ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΚΑΤΕΥΘΥΝΣΗ ΜΑΘΗΜΑΤΙΚΩΝ ΕΦΑΡΜΟΓΩΝ

Επιλογή μοντέλων βάσει μεθόδων Bootstrap και Jackknife

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΝΙΚΟΛΑΟΥ Α. ΕΛΕΥΘΕΡΙΟΥ

Αθήνα, Φεβρουάριος 2015
Επιβλέπουσα Καθηγήτρια : Φιλία Βόντα

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΚΑΤΕΥΘΥΝΣΗ ΜΑΘΗΜΑΤΙΚΩΝ ΕΦΑΡΜΟΓΩΝ

Επιλογή μοντέλων βάσει μεθόδων Bootstrap και Jackknife

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΝΙΚΟΛΑΟΥ Α. ΕΛΕΥΘΕΡΙΟΥ

Επιβλέπουσα : Φιλία Βόντα
Επικ. Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 5η Φεβρουαρίου 2015.

Φιλία Βόντα	Χρυσηίς Καρώνη	Χρήστος Κουκουβίνος
Επικ. Καθηγήτρια Ε.Μ.Π.	Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2015

Περίληψη

Η ανάλυση παλινδρόμησης αποτελεί ένα στατιστικό εργαλείο για την διερεύνηση των σχέσεων μεταξύ μεταβλητών και ιδιαίτερα στην περιγραφή του τρόπου μεταβολής της σχέσης της εξαρτημένης μεταβλητής με μια ή περισσότερες ανεξάρτητες μεταβλητές. Η διερεύνηση γίνεται συνήθως με την συλλογή δεδομένων για τις υπό εξέταση μεταβλητές και εκτιμώντας την ποσοτική επίδραση που έχουν σε άλλες μεταβλητές. Οι σχέσεις αυτές μεταξύ των μεταβλητών παρουσιάζονται με κάποια μαθηματική εξίσωση και αποτελούν ένα στατιστικό μοντέλο. Συχνά όμως υπάρχει μεγάλη δυσκολία στην εύρεση εκείνου του μοντέλου στο οποίο όλα τα δεδομένα μας θα προσαρμόζονται. Για την εύρεση εκείνου του βέλτιστου μοντέλου και των μεταβλητών του που θα περιγράφει καλύτερα τα δεδομένα μας, έχουν αναπτυχθεί διάφορα κριτήρια επιλογής μοντέλων τα οποία συγκρίνοντας όλα τα δυνατά μοντέλα προσπαθούν να επιλέξουν το βέλτιστο από αυτά.

Σκοπός της εργασίας αυτής είναι η σύγκριση δύο κριτηρίων επιλογής μοντέλων, το AIC (Akaike's information criterion) και το BIC (Bayesian information criterion). Επιπλέον, συγκρίνεται η απόδοση των δύο κριτηρίων όταν αυτά συνδυάζονται με μεθόδους αναδειγματοληψίας όπως η Bootstrap και η Jackknife.

Πιο αναλυτικά στο πρώτο κεφάλαιο περιγράφεται η έννοια και ο ρόλος του στατιστικού μοντέλου. Ενώ στο δεύτερο κεφάλαιο παρουσιάζονται βασικές έννοιες που χρησιμοποιούνται στην εργασία, όπως για παράδειγμα της ανάλυσης παλινδρόμησης, της λογαριθμικής πιθανοφάνειας και του συντελεστή συσχέτισης.

Στο τρίτο κεφάλαιο αναλύονται τα κριτήρια επιλογής μοντέλων AIC (Akaike's Information Criterion) και BIC (Bayesian Information Criterion). Αναλύεται η έννοια της Kullback-Leibler πληροφορίας και δίνονται οι περιγραφές των αποδείξεων των δύο κριτηρίων.

Στο τέταρτο κεφάλαιο αναλύονται οι μέθοδοι αναδειγματοληψίας Bootstrap και Jackknife, οι βασικές τους έννοιες και δίνονται μερικά παραδείγματα εφαρμογής τους.

Στο τελευταίο κεφάλαιο παρουσιάζονται οι προσομοιώσεις που συγκρίνουν τα κριτήρια πληροφορίας για τις δύο μεθόδους αναδειγματοληψίας. Υπολογίζονται οι φορές που επιλέγεται το σωστό μοντέλο καθώς και τα σφάλματα τύπου I και II σε διάφορες προσομοιώσεις. Τα προγράμματα που χρησιμοποιήθηκαν κατασκευάστηκαν στο R 3.0.2 πακέτο.

Λέξεις κλειδιά: «Ανάλυση Παλινδρόμησης, AIC, BIC, Bootstrap, Jackknife, σφάλματα τύπου I and II»

Abstract

Regression analysis is a statistical tool for investigating the relationships between variables and especially for specifying how the relationship between a dependent variable and one or more independent variables varies. The investigation, often, involves the assembly of data on the variables under study and estimating the quantitative effect that they have on the other variables. These relationships are represented as mathematical equations and depict a statistical model. Though often, there is great difficulty in finding that model that best fits all our data. In order to find that optimal model and the variables that will best describe our data, different model selection criteria have been developed which by comparing all possible models they try to select the optimal one.

The purpose of this thesis is the comparison of two model selection criteria, the AIC (Akaike's Information Criterion) and BIC (Bayesian Information Criterion). Moreover, the performance of the two criteria is compared in cases where they are combined with resampling methods like the Bootstrap and the Jackknife.

More specifically, in the first chapter the concept and role of the statistical model is described. While in the second one basic concepts that are used in the thesis are introduced, like for example the theory of regression analysis, the log likelihood and the correlation coefficient.

In the third chapter, the model selection criteria AIC and BIC are analysed. The theory of Kullback-Leibler is presented as well as descriptions of the proofs of the two criteria.

In the fourth chapter the resampling methods Bootstrap and Jackknife are being introduced. The basic concepts behind the methods and a few examples with applications are given.

In the last chapter, simulations through which we compare the criteria for both resampling methods are being presented. The percentage of times that a correct model is selected is being computed as well as the errors of type I and II in different cases. The code of the programs used has been written in R 3.0.2.

Keywords: «Regression Analysis, AIC, BIC, Bootstrap, Jackknife, errors type I and II»

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα καθηγήτρια κ. Φιλία Βόντα για την εμπιστοσύνη που μου έδειξε αναθέτοντας μου την εργασία αυτή. Την ευχαριστώ για την υπομονή και την αμέριστη καθοδήγησή της με καίριες υποδείξεις καθ'όλη την διάρκεια της εκπόνησης της διπλωματικής μου εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω ιδιαίτερα την οικογενειά μου για την υποστήριξη τους και την πολύτιμη βοήθεια τους σε όλη την διάρκεια των σπουδών μου.

Πίνακας περιεχομένων

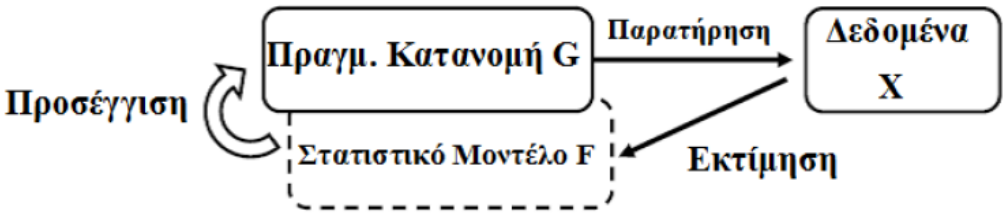
ΚΕΦΑΛΑΙΟ 1

ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ

1.1 Η Έννοια και ο Ρόλος της Στατιστικής Μοντελοποίησης

Η στατιστική μοντελοποίηση είναι ένα κρίσιμο ζήτημα στην ανάλυση δεδομένων. Τα μοντέλα χρησιμοποιούνται για να αναπαραστήσουν στοχαστικές δομές, να προβλέψουν μελλοντικές συμπεριφορές, και να εξάγουν χρήσιμες πληροφορίες από τα δεδομένα. Επομένως, διαδραματίζουν έναν κρίσιμο ρόλο στην ανάλυση των στατιστικών δεδομένων. Με την κατασκευή ενός μοντέλου, διάφορα συμπεράσματα, όπως πρόβλεψη, έλεγχος, εξαγωγή πληροφοριών, αξιολόγηση κινδύνων και λήψεις αποφάσεων μπορούν να πραγματοποιηθούν στο πλαίσιο της στατιστικής έρευνας. Επομένως το κλειδί για την επίλυση σύνθετων προβλημάτων του πραγματικού κόσμου έγκειται στην ανάπτυξη και την κατασκευή ενός κατάλληλου μοντέλου.

Ένα στατιστικό μοντέλο, σύμφωνα με Konishi and Kitagawa (2008), είναι μια κατανομή πιθανότητας που χρησιμοποιεί παρατηρούμενα δεδομένα, δηλαδή το δείγμα, ώστε να προσεγγίσει την αληθινή κατανομή αυτών που θέλει ο ερευνητής να αναλύσει. Άρα ο ρόλος της στατιστικής μοντελοποίησης είναι η κατασκευή ενός μοντέλου που να προσεγγίζει την πραγματική δομή, με όσο μεγαλύτερη ακρίβεια με την χρήση των διαθέσιμων δεδομένων. Όπως φαίνεται και στο παρακάτω σχήμα:



Σχήμα 1.1 Εκτίμηση πραγματικής κατανομής με βάση την στατιστική μοντελοποίηση

Πηγή: Konishi S. & Kitagawa G. (2008)

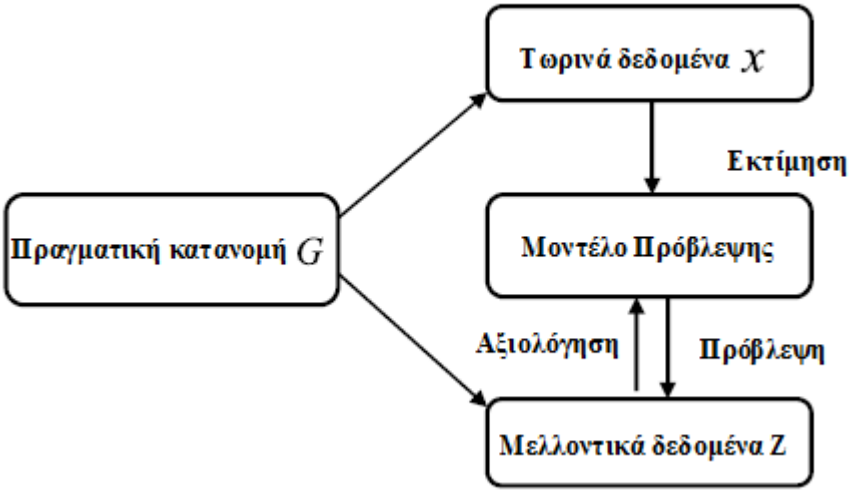
Ένα χαρακτηριστικό παράδειγμα, είναι η προσαρμογή ενός μοντέλου παλινδρόμησης, μια διαδικασία που περιλαμβάνει τον εντοπισμό του «πραγματικού συνόλου των επεξηγηματικών μεταβλητών». Ενώ για παράδειγμα στην προσαρμογή ενός πολυωνυμικού μοντέλου παλινδρόμησης ή αυτοπαλίνδρομων μοντέλων, απαιτείται η επιλογή της πραγματικής τάξης. Ωστόσο, είναι σπάνιο γραμμικά μοντέλα παλινδρόμησης με έναν πεπερασμένο αριθμό επεξηγηματικών μεταβλητών ή AR (αυτοπαλίνδρομα) μοντέλα με πεπερασμένη τάξη να μπορούν να εκφράσουν την πραγματική δομή. Ως εκ τούτου, αυτά τα μοντέλα πρέπει να θεωρηθούν ως προσεγγίσεις που αντιπροσωπεύουν μία μόνο πτυχή των πολύπλοκων φαινομένων. Το σημαντικό ζήτημα εδώ είναι η επιδίωξη μιας δομής που να είναι όσο το δυνατόν «πλησιέστερη» προς το πραγματικό μοντέλο.

1.1.1 Προβλέψεις σύμφωνα με τα Στατιστικά Μοντέλα

Ο Akaike ξεχώρισε ένα ιδιαίτερο πρόβλημα στην επιλογή ενός τέτοιου «καλού» μοντέλου, αυτό της πρόβλεψης [Akaike (1974, 1985)]. Έκρινε ότι ο σκοπός της στατιστικής μοντελοποίησης δεν είναι να περιγράψει με ακρίβεια τωρινά δεδομένα ή να συναγάγει την "αληθινή κατανομή". Αντιθέτως, θεώρησε ότι ο σκοπός των στατιστικών μοντέλων είναι να προβλέψουν τα μελλοντικά δεδομένα, όσο το δυνατόν ακριβέστερα.

Μπορεί να μην υπάρχει σημαντική διαφορά μεταξύ της συμπερασματολογίας της πραγματικής δομής και της πρόβλεψης, εάν μια μεγάλη ποσότητα δεδομένων είναι διαθέσιμη ή εάν τα δεδομένα είναι χωρίς θόρυβο. Ωστόσο, στη μοντελοποίηση βασιζόμενοι σε μια πεπερασμένη ποσότητα πραγματικών δεδομένων, υπάρχει ένα σημαντικό χάσμα μεταξύ αυτών των δύο σημείων, δεδομένου ότι ένα βέλτιστο μοντέλο με στόχο την πρόβλεψη (*Σχήμα 1.2*) μπορεί να διαφέρει από ένα μοντέλο που λαμβάνεται εκτιμώντας το «πραγματικό μοντέλο».

Στην πραγματικότητα, όπως προκύπτει από τα κριτήρια πληροφοριών που υπάρχουν για την αξιολόγηση των μοντέλων που προορίζονται για την κατασκευή προβλέψεων, απλά μοντέλα, ακόμα και αυτά που περιέχουν μεροληψία, είναι συχνά ικανά να δώσουν καλύτερες κατανομές πρόβλεψης από τα μοντέλα που λαμβάνονται με την εκτίμηση της πραγματικής δομής.



Σχήμα 1.2 Στατιστική μοντελοποίηση και το μοντέλο πρόβλεψης

Πηγή: Konishi S. & Kitagawa G. (2008)

1.1.2 Εξαγωγή Πληροφοριών

Μια άλλη σημαντική χρήση των στατιστικών μοντέλων είναι η εξαγωγή πληροφοριών. Πολλά συμβατικά στατιστικά συμπεράσματα υποθέτουν ότι το "πραγματικό" μοντέλο που διέπει το αντικείμενο της μοντελοποίησης είναι μια γνωστή οντότητα, ή, τουλάχιστον, ότι ένα «πραγματικό» μοντέλο υπάρχει. Ενώ επιπλέον, έχουν υιοθετήσει την προσέγγιση του ορισμού ενός προβλήματος, ως τον υπολογισμό ενός μικρού αριθμού άγνωστων παραμέτρων, με δεδομένο ότι υπάρχει το "πραγματικό" μοντέλο και ότι αυτές οι παράμετροι περιέχονται στο μοντέλο. Ωστόσο, στην πρόσφατη τάση που κυριαρχεί, τα μοντέλα είναι εργαλεία ευκολίας που χρησιμοποιούνται για την εξαγωγή πληροφοριών και την συμπερασματολογία.

Σύμφωνα με αυτή την άποψη, ένα στατιστικό μοντέλο δεν είναι κάτι που υπάρχει στον αντικειμενικό κόσμο. Αντιθέτως, είναι κάτι που έχει κατασκευαστεί με βάση την γνώση και τις προσδοκίες των αναλυτών. Δηλαδή με την χρήση των γνώσεων του αναλυτή και με βάση την εμπειρία του παρελθόντος και τα δεδομένα και βασιζόμενος στο σκοπό της ανάλυσης, όπως για παράδειγμα ο ειδικός τύπος των πληροφοριών που πρέπει να εξαχθεί από τα δεδομένα και τι θα πρέπει να επιτευχθεί με την ανάλυση, επιτυγχάνεται η κατασκευή ενός μοντέλου. Ως εκ τούτου, αν ένα συγκεκριμένο μοντέλο λαμβάνεται ως αποτέλεσμα της στατιστικής μοντελοποίησης, αυτό δεν σημαίνει αναγκαστικά ότι το πραγματικό φαινόμενο συμπεριφέρεται σύμφωνα με το μοντέλο, με την αυστηρή έννοια του όρου. Τα πραγματικά γεγονότα είναι πολύπλοκα, περιέχουν διαφόρων ειδών μη γραμμικότητες και μη στασιμότητες. Επιπλέον, σε πολλές περιπτώσεις θα πρέπει να θεωρηθούν ότι υπόκεινται στην επιρροή άλλων μεταβλητών. Ακόμη και σε τέτοιες καταστάσεις, ωστόσο, ένα σχετικά απλό μοντέλο συχνά αποδεικνύεται ότι είναι πιο κατάλληλο για την επίτευξη ενός συγκεκριμένου σκοπού. Επομένως, η ουσία του θέματος δεν είναι, εάν δεδομένου ενός στατιστικού μοντέλου αντιπροσωπεύεται με ακρίβεια η πραγματική δομή ενός φαινομένου, αλλά εάν είναι κατάλληλο ως ένα εργαλείο για την εξαγωγή χρήσιμων πληροφοριών από τα δεδομένα.

1.2 Κατασκευή Στατιστικών Μοντέλων

1.2.1 Αξιολόγηση μοντέλων

Αν ο ρόλος ενός στατιστικού μοντέλου εκφράζεται ως ένα εργαλείο για την εξαγωγή πληροφοριών, προκύπτει ότι ένα μοντέλο δεν είναι κάτι που καθορίζεται μοναδικά για ένα δεδομένο αντικείμενο αλλά μπορεί να αναλάβει μία ποικιλία μορφών ανάλογα με την οπτική γωνία του δημιουργού του μοντέλου και των διαθέσιμων πληροφοριών. Δηλαδή, ο σκοπός της στατιστικής μοντελοποίησης δεν είναι να εκτιμήσει ή να προσδιορίσει το «μοναδικό» ή το «τέλειο » μοντέλο , αλλά να κατασκευάσει ένα «καλό» μοντέλο, ως ένα εργαλείο για την εξαγωγή πληροφοριών, σύμφωνα με τα χαρακτηριστικά του αντικειμένου και του σκοπού της μοντελοποίησης [Akaike and Kitagawa (1998), Chapter 23].

Αυτό σημαίνει ότι, κατά γενικό κανόνα, τα αποτελέσματα της αξιολόγησης και συμπερασματολογίας θα ποικίλουν ανάλογα με το συγκεκριμένο μοντέλο. Ένα καλό μοντέλο θα δώσει γενικά καλά αποτελέσματα. Ωστόσο, δεν μπορεί κανείς να αναμένει να επιτύχει καλά αποτελέσματα, όταν χρησιμοποιείται ένα ακατάλληλο μοντέλο. Εδώ έγκειται η σημασία των κριτηρίων αξιολόγησης μοντέλων για την αξιολόγηση της « καλής προσαρμογής » ενός μοντέλου.

Κατά την εξέταση των συνθηκών υπό των οποίων χρησιμοποιούνται τα στατιστικά μοντέλα, ο Akaike θεώρησε ότι ένα μοντέλο πρέπει να αξιολογηθεί σε σχέση με την καλή προσαρμογή των αποτελεσμάτων, όταν το μοντέλο χρησιμοποιείται για την πρόβλεψη. Επιπλέον, για την γενική αξιολόγηση της καλής προσαρμογής ενός στατιστικού μοντέλου, θεώρησε ότι είναι σημαντικό να αξιολογήσει την εγγύτητα μεταξύ της κατανομής πρόβλεψης f(x), η οποία ορίζεται από το μοντέλο και την πραγματική κατανομή g(x) και όχι απλώς από την ελαχιστοποίηση του σφάλματος πρόβλεψης. Με βάση την έννοια αυτή, πρότεινε την αξιολόγηση των στατιστικών μοντέλων, σύμφωνα με την πληροφορία Kullback – Leibler ή αλλιώς απόκλιση [Akaike (1973)].

Το κριτήριο αξιολόγησης μοντέλων που προέρχεται από αυτή τη θεμελιώδη ιδέα που βασίζεται στην πληροφορία Kullback-Leibler, αποτελεί το κριτήριο πληροφορίας που θα αναλυθεί παρακάτω. Το κριτήριο πληροφορίας αυτό, εξάγεται από τρεις θεμελιώδεις έννοιες: (1) από την μοντελοποίηση που βασίζεται στην πρόβλεψη (2) από την αξιολόγηση της ακρίβειας της πρόβλεψης σύμφωνα με τις κατανομές και (3) από την αξιολόγηση της εγγύτητας των κατανομών σύμφωνα με την πληροφορία Kullback-Leibler.

Από το κριτήριο πληροφορίας κατασκευάζονται αρκετές μέθοδοι για την ανάπτυξη καλών μοντέλων, τα οποία βασίζονται σε περιορισμένο αριθμό δεδομένων. Αρχικά, είναι προφανές ότι όσο μεγαλύτερη η λογαριθμική πιθανοφάνεια (log-likelihood) ενός μοντέλου τόσο καλύτερο είναι το μοντέλο. Το κριτήριο πληροφορίας δηλώνει, ωστόσο, ότι με δεδομένη μια πεπερασμένη ποσότητα διαθέσιμων δεδομένων για μοντελοποίηση, ένα μοντέλο που έχει υπερβολικά υψηλούς βαθμούς ελευθερίας θα οδηγήσει σε αύξηση της αστάθειας του εκτιμώμενου μοντέλου, και αυτό θα έχει ως αποτέλεσμα την μείωση της ικανότητας πρόβλεψης, όπως περιγράφεται και παρακάτω.

ΚΕΦΑΛΑΙΟ 2

ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

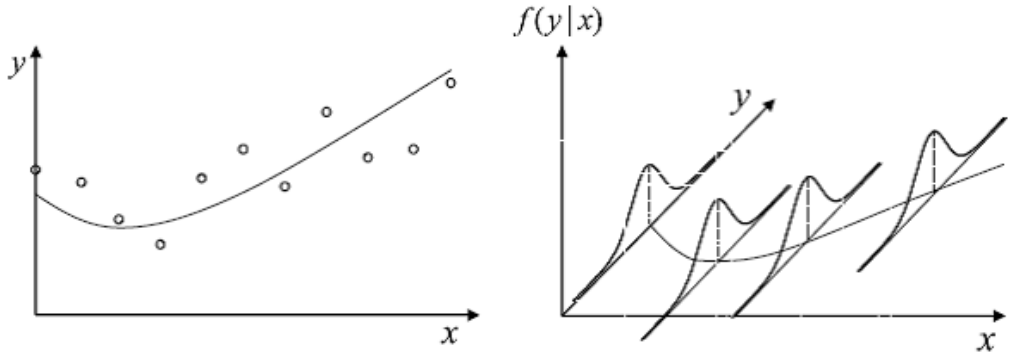
2.1 Μοντέλο Παλιδρόμησης

Στην στατιστική, σύμφωνα με Konishi and Kitagawa (2008) και Xin Yan and Xiao Gang Su (2009), η ανάλυση παλινδρόμησης αποτελείται από τεχνικές για την μοντελοποίηση της σχέσης μιας βαθμωτής εξαρτημένης μεταβλητής ή αλλιώς και μεταβλητής απόκρισης y με μία ή περισσότερες ανεξάρτητες ή αλλιώς και επεξηγηματικές μεταβλητές $\vec{x} = (x_1, x_2, .. x_p)^T$. Δηλαδή η κατανομή πιθανότητας της μεταβλητής απόκρισης y μεταβάλλεται εξαρτώμενη από τις επεξηγηματικές μεταβλητές $\vec{x}$. Αυτό έχει ως αποτέλεσμα η δεσμευμένη κατανομή να δίνεται με την μορφή f(y|$\vec{x}$). Η εξαρτημένη μεταβλητή y επομένως, διαμορφώνεται ως συνάρτηση των ανεξάρτητων μεταβλητών, των αντίστοιχων παραμέτρων παλινδρόμησης (συντελεστές), και ενός τυχαίου όρου σφάλματος που εκφράζει την διακύμανση της εξαρτώμενης μεταβλητής εξαιτίας παραγόντων που δεν είναι συναρτήσει των εξαρτώμενων μεταβλητών ή των αντίστοιχων συντελεστών.

Αν {(y_α ,$\vec{x}_\alpha$), α=1,2..n} είναι ένα σύνολο n δεδομένων της y μεταβλητής και του $\vec{x} = (x_1, x_2, .. x_p)^T$, τότε το μοντέλο:

$$y_\alpha = u(\vec{x}_\alpha) + \varepsilon_\alpha , \qquad \alpha = 1,2,\ldots n$$

των παρατηρούμενων δεδομένων, ονομάζεται μοντέλο παλινδρόμησης. Το u($\vec{x}$) είναι μια συνάρτηση της επεξηγηματικής μεταβλητής $\vec{x}$ και οι όροι σφάλματος ή «θόρυβος» ε_α θεωρούνται ανεξάρτητα κατενεμημένοι με μέση τιμή E[ε_α]=0 και διακύμανση V(ε_α)=σ^2. Θεωρείται ότι ο θόρυβος ε_α ακολουθεί την κανονική κατανομή N(0,σ^2).



Σχήμα 2.1 Μοντέλο παλινδρόμησης (αριστερά) και μοντέλο δεσμευμένης κατανομής (δεξιά) στο οποίο η μέση τιμή της μεταβλητής απόκρισης κυμαίνεται ως συνάρτηση της επεξηγηματικής μεταβλητής x.

Πηγή: Konishi S. & Kitagawa G. (2008)

Σε αυτή την περίπτωση το y_α ακολουθεί την κανονική κατανομή με $N(u(\vec{x}_\alpha),\sigma^2)$ και η συνάρτηση πυκνότητας δίνεται από τον τύπο:

$$f(y_\alpha|\vec{x}_\alpha) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y_\alpha - u(\vec{x}_\alpha)^2}{2\sigma^2}\right\}, \quad \alpha = 1,2,\dots n$$

Αυτή η κατανομή είναι ένα είδος δεσμευμένης κατανομής, όπου η μέση τιμή είναι $E[Y|\vec{x}]=u(\vec{x})$, εξαρτώμενη δηλαδή από τις τιμές των μεταβλητών $\vec{x}$.

Ο αριστερός πίνακας στο Σχήμα 2.1 δείχνει 11 παρατηρήσεις και την μέση συνάρτηση u(x) της μονοδιάστατης επεξηγηματικής μεταβλητής x και την μεταβλητή απόκρισης y. Τα δεδομένα y_α στο σημείο x_α παρατηρούνται ως:

$$y_\alpha = \mu_\alpha + \varepsilon_\alpha, \quad \alpha = 1,2,..,11$$

με μέση τιμή $E[Y_\alpha|x_\alpha]= \mu_\alpha$ και ε_α ο θόρυβος. Ο δεξιός πίνακας δείχνει μια δεσμευμένη κατανομή που προσδιορίζεται χρησιμοποιώντας ένα μοντέλο παλινδρόμησης. Καθορίζοντας δεδομένη τιμή της επεξηγηματικής μεταβλητής x κατασκευάζεται η κατανομή πιθανότητας f(y|x), της οποίας η μέση τιμή είναι u(x). Οπότε το μοντέλο παλινδρόμησης καθορίζει μια κλάση κατανομών όπως φαίνεται και στο σχήμα.

2.1.1 Ο ρόλος και οι τύποι της ανάλυσης παλινδρόμησης

Η ανάλυση παλινδρόμησης είναι μια διαδικασία που χρησιμοποιείται για την εκτίμηση μιας συνάρτησης, η οποία προβλέπει την τιμή της μεταβλητής απόκρισης σε σχέση με τις άλλες ανεξάρτητες μεταβλητές. Εάν η συνάρτηση παλινδρόμησης προσδιορίζεται μόνο μέσω ενός συνόλου παραμέτρων, το είδος της παλινδρόμησης ονομάζεται παραμετρική παλινδρόμηση. Πολλές μέθοδοι έχουν αναπτυχθεί για τον προσδιορισμό διαφόρων παραμετρικών σχέσεων μεταξύ της μεταβλητής απόκρισης

και των ανεξάρτητων μεταβλητών. Αυτές οι μέθοδοι τυπικά εξαρτώνται από την μορφή της παραμετρικής συνάρτησης παλινδρόμησης και την κατανομή του όρου σφάλματος σε ένα μοντέλο παλινδρόμησης. Τέτοια μοντέλα για παράδειγμα, είναι η γραμμική παλινδρόμηση, λογιστική παλινδρόμηση, ή παλινδρόμηση Poisson. Αυτά τα μοντέλα παλινδρόμησης εκφράζονται με διαφορετικές συναρτήσεις παλινδρόμησης και όρους σφάλματος από τις αντίστοιχες κατανομές τους.

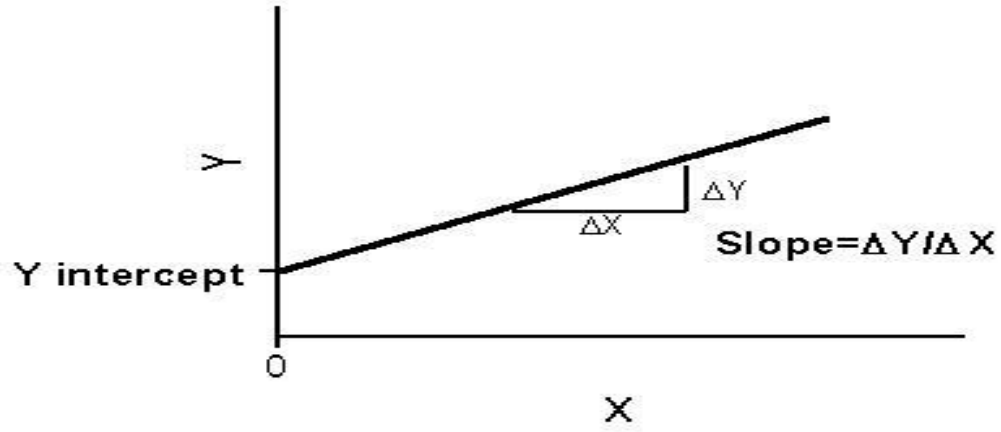
Στην γραμμική παλινδρόμηση, η εξαρτημένη μεταβλητή μοντελοποιείται ως μια γραμμική συνάρτηση ενός συνόλου παραμέτρων παλινδρόμησης και ενός τυχαίου σφάλματος. Οι παράμετροι πρέπει να εκτιμηθούν ώστε το μοντέλο να δώσει την «καλύτερη προσαρμογή» στα δεδομένα. Αυτή η εκτίμηση γίνεται με βάση κάποιο προκαθορισμένο κριτήριο. Αν ένα μοντέλο παλινδρόμησης επαρκώς εκφράζει την πραγματική σχέση μεταξύ της μεταβλητής απόκρισης και των ανεξάρτητων μεταβλητών, αυτό το μοντέλο μπορεί να χρησιμοποιηθεί για την πρόβλεψη εξαρτώμενων μεταβλητών, προσδιορίζοντας τις σημαντικές ανεξάρτητες μεταβλητές, και να εδραιώσει την ύπαρξη αιτιώδους σχέσης μεταξύ της μεταβλητής απόκρισης και των ανεξάρτητων μεταβλητών. Στην ανάλυση παλινδρόμησης, επομένως, ο ερευνητής συγκεντρώνει συχνά τα δεδομένα των υπό εξέταση μεταβλητών και χρησιμοποιεί το μοντέλο παλινδρόμησης για την εκτίμηση της ποσοτικής αιτιατής επίδρασης μεταξύ των ανεξάρτητων μεταβλητών και της μεταβλητής απόκρισης. Συνήθως αξιολογεί την «στατιστική σημαντικότητα» της εκτιμώμενης σχέσης μεταξύ τους, δηλαδή, τον βαθμό εμπιστοσύνης σχετικά με το πώς η πραγματική σχέση είναι κοντά στην εκτιμώμενη στατιστική σχέση.

Η γραμμική παλινδρόμηση προϋποθέτει ότι το μοντέλο είναι γραμμικό σε σχέση με τις παραμέτρους παλινδρόμησης. Στόχος είναι να καταγράψει πώς η αναμενόμενη μέση τιμή του Y μεταβάλλεται όταν μεταβάλλονται μια ή περισσότερες ανεξάρτητες μεταβλητές X. Όπως όλες οι μορφές ανάλυσης παλινδρόμησης, η γραμμική παλινδρόμηση εστιάζει στην υποθετική κατανομή πιθανότητας του Y με δεδομένο το X, αντί για την από κοινού κατανομή πιθανότητας των Y και X, το οποίο είναι τομέας της πολυμεταβλητής ανάλυσης (όταν υπάρχουν παραπάνω από μια εξαρτημένες μεταβλητές).

Υπάρχουν τρεις τύποι παλινδρόμησης. Η πρώτη είναι η απλή γραμμική παλινδρόμηση. Η απλή γραμμική παλινδρόμηση χρησιμοποιείται για την μοντελοποίηση της γραμμικής σχέσης μεταξύ δύο μεταβλητών. Μια από αυτές είναι η εξαρτημένη μεταβλητή Y και η άλλη είναι η ανεξάρτητη μεταβλητή x. Το απλό μοντέλο παλινδρόμησης γράφεται συχνά στην παρακάτω μορφή:

$$y=\beta_0+\beta_1 x+\varepsilon,$$

όπου y είναι η εξαρτημένη μεταβλητή, β_0 είναι η τεταγμένη του y όταν το x=0 (intercept), β_1 είναι η κλίση (slope) της ευθείας παλινδρόμησης και εκφράζει την μεταβολή κατά β_1 μονάδες του y για κάθε μονάδα που αυξάνεται το x, x είναι η ανεξάρτητη μεταβλητή, και ε είναι το τυχαίο σφάλμα (Σχήμα 2.2). Το ε, θεωρείται συνήθως ότι ακολουθεί την κανονική κατανομή με $E(\varepsilon) = 0$ και διακύμανση ίση με $Var(\varepsilon) = \sigma^2$ στην απλή γραμμική παλινδρόμηση.



Σχήμα 2.2 Μοντέλο Γραμμικής Παλινδρόμησης

Πηγή: http://www.camo.com/rt/Resources/linear_regression_model.html

Ο δεύτερος τύπος παλινδρόμησης είναι η πολλαπλή γραμμική παλινδρόμηση το οποίο είναι ένα γραμμικό μοντέλο παλινδρόμησης με μία εξαρτημένη μεταβλητή και περισσότερες από μία ανεξάρτητες μεταβλητές. Η πολλαπλή γραμμική παλινδρόμηση υποθέτει ότι η μεταβλητή απόκρισης είναι μία γραμμική συνάρτηση των παραμέτρων του μοντέλου και υπάρχουν περισσότερες από μία ανεξάρτητες μεταβλητές στο μοντέλο. Η γενική μορφή του μοντέλου πολλαπλής γραμμικής παλινδρόμησης έχει ως εξής:

$$y=\beta_0+\beta_1x_1+...+\beta_px_p+\varepsilon,$$

όπου y είναι η εξαρτημένη μεταβλητή, β_0, β_1, β_2, ..., β_p είναι οι συντελεστές παλινδρόμησης, και x_1, x_2, ..., x_p είναι οι ανεξάρτητες μεταβλητές στο μοντέλο. Στην κλασική παλινδρόμηση συνήθως θεωρείται ότι ο όρος σφάλματος ε ακολουθεί την κανονική κατανομή με $E(\varepsilon) = 0$ και σταθερή διακύμανση $Var(\varepsilon) = \sigma^2$.

Η διαφορά των δύο αυτών τύπων έγκειται στο γεγονός ότι απλή γραμμική παλινδρόμηση, όπως αναφέρθηκε, χρησιμοποιείται για την διερεύνηση της γραμμικής σχέσης ανάμεσα σε μια εξαρτημένη μεταβλητή και μία ανεξάρτητη μεταβλητή, ενώ η πολλαπλή γραμμική παλινδρόμηση επικεντρώνεται στην γραμμική σχέση ανάμεσα σε μια εξαρτημένη μεταβλητή και περισσότερες από μία ανεξάρτητες μεταβλητές.

Ο τρίτος τύπος της παλινδρόμησης είναι η μη γραμμική παλινδρόμηση, η οποία υποθέτει ότι η σχέση μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών δεν είναι γραμμική. Ένα παράδειγμα μοντέλου μη γραμμικής παλινδρόμησης (μοντέλο ανάπτυξης) μπορεί να γραφτεί ως εξής:

$$y = \frac{a}{1 + e^{\beta t}} + \varepsilon$$

όπου Y είναι η ανάπτυξη ενός συγκεκριμένου οργανισμού ως συνάρτηση του χρόνου t, α και β είναι οι παράμετροι του μοντέλου, και ε είναι το τυχαίο σφάλμα. Το μοντέλο μη γραμμικής παλινδρόμησης είναι πιο περίπλοκο από το μοντέλο

γραμμικής παλινδρόμησης από την άποψη της εκτίμησης των παραμέτρων του μοντέλου, της επιλογής του μοντέλου καθώς και της επιλογής των μεταβλητών.

Παράδειγμα Γραμμικού μοντέλου παλινδρόμησης

Αν η συνάρτηση παλινδρόμησης ή η συνάρτηση της μέσης τιμής u($\vec{x}$) μπορούν να προσεγγιστούν από μια γραμμική συνάρτηση του $\vec{x}$ τότε το μοντέλο μπορεί να εκφραστεί ως:

$$y_a = \beta_0 + \beta_1 x_{\alpha 1} + \cdots \beta_p x_{\alpha p} + \varepsilon_\alpha = \vec{x}_\alpha^T \vec{\beta} + \varepsilon_\alpha, \quad \alpha = 1,2,..,n$$

με $\vec{\beta} = (\beta_0, \beta_1, \ldots \beta_p)^T, x_\alpha = \left(1, x_{\alpha 1}, \ldots x_{\alpha p}\right)^T$να αποτελεί το γραμμικό μοντέλο παλινδρόμησης. Με την προσθήκη του θορύβου, η συνάρτηση πυκνότητας μπορεί να εκφραστεί ως :

$$f(y_\alpha | \vec{x}_\alpha; \vec{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_\alpha - x_\alpha^T \beta)^2}{2\sigma^2}\right\}, \quad \alpha = 1,2, \ldots, n$$

όπου οι άγνωστες παράμετροι στο μοντέλο είναι οι $\vec{\theta} = (\beta^T, \sigma^2)^T$. Το κύριο σημείο είναι ο προσδιορισμός του συνόλου των επεξηγηματικών μεταβλητών που θα περιγράφει κατάλληλα τις αλλαγές στην κατανομή της μεταβλητής απόκρισης. Αυτό αποτελεί το πρόβλημα επιλογής μεταβλητών (variable selection problem).

Στην περίπτωση που το μοντέλο παλινδρόμησης μπορεί να εκφραστεί από μια συνάρτηση πυκνότητας, εκτιμάται το παραμετρικό διάνυσμα $\vec{\theta}$ του μοντέλου χρησιμοποιώντας την μέθοδο μέγιστης πιθανοφάνειας και ο εκτιμητής συμβολίζεται ως $\hat{\vec{\theta}} = (\hat{\beta}^T, \hat{\sigma}^2)^T$. Τότε η παραπάνω εξίσωση, στην οποία οι άγνωστες παράμετροι αντικαθίστανται με τις αντίστοιχες εκτιμήτριες τους γίνεται:

$$f(y_\alpha | \vec{x}_\alpha; \hat{\vec{\theta}}) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left\{-\frac{(y_\alpha - x_\alpha^T \hat{\beta})^2}{2\hat{\sigma}^2}\right\}, \quad \alpha = 1,2, \ldots, n$$

Η παραπάνω εξίσωση αποτελεί ένα στατιστικό μοντέλο.

2.2 Μέθοδος Μεγίστης Πιθανοφάνειας και οι Εκτιμητές Μεγίστης Πιθανοφάνειας

Ανάλυση της συνάρτησης λογαριθμικής πιθανοφάνειας και των εκτιμητών μεγίστης πιθανοφάνειας.

Έστω δοθέν μοντέλο με κατανομή πιθανότητας f(x|$\vec{\theta}$) ($\vec{\theta} \in \Theta \subset R^p$) με άγνωστη p-διάστατη παράμετρο $\vec{\theta} = \left(\theta_1, \theta_2, \ldots \theta_p\right)^T$. Σε αυτή την περίπτωση, σύμφωνα με Konishi and Kitagawa (2008), για δοθέντα δεδομένα $\vec{x}_n = \{x_1, x_2, \ldots, x_n\}$, η λογαριθμική πιθανοφάνεια μπορεί να προσδιορισθεί για κάθε $\vec{\theta} \in \Theta$. Επομένως,

θεωρώντας την λογαριθμική πιθανοφάνεια ως συνάρτηση του $\vec{\theta} \in \Theta$ μπορεί να εκφραστεί ως:

$$l(\vec{\theta}) = \sum_{\alpha=1}^{n} logf(x_\alpha|\vec{\theta}),$$

που αποτελεί την συνάρτηση λογαριθμικής πιθανοφάνειας. Ένας φυσικός εκτιμητής του $\vec{\theta}$ προσδιορίζεται υπολογίζοντας το μέγιστο $\vec{\theta} \in \Theta$ του $l(\vec{\theta})$. Δηλαδή, εκείνο το $\vec{\theta}$ που ικανοποιεί την εξίσωση:

$$l\left(\hat{\vec{\theta}}\right) = \max_{\vec{\theta}\in\Theta} l(\vec{\theta})$$

Αυτή η μέθοδος ονομάζεται μέθοδος μεγίστης πιθανοφάνειας, και το $\hat{\vec{\theta}}$ αποτελεί τον εκτιμητή μεγίστης πιθανοφάνειας. Το μοντέλο f(x|$\hat{\vec{\theta}}$) ονομάζεται μοντέλο μεγίστης πιθανοφάνειας και ο όρος $l\left(\hat{\vec{\theta}}\right) = \sum_{\alpha=1}^{n} logf\left(x_\alpha\middle|\hat{\vec{\theta}}\right)$ αποτελεί την μέγιστη λογαριθμική πιθανοφάνεια.

2.2.1 Εφαρμογή της μεθόδου μεγίστης πιθανοφάνειας.

Εάν η συνάρτηση λογαριθμικής πιθανοφάνειας $l(\vec{\theta})$ είναι συνεχώς διαφορίσιμη, τότε η εκτιμήτρια μεγίστης πιθανοφάνειας $\hat{\vec{\theta}}$ δίνεται ως λύση της εξίσωσης πιθανοφάνειας

$$\frac{\partial l(\vec{\theta})}{\partial \theta_i} = 0, \qquad i = 1,2,\dots,p \quad \text{ή} \quad \frac{\partial l(\vec{\theta})}{\partial \vec{\theta}} = \vec{0}$$

όπου $\frac{\partial l(\vec{\theta})}{\partial \vec{\theta}}$ είναι ένα p-διάστατο διάνυσμα, όπου το ι-οστό στοιχείο δίνεται από το $\frac{\partial l(\vec{\theta})}{\partial \theta_i}$ και $\vec{0}$ είναι το p-διάστατο μηδενικό διάνυσμα, του οποίου όλα τα στοιχεία είναι 0. Εάν η συνάρτηση πιθανοφάνειας είναι γραμμική εξίσωση με p-διάστατες παραμέτρους τότε ο εκτιμητής μεγίστης πιθανοφάνειας μπορεί να εκφραστεί ρητά.

Παράδειγμα (μοντέλο Bernoulli)

Η συνάρτηση λογαριθμικής πιθανοφάνειας βασιζόμενη σε n παρατηρήσεις $\{x_1, x_2, \dots, x_n\}$ από την κατανομή Bernoulli με $f(x|p) = p^x(1-p)^{1-x}, (x = 0,1)$ δίνεται από:

$$l(p) = \log\{\prod_{\alpha=1}^{n} p^{x_\alpha}(1-p)^{1-x_\alpha}\} = \sum_{\alpha=1}^{n} x_\alpha logp + (n - \sum_{\alpha=1}^{n} x_\alpha)\log(1-p)$$

Οπότε η εξίσωση πιθανοφάνειας είναι η :

$$\frac{\partial l(p)}{\partial p} = \frac{1}{p}\sum_{\alpha=1}^{n} x_\alpha - \frac{1}{1-p}\left(n - \sum_{\alpha=1}^{n} x_\alpha\right) = 0$$

Ο εκτιμητής μεγίστης πιθανοφάνειας για το p δίνεται από το

$$\hat{p} = \frac{1}{n}\sum_{\alpha=1}^{n} x_\alpha$$

Παράδειγμα (Κανονικό μοντέλο)

Έστω το μοντέλο κανονικής κατανομής με δεδομένα $\{x_1, x_2, \dots, x_n\}$. Αφού η λογαριθμική συνάρτηση δίνεται από τον τύπο

$$l(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{\alpha=1}^{n}(x_\alpha - \mu)^2$$

η λογαριθμική συνάρτηση θα έχει την μορφή:

$$\frac{\partial l(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{\alpha=1}^{n}(x_\alpha - \mu) = \frac{1}{\sigma^2}(\sum_{\alpha=1}^{n} x_\alpha - n\mu) = 0$$

$$\frac{\partial l(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{\alpha=1}^{n}(x_\alpha - \mu)^2 = 0$$

Με αποτέλεσμα οι εκτιμήτριες μεγίστης πιθανοφάνειας των μ και σ^2 να είναι

$$\hat{\mu} = \frac{1}{n}\sum_{\alpha=1}^{n} x_\alpha \ , \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{\alpha=1}^{n}(x_\alpha - \hat{\mu})^2$$

Παράδειγμα 3 (Απλό γραμμικό μοντέλο παλινδρόμησης)

Οι εκτιμητές μεγίστης πιθανοφάνειας της απλής γραμμικής παλινδρόμησης μπορούν να κατασκευαστούν, αν θεωρηθεί ότι η εξαρτημένη μεταβλητή y_i ακολουθεί την κανονική κατανομή: $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Η συνάρτηση πιθανοφάνειας για $(y_1, y_2, \dots, y_n)$ δίνεται από τον τύπο:

$$L = \prod_{i=1}^{n} f(y_i) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{\left(-\frac{1}{2\sigma^2}\right)\Sigma_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}$$

Οι εκτιμητές των β_0 και β_1 που μεγιστοποιούν τη συνάρτηση πιθανοφάνειας L είναι ισοδύναμες με τους εκτιμητές που ελαχιστοποιούν το εκθετικό μέρος της συνάρτησης πιθανοφάνειας.

Αφού έχουν αποκτηθεί οι b_1 και b_0, οι εκτιμήτριες μεγίστης πιθανοφάνειας των παραμέτρων β_0 και β_1, μπορεί να υπολογιστεί η προσαρμοσμένη τιμή $\hat{y}_i$, και η συνάρτηση πιθανοφάνειας σε σχέση με τις προσαρμοσμένες τιμές.

$$L = \prod_{i=1}^{n} f(y_i) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{\left(-\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

Στη συνέχεια λαμβάνεται η μερική παράγωγος σε σχέση με το σ^2 στην λογαριθμική συνάρτηση πιθανοφάνειας log(L) και ορίζεται ίση με το μηδέν:

$$\frac{\partial \log(L)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$$

Η εκτιμήτρια μεγίστης πιθανοφάνειας του σ^2 είναι $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Η εκτιμήτρια είναι μεροληπτική αφού το $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ είναι αμερόληπτη εκτιμήτρια του σ^2. Επίσης το $\frac{n}{n-2}\hat{\sigma}^2$ είναι ένας αμερόληπτος εκτιμητής του σ^2.

Παράδειγμα 4 (Γραμμικό μοντέλο παλινδρόμησης)

Έστω $\{y_\alpha, x_{\alpha 1}, x_{\alpha 2}, \dots, x_{\alpha p}\}$ (α=1,2,...,n) n σύνολα δεδομένων που παρατηρούνται αναφορικά με μια μεταβλητή απόκρισης y και p επεξηγηματικές μεταβλητές $\{x_1, x_2, \dots, x_p\}$. Για να περιγραφεί η σχέση μεταξύ των μεταβλητών, θεωρείται το παρακάτω μοντέλο γραμμικής παλινδρόμησης με Gaussian θόρυβο:

$$y_\alpha = \vec{x}_\alpha{}^T \vec{\beta} + \varepsilon_\alpha, \qquad \varepsilon_\alpha \sim N(0, \sigma^2), \quad \alpha = 1,2,\dots,n$$

όπου $\vec{x}_\alpha = (1, x_{\alpha 1}, x_{\alpha 2}, \dots, x_{\alpha p})^T$ και $\vec{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$. Αφού η συνάρτηση πυκνότητας πιθανότητας του y_α είναι

$$f(y_\alpha | \vec{x}_\alpha; \vec{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_\alpha - \vec{x}_\alpha{}^T \vec{\beta})^2\right\}$$

και η συνάρτηση λογαριθμικής πιθανοφάνειας μπορεί να εκφραστεί ως:

$$\begin{aligned} l(\vec{\theta}) &= \sum_{\alpha=1}^{n} log f(y_\alpha | \vec{x}_\alpha; \vec{\theta}) = \\ &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{\alpha=1}^{n}(y_\alpha - \vec{x}_\alpha{}^T \vec{\beta})^2 = \\ &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\vec{y} - \vec{X}\vec{\beta})^T(\vec{y} - \vec{X}\vec{\beta}) \end{aligned}$$

όπου $\vec{y} = (y_1, y_2, \dots, y_n)^T$ και $\vec{X} = (\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n)^T$. Υπολογίζοντας τις μερικές παραγώγους της παραπάνω εξίσωσης αναφορικά με το παραμετρικό διάνυσμα $\vec{\theta} = (\vec{\beta}^T, \sigma^2)^T$, η εξίσωση πιθανοφάνειας δίνεται από τον τύπο:

$$\frac{\partial l(\vec{\theta})}{\partial \vec{\beta}} = -\frac{1}{2\sigma^2}\left(-2\vec{X}^T\vec{y} + 2\vec{X}^T\vec{X}\vec{\beta}\right) = \vec{0},$$

$$\frac{\partial l(\vec{\theta})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\vec{y} - \vec{X}\vec{\beta}\right)^T\left(\vec{y} - \vec{X}\vec{\beta}\right) = 0$$

Άρα οι εκτιμήτριες μεγίστης πιθανοφάνειας για τα $\vec{\beta}$ και σ^2 θα δίνονται από:

$$\hat{\vec{\beta}} = (\vec{X}^T\vec{X})^{-1}\vec{X}^T\vec{y}, \qquad \hat{\sigma}^2 = \frac{1}{n}(\vec{y} - \vec{X}\hat{\vec{\beta}})^T(\vec{y} - \vec{X}\hat{\vec{\beta}})$$

2.3 Συντελεστής Συσχέτισης

Όπως αναφέρθηκε η ανάλυση παλινδρόμησης προσπαθεί να εντοπίσει την αιτιατή σχέση μεταξύ μεταβλητών. Ο συντελεστής συσχέτισης αποτελεί ένα μέτρο που μετράει το μέγεθος ή εναλλακτικά την ένταση της σχέσης μεταξύ δύο ποσοτικών μεταβλητών. Ένας τέτοιος συντελεστής που περιγράφει την γραμμική σχέση των μεταβλητών είναι και ο γραμμικός συντελεστής συσχέτισης Pearson. Ορίζεται ως το πηλίκο της συνδιασποράς δια τις τυπικές αποκλίσεις των μεταβλητών, συμβολίζεται με ρ ενώ η αντίστοιχη εκτίμηση του με r και γραφικά περιγράφει την συγκέντρωση των σημείων ενός διαγράμματος διασποράς γύρω από την ευθεία παλινδρόμησης.

Ο τύπος του συντελεστή συσχέτισης πληθυσμού για δύο μεταβλητές X, Y είναι:

$$\rho = \frac{cov(X,Y)}{\sigma_X\sigma_Y}$$

$$\text{όπου η συνδιακύμανση } cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Στην περίπτωση όμως, που ο συντελεστής γραμμικής συσχέτισης είναι άγνωστος, χρησιμοποιείται ο συντελεστής συσχέτισης δείγματος δύο μεταβλητών X και Y που ορίζεται με βάση ένα δείγμα n ζευγών παρατηρήσεων (x_i, y_i), i=1,2,…n και συμβολίζεται με r(X, Y).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^{n}(x_i - \bar{x})^2][\sum_{i=1}^{n}(y_i - \bar{y})^2]}}$$

Πιο συγκεκριμένα αν το ζεύγος των παρατηρήσεων προέρχεται από το διμεταβλητό Κανονικό πληθυσμό και το δείγμα είναι αρκετά μεγάλο, τότε ο συντελεστής συσχέτισης δείγματος αποτελεί την αμερόληπτη εκτιμήτρια μεγίστης πιθανοφάνειας του συντελεστή συσχέτισης του πληθυσμού. Παρατηρείται επίσης ότι για μεγάλες τιμές x_i και y_i τα $(x_i - \bar{x})(y_i - \bar{y})$ είναι θετικά και άρα και το γινομενό τους. Ομοίως

για μικρές τιμές των x_i και y_i τα $(x_i - \bar{x})$ $και$ $(y_i - \bar{y})$ είναι αρνητικά, άρα το γινόμενο τους θετικό. Άρα η συνδιακύμανση σε αυτές τις περιπτώσεις θα είναι θετικός αριθμός και οι μεταβλητές θα μεταβάλλονται ομόρροπα. Δηλαδή αύξηση σε μια μεταβλητή σχετίζεται με αύξηση στην άλλη μεταβλητή, τότε οι μεταβλητές θα είναι θετικά συσχετισμένες. Αν αύξηση σε μια μεταβλητή επιφέρει μείωση στην άλλη, τότε οι μεταβλητές θα είναι αρνητικά συσχετισμένες. Ενώ αν δύο μεταβλητές δεν έχουν σχέση μεταξύ τους, θα είναι ασυσχέτιστες. Όπως φαίνεται και στο παρακάτω σχήμα:

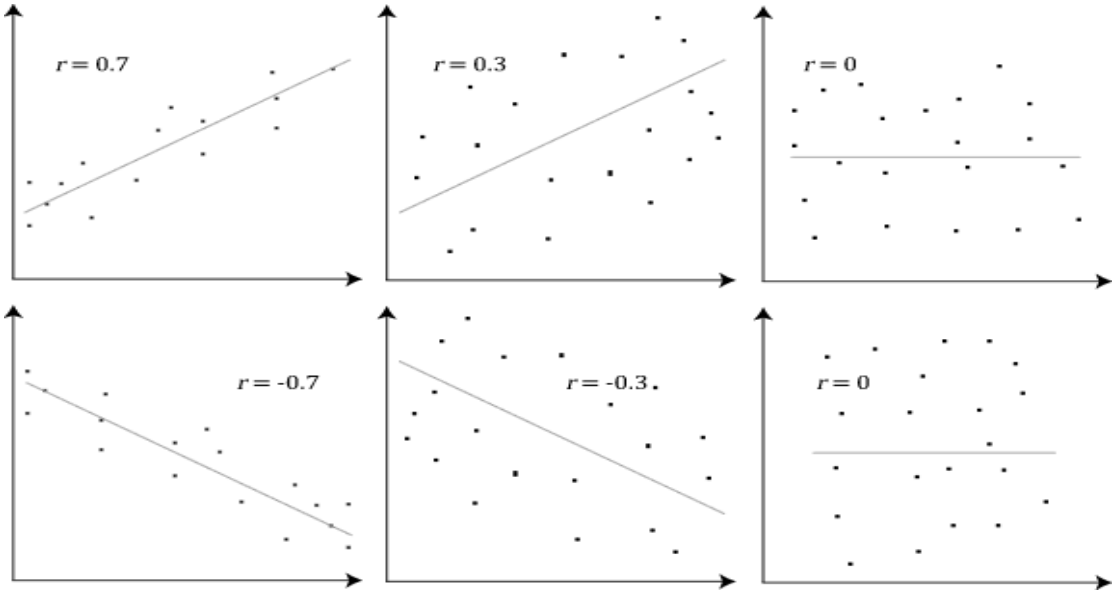
***Πηγή:* McLeod, S. A. (2008). Correlation. Retrieved from http://www.simplypsychology.org/correlation.html**

Ιδιότητες συντελεστή συσχέτισης.

1. Η συσχέτιση αφορά ποσοτικές μεταβλητές.
2. Η r δεν αλλάζει όταν οι μονάδες μέτρησης αλλάζουν. Είναι καθαρός αριθμός.
3. Οι τιμές του συντελεστή συσχέτισης κυμαίνονται από -1 έως 1. Τιμές του r κοντά στο 0 υποδηλώνουν πολύ ασθενή σχέση. Όσο πιο κοντά το r είναι στα -1 και 1 τόσο πιο γραμμικά είναι τα δεδομένα. Οι τιμές -1 και 1 υποδηλώνουν τέλεια αρνητική ή θετικά αντίστοιχα γραμμική σχέση.

Οι διάφορες τιμές του r υποδηλώνουν και το μέγεθος της σχέσης των μεταβλητών:

1. $Αν\ |r| \leq 0.1\ τότε\ υπάρχει\ πολύ\ ασθενής$ ή $καμία\ συσχέτιση$
2. $Αν\ 0.1 < |r| \leq 0.3\ τότε\ υπάρχει\ ασθενής\ συσχέτιση$
3. $Αν\ 0.3 < |r| \leq 0.5\ τότε\ υπάρχει\ μέτρια\ συσχέτιση$
4. $Αν\ 0.5 < |r| \leq 1.0\ τότε\ υπάρχει\ ισχυρή\ συσχέτιση$
5. $Αν\ |r| = 1.0\ τότε\ υπάρχει\ τέλεια\ συσχέτιση$



Πηγή: https://statsmethods.wordpress.com/2013/05/10/pearson-correlation-coefficient-r/

ΚΕΦΑΛΑΙΟ 3

ΚΡΙΤΗΡΙΑ ΠΛΗΡΟΦΟΡΙΑΣ

3.1 Kullback-Leibler πληροφορία

Στο κεφάλαιο αυτό περιγράφεται η πληροφορία Kullback-Leibler ως ένα κριτήριο για την αξιολόγηση στατιστικών μοντέλων που προσεγγίζουν την πραγματική κατανομή των δεδομένων και των ιδιοτήτων τους. Με βάση το κριτήριο αυτό περιγράφεται η απόδειξη του κριτηρίου πληροφορίας του Akaike, και τα βασικά σημεία επιλογής μοντέλων. Επίσης περιγράφεται σε αντιπαράθεση το Μπεϋζιανό κριτήριο πληροφορίας, το οποίο αποτελεί αυστηρότερο κριτήριο επιλογής μοντέλων.

Σύμφωνα με Konishi and Kitagawa (2008), η πληροφορία Kullback-Leibler δημοσιεύτηκε το 1951 από τους Solomon Kullback και Richard Leibler, έχοντας κάνει την έρευνα κατά την διάρκεια του 2^{ου} παγκοσμίου πολέμου. Είναι μια συνάρτηση που συμβολίζεται με "I" και περιλαμβάνει δύο ορίσματα την πραγματική κατανομή και το μοντέλο που προσπαθεί να την προσεγγίσει. Η πληροφορία Kullback-Leibler ή K-L για συντομία, είναι η "πληροφορία" που χάνεται όταν το μοντέλο προσπαθεί να προσεγγίσει την πραγματική κατανομή ή αλλιώς είναι η «απόσταση» του μοντέλου από το πραγματικό. Η «απόσταση» αυτή όμως δεν έχει ακριβώς την γνωστή έννοια, αφού δεν ισχύει η συμμετρία, δηλαδή, συνήθως η απόσταση του μοντέλου από το πραγματικό είναι διαφορετική από την απόσταση του πραγματικού μοντέλου από το μοντέλο που την προσεγγίζει. Αυτό έχει σαν αποτέλεσμα η έννοια της απόστασης να εμπεριέχει την έννοια της προσανατολισμένης ή με κατεύθυνση απόστασης.

3.1.1 Ορισμός και Ιδιότητες

Έστω $x_n = \{x_1, x_2, ..., x_n\}$ είναι ένα σύνολο από n παρατηρήσεις που επιλέγονται τυχαία (ανεξάρτητα) από μια άγνωστη συνάρτηση κατανομής πιθανότητας G(x). Η συνάρτηση κατανομής πιθανότητας G(x) που παράγει τα δεδομένα αποτελεί το πραγματικό μοντέλο ή την πραγματική κατανομή. Σε αντίθεση, έστω F(x) αυθαίρετα ορισμένο μοντέλο. Αν οι συναρτήσεις κατανομής πιθανότητας G(x) και F(x) έχουν συναρτήσεις πυκνότητας g(x) και f(x), αντίστοιχα, τότε αποτελούν συνεχή μοντέλα (ή μοντέλα συνεχούς κατανομής). Εάν, δοθέντος είτε ενός πεπερασμένου συνόλου είτε ενός αριθμήσιμου άπειρου συνόλου διακριτών σημείων $\{x_1, x_2, ..., x_k, ...\}$, εκφράζονται ως πιθανότητες γεγονότων

$$g_i = g(x_i) \equiv Pr(\{\omega; X(\omega) = x_i\}),$$
$$f_i = f(x_i) \equiv Pr(\{\omega; X(\omega) = x_i\}), i = 1,2, ..., \quad (1)$$

τότε αυτά τα μοντέλα ονομάζονται διακριτά μοντέλα (μοντέλα διακριτής κατανομής).

Θεωρείται ότι η καλή προσαρμογή του μοντέλου f(x) εκτιμάται από την άποψη της εγγύτητας ως κατανομής πιθανότητας σε σχέση με την πραγματική κατανομή g(x). Η K-L πληροφορία ή αλλιώς η Kullback-Leibler απόκλιση εκφράζεται με τον παρακάτω τύπο:

$$I(G;F) = E_G[log\left\{\frac{G(X)}{F(X)}\right\}] \quad (2)$$

όπου το E_G αναπαριστά την αναμενόμενη τιμή ως προς την συνάρτηση κατανομής G.
Αν οι συναρτήσεις κατανομής πιθανότητας είναι συνεχείς, με συναρτήσεις πυκνότητας τις g (x) και f (x), τότε η πληροφορία K-L μπορεί να εκφραστεί ως:

$$I(g;f) = \int_{-\infty}^{\infty} log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx. \quad (3)$$

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

$$I(g;f) = \sum_{i=1}^{\infty} g(x_i)\log\left\{\frac{g(x_i)}{f(x_i)}\right\} \quad (4)$$

Ενώνοντας τα συνεχή και διακριτά μοντέλα, μπορούμε να εκφράσουμε την K-L πληροφορία ως:

$$I(g;f) = \int log\left\{\frac{g(x)}{f(x)}\right\} dG(x) = \begin{cases} \int_{-\infty}^{\infty} log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx, & \textit{για συνεχές μοντέλο} \\ \sum_{i=1}^{\infty} g(x_i) log\left\{\frac{g(x_i)}{f(x_i)}\right\}, & \textit{για διακριτό μοντέλο} \end{cases} \quad (5)$$

Ιδιότητες της K-L πληροφορίας

Η K-L πληροφορία έχει τις ακόλουθες ιδιότητες

(i) I(g;f)≥0

(ii) I(g;f)=0 ↔g(x)=f(x)

Εν όψει των ιδιοτήτων αυτών, θεωρούμε ότι όσο μικρότερη είναι η ποσότητα της K-L πληροφορίας, τόσο πιο κοντά το μοντέλο f(x) είναι στο g(x).

Απόδειξη.

Έστω η συνάρτηση $K(t) = logt - t + 1$, με t > 0. Σε αυτή την περίπτωση, η παράγωγος του K(t), $K'(t) = t^{-1} - 1$ ικανοποιεί την συνθήκη K΄(1) = 0 και το K(t) έχει μέγιστο K(1)=0, για t=1. Για αυτό η ανισότητα K(t) ≤ 0 ισχύει για όλα τα t τέτοια ώστε t > 0. Η ισότητα ισχύει μόνο για t=1 που σημαίνει ότι και η παρακάτω σχέση ισχύει:

$$logt \leq t - 1 \quad \text{(η ισότητα ισχύει μόνο όταν t=1)}$$

Για το συνεχές μοντέλο, αντικαθιστώντας t=f(x)/g(x) σε αυτή την παράσταση έχουμε:

$$log\frac{f(x)}{g(x)} \leq \frac{f(x)}{g(x)} - 1$$

Πολλαπλασιάζοντας και τα δύο μέλη της εξίσωσης με g(x) και ολοκληρώνοντας θα ισχύει:

$$\int log\left\{\frac{f(x)}{g(x)}\right\} g(x)dx \leq \int \left\{\frac{f(x)}{g(x)} - 1\right\} g(x)dx$$

$$= \int f(x)dx - \int g(x)dx = 0$$

Αυτό δίνει:

$$\int log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx = -\int log\left\{\frac{f(x)}{g(x)}\right\} g(x)dx \geq 0$$

Άρα ισχύει το (i). Η ισότητα ισχύει μόνο όταν g(x)=f(x).

Για το διακριτό μοντέλο, αρκεί μόνο να αντικατασταθούν οι συναρτήσεις πυκνότητας πιθανότητας g(x) και f(x) με τις συναρτήσεις μάζας πιθανότητας $g(x_i)$ και $f(x_i)$, αντίστοιχα, και να προστεθούν οι όροι για i=1,2… αντί να γίνει ολοκλήρωση.

Παράδειγμα της πληροφορίας K-L για κανονικό μοντέλο

Έστω ότι το πραγματικό μοντέλο g (x) και το καθορισμένο μοντέλο f (x) ακολουθούν αντίστοιχα, τις κανονικές κατανομές $N(\xi, \tau^2)$ και $N(\mu, \sigma^2)$. Αν E_G είναι η αναμενόμενη

τιμή (μέσος όρος) αναφορικά με το πραγματικό μοντέλο και η τυχαία μεταβλητή X ακολουθεί την N(ξ, τ^2) τότε θα ισχύει:

$$E_G[(X-\mu)^2] = E_G[(X-\xi)^2 + 2(X-\xi)(\xi-\mu) + (\xi-\mu)^2] = \tau^2 + (\xi-\mu)^2$$

Οπότε για την κανονική κατανομή $f(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$ θα ισχύει:

$$E_G[\log f(X)] = E_G[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(X-\mu)^2}{2\sigma^2}]$$

$$= -\frac{1}{2}log(2\pi\sigma^2) - \frac{\tau^2 + (\xi-\mu)^2}{2\sigma^2}$$

Εάν τεθεί μ = ξ και $\sigma^2 = \tau^2$ στην εξίσωση αυτή θα ισχύει:

$$E_G[log\ g(X)] = -\frac{1}{2}log(2\pi\tau^2) - \frac{1}{2}$$

Οπότε η K-L πληροφορία του μοντέλου f(x) ως προς το g(x) δίνεται από:

$$I(g;f) = E_G[log\ g(X)] - E_G[log\ f(X)]$$

$$= \frac{1}{2}\left\{log\frac{\sigma^2}{\tau^2} + \frac{\tau^2 + (\xi-\mu)^2}{2\sigma^2} - 1\right\}$$

3.1.2 Αναμενόμενη Λογαριθμική Πιθανοφάνεια και αντίστοιχη Εκτιμήτρια

Ένας τρόπος για την αξιολόγηση της προσαρμοστικότητας (goodness of fit) ενός μοντέλου γίνεται με την πληροφορία K-L. Παρ'όλα αυτά, η K-L πληροφορία μπορεί να χρησιμοποιηθεί στην μοντελοποίηση μόνο για συγκεκριμένες περιπτώσεις, αφού περιέχει την άγνωστη κατανομή g, με αποτέλεσμα η τιμή της να μην μπορεί να υπολογιστεί άμεσα. Η K-L μπορεί να αναλυθεί σε:

$$I\ (g;\ f) =\ E_G\left[log\left\{\frac{g(X)}{f(X)}\right\}\right] = E_G[log\ g(X)] - E_G[log\ f(X)] \quad (6)$$

Επιπλέον, επειδή ο πρώτος όρος στο δεξιό μέλος είναι σταθερά, η οποία βασίζεται μόνο στο πραγματικό μοντέλο g, την σταθερή κατανομή που ακολουθούν τα δεδομένα, επομένως για να συγκριθούν διαφορετικά μοντέλα, αρκεί να ληφθεί υπόψη μόνο ο δεύτερος όρος στο δεξιό μέλος. Αυτός ο όρος ονομάζεται αναμενόμενη λογαριθμική πιθανοφάνεια. Όσο μεγαλύτερη τιμή έχει για το μοντέλο, τόσο μικρότερη η τιμή της K-L πληροφορίας και άρα τόσο μικρότερη η "απόσταση" των μοντέλων, επομένως τόσο καλύτερο θα είναι το μοντέλο που προσεγγίζει το πραγματικό.

Αφού η αναμενόμενη λογαριθμική πιθανοφάνεια μπορεί να εκφραστεί ως:

$$E_G[\log f(X)] = \int \log f(x) dG(x) = \tag{7}$$

$$= \begin{cases} \int_{-\infty}^{\infty} g(x) \log f(x) dx, & \text{για συνεχές μοντέλο} \\ \sum_{i=1}^{\infty} g(x_i) \log f(x_i), & \text{για διακριτό μοντέλο} \end{cases}$$

ακόμα βασίζεται στην πραγματική κατανομή g και είναι μια άγνωστη ποσότητα που διαφεύγει ακριβή υπολογισμό. Παρ'όλα αυτά, αν μια καλή εκτίμηση για την αναμενόμενη λογαριθμική πιθανοφάνεια μπορεί να αποκτηθεί από τα δεδομένα, αυτή η εκτίμηση μπορεί να χρησιμοποιηθεί ως ένα κριτήριο για την σύγκριση μοντέλων.

3.2 Κριτήριο πληροφορίας του Akaike

3.2.1 Ανάλυση του κριτηρίου πληροφορίας του Akaike AIC

Κατά τη διάρκεια των τελευταίων δεκαπέντε ετών, σύμφωνα με Hamparsum Bozdogan (1987) και Kenneth P. Burnham and David R. Anderson (2004) το εντροπικό κριτήριο πληροφορίας του Akaike (1973) το οποίο είναι γνωστό ως AIC, είχε σημαντικότατες επιρροές στα στατιστικά προβλήματα αξιολόγησης μοντέλων. Η εισαγωγή του AIC προώθησε την αναγνώριση της σημασίας της «καλής» μοντελοποίησης στην στατιστική. Ως αποτέλεσμα, πολλές σημαντικές τεχνικές στατιστικής μοντελοποίησης έχουν αναπτυχθεί σε διάφορους τομείς της στατιστικής, θεωρίας ελέγχου, της οικονομετρίας, της μηχανικής, ψυχομετρίας, και σε πολλούς άλλους τομείς. Παρακάτω αναλύεται η γενική θεωρία του κριτηρίου και περιγράφεται η απόδειξή του.

Στην στατιστική ανάλυση, όπως αναφέρθηκε, είναι ιδιαίτερα δύσκολη η επιλογή ενός κατάλληλου μοντέλου, καθώς και ο προσδιορισμός και η εκτίμηση της διάστασης του μοντέλου. Ένα πρόβλημα, ιδιαίτερα συχνό, είναι όταν υπάρχουν πολλές παράμετροι στο στατιστικό μοντέλο. Σκοπός της στατιστικής μοντελοποίησης είναι η εύρεση ενός μοντέλου το οποίο θα προσαρμόζεται στα δεδομένα, χωρίς να είναι γνωστό ποιο είναι το πραγματικό μοντέλο.

Αυτό έχει ως αποτέλεσμα, ο σκοπός των ερευνητών να είναι η εύρεση τέτοιων μοντέλων, καθώς και η επιλογή του «καλύτερου» από αυτά που μπορεί να προσαρμοστεί στα δεδομένα, επιλεγμένο μεταξύ μοντέλων με διαφορετικό αριθμό παραμέτρων έχοντας γνωστό ένα σύνολο δεδομένων. Επίσης βασικός στόχος είναι η επιλογή με την χρήση απλών κριτηρίων που διαλέγουν από ένα σύνολο μοντέλων εκείνο με το μικρότερο αριθμό παραμέτρων που μπορεί να περιγράψει τα δεδομένα καλύτερα. Αυτό είναι γνωστό ως η αρχή της φειδωλότητας. Τα μοντέλα επιλέγονται με τέτοιο τρόπο ώστε για συγκεκριμένη ακρίβεια, ένα πιο απλό ή φειδωλό μοντέλο να είναι προτιμότερο από ένα πολύπλοκο. Στόχος είναι η επιλογή του πιο ακριβούς και φειδωλού μοντέλου, εκείνου που με την λιγότερη πολυπλοκότητα ή ισοδύναμα εκείνο με το μεγαλύτερο κέρδος πληροφορίας να περιγράφει καλύτερα την πραγματικότητα.

Ο Akaike έθεσε τα θεμέλια της σύγχρονης στατιστικής μοντελοποίησης. Ανέπτυξε το κριτήριο πληροφορίας AIC για την εύρεση ενός βέλτιστου και φειδωλού μοντέλου από μια κλάση μοντέλων, έχοντας υπόψη του την πολυπλοκότητα των μοντέλων. Το κριτήριο αυτό είναι ευπροσάρμοστο και βασίζεται στην προαναφερθέντα Kullback Leibler πληροφορία.

3.2.2 Περιγραφή απόδειξης

Έστω $\vec{X}$ ένα συνεχές τυχαίο διάνυσμα, με συνάρτηση πυκνότητας πιθανότητας $f(\vec{x}|\vec{\theta})$ όπου $\vec{\theta}$ το K-διάστασης διάνυσμα παραμέτρων $\vec{\theta}=\vec{\theta}_K=(\theta_1,\theta_2,...\theta_K)$, $\vec{\theta}_K \in R^K$. Έστω ότι υπάρχει πραγματικό διάνυσμα παραμέτρων $\vec{\theta}^*$ του $\vec{\theta}$ και με την σ.π.π. να συμβολίζεται ως $f(\vec{x}|\vec{\theta}^*)$. Το $\vec{\theta}$ είναι διαλεγμένο με τέτοιο τρόπο ώστε να πλησιάζει όσο το δυνατόν περισσότερο το $\vec{\theta}^*$.
Από την (6) η K-L είναι:

$$I\ (g;\ f) = E_G[log\ g(\vec{X})] - E_G[log\ f(\vec{X})]$$

Άρα για την παραπάνω περίπτωση γίνεται:

$$\begin{aligned} I(\vec{\theta}^*;\vec{\theta}) &= E\left[logf\left(\vec{X}|\vec{\theta}^*\right)\right] - E\left[logf\left(\vec{X}|\vec{\theta}\right)\right] \\ &= \int f\left(\vec{x}|\vec{\theta}^*\right) logf\left(\vec{x}|\vec{\theta}^*\right) d\vec{x} - \int f\left(\vec{x}|\vec{\theta}^*\right) logf\left(\vec{x}|\vec{\theta}\right) d\vec{x} \\ &= H\left(\vec{\theta}^*;\vec{\theta}^*\right) - H(\vec{\theta}^*;\vec{\theta}), \end{aligned} \tag{8}$$

όπου $H(\vec{\theta}^*;\vec{\theta}) = \int f(\vec{x}|\vec{\theta}^*) logf(\vec{x}|\vec{\theta}) d\vec{x}$ είναι η αναμενόμενη λογαριθμική πιθανοφάνεια που μετράει την καλή προσαρμογή (goodness of fit) του $f(\vec{x}/\vec{\theta})$ στο $f(\vec{x}/\vec{\theta}^*)$ και το $H(\vec{\theta}^*;\vec{\theta}^*) \equiv H(\vec{\theta}^*)$ είναι η αρνητική εντροπία κατά Shannon που είναι σταθερή για δοθέν $f(\vec{x}/\vec{\theta}^*)$ και log ο νεπέριος λογάριθμος.
Αφού $H(\vec{\theta}^*;\vec{\theta}^*) \equiv H(\vec{\theta}^*)$ είναι σταθερά, πρέπει να υπολογιστεί μόνο η αναμενόμενη λογαριθμική πιθανοφάνεια

$$H(\vec{\theta}^*;\vec{\theta}) = E[logf(\vec{X}|\vec{\theta})] = \int f(\vec{x}|\vec{\theta}^*) logf(\vec{x}|\vec{\theta}) d\vec{x} \tag{9}$$

3.2.3 Μέση λογαριθμική πιθανοφάνεια ως ένας εκτιμητής για την K-L πληροφορία.

Όπως αναφέρθηκε η πληροφορία K-L δεν μπορεί να υπολογιστεί ρητά, αφού εξαρτάται από την πραγματική κατανομή για αυτό πρέπει να εκτιμηθεί από τα δεδομένα.
Έστω ότι η συνάρτηση πιθανοφάνειας για n ανεξάρτητες παρατηρήσεις από ένα μοντέλο με σ.π.π. $f(\vec{x}/\vec{\theta})$ και $\vec{\theta}$ $=(\theta_1,\theta_2,...\theta_k)$, k=1,2,…K είναι:

$$L(\vec{\theta}) = f(x_1, x_2, \ldots, x_n \,|\vec{\theta}) = \prod_{i=1}^{n} f(x_i|\vec{\theta}) \tag{10}$$

Η συνάρτηση λογαριθμικής πιθανοφάνειας $l(\vec{\theta})$ είναι ο λογάριθμος του $L(\vec{\theta})$ και ορίζεται ως:

$$l(\vec{\theta}) \equiv logL(\vec{\theta}) = \sum_{i=1}^{n} logf(x_i|\vec{\theta}) \tag{11}$$

Η μέση λογαριθμική πιθανοφάνεια του δείγματος υπολογίζεται σύμφωνα με την σχέση:

$$\frac{1}{n} l(\vec{\theta}) \equiv \frac{1}{n} logL(\vec{\theta}) = \frac{1}{n} \sum_{i=1}^{n} logf(x_i|\vec{\theta}) = l_n(\vec{\theta}) \tag{12}$$

η οποία αποτελεί εκτιμήτρια της "απόστασης" μεταξύ της πραγματικής σ.π.π. $f(\vec{x}|\vec{\theta}^*)$ και του μοντέλου $f(\vec{x}/\vec{\theta})$. Η διαδικασία υπολογισμού εκτίμησης της K-L πληροφορίας από τα δεδομένα παρουσιάζεται παρακάτω.

Έστω ότι το $\tilde{I}(\vec{\theta}^*;\vec{\theta})$ είναι εκτιμητής της K-L πληροφορίας $I(\vec{\theta}^*;\vec{\theta})$. Τότε η (8) γίνεται:

$$\tilde{I}\,(\vec{\theta}^*;\vec{\theta}) = \tilde{H}(\vec{\theta}^*;\vec{\theta}^*) - \tilde{H}(\vec{\theta}^*;\vec{\theta}), \tag{13}$$

$$\tilde{H}(\vec{\theta}^*;\vec{\theta}) = -\tilde{I}(\vec{\theta}^*;\vec{\theta}) + \tilde{H}(\vec{\theta}^*;\,\vec{\theta}^*) \tag{14}$$

Η μεγιστοποιήση της αναμενόμενης λογαριθμικής πιθανοφάνειας $\tilde{H}(\vec{\theta}^*;\vec{\theta})$ είναι ασυμπτωτικά ίση με την ελαχιστοποίηση της K-L πληροφορίας $\tilde{I}(\vec{\theta}^*;\vec{\theta})$ ενώ η $\tilde{H}\,(\vec{\theta}^*;\vec{\theta}^*) \equiv \tilde{H}\,(\vec{\theta}^*)$ είναι σταθερά.

Χρησιμοποιώντας ένα δείγμα με n παρατηρήσεις $\vec{x}$=(x_1,x_2,...x_n) ώστε να βρεθεί μια εκτιμήτρια $\hat{\vec{\theta}} = \hat{\vec{\theta}}(\vec{x})$ του $\vec{\theta}$, παρατηρούμε ότι η μέση λογαριθμική πιθανοφάνεια $l_n(\vec{\theta})$ αποτελεί εκτιμήτρια του $\tilde{H}(\vec{\theta}^*;\vec{\theta})$ της αναμενόμενης λογαριθμικής πιθανοφάνειας. Οπότε θα ισχύει:

$$E\left[\frac{1}{n} l(\vec{\theta})\right] = \tilde{H}(\vec{\theta}^*;\vec{\theta}) = E[logf(\vec{X}|\vec{\theta})] \tag{15}$$

Η αναμενόμενη τιμή είναι ως προς την πραγματική κατανομή $f(\vec{x}|\vec{\theta}^*)$ και ο εκτιμητής μέγιστης πιθανοφάνειας $\hat{\vec{\theta}}$ είναι εκτιμητής του $\vec{\theta}^*$ οπότε η (13) γίνεται:

$$\tilde{I}\left(\vec{\theta}^{*};\hat{\vec{\theta}}\right)=\tilde{H}(\vec{\theta}^{*})-\frac{1}{n}\sum_{i=1}^{n}logf\left(x_i\middle|\hat{\vec{\theta}}\right)=\tilde{H}\left(\vec{\theta}^{*}\right)-\frac{1}{n}l\left(\hat{\vec{\theta}}\right)$$
$$=\tilde{H}\left(\vec{\theta}^{*}\right)-l_n\left(\hat{\vec{\theta}}\right) \quad (16)$$

Άρα
$$l_n\left(\hat{\vec{\theta}}\right)=\tilde{H}\left(\vec{\theta}^{*}\right)+\tilde{I}\left(\vec{\theta}^{*};\hat{\vec{\theta}}\right) \quad (17)$$

Από τις ιδιότητες της K-L πληροφορίας για να ισχύει $I(\vec{\theta}^{*};\vec{\theta})=0$ πρέπει να ισχύει f($\vec{x}|\vec{\theta}^{*}$)= f($\vec{x}|\vec{\theta}$). Άρα ασυμπτωτικά το μέγιστο του $l_n(\vec{\theta})=\frac{1}{n}logL(\vec{\theta})$ (μέση λογαριθμική πιθανοφάνεια) θα είναι $H(\vec{\theta}^{*})\equiv\int f(\vec{x}|\vec{\theta}^{*})logf(\vec{x}|\vec{\theta}^{*})d\vec{x}$, η αρνητική εντροπία Shannon, το οποίο ισχύει όταν f($\vec{x}|\vec{\theta}^{*}$)= f($\vec{x}|\vec{\theta}$). Η μεγιστοποίηση της $\tilde{H}(\vec{\theta}^{*};\vec{\theta})$ επομένως γίνεται για $\vec{\theta}=\vec{\theta}^{*}=\theta_0$ το οποίο αποτελεί το σημείο ελάχιστης πληροφορίας, αφού ελαχιστοποιεί την K-L πληροφορία.

Επειδή η εκτιμήτρια της K-L πληροφορίας βασίζεται στην μέση λογαριθμική πιθανοφάνεια, η οποία είναι και εκτιμήτρια της αναμενόμενης λογαριθμικής πιθανοφάνειας, και αφού οι εκτιμήτριες μεγίστης πιθανοφάνειας είναι μεροληπτικές, είναι αναπόφευκτη η εισαγωγή ενός σφάλματος εκτίμησης της K-L πληροφορίας, όταν χρησιμοποιούνται εκτιμητές μεγίστης πιθανοφάνειας.

Το $l_n(\vec{\theta})$ έχει την μέγιστη τιμή του σε ένα σημείο κοντά στο $\vec{\theta}^{*}$ και αυτό είναι το $\hat{\vec{\theta}}$. Επειδή η ποσότητα $H(\vec{\theta}^{*};\vec{\theta})$ δεν είναι άμεσα υπολογίσημη χρησιμοποιείται η μεγιστοποίηση της μέσης λογαριθμικής πιθανοφάνειας και ασυμπτωτικά προσπαθούμε να βρεθεί ένας αμερόληπτος εκτιμητής της μέσης λογαριθμικής πιθανοφάνειας, διορθώνοντας την μεροληψία της μέσης λογαριθμικής πιθανοφάνειας $l_n\left(\hat{\vec{\theta}}\right)$ με την χρήση κάποιου όρου ποινικοποίησης. Αυτή την διαδικασία είχε υπόψη του ο Akaike, όταν κατασκεύαζε το κριτήριο, δηλαδή την ποινοκοποίηση των επιπλέον παραμέτρων όταν χρησιμοποιούνται οι εκτιμήτριες μεγίστης πιθανοφάνειας για την εκτίμηση της αναμενόμενης λογαριθμικής πιθανοφάνειας από την μέση λογαριθμική πιθανοφάνεια.

3.2.4 Αποτέλεσμα της συλλογιστικής πορείας του Akaike

Ο Akaike (1973) αναγνωρίζοντας ότι οι παράμετροι των μοντέλων όντας άγνωστες πρέπει να εκτιμηθούν από τα δεδομένα, το αποτέλεσμα ήταν ότι βρήκε μια σχέση μεταξύ της K-L πληροφορίας και της μέγιστης πιθανοφάνειας. Έστω ότι η τιμή του θ που ελαχιστοποιεί την K-L πληροφορία I(f,g) είναι η θ_0. Έστω, επίσης, ότι έχει σαν εκτιμήτρια μεγίστης πιθανοφάνειας την $\hat{\theta}$ που εκτιμά το θ_0 , στην περίπτωση που το g μοντέλο είναι το καλύτερο μοντέλο σύμφωνα με την K-L. Επειδή το θ_0 είναι άγνωστο, το γεγονός ότι χρησιμοποιείται η εκτίμηση $\hat{\theta}$, ουσιαστικά επιτρέπει, χωρίς να απαιτείται η ελαχιστοποίηση της K-L απόστασης, να ελαχιστοποιείται η αναμενόμενη εκτίμηση της K-L. Ο Akaike απέδειξε ότι για την δημιουργία ενός εφαρμοσμένου κριτηρίου επιλογής μοντέλων χρειάζεται η εκτίμηση της σχέσης:

$$E_y E_x[log(g(\vec{x}|\hat{\vec{\theta}}(\vec{y})))] \quad (18)$$

όπου x και y είναι ανεξάρτητα τυχαία δείγματα από την ίδια κατανομή και οι αναμενόμενες τιμές είναι ως προς την πραγματική κατανομή f. Ενώ το E_y αναφέρεται στα δεδομένα (y) που επιτρέπουν την κατασκευή εκτιμητριών των άγνωστων παραμέτρων του μοντέλου.

Ένας τρόπος εκτίμησης του $E_y E_x[\log(g(\vec{x}|\hat{\vec{\theta}}(\vec{y})))]$ θα ήταν με την μεγιστοποίηση του $\log L(\hat{\vec{\theta}}|data)$ για κάθε προσεγγιστικό μοντέλο g. Ο Akaike όμως έδειξε ότι η μέγιστη λογαριθμική πιθανοφάνεια είναι μεροληπτική ως εκτιμήτρια της παραπάνω σχέσης. Έδειξε ότι σε ορισμένες συνθήκες, η μεροληψία είναι προσεγγιστικά ίση με K, ο αριθμός των παραμέτρων προς εκτίμηση στο προσεγγιστικό μοντέλο, και αποτελεί τον ασυμπτωτικό όρο διόρθωσης μεροληψίας. Οπότε ένας ασυμπτωτικά αμερόληπτος εκτιμητής του $E_y E_x[\log(g(\vec{x}|\hat{\vec{\theta}}(\vec{y})))]$ για μεγάλα δείγματα είναι ο $\log L(\hat{\vec{\theta}}|data) - K$.

Το οποίο είναι ίσο με:

$$logL\left(\hat{\vec{\theta}}|data\right) - K = σταθερά - \hat{E}_{\hat{\theta}}[I(f, \hat{g})], \text{ όπου } \hat{g} = g(.|\hat{\theta}) \quad (29)$$

Ο τύπος αυτός επιτρέπει τον συνδυασμό της εκτίμησης με την χρήση της θεωρίας της μέγιστης πιθανοφάνειας και την επιλογή μοντέλων.
Το τελευταίο βήμα που χρησιμοποίησε ο Akaike ώστε να ορίσει το κριτήριο πληροφορίας AIC (Akaike's information criterion) ήταν να πολλαπλασιάσει το $logL\left(\hat{\vec{\theta}}|data\right) - K$ με -2 (η συγκεκριμένη τιμή έχει επιλεγεί για ιστορικούς λόγους) καταλήγοντας στον παρακάτω τύπο:

$$AIC = -2log(L(\hat{\vec{\theta}}|\vec{y})) + 2K \quad (20)$$

Με αυτόν τον τρόπο αντί να χρησιμοποιείται η προσανατολισμένη απόσταση (K-L) μεταξύ δύο μοντέλων, χρησιμοποείται μια εκτίμηση της αναμενόμενης απόστασης μεταξύ του προσαρμοσμένου μοντέλου και του πραγματικού μοντέλου που παρήγαγε τα δεδομένα.

Η έκφραση $log(L(\hat{\vec{\theta}}|\vec{y}))$ είναι η τιμή της λογαριθμικής πιθανοφάνειας στο μέγιστο σημείο, που αντιστοιχεί στις τιμές των εκτιμητριών μέγιστης πιθανοφάνειας. Το AIC δεν απαιτεί την αναλυτική κατασκευή όρων διόρθωσης της μεροληψίας σε κάθε πρόβλημα και δεν βασίζεται στην άγνωστη κατανομή πιθανότητας, αφαιρώντας με αυτόν τον τρόπο διακυμάνσεις λόγω της εκτίμησης της μεροληψίας. Επίσης, ο Akaike ορίζει ότι αν η πραγματική κατανομή που παρήγαγε τα δεδομένα είναι σε "κοντινή απόσταση" από το παραμετρικό μοντέλο, τότε η μεροληψία που συνδέεται με την λογαριθμική πιθανοφάνεια του μοντέλου που βασίζεται στην μέθοδο μέγιστης πιθανοφάνειας, μπορεί να προσεγγιστεί από τον αριθμό των παραμέτρων.
Σε εφαρμογή, ουσιαστικά υπολογίζεται η τιμή του AIC για τα πιθανά μοντέλα και επιλέγεται αυτό με την μικρότερη τιμή. Αυτό το μοντέλο εκτιμάται ότι είναι πιο

κοντά στο πραγματικό. Παρ'όλα αυτά στην περίπτωση που τα μοντέλα προς επιλογή δεν είναι «κοντά» στο πραγματικό, τότε οποιαδήποτε επιλογή από το AIC, μπορεί να είναι η καλύτερη, αλλά στην πραγματικότητα δεν θα προσεγγίζει την πραγματική κατανομή.

Όπως αναφέρθηκε και στην (20), ο Akaike πολλαπλασίασε με -2, για ιστορικούς λόγους. Είναι γνωστό ότι -2 φορές τον λογάριθμο του λόγου δύο μέγιστων τιμών πιθανοφάνειας ακολουθεί ασυμπτωτικά την X^2-κατανομή. Επειδή το -2 έχει χρησιμοποιηθεί και σε άλλες στατιστικές συμπερασματολογίες, δεν είναι παράλογο που και ο Akaike το χρησιμοποιεί. Επισημαίνεται επίσης ότι το μοντέλο με τη μικρότερη τιμή AIC παραμένει σταθερό ακόμα και αν το log(L)-K είχε πολλαπλασιαστεί με οποιοδήποτε αρνητικό αριθμό. Επίσης το -2Log(L) αποτελεί στην στατιστική την απόκλιση, την ποσοτικοποίηση δηλαδή της έλλειψης προσαρμοστικότητας. Επομένως η (20) μπορεί να θεωρηθεί και ως η απόκλιση με όρο ποινικοποίησης 2K ώστε να υπάρξει διόρθωση για την ασυμπτωτική μεροληψία.

Ο όρος της απόκλισης, της έλλειψης προσαρμοστικότητας μπορεί να μικρύνει προσθέτοντας περισσότερες γνωστές (όχι εκτιμώμενες) παραμέτρους στο προσεγγιστικό μοντέλο g, επιτρέποντας σε αυτό να έρθει πιο "κοντά" με το f. Παρ'όλα αυτά, όταν οι παράμετροι αυτοί θα πρέπει να εκτιμηθούν, η προσθήκη τους αντιθέτως αυξάνει επιπλέον την αβεβαιότητα της εκτίμησης της αναμενόμενης K-L πληροφορίας. Η επιπλέον προσθήκη παραμέτρων που θα πρέπει να εκτιμηθούν, θα αυξήσει την τιμή της K-L απόστασης γιατί ο "θόρυβος" τότε μοντελοποιείται ως δομικό στοιχείο της συμπερασματολογίας. Αυτό επιβεβαιώνεται και από την εξίσωση AIC όπου ο πρώτος όρος στο δεξί μέλος τείνει να ελαττωθεί όσο περισσότερες παράμετροι προστίθενται στο g, ενώ ο δεύτερος όρος ως όρος "ποινικοποίησης" αυξάνεται ανάλογα. Χωρίς αυτόν τον όρο το καλύτερο μοντέλο θα ήταν σχεδόν πάντα το μεγαλύτερο μοντέλο στο σύνολο, αφού κάθε προσθήκη παραμέτρων θα ήταν χωρίς κόστος (ποινικοποίηση). Τα μοντέλα, επομένως, θα υφίστανται υπερπροσαρμογή, θα είχαν χαμηλή ακρίβεια, και θα υπήρχε το ρίσκο κίβδηλων αποτελεσμάτων λόγω της εισαγωγής του «θορύβου». Παρ'όλα αυτά, το πλεονέκτημα της προσθήκης παραμέτρων και το επακόλουθο μειονέκτημα στην προσθήκη ακόμα παραπάνω αποτελεί ένα είδος ανταλλαγής. Αυτή είναι η ανταλλαγή μεταξύ της μεροληψίας και της διασποράς ή αλλιώς η ανταλλαγή μεταξύ της υποπροσαρμογής και της υπερπροσαρμογής μοντέλων που προδιαθέτει την σημασία της αρχής της φειδωλότητας που αναφέρθηκε παραπάνω.

Παράδειγμα επιλογής μεταβλητών για το μοντέλο παλινδρόμησης

Έστω η μεταβλητή απόκρισης y και m επεξηγηματικές μεταβλητές x_1, .., x_m. Το μοντέλο γραμμικής παλινδρόμησης είναι:

$$y = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_m x_m + \varepsilon$$

όπου το ε τυχαία μεταβλητή και θεωρείται ότι ακολουθεί την κανονική κατανομή $N(0,\sigma^2)$. Η δεσμευμένη κατανομή της μεταβλητής απόκρισης δεδομένου των επεξηγηματικών μεταβλητών είναι

$$p(y, x_1, \dots, x_m) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y - \alpha_0 - \sum_{j=1}^{m} \alpha_j x_j)^2\right\}$$

Δεδομένου ενός συνόλου από n ανεξάρτητες παρατηρήσεις $\{(y_i, x_{i1}, \dots, x_{im}); i = 1, \dots n\}$, η πιθανοφάνεια του μοντέλου παλινδρόμησης είναι

$$L(\alpha_0, \alpha_1, \dots, \alpha_m, \sigma^2) = \prod_{i=1}^{n} p(y_i | x_{i1}, \dots, x_{im})$$

Οπότε η λογαριθμική πιθανοφάνεια δίνεται από τον τύπο:

$$l(\alpha_0, \alpha_1, \dots, \alpha_m, \sigma^2) = -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha_0 - \sum_{j=1}^{m} \alpha_j x_{ij})^2$$

Και οι εκτιμήτριες μέγιστης πιθανοφάνειας $\hat{a}_0, \hat{a}_1, \dots, \hat{a}_m$ των συντελεστών παλινδρόμησης $\alpha_0, \alpha_1, \dots, \alpha_m$, υπολογίζονται ως λύση του συστήματος γραμμικών εξισώσεων

$$X^T X \vec{\alpha} = X^T \vec{y}$$

όπου $\vec{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_m)^T$ και ο n × (m+1) πίνακας X και το n-διάστατο διάνυσμα $\vec{y}$ προσδιορίζονται από τους παρακάτω τύπους:

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nm} \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Η εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\sigma}^2$ είναι:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \{y_i - (\hat{a}_0 + \hat{\alpha}_1 x_{i1} + \dots + \hat{\alpha}_m x_{im})\}^2$$

Άρα αντικαθιστώντας τις τιμές αυτές παραπάνω, η μέγιστη λογαριθμική πιθανοφάνεια γίνεται:

$$l(\hat{a}_0, \hat{a}_1, \dots, \hat{a}_m, \hat{\sigma}^2) = -\frac{n}{2} log(2\pi) - \frac{n}{2} logd(x_1, \dots, x_m) - \frac{n}{2}$$

όπου $d(x_1, \dots, x_m)$ είναι ο εκτιμητής του σ^2 που δίνεται από τον τύπο του $\hat{\sigma}^2$.

Αφού ο αριθμός των ελευθέρων παραμέτρων που περιέχονται στο μοντέλο πολλαπλής παλινδρόμησης είναι m+2, το AIC για αυτό το μοντέλο είναι

$$AIC = n(log2\pi + 1) + nlogd(x_1, \dots, x_m) + 2(m + 2)$$

Στην ανάλυση της πολλαπλής παλινδρόμησης, όλες οι επεξηγηματικές μεταβλητές μπορεί να μην είναι απαραίτητες για την πρόβλεψη της μεταβλητής απόκρισης. Ένα εκτιμώμενο μοντέλο με υπερβολικά μεγάλο αριθμό επεξηγηματικών μεταβλητών μπορεί να είναι ασταθές. Επιλέγοντας το μοντέλο με το μικρότερο AIC για διαφορετικούς συνδυασμούς των μεταβλητών, επιτυγχάνεται η εύρεση ενός σχετικά «καλού» μοντέλου.

3.3 Το Μπεϋζιανό κριτήριο πληροφορίας BIC

Το κριτήριο αυτό περιέχει την έννοια της ποινικοποίησης των παραμέτρων, όπως και το AIC, με μεγαλύτερη ακρίβεια. Για την απόδειξη του κριτηρίου απαιτούνται κάποιες βασικές έννοιες που παρουσιάζονται παρακάτω.

3.3.1 Η προσέγγιση Laplace για ολοκληρώματα.

Για την απόδειξη της Λαπλασιανής μεθόδου προσεγγίσεως σύμφωνα με Tierney and Kadane (1986), Davison (1986), and Barndorff-Nielsen and Cox (1989, p. 169), Konishi and Kitagawa (2008) θεωρείται η προσέγγιση ενός απλού ολοκληρώματος το οποίο δίνεται από τον τύπο:

$$\int \exp\{nq(\vec{\theta})\}\, d\vec{\theta},$$

όπου $\vec{\theta}$ είναι ένα p-διάστατο διάνυσμα παραμέτρων. Η απόδειξη βασίζεται στο γεγονός ότι όταν ο αριθμός των παρατηρήσεων n είναι μεγάλος, η προς ολοκλήρωση συνάρτηση προσεγγίζεται στην γειτονιά της κορυφής $\hat{\vec{\theta}}$ του q($\vec{\theta}$) και άρα η τιμή της εξαρτάται μόνο από την συμπεριφορά της συνάρτησης στην γειτονιά αυτή.

Με $\partial q(\vec{\theta})/\partial\vec{\theta}|_{\vec{\theta}=\hat{\vec{\theta}}}$ το ανάπτυγμα Taylor για το q($\vec{\theta}$) γύρω από το $\hat{\vec{\theta}}$ γίνεται :

$$q(\vec{\theta}) = q\left(\hat{\vec{\theta}}\right) - \frac{1}{2}\left(\vec{\theta} - \hat{\vec{\theta}}\right)^T J_q\left(\hat{\vec{\theta}}\right)\left(\vec{\theta} - \hat{\vec{\theta}}\right) + \cdots,$$

όπου :

$$J_q\left(\hat{\vec{\theta}}\right) = -\frac{\partial^2 q(\vec{\theta})}{\partial\vec{\theta}\partial\vec{\theta}^T}|_{\vec{\theta}=\hat{\vec{\theta}}}$$

Αντικαθιστώντας το ανάπτυγμα του Taylor του q($\vec{\theta}$) στο πρώτο ολοκλήρωμα θα ισχύει :

$$\int \exp\left[n\left\{q\left(\hat{\vec{\theta}}\right) - \frac{1}{2}\left(\vec{\theta} - \hat{\vec{\theta}}\right)^T J_q\left(\hat{\vec{\theta}}\right)\left(\vec{\theta} - \hat{\vec{\theta}}\right) + \cdots,\right\}\right] d\vec{\theta}$$

$$\approx exp\{nq\left(\hat{\vec{\theta}}\right)\} \int exp\left\{-\frac{n}{2}\left(\vec{\theta} - \hat{\vec{\theta}}\right)^T J_q\left(\hat{\vec{\theta}}\right)\left(\vec{\theta} - \hat{\vec{\theta}}\right)\right\} d\vec{\theta}$$

Επειδή το p-διάστατο τυχαίο διάνυσμα $\vec{\theta}$ ακολουθεί την κανονική κατανομή με p-μεταβλητές, με μέση τιμή το διάνυσμα $\hat{\vec{\theta}}$ και πίνακα διακύμανσης-συνδιακύμανσης $n^{-1}J_q\left(\hat{\vec{\theta}}\right)^{-1}$, ο υπολογισμός του παραπάνω ολοκληρώματος στο δεξιό μέλος δίνει:

$$\int \exp\left\{-\frac{n}{2}\left(\vec{\theta} - \hat{\vec{\theta}}\right)^T J_q\left(\hat{\vec{\theta}}\right)\left(\vec{\theta} - \hat{\vec{\theta}}\right)\right\} d\vec{\theta} = \frac{(2\pi)^{p/2}}{n^{p/2}|J_q\left(\hat{\vec{\theta}}\right)|^{1/2}}$$

Άρα η προσέγγιση του Laplace για το ολοκλήρωμα έχει την ακόλουθη μορφή.

Έστω q($\vec{\theta}$) μια πραγματική συνάρτηση ενός p-διάστατου διανύσματος παραμέτρων $\vec{\theta}$ και έστω $\hat{\vec{\theta}}$ η επικρατούσα τιμή του q($\vec{\theta}$). Τότε η προσέγγιση Laplace του ολοκληρώματος δίνεται από:

$$\int exp\{nq(\vec{\theta})\}d\vec{\theta} \approx \frac{(2\pi)^{p/2}}{n^{p/2}|J_q\left(\hat{\vec{\theta}}\right)|^{1/2}} exp\{nq\left(\hat{\vec{\theta}}\right)\}$$

3.3.2 Περιγραφή της απόδειξης του BIC

Το Μπεϋζιανό Κριτήριο Πληροφορίας (BIC) ή κριτήριο πληροφορίας του Schwarz (SIC) το οποίο πρότεινε ο Schwarz (1978) είναι ένα κριτήριο αξιολόγησης για τα μοντέλα που ορίζεται σύμφωνα με την εκ των υστέρων πιθανότητα τους (posterior probability).

Έστω M_1, M_2, ..., M_r r υποψήφια μοντέλα, και έστω ότι κάθε μοντέλο M_i χαρακτηρίζεται από μια παραμετρική κατανομή f_i (x | $\vec{\theta}_i$) ($\vec{\theta}_i \in \Theta_i \subset R^{k_i}$) και την εκ των προτέρων κατανομή $\pi_i(\vec{\theta}_i)$ του k_i-διαστάσεων διανύσματος παραμέτρων $\vec{\theta}_i$. Όταν δίνονται n παρατηρήσεις $\vec{x}_n = \{x_1, \ldots, x_n\}$, τότε, για το i-οστό μοντέλο M_i, η οριακή κατανομή ή πιθανότητα του $\vec{x}_n$ δίνεται από τον τύπο

$$p_i(\vec{x}_n) = \int f_i(\vec{x}_n | \vec{\theta}_i)\pi_i(\vec{\theta}_i)d\vec{\theta}_i.$$

Αυτή η ποσότητα μπορεί να θεωρηθεί ως η πιθανοφάνεια για το ι-οστό μοντέλο και αναφέρεται ως η οριακή πιθανοφάνεια των δεδομένων.

Η οριακή πιθανοφάνεια ή οριακή κατανομή των δεδομένων $\vec{x}_n$ μπορεί να προσεγγιστεί χρησιμοποιώντας την μέθοδο Laplace για ολοκληρώματα. Αφαιρώντας την συμβολική εξάρτηση στο μοντέλο M_i, τότε η οριακή πιθανοφάνεια εκφράζεται ως:

$$p(\vec{x}_n) = \int f(\vec{x}_n|\vec{\theta})\pi(\vec{\theta})d\vec{\theta},$$

όπου $\vec{\theta}$ είναι ένα p-διάστατο διάνυσμα παραμέτρων. Αυτή η εξίσωση μπορεί να ξαναγραφεί ως:

$$p(\vec{x}_n) = \int exp\{log\, f(\vec{x}_n|\vec{\theta})\}\pi(\vec{\theta})d\vec{\theta} = \int exp\{l(\vec{\theta})\}\pi(\vec{\theta})d\vec{\theta}$$

όπου $l(\vec{\theta})$ είναι η συνάρτηση λογαριθμικής πιθανοφάνειας.

Σύμφωνα με την προσέγγιση Laplace, όταν ο αριθμός n των παρατηρήσεων είναι επαρκώς μεγάλος, τότε το ολοκλήρωμα προσεγγίζεται σε μια περιοχή της κορυφής ή επικρατούσας τιμής (mode) του $l(\vec{\theta})$ ή σε αυτήν την περίπτωση σε μια περιοχή του εκτιμητή μεγίστης πιθανοφάνειας $\hat{\vec{\theta}}$, με αποτέλεσμα η τιμή του ολοκληρώματος να βασίζεται στην συμπεριφορά της συνάρτησης στην περιοχή αυτή.
Αφού το $\partial l(\vec{\theta})/\partial\vec{\theta}|_{\vec{\theta}=\hat{\vec{\theta}}} = 0$ ισχύει για την εύρεση του εκτιμητή μέγιστης πιθανοφάνειας $\hat{\vec{\theta}}$ της παραμέτρου $\vec{\theta}$, το ανάπτυγμα Taylor της λογαριθμικής πιθανοφάνειας $l(\vec{\theta})$ γύρω από το $\hat{\vec{\theta}}$ δίνει

$$l(\vec{\theta}) = l\left(\hat{\vec{\theta}}\right) - \frac{n}{2}\left(\vec{\theta} - \hat{\vec{\theta}}\right)^T J(\hat{\vec{\theta}})(\vec{\theta} - \hat{\vec{\theta}}) + \ldots,$$

όπου

$$J\left(\hat{\vec{\theta}}\right) = -\frac{1}{n}\frac{\partial^2 l(\vec{\theta})}{\partial\vec{\theta}\partial\vec{\theta}^T}|_{\vec{\theta}=\hat{\vec{\theta}}} = -\frac{1}{n}\frac{\partial^2 log f(\vec{x}_n|\vec{\theta})}{\partial\vec{\theta}\partial\vec{\theta}^T}|_{\vec{\theta}=\hat{\vec{\theta}}}$$

Παρόμοια, αναπτύσσοντας την εκ των προτέρων κατανομή $\pi(\vec{\theta})$ σε σειρά Taylor γύρω από τον εκτιμητή μέγιστης πιθανοφάνειας $\hat{\vec{\theta}}$ θα ισχύσει

$$\pi(\vec{\theta}) = \pi\left(\hat{\vec{\theta}}\right) + \left(\vec{\theta} - \hat{\vec{\theta}}\right)^T \frac{\partial\pi(\vec{\theta})}{\partial\vec{\theta}}|_{\vec{\theta}=\hat{\vec{\theta}}} + \cdots$$

Αντικαθιστώντας και απλοποιώντας τα αποτελέσματα, η προσέγγιση της οριακής πιθανοφάνειας θα είναι:

$$p(\vec{x}_n) = \int exp\left\{l\left(\hat{\vec{\theta}}\right) - \frac{n}{2}\left(\vec{\theta} - \hat{\vec{\theta}}\right)^T J\left(\hat{\vec{\theta}}\right)\left(\vec{\theta} - \hat{\vec{\theta}}\right) + \ldots\right\}\left\{\pi\left(\hat{\vec{\theta}}\right) + \left(\vec{\theta} - \hat{\vec{\theta}}\right)^T \frac{\partial\pi(\vec{\theta})}{\partial\vec{\theta}}\Big|_{\vec{\theta}=\hat{\vec{\theta}}} + \cdots\right\} d\vec{\theta} \approx$$

$$\approx \exp\left\{l\left(\hat{\vec{\theta}}\right)\right\}\pi(\hat{\vec{\theta}}) \int \exp\left\{-\frac{n}{2}\left(\vec{\theta} - \hat{\vec{\theta}}\right)^T J\left(\hat{\vec{\theta}}\right)\left(\vec{\theta} - \hat{\vec{\theta}}\right)\right\} d\vec{\theta}$$

Χρησιμοποιήθηκε το γεγονός ότι το $\hat{\vec{\theta}}$ συγκλίνει στο $\vec{\theta}$ με τάξη $\hat{\vec{\theta}} - \vec{\theta} = O_p\ (n^{-1/2})$ και το γεγονός ότι η παρακάτω εξίσωση ισχύει:

$$\int (\vec{\theta} - \hat{\vec{\theta}})\, exp\left\{-\frac{n}{2}\left(\vec{\theta} - \hat{\vec{\theta}}\right)^T J\left(\hat{\vec{\theta}}\right)\left(\vec{\theta} - \hat{\vec{\theta}}\right)\right\} d\vec{\theta} = \vec{0}$$

Υπολογίζοντας το ολοκλήρωμα ως προς την παράμετρο $\vec{\theta}$ ισχύει:

$$\int \exp\{-\frac{n}{2}\left(\vec{\theta} - \hat{\vec{\theta}}\right)^T J\left(\hat{\vec{\theta}}\right)\left(\vec{\theta} - \hat{\vec{\theta}}\right)\} d\vec{\theta} = (2\pi)^{p/2} n^{-p/2} |J\left(\hat{\vec{\theta}}\right)|^{-1/2}$$

επειδή η προς ολοκλήρωση συνάρτηση είναι η συνάρτηση πυκνότητας της p-διάστατης κανονικής κατανομής με μέσο το διάνυσμα $\hat{\vec{\theta}}$ και πίνακα διασποράς-συνδιασποράς $J^{-1}(\hat{\vec{\theta}})/n$. Με αποτέλεσμα όταν το δείγμα είναι αρκετά μεγάλο, η οριακή πιθανοφάνεια να μπορεί να προσεγγιστεί από την σχέση

$$p(\vec{x}_n) \approx \exp\left\{l\left(\hat{\vec{\theta}}\right)\right\}\pi(\hat{\vec{\theta}})(2\pi)^{p/2} n^{-p/2} |J\left(\hat{\vec{\theta}}\right)|^{-1/2}$$

Λογαριθμίζοντας τα δύο μέλη και πολλαπλασιάζοντας με -2 θα ισχύει:

$$-2logp(\vec{x}_n) = -2\log\left\{\int f(\vec{x}_n|\vec{\theta})\,\pi(\vec{\theta})d\vec{\theta}\right\} \approx -2\, l\left(\hat{\vec{\theta}}\right) + plog(n) + log\left|J\left(\hat{\vec{\theta}}\right)\right| - plog(2\pi) - 2\log\pi(\hat{\vec{\theta}})$$

Αγνοώντας τους όρους με τάξη μικρότερη από $O(1)$ σε σχέση με το δείγμα θα ισχύει ο παρακάτω ορισμός.

Γενικός τύπος του Μπεϋζιανού Κριτηρίου Πληροφορίας (BIC).

Έστω $f(\vec{x}_n|\hat{\vec{\theta}})$ ένα στατιστικό μοντέλο εκτιμώμενο με την μέθοδο μέγιστης πιθανοφάνειας. Τότε το BIC δίνεται από:

$$BIC = -2logf\left(\vec{x}_n\middle|\hat{\vec{\theta}}\right) + plogn$$

όπου το p αντιπροσωπεύει την διάσταση του μοντέλου και n το μέγεθος του δείγματος. Επιλέγεται το μοντέλο με την μικρότερη τιμή BIC. Ο πρώτος όρος στον παραπάνω τύπο είναι, ο όρος ακρίβειας και δείχνει πόσο καλά ένα μοντέλο προσαρμόζεται στα δεδομένα. Ο δεύτερος όρος είναι ο όρος πολυπλοκότητας και είναι αυτός που ποινικοποιεί την πολυπλοκότητα του επιλεγμένου μοντέλου. Ο πρώτος όρος αυξάνεται, όσο το μοντέλο γίνεται πιο πολύπλοκο, ενώ ο δεύτερος μειώνεται αντίστοιχα. Θεωρητικά ο δεύτερος όρος ποινικοποιεί τα πολύπλοκα μοντέλα περισσότερο από ότι τα απλά. Οπότε σε συνδυασμό αυτοί οι δύο όροι προσπαθούν να εντοπίσουν απλά και ακριβή μοντέλα.
Το BIC είναι ένα κριτήριο αξιολόγησης μοντέλων που εκτιμάται χρησιμοποιώντας την μέθοδο της μέγιστης πιθανοφάνειας, ενώ το κριτήριο ισχύει υπό την προυπόθεση ότι το μέγεθος του δείγματος n είναι επαρκώς μεγάλο. Επίσης το κριτήριο υπολογίζεται προσεγγίζοντας την οριακή πιθανοφάνεια, η οποία συνδέεται με την εκ των προτέρων πιθανότητα του μοντέλου χρησιμοποιώντας την μέθοδο του Laplace για ολοκληρώματα.
Αν και υπάρχουν διαφορές στον τρόπο επιλογής των μοντέλων από τα κριτήρια AIC και BIC, η σημαντικότερη διαφορά τους που εμφανίζεται στις πρακτικές εφαρμογές είναι ότι το BIC συγκλίνει στο «καλύτερο» μοντέλο με μεγαλύτερη ακρίβεια, όταν διατίθενται μεγάλα δείγματα.

3.4 Βηματική Παλινδρόμηση (Stepwise Regression)

Στην στατιστική, η βηματική παλινδρόμηση (stepwise regression) περιλαμβάνει παλινδρομικά μοντέλα στα οποία η επιλογή των μεταβλητών πρόβλεψης πραγματοποιείται από μια αυτόματη διαδικασία. Ο παρακάτω αλγόριθμος είχε αρχικά προταθεί από τον Efroymson (1960) και περιγράφει την αυτόματη διαδικασία για την επιλογή μοντέλων σε περιπτώσεις που υπάρχει ένας μεγάλος αριθμός πιθανών επεξηγηματικών μεταβλητών. Οι βασικές προσεγγίσεις της διαδικασίας είναι:

- Η προς τα εμπρός επιλογή (Forward Selection) περιλαμβάνει ένα αρχικό μοντέλο χωρίς μεταβλητές και με την πρόσθεση κάθε μια μεταβλητής ελέγχεται με ένα κριτήριο επιλογής μοντέλων οι μεταβλητές που βελτιώνουν το μοντέλο περισσότερο. Η διαδικασία αυτή επαναλαμβάνεται μέχρι καμία από τις υπόλοιπες μεταβλητές να μην βελτιώνει άλλο το μοντέλο.
- Η προς τα πίσω επιλογή (Backward Selection) περιλαμβάνει ένα αρχικό μοντέλο με όλες τις υποψήφιες μεταβλητές και αφαιρώντας κάθε μια μεταβλητή ερευνάται με ένα κριτήριο επιλογής μοντέλων ποια μεταβλητή από αυτές που αφαιρούνται βελτιώνει με την διαγραφή της το μοντέλο περισσότερο. Η διαδικασία αυτή επαναλαμβάνεται μέχρι καμία από τις υπόλοιπες μεταβλητές να μην βελτιώνει άλλο το μοντέλο.

- Η αμφίδρομη επιλογή (Bidirectional) αποτελεί ένα συνδυασμό των δύο μεθόδων, όπου ελέγχεται σε κάθε βήμα οι μεταβλητές που θα προσθεθούν ή θα αφαιρεθούν.

Η μέθοδος αυτή όπως προαναφέρθηκε χρησιμοποιείται κυρίως στην ανάλυση παλινδρόμησης και έχει την δυνατότητα να ψάχνει από ένα μεγάλο αριθμό πιθανών μοντέλων. Αυτό έχει σαν αποτέλεσμα συχνά την υπερπροσαρμογή των δεδομένων. Ένας έλεγχος σφαλμάτων σε μοντέλα που έχουν κατασκευαστεί με την βηματική παλινδρόμηση είναι η αξιολόγηση του μοντέλου με ένα σύνολο δεδομένων που δεν έχει χρησιμοποιηθεί για την κατασκευή του. Συχνά, η κατασκευή του μοντέλου γίνεται με την χρήση ενός δείγματος του συνόλου των δεδομένων (π.χ. 70%) ενώ το υπόλοιπο σύνολο δεδομένων (30%) χρησιμοποιείται για την αξιολόγηση της ακρίβειας του μοντέλου. Η μέθοδος αυτή θα χρησιμοποιηθεί σε συνδυασμό με τα παραπάνω κριτήρια στις προσομοιώσεις της εργασίας.

ΚΕΦΑΛΑΙΟ 4

Μέθοδοι Αναδειγματοληψίας

Από τα σημαντικότερα προβλήματα στην εφαρμοσμένη στατιστική είναι ο προσδιορισμός μιας εκτιμήτριας για μια συγκεκριμένη παράμετρο και η αξιολόγηση της ακρίβειας της εκτιμήτριας αυτής με την εύρεση εκτιμήσεων του τυπικού της σφάλματος. Για την ανάπτυξη της μεθοδολογίας των κριτηρίων πληροφορίας, όπως προαναφέρθηκε, χρησιμοποιήθηκε η έννοια της KL απόστασης και η εκτίμηση της αναμενόμενης λογαριθμικής πιθανοφάνειας, η οποία είχε ως αποτέλεσμα την ύπαρξη μεροληψίας. Η μεροληψία αυτή εμφανίζεται ως όρος ποινής, για παράδειγμα για το AIC, ανάλογα με τον αριθμό των παραμέτρων στο μοντέλο. Ένας τρόπος ελάττωσης της μεροληψίας μπορεί να γίνει με την χρήση στατιστικών μεθόδων αναδειγματοληψίας όπως είναι η Jackknife και η Bootstrap μέθοδοι.

Το Bootstrap αποτελεί μια μέθοδο αναδειγματοληψίας από το αρχικό σύνολο δεδομένων. Με τα άρθρα από τους Efron και Gong (1983), Efron και Tibshirani (1986) και τη μονογραφία του Efron (1982) άρχισε η στατιστική έρευνα για το bootstrap να εξελίσσεται ταχύτατα και να αναγνωρίζεται η σημασία και η ευρεία εφαρμογή του. Η μέθοδος Jackknife προυπήρχε της Bootstrap και η βασική τεχνική αναπτύχθηκε από τον Maurice Quenouille (1949,1956). Ο Tukey (1958) ανέπτυξε περαιτέρω την τεχνική και πρότεινε το όνομα jackknife που δόθηκε στην μέθοδο.

4.1 Η Bootstrap Μέθοδος

Σύμφωνα με Abdelhak M. Zoubir, D.Robert Iskander (2004) και Bradley Efron, R. J. Tibshirani (1993) έστω $X = \{X_1, X_2, \dots, X_n\}$ ένα δείγμα, δηλαδή μια συλλογή από n τυχαίους αριθμούς που λαμβάνονται από μια άγνωστη κατανομή F. Τα X_i υποθέτονται ανεξάρτητα και ισόνομα. Έστω θ άγνωστη παράμετρος που μπορεί να

αποτελεί την μέση τιμή ή την διασπορά, για παράδειγμα, της κατανομής F. Σκοπός είναι να εντοπιστεί η κατανομή της $\hat{\theta}$ εκτιμήτριας του θ από το δείγμα X. Ένας τρόπος ώστε να βρεθεί η κατανομή του $\hat{\theta}$ είναι η επανάληψη του πειράματος από το οποίο λήφθηκε το δείγμα αρκετές φορές ώστε να προσεγγιστεί η αντίστοιχη κατανομή από την εμπειρική κατανομή. Αυτό βέβαια δεν είναι πρακτικό, αφού το εν λόγω πείραμα μπορεί να μην μπορεί να επαναληφθεί, οπότε και να μην υπάρχουν αρκετά δείγματα. Παρακάτω παρουσιάζονται οι βασικές έννοιες της εμπειρικής κατανομής και της plug-in μεθόδου που χρησιμοποιούνται από την bootstrap μέθοδο αναδειγματοληψίας.

4.1.1 Η Συνάρτηση Εμπειρικής Κατανομής

Έστω $X_1, \dots, X_n \sim F$ $με$ $F(x) = P(X \leq x)$. Η F μπορεί να εκτιμηθεί με την χρήση της εμπειρικής κατανομής $\widehat{F_n}$, η οποία αποτελεί την αθροιστική συνάρτηση κατανομής που κατανέμει κάθε X_i με πιθανότητα 1/n. Η $\widehat{F_n}$ εκφράζεται ως:

$$\widehat{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) = \frac{\#\{X_i \leq x\}}{n}$$

με

$$I(X_i \leq x) = \begin{cases} 1 \ αν \ X_i \leq x \\ 0 \ αν \ X_i > x \end{cases}$$

Η εμπειρική πιθανότητα ενός συνόλου Α είναι:

$$\widehat{P_n}(A) = \frac{\#\{X_i \in A\}}{n}$$

Στην περίπτωση που στο δείγμα υπάρχουν επαναλαμβανόμενες τιμές η πιθανότητα καθενός X_i που κατανέμεται από την $\widehat{F_n}$ είναι ίση με την πιθανότητα εμφάνισής του στο δείγμα.

Έστω παράμετρος που είναι συνάρτηση του F,

$$\theta = t(F)$$

δηλαδή η τιμή θ της παραμέτρου μπορεί να ορισθεί εφαρμόζοντας κάποια αριθμητική διαδικασία $t(.)$ στην συνάρτηση κατανομής F. Η εμπειρική κατανομή είναι ένας τρόπος εκτίμησης της πραγματικής κατανομής από την οποία λαμβάνεται το δείγμα. Η εμπειρική κατανομή είναι μια τεχνική που χρησιμοποιείται και από την μέθοδο plug-in.

4.1.2 Η plug-in διαδικασία

Προβλήματα στατιστικής συμπερασματολογίας συχνά περιλαμβάνουν την εκτίμηση $\theta = t(F)$ μιας πιθανότητας κατανομής F, για την οποία μπορεί να μην έχουμε καμία πληροφορία. Για αυτό το λόγο επιλέγεται η εκτίμησή της, η οποία συμβολίζεται με $\hat{F}$. Τότε η εκτίμηση της άγνωστης παραμέτρου θα είναι συνάρτηση της εκτιμήτριας $\hat{F}$. Η διαδικασία αυτή είναι μια μέθοδος εκτίμησης παραμέτρων από δείγματα και ονομάζεται plug-in. Σύμφωνα με Michael W. Trosset (2009), η εκτίμηση plug-in μιας παραμέτρου $\theta = t(F)$ εκφράζεται επομένως ως:

$$\hat{\theta} = t(\hat{F})$$

Πολλές φορές επιλέγεται για την εκτίμηση της συνάρτησης $\theta = t(F)$ της κατανομής πιθανότητας F, η χρήση της συνάρτησης εμπειρικής κατανομής $\widehat{F_n}$, $δηλαδή$

$$\hat{\theta} = t(\widehat{F_n}).$$

Χαρακτηριστικά παραδείγματα της εκτίμησης plug-in είναι:

Παράδειγμα 1

Για

$\theta = E_F(x) = \int x dF(x) = \int x f(x) dx$ θα ισχύει:

$$\hat{\theta} = E_{\widehat{F_n}}(x) = \int x d\widehat{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X_n}$$

Το τυπικό σφάλμα εκφράζεται ως:

$$se = \sqrt{Var(\overline{X_n})} = \frac{\sigma}{\sqrt{n}}$$

Παράδειγμα 2

Έστω $\sigma^2 = Var(X) = \int x^2 dF(x) - (\int x dF(x))^2$. Η plug-in εκτιμήτρια θα είναι:

$$\hat{\sigma}^2 = \int x^2 d\hat{F}_n(x) - (\int x d\hat{F}_n(x))^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$$

Η διαδικασία plug-in χρησιμοποιείται ιδιαίτερα από την bootstrap μέθοδο. Παρακάτω θα χρησιμοποιηθεί η μέθοδος bootstrap για τον υπολογισμό της μεροληψίας και του τυπικού σφάλματος της εκτιμήτριας $\hat{\theta} = t(\hat{F})$. Η bootstrap μπορεί να υπολογίσει μεροληψίες και τυπικά σφάλματα, όσο πολύπλοκη και αν είναι η συνάρτηση $\theta = t(F)$.

4.1.3 Εύρεση του Τυπικού Σφάλματος μέσης Τιμής

Μια σημαντική πληροφορία στην ανάλυση δεδομένων είναι η ακρίβεια της εκτιμήτριας $\hat{\theta}$. Η bootstrap μέθοδος μπορεί να κατασκευάσει εκτιμήσεις της ακρίβειας με την χρήση της plug-in μεθόδου, ώστε να εκτιμηθεί το τυπικό σφάλμα της εκτιμήτριας. Αρχικά αναλύεται η εκτίμηση του τυπικού σφάλματος της μέσης τιμής, σε περιπτώσεις που η plug-in μέθοδος μπορεί να υπολογιστεί αναλυτικά.

Έστω x τυχαία μεταβλητή με πιθανότητα κατανομής F. Έστω η αναμενόμενη τιμή και η διασπορά ως προς την F να εκφράζονται ως:

$$\mu_F = E_F(x), \quad \sigma_F^2 = var_F(x) = E_F[(x - \mu_F)^2]$$

Και

$$x \sim (\mu_F, \sigma_F^2)$$

Έστω $(x_1, \dots, x_n)$ ένα τυχαίο δείγμα μεγέθους n από την κατανομή F. Η μέση τιμή του δείγματος $\bar{x} = \sum_{i=1}^{n} x_i/n$ έχει αναμενόμενη τιμή μ_F και διασπορά σ_F^2/n,

$$\bar{x} \sim (\mu_F, \frac{\sigma_F^2}{n})$$

Επομένως η αναμενόμενη τιμή του $\bar{x}$ είναι ίδια με του x, αλλά η διασπορά του $\bar{x}$ είναι 1/n φορές η διασπορά του x. Όσο μεγαλύτερο το n τόσο μικρότερη η $var(\bar{x})$, οπότε τόσο καλύτερη η εκτίμηση της μ_F.

Το τυπικό σφάλμα της μέσης τιμής $\bar{x}$, συμβολίζεται με $se_F(\bar{x})$ ή $se(\bar{x})$ και εκφράζεται ως:

$$se_F(\bar{x}) = [var_F(\bar{x})]^{1/2} = \frac{\sigma_F}{\sqrt{n}}$$

Σύμφωνα με το κεντρικό οριακό θεώρημα για $n \to \infty$ το $\bar{x}$ ακολουθεί την κανονική κατανομή, δηλαδή:

$$\bar{x} \sim N(\mu_F, \frac{\sigma_F^2}{n})$$

4.1.4 Εκτίμηση του Τυπικού Σφάλματος της Μέσης Τιμής.

Έστω τυχαίο δείγμα από την F και έστω $\bar{x}$ η εκτιμήτρια της μ_F. Για την εύρεση του τυπικού σφάλματος, παρατηρείται ότι $se_F(\bar{x}) = \sigma_F/\sqrt{n}$ η οποία εξαρτάται από την άγνωστη κατανομή F και άρα δεν μπορεί να χρησιμοποιηθεί. Για αυτό θα χρησιμοποιηθεί η plug-in μέθοδος. Αντικαθιστώντας την F *για την* $\hat{F}$, η plug-in εκτιμήτρια της $\sigma_F = [E_F(x - \mu_F)^2]^{1/2}$ είναι:

$$\hat{\sigma} = \sigma_{\hat{F}} = \left\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^{1/2}$$

αφού $\mu_{\hat{F}} = \bar{x}$ και $E_{\hat{F}}g(x) = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$ για οποιαδήποτε συνάρτηση g.

Οπότε η εκτιμήτρια τυπικού σφάλματος θα είναι $\widehat{se}(\bar{x}) = se_{\hat{F}}(\bar{x})$,

$$\widehat{se}(\bar{x}) = \frac{\sigma_{\hat{F}}}{\sqrt{n}} = \left\{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n^2}\right\}^{1/2}$$

Όπως φαίνεται χρησιμοποιήθηκε η plug-in μέθοδος, αρχικά για την εκτίμηση της μ_F από την $\mu_{\hat{F}} = \bar{x}$ και στην συνέχεια για την εκτίμηση του τυπικού σφάλματος $se_F(\bar{x})$ από το $se_{\hat{F}}(\bar{x})$. Η bootstrap εκτίμηση του τυπικού σφάλματος χρησιμοποιεί ουσιαστικά την plug-in μέθοδο για να εκτιμήσει το τυπικό σφάλμα μιας τυχαίας στατιστικής συνάρτησης. Παραπάνω αναλύθηκε η περίπτωση που η τυχαία συνάρτηση είναι η $\hat{\theta} = \bar{x}$, και η οποία κατέληξε στην γνωστή εκτίμηση του τυπικού σφάλματος. Το πλεονέκτημα της bootstrap είναι ότι μπορεί να εφαρμοστεί για οποιαδήποτε τυχαία συνάρτηση.

4.1.5 Η Bootstrap Εκτιμήτρια του Τυπικού Σφάλματος και η Έννοια της Μη Παραμετρικής Bootstrap Μεθόδου

Έστω ένα τυχαίο δείγμα $x = (x_1, \dots, x_n)$ το οποίο έχει ληφθεί από μια άγνωστη κατανομή πιθανότητας F. Σκοπός είναι η εκτίμηση μιας παραμέτρου $\theta = t(F)$ από το δείγμα x. Για το λόγο αυτό υπολογίζεται μια εκτιμήτρια $\hat{\theta} = s(x)$ από το δείγμα x (η $s(x)$ μπορεί να είναι και η plug-in εκτιμήτρια $t(\hat{F})$). Η μέθοδος bootstrap παρουσιάστηκε το 1979 ως μια υπολογιστική μέθοδος για την εκτίμηση του τυπικού σφάλματος της εκτιμήτριας $\hat{\theta}$.

Η bootstrap εκτίμηση του τυπικού σφάλματος δεν απαιτεί θεωρητικούς υπολογισμούς και μπορεί να υπολογιστεί ακόμα και αν η εκτιμήτρια $\hat{\theta}$ είναι μαθηματικώς πολύπλοκη.
Ένα δείγμα bootstrap ορίζεται ως το τυχαίο δείγμα μεγέθους n, το οποίο έχει ληφθεί από μια εκτίμηση $\hat{F}$ μιας τυχαίας κατανομής F. Δηλαδή αν $x^* = (x_1^*, \dots, x_n^*)$

$$\hat{F} \rightarrow (x_1^*, \dots, x_n^*)$$

Η $\hat{F}$ θα μπορούσε να αποτελεί και την εμπειρική κατανομή, όπως αναφέρθηκε παραπάνω. Ο συμβολισμός * δηλώνει ότι το x^* δεν είναι το αρχικό δείγμα αλλά ένα τυχαίο δείγμα, το οποίο προήλθε από την εφαρμογή της αναδειγματοληψίας με επανάθεση στον αρχικό πληθυσμό x. Αντίστοιχα με το bootstrap σύνολο δεδομένων x^* είναι και η bootstrap τιμή του $\hat{\theta}$ (ή αλλιώς bootstrap replication),

$$\hat{\theta}^* = s(x^*)$$

Το $s(x^*)$ είναι αποτέλεσμα της εφαρμογής της ίδιας συνάρτησης $s(.)$ στο x^* όπως είχε εφαρμοστεί και στο x. Για παράδειγμα αν $s(x)$ είναι η δειγματική μέση τιμή $\bar{x}$ τότε το $s(x^*)$ είναι η μέση τιμή του bootstrap συνόλου δεδομένων, $\bar{x}^* = \sum_{i=1}^{n} x_i^*/n$.

Στην περίπτωση που η $\hat{F}$ είναι η εμπειρική κατανομή, η διενέργια αναδειγματοληψίας με την μέθοδο bootstrap από την κατανομή αυτή αναφέρεται ως μη παραμετρικό bootstrap. Τα βήματα της μεθόδου παρουσιάζονται στον παρακάτω πίνακα:

Βήμα 0. Πραγματοποιείται το πείραμα που θα παράγει το τυχαίο δείγμα $X = \{X_1, X_2, \dots, X_n\}$ Και θα υπολογιστεί η εκτιμήτρια $\hat{\theta}$ από το δείγμα X

Βήμα 1. Κατασκευή της εμπειρικής κατανομής $\hat{F}$, όπου κάθε παρατήρηση θα έχει ίδια μάζα πιθανότητας

$$X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$$

Βήμα 2. Από την $\hat{F}$ παράγεται ένα δείγμα

$$X^* = \{X_1^*, X_2^*, \dots, X_n^*\},$$

το οποίο αποτελεί το δείγμα Bootstrap.

Βήμα 3. Προσεγγίζεται η κατανομή του $\hat{\theta}$ από την κατανομή του $\hat{\theta}^*$, το οποίο προέρχεται από το Bootstrap δείγμα X^*.

Αν και ο παραπάνω πίνακας προυποθέτει την κατασκευή της $\hat{F}$ και την αναδειγματοληψία από αυτή, στην πραγματικότητα δεν χρειάζεται η άμεση εκτίμηση της F. Αυτό σημαίνει ότι η αναδειγματοληψία Bootstrap δεν απαιτεί τον άμεσο υπολογισμό της συνάρτησης εμπειρικής κατανομής ή της εκτίμησης της μέσης τιμής, της διασποράς ή των ροπών της F. Με την χρήση του μη παραμετρικού bootstrap, χρησιμοποιείται το τυχαίο δείγμα $X = \{X_1, X_2, \dots, X_n\}$ και παράγεται ένα νέο δείγμα με την αναδειγματοληψία με επανάθεση από το X. Το νέο αυτό δείγμα αποτελεί ένα bootstrap δείγμα. Στην συνέχεια παρουσιάζεται η bootstrap εκτιμήτρια του τυπικού σφάλματος και της μεροληψίας.

Η bootstrap εκτιμήτρια του $se_F(\hat{\theta})$, του τυπικού σφάλματος της στατιστικής συνάρτησης $\hat{\theta}$, είναι μια εκτιμήτρια plug-in που χρησιμοποιεί την εμπειρική συνάρτηση κατανομής $\hat{F}_n = \hat{F}$ αντί της άγνωστης κατανομής F. Η bootstrap εκτιμήτρια του $se_F(\hat{\theta})$ ορίζεται από το

$$se_{\hat{F}}(\hat{\theta}^*)$$

Δηλαδή η εκτίμηση bootstrap του $se_F(\hat{\theta})$ είναι το τυπικό σφάλμα του $\hat{\theta}$ για σύνολα δεδομένων μεγέθους n, τα οποία έχουν τυχαία ληφθεί από την $\hat{F}$, η οποία έστω αποτελεί την εμπειρική συνάρτηση κατανομής.

Το $se_{\hat{F}}(\hat{\theta}^*)$ αποτελεί την ιδανική bootstrap εκτιμήτρια (ideal bootstrap estimate) του τυπικού σφάλματος $\hat{\theta}$. Παρ'όλα αυτά για οποιαδήποτε εκτιμήτρια εκτός από την μέση τιμή, η πολυπλοκότητα του υπολογισμού της αριθμητικής τιμής της ιδανικής εκτιμήτριας είναι πολύ δύσκολη. Ένας τρόπος προσέγγισης της τιμής του $se_{\hat{F}}(\hat{\theta}^*)$ είναι ο αλγόριθμος Bootstrap. Ο αλγόριθμος αυτός λαμβάνει πολλά ανεξάρτητα bootstrap δείγματα, αξιολογώντας τις αντίστοιχες bootstrap τιμές του $\hat{\theta}^*$ (ή αλλιώς bootstrap replications) και εκτιμώντας το τυπικό σφάλμα του $\hat{\theta}$ από την εμπειρική τυπική απόκλιση των replications. Η bootstrap εκτιμήτρια του τυπικού σφάλματος συμβολίζεται με $\widehat{se}_B$, όπου Β ο αριθμός των bootstrap δειγμάτων που λήφθηκαν. Ο παρακάτω αλγόριθμος αποτελεί μια αναλυτική αναπαράσταση της διαδικασίας υπολογισμού.

Ο Bootstrap αλγόριθμος για την εκτίμηση τυπικών σφαλμάτων.

Βήμα 1. Επιλογή Β ανεξάρτητων bootstrap δειγμάτων $x^{*1}, \dots, x^{*B}$ με το κάθε ένα να περιέχει n τιμές δεδομένων, οι οποίες έχουν ληφθεί με επανάθεση από το x.

Βήμα 2. Εύρεση του bootstrap replication που αντιστοιχεί σε κάθε δείγμα,

$$\hat{\theta}^*(b) = s(x^{*b}), b = 1,2,..,B$$

Βήμα 3. Εύρεση της εκτιμήτριας του τυπικού σφάλματος $se_F(\hat{\theta})$ από την δειγματική τυπική απόκλιση των B replications.

$$\widehat{se}_B = \left\{ \frac{\sum_{b=1}^{B} [\hat{\theta}^*(b) - \hat{\theta}^*(.)]^2}{B-1} \right\}^{1/2}$$

$$\text{όπου } \hat{\theta}^*(.) = \sum_{b=1}^{B} \frac{\hat{\theta}^*(b)}{B}$$

Όταν το B τείνει στο άπειρο το όριο του $\widehat{se}_B$ είναι η ιδανική bootstrap εκτίμηση του $se_F(\hat{\theta})$,

$$\lim_{B\to\infty} \widehat{se}_B = se_{\hat{F}} = se_{\hat{F}}(\hat{\theta}^*)$$

Το γεγονός ότι το $\widehat{se}_B$ προσεγγίζει το $se_{\hat{F}}$ καθώς το B τείνει στο άπειρο ισοδυναμεί με την εμπειρική τυπική απόκλιση να προσεγγίζει την τυπική απόκλιση του πληθυσμού όσο το μέγεθος των επαναλήψεων μεγαλώνει. Ο πληθυσμός σε αυτή την περίπτωση, είναι ο πληθυσμός των τιμών $\hat{\theta}^* = s(x^*)$, $\hat{F} \to (x_1^*, \dots, x_n^*) = x^*$.

Η ιδανική Bootstrap εκτιμήτρια $se_{\hat{F}}(\hat{\theta}^*)$ και η προσέγγισή της $\widehat{se}_B$ επειδή βασίζονται στην $\hat{F}$ την μη παραμετρική εκτιμήτρια του πληθυσμού F, αποτελούν τις εκτιμήτριες του μη παραμετρικού Bootstrap.

4.1.6 Παραμετρική Εκτιμήτρια Τυπικού Σφάλματος

Εν αντιθέσει, η παραμετρική εκτιμήτρια Bootstrap για το τυπικό σφάλμα ορίζεται ως:

$$se_{\hat{F}_{par}}(\hat{\theta}^*)$$

Όπου $\hat{F}_{par}$ είναι μια εκτιμήτρια της F, η οποία προέρχεται από ένα παραμετρικό μοντέλο για τα δεδομένα. Όπως και στο μη παραμετρικό μοντέλο η ιδανική παραμετρική bootstrap εκτιμήτρια δεν είναι εύκολα υπολογίσιμη παρά μόνο στην περίπτωση που το $\hat{\theta}$ είναι η μέση τιμή. Για αυτό το λόγο προσεγγίζεται η ιδανική εκτιμήτρια με την bootstrap αναδειγματοληψία. Αντί όμως να γίνεται η

αναδειγματοληψία με επανάθεση από τα δεδομένα, λαμβάνονται B δείγματα μεγέθους n από την παραμετρική εκτιμήτρια του πληθυσμού $\hat{F}_{par}$:

$$\hat{F}_{par} \rightarrow (x_1^*, \dots, x_n^*)$$

Με την κατασκευή των bootstrap δειγμάτων, εκτελούνται τα βήματα 2 και 3 του παραπάνω αλγόριθμου.

Η μέθοδος Bootstrap παρουσιάζει κάποια πλεονεκτήματα σε σχέση με τις υπόλοιπες μαθηματικές μεθόδους:

1. Το μη παραμετρικό Bootstrap επιτρέπει στον αναλυτή την δυνατότητα να μην χρειάζεται να κάνει παραμετρικές υποθέσεις για τον υποκείμενο πληθυσμό.
2. Το παραμετρικό Bootstrap παρέχει πιο ακριβείς απαντήσεις και μπορεί να εφαρμοστεί σε προβλήματα που μαθηματικές φόρμουλες δεν μπορούν να εφαρμοστούν.

Υπάρχουν περιπτώσεις που το μη παραμετρικό bootstrap μπορεί να αποτύχει να δώσει σωστά αποτελέσματα. Μερικές τέτοιες περιπτώσεις που παρουσιάζουν αυξημένη δυσκολία είναι όταν η εμπειρική συνάρτηση κατανομής $\hat{F}$ δεν είναι καλή εκτιμήτρια της πραγματικής κατανομής στην βαριά ουρά (extreme tail). Σε τέτοιες περιπτώσεις για να βελτιωθεί το αποτέλεσμα είτε μερική παραμετρική γνώση της F χρειάζεται ή απαιτείται η ομαλοποίηση της $\hat{F}$.

4.1.7 Η Bootstrap Εκτιμήτρια της Μεροληψίας.

Όπως παρουσιάστηκε, ένα μέγεθος της ακρίβειας για έναν εκτιμητή $\hat{\theta}$, αποτελεί το τυπικό σφάλμα. Υπάρχουν και άλλα μέτρα της στατιστικής ακρίβειας που μετρούν διαφορετικές όψεις της συμπεριφοράς του $\hat{\theta}$. Ένα τέτοιο μέγεθος είναι και η μεροληψία, η διαφορά μεταξύ της αναμενόμενης τιμής μιας εκτιμήτριας $\hat{\theta}$ και της ποσότητας θ, που είναι προς εκτίμηση. Ο αλγόριθμος Bootstrap προσαρμόζεται ώστε να υπολογίζει εκτιμήτριες της μεροληψίας, όπως και στο τυπικό σφάλμα.

Έστω η μη παραμετρική περίπτωση και έστω η άγνωστη κατανομή πιθανότητας F και $x = (x_1, \dots, x_n)$ δεδομένα από τυχαία αναδειγματοληψία $F \rightarrow x$. Θα γίνει η εκτίμηση της παραμέτρου $\theta = t(F)$. Έστω ότι μια εκτιμήτρια είναι η τυχαία στατιστική συνάρτηση $\hat{\theta} = s(x)$.

Η μεροληψία του $\hat{\theta} = s(x)$ ως μια εκτιμήτρια του θ, ορίζεται ως η διαφορά μεταξύ της αναμενόμενης τιμής του $\hat{\theta}$ και της τιμής της παραμέτρου θ,

$$bias_F = bias_F(\hat{\theta}, \theta) = E_F[s(x)] - t(F)$$

Συνήθως η μεγάλη μεροληψία δεν είναι επιθυμητή για την απόδοση της εκτιμήτριας. Αμερόληπτες εκτιμήτριες, εκείνες δηλαδή για τις οποίες ισχύει $E_F(\hat{\theta}) = \theta$ είναι αυτές που προτιμούνται στην στατιστική. Οι plug-in εκτιμήτριες $\hat{\theta} = t(\hat{F})$ δεν είναι απαραίτητα αμερόληπτες αλλά τείνουν να έχουν μικρότερες μεροληψίες σε σύγκριση με την τάξη του τυπικού σφάλματός τους. Αυτό αποτελεί ένα από τα μεγάλα πλεονεκτήματα της μεθόδου plug-in.

Για την εκτίμηση της μεροληψίας οποιασδήποτε εκτιμήτριας $\hat{\theta} = s(x)$ θα χρησιμοποιηθεί η μέθοδος bootstrap. Η Bootstrap εκτιμήτρια της μεροληψίας ορίζεται ως η εκτιμήτρια που λαμβάνεται με την αντικατάσταση του F από το $\hat{F}$ στην εξίσωση του $bias_F$, δηλαδή:

$$bias_{\hat{F}} = E_{\hat{F}}[s(x^*)] - t(\hat{F})$$

Το $t(\hat{F})$ αποτελεί την plug-in εκτιμήτρια του θ και μπορεί να είναι διαφορετική από τη $\hat{\theta} = s(x)$. Δηλαδή, $bias_{\hat{F}}$ είναι η plug-in εκτιμήτρια του $bias_F$ είτε το $\hat{\theta}$ είναι η plug-in εκτιμήτρια του θ είτε όχι.

Έαν η στατιστική συνάρτηση $s(x)$ είναι η μέση τιμή και η $t(F)$ είναι η μέση τιμή του πληθυσμού τότε $bias_{\hat{F}} = 0$. Το οποίο ισχύει αφού η μέση τιμή είναι η αμερόληπτη εκτιμήτρια της πληθυσμιακής μέσης τιμής $bias_F = 0$. Παρ'όλα αυτά συνήθως μια στατιστική συνάρτηση έχει μεροληψία και μια εκτιμήτρια αυτής της μεροληψίας είναι το $bias_{\hat{F}}$. Ένα χαρακτηριστικό παράδειγμα αποτελεί και η δειγματική διασπορά $s(x) = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$, της οποίας η μεροληψία είναι $(-1/n)$ φορές η διασπορά του πληθυσμού. Άρα το $bias_{\hat{F}}$ θα έχει την μορφή

$$bias_{\hat{F}} = (-\frac{1}{n^2})\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Μερικές χαρακτηριστικές εφαρμογές του μη παραμετρικού Bootstrap.

Παράδειγμα 1

Έστω το πρόβλημα της εκτίμησης της διασποράς άγνωστης κατανομής με παραμέτρους μ και σ, οι οποίες δηλώνονται με $F_{\mu,\sigma}$ δεδομένου τυχαίου δείγματος $X = \{X_1, X_2, \dots, X_n\}$. Δύο διαφορετικές εκτιμήτριες είναι οι:

$$\hat{\sigma}_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)^2$$

Και

$$\hat{\sigma}_b^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)^2$$

Θα ισχύει για τις δύο εκτιμήτριες ότι:

$$E[\hat{\sigma}_u^2] = \sigma^2 \quad και \quad E[\hat{\sigma}_b^2] = (1 - \frac{1}{n})\sigma^2$$

Οπότε η $\hat{\sigma}_b^2$ είναι μια μεροληπτική εκτιμήτρια του σ^2 ενώ το $\hat{\sigma}_u^2$ μια αμερόληπτη. Με την χρήση του Bootstrap, μπορεί να εκτιμηθεί η παρακάτω μεροληψία:

$$b(\hat{\sigma}^2) = E[\hat{\sigma}^2 - \sigma^2]$$

από

$$E_*[\hat{\sigma}^{*2} - \sigma^2],$$

όπου $\hat{\sigma}^2$ είναι η εκτιμήτρια μέγιστης πιθανοφάνειας του σ^2 και $E_*[.]$ η αναμενόμενη τιμή ως προς την δειγματοληψία bootstrap. Τα βήματα στον παραπάνω πίνακα για το μη παραμετρικό bootstrap μπορούν να προσαρμοστούν ώστε να περιγράφουν την διαδικασία για την εκτίμηση της μεροληψίας. Δηλαδή:

Βήμα 0. Πραγματοποιείται το πείραμα που θα παράγει το τυχαίο δείγμα
$X = \{X_1, X_2, \dots, X_n\}$
Και θα υπολογιστούν οι εκτιμήτριες $\hat{\sigma}_u^2$ και $\hat{\sigma}_b^2$ σύμφωνα με τις παραπάνω εξισώσεις.

Βήμα 1. Αναδειγματοληψία. Έστω τυχαίο δείγμα μεγέθους n, το οποίο έχει ληφθεί με επανάθεση από το X.

Βήμα 2. Υπολογισμός της bootstrap εκτιμήτριας. Υπολογισμός των bootstrap Εκτιμητριών $\hat{\sigma}_u^{*2}$ και $\hat{\sigma}_b^{*2}$ από το X^* με τον ίδιο τρόπο που τα $\hat{\sigma}_u^2$ και $\hat{\sigma}_b^2$ υπολογίστηκαν αλλά με την αναδειγματοληψία X^*.

Βήμα 3. Επανάληψη. Τα βήματα 1 και 2 επαναλαμβάνονται ώστε να αποκτηθεί ένα σύνολο B bootstrap εκτιμητριών $\hat{\sigma}_{u,1}^{*2}, \dots, \hat{\sigma}_{u,B}^{*2}$ και $\hat{\sigma}_{b,1}^{*2}, \dots, \hat{\sigma}_{b,B}^{*2}$.

Βήμα 4. Η εκτίμηση της μεροληψίας. Εκτιμάται η b ($\hat{\sigma}_u^2$) από:

$$b_*(\hat{\sigma}_u^{*2}) = \frac{1}{B}\sum_{i=1}^{B} \hat{\sigma}_{u,i}^{*2} - \hat{\sigma}_b^2$$

και του b ($\hat{\sigma}_b^2$) από

$$b_*(\hat{\sigma}_b^{*2}) = \frac{1}{B}\sum_{i=1}^{B} \hat{\sigma}_{b,i}^{*2} - \hat{\sigma}_b^2$$

Παράδειγμα 2

Διάστημα εμπιστοσύνης για την μέση τιμή.

Έστω $X_1, \dots, X_n$ n ανεξάρτητες και ισόνομες τυχαίες μεταβλητές από μια άγνωστη κατανομή $F_{\mu,\sigma}$. Για την εύρεση μιας εκτιμήτριας και ενός διαστήματος 100(1-α)% για την μέση τιμή μ, χρησιμοποιείται η δειγματική μέση τιμή ως μια εκτίμηση για το μ

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Ένα διάστημα εμπιστοσύνης για το μ μπορεί να βρεθεί προσδιορίζοντας την κατανομή του $\hat{\mu}$ και βρίσκοντας τιμές $\hat{\mu}_L, \hat{\mu}_U$ τέτοιες ώστε:

$$\Pr[\hat{\mu}_L \leq \mu \leq \hat{\mu}_U] = 1 - \alpha$$

Η κατανομή του $\hat{\mu}$ εξαρτάται από την κατανομή των X_i που είναι άγνωστη. Στην περίπτωση που το n είναι μεγάλο η κατανομή του $\hat{\mu}$ μπορεί να προσεγγιστεί από την Γκαουσιανή κατανομή, σύμφωνα με το Κεντρικό Οριακό Θεώρημα, αλλά τέτοια προσέγγιση δεν μπορεί να ισχύσει όταν το n είναι μικρό.

Το bootstrap, όπως και στις παραπάνω περιπτώσεις, υποθέτει ότι το τυχαίο δείγμα $X = \{X_1, X_2, \dots, X_n\}$ αποτελεί την κατανομή που ψάχνουμε. Οπότε με την αναδειγματοληψία απο το X αρκετές φορές και υπολογίζοντας το $\hat{\mu}$ για κάθε ένα από τα νέα δείγματα, λαμβάνεται μια bootstrap κατανομή για το $\hat{\mu}$ που προσεγγίζει την κατανομή του $\hat{\mu}$ και από το οποίο μπορεί να κατασκευαστεί ένα διάστημα εμπιστοσύνης για το μ. Αυτή η διαδικασία παρουσιάζεται και στον παρακάτω πίνακα.

Βήμα 0. Πραγματοποιείται το πείραμα που θα παράγει το τυχαίο δείγμα $X = \{X_1, X_2, \dots, X_n\}$

Βήμα 1. Αναδειγματοληψία. Έστω τυχαίο δείγμα μεγέθους n, το οποίο έχει ληφθεί με επανάθεση από το X.

Βήμα 2. Υπολογισμός της bootstrap εκτίμησης. Υπολογισμός της μέσης τιμής όλων των τιμών στο X^*.

Βήμα 3. Επανάληψη. Τα βήματα 1 και 2 επαναλαμβάνονται αρκετές φορές ώστε να αποκτηθεί ένα σύνολο B bootstrap εκτιμητριών $\hat{\mu}_1^*, \dots, \hat{\mu}_B^*$.

Βήμα 4. Η προσέγγιση της κατανομής του $\hat{\mu}$. Ταξινομούνται οι bootstrap εκτιμήσεις σε αύξουσα σειρά ώστε να αποκτηθούν τα $\hat{\mu}_{(1)}^* \leq \hat{\mu}_{(2)}^* \leq \cdots \leq \hat{\mu}_{(B)}^*$, όπου $\hat{\mu}_{(i)}^*$ είναι η ιοστή μικρότερη τιμή των $\hat{\mu}_1^*, \dots, \hat{\mu}_B^*$.

Βήμα 5. Διάστημα Εμπιστοσύνης. Το ζητούμενο 100(1-α)% Bootstrap διάστημα εμπιστοσύνης είναι $(\hat{\mu}_{(q_1)}^*, \hat{\mu}_{(q_2)}^*)$, $όπου\ q_1 = \lfloor Ba/2 \rfloor$ αποτελεί το ακέραιο μέρος του Bα/2 και $q_2 = B - q_1 + 1$.

4.1 Διάφορες παραλλαγές του Bootstrap

Έχουν αναπτυχθεί διάφορες παραλλαγές του Bootstrap. Σύμφωνα με Michael R. Chernick, Robert A. LaBudde (2011) και Michael R. Chernick (2008) μερικές χαρακτηριστικές περιπτώσεις είναι:

Η Μπεϋζιανή Bootstrap Μέθοδος

Έστω το δείγμα των n ανεξάρτητων και ισόνομων $x_1, \dots, x_n$ των μεταβλητών $X_1, \dots, X_n$ το καθένα με κατανομή F και έστω η εμπειρική κατανομή $\hat{F}$. Η μη παραμετρική μέθοδος Bootstrap, όπως αναφέρθηκε πραγματοποιεί δειγματοληψία με

επανάθεση από την $\hat{F}$. Έστω θ η παράμετρος της κατανομής F. Έστω $\hat{\theta}$ εκτιμήτρια του θ από το παραπάνω δείγμα. Το μη παραμετρικό Bootstrap μπορεί να χρησιμοποιηθεί για την προσέγγιση της κατανομής του $\hat{\theta}$.

Αντί να γίνεται η δειγματοληψία με επανάθεση για κάθε x_i και ίση πιθανότητα 1/n, η Μπεϋζιανή μέθοδος χρησιμοποιεί μια εκ των υστέρων κατανομή πιθανότητας για τα X_i. Αυτή η εκ των υστέρων κατανομή πιθανότητας είναι επικεντρωμένη στο 1/n για κάθε X_i, αλλά η πιθανότητα αλλάζει από δείγμα σε δείγμα. Δηλαδή, έστω ένα n-διάστατο διάνυσμα με μέση τιμή που κατανέμει ίσο βάρος 1/n σε κάθε X_i αλλά το πραγματικό ποσοστό για κάθε X_i στην κ-οστή λήψη λαμβάνεται από την εκ των υστέρων κατανομή.

Πιο συγκεκριμένα η μέθοδος ορίζεται ως εξής: Λαμβάνονται n-1 ομοιόμορφες τυχαίες μεταβλητές από το διάστημα [0,1]. Ορίζοντας τις ταξινομημένες τιμές από ελάχιστη σε μέγιστη ως $u_1, \dots, u_{n-1}$ με $u_0 = 0 \; και \; u_n = 1$. Έστω $g_i = u_i - u_{i-1}, i = 1,2, \dots n$, το οποίο ορίζεται ως εξής:

$$\sum_{i=1}^{n} g_i = 1$$

Τα g_i αποτελούν τα ομοιογενή διάκενα, χρησιμοποιούνται ώστε να προσδίδουν πιθανότητες στο Μπεϋζιανό Bootstrap δείγμα και η θεωρία κατανομής τους παρουσιάζεται στο έργο του David H. A. (1981). Πιο συγκεκριμένα n παρατηρήσεις επιλέγονται με δειγματοληψία με επανάθεση από το $x_1, \dots, x_n$ αλλά αντί κάθε x_i να έχει ακριβώς 1/n πιθανότητα να επιλεχθεί κάθε φορά, το x_1 επιλέγεται με πιθανότητα g_1, το x_2 με πιθανότητα g_2 και ούτω καθ'εξής. Μια δεύτερη Μπεϋζιανή Bootstrap επανάληψη κατασκευάζεται παρόμοια αλλά με ένα νέο σύνολο n-1 ομοιόμορφων τυχαίων αριθμών και οπότε ένα νέο σύνολο g_i. Σύμφωνα με Rubin, D. B. (1981) The Bayesian bootstrap, η bootstrap μέθοδος και η αντίστοιχη Μπεϋζιανή είναι παρόμοιες μέθοδοι με κοινές ιδιότητες. Για αυτό το λόγο θεωρεί ότι όποιοι περιορισμοί υφίστανται στη Μπεϋζιανή μέθοδο αντιστοιχούν και στην μη παραμετρική περίπτωση. Ένα πλεονέκτημα της Μπεϋζιανής μεθόδου είναι ότι επιτρέπει Μπεϋζιανά τύπου συμπεράσματα σχετικά με την παράμετρο θ με βάση την εκτιμώμενη εκ των υστέρων κατανομή του.

Έστω n Μπευζιανές bootstrap επαναλήψεις και έστω ${g_i}^{(k)}$ είναι η πιθανότητα του X_i στην κ-οστή Μπευζιανή bootstrap επανάληψη.
Από τον David , H. A. (1981) ισχύει ότι:

$E\left[{g_i}^{(k)}\right] = 1/n$ για κάθε $i \; και \; k$.

$$var\left({g_i}^{(k)}\right) = \frac{(n-1)}{n^3} \; και \; cov\left({g_i}^{(k)}, {g_j}^{(k)}\right) = -\frac{1}{(n-1)} \; για \; κάθε \; k$$

με E[.], var(.) και cov(.) να αποτελούν την μέση τιμή, την διασπορά και την συνδιακύμανση. Λόγω αυτών των ιδιοτήτων η bootstrap κατανομή του $\hat{\theta}$ είναι παρόμοια με την Μπεϋζιανή bootstrap εκ των υστέρων κατανομή του θ. Σύμφωνα με τον Rubin, D. B. (1981) παρουσιάζονται κάποια παραδείγματα που δείχνουν ότι η

Μπεϋζιανή bootstrap διαδικασία καταλήγει σε μια εκ των υστέρων κατανομή, η οποία είναι η Dirichlet και η οποία βασίζεται σε μια συζυγή εκ των προτέρων Dirichlet κατανομή (conjucate prior).

Επειδή το Μπεϋζιανό bootstrap είναι κατάλληλο για συγκεκριμένα προβλήματα, ο Rubin πιστεύει ότι η μέθοδος είναι αρκετά περιοριστική, καθιστώντας την να μην προτιμάται ως εργαλείο για γενική συμπερασματολογία. Αυτό ισχύει όμως σε κάποιες περιπτώσεις και όχι γενικά. Εάν υπάρχει λόγος αμφισβήτησης της μεθόδου σε κάποιες περιπτώσεις, αντίστοιχα και η μη παραμετρική bootstrap που είναι τόσο παρόμοια θα έπρεπε και αυτή να αμφισβητηθεί. Η κριτική αυτή βασίζεται στην έλειψη ομαλότητας της εμπειρικής κατανομής. Οπότε ομαλοποιώντας την bootstrap το πρόβλημα αυτό μπορεί να διορθωθεί.
Η Μπεϋζιανή bootstrap μέθοδος έχει γενικοποιηθεί ώστε να επιδέχεται εκ των υστέρων κατανομές που δεν ανήκουν στην Dirichlet.

Ομαλοποιημένη Bootstrap Μέθοδος (The Smoothed Bootstrap Method).

Ένας λόγος επιλογής της μη παραμετρικής μεθόδου είναι ότι η $\hat{F}$ εμπειρική κατανομή είναι η εκτιμήτρια μέγιστης πιθανοφάνειας της F, χωρίς να έχουν γίνει υποθέσεις για την F. Για το λόγο αυτο οι bootstrap εκτιμήτριες των παραμέτρων της F μπορούν να θεωρηθούν ως μη παραμετρικές εκτιμήτριες μέγιστης πιθανοφάνειας. Όμως σε αρκετές εφαρμογές θεωρείται ότι η κατανομή είναι απολύτως συνεχής. Σε αυτή την περίπτωση μπορεί να υπολογιστεί μια εκτιμήτρια πυκνότητας ή μια «ομαλή εκδοχή» της εκτιμήτριας της F. Ένας τρόπος κατασκευής της εκτιμήτριας αυτής πραγματοποιείται με την χρήση μεθόδων εξομάλυνσης πυρήνων (kernel smoothing methods). Ένα χαρακτηριστικό παράδειγμα αποτελεί η αντικατάσταση της $\hat{F}$ με μια ομαλή κατανομή η οποία για παράδειγμα μπορεί να βασίζεται σε μια εκτιμήτρια πυκνότητας πυρήνα της $F^{'}$, η οποία είναι η παράγωγος της F ως προς x στην περίπτωση της μονομεταβλητής κατανομής F.

Ο Efron (1982) παρουσίασε την χρήση της ομαλής εκδοχής του bootstrap για τον συντελεστή συσχέτισης χρησιμοποιώντας Γκαουσιανές και ομοιόμορφες συναρτήσεις πυρήνων. Οι παρατηρήσεις στην προσομοίωση του κατά την διάρκεια της έρευνάς του ήταν Γκαουσιανές και τα αποτελέσματα έδειξαν ότι το smoothed bootstrap είναι λίγο καλύτερο από το αρχικό μη παραμετρικό bootstrap στην εκτίμηση του τυπικού σφάλματος του συντελεστή συσχέτισης.

Από την αρχή της ανάπτυξης του bootstrap είχε θεωρηθεί η χρήση των ομαλών εκδοχών του F. Πρόσφατα ο Dudewicz και άλλοι έχουν κάνει εκτενείς συγκρίσεις και με την μη ομαλή εκδοχή. Ο Dudewicz (1992) πρότεινε μια Monte Carlo προσέγγιση η οποία βασίζεται στην δειγματοληψία από μια εκτιμήτρια πυρήνα ή από μια παραμετρική εκτίμηση του F. Η μέθοδος αυτή αποτελεί την γενικευμένη bootstrap μέθοδο (generalized bootstrap method) και επιτρέπει ευρύτερη επιλογή εκτιμητριών χωρίς να παραβαίνει τις παραδοχές σχετικά με την κατανομή. Αντί να χρησιμοποιηθούν οι μέθοδοι πυρήνων ώστε να εκτιμηθεί η πυκνότητα, ο Dudewicz πρότεινε την χρήση μιας μεγαλύτερης κλάσης κατανομών για την προσαρμογή των παρατηρούμων δεδομένων. Η προσαρμοσμένη αυτή κατανομή χρησιμοποιείται για την λήψη των bootstrap δειγμάτων.

Μια τέτοια οικογένεια κατανομών αποτελεί και η γενικευμένη λάμδα κατανομή. Αυτή η κατανομή είναι μια οικογένεια κατανομών τεσσάρων παραμέτρων που μπορεί να προσδιοριστεί από την μέση τιμή, την διασπορά, την ασυμμετρία και την κυρτότητα. Η μέθοδος των ροπών εκτιμητριών για τις τέσσερις παραμέτρους είναι μια από τις μεθόδους που μπορεί να χρησιμοποιηθεί για την προσαρμογή της κατανομής. Σε μια συγκεκριμένη εφαρμογή οι Sun and Muller-Schwarze (1996) συνέκριναν την γενικευμένη bootstrap μέθοδο, όπως περιγράφεται παραπάνω, με την μη παραμετρική bootstrap μέθοδο.

Η σύγκριση έδειξε οτι η γενικευμένη μέθοδος ίσως είναι μια αποτελεσματική εναλλακτική μέθοδος σε σύγκριση με την μη παραμετρική, αφού έχει το πλεονέκτημα τα δεδομένα να λαμβάνονται από μια συνεχή κατανομή, ενώ παράλληλα δεν περιορίζεται όπως το smoothed bootstrap.

Παραμετρική μέθοδος.

Ο Efron (1982) θεώρησε την αρχική bootstrap μέθοδο ως μια μη παραμετρική προσέγγιση μέγιστης πιθανοφάνειας. Σε αυτή την περίπτωση μπορεί να θεωρηθεί ως μια επέκταση της προσέγγισης της μέγιστης πιθανοφάνειας του Fisher σε ένα μη παραμετρικό πλαίσιο. Επίσης αν περαιτέρω υποτεθεί ότι η κατανομή F για τα δείγματα είναι απολύτως συνεχής τότε μπορεί να ομαλοποιηθεί η F. Επίσης αν υποτεθεί ότι η F προέρχεται από μια παραμετρική οικογένεια, για παράδειγμα την Γκαουσιανή κατανομή, τότε η θεωρία μέγιστης πιθανοφάνειας του Fisher θα είναι εφαρμόσιμη για την εκτίμηση ενός μικρού αριθμού παραμέτρων. Δηλαδή η κατάλληλη εκτιμήτρια για την F θα ήταν μια Γκαουσιανή κατανομή με τις εκτιμήτριες μέγιστης πιθανοφάνειας των μ και σ^2 να χρησιμοποιούνται για τις άγνωστες παραμέτρους.

Οπότε αν μια εκτιμήτρια της F επιλέγεται από μια παραμετρική οικογένεια, τότε η δειγματοληψία με επανάθεση από αυτή την κατανομή μπορεί να θεωρηθεί ως η παραμετρική bootstrap μέθοδος. Αν χρησιμοποιείται η μέθοδος μέγιστης πιθανοφάνειας για την εκτίμηση των παραμέτρων της κατανομής F τότε η προσέγγιση είναι ουσιαστικά ίδια με την μέγιστη πιθανοφάνεια. Οπότε η bootstrap μέθοδος συνήθως δεν προσφέρει κάτι παραπάνω στα παραμετρικά προβλήματα. Παρ'όλα αυτά σε πολύπλοκα προβλήματα συνήθως θα πρέπει να υπάρχει μερική παραμετροποίηση.

Οι Davison and Hinkley (1997) παρείχαν και μια άλλη εξήγηση για το παραμετρικό bootstrap. Μέσω μιας σύγκρισης μεταξύ της παραμετρικής και της μη παραμετρικής bootstrap μεθόδου μπορούν να ελεγχθούν οι παραμετρικές υποθέσεις. Χρησιμοποιώντας την εκθετική κατανομή ώστε να επεξηγήσουν την παραμετρική μέθοδο bootstrap, επισήμαναν την σημασία της μεθόδου όταν η παραμετρική κατανομή είναι δύσκολο να εξαχθεί ή έχει μια ασυμπτωτική προσέγγιση που δεν είναι ακριβής σε δείγματα μικρού μεγέθους.

Double Bootstrap

Το double bootstrap είναι μια ιδιαίτερη περίπτωση της bootstrap μεθόδου. Επειδή απαιτεί μεγάλη υπολογιστική ισχύ, συχνά αποφεύγεται η χρήση της. Παρ'όλα αυτά μερικές φορές τα πλεονεκτήματα είναι αρκετά ώστε να επιτρέπεται η χρήση της. Ο

Efron (1983) χρησιμοποίησε το double bootstrap ως μια μέθοδο ώστε να προσαρμόσει την μεροληψία του ρυθμού σφαλμάτων στο πρόβλημα ταξινόμησης (classification problem).

Η χρήση της brute force προσέγγισης, μιας τεχνικής επίλυσης προβλημάτων στο double bootstrap απαιτεί την δημιουργία B^2 bootstrap δειγμάτων, με B να είναι ο αριθμός των bootstrap δειγμάτων που λαμβάνονται από το αρχικό δείγμα και ο αριθμός των bootstrap δειγμάτων που αποκτώνται για κάθε προσαρμογή στην εκτιμήτρια από το αρχικό bootstrap δείγμα. Η brute force μέθοδος ουσιαστικά πραγματοποιεί την συστημική απαρίθμηση όλων των πιθανών λύσεων ενός προβλήματος ελέγχοντας αν κάθε μια από αυτές τις λύσεις ικανοποιεί το πρόβλημα. Στην προκειμένη περίπτωση, αν ο αριθμός των αρχικών bootstrap δειγμάτων είναι B_1και κάθε δείγμα προσαρμόζεται χρησιμοποιώντας B_2 bootstrap δείγματα, τότε ο συνολικός αριθμός των bootstrap δειγμάτων που θα παραχθούν είναι B_1B_2. Αλλά ο Efron (1983) παρείχε μια τεχνική μείωσης της διασποράς που επιτρέπει την ίδια ακρίβεια όπως η brute force μέθοδος, όπου αν B^2 τα bootstrap δείγματα που απαιτούνται για την εύρεση της ακρίβειας, να χρησιμοποιούνται στην πραγματικότητα μόνο $2B$ bootstrap δείγματα.

Η επανάληψη της bootstrap μεθόδου (bootstrap iteration) βελτιώνει την τάξη της ακρίβειας των bootstrap διαστημάτων εμπιστοσύνης, όπως τονίζεται και στο έργο των Hall, Beran, Martin και ιδιαίτερα από τον Hall (1992). Παρ'όλα αυτά επισημαίνεται ότι η βελτίωση της ακρίβειας έχει ως αποτέλεσμα την αύξηση του χρόνου υπολογισμού.

The M-out-of-N Bootstrap

Όταν η μη παραμετρική μέθοδος Bootstrap παρουσιάστηκε αρχικά, ο Efron πρότεινε την επιλογή του μεγέθους δείγματος του bootstrap δείγματος να είναι ίδιο με το μέγεθος δείγματος n από το αρχικό δείγμα. Αυτή η μέθοδος είχε αρκετές εφαρμογές με καλά αποτελέσματα. Αν και υπάρχει η αντίληψη ότι αφού η κεντρική θεωρία του δείγματος bootstrap στηρίζεται στην μίμηση της συμπεριφοράς του αρχικού δείγματος και η ακρίβεια των εκτιμητριών του δείγματος γενικώς στηρίζονται στο μέγεθος δείγματος n, επιλέγοντας ένα μέγεθος δείγματος m μεγαλύτερο από το n θα οδηγούσε σε μια εκτιμήτρια με λιγότερη μεταβλητότητα από ότι το αρχικό δείγμα. Ενώ ένα μέγεθος δείγματος m μικρότερο από το n θα αύξανε την διασπορά της εκτιμήτριας.

Παρ'όλα αυτά, αρχικά υποστηρίχθηκε η έρευνα για την περίπτωση $m < n$. Αυτό έγινε κυρίως για την δειγματοληψία από ένα πεπερασμένο πληθυσμό ή στην περίπτωση εξαρτήσεων-αλληλοσχετίσεων (dependencies) όπως οι χρονοσειρές ή τα χωρικά δεδομένα. Η ανεξάρτητη δειγματοληψία με επανάθεση στις εξαρτώμενες περιπτώσεις δημιουργεί εκτιμήτριες με διασπορά μικρότερη από αυτή του αρχικού συνόλου δεδομένων, αφού n εξαρτώμενες παρατηρήσεις περιέχουν μόνο πληροφορία ισοδύναμη σε ένα μικρότερο αριθμό ανεξάρτητων παρατηρήσεων. Οπότε αν η εφαρμογή της μεθόδου bootstrap συμπεριφέρεται με μικρότερη εξάρτηση από ότι το αρχικό δείγμα, ένας μικρότερος αριθμός παρατηρήσεων θα μιμούνταν την διασπορά της εκτιμήτριας καλύτερα.

Η m-out-of-n bootstrap μέθοδος με $m \ll n$ ελαττώνει τα προβλήματα ασυνέπειας σχετιζόμενα με την δειγματοληπτική έρευνα. Επίσης εφαρμόζεται και σε ανεξάρτητες περιπτώσεις, όπως όταν η μέση τιμή συγκλίνει σε μια σταθερή κατανομή εκτός της κανονικής ή όταν εκτιμάται η ασυμπτωτική συμπεριφορά του μεγίστου μιας ανεξάρτητης και ισόνομης ακολουθίας τυχαίων μεταβλητών. Αυτή η μέθοδος έχει αναλυθεί από αρκετούς ερευνητές και έχει αναπτυχθεί κατάλληλη ασυμπτωτική θεωρία που δείχνει ότι αν $m \to \infty \; και \; n \to \infty$ με ένα ρυθμό ώστε m/n→ 0, τότε η m-out-of-n bootstrap μέθοδος είναι συνεπής στις περιπτώσεις που η n-out-of-n δεν είναι.

Πρακτικά η μέθοδος αυτή θα μπορούσε να χρησιμοποιηθεί σε περιπτώσεις που η κανονική bootstrap μέθοδος είναι γνωστό ότι δεν είναι συνεπής. Αλλιώς τα διαστήματα εμπιστοσύνης θα ήταν ευρύτερα.

Το Wild Bootstrap

Παρουσιάζεται η περιγραφή του wild bootstrap και της εφαρμογής του στην γραμμική παλινδρόμηση με ετεροσκεδαστικούς ορους σφάλματος. Έστω το γραμμικό μοντέλο $Y = X\beta + u$ και σύμφωνα με το μοντέλο αυτό, οι παρατηρούμενες τιμές

$$y_i = X_i\beta + u_i$$

όπου τα u_i είναι οι όροι σφάλματος ανεξάρτητοι μεταξύ τους, τα X_i είναι οι τιμές των k συμμεταβλητές και β είναι ένα διάνυσμα k παραμέτρων που αντιστοιχεί στις ανάλογες συμμεταβλητές. Έστω $\hat{\beta}$ ο εκτιμητής ελαχίστων τετραγώνων του β και $\widehat{u_i}$ τα υπόλοιπα των ελαχίστων τετραγώνων. Κατασκευάζεται ένα bootstrap δείγμα από τα υπόλοιπα, χρησιμοποιώντας το μοντέλο με τις εκτιμήτριες ελαχίστων τετραγώνων του β ώστε να κατασκευαστεί το bootstrap δείγμα των παρατηρήσεων και στην συνέχεια χρησιμοποιείται η μέθοδος των ελαχίστων τετραγώνων στις bootstrap παρατηρήσεις ώστε να αποκτηθεί η bootstrap εκτιμήτρια του β. Αυτή η διαδικασία επαναλαμβάνεται B φορές ώστε να παραχθεί μια προσέγγιση στην bootstrap κατανομή του $\hat{\beta}$. Από αυτή την κατανομή, μπορεί να αποκτηθεί ένας εκτιμώμενος πίνακας συνδιασποράς για τις παραμέτρους από τον οποίο λαμβάνονται οι εκτιμήτριες των τυπικών σφαλμάτων για τις παραμέτρους παλινδρόμησης.

Το wild bootstrap πραγματοποιεί παρόμοια διαδικασία εκτός όμως ότι τα υπόλοιπα ελαχίστων τετραγώνων είναι τροποποιημένα και σε αυτά πραγματοποιείται η bootstrap μέθοδος. Πιο συγκεκριμένα τα υπόλοιπα του wild bootstrap αποκτώνται ως εξής:

$$u_i^* = h_i(\hat{u}_i)\varepsilon_i$$

όπου το ε_i έχει μέση τιμή 0 και διασπορά 1 και το h_i αποτελεί ένα μετασχηματισμό. Στο ετεροσκεδαστικό μοντέλο προτιμάται ο τύπος:

$$h_i(\hat{u}_i) = \frac{\hat{u}_i}{(1-H_i)^{\frac{1}{2}}} \text{ ή } \frac{\hat{u}_i}{(1-H_i)}$$

όπου $H_i = X_i(X^TX)^{-1}X_i^T$ το ιοστό διαγώνιο στοιχείο του ορθογωνικού πίνακα προβολής. Οπότε το wild bootstrap στηρίζεται στην εφαρμογή διαφορετικών επιλογών για το h_i και την κατανομή των ε_i.

4.2 Μέθοδος Jackknife

Ο Quenouille (1949) ανέπτυξε την μέθοδο jackknife ως μια μέθοδο για την εκτίμηση της μεροληψίας μιας εκτιμήτριας διαγράφοντας ένα δεδομένο κάθε φορά από το αρχικό σύνολο δεδομένων και υπολογίζοντας κάθε φορά την εκτιμήτρια βασιζόμενος στα υπόλοιπα δεδομένα. Ο Tukey (1958) επισήμανε ότι η μέθοδος μπορούσε να προσαρμοστεί για την εκτίμηση διασπορών χρησιμοποιώντας ψευδοτιμές, εν αντιθέσει με την μέθοδο bootstrap που χρησιμοποιεί τα bootstrap δείγματα για την εκτίμηση της διασποράς.

Σύμφωνα με τους Jun Shao και Dongsheng Tu (1995) έστω $T_n = T(X_1, X_2, \dots, X_n)$ μια εκτιμήτρια μιας άγνωστης παραμέτρου θ. Η μεροληψία του T_n ορίζεται ως:

$$bias(T_n) = E(T_n) - \theta.$$

Έστω $T_{n-1,i} = T_{n-1}(X_1, \dots, X_{i-1}, X_{i+1} \dots, X_n)$ το στατιστικό βασιζόμενο σε n-1 παρατηρήσεις και του οποίου η ι-οστή παρατήρηση έχει αφαιρεθεί. Η εκτιμήτρια μεροληψίας του jackknife είναι:

$$b_{jack} = (n-1)(\overline{T}_n - T_n)$$

όπου

$$\overline{T}_n = \frac{1}{n}\sum_i T_{n-1,i}$$

Η διόρθωση μεροληψίας είναι:

$$T_{jack} = T_n - b_{jack} = nT_n - (n-1)\overline{T}_n$$

Για τις jackknife εκτιμήτριες ισχύει η σχέση:

$$bias(T_n) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right), \text{για τυχαία } a, b$$

Επειδή τα $T_{n-1,i}, i = 1, \dots, n$ είναι ισόνομα θα ισχύει:

$$b(T_{n-1,i}) = \frac{a}{n-1} + \frac{b}{(n-1)^2} + O(\frac{1}{(n-1)^3})$$

Την παραπάνω εξίσωση θα ικανοποιεί και η bias($\overline{T}_n$). Επομένως:

$$\begin{aligned} E(b_{jack}) &= (n-1)[bias(\overline{T}_n) - bias(T_n)] \\ &= (n-1)\left[\left(\frac{1}{n-1} - \frac{1}{n}\right)a + \left(\frac{1}{(n-1)^2} - \frac{1}{n^2}\right)b + O\left(\frac{1}{n^3}\right)\right] \\ &= \frac{a}{n} + \frac{(2n-1)b}{n^2(n-1)} + O\left(\frac{1}{n^2}\right) = bias(T_n) + O(\frac{1}{n^2}) \end{aligned}$$

το οποίο δείχνει ότι το b_{jack} είναι μια αμερόληπτη εκτιμήτρια της μεροληψίας του T_n με τάξη O(n^{-2}). Άρα:

$$bias(T_{jack}) = bias(T_n) - E(b_{jack}) = -\frac{b}{n(n-1)} + O\left(\frac{1}{n^2}\right)$$

Δηλαδή η μεροληψία του T_{jack} είναι τάξης n^{-2}. Η jackknife παράγει μια εκτιμήτρια με διόρθωση μεροληψίας (bias-reduced estimator) αφαιρώντας τον όρο πρώτης τάξης από το $bias(T_n)$.

Η μέθοδος jackknife έγινε ένα ακόμα πιο πολύτιμο εργαλείο αφού σύμφωνα με τον Tukey (1958) μπορεί να χρησιμοποιηθεί για την κατασκευή εκτιμητριών διασποράς. Το T_{jack} μπορεί να εκφραστεί ως:

$$T_{jack} = \frac{1}{n}\sum_{i=1}^{n}[n\,T_n - (n-1)T_{n-1,i}]$$

Ο Tukey όρισε την παρακάτω εξίσωση

$$\tilde{T}_{n,i} = nT_n - (n-1)T_{n-1,i} \quad i = 1, \dots, n$$

ως τις jackknife ψευδοτιμές και είκασε ότι:

- Οι ψευδοτιμές $\tilde{T}_{n,i}\,, i = 1, \dots, n$ μπορούν να θεωρηθούν ως ανεξάρτητες και ισόνομες.
- Το $\tilde{T}_{n,i}$ έχει περίπου την ίδια διασπορά με το $\sqrt{n}T_n$

Σύμφωνα με τα παραπάνω, εκτιμάται η $var(\sqrt{n}T_n)$ από την δειγματική διασπορά που βασίζεται στο $\tilde{T}_{n,1}, \dots, \tilde{T}_{n,n}$ ώστε να εκτιμηθεί η διασπορά $var(T_n)$ από:

$$\begin{aligned} v_{jack} &= \frac{1}{n(n-1)}\sum\nolimits_{i=1}^{n}\left(\tilde{T}_{n,i} - \frac{1}{n}\sum\nolimits_{j=1}^{n}\tilde{T}_{n,j}\right)^2 \\ &= \frac{n-1}{n}\sum\nolimits_{i=1}^{n}\left(T_{n-1,i} - \frac{1}{n}\sum\nolimits_{j=1}^{n}T_{n-1,j}\right)^2 \end{aligned}$$

Αυτή η εκτιμήτρια αποτελεί την (delete-1) jackknife εκτιμήτρια διασποράς για το T_n.

Παράδειγμα

Έστω ότι ισχύει

$$T_n = \bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

από τους παραπάνω τύπους θα ισχύουν τα ακόλουθα:

$$T_{n-1,i} = \bar{X}_{n-1,i} = \frac{(n\bar{X}_n - X_i)}{n-1}$$

Και $\bar{T}_n = \bar{X}_n,\ b_{jack} = 0\ και\ T_{jack} = T_n = \bar{X}_n$. Επίσης από το παρακάτω τύπο ισχύει:

$$\tilde{T}_{n,i} = nT_n - (n-1)T_{n-1,i} = n\bar{X}_n - (n-1)\frac{(n\bar{X}_n - X_i)}{n-1}$$
$$= n\bar{X}_n - (n\bar{X}_n - X_i) = X_i$$

Και διασπορά ίση με $var(\sqrt{n}\bar{X}_n)$. Οπότε οι υποθέσεις του Tukey ισχύουν. Σε αυτή την περίπτωση ισχύει:

$$v_{jack} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n(n-1)}$$

Αν και το jackknife είναι ένα σημαντικό εργαλείο για την εκτίμηση για παράδειγμα της διασποράς εκτιμητριών, προσομοιώσεις του Efron έχουν δείξει ότι το bootstrap υπολογίζει καλύτερα την εκτιμήτρια της τυπικής απόκλισης και ότι η jackknife εκτιμήτρια του τυπικού σφάλματος είναι μια bootstrap εκτιμήτρια με αντικατάσταση του $\hat{\theta}$ από μια γραμμική προσέγγιση. Αυτό έχει σαν αποτέλεσμα η jackknife να θεωρείται ως προσέγγιση της bootstrap. Επίσης, το γεγονός ότι η jackknife κάνει υπολογισμούς μόνο για n jackknife σύνολα δεδομένων σημαίνει ότι δεν είναι τόσο αποδοτική όσο η bootstrap αφού έχει περιορισμένη πληροφόρηση για το στατιστικό $\hat{\theta}$.

ΚΕΦΑΛΑΙΟ 5

Η συμπεριφορά των μεθόδων Bootstrap και Jackknife στην επιλογή μεταβλητών

5.1 Προσομοιώσεις

Στο κεφάλαιο αυτό εξετάζεται η συμπεριφορά των μεθόδων Bootstrap και Jackknife σε συνδυασμό με τα κριτήρια πληροφορίας AIC και BIC για την επιλογή μοντέλων, μέσω προσομοιώσεων. Σκοπός μας είναι η βελτίωση της απόδοσης των κριτηρίων AIC και BIC όταν αυτά συνδυαστούν με τις μεθόδους Bootstrap και Jackknife. Οι μέθοδοι και τα κριτήρια που χρησιμοποιούνται σε αυτό το κεφάλαιο βασίζονται στα προηγούμενα κεφάλαια στα οποία γίνεται η θεωρητική ανάλυση τους. Η απόδοση των AIC και BIC σε συνδυασμό με τις δύο μεθόδους συγκρίνεται χρησιμοποιώντας το παρακάτω γραμμικό παλινδρομικό μοντέλο για τέσσερεις ομάδες προσομοιώσεων

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + e_i \, , i = 1, \dots , n \qquad (5.1.1)$$

όπου y_i η μεταβλητή απόκρισης, τα $b_j \, για \, j = 1, \dots , 4$ είναι οι συντελεστές παλινδρόμησης και τα e_i όροι σφάλματος που ακολουθούν την $N(0,1)$ κατανομή. Οι ανεξάρτητες μεταβλητές x_j για $j = 1, \dots , 4$ ακολουθούν την πολυμεταβλητή κανονική κατανομή με μέση τιμή 0 και διασπορά 1 η κάθε μια και συντελεστή συσχετίσεως ρ. Για τις προσομοιώσεις τα $(b_0, b_1, b_2, b_3, b_4)$ έχουν την τιμή (1,0.8,0,0,1). Οπότε η πρώτη και τέταρτη συμμεταβλητή είναι σημαντικές και πρέπει τα κριτήρια πληροφορίας να τις επιλέγουν πιο συχνά. Οι προσομοιώσεις που παρουσιάζονται σε αυτό το κεφάλαιο συγκρίνουν τις διαφορετικές μεθόδους για διαφορετικά μεγέθη δείγματος και για διαφορετικό συντελεστή συσχέτισης ρ,

υπολογίζοντας αφ' ενός μεν το ποσοστό των φορών που επιλέγεται το σωστό μοντέλο, και αφ' ετέρου το ποσοστό των φορών που επιλέγεται μια τουλάχιστον μη σημαντική μεταβλητή, καθώς και το ποσοστό των φορών που δεν επιλέγεται μια σημαντική μεταβλητή.

Προσομοίωση 1

Στην πρώτη ομάδα προσομοιώσεων δημιουργούνται 100 αρχικά δείγματα μεγέθους *sample*1 με τις εντολές:

$$for(i\ in\ 1{:}\ iter)\{out\ <-\ mvrnorm(n = sample1, mu\ =\ c(0{,}0{,}0{,}0), Sigma\ =\ matrix(c(1{,}0.5{,}0.25{,}0.125{,}0.5{,}1{,}0.5{,}0.25{,}0.25{,}0.5{,}1{,}0.5{,}0.125{,}0.25{,}0.5{,}1),4{,}4),\ empirical\ =\ FALSE)$$

Με την εντολή $mvrnorm$ κατασκευάζονται οι ανεξάρτητες συμμεταβλητές των δειγμάτων από την πολυμεταβλητή κανονική κατανομή. Το mu είναι διάνυσμα μεγέθους 4 και αποτελεί τις μέσες τιμές των ανεξάρτητων συμμεταβλητών ενώ το $Sigma$ αποτελεί τον πίνακα διασπορών-συνδιασπορών τους. Ο πίνακας αυτός κατασκευάστηκε θεωρώντας ότι ο τύπος της συσχέτισης μεταξύ των i και j συμμεταβλητών είναι ίσος με $rho^{|i-j|}$ και το $rho = 0.5$. Οι μεταβλητές e_i όπως είπαμε κατασκευάζονται από την N(0,1) κατανομή και οι μεταβλητές y_i από το μοντέλο (5.1.1).

Οι συγκρίσεις των δύο μεθόδων και των δύο κριτηρίων γίνονται για διαφορετικά μεγέθη δείγματος έτσι ώστε να διαπιστωθεί η συμπεριφορά τους για μικρό δείγμα και για μεγάλο δείγμα. Συγκεκριμένα, εξετάζονται οι περιπτώσεις που το δείγμα είναι ίσο με 20, 50, 100 και 200. Για αυτές τις περιπτώσεις χρησιμοποιείται η μέθοδος bootstrap και η μέθοδος jackknife σε συνδυασμό με τα κριτήρια AIC και BIC καθώς και τα κριτήρια AIC και BIC ανεξάρτητα από τις μεθόδους. Παρακάτω παρατίθενται κομμάτια του κώδικα που χρησιμοποιήθηκαν.

Στο bootstrap έχοντας ως γνωστά τα αρχικά δείγματα η διαδικασία που ακολουθήθηκε ήταν:

$$q2 < -q21[sample(sample1, replace = T),]$$

$$p1 < -lm(y1 \sim f1 + f2 + f3 + f4, data = q23)$$

$$q3 < -stepAIC(p1, data = q23, k = 2, direction\ =\ c(\text{"}backward\text{"}),\ trace = FALSE)$$

$$q4 < -stepAIC(p1, data = q23, k = \log(sample1),\ direction\ =\ c(\text{"}backward\text{"}), trace = FALSE)$$

Για κάθε ένα από τα αρχικά δείγματα, πρώτα εκτελείται η αναδειγματοληψία, όπως προβλέπει η μέθοδος bootstrap, κατασκευάζοντας με την εντολή sample ένα καινούργιο bootstrap δείγμα και μετά προσαρμόζεται σε αυτό το γραμμικό μοντέλο.

Η $q3$ εντολή εκτελεί την βηματική επιλογή μοντέλων (stepwise model selection) με βάση το κριτήριο AIC (για $k = 2$) ενώ το $q4$ αντίστοιχα με βάση το κριτήριο BIC ($k = \log(sample1)$).

Στην συνέχεια χρησιμοποιείται η εντολή summary για την εύρεση της ποιότητας της προσαρμογής. Από το summary καταγράφονται οι συντελεστές των συμμεταβλητών που έχουν επιλεγεί από τα δύο κριτήρια για το bootstrap δείγμα. Η διαδικασία αυτή επαναλαμβάνεται b_iter φορές, δηλαδή κατασκευάζουμε στο σύνολο b_iter bootstrap δείγματα για κάθε αρχικό δείγμα. Στη συνέχεια μπορούμε να βρούμε πόσες φορές έχει επιλεχθεί η κάθε συμμεταβλητή καθώς και πόσες φορές είναι στατιστικά σημαντικός ο αντίστοιχος συντελεστής. Πιο συγκεκριμένα, cov1 αποτελεί τον αριθμό των φορών που έχει επιλεχθεί από το AIC η μεταβλητή x_1 σε κάθε bootstrap επανάληψη b_iter και sig1 ο αριθμός που δείχνει σε πόσες από τις φορές που έχει επιλεγεί η πρώτη μεταβλητή είναι στατιστικά σημαντική, το οποίο ελέγχεται από το αν η τιμή της p-value είναι μικρότερη από το a=0.05 επίπεδο σημαντικότητας, όταν εκτελείται η summary. Προτείνουμε ένα κατώτατο όριο, έστω 30%, για το οποίο αν ισχύει:

$$F_1 = cov1 * sig1/100 \geq 30,$$

η μεταβλητή x_1 θα συμπεριλαμβάνεται στο τελικό μοντέλο. Το ίδιο κάτω όριο υιοθετείται για όλες τις ανεξάρτητες μεταβλητές. Εχουμε επίσης εξετάσει τη συμπεριφορά της μεθόδου bootstrap για διάφορες τιμές των επαναλήψεων b_iter, δηλαδή για τις τιμές 20, 50, 100, 200 και 500.

Ομοίως στην περίπτωση του jackknife χρησιμοποιώντας το αρχικό δείγμα και αφαιρώντας κάθε φορά μια παρατήρηση από αυτό, δημιουργούμε ένα νέο δείγμα, και για την ακρίβεια δημιουργούμε *sample*1 καινούργια jackknife δείγματα. Το γραμμικό μοντέλο (5.1.1) προσαρμόζεται σε κάθε jackknife δείγμα και στην συνέχεια ακολουθείται η ίδια διαδικασία που περιγράφηκε παραπάνω για την εύρεση του τελικού μοντέλου.

Στο τέλος, υπολογίζονται για το κάθε τελικό μοντέλο από τα δύο κριτήρια πληροφορίας και για τις δύο μεθόδους δειγματοληψίας, το ποσοστό των φορών που επιλέγεται το σωστό μοντέλο (5.1.1) που έχουμε υποθέσει από την αρχή, δηλαδή το ποσοστό των φορών από τις 100 (για τα 100 αρχικά δείγματα αντίστοιχα) που επιλέγεται μόνο η πρώτη και η τέταρτη μεταβλητή. Το ίδιο ποσοστό υπολογίζεται από την εφαρμογή του AIC και του BIC στα αρχικά δείγματα ανεξάρτητα από τις μεθόδους bootstrap και jackknife. Επίσης υπολογίζεται και για τις 6 περιπτώσεις το ποσοστό των φορών που επιλέγεται μια τουλάχιστον μη σημαντική μεταβλητή καθώς και το ποσοστό των φορών που δεν επιλέγεται μια σημαντική μεταβλητή.

Τα αποτελέσματα των προσομοιώσεων, οι οποίες εκτελούνται για set.seed(9), παρουσιάζονται για κάθε τιμή του b_iter και για διάφορα μεγέθη δειγμάτων στους Πίνακες 1 έως 4:

Στους πίνακες παρακάτω καθώς και για τους πίνακες μέχρι και την προσομοίωση 3 αναγράφονται τα εξής αποτελέσματα με την ακόλουθη σειρά:

Το Sample1 αποτελεί το μέγεθος δείγματος των αρχικών δειγμάτων που χρησιμοποιούνται από το πρόγραμμα. Το b_iter είναι ο αριθμός των επαναλήψεων που πραγματοποιεί η bootstrap μέθοδος δεδομένων των αρχικών δειγμάτων. Το COR_AIC είναι το ποσοστό των φορών από τις 100 (δηλαδή από τα 100 αρχικά δείγματα) που έχει το κριτήριο AIC επιλέξει το σωστό μοντέλο, δηλαδή την πρώτη και την τέταρτη μεταβλητή. Ομοίως το COR_BIC αποτελεί το ποσοστό των φορών από τις 100 που έχει το κριτήριο BIC επιλέξει το σωστό μοντέλο.

Το COR_BOOTAIC είναι το ποσοστό των φορών από τα 100 τελικά μοντέλα, που έχει το κριτήριο AIC σε συνδυασμό με τη μέθοδο bootstrap, επιλέξει το σωστό μοντέλο. Για την κατασκευή των τελικών μοντέλων, όπως προαναφέρθηκε χρησιμοποιήθηκε η παραπάνω μεθοδολογία με το κατώτατο όριο ίσο με 30%, για 5 περιπτώσεις, δηλαδή για b_iter=20, 50, 100, 200, 500 bootstrap επαναλήψεις. Το COR_BOOTBIC, ομοίως με την μεθοδολογία του COR_BOOTAIC, είναι το ποσοστό των φορών από τις 100 που έχει επιλέξει το BIC σε συνδυασμό με τη μέθοδο bootstrap το σωστό μοντέλο.

Το COR_JACKAIC αναφέρεται στην δειγματοληψία jackknife. Αποτελεί το ποσοστό των φορών από τις 100 που το κριτήριο AIC σε συνδυασμό με τη μέθοδο jackknife έχει επιλέξει το σωστό μοντέλο. Ομοίως με την μέθοδο bootstrap ακολουθείται η παραπάνω μεθοδολογία για την επιλογή των 100 τελικών μοντέλων από τα οποία θα ελεγχθεί αν έχει επιλεγεί το σωστό μοντέλο. Το COR_JACKBIC παρόμοια υπολογίζει ότι και το COR_JACKAIC όταν χρησιμοποιείται όμως το κριτήριο BIC.

Στην συνέχεια αναγράφονται οι δυο περιπτώσεις σφαλμάτων που υπολογίζονται από το πρόγραμμα. Το σφάλμα ERROR1AIC είναι το ποσοστό των φορών από τις 100 (100 αρχικά δείγματα) που το κριτήριο AIC έχει επιλέξει μοντέλο με μια τουλάχιστον μη σημαντική μεταβλητή. Ομοίως το ERROR1BIC αποτελεί το ποσοστό των φορών από τις 100 που το κριτήριο BIC έχει επιλέξει μοντέλο με μια τουλάχιστον μη σημαντική μεταβλητή.

Το σφάλμα ERROR1_BOOTAIC είναι το ποσοστό των φορών από τις 100, δηλαδή τα 100 τελικά μοντέλα, που το κριτήριο AIC σε συνδυασμό με τη μέθοδο bootstrap έχει επιλέξει μοντέλο με μια τουλάχιστον μη σημαντική μεταβλητή. Το ERROR1_BOOTBIC υπολογίζει ότι και το ERROR1_BOOTAIC όταν χρησιμοποιείται όμως το κριτήριο BIC.

Όμοια το ERROR1_JACKAIC για την μέθοδο jackknife υπολογίζει το ποσοστό των φορών από τις 100 που το κριτήριο AIC σε συνδυασμό με τη μέθοδο jackknife έχει επιλέξει μοντέλο με μια τουλάχιστον μη σημαντική μεταβλητή. Το ERROR1_JACKBIC υπολογίζει ότι και το ERROR1_JACKAIC όταν χρησιμοποιείται όμως το κριτήριο BIC.

Η δεύτερη περίτωση σφάλματος που υπολογίζεται είναι το σφάλμα ERROR2AIC το οποίο αποτελεί το ποσοστό των φορών από τις 100 που το κριτήριο AIC έχει επιλέξει μοντέλο στο οποίο δεν επιλέγεται μια σημαντική μεταβλητή. Όμοια το ERROR2BIC

υπολογίζει το ποσοστό των φορών από τις 100 που το κριτήριο BIC έχει επιλέξει μοντέλο στο οποίο δεν επιλέγεται μια σημαντική μεταβλητή.

Το ERROR2_BOOTAIC αποτελεί το ποσοστό των φορών από τις 100, 100 τελικά μοντέλα, που το κριτήριο AIC σε συνδυασμό με τη μέθοδο bootstrap έχει επιλέξει μοντέλο στο οποίο δεν επιλέγεται μια σημαντική μεταβλητή. Το ERROR2_BOOTBIC, όπως και το ERROR2_BOOTAIC αποτελεί το ποσοστό των φορών από τις 100, 100 τελικά μοντέλα, που το κριτήριο BIC σε συνδυασμό με τη μέθοδο bootstrap έχει επιλέξει μοντέλο στο οποίο δεν επιλέγεται μια σημαντική μεταβλητή.

Το ERROR2_JACKAIC είναι το ποσοστό των φορών από τις 100 που το κριτήριο AIC σε συνδυασμό με τη μέθοδο jackknife έχει επιλέξει μοντέλο στο οποίο δεν επιλέγεται μια σημαντική μεταβλητή. Το ERROR2_JACKBIC είναι το ποσοστό των φορών από τις 100 που το κριτήριο BIC σε συνδυασμό με τη μέθοδο jackknife έχει επιλέξει μοντέλο στο οποίο δεν επιλέγεται μια σημαντική μεταβλητή.

Πίνακας 1					
Sample1=20					
b_iter	20	50	100	200	500
COR_AIC	0.6				
COR_BIC	0.71				
COR_BOOTAIC	0.63	0.67	0.66	0.66	0.68
COR_BOOTBIC	0.65	0.66	0.68	0.65	0.67
COR_JACKAIC	0.71				
COR_JACKBIC	0.71				
ERROR1AIC	0.35				
ERROR1BIC	0.2				
ERROR1_BOOTAIC	0.23	0.18	0.18	0.19	0.17
ERROR1_BOOTBIC	0.19	0.18	0.15	0.15	0.16
ERROR1_JACKAIC	0.09				
ERROR1_JACKBIC	0.1				
ERROR2AIC	0.08				
ERROR2BIC	0.13				
ERROR2_BOOTAIC	0.21	0.19	0.19	0.21	0.19
ERROR2_BOOTBIC	0.21	0.21	0.19	0.23	0.2
ERROR2_JACKAIC	0.21				
ERROR2_JACKBIC	0.2				

Πίνακας 2					
Sample1=50					
b_iter	20	50	100	200	500
Cor_AIC	0.64				
COR_BIC	0.88				
COR_BOOTAIC	0.8	0.8	0.81	0.81	0.8
COR_BOOTBIC	0.86	0.92	0.89	0.89	0.9
COR_JACKAIC	0.87				
COR_JACKBIC	0.91				
ERROR1AIC	0.36				
ERROR1BIC	0.12				
ERROR1_BOOTAIC	0.2	0.2	0.19	0.19	0.2
ERROR1_BOOTBIC	0.14	0.08	0.11	0.11	0.1
ERROR1_JACKAIC	0.13				
ERROR1_JACKBIC	0.09				
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0				
ERROR2_JACKBIC	0				

Πίνακας 3					
Sample1=100					
b_iter	20	50	100	200	500
Cor_AIC	0.71				
COR_BIC	0.96				
COR_BOOTAIC	0.9	0.87	0.91	0.91	0.91
COR_BOOTBIC	0.94	0.96	0.97	0.96	0.96
COR_JACKAIC	0.94				
COR_JACKBIC	0.96				
ERROR1AIC	0.29				
ERROR1BIC	0.04				
ERROR1_BOOTAIC	0.1	0.13	0.09	0.09	0.09
ERROR1_BOOTBIC	0.06	0.04	0.03	0.04	0.04
ERROR1_JACKAIC	0.06				
ERROR1_JACKBIC	0.04				
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0				
ERROR2_JACKBIC	0				

Πίνακας 4					
Sample1=200					
b_iter	20	50	100	200	500
Cor_AIC	0.77				
COR_BIC	0.96				
COR_BOOTAIC	0.91	0.88	0.92	0.87	0.9
COR_BOOTBIC	0.96	0.97	0.96	0.96	0.98
COR_JACKAIC	0.91				
COR_JACKBIC	0.96				
ERROR1AIC	0.23				
ERROR1BIC	0.04				
ERROR1_BOOTAIC	0.09	0.12	0.08	0.13	0.1
ERROR1_BOOTBIC	0.04	0.03	0.04	0.04	0.02
ERROR1_JACKAIC	0.09				
ERROR1_JACKBIC	0.04				
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0				
ERROR2_JACKBIC	0				

Όπως φαίνεται και από τους Πίνακες 1-4 το κριτήριο BIC δίνει καλύτερα αποτελέσματα από το κριτήριο AIC. Πρέπει να επισημάνουμε ότι το κριτήριο AIC σε συνδυασμό με τη μέθοδο bootstrap δίνει πολύ καλύτερα αποτελέσματα από το AIC καθώς επίσης ότι και το κριτήριο AIC σε συνδυασμό με τη μέθοδο jackknife δίνει πολύ καλύτερα αποτελέσματα από το AIC ειδικά όσο το μέγεθος του δείγματος αυξάνει. Για μικρό μέγεθος δείγματος η μέθοδος jackknife παρουσιάζει καλύτερα αποτελέσματα από την bootstrap. Ενώ με την αύξηση του δείγματος, η απόδοση του bootstrap βελτιώνεται παρουσιάζοντας καλύτερα αποτελέσματα από την jackknife. Το κριτήριο BIC σε συνδυασμό με τη μέθοδο bootstrap ή jackknife δίνει παρόμοια αποτελέσματα με το BIC αλλά καλύτερα αποτελέσματα από το κριτήριο AIC σε συνδυασμό με τη μέθοδο bootstrap ή jackknife.

Προσομοίωση 2

Στην δεύτερη προσομοίωση ακολουθείται η ίδια διαδικασία που περιγράφηκε παραπάνω με μόνη διαφορά ότι η συσχέτιση των συμμεταβλητών υπολογίζεται για ρ=0.8 με σκοπό να εξετάσουμε την περίπτωση των ισχυρά συσχετισμένων μεταβλητών και την επίπτωση που θα έχει αυτό στα αποτελέσματα των μεθόδων. Η εντολή που δημιουργεί τα δείγματα και αλλάζει στο πρόγραμμα είναι η:

$$out <- mvrnorm(n = sample1, mu = c(0,0,0,0), Sigma = matrix(c(1,0.8,0.64,0.512,0.8,1,0.8,0.64,0.64,0.8,1,0.8,0.512,0.64,0.8,1),4,4),$$

$$empirical = FALSE)$$

Τα αποτελέσματα της εφαρμογής για set.seed(9) παρουσιάζονται στους Πίνακες 5-8 με τα ποσοστά να αναγράφονται όπως περιγράφτηκαν στην προσομοίωση 1:

Πίνακας 5					
Sample1=20					
b_iter	20	50	100	200	500
Cor_AIC	0.54				
COR_BIC	0.55				
COR_BOOTAIC	0.47	0.48	0.5	0.5	0.53
COR_BOOTBIC	0.46	0.45	0.49	0.47	0.48
COR_JACKAIC	0.51				
COR_JACKBIC	0.5				
ERROR1AIC	0.39				
ERROR1BIC	0.32				
ERROR1_BOOTAIC	0.34	0.27	0.25	0.24	0.25
ERROR1_BOOTBIC	0.29	0.26	0.22	0.22	0.21
ERROR1_JACKAIC	0.2				
ERROR1_JACKBIC	0.22				
ERROR2AIC	0.29				
ERROR2BIC	0.33				
ERROR2_BOOTAIC	0.4	0.42	0.4	0.4	0.37
ERROR2_BOOTBIC	0.44	0.46	0.43	0.44	0.44
ERROR2_JACKAIC	0.43				
ERROR2_JACKBIC	0.43				

Πίνακας 6					
Sample1=50					
b_iter	20	50	100	200	500
Cor_AIC	0.62				
COR_BIC	0.8				
COR_BOOTAIC	0.67	0.73	0.77	0.76	0.74
COR_BOOTBIC	0.74	0.79	0.8	0.81	0.81
COR_JACKAIC	0.79				
COR_JACKBIC	0.82				
ERROR1AIC	0.38				
ERROR1BIC	0.2				
ERROR1_BOOTAIC	0.3	0.24	0.23	0.22	0.25
ERROR1_BOOTBIC	0.23	0.16	0.15	0.16	0.15
ERROR1_JACKAIC	0.2				
ERROR1_JACKBIC	0.18				
ERROR2AIC	0.03				
ERROR2BIC	0.03				
ERROR2_BOOTAIC	0.07	0.07	0.04	0.06	0.04
ERROR2_BOOTBIC	0.08	0.09	0.08	0.07	0.07
ERROR2_JACKAIC	0.04				
ERROR2_JACKBIC	0.04				

Πίνακας 7					
Sample1=100					
b_iter	20	50	100	200	500
Cor_AIC	0.72				
COR_BIC	0.95				
COR_BOOTAIC	0.89	0.89	0.92	0.93	0.94
COR_BOOTBIC	0.96	0.98	0.96	0.96	0.96
COR_JACKAIC	0.93				
COR_JACKBIC	0.95				
ERROR1AIC	0.28				
ERROR1BIC	0.05				
ERROR1_BOOTAIC	0.11	0.11	0.08	0.07	0.06
ERROR1_BOOTBIC	0.04	0.01	0.03	0.03	0.03
ERROR1_JACKAIC	0.07				
ERROR1_JACKBIC	0.05				
ERROR2AIC	0				
ERROR2BIC	0.01				
ERROR2_BOOTAIC	0.01	0.01	0	0	0
ERROR2_BOOTBIC	0.01	0.02	0.01	0.02	0.02
ERROR2_JACKAIC	0				
ERROR2_JACKBIC	0.01				

Πίνακας 8					
Sample1=200					
b_iter	20	50	100	200	500
Cor_AIC	0.72				
COR_BIC	0.96				
COR_BOOTAIC	0.88	0.9	0.89	0.88	0.9
COR_BOOTBIC	0.95	0.96	0.97	0.96	0.96
COR_JACKAIC	0.9				
COR_JACKBIC	0.96				
ERROR1AIC	0.28				
ERROR1BIC	0.04				
ERROR1_BOOTAIC	0.12	0.1	0.11	0.12	0.1
ERROR1_BOOTBIC	0.05	0.04	0.03	0.04	0.04
ERROR1_JACKAIC	0.1				
ERROR1_JACKBIC	0.04				
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0				
ERROR2_JACKBIC	0				

Από τα αποτελέσματα των Πινάκων 5-8 φαίνεται ότι με την αύξηση του συντελεστή συσχέτισης, δηλαδή για ισχυρά συσχετισμένες μεταβλητές, υπάρχει και αύξηση του ποσοστού των φορών που μη σημαντικές μεταβλητές επιλέγονται από τα δύο κριτήρια με όλες τις μεθόδους, καθώς και αύξηση του ποσοστού των φορών που δεν επιλέγεται μια σημαντική μεταβλητή. Με την αύξηση της συσχέτισης οι μεταβλητές είναι δύσκολο να διακριθούν στατιστικά, αφού η σημασία κάθε συσχετιζόμενης μεταβλητής αυξάνεται ενώ το ατομικό βάρος κάθε μεταβλητής μειώνεται. Το αποτέλεσμα είναι η σημασία των σημαντικών μεταβλητών που έχουμε θεωρήσει από την αρχή των προσομοιώσεων, να εξασθενήσει. Μια λύση για την βελτίωση των σφαλμάτων τύπου I και II, όπως φαίνεται και από τους πίνακες και ιδιαίτερα από τον Πίνακα 8, είναι η αύξηση του μεγέθους δείγματος. Επίσης, επισημαίνεται ότι το κριτήριο BIC προσφέρει καλύτερα αποτελέσματα από το AIC στις περισσότερες περιπτώσεις ενώ για μέγεθος δείγματος 100 και 200 το BIC παρουσιάζει μέγεθος σφάλματος τύπου I και II κάτω από 0.05. Και πάλι θα πρέπει να πούμε ότι το κριτήριο AIC σε συνδυασμό με τις μεθόδους bootstrap και jackknife δίνει καλύτερα αποτελέσματα από το AIC ειδικά όσο το μέγεθος του δείγματος αυξάνει. Το κριτήριο BIC σε συνδυασμό με τις μεθόδους bootstrap και jackknife δεν καταφέρνει να δώσει σημαντικά βελτιωμένα αποτελέσματα από το BIC.

Προσομοίωση 3

Σε αυτή την προσομοίωση υποθέτουμε ότι ο συντελεστής συσχετίσεως μεταξύ των συμμεταβλητών είναι ίσος με 0, δηλαδή εξετάζουμε την περίπτωση των ανεξάρτητων συμμεταβλητών. Η αλλαγή στον κώδικα εφαρμόζεται μόνο στον πίνακα διασπορών-συνδιασπορών. Δηλαδή:

$$out <- mvrnorm(n = sample1, mu = c(0,0,0,0), Sigma = diag(4), empirical = FALSE)$$

Τα αποτελέσματα παρουσιάζονται στους Πίνακες 9-12:

Πίνακας 9					
Sample1=20					
b_iter	20	50	100	200	500
Cor_AIC	0.55				
COR_BIC	0.68				
COR_BOOTAIC	0.69	0.7	0.7	0.72	0.72
COR_BOOTBIC	0.67	0.71	0.73	0.73	0.73
COR_JACKAIC	0.72				
COR_JACKBIC	0.72				
ERROR1AIC	0.39				
ERROR1BIC	0.22				
ERROR1_BOOTAIC	0.19	0.16	0.18	0.16	0.15
ERROR1_BOOTBIC	0.18	0.14	0.14	0.13	0.14
ERROR1_JACKAIC	0.09				
ERROR1_JACKBIC	0.09				
ERROR2AIC	0.14				
ERROR2BIC	0.17				
ERROR2_BOOTAIC	0.19	0.17	0.17	0.17	0.17
ERROR2_BOOTBIC	0.21	0.19	0.17	0.18	0.17
ERROR2_JACKAIC	0.21				
ERROR2_JACKBIC	0.22				

Πίνακας 10					
Sample1=50					
b_iter	20	50	100	200	500
Cor_AIC	0.65				
COR_BIC	0.84				
COR_BOOTAIC	0.82	0.84	0.8	0.77	0.81
COR_BOOTBIC	0.89	0.88	0.87	0.87	0.87
COR_JACKAIC	0.86				
COR_JACKBIC	0.86				
ERROR1AIC	0.35				
ERROR1BIC	0.16				
ERROR1_BOOTAIC	0.18	0.16	0.2	0.23	0.19
ERROR1_BOOTBIC	0.11	0.12	0.13	0.13	0.13
ERROR1_JACKAIC	0.14				
ERROR1_JACKBIC	0.14				
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0				
ERROR2_JACKBIC	0				

Πίνακας 11					
Sample1=100					
b_iter	20	50	100	200	500
Cor_AIC	0.64				
COR_BIC	0.91				
COR_BOOTAIC	0.78	0.81	0.84	0.85	0.83
COR_BOOTBIC	0.9	0.94	0.92	0.94	0.94
COR_JACKAIC	0.86				
COR_JACKBIC	0.92				
ERROR1AIC	0.36				
ERROR1BIC	0.09				
ERROR1_BOOTAIC	0.22	0.19	0.16	0.15	0.17
ERROR1_BOOTBIC	0.1	0.06	0.08	0.06	0.06
ERROR1_JACKAIC	0.14				
ERROR1_JACKBIC	0.08				
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0				
ERROR2_JACKBIC	0				

Πίνακας 12					
***Sample*1=200**					
b_iter	20	50	100	200	500
Cor_AIC	0.69				
COR_BIC	0.98				
COR_BOOTAIC	0.85	0.86	0.87	0.87	0.9
COR_BOOTBIC	0.99	0.98	0.99	0.98	1
COR_JACKAIC	0.92				
COR_JACKBIC	0.98				
ERROR1AIC	0.31				
ERROR1BIC	0.02				
ERROR1_BOOTAIC	0.15	0.14	0.13	0.13	0.1
ERROR1_BOOTBIC	0.01	0.02	0.01	0.02	0
ERROR1_JACKAIC	0.08				
ERROR1_JACKBIC	0.02				
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0				
ERROR2_JACKBIC	0				

Όπως φαίνεται από τους Πίνακες 9-12 και σε αυτή την περίπτωση η μέθοδος BIC δίνει καλύτερα αποτελέσματα από το AIC ιδιαίτερα όσο το μέγεθος του δείγματος αυξάνει. Επίσης το κριτήριο BIC σε συνδυασμό με τις μεθόδους bootstrap και jackknife δίνει τα καλύτερα αποτελέσματα ειδικά όσο αυξάνει ο αριθμός των bootstrap επαναλήψεων. Εδώ που δεν υπάρχει συσχέτιση μεταξύ των μεταβλητών, για μεγάλο μέγεθος δείγματος και μεγάλο αριθμό bootstrap δειγμάτων παρατηρούμε ότι το ποσοστό επιτυχίας πλησιάζει και πετυχαίνει ενίοτε και τη μονάδα.

Προσομοίωση 4

Στην τελευταία προσομοίωση επιλέγονται διαφορετικές τιμές για το κατώτατο όριο που χρησιμοποιήθηκε για την επιλογή των μεταβλητών στο τελικό μοντέλο, όπως περιγράφηκε στην αρχή του κεφαλαίου. Για την επιλογή της πρώτης μεταβλητής είχαμε θεωρήσει το κάτω όριο

$$F_1 = cov1 * \frac{sig1}{100} \geq v = 30$$

αλλά το ίδιο κάτω όριο, δηλαδη το 30 είχε υιοθετηθεί για όλες τις συμμεταβλητές.

Εδώ για σκοπούς σύγκρισης θα θεωρήσουμε διαφορετικές τιμές του ν για να δούμε πως αυτές μπορεί να επηρεάσουν τα αποτελέσματα και πιο είναι το κάτω όριο που δίνει τα καλύτερα αποτελέσματα. Στους Πίνακες 13-17 δίνονται τα αποτελέσματα για μέγεθος δείγματος 20 και για διαφορετικές τιμές του b_iter και του ν. Συγκεκριμένα, η σύγκριση γίνεται για $v = 10, 15, 20, 25, 30$.

Όπως και στις προηγούμενες περιπτώσεις, οι συμβολισμοί για τα σφάλματα και τα ποσοστά σωστής επιλογής μοντέλου που υπολογίζονται, είναι οι ίδιοι.

Πίνακας 13					
Sample1=20					
b_iter=20					
v	10	15	20	25	30
Cor_AIC	0.6				
COR_BIC	0.71				
COR_BOOTAIC	0.41	0.48	0.59	0.59	0.63
COR_BOOTBIC	0.45	0.56	0.62	0.65	0.65
COR_JACKAIC	0.71	0.71	0.7	0.71	0.71
COR_JACKBIC	0.7	0.71	0.71	0.71	0.71
ERROR1AIC	0.35				
ERROR1BIC	0.2				
ERROR1_BOOTAIC	0.56	0.48	0.33	0.28	0.23
ERROR1_BOOTBIC	0.52	0.38	0.29	0.23	0.19
ERROR1_JACKAIC	0.16	0.13	0.12	0.1	0.09
ERROR1_JACKBIC	0.16	0.12	0.11	0.1	0.1
ERROR2AIC	0.08				
ERROR2BIC	0.13				
ERROR2_BOOTAIC	0.05	0.09	0.11	0.21	0.21
ERROR2_BOOTBIC	0.07	0.11	0.15	0.2	0.21
ERROR2_JACKAIC	0.15	0.17	0.19	0.2	0.21
ERROR2_JACKBIC	0.16	0.18	0.19	0.2	0.2

Πίνακας 14					
Sample1=20					
b_iter=50					
v	10	15	20	25	30
Cor_AIC	0.6				
COR_BIC	0.71				
COR_BOOTAIC	0.35	0.48	0.62	0.65	0.67
COR_BOOTBIC	0.43	0.54	0.65	0.68	0.66
COR_JACKAIC	0.71	0.71	0.7	0.71	0.71
COR_JACKBIC	0.7	0.71	0.71	0.71	0.71
ERROR1AIC	0.35				
ERROR1BIC	0.2				
ERROR1_BOOTAIC	0.63	0.46	0.31	0.23	0.18
ERROR1_BOOTBIC	0.54	0.39	0.26	0.19	0.18
ERROR1_JACKAIC	0.16	0.13	0.12	0.1	0.09
ERROR1_JACKBIC	0.16	0.12	0.11	0.1	0.1
ERROR2AIC	0.08				
ERROR2BIC	0.13				
ERROR2_BOOTAIC	0.06	0.1	0.13	0.18	0.19
ERROR2_BOOTBIC	0.08	0.12	0.16	0.17	0.21
ERROR2_JACKAIC	0.15	0.17	0.19	0.2	0.21
ERROR2_JACKBIC	0.16	0.18	0.19	0.2	0.2

Πίνακας 15					
Sample1=20					
b_iter=100					
v	10	15	20	25	30
Cor_AIC	0.6				
COR_BIC	0.71				
COR_BOOTAIC	0.4	0.54	0.61	0.63	0.66
COR_BOOTBIC	0.5	0.6	0.66	0.69	0.68
COR_JACKAIC	0.71	0.71	0.7	0.71	0.71
COR_JACKBIC	0.7	0.71	0.71	0.71	0.71
ERROR1AIC	0.35				
ERROR1BIC	0.2				
ERROR1_BOOTAIC	0.58	0.39	0.29	0.24	0.18
ERROR1_BOOTBIC	0.45	0.31	0.22	0.17	0.15
ERROR1_JACKAIC	0.16	0.13	0.12	0.1	0.09
ERROR1_JACKBIC	0.16	0.12	0.11	0.1	0.1
ERROR2AIC	0.08				
ERROR2BIC	0.13				
ERROR2_BOOTAIC	0.06	0.1	0.13	0.16	0.19
ERROR2_BOOTBIC	0.08	0.12	0.14	0.17	0.19
ERROR2_JACKAIC	0.15	0.17	0..19	0.2	0.21
ERROR2_JACKBIC	0.16	0.18	0.19	0.2	0.2

Πίνακας 16					
Sample1=20					
b_iter=200					
v	10	15	20	25	30
Cor_AIC	0.6				
COR_BIC	0.71				
COR_BOOTAIC	0.36	0.52	0.59	0.64	0.66
COR_BOOTBIC	0.45	0.58	0.62	0.66	0.65
COR_JACKAIC	0.71	0.71	0.7	0.71	0.71
COR_JACKBIC	0.7	0.71	0.71	0.71	0.71
ERROR1AIC	0.35				
ERROR1BIC	0.2				
ERROR1_BOOTAIC	0.62	0.43	0.31	0.23	0.19
ERROR1_BOOTBIC	0.52	0.35	0.25	0.19	0.15
ERROR1_JACKAIC	0.16	0.13	0.12	0.1	0.09
ERROR1_JACKBIC	0.16	0.12	0.11	0.1	0.1
ERROR2AIC	0.08				
ERROR2BIC	0.13				
ERROR2_BOOTAIC	0.06	0.07	0.13	0.17	0.21
ERROR2_BOOTBIC	0.07	0.1	0.17	0.2	0.23
ERROR2_JACKAIC	0.15	0.17	0.19	0.2	0.21
ERROR2_JACKBIC	0.16	0.18	0.19	0.2	0.2

Πίνακας 17					
Sample1=20					
b_iter=500					
v	10	15	20	25	30
Cor_AIC	0.6				
COR_BIC	0.71				
COR_BOOTAIC	0.34	0.54	0.61	0.63	0.68
COR_BOOTBIC	0.45	0.6	0.64	0.68	0.67
COR_JACKAIC	0.71	0.71	0.7	0.71	0.71
COR_JACKBIC	0.7	0.71	0.71	0.71	0.71
ERROR1AIC	0.35				
ERROR1BIC	0.2				
ERROR1_BOOTAIC	0.63	0.41	0.29	0.24	0.17
ERROR1_BOOTBIC	0.52	0.31	0.24	0.18	0.16
ERROR1_JACKAIC	0.16	0.13	0.12	0.1	0.09
ERROR1_JACKBIC	0.16	0.12	0.11	0.1	0.1
ERROR2AIC	0.08				
ERROR2BIC	0.13				
ERROR2_BOOTAIC	0.06	0.09	0.13	0.17	0.19
ERROR2_BOOTBIC	0.06	0.11	0.15	0.18	0.2
ERROR2_JACKAIC	0.15	0.17	0.19	0.2	0.21
ERROR2_JACKBIC	0.16	0.18	0.19	0.2	0.2

Από τους Πίνακες 13-17 για μέγεθος δείγματος 20, παρατηρούμε ότι για διαφορετικό αριθμό bootstrap επαναλήψεων παίρνουμε περίπου τα ίδια αποτελέσματα. Καλύτερα αποτελέσματα παίρνουμε όσο αυξάνεται το κατώτατο όριο ν και γίνεται ίσο με 25 ή 30. Σε αντιδιαστολή τα σφάλματα τύπου Ι είναι πολύ μεγαλύτερα όσο μικρότερο είναι το κατώτατο όριο ν που επιλέγουμε. Τα σφάλματα τύπου ΙΙ για το bootstrap είναι πολύ μικρότερα για μικρότερες τιμές του κατώτατου ορίου ν.

Η μέθοδος jackknife δεν φαίνεται να επηρεάζεται δραστικά από το ν και δίνει ένα ποσοστό επιτυχίας περίπου σταθερό στο 70% για όλα τα ν. Για μικρότερο ν παρουσιάζει όμως μικρότερα σφάλματα τύπου ΙΙ.

Στους Πίνακες 18-22 παρουσιάζονται τα αποτελέσματα για μέγεθος δείγματος 50.

Πίνακας 18					
Sample1=50					
b_iter=20					
v	10	15	20	25	30
Cor_AIC	0.64				
COR_BIC	0.88				
COR_BOOTAIC	0.46	0.55	0.65	0.77	0.8
COR_BOOTBIC	0.65	0.73	0.77	0.84	0.86
COR_JACKAIC	0.82	0.84	0.85	0.87	0.87
COR_JACKBIC	0.88	0.9	0.9	0.91	0.91
ERROR1AIC	0.36				
ERROR1BIC	0.12				
ERROR1_BOOTAIC	0.54	0.45	0.35	0.23	0.2
ERROR1_BOOTBIC	0.35	0.27	0.23	0.16	0.14
ERROR1_JACKAIC	0.18	0.16	0.15	0.13	0.13
ERROR1_JACKBIC	0.12	0.1	0.1	0.09	0.09
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0	0	0	0	0
ERROR2_JACKBIC	0	0	0	0	0

Πίνακας 19					
Sample1=50					
b_iter=50					
v	10	15	20	25	30
Cor_AIC	0.64				
COR_BIC	0.88				
COR_BOOTAIC	0.48	0.62	0.69	0.75	0.8
COR_BOOTBIC	0.66	0.72	0.79	0.86	0.92
COR_JACKAIC	0.82	0.84	0.85	0.87	0.87
COR_JACKBIC	0.88	0.9	0.9	0.91	0.91
ERROR1AIC	0.36				
ERROR1BIC	0.12				
ERROR1_BOOTAIC	0.52	0.38	0.31	0.25	0.2
ERROR1_BOOTBIC	0.34	0.28	0.21	0.14	0.08
ERROR1_JACKAIC	0.18	0.16	0.15	0.13	0.13
ERROR1_JACKBIC	0.12	0.1	0.1	0.09	0.09
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0	0	0	0	0
ERROR2_JACKBIC	0	0	0	0	0

Πίνακας 20					
Sample1=50					
b_iter=100					
v	10	15	20	25	30
Cor_AIC	0.64				
COR_BIC	0.88				
COR_BOOTAIC	0.49	0.63	0.71	0.76	0.81
COR_BOOTBIC	0.66	0.77	0.78	0.89	0.89
COR_JACKAIC	0.82	0.84	0.85	0.87	0.87
COR_JACKBIC	0.88	0.9	0.9	0.91	0.91
ERROR1AIC	0.36				
ERROR1BIC	0.12				
ERROR1_BOOTAIC	0.51	0.37	0.29	0.24	0.19
ERROR1_BOOTBIC	0.34	0.23	0.22	0.11	0.11
ERROR1_JACKAIC	0.18	0.16	0.15	0.13	0.13
ERROR1_JACKBIC	0.12	0.1	0.1	0.09	0.09
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0	0	0	0	0
ERROR2_JACKBIC	0	0	0	0	0

Πίνακας 21					
Sample1=50					
b_iter=200					
v	10	15	20	25	30
Cor_AIC	0.64				
COR_BIC	0.88				
COR_BOOTAIC	0.47	0.61	0.69	0.74	0.81
COR_BOOTBIC	0.63	0.73	0.81	0.84	0.89
COR_JACKAIC	0.82	0.84	0.85	0.87	0.87
COR_JACKBIC	0.88	0.9	0.9	0.91	0.91
ERROR1AIC	0.36				
ERROR1BIC	0.12				
ERROR1_BOOTAIC	0.53	0.39	0.31	0.26	0.19
ERROR1_BOOTBIC	0.37	0.27	0.19	0.16	0.11
ERROR1_JACKAIC	0.18	0.16	0.15	0.13	0.13
ERROR1_JACKBIC	0.12	0.1	0.1	0.09	0.09
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0	0	0	0	0
ERROR2_JACKBIC	0	0	0	0	0

Πίνακας 22					
Sample1=50					
b_iter=500					
v	10	15	20	25	30
Cor_AIC	0.64				
COR_BIC	0.88				
COR_BOOTAIC	0.52	0.62	0.68	0.77	0.8
COR_BOOTBIC	0.64	0.76	0.81	0.87	0.9
COR_JACKAIC	0.82	0.84	0.85	0.87	0.87
COR_JACKBIC	0.88	0.9	0.9	0.91	0.91
ERROR1AIC	0.36				
ERROR1BIC	0.12				
ERROR1_BOOTAIC	0.48	0.38	0.32	0.23	0.2
ERROR1_BOOTBIC	0.36	0.24	0.19	0.13	0.1
ERROR1_JACKAIC	0.18	0.16	0.15	0.13	0.13
ERROR1_JACKBIC	0.12	0.1	0.1	0.09	0.09
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0	0	0	0	0
ERROR2_JACKBIC	0	0	0	0	0

Από τους Πίνακες 18-22 φαίνεται ότι για μέγεθος δείγματος 50, για $\nu = 30\ και\ 25$, η μέθοδος jackknife παρουσιάζει τα καλύτερα αποτελέσματα, κυρίως σε συνδυασμό με το BIC. Το bootstrap για αυξανόμενο αριθμό επαναλήψεων βελτιώνει τα ποσοστά του σφάλματος τύπου Ι και κατά συνέπεια αυξάνει τα ποσοστά της επιτυχίας ενώ κυρίως το $\nu = 30$ αποτελεί το κατώτατο όριο με τα καλύτερα αποτελέσματα. Οι τιμές του σφάλματος τύπου ΙΙ είναι 0 και για τις δύο μεθόδους.

Στη συνέχεια, στους Πίνακες 23-27 παρατίθενται τα αποτελέσματα για δείγμα μεγέθους 100.

Πίνακας 23					
Sample1=100					
b_iter=20					
v	10	15	20	25	30
Cor_AIC	0.71				
COR_BIC	0.96				
COR_BOOTAIC	0.56	0.7	0.79	0.85	0.9
COR_BOOTBIC	0.8	0.84	0.92	0.93	0.94
COR_JACKAIC	0.92	0.92	0.94	0.94	0.94
COR_JACKBIC	0.96	0.96	0.96	0.96	0.96
ERROR1AIC	0.29				
ERROR1BIC	0.04				
ERROR1_BOOTAIC	0.44	0.3	0.21	0.15	0.1
ERROR1_BOOTBIC	0.2	0.16	0.08	0.07	0.06
ERROR1_JACKAIC	0.08	0.08	0.06	0.06	0.06
ERROR1_JACKBIC	0.04	0.04	0.04	0.04	0.04
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0	0	0	0	0
ERROR2_JACKBIC	0	0	0	0	0

Πίνακας 24					
Sample1=100					
b_iter=50					
v	10	15	20	25	30
Cor_AIC	0.71				
COR_BIC	0.96				
COR_BOOTAIC	0.58	0.71	0.81	0.86	0.87
COR_BOOTBIC	0.82	0.86	0.93	0.94	0.96
COR_JACKAIC	0.92	0.92	0.94	0.94	0.94
COR_JACKBIC	0.96	0.96	0.96	0.96	0.96
ERROR1AIC	0.29				
ERROR1BIC	0.04				
ERROR1_BOOTAIC	0.42	0.29	0.19	0.14	0.13
ERROR1_BOOTBIC	0.18	0.14	0.07	0.06	0.04
ERROR1_JACKAIC	0.08	0.08	0.06	0.06	0.06
ERROR1_JACKBIC	0.04	0.04	0.04	0.04	0.04
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0	0	0	0	0
ERROR2_JACKBIC	0	0	0	0	0

Πίνακας 25					
Sample1=100					
b_iter=100					
v	10	15	20	25	30
Cor_AIC	0.71				
COR_BIC	0.96				
COR_BOOTAIC	0.61	0.68	0.78	0.86	0.91
COR_BOOTBIC	0.82	0.9	0.92	0.95	0.97
COR_JACKAIC	0.92	0.92	0.94	0.94	0.94
COR_JACKBIC	0.96	0.96	0.96	0.96	0.96
ERROR1AIC	0.29				
ERROR1BIC	0.04				
ERROR1_BOOTAIC	0.39	0.32	0.22	0.14	0.09
ERROR1_BOOTBIC	0.18	0.1	0.08	0.05	0.03
ERROR1_JACKAIC	0.08	0.08	0.06	0.06	0.06
ERROR1_JACKBIC	0.04	0.04	0.04	0.04	0.04
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0	0	0	0	0
ERROR2_JACKBIC	0	0	0	0	0

Πίνακας 26					
Sample1=100					
b_iter=200					
v	10	15	20	25	30
Cor_AIC	0.71				
COR_BIC	0.96				
COR_BOOTAIC	0.57	0.69	0.78	0.85	0.91
COR_BOOTBIC	0.84	0.92	0.93	0.94	0.96
COR_JACKAIC	0.92	0.92	0.94	0.94	0.94
COR_JACKBIC	0.96	0.96	0.96	0.96	0.96
ERROR1AIC	0.29				
ERROR1BIC	0.04				
ERROR1_BOOTAIC	0.43	0.31	0.22	0.15	0.09
ERROR1_BOOTBIC	0.16	0.08	0.07	0.06	0.04
ERROR1_JACKAIC	0.08	0.08	0.06	0.06	0.06
ERROR1_JACKBIC	0.04	0.04	0.04	0.04	0.04
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0	0	0	0	0
ERROR2_JACKBIC	0	0	0	0	0

Πίνακας 27					
Sample1=100					
b_iter=500					
v	10	15	20	25	30
Cor_AIC	0.71				
COR_BIC	0.96				
COR_BOOTAIC	0.56	0.7	0.79	0.85	0.91
COR_BOOTBIC	0.83	0.91	0.93	0.95	0.96
COR_JACKAIC	0.92	0.92	0.94	0.94	0.94
COR_JACKBIC	0.96	0.96	0.96	0.96	0.96
ERROR1AIC	0.29				
ERROR1BIC	0.04				
ERROR1_BOOTAIC	0.44	0.3	0.21	0.15	0.09
ERROR1_BOOTBIC	0.17	0.09	0.07	0.05	0.04
ERROR1_JACKAIC	0.08	0.08	0.06	0.06	0.06
ERROR1_JACKBIC	0.04	0.04	0.04	0.04	0.04
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0	0	0	0	0
ERROR2_JACKBIC	0	0	0	0	0

Παρατηρούμε ότι για μέγεθος δείγματος 100 η jackknife μέθοδος έχει πολύ καλά αποτελέσματα και ιδιαίτερα το κριτήριο BIC που παρουσιάζει και το χαμηλότερο σφάλμα τύπου Ι.

Παρόμοια η μέθοδος bootstrap όσο αυξάνεται ο αριθμός των επαναλήψεων επιλέγει το σωστό μοντέλο τις περισσότερες φορές ενώ το σφάλμα τύπου Ι μειώνεται. Αυτό συμβαίνει κυρίως για $v = 30$.

Στους Πίνακες 28-32 δίνονται τέλος τα αποτελέσματα για δείγμα μεγέθους 200.

Πίνακας 28					
Sample1=200					
b_iter=20					
v	10	15	20	25	30
Cor_AIC	0.77				
COR_BIC	0.96				
COR_BOOTAIC	0.61	0.72	0.81	0.87	0.91
COR_BOOTBIC	0.87	0.91	0.95	0.96	0.96
COR_JACKAIC	0.91	0.91	0.91	0.91	0.91
COR_JACKBIC	0.96	0.96	0.96	0.96	0.96
ERROR1AIC	0.23				
ERROR1BIC	0.04				
ERROR1_BOOTAIC	0.39	0.28	0.19	0.13	0.09
ERROR1_BOOTBIC	0.13	0.09	0.05	0.04	0.04
ERROR1_JACKAIC	0.09	0.09	0.09	0.09	0.09
ERROR1_JACKBIC	0.04	0.04	0.04	0.04	0.04
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0	0	0	0	0
ERROR2_JACKBIC	0	0	0	0	0

Πίνακας 29					
Sample1=200					
b_iter=50					
v	10	15	20	25	30
Cor_AIC	0.77				
COR_BIC	0.96				
COR_BOOTAIC	0.64	0.77	0.82	0.86	0.88
COR_BOOTBIC	0.9	0.93	0.94	0.95	0.97
COR_JACKAIC	0.91	0.91	0.91	0.91	0.91
COR_JACKBIC	0.96	0.96	0.96	0.96	0.96
ERROR1AIC	0.23				
ERROR1BIC	0.04				
ERROR1_BOOTAIC	0.36	0.23	0.18	0.14	0.12
ERROR1_BOOTBIC	0.1	0.07	0.06	0.05	0.03
ERROR1_JACKAIC	0.09	0.09	0.09	0.09	0.09
ERROR1_JACKBIC	0.04	0.04	0.04	0.04	0.04
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0	0	0	0	0
ERROR2_JACKBIC	0	0	0	0	0

Πίνακας 30					
Sample1=200					
b_iter=100					
v	10	15	20	25	30
Cor_AIC	0.77				
COR_BIC	0.96				
COR_BOOTAIC	0.67	0.76	0.83	0.89	0.92
COR_BOOTBIC	0.9	0.93	0.94	0.95	0.96
COR_JACKAIC	0.91	0.91	0.91	0.91	0.91
COR_JACKBIC	0.96	0.96	0.96	0.96	0.96
ERROR1AIC	0.23				
ERROR1BIC	0.04				
ERROR1_BOOTAIC	0.33	0.24	0.17	0.11	0.08
ERROR1_BOOTBIC	0.1	0.07	0.06	0.05	0.04
ERROR1_JACKAIC	0.09	0.09	0.09	0.09	0.09
ERROR1_JACKBIC	0.04	0.04	0.04	0.04	0.04
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0	0	0	0	0
ERROR2_JACKBIC	0	0	0	0	0

Πίνακας 31					
Sample1=200					
b_iter=200					
v	10	15	20	25	30
Cor_AIC	0.77				
COR_BIC	0.96				
COR_BOOTAIC	0.64	0.77	0.83	0.84	0.87
COR_BOOTBIC	0.88	0.92	0.95	0.96	0.96
COR_JACKAIC	0.91	0.91	0.91	0.91	0.91
COR_JACKBIC	0.96	0.96	0.96	0.96	0.96
ERROR1AIC	0.23				
ERROR1BIC	0.04				
ERROR1_BOOTAIC	0.36	0.23	0.17	0.16	0.13
ERROR1_BOOTBIC	0.12	0.08	0.05	0.04	0.04
ERROR1_JACKAIC	0.09	0.09	0.09	0.09	0.09
ERROR1_JACKBIC	0.04	0.04	0.04	0.04	0.04
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0	0	0	0	0
ERROR2_JACKBIC	0	0	0	0	0

Πίνακας 32					
Sample1=200					
b_iter=500					
v	10	15	20	25	30
Cor_AIC	0.77				
COR_BIC	0.96				
COR_BOOTAIC	0.65	0.78	0.81	0.87	0.9
COR_BOOTBIC	0.92	0.92	0.94	0.96	0.98
COR_JACKAIC	0.91	0.91	0.91	0.91	0.91
COR_JACKBIC	0.96	0.96	0.96	0.96	0.96
ERROR1AIC	0.23				
ERROR1BIC	0.04				
ERROR1_BOOTAIC	0.35	0.22	0.19	0.13	0.1
ERROR1_BOOTBIC	0.08	0.08	0.06	0.04	0.02
ERROR1_JACKAIC	0.09	0.09	0.09	0.09	0.09
ERROR1_JACKBIC	0.04	0.04	0.04	0.04	0.04
ERROR2AIC	0				
ERROR2BIC	0				
ERROR2_BOOTAIC	0	0	0	0	0
ERROR2_BOOTBIC	0	0	0	0	0
ERROR2_JACKAIC	0	0	0	0	0
ERROR2_JACKBIC	0	0	0	0	0

Για μέγεθος δείγματος 200 παρατηρούμε από τους Πίνακες 28-32 ότι το κριτήριο BIC σε συνδυασμό με την μέθοδο jackknife παρουσιάζει πολύ καλά αποτελέσματα ενώ το ποσοστό που επιλέγεται μια τουλάχιστον μη σημαντική μεταβλητή είναι 0.04. Γενικώς η μέθοδος jackknife δεν φαίνεται να επηρεάζεται πολύ από το ν. Ομοίως για την μέθοδο bootstrap φαίνεται ότι τα καλύτερα αποτελέσματα δίνονται από το κριτήριο BIC με το ποσοστό σφάλματος τύπου I να είναι 0.02 για 500 bootstrap επαναλήψεις. Όπως φαίνεται ο καλύτερος συνδυασμός των αποτελεσμάτων των κριτηρίων δίνεται για $v = 30$.

5.2 Συνοπτικά αποτελέσματα

Συμπεράσματα

- Όσο αυξάνει το μέγεθος του δείγματος τόσο καλύτερα τα αποτελέσματα των όλων των μεθόδων.

- Όσο αυξάνει ο αριθμός των bootstrap επαναλήψεων τόσο καλύτερα και τα αποτελέσματα των δύο κριτηρίων σε συνδυασμό με τη μέθοδο bootstrap.

- Γενικώς το κριτήριο BIC δίνει καλύτερα αποτελέσματα από το κριτήριο AIC.

- Η μέθοδος jackknife παρουσιάζει πολύ καλά αποτελέσματα σε συνδυασμό με το κριτήριο BIC.

- Η αύξηση της συσχέτισης των μεταβλητών για μικρό μέγεθος δείγματος φαίνεται να ελαττώνει την απόδοση όλων των μεθόδων. Ενώ με την αύξηση του μεγέθους του δείγματος η απόδοση αυξάνει και πάλι. Παρόμοια στην περίπτωση της μηδενικής συσχέτισης με την αύξηση του μέγεθους του δείγματος παρουσιάζεται μηδενικό σφάλμα τύπου ΙΙ και αύξηση του ποσοστού των φορών που επιλέγεται το σωστό μοντέλο.

- Από την προσομοίωση 4 φαίνεται ότι για $v = 30$ κατώτατο όριο δίνονται τα καλύτερα αποτελέσματα για την επιλογή των μεταβλητών του τελικού μοντέλου από κάθε κριτήριο. Το ν πάντως δε φαίνεται να επηρεάζει σημαντικά τη μέθοδο jackknife.

Βιβλιογραφία

Akaike H. (1973*), Information theory and an extension of the maximum likelihood principle,* 2nd International Symposium on Information Theory (Petrov B. N. and Csaki F., eds.), Akademiai Kiado, Budapest, 267–281. (Reproduced in Breakthroughs in Statistics, **1**, S. Kotz and N. L. Johnson, eds., Springer-Verlag, New York, 1992.)

Akaike H. (1974), *A new look at the statistical model identification*, IEEE Transactions on Automatic Control AC-19, 716–723.

Akaike H. (1985), *Prediction and entropy*, In A Celebration of Statistics, A. C. Atkinson and E. Fienberg. eds., Springer-Verlag, New York, 1–24

Akaike H. and Kitagawa G. (eds.) (1998). *The Practice of Time Series Analysis*. Springer-Verlag, New York

Anderson D.R. (2008), *Model-based Inference in the Life Sciences: A Primer on Evidence*, Springer New York

Barndorff-Nielsen O. E. and Cox D. R. (1989), *Asymptotic Techniques for Use in Statistics*, Chapman and Hall, New York

Bozdogan H. (1987), *Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions*, PSYCHOMETRIKA--VOL. 52, NO. 3, 345-370

Bozdogan H. (2000), *Akaike's Information Criterion and Recent Developments in Information Complexity*, Journal of Mathematical Psychology 44, 62-91

Boos D.D. and Stefanski L.A. (2013), *Essential Statistical Inference Theory and Methods*, Springer New York

Burnham K. P. & Anderson D. R. (2002), *Information and Likelihood Theory: A Basis for Model Selection and Inference*, Springer New York

Burnham K. P. & Anderson D. R. (2004), *Multimodel Inference Understanding AIC and BIC in Model Selection*, SOCIOLOGICAL METHODS & RESEARCH, Vol. 33, No. 2, November 2004 261-304, Sage Publications

Chernick M. R. and Murthy V. K. and Nealy C. D. (1985), *Applications of bootstrap and other resampling techniques: Evaluation of classifier performance.* Pattern Recog. Lett. 3 , 167 – 178

Chernick M. R. (2007), *Bootstrap Methods: A Guide for Practitioners and Researchers*, Second Edition, Wiley, Hoboken

Chernick M. R. and LaBudde R. A., (2011), *An Introduction to Statistical Inference and its Applications with R*, John Wiley & Sons, Inc., Hoboken, New Jersey.

David H. A. (1981). *Order Statistics*. Wiley, New York

Davison A. C. (1986), *Approximate predictive likelihood,* Biometrika 73, 323–332.

Davison A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*, Cambridge University Press , Cambridge .

Dudewicz E. J. (1992). *The generalized bootstrap. Bootstrapping and Related Techniques,* edited by (K.-H. Jöckel, G. Rothe and W. Sendler), Lecture Notes in Economics and Mathematical Systems, Vol. 376, pp. 31 – 37. Springer-Verlag, Berlin

Efron B. (1979), *Bootstrap Methods: Another Look at the Jackknife*, Annals of Statistics 7:1—26

Efron B. (1982), *The jackknife, the bootstrap and other resampling plans*, Society for Industrial & Applied Mathematics, Philadelphia

Efron B. (1983). *Estimating the error rate of a prediction rule: improvements on cross- validation*, J. Am. Statist. Assoc. **78**, 316 – 331

Efron B. and Gong G. (1983). *A leisurely look at the bootstrap, the jackknife and cross - validation*. Am. Stat. 37, 36 – 48

Efron B. and Tibshirani R. (1986). *Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy*, Stat. Sci. 1, 54 – 77

Efron B. & R. J. Tibshirani. (1993), *An Introduction to the Bootstrap*, New York, Chapman and Hall, New York

Efroymson M. A. (1960), *Multiple regression analysis*, In A. Ralston and Wilf H.S. (editors), Mathematical Methods for Digital Computers. Wiley.

Hall P. (1992). *The Bootstrap and Edgeworth Expansion*, Springer - Verlag, New York

Kullback S. (1978), *Information Theory and Statistics*, Dover Publications, Inc.

Konishi S. & Kitagawa G. (2008), *Information Criteria and Statistical Modeling*, Springer New York.

Quenouille, M. H. (1949), *Problems in Plane Sampling*, The Annals of Mathematical Statistics 20 (3): 355–375

Quenouille, M. H. (1956), *Notes on Bias in Estimation*, Biometrika 43 (3-4): 353–360

Rubin D. B. (1981), *The Bayesian bootstrap*, Ann. Statist. **9**, 130 – 134

Schwarz G. (1978), *Estimating the dimension of a model*, Annals of Statistics 6, 461–464.

Shao J. and Tu D. (1995), *The Jackknife and Bootstrap*, Springer-Verlag New York, Inc

Sun L. and Muller-Schwarze D. (1996), *Statistical resampling methods in biology: A case study of beaver dispersal patterns*, Am. J. Math. Manage. Sci. 16, 463 – 502

Shang J & Cavanaugh J. E. (2008), *Bootstrap Variants of the Akaike Information Criterion for Mixed Model Selection*, Statistics & Probability Letters, Elsevier

Tukey J.W. (1958), *Bias and confidence in not-quite large samples*, Annals of Mathematical Statistics 29: 614

Tierney L. and Kadane J. B. (1986), *Accurate approximations for posterior moments and marginal densities*, Journal of the American Statistical Association 81, 82–86

Trosset M. W. (2009), *An Introduction to Statistical Inference and Its Applications with R*, Chapmann and Hall/CRC.

Wilcox R. R. (2010), *Fundamentals of Modern Statistical Methods*, Second Edition, Springer New York

Xin Yan & Xiao Gang Su (2009), *Linear Regression Analysis Theory and Computing*, World Scientific Publishing Co. Pte. Ltd

Zoubir M. A. and Iskander R. D. (2004), *Bootstrap Techniques for Signal Processing*, Cambridge University Press, Cambridge

Ιστοσελίδες, pdf

Exploratory Data Analysis: Conceptual Foundations of Empirical Cumulative Distribution Functions derived from https://chemicalstatistician.wordpress.com/2013/06/24/exploratory-data-analysis-conceptual-foundations-of-empirical-cumulative-distribution-functions/

McLeod, S. A. (2008). Correlation. Retrieved from http://www.simplypsychology.org/correlation.html

https://statsmethods.wordpress.com/2013/05/10/pearson-correlation-coefficient-r/

http://www.strath.ac.uk/aer/materials/4dataanalysisineducationalresearch/unit4/correlationsdirectionandstrength/

http://www.creative-wisdom.com/computer/sas/collinear_stepwise.html

http://www.camo.com/rt/Resources/linear_regression_model.html

http://en.wikipedia.org/wiki/Maximum_likelihood

https://www.udemy.com/blog/regression-analysis/

http://en.wikipedia.org/wiki/Akaike_information_criterion

http://en.wikipedia.org/wiki/Regression_analysis

en.wikipedia.org/wiki/Stepwise_regression

http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

http://en.wikipedia.org/wiki/Correlation_and_dependence

Simpson G. (2013), *An Introduction to R for the Geosciences: Regression (pdf)*, Institute of Environmental Change & Society and Department of Biology University of Regina

http://mathbits.com/MathBits/TISection/Statistics2/correlation.htm

http://sites.stat.psu.edu/~jls/stat100/lectures/lec16.pdf

http://en.wikipedia.org/wiki/Linear_regression

http://en.wikipedia.org/wiki/Bootstrapping_%28statistics%29

http://en.wikipedia.org/wiki/Statistical_model

http://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence

http://en.wikipedia.org/wiki/Statistical_inference

http://en.wikipedia.org/wiki/Bayesian_information_criterion