f του πραγματικού μοντέλου. Αποτελεί μία διόρθωση του AIC και υποστηρίζει ότι η πραγματική κατανομή ανήκει στο σύνολο των υποψήφιων μοντέλων. Με βάση λοιπόν την σχέση (2.4)

$$I(f;g) = C - E_f\left[\log\left(g(x|\theta)\right)\right] \Leftrightarrow I(f;g) = C - E_f\{\log\left[\mathrm{L}\left(\hat{\theta}|y\right)\right]\}$$

όπου $\hat{\theta}$ είναι η εκτιμήτρια μέγιστης πιθανοφάνειας των k- παραμέτρων των μοντέλων υπό το μοντέλο g. Ο Taukechi λοιπόν πρόσθεσε στην παραπάνω σχέση την εξής ποσότητα:

$$-E_f\{\log\left[\mathrm{L}\left(\hat{\theta}|y\right)\right]\} = \mathrm{I(f,g)} - \mathrm{C} - \mathrm{trace}\{\mathrm{J}(\theta_0), [\mathrm{I}(\theta_0)]^{-1}\}$$

ή ισοδύναμα:

$$-E_f\{\log\left[\mathrm{L}\left(\hat{\theta}|y\right)\right]\} + \mathrm{trace}\{\mathrm{J}(\theta_0), [\mathrm{I}(\theta_0)]^{-1}\} = \mathrm{I(f,g)} - \mathrm{C}$$

Όπου J,I είναι k×k πίνακες οι οποίοι ορίζονται ως:

$$J(\theta_0) = E_f\left[\left(\frac{\partial \log\left(g(x|\theta)\right)}{\partial\theta}\right)\left(\frac{\partial \log\left(g(x|\theta)\right)}{\partial\theta}\right)^{\mathrm{T}}\right]_{\theta=\theta_0}$$

$$\mathrm{I}(\theta_0) = E_f\left[-\frac{\partial^2 \log\left(g(x|\theta)\right)}{\partial\theta_\iota\theta_j}\right]_{\theta=\theta_0}$$

Μία αμερόληπτη εκτιμήτρια της ποσότητας $E_f\{\log\left[\mathrm{L}\left(\hat{\theta}|y\right)\right]\}$ είναι απλά η $\log[\mathrm{L}\left(\hat{\theta}|y\right)]$. Συνεπώς αρκεί να υπολογιστεί το παραπάνω ίχνος των πινάκων προκειμένου να διαμορφωθεί το κριτήριο TIC το οποίο τελικά ορίζεται ως:

$$\mathbf{TIC} = -\mathbf{2}\log\left[\boldsymbol{L}\left(\widehat{\boldsymbol{\theta}}|\boldsymbol{y}\right)\right] + \mathbf{2}\boldsymbol{tr}\left(\widehat{\boldsymbol{J}}\left(\widehat{\boldsymbol{\theta}}\right)\left[\widehat{\boldsymbol{I}}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-\mathbf{1}}\right)$$

όπου $\hat{I}(\hat{\theta}) = -\frac{\partial^2 \log(g(x|\hat{\theta})}{\partial \theta^2}$, η εκτίμηση του $I(\theta_0)$

$\hat{J}(\hat{\theta}) = \sum_{i=1}^{n}(\frac{\partial \log[g(x_i|\hat{\theta})]}{\partial \theta})(\frac{\partial \log[g(x_i|\hat{\theta})]}{\partial \theta})^T$, η εκτίμηση του $J(\theta_0)$

Το AIC κριτήριο μπορεί να χαρακτηριστεί φτωχό εάν υπάρχουν πολλές παράμετροι σε σχέση με το μέγεθος του δείγματος. Σύμφωνα με την αρχική με την έρευνα του Sugiura, οι Hurvich και Tsai το 1989 οδηγήθηκαν στην έκφραση του κριτηρίου του AICc το οποίο θεωρείται ως επέκταση ή διόρθωση του κριτηρίου AIC (second order Information criterion) το οποίο χρησιμοποιεί αμερόληπτες εκτιμήτριες προκειμένου να διορθώσει την μεροληψία ως προς το μέγεθος του δείγματος. Ορίζεται ως:

$$\mathbf{AICc} = -\mathbf{2}\,\mathbf{l}(\widehat{\boldsymbol{\theta}}) + \mathbf{2}\boldsymbol{k}(\frac{\boldsymbol{n}}{\boldsymbol{n}-\boldsymbol{k}-\mathbf{1}})$$

όπου παρατηρούμε ότι ο όρος $\mathbf{2}\boldsymbol{k}$ ο οποίος προέρχεται από το κριτήριο AIC πολλαπλασιάζεται με τον παράγοντα διόρθωσης $\frac{\boldsymbol{n}}{\boldsymbol{n}-\boldsymbol{k}-\mathbf{1}}$**.** Μία άλλη ισοδύναμη έκφραση του κριτηρίου είναι:

$$\mathbf{AICc} = -\mathbf{2}\,\mathbf{l}(\widehat{\boldsymbol{\theta}}) + \mathbf{2}\boldsymbol{k} + (\frac{\mathbf{2}\boldsymbol{k}(\boldsymbol{k}+\mathbf{1})}{\boldsymbol{n}-\boldsymbol{k}-\mathbf{1}})$$

ή

$$\mathbf{AICc} = \boldsymbol{AIC} + (\frac{\mathbf{2}\boldsymbol{k}(\boldsymbol{k}+\mathbf{1})}{\boldsymbol{n}-\boldsymbol{k}-\mathbf{1}})$$

όπου n το μέγεθος του δείγματος. Το AICc ενδείκνυται κυρίως όταν το μέγεθος του δείγματος δεν είναι αρκετά μεγάλο. Το AICc λοιπόν έχει ένα προσθετικό όρο διόρθωσης της μεροληψίας. Εάν το μέγεθος του δείγματος όμως είναι μεγάλο τότε το AIC αρμόζει κατάλληλα και το AICc μπορεί εύκολα να θεωρηθεί αμελητέο. Ο Findley το 1985 σημείωσε το πόσο σημαντική είναι η μελέτη της μεροληψίας. Ο όρος διόρθωσης της μεροληψίας βέβαια ποικίλει ανάλογα με τον τύπο μοντέλου που χρησιμοποιούμε (π.χ. μπορεί το μοντέλο να προέρχεται από κανονική κατανομή, εκθετική ή Poisson). Ενώ το AICc προέρχεται από Γκαουσιανές υποθέσεις για γραμμικά μοντέλα, ο Burnham το 1994 ανακάλυψε ότι αυτή η επέκταση της προσέγγισης στην K-L απόσταση είναι χρήσιμη σε πολυγραμμικά μοντέλα. Γενικότερα υποστηρίζουμε τη χρήση του AICc όταν ο λόγος $\frac{n}{k}$ είναι μικρός (περίπου μικρότερο του 40). Εάν όμως ο λόγος είναι μεγάλος τότε τα κριτήρια AIC και AICc είναι παρόμοια και συνήθως οδηγούν στο ίδιο μοντέλο.

Υπάρχουν αρκετά λογισμικά που παρέχουν το AICc, παρ' όλα αυτά μπορεί εύκολα να υπολογιστεί με το χέρι.

Όταν όμως στο σύνολο των δεδομένων υπάρχει μεγαλύτερη διασπορά από αυτή που θα αναμενόταν σε ένα δεδομένο στατιστικό μοντέλο τότε οι παρακάτω μορφές κριτηρίων θεωρούνται κατάλληλες(Cox and Snell,1989).

$$\boldsymbol{QAIC} = -\left[\frac{\mathbf{2}\log\left(\boldsymbol{L}(\widehat{\boldsymbol{\theta}})\right)}{\hat{\boldsymbol{c}}}\right] + \mathbf{2}\boldsymbol{k}$$

και

$$\boldsymbol{QAICc} = -\left[\frac{\mathbf{2}\log\left(\boldsymbol{L}(\widehat{\boldsymbol{\theta}})\right)}{\hat{\boldsymbol{c}}}\right] + \mathbf{2}\boldsymbol{k} + \frac{\mathbf{2}\boldsymbol{k}(\boldsymbol{k}+\mathbf{1})}{\boldsymbol{n}-\boldsymbol{k}-\mathbf{1}}$$

όπου ο συντελεστής μεταβλητής c μπορεί να εκτιμηθεί μέσω ενός αξιόλογου τεστ καλής προσαρμογής, του στατιστικού ελέγχου χ^2 το οποίο ανακαλύφθηκε από τον Pearson το 1900. Ο έλεγχος αξιολογεί κατά πόσο πολυωνυμικές πιθανότητες είναι ίσες με κάποιες υποθετικές τιμές ή μία παρατηρούμενη κατανομή προσαρμόζεται κατάλληλα στη θεωρητική. Η εκτίμηση του c συντελεστή υπολογίζεται ως:

$$\hat{c} = \frac{\chi^2}{df}$$

και προέρχεται κυρίως από το μοντέλο με τη μεγαλύτερη διάσταση. Ο όρος χ^2 είναι ο αρχαιότερος και περισσότερο γνωστός έλεγχος καλής προσαρμογής , ο οποίος προτάθηκε από τον Karl Pearson το 1900. Ο έλεγχος αυτός χρησιμοποιείται σε περιπτώσεις προβλημάτων στα οποία ενδιαφερόμαστε να εξετάσουμε αν τα δεδομένα μας προέρχονται από μια ορισμένη κατανομή. Ενώ η ποσότητα df αποτελεί τους βαθμούς ελευθερίας. Προφανώς, όταν δεν υπάρχει τεράστια διασπορά, ο συντελεστής ισούται με 1 $(c = 1)$ και τότε τα κριτήρια QAIC και QAICc ταυτίζονται με τα AIC και AICc αντίστοιχα.

Τα AIC, AICc,TIC, QAIC, QAICc θεωρούνται εκτιμήσεις της σχετικής K-L απόστασης μεταξύ της κατανομής f και καθενός από τα M προσεγγιστικά μοντέλα $g_i(x)$ και βασίζονται στην ιδέα ότι η πραγματικότητα είναι αρκετά πολύπλοκη και δεν υπάρχει «σωστό μοντέλο». Ένα κριτήριο μπορεί να προσεγγίσει τη πραγματικότητα με ένα μοντέλο $g(x)$.Το πλήθος K των μοντέλων βασίζεται στο μέγεθος του δείγματος. Όσο περισσότερα δεδομένα είναι διαθέσιμα τόσο περισσότερα μοντέλα προσδιορίζονται.

2.5 Σύγκριση Μοντέλων

Το Δ_i (Delta AIC) και το w_i (Akaike weights) είναι δύο μέτρα που σχετίζονται με το AIC τα οποία είναι εύκολο να υπολογιστούν καθώς τα αποτελέσματα παραμένουν τα ίδια ανεξάρτητα αν χρησιμοποιείται το AIC ή το AICc, καθώς επίσης μπορούν να ερμηνευτούν εύκολα.

Το Δ_i σχετίζει κάθε μοντέλο με το καταλληλότερο μοντέλο (best model) και ορίζεται ως εξής:

$$\mathbf{\Delta_i = AIC_i - minAIC}$$

όπου ο παράγοντας AIC_i εκφράζει την τιμή του AIC για κάθε μοντέλο και ο παράγοντας minAIC αναφέρεται στην μικρότερη τιμή όλων των μοντέλων. Υποστηρίζεται ότι όσο μικρότερη είναι η τιμή του Δ_i τόσο μεγαλύτερη ένδειξη υπάρχει για το μοντέλο.

Το w_i (Akaike weights) παρέχει ένα άλλο μέτρο σύγκρισης μοντέλων και αναπαριστά την αναλογία των τιμών του Δ_i του κάθε μοντέλου σε σχέση με το πλήθος r των μοντέλων και ορίζεται ως:

$$\textbf{Akaike weights} = \mathbf{w_i} = \frac{\mathbf{e}^{-\frac{\Delta_i}{2}}}{\sum_{\mathbf{r=1}}^{\mathbf{R}} \mathbf{e}^{-\frac{\Delta_r}{2}}}$$

Η ερμηνεία του w_i (Akaike weights) είναι σαφής. Υποδεικνύει δηλαδή την πιθανότητα ότι το μοντέλο είναι το καλύτερο σε σχέση με τα υπόλοιπα υποψήφια μοντέλα. Για παράδειγμα εάν η τιμή του w_i για το μοντέλο είναι 0.75 υποδεικνύει ότι υπάρχει 75% πιθανότητα να είναι το καταλληλότερο μοντέλο σε σχέση με τα υπόλοιπα. Έτσι μπορεί να θεωρηθεί ως αναλογία ένδειξης το εξής:

$$\textbf{Evidence ratio} = \frac{w_j}{w_i}$$

για τη σύγκριση του j μοντέλου με το μοντέλο i. Για παράδειγμα εάν $\frac{w_j}{w_i} = \frac{0.55}{0.40} = 1.375$, αυτό σημαίνει ότι το μοντέλο j είναι μόνο 1.375 περισσότερο πιθανό να είναι το κατάλληλο σε σχέση με το μοντέλο i (MarcJ. Mazerolle, 2007).

Τα w_i (Akaike weights) λοιπόν μπορούν να εκληφθούν ως πιθανότητες, η πιθανότητα δηλαδή ότι το δοθέν μοντέλο είναι το κατάλληλο. Εάν επιστρέψουμε στο δείγμα μας, χρησιμοποιήσουμε περισσότερα στοιχεία από τον πληθυσμό μας και τα προσαρμόσουμε ξανά στα ίδια μοντέλα τότε το w_i θα έδινε πάλι την πιθανότητα ότι το δοθέν μοντέλο θεωρείται το κατάλληλο ακόμη και σε επαναλαμβανόμενη δειγματοληψία. (Burnham, AndersonD.R. 2002)

Κεφάλαιο 3ο

Μπεϋζιανό Κριτήριο Πληροφορίας (BIC)

3.1 Εισαγωγή στη Μπεϋζιανή Στατιστική

Μετά την εποχή του Fisher, ο οποίος εισήγαγε βασικές έννοιες στην κλασσική Στατιστική ακολουθεί η Μπεϋζιανή Στατιστική. Η Μπεϋζιανή προσέγγιση προσπαθεί να εξηγήσει και να ελέγξει, μέσω πιθανοτήτων, την αβεβαιότητα. Σύμφωνα με τον Bayes η μόνη ικανοποιητική περιγραφή της αβεβαιότητας επιτυγχάνεται μέσω της πιθανότητας. Η συμπερασματολογία λοιπόν κατά Bayes συνδέεται άμεσα με την κλασσική συμπερασματολογία της οποίας στόχος είναι να εξάγει συμπεράσματα για ένα πληθυσμό, μελετώντας όμως ένα δείγμα. Όπως έχει αναφερθεί και στο 2ο κεφάλαιο προκειμένου να εκτιμηθεί μία παράμετρος θ του πληθυσμού εξετάζεται και παρατηρείται η τιμή της μεταβλητής X με τη βοήθεια κάποιου μοντέλου πιθανότητας $f(x|\theta)$. Στη συμπερασματολογία κατά Bayes όμως χρησιμοποιείται το μοντέλο πιθανότητας $f(\theta|x)$, βασίζεται δηλαδή στη πιθανότητα της κατανομής της παραμέτρου δεδομένης της x, (α- posterior κατανομή). Βασική διαφορά της μπευζιανής με την κλασσική στατιστική είναι ότι στην πρώτη οι άγνωστες παράμετροι θ χρησιμοποιούνται ως τυχαίες μεταβλητές. Γι' αυτό το λόγο απαραίτητος είναι ο καθορισμός της λεγόμενης a-priori κατανομής f(θ) (prior probability distribution). Η χρήση της υποκειμενικής πιθανότητας, η δυνατότητα δηλαδή του κάθε ατόμου να επιλέγει με βάση τις γνώσεις του και η συνέπεια είναι κάποια βασικά χαρακτηριστικά της Μπεϋζιανή στατιστικής.

Υπάρχει γενικότερα ένας μεγάλος αντίλογος ανάμεσα στους δύο αυτούς κόσμους της Στατιστικής. Στη Μπεϋζιανή θεωρία έχει προκαλέσει αρκετά το γεγονός ότι τα συμπεράσματα εξαρτώνται από την επιλογή της a-prior κατανομής. Η συμπερασματολογία για την άγνωστη ποσότητα γίνεται με τη βάση του κανόνα του Bayes.

<u>Το Θεώρημα του Bayes:</u>

Η βασική μορφή του θεωρήματος είναι η εξής:

Έστω A και B είναι δύο ενδεχόμενα με P(A)>0, τότε ισχύει

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Σε όρους τυχαίων μεταβλητών το θεώρημα παίρνει την εξής μορφή:

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{\int f(\theta)f(x|\theta)d\theta}$$

ή σε περίπτωση διακριτών μεταβλητών το θεώρημα γίνεται:

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{\sum f(\theta_j)f(x|\theta_j)}$$

όπου f θεωρείται η α-priori συνάρτηση πυκνότητας πιθανότητας.

Μία βασική διαφορά ανάμεσα στη Κλασσική Στατιστική και στη Μπεϋζιανή είναι η χρήση της a-priori κατανομής (Prior Distribution). Η εκ των προτέρων κατανομή $f(\theta)$ καθορίζει τις τιμές του θ οι οποίες είναι περισσότερο πιθανές πριν την παρατήρηση δεδομένων. Ενώ το αποτέλεσμα που προκύπτει από την χρήση του κανόνα του Bayes, είναι γνωστό ως η εκ των υστέρων κατανομή (Posterior Distribution) του θ και περιγράφει την κατάσταση της παραμέτρου θ αφού παρατηρηθούν τα δεδομένα. Η a-prior κατανομή είναι αυτή όμως που προβληματίζει περισσότερο τους Στατιστικούς και κατά πόσο η επιλογή αυτής της κατανομής μπορεί να επηρεάσει τα συμπεράσματα. (Ι. Παναρέτου & Ε.Ξεκαλάκη,2000, Chapter24, Aueb).

Συνήθως πριν δοθεί μία εκτίμηση της παραμέτρου θ θα πρέπει να υπάρχει κάποια εκτίμηση για την τιμή της. Στο κόσμο της Μπεϋζιανή Στατιστικής, η παράμετρος θ θεωρείται τυχαία και η μελέτη μας βασίζεται στη πιθανότητα της κατανομής της παραμέτρου θ δεδομένου του x, δηλαδή στη συνάρτηση $f(\theta|x)$. Το γεγονός όμως ότι τα συμπεράσματα εξαρτώνται από την a-prior κατανομή έχει δημιουργήσει τη μεγαλύτερη αντίρρηση ανάμεσα σε αυτούς τους δύο κόσμους της Στατιστικής. (Aueb, 2006)

3.1.1 Περιγραφή Απόδειξης του BIC

Ένα σημαντικό εργαλείο της Μπεϋζιανή Στατιστικής είναι το μπεϋζιανό κριτήριο πληροφορίας BIC (Bayesian Information Criterion) το οποίο παρουσιάστηκε από τον Akaike και τον Schwarz (1977-1978) ως βελτίωση του κριτηρίου AIC. Το BIC όπως και το AIC χρησιμοποιείται προκειμένου να επιλεχθεί το κατάλληλο μοντέλο ανάμεσα από ένα πλήθος υποψήφιων μοντέλων και το μοντέλο με τη μικρότερη τιμή του κριτηρίου θεωρείται το καλύτερο, εκείνο δηλαδή που προσαρμόζεται καλύτερα στα δεδομένα. Η γενική μορφή του BIC διαφέρει με αυτή του AIC ως προς τον όρο ποινικοποίησης. Θέτοντας λοιπόν στη γενική εξίσωση κριτηρίων πληροφορίας που ορίστηκε στο 2ο κεφάλαιο:

$$\boldsymbol{IC = -2l + A_n k}$$

ως $A_n = \log(n)$ τότε προκύπτει το BIC ,το οποίο έχει την εξής μορφή:

$$\mathbf{BIC} = -\mathbf{2}(\mathbf{maximum\,log\,likelihood}) + \mathbf{log}(\mathbf{n})(\mathbf{number\ of\ parameters})$$

$$\mathbf{BIC} = -\mathbf{2}\,\mathbf{l}(\widehat{\boldsymbol{\theta}}) + \mathbf{log}(\mathbf{n})\boldsymbol{k} \qquad (3.1)$$

Έστω $y = (\mathrm{y}_1, \ldots\ldots, y_n)$ τα παρατηρούμενα δεδομένα ,$\theta = (\theta_1, \ldots \theta_k)$ διάνυσμα παραμέτρων k-διάστασης και $M_1, M_2, \ldots, M_m$ τα υποψήφια μοντέλα. Επίσης θεωρείται ως $f_i(y_i|\theta_i)$ η συνάρτηση πυκνότητας πιθανότητας για κάθε μοντέλο και $g_i(\theta_i)$ η prior πυκνότητα των παραμέτρων θ_i. Για την απόδειξη της (3.1) απαιτείται η χρήση βασικών εννοιών όπως είναι η προσέγγιση Laplace για ολοκληρώματα σύμφωνα με την οποία ισχύει:

$$\int_a^b e^M f(x)dx \approx \sqrt{\frac{2\pi}{M \mid f''(x_0)\mid}} e^{Mf(x_0)}, \quad \text{καθώς } M \to +\infty.$$

Το επόμενο απαραίτητο στοιχείο είναι να υπολογιστεί ο παράγοντας κατά Bayes ο οποίος χρησιμοποιείται για να επιλεγεί ένα μοντέλο ανάμεσα από δύο($\acute{\epsilon}\sigma\tau\omega\ M_0\ \kappa\alpha\iota\ M_1$) και ορίζεται ως:

$$B_{01}(y) = \frac{P(y|M_0)}{P(y|M_1)}$$

Η συνάρτηση πιθανοφάνειας για το i-οστό μοντέλο ορίζεται ως:

$$P(y|M_i) = \int f_i(y_i|\theta_i)g(\theta_i)d\theta_i \qquad (3.2)$$

Η εκτίμηση της παραπάνω ποσότητας μπορεί να οδηγήσει στην εκτίμηση του κριτηρίου BIC. Μία προσέγγιση της συνάρτησης πιθανοφάνειας $P(y|M_i)$ γίνεται με τη χρήση της μεθόδου Laplace για ολοκληρώματα. Η (3.2) μπορεί επίσης να γραφεί ως:

$$P(y|M_i) = \int \exp\{\log[f_i(y_i|\theta_i)]\}\, g(\theta_i)d\theta_i = \int \exp(l(\theta_i))g(\theta_i)d\theta_i$$

όπου $l(\theta_i)$ η λογαριθμική συνάρτηση πιθανοφάνειας. Προκειμένου να βρεθεί ο εκτιμητής της μέγιστης πιθανοφάνειας της $l(\theta_i)$, χρησιμοποιείται το ανάπτυγμα Taylor:

$$l(\theta_i) = \log(f(y|\theta_i)g(\theta_i) \approx \log(f(y|\tilde{\theta}_i)g(\tilde{\theta}_i) + (\theta_i - \tilde{\theta}_i)\nabla_{\theta_i}Q|_{\theta_i} + \frac{1}{2}(\theta_i - \tilde{\theta}_i)^T H_{\theta_i}(\theta_i - \tilde{\theta}_i)$$

όπου H_{θ_i} Εσσιανός πίνακας k x k, όπου k η διάσταση των παραμέτρων:

$$H_{mn} = \frac{\partial^2 Q}{\partial\theta_m \partial\theta_n}|_{\tilde{\theta}_i}$$

Από την στιγμή που το Q μεγιστοποιείται στο $\tilde{\theta}_i$, ο Εσσιανός πίνακας H_{θ_i} είναι αρνητικά ορισμένος.

Θέτοντας $H_{\theta_i} = -\widetilde{H_{\theta_\iota}}$ τότε μια προσέγγιση της συνάρτησης πιθανοφάνειας γίνεται:

$$P(y \mid M_i) \approx \int exp\left\{ Q|_{\tilde{\theta}_i} + (\theta_i - \tilde{\theta}_i)\nabla_{\theta_i} Q|_{\tilde{\theta}_i} - \frac{1}{2}(\theta_i - \tilde{\theta}_i)^T \tilde{H}_{\theta_i}(\theta_i - \tilde{\theta}_i)\right\} d\theta_i$$

Από την στιγμή που το Q μεγιστοποιείται στο $\tilde{\theta}_i$ έχουμε ότι:

$$\nabla_{\theta_i} Q|_{\tilde{\theta}_i} = 0$$

Αντικαθιστώντας στη παραπάνω σχέση και θεωρώντας ως $(\theta_i - \widetilde{\theta_\iota}) = X$:

$$P(y \mid M_i) \approx \exp(\mathrm{Q}|_{\tilde{\theta}_i}) \int exp\left\{ -\frac{1}{2}(\theta_i - \tilde{\theta}_i)^T \tilde{H}_{\theta_i}(\theta_i - \tilde{\theta}_i)\right\} d\theta_i =$$

$$= \exp(\mathrm{Q}|_{\tilde{\theta}_i}) \int exp\left\{ -\frac{1}{2} X^T \tilde{H}_{\vartheta_i} X\right\} dX$$

Επιπλέον ο πίνακας $\widetilde{H_{\theta_\iota}}$ είναι συμμετρικός, συνεπώς μπορεί να διαγωνιοποιηθεί έτσι ώστε:

$$\widetilde{H_{\theta_\iota}} = S^T \Lambda S = S^T U$$

Ο Ιακωβιανός πίνακας $J_{mn}(U) = \partial X_m / \partial U_n \Rightarrow J(U) = S^T$ και δεδομένου ότι $\det J(U) = 1$ και

$$P(y \mid M_i) \approx \exp(Q|_{\tilde{\theta}_\iota}) \int exp\{-\frac{1}{2} U^T \Lambda U\}(\det J(U)) dU = \exp\left(Q|_{\tilde{\theta}_\iota}\right) \int exp\left\{ -\frac{1}{2}\sum_{j=1}^{|\theta_i|} \lambda_j U_j^2 \right\} dU$$

$$= \exp\left(Q|_{\tilde{\theta}_\iota}\right) \prod_{j=1}^{|\theta_i|} \sqrt{\frac{2\pi}{\lambda_j}} = \exp\left(Q|_{\tilde{\theta}_\iota}\right) \frac{(2\pi)^{\frac{|\theta_i|}{2}}}{\prod_{j=1}^{|\theta_i|} \lambda_j^{\frac{1}{2}}} = f\left(y|\tilde{\theta}_\iota\right) g_i(\tilde{\theta}) \frac{(2\pi)^{\frac{|\theta_i|}{2}}}{\sqrt{|\widetilde{H_{\theta_\iota}}|}}$$

όπου λ_j η j-οστή ιδιοτιμή του πίνακα $\widetilde{H_{\theta_\iota}}$. Λογαριθμώντας την παραπάνω σχέση και πολλαπλασιάζοντας με 2 προκύπτει:

$$2\log P(y \mid M_i) = 2\log f(y \mid \tilde{\theta}_i) + 2\log g_i(\tilde{\theta}_i) + \mid \theta_i \mid \log(2\pi) + \log \mid \tilde{H}_{\theta_i}^{-1} \mid\mid \qquad (3.3)$$

Από την στιγμή που τα παρατηρούμενα μεγέθη δίνονται και υποθέτοντας ότι η εκτίμηση της μέγιστης πιθανοφάνειας είναι $\widehat{\theta_i} = \theta_i$ τότε αν θέσουμε $g_i(\theta i) = 1$ συνεπάγεται ότι:

$$\tilde{H}_{mn} = -\frac{\partial^2 \log L(\theta_i \mid y)}{\partial\theta_m \partial\theta_n}\Big|_{\theta_i=\hat{\theta}_i}$$

$$\tilde{H}_{mn} = -\frac{\partial^2 \log L(\theta_i \mid y)}{\partial\theta_m \partial\theta_n}\Big|_{\theta_i=\hat{\theta}_i} =$$

$$= -\frac{\partial^2 \log \prod_{j=1}^{n} L(\theta_i \mid y_j)}{\partial\theta_m \partial\theta_n}\Big|_{\theta_i=\hat{\theta}_i} =$$

$$= -\frac{\partial^2 \sum_{j=1}^{n} \log L(\theta_i \mid y_j)}{\partial\theta_m \partial\theta_n}\Big|_{\theta_i=\hat{\theta}_i} =$$

$$= -\frac{\partial^2 \, 1/n \left(\sum_{j=1}^{n} \mathrm{n} \log L(\theta_i \mid y_j) \right)}{\partial\theta_m \partial\theta_n}\Big|_{\theta_i=\hat{\theta}_i}$$

Σύμφωνα με τον ασθενή νόμο των μεγάλων αριθμών για τις τυχαίες μεταβλητές $X_j = n\log L(\theta_i \mid y_j)$ ισχύει :

$$\frac{1}{n}\sum_{j=1}^{n} n\log L(\theta_i \mid y_i) \xrightarrow{P} E\left[n\log L(\theta_i \mid y_j) \right]$$

Χρησιμοποιώντας την πιο πάνω ιδιότητα έχουμε ότι :

$$\tilde{H}_{mn} = -\frac{\partial^2 \, 1/n \left(\sum_{j=1}^{n} \mathrm{n} \log L(\theta_i \mid y_j) \right)}{\partial\theta_m \partial\theta_n}\Big|_{\theta_i=\hat{\theta}_i} =$$

$$= -\frac{\partial^2 E[\mathrm{n}\log L(\theta_i \mid y_j)]}{\partial\theta_m \partial\theta_n}\Big|_{\theta_i=\hat{\theta}_i} =$$

$$= -n\frac{\partial^2 E[\log L(\theta_i \mid y_1)]}{\partial\theta_m \partial\theta_n}\Big|_{\theta_i=\hat{\theta}_i} =$$

$$= nI_{mn}$$

Επομένως ,

$$|\tilde{H}_{\theta_i}| = n^{|\theta_i|} \, |I_{\theta_i}|$$

όπου I_{θ_i} είναι η πληροφορία κατά Fisher για ένα συγκεκριμένο σημείο y_1. Έτσι η (3.3) γίνεται:

$$\boldsymbol{BIC = -2logP(y|M_i) = -2\log L(\theta_i|y) + |\theta_i| logn} \qquad (3.4)$$

όπου k είναι η διάσταση του μοντέλου και n το μέγεθος του δείγματος. Ο δεύτερος όρος αποτελεί τον όρο ποινικοποίησης. Κατά την προσαρμογή μοντέλων είναι λογικό να αυξηθεί η πιθανοφάνεια καθώς προστίθενται και άλλες παράμετροι στο μοντέλο. Αυτό όμως μπορεί να οδηγήσει σε υπερπροσαρμογή (overfitting). Έτσι και στο AIC και στο BIC ο όρος ποινικοποίησης προσπαθεί να μειώσει αυτή την πολυπλοκότητα του μοντέλου. Ο όρος ποινής του BIC είναι μεγαλύτερος από εκείνον του AIC και επιπλέον πιο αυστηρός ως προς την εισαγωγή μεταβλητών. Και τα δύο κριτήρια αξιολόγησης μοντέλων εκτιμώνται μέσα από τη μέθοδο εκτίμησης της μέγιστης πιθανοφάνειας. Για τη χρήση του BIC απαραίτητος είναι ο προσδιορισμός της οριακής πιθανοφάνειας όπου η προσέγγισή της γίνεται με τη βοήθεια της μεθόδου Laplace. Αξιοσημείωτο είναι ότι το BIC χρησιμοποιείται για δείγματα μεγάλου μεγέθους n.

3.2 Τρόποι αξιολόγησης των κριτηρίων AIC και BIC

Κλασικοί τρόποι αξιολόγησης των κριτηρίων πληροφορίας με βάση τη διεθνή βιβλιογραφία είναι η συνέπεια, η ασθενής και ισχυρή συνέπεια και η αποδοτικότητα των κριτηρίων. Όλες αυτές οι έννοιες συνοδεύονται από θεωρήματα και αποδείξεις προκειμένου να γίνουν κατανοητές.

3.2.1 Συνέπεια - Φειδωλότητα

Ένας τρόπος αξιολόγησης των κριτηρίων είναι η συνέπεια τους η οποία χωρίζεται σε δύο περιπτώσεις. Εξετάζεται η ασθενής και ισχυρή συνέπεια των κριτηρίων η οποία αξιολογεί την ικανότητα του κριτηρίου να συγκλίνει κατά πιθανότητα ή να συγκλίνει σχεδόν βεβαίως στο πραγματικό μοντέλο όταν το πραγματικό μοντέλο περιλαμβάνεται μεταξύ των υποψήφιων μοντέλων. Συχνά όμως δεν θέλουμε να κάνουμε την υπόθεση ότι το πραγματικό μοντέλο ανήκει

μέσα στα υποψήφια μοντέλα. Σε αυτή την περίπτωση θέλουμε να υποθέσουμε ότι υπάρχει ένα υποψήφιο μοντέλο που είναι πιο κοντά από όλα στο πραγματικό μοντέλο υπό την έννοια της Kullback-Leibler απόστασης. Ένα κριτήριο λέγεται ασθενώς συνεπές αν με πιθανότητα που τείνει στο 1 επιλέγει αυτό το κοντινότερο (στο πραγματικό) μοντέλο ενώ ισχυρώς συνεπές θεωρείται εκείνο που συγκλίνει σχεδόν βεβαίως στο πραγματικό. Μας ενδιαφέρει επίσης ένα κριτήριο όχι μόνο να επιλέγει εκείνο το μοντέλο που ελαχιστοποιεί την απόσταση K-L αλλά να επιλέγει το μοντέλο εκείνο που είναι πιο απλό, δηλαδή με τις λιγότερες παραμέτρους. Αυτή είναι η αρχή της φειδωλότητας.

Όπως αναφέρθηκε και στο 2ο κεφάλαιο σύμφωνα με τον Atkinson(1980) η γενική εξίσωση του κριτηρίου πληροφορίας του (IC) είναι η ακόλουθη:

$$\boldsymbol{IC = -2l + A_n k}$$

Όπου

l: λογαριθμική συνάρτηση πιθανοφάνειας

$\mathrm{A_n}$: συνάρτηση του μεγέθους n του δείγματος

k: ο αριθμός των παραμέτρων στο μοντέλο

Έστω θ_k το διάνυσμα παραμέτρων και $f_{k,i}$ η συνάρτηση πυκνότητας για την i παρατήρηση για το k-οστό μοντέλο από το σύνολο υποψήφιων μοντέλων (k=1,..,K). Προκειμένου να αποδειχθούν παρακάτω, τα θεωρήματα συνέπειας, η γενική εξίσωση και για τα δύο κριτήρια παίρνει τη μορφή:

$$IC(M_k) = 2\sum_{i=1}^{n} \log f_{k,i}\left(y_i, x_i; \hat{\theta}_k\right) - C_{n,k}$$

όπου $C_{n,k}$ ο όρος ποινής για το μοντέλο M_k ο οποίος είναι πάντα θετικός και συγκεκριμένα ισχύει:

$$C_{n,k} = \begin{cases} 2dim\theta \ για \ το \ AIC \\ logn \ dim(\theta) \ για \ το \ BIC \end{cases}$$

και $\hat{\theta}_k$ η εκτιμήτρια μέγιστης πιθανοφάνειας. Αξιοσημείωτο είναι ότι ο όρος ποινικοποίησης στο BIC είναι μεγαλύτερος για $n \geq 9$. Αυτό εποπτικά τονίζει ότι το μοντέλο BIC δεν προτιμά την επιλογή μοντέλων με πολλές μεταβλητές σε σχέση με το AIC.

Παρακάτω παρουσιάζονται τέσσερα θεωρήματα για τη συνέπεια. Τα Θεωρήματα 3.1 και 3.2 αφορούν την περίπτωση που το πραγματικό μοντέλο δεν περιλαμβάνεται μεταξύ των

υποψηφίων μοντέλων ενώ υπάρχει μόνο ένα μοντέλο που είναι κοντύτερα στο πραγματικό με βάση την απόσταση KL. Τα Θεωρήματα 3.3 και 3.4 αφορούν την περίπτωση που υπάρχουν περισσότερα από ένα μοντέλα που ελαχιστοποιούν την απόσταση KL.

Θεώρημα 3.1 (Ασθενής Συνέπεια-Weak Consistency)

Έστω ότι υπάρχει ένα ακριβώς μοντέλο M_{k_0} ανάμεσα στα υποψήφια μοντέλα το οποίο ελαχιστοποιεί την K-L απόκλιση. Γι' αυτό το μοντέλο ισχύει δηλαδή ότι

$$\liminf_{n\to\infty} \min_{k\neq k_0} \frac{1}{n}\sum_{i=1}^{n}\left(KL(g;\, f_{k,i}) - KL(g;\, f_{k_0,i}\right) > 0$$

Έστω ότι ο θετικός όρος ποινικοποίησης είναι $o_p(n)$.

Τότε το κριτήριο πληροφορίας επιλέγει με πιθανότητα που τείνει στο 1 το μοντέλο M_{k_0} ως το βέλτιστο.

Έτσι εφαρμόζοντας το παραπάνω θεώρημα για τα κριτήρια AIC και BIC προκύπτει:

$$AIC:\ \frac{C_{n,k}}{n} = \lim_{n\to+\infty}\frac{2}{n}\dim(\theta) = 0$$

$$BIC:\ \frac{C_{n,k}}{n} = \lim_{n\to+\infty}\frac{logn}{n}\dim(\theta) = 0$$

Συνεπώς και τα δύο κριτήρια είναι ασθενώς συνεπή (Claeskens, Gerda, Nils Lid Hjort, 2008).

Θεώρημα 3.2 (Ισχυρή συνέπεια-Strong Consistency)

Έστω ότι υπάρχει ακριβώς ένα μοντέλο M_{k_0} ανάμεσα στα υποψήφια μοντέλα το οποίο ελαχιστοποιεί την K-L απόκλιση. Δεδομένου ότι ικανοποιούνται οι παρακάτω προϋποθέσεις

1. $\liminf_{n\to\infty} \min_{k\neq k_0} \frac{1}{n}\sum_{i=1}^{n}\left(KL(g;\, f_{k,i}) - KL(g;\, f_{k_0,i}\right) > 0$
2. Ο θετικός όρος ποινικοποίησης είναι $o(n)$ σχεδόν βεβαίως

Τότε $P\left(\min_{i\neq k_0}(IC(M_{k_0}) - IC(M_i)) > 0,\ \text{για σχεδόν όλα τα n}\right) = 1$.

Όπως αποδείχθηκε προηγουμένως η σχέση για τους όρους ποινικοποίησης ικανοποιείται και για τα δύο κριτήρια. Άρα τα AIC και BIC τα οποία προσπαθούν να επιλέξουν το μοντέλο με τη μικρότερη απόσταση K-L είναι **Ισχυρά Συνεπή** (Claeskens, Gerda, Nils Lid Hjort, 2008).

Θεώρημα 3.3- Ασθενής Συνέπεια (Weak Consistency)

Έστω I το σύνολο των μοντέλων που αντιστοιχούν στην ελάχιστη απόσταση K-L και I_0 το υποσύνολο του I που αντιστοιχεί στα μοντέλα που έχουν τις λιγότερες παραμέτρους. Κάτω από την υπόθεση (1) ή (2) όπου

1. Για κάθε $k_0 \neq i_0 \in I$ ισχύει ότι:

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(KL(g; f_{k_0,i}) - KL(g; f_{i_0,i}) \right) < \infty$$

Για $\forall i_0 \in I_0, \forall k \in I \setminus I_0,\ P\left(\frac{C_{n,i} - C_{n,i_0}}{\sqrt{n}} \to \infty \right) = 1$

2. Για κάθε $k_0 \neq i_0 \in I$ έστω ότι ισχύει ότι ο λόγος της λογαριθμικής πιθανοφάνειας:

$$\sum_{i=1}^{n} log \frac{f_{k0,i}(y_i;\theta^*_{k0})}{f_{i_0,i}\left(y_i;\theta^*_{i_0}\right)} = O_p(1)$$

και για τους όρους ποινικοποίησης ισχύει ότι για:

και για κάθε $i_0 \in I_0, για\ κάθε\ i \in I \setminus I_0\ τότε\ P\left((C_{n,i} - C_{n,i_0}) \to \infty\right) = 1$

Τότε με πιθανότητα που τείνει στο 1 το κριτήριο πληροφορίας θα διαλέξει ένα μοντέλο που ελαχιστοποιεί την απόσταση K-L και θα χαρακτηρίζεται ως φειδωλό, δηλαδή θα έχει τις λιγότερες μεταβλητές. Άρα θα ισχύει:

$$\lim_{n\to+\infty} P\{min_{i \in I \setminus I_0}\left(IC(M_{i0}) - IC(M_i)\right) > 0\} = 1$$

Έστω $C_{n,i}, C_{n,i_0}$ οι όροι ποινικοποίησης για τα μοντέλα M_i, M_{i0} αντίστοιχα οι οποίοι για το κριτήριο AIC έχουν τις εξής τιμές:

$$C_{n,i} = 2\dim(\theta_i) \ \ \& \ \ C_{n,i_0} = 2\dim(\theta_{i_0})$$

Τότε το όριο της διαφοράς των δύο όρων ποινής ισούται με μία σταθερά c:

$$\lim_{n\to+\infty} (C_{n,i} - C_{n,i_0}) = \lim_{n\to+\infty} (dim\,\theta_i - dim\theta_{i_0}) = c$$

Άρα,

$$P(\lim_{n\to+\infty}(dim\,\theta_i - dim\theta_{i_0})) < 1$$

Επομένως το κριτήριο πληροφορίας AIC δεν είναι συνεπές.

Για το BIC έχουμε:

$$C_{n,i} = logn2\dim(\theta_i)\ \&\ C_{n,i_0} = logn2\dim(\theta_{i_0})$$

Με $2\dim(\theta_i) > \dim(\theta_{i_0})$ και $\lim_{n\to+\infty}(C_{n,i} - C_{n,i_0}) = \lim_{n\to+\infty}(dim\,\theta_i - dim\theta_{i_0}) = +\infty$

Επομένως $P(\lim_{n\to+\infty}(C_{n,i} - C_{n,i_0}) = +\infty) = 1$. Άρα το BIC είναι συνεπές (Claeskens, Gerda, Nils Lid Hjort, 2008).

Θεώρημα 3.4- Ισχυρή Συνέπεια (Strong Consistency)

Έστω I το σύνολο των μοντέλων που αντιστοιχούν στην ελάχιστη απόσταση K-L και I_0 το υποσύνολο του I που αντιστοιχεί στα μοντέλα που έχουν τις λιγότερες παραμέτρους.

1. Για κάθε $k_0 \neq i_0 \in I$ ισχύει ότι:

$$\limsup_{n\to\infty}\frac{1}{\sqrt{n\log\log n}}\sum_{i=1}^{n}\left(KL(g;f_{k_0,i}) - KL(g;f_{i_0,i})\right) \leq 0,$$

2. Για κάθε $k_0 \neq i_0 \in I$ ισχύει ότι ο λόγος της λογαριθμικής πιθανοφάνειας:

$$\sum_{i=1}^{n} log\frac{f_{k0,i}(y_i;\theta^*_{k0})}{f_{i_0,i}\left(y_i;\theta^*_{i_0}\right)} = o(loglogn)\ \sigma.\beta.$$

Τότε η ισχυρή συνέπεια του IC εξασφαλίζεται αρκεί ο όρος ποινικοποίησης να ικανοποιεί την ακόλουθη σχέση όπου b_n μια τυχαία ακολουθία φραγμένη από κάτω από θετικό αριθμό:

$$P\left(C_{n,k} \geq b_n \log\log n\ \ for\ almost\ all\ n\right) = 1.$$

3.2.2 Αποδοτικότητα

Ένα τελευταίο κριτήριο που χρησιμοποιείται για την αξιολόγηση των δύο κριτηρίων πληροφορίας AIC και BIC είναι η αποδοτικότητα. Συγκεκριμένα γίνεται σύγκριση με το θεωρητικά καλύτερο μοντέλο που επιλέχθηκε με βάση το τετραγωνικό σφάλμα θεωρητικής πρόβλεψης. Εάν η επιλογή το μοντέλου πρόβλεψης που γίνεται με τη βοήθεια των κριτηρίων πληροφορίας συμβαδίζει με την επιλογή του θεωρητικά κατάλληλου μοντέλου που προτείνει

το μέσο τετραγωνικό σφάλμα πρόβλεψης τότε το κριτήριο πληροφορίας μπορεί να χαρακτηριστεί ως αποδοτικό (efficient).

Έστω ότι μελετάμε το γενικό γραμμικό μοντέλο:

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad i = 1, \ldots, n}$$

Θέτοντας $S = P(x)$ το δυναμοσύνολο του συνόλου των επεξηγηματικών μεταβλητών, στόχος μας είναι να διαλέξουμε από το σύνολο των επεξηγηματικών μεταβλητών, ένα υποσύνολο $s \subset S$ το οποίο θα ελαχιστοποιεί το μέσο τετραγωνικό σφάλμα πρόβλεψης. Έστω $\widehat{Y_i}$ η προβλεπόμενη τιμή της μεταβλητής απόκρισης Y_i. Επιδιώκουμε να ελαχιστοποιήσουμε την ποσότητα:

$$\sum_{i=1}^{n} E[(\hat{Y}_{s,i} - Y_{t,i}) | \, y_{1,\ldots,n}]$$

Όπου $\hat{Y}_{s,i}$: οι προβλεπόμενες τιμές για τα διάφορα s υποσύνολα.

$Y_{t,i}$: οι πραγματικές προβλεπόμενες τιμές

Έστω $s_0 \subset S$ το σύνολο των επεξηγηματικών μεταβλητών του μοντέλου που επιλέχθηκαν από το κριτήριο πληροφορίας και έστω $s_0^* \subset S$ το σύνολο των επεξηγηματικών μεταβλητών όπου ελαχιστιποιείται από το μέσο σφάλμα πρόβλεψης. Σύμφωνα με τον ορισμό της αποδοτικότητας, ένα κριτήριο πληροφορίας καλείται αποδοτικό εάν επιλέγει εκείνο το μοντέλο όπου ο λόγος της αναμενόμενης συνάρτησης απώλειας ως προς την ελάχιστη θεωρητικά αναμενόμενη απώλεια τείνει κατά πιθανότητα στο 1. Δηλαδή

$$\frac{\sum_{i=1}^{n} E[\left((\hat{Y}_{s_0,i} - Y_{t,i}\right)^2]}{\sum_{i=1}^{n} E[\left((\hat{Y}_{s_0^*,i} - Y_{t,i}\right)^2]} = \frac{Ln(s_0)}{Ln(s_0^*)} \to 1 \; \gamma\iota\alpha \; n \to +\infty$$

Σύμφωνα με το παρακάτω θεώρημα φαίνεται ότι τα κριτήρια AIC, AICc είναι αποδοτικά ενώ το BIC δεν είναι.

Θεώρημα (Αποδοτικότητα)

Έστω s_0 το σύνολο επεξηγηματικών μεταβλητών που επιλέχθηκαν για την ελαχιστοποίηση του κριτηρίου όταν αυτό παίρνει τη μορφή

$$IC(s) = (n + 2 \,|\, s \,|)\sigma_s^2$$

αν ισχύει ότι:

1. Για όλα τα διαφορετικά υποσύνολα επεξηγηματικών μεταβλητών s του δυναμοσυνόλου S ο αντίστοιχος πίνακας σχεδιασμού X_s έχει μέγιστο βαθμό |s|, όπου |s| αντιστοιχεί στον αριθμό μεταβλητών του συνόλου s και $\max_s |s| = O(n^b), b \in [0,1)$.

2. Για κάθε $a \in \left[0, \frac{1}{2}\right)$ και για κάθε $c > 0$ πρέπει να ισχύει

$$\lim_{n\to\infty} \sum_s e^{n^{-2a} L_n(s)} \to 0,$$

κάτι που υποδεικνύει την ύπαρξη μοναδικότητας.

3. Τα κριτήρια πληροφορίας έχουν τη μορφή

$$IC(s) = (n + 2|s|)\sigma_s^2$$

με την εκτίμηση της διασποράς να είναι:

$$\frac{SSE}{n} = \frac{1}{n}(Y - X_s b_s)^T (Y - X_s b_s)$$

Τα κριτήρια AIC και BIC μπορούν να γραφούν σε αυτή τη μορφή.

Εάν ισχύουν οι παραπάνω τρεις ιδιότητες τότε για το σύνολο των επεξηγηματικών μεταβλητών s_0^* που ελαχιστοποιούν την $L_n(s)$ και θεωρώντας ως $C = \min\{\frac{1-b}{2}, a\}$ έχουμε ότι:

$$\frac{L_n(s_0)}{L_n(s_0^*)} - 1 = O(\frac{1}{n^c})$$

Εξετάζοντας τα κριτήρια ως προς το θεώρημα της αποδοτικότητας προκύπτει:

Τα κριτήρια AIC και $AICc(s) = AIC(s) + \frac{2(|s|+1)(|s|+2)}{n-|s|+2}$ σύμφωνα με τις παραπάνω υποθέσεις είναι ασυμπτωτικά αποδοτικά ενώ το BIC δεν είναι. Τα AIC, AICc λοιπόν τείνουν να επιλέξουν ως στατιστικά καλύτερο μοντέλο εκείνο που επιλέγεται και σαν καταλληλότερο από την ελαχιστοποίηση του αναμενόμενου τετραγωνικού σφάλματος πρόβλεψης (Claeskens, Gerda, Nils Lid Hjort, 2008).

Κεφάλαιο 4ο

Εφαρμογή των κριτηρίων AIC και BIC με τη χρήση της R

4.1 Μέθοδοι Επιλογής Μεταβλητών

Η επιλογή μεταβλητών (variable selection) θεωρείται ένα απαραίτητο μέσο προκειμένου να επιτευχθεί το βέλτιστο μοντέλο. Μεταξύ ενός πλήθους ανεξάρτητων μεταβλητών προσδιορίζεται το καλύτερο υποσύνολο μεταβλητών οι οποίες θεωρούνται στατιστικά σημαντικές. Όπως και στην ανάλυση των κριτηρίων AIC και BIC ισχύει η αρχή της φειδωλότητας (law of parsimony) η οποία υποστηρίζει ότι η απλούστερη λύση είναι και η καλύτερη, συνεπώς το απλούστερο μοντέλο είναι και το πιο κατάλληλο. Στην πολλαπλή παλινδρόμηση στόχος των αναλυτών είναι να αποφασιστεί ποιες ανεξάρτητες μεταβλητές συνεισφέρουν σημαντικά στη μεταβλητότητα της εξαρτημένης μεταβλητής. Για την επιλογή του καλύτερου μοντέλου εκτός από τα κριτήρια πληροφορίας μπορεί να εξεταστεί το άθροισμα τετραγωνικών υπολοίπων (residual sum of squares-RSS) ή ο συντελεστής προσδιορισμού r^2 ο οποίος μετρά το ποσοστό της συνολικής μεταβλητότητας των y_i που εξηγείται από τις ανεξάρτητες μεταβλητές. Ανάμεσα σε ένα πλήθος μοντέλων καλύτερο θεωρείται εκείνο με τη μικρότερη τιμή RSS, το οποίο οδηγεί στη μεγιστοποίηση του r^2. Το μοντέλο με τις περισσότερες ανεξάρτητες μεταβλητές συνοδεύεται και με τη μεγαλύτερη τιμή του συντελεστή προσδιορισμού, r^2 χωρίς όμως όλες αυτές οι μεταβλητές να είναι σημαντικές.

Παρακάτω αναλύονται τρεις μέθοδοι επιλογής μεταβλητών που χρησιμοποιούνται στη Στατιστική προκειμένου να καθοριστεί ένα υποσύνολο ανεξάρτητων μεταβλητών που θα είναι χρήσιμο για τη πρόβλεψη της εξαρτημένης μεταβλητής. Η μέθοδος προσθήκης μεταβλητών (forward selection), η μέθοδος αφαίρεσης μεταβλητών (backward selection) και η μέθοδος βηματικής παλινδρόμησης (stepwise regression) μπορεί να έχουν κοινό στόχο αλλά είναι φυσικό να οδηγήσουν σε διαφορετικά υποσύνολα μεταβλητών που πιθανόν δεν θα διαφέρουν σημαντικά.

4.1.1 Μέθοδος Αποκλεισμού Μεταβλητών (Backward Elimination Procedure)

Μία απλή μέθοδος για την επιλογή του καλύτερου υποσυνόλου μεταβλητών σε ένα μοντέλο πολλαπλής παλινδρόμησης είναι η μέθοδος αποκλεισμού μεταβλητών. Η μέθοδος αυτή στηρίζεται στο γεγονός ότι αρχικά εξετάζεται ολόκληρο το μοντέλο και σε κάθε βήμα ξεχωριστά αποκλείεται μία μεταβλητή η οποία δεν συνεισφέρει σημαντικά σε αυτό δηλαδή

εκείνη που έχει το μεγαλύτερο επίπεδο σημαντικότητας (p-value) με την προϋπόθεση ότι υπερβαίνει το επίπεδο σημαντικότητας α που έχει καθοριστεί. Στον παρακάτω πίνακα παρατίθεται ένα μέρος των προσομοιώσεων που θα εξετάσουμε αναλυτικά στη συνέχεια. Δεδομένου ότι το επίπεδο σημαντικότητας είναι $\alpha = 0.05$, στο πρώτο στάδιο της μεθόδου έχει αποκλειστεί η μεταβλητή z_3 καθώς έχει τη μεγαλύτερη τιμή σημαντικότητας (p-value=0.4362) από αυτές που είναι μεγαλύτερες από το επίπεδο σημαντικότητας α. Ανάλογα στο δεύτερο βήμα αφαιρείται η μεταβλητή z_2 κι έτσι προκύπτει το τελικό βέλτιστο μοντέλο το οποίο περιέχει τις ανεξάρτητες μεταβλητές z_1,z_4 των οποίων οι τιμές των παρατηρούμενων επιπέδων σημαντικότητας είναι μικρότερες από 0.05. Ένας άλλος τρόπος για να αποφασιστεί ποια μεταβλητή θα αφαιρεθεί σε κάθε βήμα είναι η τιμή του AIC. Η τιμή του AIC που αναγράφεται σε κάθε μεταβλητή ξεχωριστά αντιπροσωπεύει την τιμή που θα είχε το μοντέλο αν αφαιρούνταν η συγκεκριμένη μεταβλητή. Δεδομένου ότι το μοντέλο με τη μικρότερη τιμή AIC θεωρείται το βέλτιστο, είναι λογικό σε κάθε βήμα να αφαιρείται εκείνη η μεταβλητή στην οποία αντιστοιχεί η μικρότερη τιμή του AIC. Ενδεχομένως διαφορετικές μέθοδοι επιλογής μεταβλητών να καταλήξουν σε διαφορετικό υποσύνολο μεταβλητών που θα περιληφθούν στο τελικό μοντέλο. Συνήθως όμως επικρατεί συμφωνία μεταξύ των διάφορων μεθόδων. Αξιοσημείωτο στη συγκεκριμένη μέθοδο είναι ότι υπάρχει σημαντική διαφορά μεταξύ των υποψήφιων μεταβλητών στο πρώτο βήμα της μεθόδου και αυτών που τελικά επικρατούν στο τελικό μοντέλο. Αυτό εξηγείται καθώς η αφαίρεση μιας μεταβλητής μεταβάλλει αρκετά τα παρατηρούμενα επίπεδα σημαντικότητας των υπολοίπων μεταβλητών. Η μέθοδος αυτή χρησιμοποιήθηκε στις προσομοιώσεις αυτού του κεφαλαίου και στην εικόνα 4.1 παρουσιάζεται ένα απόσπασμα από την R όπου φαίνεται αναλυτικά η χρήση της μεθόδου:

```
> step<-stepAIC(fit,direction="backward",k=2)
Start:  AIC=18.66
y ~ z1 + z2 + z3 + z4

       Df Sum of Sq    RSS    AIC
- z2    1       0.7  109.8  17.32
- z3    1       1.1  110.1  17.65
<none>              109.0  18.66
- z4    1      16.7  125.8  30.92
- z1    1    9083.3 9192.3 460.10

Step:  AIC=17.32
y ~ z1 + z3 + z4

       Df Sum of Sq    RSS    AIC
- z3    1       0.7  110.5  15.97
<none>              109.8  17.32
- z4    1      16.0  125.8  28.96
- z1    1    9082.6 9192.3 458.10

Step:  AIC=15.97
y ~ z1 + z4

       Df Sum of Sq    RSS    AIC
<none>              110.5  15.97
- z4    1      15.9  126.4  27.41
- z1    1    9082.0 9192.5 456.10
```

Εικόνα 4.1: Η επιλογή ενός μοντέλου μέσα από τη μέθοδο backward με τη χρήση του κριτηρίου AIC και του στατιστικού πακέτου R

4.1.2 Μέθοδος Προσθήκης Μεταβλητών(Forward Procedure)

Η μέθοδος αυτή βασίζεται σε ακριβώς αντίθετη λογική της μεθόδου απαλοιφής μεταβλητών που αναπτύχθηκε παραπάνω. Το μοντέλο τώρα αναπτύσσεται με την προσθήκη μίας ανεξάρτητης μεταβλητής σε κάθε βήμα. Στο πρώτο βήμα της μεθόδου επιλέγεται ανάμεσα από ένα πλήθος ανεξάρτητων μεταβλητών, εκείνη που έχει το μεγαλύτερο συντελεστή συσχέτισης με την εξαρτημένη μεταβλητή Y. Στη συνέχεια υπολογίζεται η τιμή της στατιστικής συνάρτησης T για τον εξής έλεγχο:

$$H_0: b_1 = 0 \; \text{έναντι} \; H_1: b_1 \neq 0$$

όπου b_1 είναι ο συντελεστής της μεταβλητής που έχει προστεθεί στο μοντέλο. Εάν η τιμή της στατιστικής συνάρτησης βρίσκεται μέσα στη κρίσιμη περιοχή τότε υπάρχουν ενδείξεις για να απορριφθεί η H_0. Δηλαδή αν $|T| > t_{n-2,1-\frac{a}{2}}$ τότε η διαδικασία συνεχίζεται αλλιώς σταματάει εδώ. Στο δεύτερο βήμα επιλέγεται η ανεξάρτητη μεταβλητή με το μεγαλύτερο συντελεστή μερικής συσχέτισης κατά απόλυτη τιμή και υπολογίζεται όπως και πριν η τιμή της ελεγχοσυνάρτησης T μέσα από τον κατάλληλο έλεγχο υποθέσεων και εξετάζεται αν θα

προστεθεί η μεταβλητή στο συγκεκριμένο μοντέλο η θα διακοπεί η διαδικασία. Έτσι υπολογίζονται διαδοχικά συντελεστές μερικής συσχέτισης μεγαλύτερης τάξης σε κάθε βήμα, προσθέτοντας κάθε φορά στο μοντέλο τη μεταβλητή με την μεγαλύτερη απόλυτη τιμή συντελεστή μερικής συσχέτισης με το Y διατηρώντας σταθερές όλες τις μεταβλητές που έχουν προστεθεί μέχρι εκείνη τη στιγμή στο μοντέλο. Η διαδικασία σταματάει όταν η τιμή της στατιστικής συνάρτησης $|T|$ για την εξεταζόμενη μεταβλητή δεν υπερβεί κάποιο προκαθορισμένο κρίσιμο σημείο.

4.1.3 Μέθοδος της Βηματικής Παλινδρόμησης (Stepwise Regression)

Η μέθοδος της βηματικής παλινδρόμησης είναι παρόμοια με τη μέθοδο προσθήκης μεταβλητών με τη διαφορά ότι τώρα σε κάθε διαδοχικό βήμα ο έλεγχος υποθέσεων $H_0: b_i = 0$ γίνεται για όλες τις ανεξάρτητες μεταβλητές και όχι μόνο για μία. Από ένα σύνολο λοιπόν ανεξάρτητων μεταβλητών που είναι υποψήφιες να συμπεριληφθούν στο μοντέλο, επιλέγουμε εκείνη που έχει το μεγαλύτερο συντελεστή συσχέτισης με την εξαρτημένη μεταβλητή. Αξιοσημείωτο είναι ότι εκτός από τις ανεξάρτητες μεταβλητές που θεωρήθηκαν αρχικά, μπορεί να προκύψουν και συναρτήσεις μεταβλητών όπως είναι η Z_i^2 ή η αλληλεπίδραση $Z_i Z_j$. Στο πρώτο βήμα λοιπόν της μεθόδου επιλέγεται ως ανεξάρτητη μεταβλητή εκείνη που έχει κατά απόλυτη τιμή μεγαλύτερο συντελεστή συσχέτισης με την εξαρτημένη μεταβλητή. Προκειμένου να συνεχιστεί η διαδικασία θα πρέπει η τιμή του παρατηρούμενου επίπεδου σημαντικότητας να είναι μικρότερη από το καθορισμένο επίπεδο σημαντικότητας (α=0.05) αλλιώς η ανάλυση σταματάει σε αυτό το σημείο. Στο επόμενο βήμα, αφού λοιπόν έχει προστεθεί μία μεταβλητή στο μοντέλο εξετάζουμε αν υπάρχει μοντέλο με δύο ανεξάρτητες μεταβλητές, η μία από τις οποίες είναι αυτή που επιλέχθηκε από το πρώτο βήμα και ως δεύτερη θα εξεταστεί κάθε μεταβλητή ξεχωριστά από αυτές που έχουν απομείνει. Για κάθε μοντέλο από αυτά ελέγχεται η υπόθεση:

$$H_0: b_i = 0 \quad \text{έναντι} \quad H_1: b_i \neq 0$$

όπου b_i οι συντελεστές των μεταβλητών Z_i, μία από τις οποίες θα αποτελέσει τη δεύτερη από τις 2 ανεξάρτητες μεταβλητές που εξετάζουμε στο δεύτερο βήμα. Στη συνέχεια εξετάζεται ποια από τις Z_i μεταβλητές έχει το μεγαλύτερο επίπεδο σημαντικότητας και αν αυτό είναι μικρότερο από την τιμή του καθορισμένου επιπέδου σημαντικότητας α τότε και αυτή η μεταβλητή πρέπει να συμπεριληφθεί στο αρχικό μοντέλο αλλιώς η διαδικασία σταματάει. Εάν το 2^{ο} βήμα ολοκληρωθεί με επιτυχία, η βηματική παλινδρόμηση συνεχίζεται όπως ακριβώς στο προηγούμενο βήμα και εξετάζεται αν θα προστεθεί και άλλη μεταβλητή στο μοντέλο, ελέγχοντας ξανά την απόλυτη τιμή της ελεγχοσυνάρτησης |T| μέσα από τον παραπάνω έλεγχο υποθέσεων. Εξετάζεται δηλαδή αν υπάρχει μεταβλητή που έχει το μικρότερο επίπεδο σημαντικότητας ή αλλιώς επιλέγεται εκείνη η μεταβλητή που έχει το μεγαλύτερο συντελεστή συσχέτισης με την εξαρτημένη μεταβλητή του μοντέλου. Αξιοσημείωτο στη μέθοδο της

βηματικής παλινδρόμησης σε αντίθεση με τις μεθόδους προσθήκης και αφαίρεσης μεταβλητών είναι ότι κάθε φορά που προστίθεται μία νέα μεταβλητή στο μοντέλο γίνονται έλεγχοι υποθέσεων :

$$H_0: b_j = 0 \quad \acute{\epsilon}\nu\alpha\nu\tau\iota \quad H_1: b_j \neq 0$$

για όλες τις ανεξάρτητες μεταβλητές που έχουν ήδη συμπεριληφθεί στο μοντέλο και αν κάποια από αυτές παύει να είναι πλέον στατιστικά σημαντική υπάρχει η δυνατότητα διαγραφής της. Απαραίτητο λοιπόν είναι να ορισθούν από την αρχή δύο επίπεδα σημαντικότητας, $\alpha_1 \, \kappa\alpha\iota \, \alpha_2$, όπου το πρώτο χρησιμοποιείται για να εξεταστεί ποια νέα μεταβλητή θα προστεθεί στο μοντέλο και το δεύτερο χρησιμοποιείται για τους ελέγχους για τις μεταβλητές που έχουν ήδη συμπεριληφθεί. Αναγκαίος περιορισμός για να λειτουργήσει σωστά η διαδικασία είναι $\alpha_2 \geq \alpha_1$.

4.2 Προσομοιώσεις

Σε αυτό το κεφάλαιο μελετάται η χρήση των κριτηρίων AIC και BIC για την επιλογή μοντέλων μέσω προσομοιώσεων στο στατιστικό πακέτο R και εξάγονται συμπεράσματα για το ποιο από τα δύο κριτήρια έχει πιο αξιόπιστα αποτελέσματα. Χρησιμοποιείται το γενικό μοντέλο παλινδρόμησης της μορφής

$$\boldsymbol{y_i = b_o + b_1 \, z_{i\,1} + b_2 z_{i\,2} + b_3 z_{i\,3} + b_4 z_{i\,4} + e_i} \quad \textbf{για i=1,2,…,n}$$

όπου y_i είναι η μεταβλητή απόκρισης, b_1, b_1, b_3, b_4 είναι οι συντελεστές παλινδρόμησης και e_i το σφάλμα παλινδρόμησης που υποθέτουμε ότι ακολουθεί την κανονική κατανομή $N(0,1)$. Στις προσομοιώσεις εξετάζονται δύο σύνολα των συντελεστών με τιμές $(1, 1.8, 0, 0,1), (1, 0.8, 0, 0,1)$και$(1, 1.8, 0,0, 0.2)$ως πραγματικές τιμές. Οι συμμεταβλητές $z_1 \, \kappa\alpha\iota \, z_4$ λαμβάνονται να είναι σημαντικές οπότε θα περιμέναμε τα περισσότερα μοντέλα που επιλέγονται από τα κριτήρια να τις περιέχουν.

Επίσης δημιουργούνται τέσσερις κατηγορίες προσομοιώσεων όπου στην πρώτη, οι συμμεταβλητές είναι ανεξάρτητες μεταξύ τους, στις άλλες δύο περιπτώσεις είναι εξηρτημένες ασθενέστερα ή ισχυρότερα με συντελεστές συσχέτισης που βασίζονται σε συντελεστή συσχετίσεως rho $(rho = 0.5 \, \kappa\alpha\iota \, rho = 0.8)$ που θα εξηγήσουμε αναλυτικά πιο κάτω και στη τέταρτη εξετάζεται το σύνολο των συντελεστών $(1, 1.8, 0,0, 0.2)$ για κάθε συντελεστή συντελεστή συσχέτισης ξεχωριστά. Σε κάθε κατηγορία υπολογίζεται το ποσοστό των φορών που επιλέγεται το σωστό μοντέλο, το ποσοστό των φορών που επιλέγεται μία μη σημαντική μεταβλητή (σφάλμα τύπου I) και το ποσοστό των φορών που δεν επιλέγεται μία μη σημαντική μεταβλητή (σφάλμα τύπου II) . Έτσι συγκρίνονται τα δύο κριτήρια πληροφορίας, AIC και BIC, εξάγονται συμπεράσματα για το πιο τελικά έχει πιο αξιόπιστα αποτελέσματα για δείγματα

διαφορετικού μεγέθους, διαφορετικό συντελεστή συσχέτισης και σύνολο πραγματικών συντελεστών. Τα παραπάνω σφάλματα ορίζονται πιο συγκεκριμένα ως εξής:

$$\text{error I} = \begin{cases} 1, \text{αν επιλεγεί τουλάχιστον μία μη σημαντική μεταβλητή} \\ 0, \text{αλλιώς} \end{cases}$$

$$\text{error II} = \begin{cases} 1, \text{αν δεν επιλεγεί τουλάχιστον μία σημαντική μεταβλητή} \\ 0, \text{αλλιώς} \end{cases}$$

<u>Προσομοίωση 1η</u>

Στην πρώτη προσομοίωση οι συμμεταβλητές είναι ανεξάρτητες. Οι ανεξάρτητες μεταβλητές z_1, z_2, z_3, z_4 ακολουθούν τις εξής κατανομές:

$\boldsymbol{z_1 \sim N(0,5)}$

$\boldsymbol{z_2 \sim b(1,0.5)}$

$\boldsymbol{z_3 \sim N(1,5)}$

$\boldsymbol{z_4 \sim b(1,0.7)}$

Δημιουργούνται 100 δείγματα μεγέθους n και εξετάζονται οι περιπτώσεις όπου $n = 25, 50, 100 \; και \; 200$. Έτσι συγκρίνονται τα δύο κριτήρια πληροφορίας σε περιπτώσεις που το δείγμα είναι αρκετά μεγάλο και αντίστοιχα αρκετά μικρό. Αφού δηλωθούν οι τιμές των συντελεστών, κατασκευάζονται οι ανεξάρτητες συμμεταβλητές των δειγμάτων από την κανονική κατανομή και από τη διωνυμική κατανομή με τις εντολές rnorm και rbinom. στην R:

```
for(i in 1:100)
{
print(i)
z1<-rnorm(100,0,5)
z2<-rbinom(100,1,0.5)
z3<-rnorm(100,1,5)
z4<-rbinom(100,1,0.7)
eps<-rnorm(100,0,1)

y <- b0 + b1*z1 + b2*z2  + b3*z3 + b4*z4 + eps
#simdata<- data.frame(z1,z2,z3,z4,eps,y)
#simdata[10,]

fit<-lm(y~z1+z2+z3+z4)
summary(fit)
results <-coef(fit)
```

Εικόνα 4.2: Εντολές για την κατασκευή των τυχαίων ανεξάρτητων συμμεταβλητών, απόσπασμα από τον κώδικα στην R

Επίσης με την εντολή summary παρουσιάζεται περιληπτικά το πόσο καλά προσαρμόζεται το μοντέλο καθώς δίνονται οι εκτιμήσεις όλων των συντελεστών των συμμεταβλητών και σύμφωνα από τα p-value φαίνεται ποιες θεωρούνται στατιστικά σημαντικές. Ένα παράδειγμα των αποτελεσμάτων της εντολής summary φαίνεται αναλυτικά στην παρακάτω εικόνα:

```
> summary(fit)

Call:
lm(formula = y ~ z1 + z2 + z3 + z4)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74545 -0.45211  0.02538  0.49136  1.82340

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.6039425  0.1834755   3.292   0.0014 **
z1           1.7953974  0.0151235 118.716  < 2e-16 ***
z2          -0.0411167  0.1505826  -0.273   0.7854
z3           0.0004549  0.0157647   0.029   0.9770
z4           1.3495267  0.1804320   7.479  3.7e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7465 on 95 degrees of freedom
Multiple R-squared:  0.9934,    Adjusted R-squared:  0.9931
F-statistic:  3572 on 4 and 95 DF,  p-value: < 2.2e-16
```

Εικόνα 4.3: Αποτελέσματα της προσαρμογής του μοντέλου στην R

Στη συνέχεια με την εντολή stepAIC επιλέγεται το κατάλληλο μοντέλο μέσω της κατεύθυνσης backward ελέγχοντας τις τιμές των κριτηρίων AIC και BIC. Η διαδικασία αυτή επαναλαμβάνεται 100 φορές και υπολογίζεται το ποσοστό που υπολογίζεται το σωστό μοντέλο, αυτό δηλαδή που περιέχει τις σημαντικές μεταβλητές και τα σφάλματα τύπου I και τύπου II. Η εντολή η οποία εκτελεί τη βηματική επιλογή μοντέλων ορίζεται ως εξής:

Για το AIC : step< - stepAIC(fit, direction="backward", trace=FALSE, k=2)

Για το BIC: step< - stepAIC(fit, direction="backward", trace=FALSE, k=log(n))

Επιπλέον στον κώδικά μας έχει τεθεί ως correctAIC το ποσοστό των φορών από τις 100 που το AIC έχει επιλέξει το σωστό μοντέλο και correctBIC το ποσοστό των φορών από τις 100 που το BIC έχει επιλέξει το σωστό μοντέλο δηλαδή εκείνο που περιέχει τις z_1 $και$ z_4 μεταβλητές. Επίσης ορίζεται ως error1AIC το ποσοστό φορών από τις 100 που το AIC επιλέγει τουλάχιστον μία μη σημαντική μεταβλητή και error2AIC το ποσοστό των φορών που το AIC δεν επιλέγει μία σημαντική μεταβλητή. Αντίστοιχα ορίζεται ως error1BIC το ποσοστό των φορών από τις 100 που το BIC επιλέγει μία μη σημαντική μεταβλητή και error2BIC το ποσοστό φορών από τις 100 που το BIC δεν επιλέγει μία σημαντική μεταβλητή.

Στους παρακάτω πίνακες αναγράφονται αναλυτικά όλα τα παραπάνω αποτελέσματα των προσομοιώσεων για τα διάφορα μεγέθη δείγματος ($n = 25$, 50, 100, 200) που εκτελούνται για τις τιμές του $b1 = 0.8$ $και$ $b1 = 1.8$.

Πίνακας 1		
b1=1.8, b4=1		
	AIC	**BIC**
N=25	correctAIC error1AIC error2AIC 0.55 0.36 0.13	correctBIC error1BIC error2BIC 0.52 0.27 0.27
N=50	correctAIC error1AIC error2AIC 0.68 0.31 0.02	correctBIC error1BIC error2BIC 0.76 0.15 0.11
N=100	correctAIC error1AIC error2AIC 0.63 0.37 0	correctBIC error1BIC error2BIC 0.85 0.15 0
N=200	correctAIC error1AIC error2AIC 0.71 0.29 0	correctBIC error1BIC error2BIC 0.95 0.05 0

Πίνακας 2		
b1=0.8		
	AIC	**BIC**
N=25	correctAIC error1AIC error2AIC 0.43 0.42 0.24	correctBIC error1BIC error2BIC 0.53 0.20 0.36
N=50	correctAIC error1AIC error2AIC 0.61 0.36 0.05	correctBIC error1BIC error2BIC 0.77 0.15 0.1
N=100	correctAIC error1AIC error2AIC 0.67 0.33 0	correctBIC error1BIC error2BIC 0.92 0.08 0
N=200	correctAIC error1AIC error2AIC 0.69 0.31 0	correctBIC error1BIC error2BIC 0.93 0.07 0

Όπως φαίνεται από τους παραπάνω δύο πίνακες όσο αυξάνεται το μέγεθος του δείγματος, αυξάνεται και ο αριθμός των σωστών μοντέλων που επιλέγονται από τα δύο κριτήρια. Το κριτήριο BIC όμως δίνει πιο αξιόπιστα αποτελέσματα σε σχέση με το AIC καθώς επιλέγει πάντα μεγαλύτερο αριθμό σωστών μοντέλων και λιγότερα μοντέλα που περιέχουν μία

μη σημαντική μεταβλητή (error I). Η συμπεριφορά τους είναι περίπου παρόμοια ως προς το error II με το BIC να δίνει μεγαλύτερα σφάλματα για μικρό n. Δίνοντας τις τιμές 0.8 και 1.8 στον συντελεστή b1 τα αποτελέσματα δεν παρουσιάζουν μεγάλες διαφορές μεταξύ τους καθώς δεν επηρεάζουν τη σημαντικότητα των μεταβλητών.

Προσομοίωση 2η

Σε αυτή την προσομοίωση οι συμμεταβλητές δεν είναι ανεξάρτητες μεταξύ τους αλλά ακολουθούν πολυμεταβλητή κανονική κατανομή. Πιο συγκεκριμένα οι συμμεταβλητές των δειγμάτων σε αυτή την περίπτωση ακολουθούν πολυμεταβλητή κανονική κατανομή με μέση τιμή μηδέν και τυπική απόκλιση ένα και συνδιασπορές που δίνονται από τον τύπο $rho^{|i-j|}$ με i,j=1,..,4 όπου rho=0.5 . Στον κώδικα στην R χρησιμοποιήθηκε η παρακάτω εντολή:

```
for(i in 1:100)
{
print(i)

mu=c(0,0,0,0)
sigma=matrix(c(1,0.5,0.25,0.125,0.5,1,0.5,0.25,0.25,0.5,1,0.5,0.125,0.25,0.5,1),4,4)
z<-mvrnorm(n=100,mu,sigma,empirical=FALSE)

z1<-z[,1]
z2<-z[,2]
z3<-z[,3]
z4<-z[,4]

eps<-rnorm(100,0,1)
```

Εικόνα 4.4: Δήλωση εξαρτημένων συμμεταβλητών στην R οι οποίες ακολουθούν πολυμεταβλητή κατανομή με συντελεστές συσχέτισης που βασίζονται σε rho=0.5, απόσπασμα από τον κώδικα στην R.

Με την εντολή mvrnorm κατασκευάζονται οι συμμεταβλητές που ακολουθούν πολυμεταβλητή κανονική κατανομή με mu το διάνυσμα μέσης τιμής και για τις τέσσερις συμμεταβλητές και sigma ο πίνακας διασπορών-συνδιασπορών των συμμεταβλητών. Κάθε στοιχείο a_{ij} του πίνακα ισούται με $a_{ij} = rho^{|i-j|}$ με $rho = 0.5$. Οι όροι σφάλματος ακολουθούν όπως και πριν κανονική κατανομή $N(0,1)$.

Πίνακας 3		
B1=1.8		
	AIC	**BIC**
N=25	correctAIC error1AIC error2AIC 0.66 0.33 0.01	correctBIC error1BIC error2BIC 0.80 0.18 0.02
N=50	correctAIC error1AIC error2AIC 0.73 0.27 0	correctBIC error1BIC error2BIC 0.94 0.06 0
N=100	correctAIC error1AIC error2AIC 0.71 0.29 0	correctBIC error1BIC error2BIC 0.92 0.08 0
N=200	correctAIC error1AIC error2AIC 0.79 0.21 0	correctBIC error1BIC error2BIC 0.97 0.03 0

Πίνακας 4		
B1=0.8		
	AIC	**BIC**
N=25	correctAIC error1AIC error2AIC 0.65 0.33 0.04	correctBIC error1BIC error2BIC 0.78 0.19 0.06
N=50	correctAIC error1AIC error2AIC 0.83 0.17 0	correctBIC error1BIC error2BIC 0.69 0.31 0
N=100	correctAIC error1AIC error2AIC 0.69 0.31 0	correctBIC error1BIC error2BIC 0.96 0.04 0
N=200	correctAIC error1AIC error2AIC 0.64 0.36 0	correctBIC error1BIC error2BIC 0.95 0.05 0

Στη Προσομοίωση αυτή φαίνεται ξανά η αξιοπιστία του BIC σε σύγκριση με το AIC για όλα τα μεγέθη του δείγματος που υποθέσαμε. Όσο αυξάνεται το μέγεθος του δείγματος αυξάνεται και το ποσοστό των σωστών μοντέλων που επιλέγονται και από τα δύο κριτήρια. Οι τιμές για το error II είναι παρόμοιες για τα δύο μοντέλα αλλά το BIC υπερτερεί του AIC ως προς το error I αφού εμφανίζει συστηματικά μικρότερα σφάλματα.

Προσομοίωση 3η

Η προσομοίωση αυτή είναι παρόμοια με την παραπάνω με τη διαφορά ότι χρησιμοποιείται συντελεστής συσχέτισης $rho = 0.8$, δηλαδή υπάρχει μεγαλύτερη εξάρτηση μεταξύ τους. Λογικό είναι να αλλάζει και ο πίνακας διασπορών-συνδιασπορών ο οποίος χρησιμοποιείται στην παρακάτω εντολή προκειμένου να δημιουργηθούν τα δείγματα.

```
for (i in 1:100)
{
print("------------------")
print(i)
mu=c(0,0,0,0)
sigma=matrix(c(1,0.8,0.64,0.512,0.8,1,0.8,0.64,0.64,0.8,1,0.5,0.512,0.64,0.8,1),4,4)


z<-mvrnorm(n,mu,sigma,empirical=FALSE)
```

Εικόνα 4.5: Δήλωση εξαρτημένων συμμεταβλητών στην R οι οποίες ακολουθούν πολυμεταβλητή κατανομή με συντελεστές συσχέτισης που βασίζονται σε rho=0.8, απόσπασμα από τον κώδικα στην R.

Πίνακας 5		
B1=1.8		
	AIC	**BIC**
N=25	correctAIC error1AIC error2AIC 0.63 0.37 0.03	correctBIC error1BIC error2BIC 0.80 0.20 0.04
N=50	correctAIC error1AIC error2AIC 0.88 0.12 0	correctBIC error1BIC error2BIC 0.68 0.32 0
N=100	correctAIC error1AIC error2AIC 0.74 0.26 0	correctBIC error1BIC error2BIC 0.91 0.09 0
N=200	correctAIC ePror1AIC error2AIC 0.77 0.23 0	correctBIC error1BIC error2BIC 0.96 0.04 0

Πίνακας 6		
B1=0.8		
	AIC	**BIC**
N=25	correctAIC error1AIC error2AIC 0.56 0.41 0.12	correctBIC error1BIC error2BIC 0.69 0.25 0.20
N=50	correctAIC error1AIC error2AIC 0.68 0.32 0.04	correctBIC error1BIC error2BIC 0.84 0.16 0.06
N=100	correctAIC error1AIC error2AIC 0.69 0.31 0	correctBIC error1BIC error2BIC 0.92 0.08 0
N=200	correctAIC error1AIC error2AIC 0.68 0.32 0	correctBIC error1BIC error2BIC 0.96 0.04 0

Από τα παραπάνω αποτελέσματα φαίνεται ότι ενώ ο συντελεστής συσχέτισης αυξάνεται, το BIC εξακολουθεί να είναι περισσότερο αποδοτικό σε σχέση με το AIC. Ενώ το μέγεθος του δείγματος αυξάνεται, η πιθανότητα επιλογής μοντέλου με μη σημαντική μεταβλητή και η πιθανότητα μη επιλογής μοντέλου με σημαντική μεταβλητή μειώνεται. Σε σύγκριση με την προηγούμενη προσομοίωση παρατηρείται ότι καθώς ο συντελεστής συσχέτισης αυξάνεται η απόδοση των κριτηρίων συνήθως μειώνεται. Δηλαδή στην περίπτωση όπου rho=0.5 επιλέγονται περισσότερες φορές το σωστό μοντέλο σε αντίθεση με την περίπτωση όπου rho=0.8. Αυτό το αποτέλεσμα θεωρείται λογικό καθώς αφού αυξάνεται η συσχέτιση των μεταβλητών μεταξύ τους είναι πιο δύσκολο να επιλεχθεί η σωστή μεταβλητή κι έτσι αυξάνονται τα σφάλματα τύπου I και τύπου II.

<u>Προσομοίωση 4η</u>

Σε αυτή την προσομοίωση εξετάζεται η περίπτωση όπου οι συντελεστές έχουν τις εξής τιμές: $(1, 1.8, 0, 0, 0.2)$. Η b4 δηλαδή τώρα παίρνει την τιμή 0.2 δηλαδή ο συντελεστής της Z4 είναι μικρότερος από πριν και ίσως αυτό δυσκολέψει περισσότερο την επιλογή αυτής της μεταβλητής ως σημαντική μέσα στο μοντέλο. Οι προσομοιώσεις πραγματοποιούνται για τις περιπτώσεις όπου δεν υπάρχει εξάρτηση μεταξύ των συμμεταβλητών ή υπάρχει κάποια εξάρτηση με τις μεταβλητές να ακολουθούν πολυμεταβλητή κανονική κατανομή με τους συντελεστές συσχέτισης μεταξύ των μεταβλητών να παίρνουν τιμές όπως έχουμε περιγράψει προηγουμένως με βάση rho=0.5 και rho= 0.8. Τα σφάλμα ορίζονται όπως πριν ως error1AIC και error1BIC τα ποσοστά που επιλέγεται τουλάχιστον μία μη σημαντική μεταβλητή ενώ error2AIC και error2BIC τα ποσοστά που δεν επιλέγεται μία σημαντική μεταβλητή.

Πίνακας 7		
ανεξαρτησία		
	AIC	**BIC**
N=25	correctAIC error1AIC error2AIC 0.16 0.36 0.70	correctBIC error1BIC error2BIC 0.08 0.21 0.89
N=50	correctAIC error1AIC error2AIC 0.20 0.32 0.74	correctBIC error1BIC error2BIC 0.13 0.14 0.86
N=100	correctAIC error1AIC error2AIC 0.24 0.36 0.64	correctBIC error1BIC error2BIC 0.11 0.15 0.87
N=200	correctAIC error1AIC error2AIC 0.37 0.26 0.49	correctBIC error1BIC error2BIC 0.21 0.03 0.79

ΠΙΝΑΚΑΣ 8		
Rho=0.5		
	AIC	**BIC**
N=25	correctAIC error1AIC error2AIC 0.27 0.37 0.63	correctBIC error1BIC error2BIC 0.20 0.22 0.74
N=50	correctAIC error1AIC error2AIC 0.28 0.38 0.57	correctBIC error1BIC error2BIC 0.28 0.11 0.70
N=100	correctAIC error1AIC error2AIC 0.46 0.29 0.38	correctBIC error1BIC error2BIC 0.33 0.13 0.64
N=200	correctAIC error1AIC error2AIC 0.67 0.28 0.13	correctBIC error1BIC error2BIC 0.74 0.06 0.26

ΠΙΝΑΚΑΣ 9						
Rho=0.8						
	AIC			**BIC**		
N=25	correctAIC 0.13	error1AIC 0.40	error2AIC 0.77	correctBIC 0.07	error1BIC 0.29	error2BIC 0.89
N=50	correctAIC 0.31	error1AIC 0.36	error2AIC 0.54	correctBIC 0.23	error1BIC 0.16	error2BIC 0.74
N=100	correctAIC 0.46	error1AIC 0.24	error2AIC 0.43	correctBIC 0.29	error1BIC 0.05	error2BIC 0.71
N=200	correctAIC 0.60	error1AIC 0.27	error2AIC 0.21	correctBIC 0.54	error1BIC 0.04	error2BIC 0.46

Σε αντίθεση με τις προηγούμενες προσομοιώσεις σε αυτή την περίπτωση το AIC δίνει σχεδόν πάντα καλύτερα αποτελέσματα σε σχέση με το BIC όπως φαίνεται στους παραπάνω πίνακες. Αξιοσημείωτο είναι ότι τα σφάλματα σε αυτή την προσομοίωση και κυρίως το σφάλμα τύπου II παρουσιάζουν ιδιαίτερη αύξηση σε αντίθεση με τις προηγούμενες περιπτώσεις που ήταν συνήθως μηδέν. Αυτό συμβαίνει καθώς η τιμή του συντελεστή της μεταβλητής z4 έχει μειωθεί αρκετά με αποτέλεσμα να είναι αρκετά δύσκολο να επιλεχθεί αυτή η σημαντική μεταβλητή. Εξακολουθεί όμως να αυξάνεται ο αριθμός των σωστών επιλογών των κριτηρίων όσο αυξάνεται και το μέγεθος του δείγματος.

4.3 Επιλογή του βέλτιστου μοντέλου ύστερα από προσαρμογή σε πραγματικά δεδομένα

Σε αυτή την ενότητα χρησιμοποιούνται έτοιμα δεδομένα από στατιστικές έρευνες που διεξήχθησαν το 1977 στην Αμερική. Μελετώντας ένα δείγμα από 50 πολιτείες των Ηνωμένων Πολιτειών εξετάζουμε πως το προσδόκιμο ζωής των πολιτών μπορεί να επηρεαστεί από διάφορους παράγοντες που αποτελούν τις επεξηγηματικές μεταβλητές του μοντέλου μας. Συγκεκριμένα ως επεξηγηματικές μεταβλητές θεωρούνται ο πληθυσμός, το εισόδημα, το ποσοστό του πληθυσμού με αναλφαβητισμό, το ποσοστό εγκληματικότητας, το ποσοστό του πληθυσμού που έχει αποφοιτήσει από το Λύκειο, ο αριθμός των μερών ανά χρόνο με

υπερβολικά χαμηλές θερμοκρασίες (παγετώνας) και η έκταση ξηράς σε τετραγωνικά μίλια. Θέτοντας ως μεταβλητή απόκρισης το προσδόκιμο ζωής, στον παρακάτω πίνακα παρουσιάζονται συνοπτικά όλες οι μεταβλητές συνοδευόμενες με τους συμβολισμούς τους.

ΜΕΤΑΒΛΗΤΕΣ-ΟΝΟΜΑΣΙΑ	ΠΕΡΙΓΡΑΦΗ
y (Life.Expectation)	Προσδόκιμο ζωής σε έτη
x1 (Population)	Πληθυσμός
x2 (Income)	Κατά κεφαλήν εισόδημα (ανά άτομο)
x3 (Illiteracy)	Ποσοστό του πληθυσμού με αναλφαβητισμό
x4 (Murder)	Ποσοστό εγκληματικότητας
x5 (HS.Grad)	Ποσοστό του πληθυσμού που έχει αποφοιτήσει από το Λύκειο
x6 (Frost)	Αριθμός ημερών ανά χρόνο με παγετώνα
x7 (Area)	Έκταση ξηράς σε τετραγωνικά μίλια

Εικόνα 4.6: Περιγραφή και συμβολισμός μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών.

Χρησιμοποιώντας το στατιστικό εργαλείο της R προσαρμόζουμε τα δεδομένα με το παραπάνω γραμμικό μοντέλο και μέσω των κριτηρίων AIC και BIC, επιθυμούμε να επιλέξουμε το βέλτιστο μοντέλο για το πρόβλημά μας. Αρχικά με την εντολή:

>state.x77

>state<-as.data.frame(state.x77)

καλούμε έτοιμα στατιστικά δεδομένα τα οποία εμφανίζονται σε ένα πίνακα με 50 σειρές και 8 στήλες. Οι στήλες αντιπροσωπεύουν τις μεταβλητές και οι σειρές τις 50 πολιτείες της Αμερικής. Στη συνέχεια ορίζουμε τη μεταβλητή απόκρισης y (Life.Expectation) και τις επεξηγηματικές μεταβλητές x1 έως x7 που παρουσιάζονται στην εικόνα 4.6. Αρκετά σημαντικό είναι να ελέγξουμε τη συσχέτιση μεταξύ των επεξηγηματικών μεταβλητών η οποία φαίνεται στον παρακάτω πίνακα (εικόνα 4.7). Όσο πιο κοντά στη μονάδα είναι η τιμή του συντελεστή συσχέτισης τόσο πιο ισχυρή είναι η σχέση μεταξύ των μεταβλητών. Συγκεκριμένα ισχύει:

- Αν ρ=1 τότε υπάρχει τέλεια θετική γραμμική σχέση ανάμεσα στις μεταβλητές X,Y.
- Αν ρ=0 τότε οι μεταβλητές X,Y δεν συνδέονται γραμμικά, δηλαδή είναι ασυσχέτιστες.
- Αν ρ=-1 τότε υπάρχει τέλεια αρνητική σχέση ανάμεσα στις μεταβλητές X,Y.

Η εντολή που χρησιμοποιήθηκε στην R για να ελεγχθεί η αριθμητική συσχέτιση των μεταβλητών είναι η εξής

> cor (data.frame (x1,x2,x3,x4,x5,x6,x7))

	x1 (Population)	x2 (Income)	x3 (Illiteracy)	x4 (Murder)	x5 (HS.Grad)	x6 (Frost)	x7 (Area)
x1(Population)	1.00000000	0.2082276	0.10762237	0.3436428	-0.09848975	- 0.3321525	0.02254384
x2 (Income)	0.20822756	1.0000000	-0.4370751	-0.2300776	0.61993232	0.2262822	0.36331544
x3 (Illiteracy)	0.10762237	-0.4370752	1.00000000	0.7029752	-0.65718861	- 0.6719470	0.07726113
x4 (Murder)	0.34364275	-0.2300776	0.70297520	1.0000000	-0.48797102	- 0.5388834	0.22839021
x5 (HS.Grad)	-0.09848975	0.6199323	-0.6571886	-0.4879710	1.00000000	0.3667797	0.33354187
x6 (Frost)	-0.33215245	0.2262822	-0.6719469	-0.5388834	0.36677970	1.0000000	0.05922910
x7 (Area)	0.02254384	0.3633154	0.07726113	0.2283902	0.33354187	0.0592291	1.00000000

Εικόνα 4.7 : Πίνακας συσχέτισης των μεταβλητών

Σύμφωνα με τα αποτελέσματα του παραπάνω πίνακα είναι εμφανές ότι οι μεταβλητές x4 (Murder) και x3 (Illiteracy) έχουν τη μεγαλύτερη θετική συσχέτιση μεταξύ τους καθώς ο συντελεστής συσχέτισης συγκεκριμένα έχει την τιμή:

> cor (x4,x3)

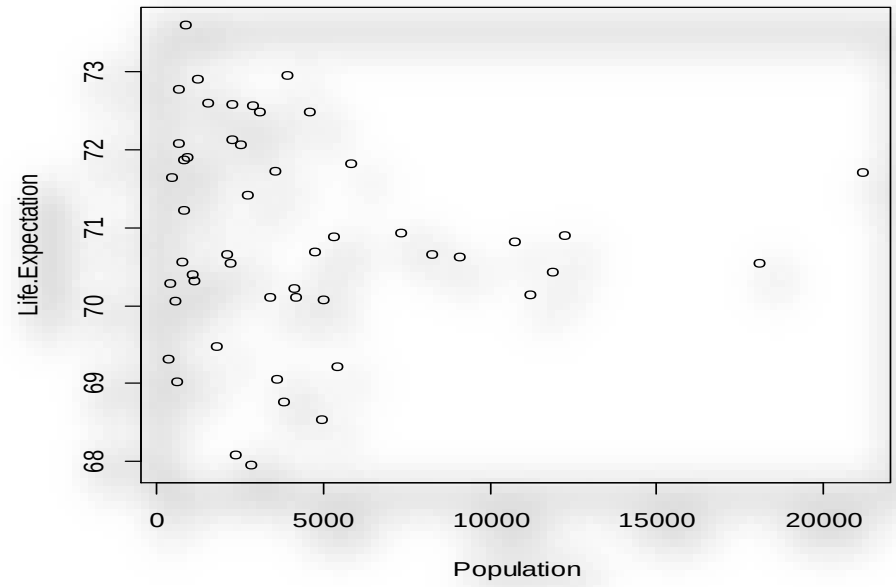
0.70297520

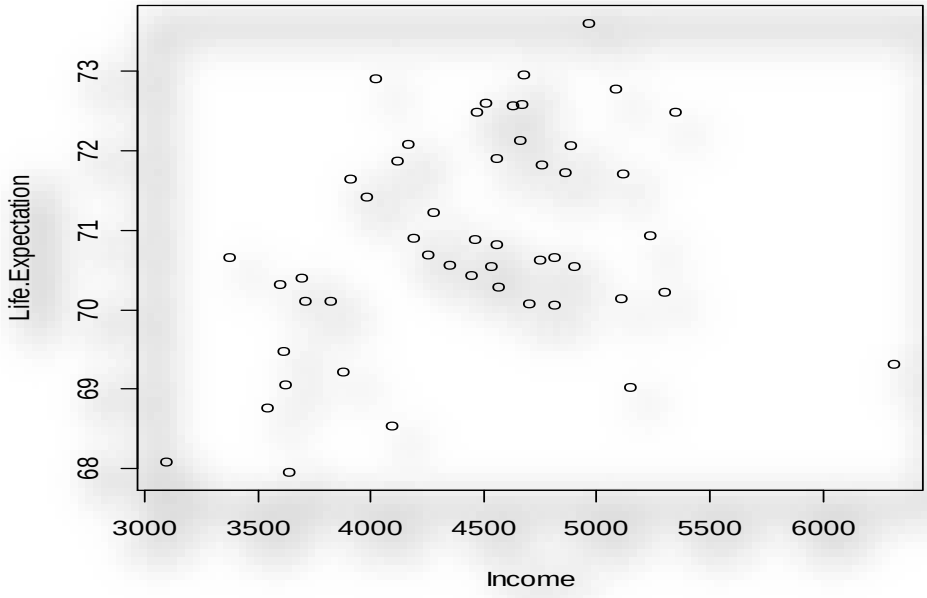
Ενώ οι μεταβλητές x6 (Frost) και x3 (Illiteracy) παρουσιάζουν τη μεγαλύτερη αρνητική συσχέτιση μεταξύ τους και συγκεκριμένα:

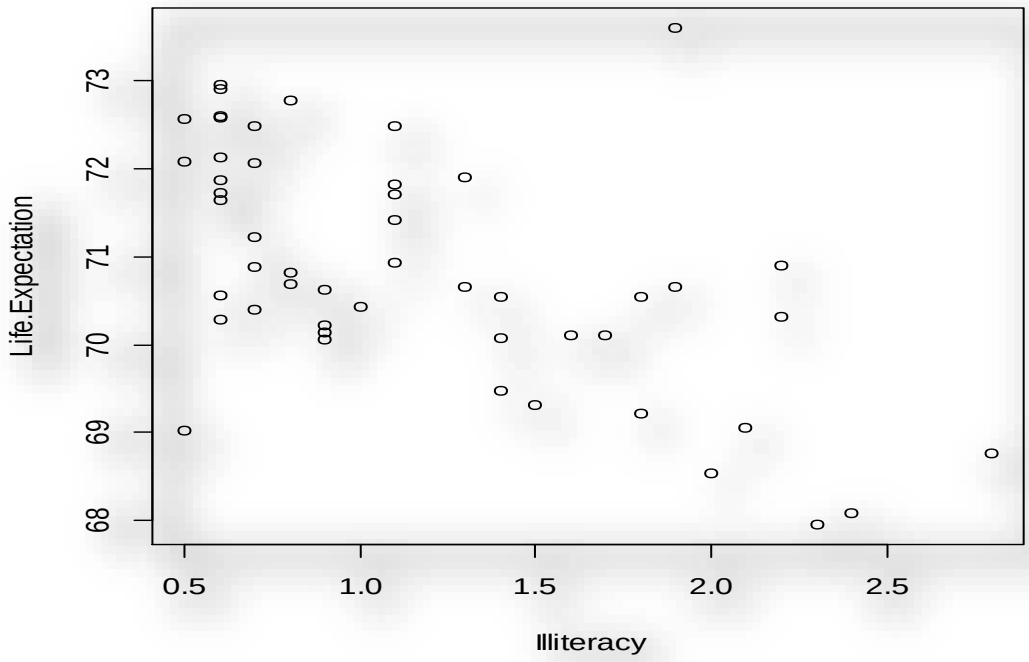
> cor(x6,x3)

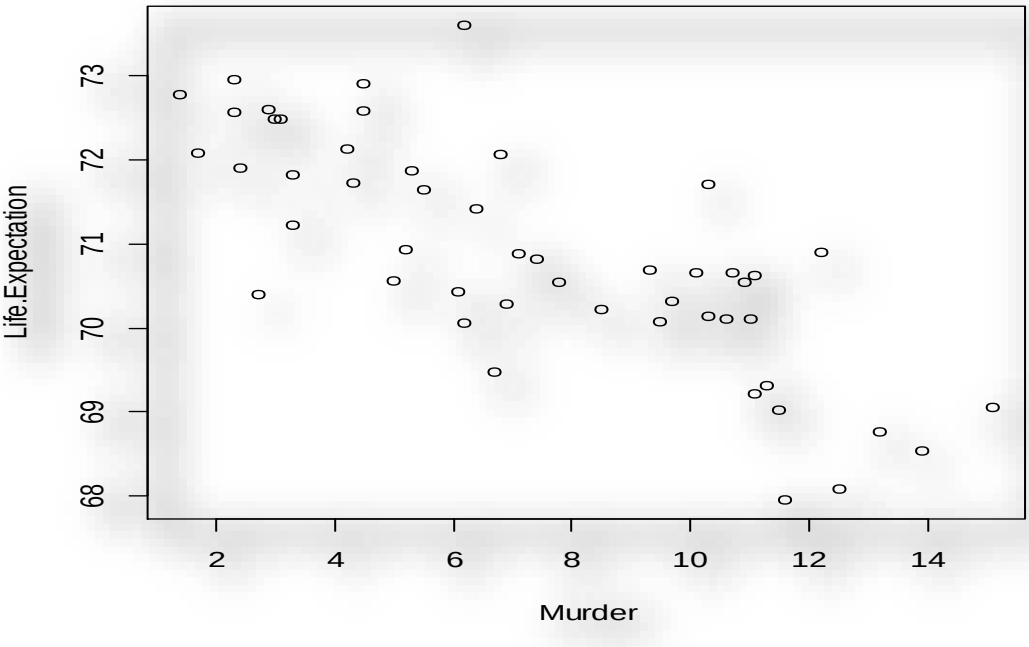
- 0.6719469

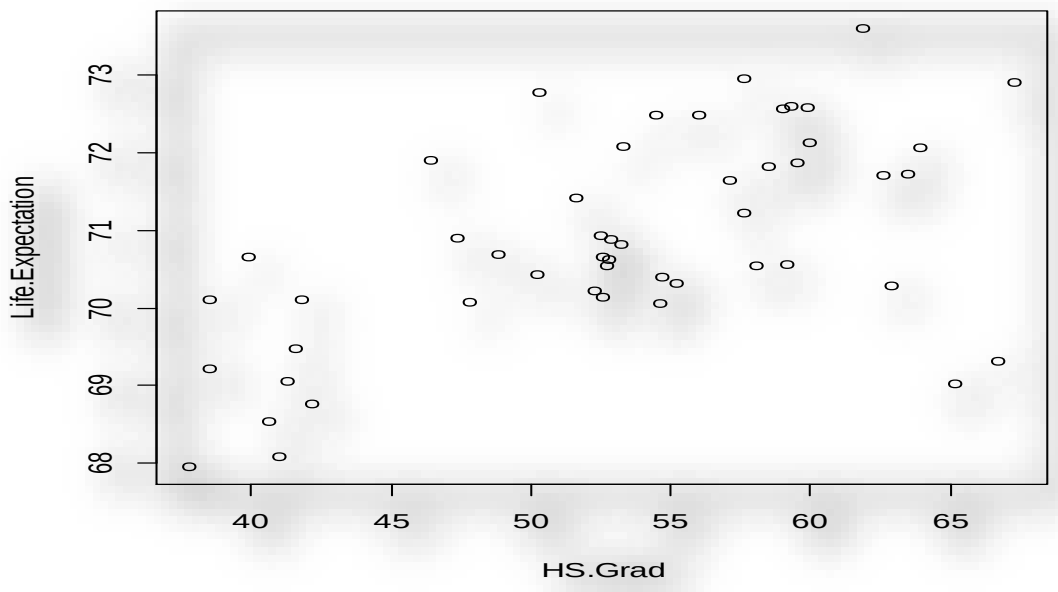
Στη συνέχεια παρουσιάζεται γραφικά η συσχέτιση της κάθε επεξηγηματικής μεταβλητής με τη μεταβλητή απόκρισης:

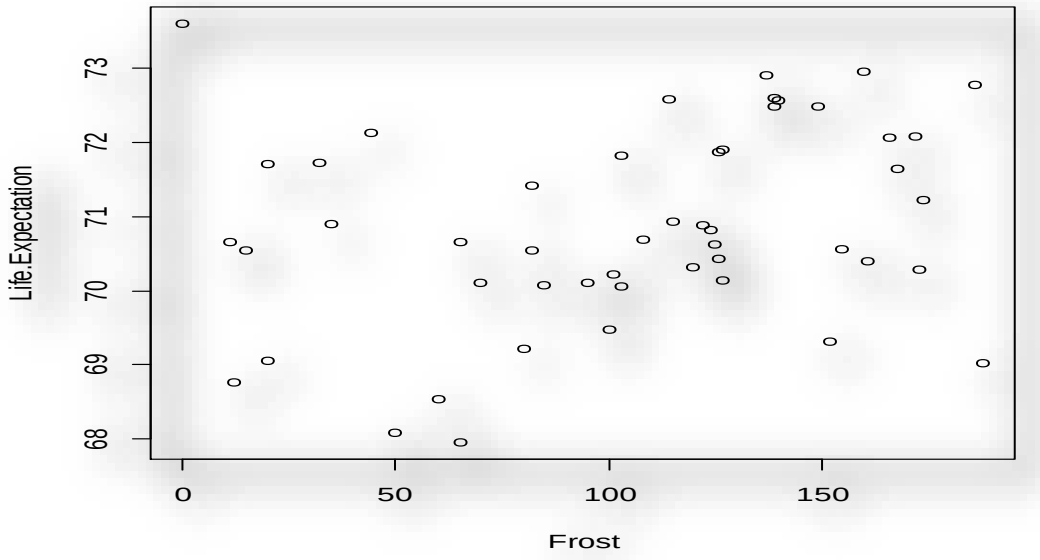

Life.Expectation
68
69
70
71
72
73
0
5000
10000
15000
20000
Population

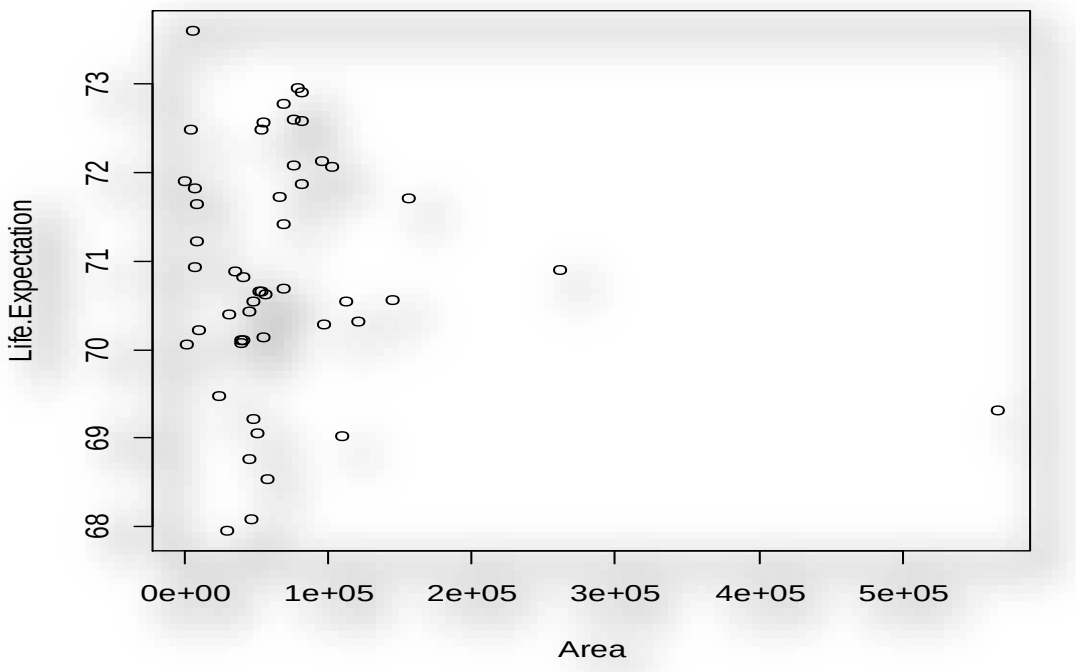

Life.Expectation
68
69
70
71
72
73
3000
3500
4000
4500
5000
5500
6000
Income


Life.Expectation
68
69
70
71
72
73
0.5
1.0
1.5
2.0
2.5
Illiteracy


Life.Expectation
68
69
70
71
72
73
2
4
6
8
10
12
14
Murder


Life.Expectation
68
69
70
71
72
73
40
45
50
55
60
65
HS.Grad


Life.Expectation
68
69
70
71
72
73
0
50
100
150
Frost



Στο 5ο διάγραμμα διασποράς φαίνεται να έχουμε ισχυρή σχέση καθώς όσο αυξάνουν οι τιμές της x5(HS.Grad) αυξάνουν και οι τιμές της Y(Life.Expectation) ενώ στο 3ο και 4ο διάγραμμα διασποράς έχουμε μια λιγότερο ισχυρή σχέση στην οποία όταν αυξάνουν οι τιμές των x3 (Illiteracy) και x4 (Murder) ελαττώνονται γενικά και οι τιμές της Y(Life.Expectation).

Στη συνέχεια γίνεται η προσαρμογή του γραμμικού μοντέλου όπου δηλώνεται ότι η Life.Expectation αποτελεί την εξαρτημένη μεταβλητή και οι υπόλοιπες τις επεξηγηματικές μεταβλητές. Συγκεκριμένα χρησιμοποιώντας στην R τις εντολές:

> fit<-lm (Life.Expectation~Population+Income+Illiteracy + Murder+ HS.Grad+Frost+Area, data=state)

>summary(fit)

Παίρνουμε τα εξής αποτελέσματα:

```
Call:
lm(formula = Life.Expectation ~ Population + Income + Illiteracy +
    Murder + HS.Grad + Frost + Area, data = state)

Residuals:
     Min       1Q   Median       3Q      Max
-1.48895 -0.51232 -0.02747  0.57002  1.49447

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
Population   5.180e-05  2.919e-05   1.775   0.0832 .
Income      -2.180e-05  2.444e-04  -0.089   0.9293
Illiteracy   3.382e-02  3.663e-01   0.092   0.9269
Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
HS.Grad      4.893e-02  2.332e-02   2.098   0.0420 *
Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .
Area        -7.383e-08  1.668e-06  -0.044   0.9649
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7448 on 42 degrees of freedom
Multiple R-squared:  0.7362,    Adjusted R-squared:  0.6922
F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

Εικόνα 4.8: Αποτελέσματα του κώδικα στην R με την εντολή summary

Σύμφωνα με τα παραπάνω αποτελέσματα του προσαρμοσμένου μοντέλου, παρατηρούμε από τις p-τιμές των ελέγχων Wald ότι οι επεξηγηματικές μεταβλητές Murder και HS.Grad είναι στατιστικά σημαντικές με p-τιμή 8.68e-08 και 0.0420 αντίστοιχα. Οριακά σημαντικές μεταβλητές επίσης μπορούν να θεωρηθούν οι Population και Frost με p-τιμή 0.0832 και 0.0752 αντίστοιχα.

Προκειμένου να επιλεχθεί το βέλτιστο μοντέλο χρησιμοποιείται στην R η εντολή stepAIC η οποία με τη μέθοδο αποκλεισμού μεταβλητών και μέσω των κριτηρίων AIC και BIC μας οδηγεί στο στατιστικά καλύτερο μοντέλο. Παρακάτω παρουσιάζονται οι εντολές που χρησιμοποιούνται στην R, καθώς επίσης τα αποτελέσματα του τελικού βήματος σε κάθε κριτήριο ξεχωριστά. Καλώντας πάλι το πακέτο library(MASS) από την R χρησιμοποιείται η εντολή stepAIC με την κατεύθυνση “backward”, η οποία αναφέρεται στη μέθοδο απόκλισης μεταβλητών και οι περιπτώσεις k=2 και k=log(n) αντιστοιχούν στα κριτήρια AIC και BIC.

> stepAIC (fit, direction="backward", k=2)

```
Step:  AIC=-28.16
Life.Expectation ~ Population + Murder + HS.Grad + Frost

             Df Sum of Sq    RSS     AIC
<none>                     23.308 -28.161
- Population  1     2.064 25.372 -25.920
- Frost       1     3.122 26.430 -23.877
- HS.Grad     1     5.112 28.420 -20.246
- Murder      1    34.816 58.124  15.528

Call:
lm(formula = Life.Expectation ~ Population + Murder + HS.Grad +
    Frost, data = state)

Coefficients:
(Intercept)   Population       Murder      HS.Grad        Frost
  7.103e+01    5.014e-05   -3.001e-01    4.658e-02   -5.943e-03
```

Εικόνα 4.9: Αποτελέσματα στην R με τη μέθοδο απόκλισης μεταβλητών και τη χρήση του κριτηρίου πληροφορίας AIC.

>stepAIC (fit,direction="backward", k=log(n))

```
Step:  AIC=-18.6
Life.Expectation ~ Population + Murder + HS.Grad + Frost

             Df Sum of Sq    RSS     AIC
<none>                     23.308 -18.601
- Population  1     2.064 25.372 -18.271
- Frost       1     3.122 26.430 -16.228
- HS.Grad     1     5.112 28.420 -12.598
- Murder      1    34.816 58.124  23.176

Call:
lm(formula = Life.Expectation ~ Population + Murder + HS.Grad +
    Frost, data = state)

Coefficients:
(Intercept)   Population       Murder      HS.Grad        Frost
  7.103e+01    5.014e-05   -3.001e-01    4.658e-02   -5.943e-03
```

Εικόνα 4.10: Αποτελέσματα στην R με τη μέθοδο απόκλισης μεταβλητών και τη χρήση του κριτηρίου πληροφορίας BIC

Από τα αποτελέσματα των παραπάνω εικόνων παρατηρούμε ότι και τα δύο κριτήρια, AIC και BIC καταλήγουν στο ίδιο τελικό βέλτιστο μοντέλο που έχει ως επεξηγηματικές μεταβλητές τις Population, Frost, HS.Grad, Murder με τη διαφορά ότι στη δεύτερη περίπτωση η τιμή του κριτηρίου του τελικού μοντέλου είναι μικρότερη κατά απόλυτη τιμή. Το BIC λοιπόν μπορεί να θεωρηθεί πιο αποδοτικό καθώς οδηγεί σε πιο φειδωλό μοντέλο σε σχέση με το κριτήριο AIC καθώς επίσης όσο μικρότερη κατά απόλυτη τιμή είναι η τιμή του κριτηρίου τόσο στατιστικά καλύτερο θεωρείται το μοντέλο. Τέλος παρουσιάζουμε την ανάλυση παλινδρόμησης του μοντέλου που επιλέχτηκε ως το βέλτιστο, με τις μεταβλητές Population, Murder, H.S.Grad, Frost και τη χρήση των εξής εντολών:

>final_fit<-lm(Life.Expectation~Population+Murder+HS.Grad+Frost, data=state)

>summary(final_fit)

Στην εικόνα 4.10 φαίνονται τα αποτελέσματα του γραμμικού μοντέλου που επιλέχτηκε με τη μέθοδο απόκλισης μεταβλητών με τη χρήση και των δύο κριτηρίων καθώς κατέληξαν στο ίδιο βέλτιστο μοντέλο. Αξιοσημείωτο είναι ότι όλες οι μεταβλητές που επιλέχτηκαν, δηλαδή οι Population, Murder, H.S.Grad, Frost είναι σημαντικές με p-τιμές 2e-16, 0.05201, 1.77e-10, 0.00297, 0.01802 αντίστοιχα.

```
Call:
lm(formula = Life.Expectation ~ Population + Murder + HS.Grad +
    Frost, data = state)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47095 -0.53464 -0.03701  0.57621  1.50683

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
Population   5.014e-05  2.512e-05   1.996  0.05201 .
Murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
HS.Grad      4.658e-02  1.483e-02   3.142  0.00297 **
Frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7197 on 45 degrees of freedom
Multiple R-squared:  0.736,     Adjusted R-squared:  0.7126
F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

Εικόνα 4.11: Αποτελέσματα του προσαρμοσμένου μοντέλου στην R

Επίλογος και Συμπεράσματα

Τα κριτήρια πληροφορίας αποτελούν μία ελκυστική βάση για την επιλογή του βέλτιστου μοντέλου ανάμεσα από ένα σύνολο υποψήφιων μοντέλων που προσπαθούν να προσεγγίσουν τα πραγματικά δεδομένα στο κόσμο της Στατιστικής. Βασικές έννοιες όπως είναι η απόσταση Kullback-Leibler και η εκτιμήτρια μέγιστης πιθανοφάνειας είναι απαραίτητες για την κατανόηση αυτών των κριτηρίων. Στη παρούσα διπλωματική εργασία μελετήθηκαν δύο τέτοια κριτήρια, τα AIC και BIC, και εξήχθησαν συμπεράσματα για το ποιο από τα δύο μπορεί να θεωρηθεί πιο αξιόπιστο. Επίσης έγινε αναφορά στις επεκτάσεις ή διορθώσεις του κριτηρίου AIC. Συγκεκριμένα αναφέρεται το κριτήριο TIC το οποίο αποτελεί διόρθωση του AIC και υποστηρίζει ότι η κατανομή που αντιστοιχεί στο μοντέλο που θέλουμε να προσεγγίσουμε βρίσκεται ανάμεσα στο πλήθος των υποψήφιων προσεγγιστικών μοντέλων. Το κριτήριο AICc, το οποίο αποτελεί και αυτό επέκταση του AIC, προτιμάται όταν το μέγεθος του δείγματος δεν είναι αρκετά μεγάλο. Έχει προσθετικό όρο μεροληψίας, δηλαδή χρησιμοποιεί αμερόληπτες εκτιμήτριες προκειμένου να διορθώσει τη μεροληψία ως προς το μέγεθος του δείγματος. Επιπλέον σε περιπτώσεις όπου υπάρχει μεγάλη διασπορά στο σύνολο των δεδομένων, προτείνεται το κριτήριο QAIC που μοιάζει με το AIC, με τη διαφορά ότι περιέχει ένα συντελεστή μεταβλητότητας c. Αναμενόμενο είναι όταν η διασπορά είναι μηδενική, ο συντελεστής μεταβλητότητας παίρνει την τιμή 1 και τότε τα δύο κριτήρια, AIC και QAIC, ταυτίζονται.

Και τα δύο κριτήρια, AIC και BIC, αποσκοπούν στο να προτείνουν το καταλληλότερο στατιστικά μοντέλο το οποίο αντιστοιχεί στη μικρότερη τιμή που παίρνουν τα κριτήρια για κάποιο συγκεκριμένο πρόβλημα. Από την άλλη πλευρά το AIC συνήθως προσπαθεί να βρει άγνωστα μοντέλα που αντιστοιχούν στη πραγματικότητα σε μεγάλες διαστάσεις. Αυτό έχει ως αποτέλεσμα το AIC να μην χρησιμοποιεί πραγματικά μοντέλα σε αντίθεση με το BIC. Επιπλέον τα δύο κριτήρια διαφέρουν ως προς τον όρο ποινικοποίησης. Ο όρος ποινής για το BIC είναι μεγαλύτερος από αυτόν του AIC, ο οποίος θεωρείται πιο αυστηρός ως προς την εισαγωγή μεταβλητών προσπαθώντας έτσι να μειώσει τη πολυπλοκότητα του μοντέλου. Δύο καταλυτικοί τρόποι αξιολόγησης της αξιοπιστίας των δύο κριτηρίων είναι η εξέταση ως προς τη συνέπεια και αποδοτικότητά τους. Και τα δύο κριτήρια χαρακτηρίζονται από ασθενή και ισχυρή συνέπεια μέσα στο πλαίσιο της συνέπειας που ορίζεται στα θεωρήματα 3.1, 3.2 καθώς επιτυγχάνουν να επιλέξουν εκείνο το μοντέλο που ελαχιστοποιεί την απόσταση K-L όταν αυτό είναι μοναδικό. Μέσω του BIC όμως προστίθεται και το χαρακτηριστικό του φειδωλού μοντέλου, κάτι που του δίνει τον τίτλο της ασθενούς και ισχυρούς συνέπειας (και φειδωλότητας) μέσα στο πλαίσιο της συνέπειας που ορίζεται στα θεωρήματα 3.3 και 3.4 σε αντίθεση με το AIC που δεν ικανοποιεί αυτού του τύπου τη συνέπεια. Από άλλη οπτική γωνία το AIC θεωρείται αποδοτικό κριτήριο ενώ το BIC όχι. Η αποδοτικότητα επιτυγχάνεται όταν το μοντέλο που έχει επιλεχθεί ως βέλτιστο συμφωνεί με αυτό που έχει επιλεχθεί από το μέσο τετραγωνικό σφάλμα πρόβλεψης.

Η μεγαλύτερη αξιοπιστία του BIC σε σχέση με το AIC φάνηκε και μέσα από τις προσομοιώσεις. Το BIC ενδείκνυται για δείγματα μεγάλου μεγέθους καθώς επίσης δίνει σχεδόν πάντα καλύτερα αποτελέσματα. Η αύξηση του μεγέθους του δείγματος συντελεί στην αύξηση του ποσοστού των φορών που θα επιλεχθεί το σωστό μοντέλο. Οι προσομοιώσεις πραγματοποιήθηκαν για τις περιπτώσεις όπου οι μεταβλητές είναι ανεξάρτητες και εξηρτημένες. Στην περίπτωση των εξηρτημένων μεταβλητών θεωρήσαμε πολυμεταβλητή κανονική κατανομή με ασθενέστερους και ισχυρότερους συντελεστές συσχετίσεως μεταξύ των μεταβλητών. Παρατηρήθηκε λοιπόν ότι όσο αυξάνονταν οι συντελεστές συσχέτισης, μειώνεται το ποσοστό επιλογής του σωστού μοντέλου καθώς αυξάνονται το ποσοστό επιλογής μοντέλου με μη σημαντική μεταβλητή καθώς και το ποσοστό μη επιλογής μοντέλου με σημαντική μεταβλητή. Αυτό είναι αναμενόμενο καθώς όσο αυξάνεται η συσχέτιση μεταξύ των συμμεταβλητών είναι πιο δύσκολο να επιλεχθούν οι σωστές μεταβλητές. Αξιοσημείωτο είναι ότι ενώ στις πρώτες τρεις προσομοιώσεις το BIC ήταν εκείνο που έδινε καλύτερα αποτελέσματα και διάλεγε περισσότερες φορές το σωστό μοντέλο, στην $4^{η}$ προσομοίωση αυτή την αξιοπιστία την παίρνει το AIC. Σε αυτή την περίπτωση ο συντελεστής μιας εκ των σημαντικών μεταβλητών έχει μειωθεί αρκετά με αποτέλεσμα να είναι δύσκολο να επιλεχθεί αυτή η μεταβλητή και ιδιαίτερα μέσω του BIC καθώς επιδιώκει να μειώνει την πολυπλοκότητα του μοντέλου και να επιλέγει εκείνο με τις λιγότερες παραμέτρους .

Κάθε Στατιστικός οφείλει να κάνει επιλογές. Οι επιλογές αυτές πραγματοποιούνται όταν έχει ολοκληρωθεί η συλλογή των δεδομένων και πρέπει να επιλεχθεί ποιο μοντέλο είναι κατάλληλο για να προσεγγίσει τα πραγματικά δεδομένα. Η χρήση των κριτηρίων πληροφορίας επιτυγχάνουν τη σωστή προσέγγιση αυτού του ζητήματος. Επιπλέον η συνέπεια και η αποδοτικότητα αποτελούν διαμορφωτικό ρόλο στην αξιοπιστία των κριτηρίων που χρησιμοποιούνται για την επιλογή του καταλληλότερου μοντέλου. Σύμφωνα με τον ορισμό της αποδοτικότητας το σύνολο των υποψήφιων μοντέλων πεπερασμένων διαστάσεων δεν περιέχει το αληθινό μοντέλο. Σκοπός είναι να επιλεχθεί εκείνο το μοντέλο που βρίσκεται πιο κοντά στο αληθινό με απαραίτητη προϋπόθεση να εκτιμηθεί αυτή η απόσταση. Από την άλλη ο όρος της συνέπειας στηρίζεται στην ύπαρξη του πραγματικού μοντέλου ανάμεσα στο σύνολο των υποψήφιων. Αξιοσημείωτο είναι ότι έχει γίνει λιγότερη μελέτη προκειμένου να διορθωθούν τα κριτήρια συνέπειας όπως είναι το BIC σε σχέση με τα κριτήρια αποδοτικότητας όπως είναι το AIC. Οι δύο αυτές έννοιες λοιπόν έχουν προκαλέσει μεγάλη διαφωνία ανάμεσα στους μελετητές καθώς αποτελεί υποκειμενικό ζήτημα, ποια από τις δύο υπερισχύει.

Βιβλιογραφία

Αγγλική

- Acquah Henry de-Graft, (2009) Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship
- Akaike, H. (1973), "Maximum likelihood identification of Gaussian autoregressive moving average models." *Biometrika* : 255-265
- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. Pages 267-281 in Second international symposium on information theory. B. N. Petrov and F. Csaki, (editors). Akademiai Kiado, Budapest
- Atkinson A. C. , (1980), Biometrika, A Note on the Generalized Information Criterion for Choice of a Model
- Bozdogan, H. (1987), Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions, 345-370
- Burnham, K. P., and Anderson, D. R. (1998), Model selection and inference: a practical information theoretic approach. Springer-Verlag, New York, NY
- Burnham, K. P., Anderson, D. R. and White, G. C. (1998), Comparison of AIC and CAIC for model selection and statistical inference from capture-recapture studies,25:263-282
- Burnham, K.P. and Anderson, D.R. (2002), Model Selection and Multimodel Inference: a practical information-theoretic approach, second edition. Springer-Verlag, New York
- Burnham, K.P. and Anderson D.R. (2004), Multimodel Inference-Understanding AIC and BIC in Model selection
- Cavanaugh Joseph E. (2012), Model selection, lecture 3, Corrected AIC and Modified AIC, AICc and MAIC
- Census Bureau, (2009),Statistical abstract of the United States. Government Printing Office
- Chi Yau (2014), R Tutorial e-book with Bayesian Statistics
- Claeskens, G., & Hjort, N. L. (2008).,Model selection and model averaging (Vol. 330). Cambridge: Cambridge University Press.
- Coffman Donna L. , Dziak John J. , Stephanie T. Lanza, and Runze Li, (2015) Sensitivity and Specificity of Information Criteria
- Collins LM, Lanza ST, (2010), Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences, New York
- Cover, T. M., and Thomas, J. A, (1991), John Elements of information theory. Wiley and Sons, New York, NY. 542pp

- Ghosh J. K., Delampady M., and Samanta T. , (2009), An Introduction to Bayesian Analysis: Theory and Methods. Springer-Verlag, New York
- Gingrich P. (2004), Chapter 11, Assosiation between variables, University of Regina
- Hurvich, C. M. and Tsai, C-L. (1989), Regression and time series model selection in small samples. Biometrika 76, 297-307
- Kadane, J. B., & Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American statistical Association*, *99*(465), 279-290.
- Kullback, S., and Leibler, R. A. (1951), On information and sufficiency, Annals of Mathematical Statistics 22 79-86.
- Kullback, S. (1959), John Wiley and Sons, Information theory and statistics New York, NY.
- Lance, Charles E., (2005), "Mallows' Cp Statistic I." *Wiley StatsRef: Statistics Reference Online*
- Mallows, Colin L. (1973), "Some comments on C p." *Technometrics* 15.4 : 661-675
- Mazerolle Marc J. (2007), Appendix 1:making sense out of AIC
- McQuarrie, A. D., & Tsai, C. L. (1998). *Regression and time series model selection*. World Scientific
- Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986), KTK Scientific Akaike information criterion statistics. Publishers, Tokyo
- Sadanorini Konishi, Genshiro Kitagawa, Information Criteria and Statistical Modeling, 2008, Springer Series in Statistics
- Sclove, S. L. (1987), Application of model-selection criteria to some problems in multivariate analysis. Psychometrika, 52, 333-343
- Schwarz, G. (1978), Estimating the dimension of a model, 461-464.
- Shuhua Hu (2007), Akaike Information Criterion, Center for research in Scientific Computation
- Takeuchi, K. (1976), Distribution of informational statistics and a criterion of model fitting. Suri-Kagaku (Mathematic Sciences)

Ελληνική

- Ζήμερας Σ. (2003) Πανεπιστήμιο Αιγαίου, Στατιστικά Πακέτα 1
- Καρώνη Χ. (2009),Μοντέλα Αξιοπιστίας Και Επιβίωσης
- Κολυβά-Μαχαίρα Φ., Μπόρα – Σέντα Ε. (1995), Θεωρία και Εφαρμογές
- Κούτρας Μ.-Μπούτσικας Μ. (2012), Σημειώσεις παραδόσεων του μαθήματος Στατιστική ΙΙ

- Πανάρετος Ι. & Ε.Ξεκαλάκη,(2000),Εισαγωγή στη Στατιστική Σκέψη,Τόμος ΙΙ, Εισαγωγή στις Πιθανότητες και την Στατιστική Συμπερασματολογία
- Παπαδόπουλος Γ.-Ανάλυση Παλινδρόμησης- Εργαστήριο. Μαθηματικών & Στατιστικής ,pdf, http://www.aua.gr/gpapadopoulos/files/regression9.pdf
- Φουσκάκης, Στατιστική Συμπερασματολογία, ΣΕΜΦΕ pdf

Ιστοσελίδες

- https://en.wikipedia.org/wiki/Statistics
- https://en.wikipedia.org/wiki/Akaike_information_criterion
- https://en.wikipedia.org/wiki/Bayesian_information_criterion
- https://en.wikipedia.org/wiki/Stepwise_regression
- https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence