ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΣΤΑΤΙΣΤΙΚΩΝ ΜΟΝΤΕΛΩΝ

ΚΟΖΥΡΑΚΗΣ ΓΙΩΡΓΟΣ

Α.Μ.:09104235

Επιβλέπων:

ΦΟΥΣΚΑΚΗΣ ΔΗΜΗΤΡΗΣ

ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ Ε.Μ.Π.

(ΑΘΗΝΑ,ΜΑΡΤΙΟΣ,2011)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΣΤΑΤΙΣΤΙΚΩΝ ΜΟΝΤΕΛΩΝ

ΚΟΖΥΡΑΚΗΣ ΓΙΩΡΓΟΣ

Α.Μ.:09104235

Επιβλέπων:

ΦΟΥΣΚΑΚΗΣ ΔΗΜΗΤΡΗΣ

ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ Ε.Μ.Π.

Επιτροπή:

.Φουσκάκης Δ.

Κοκολάκης Γ.

Σπηλιώτης Ι.

(ΑΘΗΝΑ,ΜΑΡΤΙΟΣ,2011)

Ευχαριστίες

Για την εκπόνηση της παρακάτω διπλωματικής εργασίας συνέβαλαν με τον δικό τους τρόπο οι φίλοι και συμφοιτητές μου Κοσμάς Μαντσης, Μαρία Ορφανίδου, Βασίλης Κόκκινος, και Γιάννης Μαμάης τόσο με τις συντακτικές όσο και με τις επιστημονικές διορθώσεις που μου πρότειναν. Θα ήθελα επίσης να ευχαριστήσω την φίλη μου Δρ Ινώ Αγραφιώτη για την αμέριστη βοήθεια της. Επίσης, ένα μεγάλο ευχαριστώ οφείλω και στον επιβλέπων Καθηγητή μου Επίκουρο Καθηγητή Σ.Ε.Μ.Φ.Ε. Δημήτρη Φουσκάκη για την μεγάλη βοήθεια που μου παρείχε καθ' όλη την διάρκεια εκπόνησης της διπλωματικής μου. Πολλά ευχαριστώ επίσης και στον Καθηγητή Ματθαίο Φαλάγγα, Διευθυντή του Άλφα Ινστιτούτου Βιοϊατρικών Επιστημών (AIBE) για την βοήθεια του, τόσο στην εύρεση της κατάλληλης βιογραφίας, όσο και για την θέση της πρακτικής άσκησης που μου προσέφερε στο Ινστιτούτο που διευθύνει πάνω στην ανάλυση στατιστικών δεδομένων. Τέλος θα ήθελα να ευχαριστήσω τον Αναπληρωτή Καθηγητή Ιωάννη Σπηλιώτη και τον Καθηγητή Γεώργιο Κοκολάκη, για την συμμετοχή τους στην τριμελή επιτροπή αξιολόγησης της διπλωματικής μου εργασίας.

Αφιερωμένο στην οικογένεια μου και στους ελευθεριακά σκεπτόμενους ανθρώπους.

Abstract

Information criteria are used in finding the most suitable model for the analysis or prediction of data, since they balance the explanatory power of the candidate models against the number of their parameters. Various information criteria have been developed over the years. In this thesis, an outline of the proof of the Akaike Information Criterion (AIC) is presented, thus we demonstrate its main characteristics. AIC's extensions (TIC, AICc, WleAIC) are also described since they aim at eliminating some of the latter, as well as maximum likelihood estimators and Kullback – Leibler distance.

All of these information criteria are then compared with Bayesian information criterion (BIC). Specifically the weak consistency, strong consistency, consistency and efficiency of these criteria were compared through concrete theorems. It was concluded that both AIC and BIC manage to choose the model that is closer to the true distribution of data according to the Kullback – Leibler distance as $n \to \infty$. In addition, AIC manages to choose the model also chosen through minimization of squared errors as $n \to \infty$ (efficiency), whereas BIC manages to choose the most parsimonious model from those that minimize the Kullback – Leibler distance between the adjusted and true distribution of data (consistency).

Πρόλογος

Η παρούσα Διπλωματική εργασία εκπονήθηκε στα πλαίσια των σπουδών μου για την απόκτηση του προπτυχιακού διπλώματος της σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Εφαρμογών (Σ.Ε.Μ.Φ.Ε.) του Εθνικού Μετσόβιου Πολυτεχνείου (Ε.Μ.Π.). Σκοπός της είναι η παρουσίαση του Akaike κριτηρίου πληροφορίας (AIC) και κάποιων επεκτάσεών του, TIC, AICc, WleAIC, καθώς και σύγκρισή του με το Μπεϋζιανό κριτήριο πληροφορίας (BIC) ως προς την συνέπεια και αποδοτικότητά του.

Η εργασία εκτείνεται σε τρείς ενότητες και 5 κεφάλαια. Στην πρώτη ενότητα ανήκουν τα πρώτα δύο κεφάλαια. Στο $1^{ο}$ Κεφάλαιο 'Εισαγωγικό Σημείωμα', αναφερόμαστε στο λόγο ύπαρξης των κριτηρίων πληροφορίας ο οποίος είναι η εύρεση του καταλληλότερου μοντέλου ανάλυσης ή πρόβλεψης των δεδομένων και η δημιουργία του στατιστικού μοντέλου ανάλυσης ή πρόβλεψης. Στο $2^{ο}$ κεφάλαιο 'Βασικές Έννοιες', παρουσιάζουμε το στατιστικό μοντέλο παλινδρόμησης μέσα από το γραμμικό μοντέλο παλινδρόμησης που είναι το πιο συχνά εφαρμόσιμο στατιστικό μοντέλο. Επίσης παραθέτουμε την μέθοδο ελαχίστων τετραγώνων, καθώς και τα μέτρα καταλληλότητας R^2 και $Cp-Mallows$.

Η $2^{η}$ ενότητα αποτελείται από το $3^{ο}$ και το $4^{ο}$ κεφάλαιο. Στο $3^{ο}$ κεφάλαιο 'Βασικά χαρακτηριστικά Akaike κριτηρίου πληροφορίας', παρουσιάζεται η γενική μορφή του AIC καθώς και τα κύρια εργαλεία κατασκευής του, οι εκτιμήτριες μέγιστης πιθανοφάνειας και η απόκλιση κατά Kullback – Leibler. Στο $4^{ο}$ κεφάλαιο 'Akaike κριτήριο πληροφορίας και οι επεκτάσεις του TIC, AICc, WleAIC', παρουσιάζουμε ένα σχεδιάγραμμα της απόδειξης του Akaike κριτήριου πληροφορίας καταδεικνύοντας έτσι τα πλεονεκτήματα και τα μειονεκτήματά του, καθώς και κάποια άλλα κριτήρια,TIC, AICc, WleAIC, τα οποία προσπαθούν να εξαλείψουν τα μειονεκτήματα αυτά.

Στην $3^{η}$ ενότητα ανήκει το $5^{ο}$ κεφάλαιο 'Μπεϋζιανό κριτήριο πληροφορίας (BIC)', παρουσιάζουμε μια σύγκριση του Akaike κριτήριου πληροφορίας με το Μπεϋζιανό κριτήριο πληροφορίας με βάση την συνέπεια και την αποδοτικότητά τους μέσα από αντίστοιχα θεωρήματα.

Προσπαθήσαμε καθ' όλη την έκταση της διπλωματικής να αναφέρουμε όσο το δυνατόν περισσότερα παραδείγματα έτσι ώστε να τονίσουμε τον εφαρμοσμένο ρόλο του θέματος. Τέλος για την ανάλυση των δεδομένων στα παραδείγματα χρησιμοποιήσαμε το ελεύθερο πρόγραμμα στατικής ανάλυσης R, έκδοση 2.12.0.

Περιεχόμενα

Κεφάλαιο Ι

Εισαγωγικό Σημείωμα

1.1. Εισαγωγή

Έχοντας στα χέρια μας ένα σύνολο δεδομένων μπορούμε εύκολα στις μέρες μας, με το πάτημα ενός κουμπιού να εφαρμόσουμε χιλιάδες μοντέλα. Πώς όμως επιλέγουμε το καλύτερο δυνατό; Με ποιά κριτήρια κατατάσσουμε τα διάφορα μοντέλα; Ο κίνδυνος της υπερπροσαρμογής (overfitting) ελλοχεύει πίσω από την επιλογή του κατάλληλου μοντέλου. Ένα επιμέρους πρόβλημα της επιλογής μοντέλου, είναι η επιλογή των μεταβλητών που θα εισάγουμε στο μοντέλο αυτό.

Τις τελευταίες 2 δεκαετίες έχει επέλθει ραγδαία πρόοδος τόσο στην πρακτική ικανότητα μας να προσαρμόζουμε μοντέλα όσο και στην θεωρητική κατανόηση της ανάλυσης αυτών των μοντέλων. Σε αυτή τη διπλωματική θα προσπαθήσουμε να αναλύσουμε το Akaike κριτήριο πληροφορίας και να το συγκρίνουμε με το Bayesian κριτήριο πληροφορίας (BIC). Η επιλογή μοντέλου είναι κάτι παραπάνω από την επιλογή των κατάλληλων μεταβλητών που θα εισαχθούν σε ένα μοντέλο παλινδρόμησης. Συγκεκριμένα, αποτελεί το δεύτερο επίπεδο από μια στρατηγική δύο επιπέδων.

Το πρώτο επίπεδο είναι η συλλογή των δεδομένων. Ο κλάδος της στατιστικής που ασχολείται με αυτό το επίπεδο ονομάζεται Στατιστική Δειγματοληψία. Αυτό που δεν πρέπει ποτέ να ξεχνάμε είναι ότι υπάρχει μια αβεβαιότητα (σφάλμα) από το πρώτο στάδιο, που μετακυλίεται στο δεύτερο, με αποτέλεσμα μικρές τροποποιήσεις στα δεδομένα να επηρεάζουν την επιλογή κατάλληλου μοντέλου επεξεργασίας και ανάλυσης των δεδομένων στο δεύτερο επίπεδο. Αυτή η αβεβαιότητα έγκειται στο πώς έχει δομηθεί η συλλογή των δεδομένων. Ένα απλό παράδειγμα διαφορετικής δομής είναι η συλλογή των δεδομένων τηλεφωνικά ή με προσωπική συνέντευξη.

1.2. Επιλογή Μοντέλου

Με τον όρο επιλογή μοντέλων προσπαθούμε να παντρέψουμε δύο αλληλοσυγκρουόμενα συμφέροντα. Πρώτον, την όσο τον δυνατόν καλύτερη προσαρμογή του μοντέλου στα δεδομένα, όπου συνήθως επιτυγχάνεται με την εισαγωγή μεταβλητών στο μοντέλο, και

δεύτερον την μείωση της πολυπλοκότητας του μοντέλου έτσι ώστε να είναι εύκολα παρουσιάσιμο και ερμηνεύσιμο, πράγμα που επιτυγχάνεται με την φειδωλή χρήση μεταβλητών.

Τα μοντέλα που πραγματευόμαστε και προσπαθούμε να εφαρμόσουμε ή και να κατασκευάσουμε από τα δεδομένα είναι προσεγγιστικά του πραγματικού μοντέλου που ακολουθεί ο πληθυσμός μας. Στις περισσότερες των περιπτώσεων μάλιστα, το πραγματικό μοντέλο δεν είναι καν γνωστό. Τα μοντέλα μας άλλωστε, δεν θα μπορούσαν να είναι παρά μια προσέγγιση και από τον προφανή λόγο ότι αναπτύσσονται με βάση ένα συνήθως μικρό δείγμα του πληθυσμού μας. Η προσεγγιστική αυτή φύση των μοντέλων μας είναι που δημιουργεί την αλληλοσυγκρουόμενη σχέση μεταξύ της διασποράς της μεταβλητής απόκρισης και της μεροληψία των επεξηγηματικών μεταβλητών.

Η αλληλοσυγκρουόμενη σχέση μεταξύ διασποράς και μεροληψίας έγκειται στον παράγοντα που πυροδοτεί τα δύο αυτά μέτρα. Η μεταβλητότητα και άρα η διασπορά της μεταβλητής απόκρισης, μειώνεται όταν έχουμε μικρό αριθμό παραμέτρων προς εκτίμηση, ενώ η μεροληψία μοντελοποίησης (modeling bias) αυξάνεται. Από την άλλη πλευρά, προσθέτοντας παραμέτρους στο μοντέλο παλινδρόμησης αυξάνουμε τον βαθμό μεταβλητότητας αλλά μειώνουμε την μεροληψία. Η μεγάλη μεροληψία του μοντέλου μας το καθιστά πιο «άκαμπτο» πράγμα που δεν βοηθάει στην γενίκευση του μοντέλου σε όλο τον πληθυσμό. Η αυξημένη διασπορά από την άλλη πλευρά, δε μας βοηθάει στην εξαγωγή χρήσιμων συμπερασμάτων καθώς κατακερματίζει το δείγμα μας σε πολλά κομμάτια. Η «διαμάχη» μεροληψίας – μεταβλητότητας αντικατοπτρίζει το φιλοσοφικό ερώτημα που άπτεται όλων των επιστημών, απλότητα ή πολυπλοκότητα; Την αρχαία ρήση «παν μέτρον άριστον» προσπαθούμε να εφαρμόσουμε μέσω της κατάλληλης επιλογής μοντέλου καθώς προσπαθούμε να συνδυάσουμε την υπερπροσαρμογή (overfitting) με την υποπροσαρμογή (unterfitting).

Η αρχή της *φειδωλότητας* μεταφράζεται στη στατιστική ως ο κανόνα ότι μόνο οι παράμετροι που πραγματικά χρειάζονται πρέπει να εισαχθούν σε ένα μοντέλο, έτσι ώστε να αποφεύγονται παράμετροι οι οποίοι περισσότερο περιπλέκουν, προσθέτοντας «θόρυβο», παρά βοηθούν τον ερευνητή στην εξαγωγή χρήσιμων συμπερασμάτων.

Με τον όρο *πλαίσιο* (the context) εννοούμε την όλη διαδικασία της στατιστικής έρευνας που αλλάζει από ερευνητή σε ερευνητή. Για παράδειγμα η άποψη πλαίσιο, ο λόγος δημιουργίας ενός μοντέλου δεν είναι να ταιριάζει στα δεδομένα αλλά να απαντάει στα ερωτήματα για τα οποία δημιουργήθηκε, αποτελεί μια διαφορετική σκοπιά σε σχέση με αυτούς που πιστεύουν ότι, προσαρμόζοντας κατάλληλα ένα μοντέλο στα δεδομένα αντλείς τις πληροφορίες που χρειάζεσαι από αυτό. Όσον αφορά τα μοντέλα πρόβλεψης ο Akaike (1969) είχε την άποψη ότι «ο στόχος της στατιστικής μοντελοποίησης δεν πρέπει να είναι η ακριβής περιγραφή των συγκεκριμένων δεδομένων ή της κατανομής από την οποία προέρχονται αλλά η πρόβλεψη των μελλοντικών δεδομένων με όσο το

δυνατόν μεγαλύτερη ακρίβεια». Αυτό επικράτησε να αναφέρεται στη βιβλιογραφία ως *predictive point of view*. Μπορεί να μην υπάρχει διαφορά μεταξύ των δύο εκδοχών όταν έχουμε να κάνουμε με ένα πολύ μεγάλο αριθμό δεδομένων ή με δεδομένα που δεν έχουν «θόρυβο», παρ' όλα αυτά όταν η μοντελοποίηση βασίζεται σε μια πεπερασμένη ποσότητα πραγματικών δεδομένων, όπως στις περισσότερες περιπτώσεις, υπάρχει ένα σημαντικό χάσμα μεταξύ των δύο αυτών εκδοχών. Όπως θα δούμε και παρακάτω μέσα από τα κριτήρια πληροφορίας, μοντέλα που σκοπός τους είναι η πρόβλεψη, ακόμα και απλά μοντέλα που περιέχουν αρκετά μεγάλη μεροληψία, είναι συχνά καλύτερα ως προς την ικανότητα πρόβλεψης, από τα μοντέλα όπου στόχος τους είναι να εκτιμήσουν την «πραγματική» κατανομή.

Η *στάθμιση μοντέλων* αποτελεί ένα διαφορετικό τρόπο επιλογής μοντέλου καθώς δεν επιλέγει το καλύτερο μοντέλο άλλα προσπαθεί να συνδυάσει τα καλύτερα μοντέλα. Τα κριτήρια πληροφορίας όπως θα δούμε και παρακάτω, αξιολογούν και βαθμολογούν τα διάφορα μοντέλα καταλήγοντας έτσι να επιλέξουμε κάποιο από αυτό. Τι γίνετε όμως όταν ένας αριθμός μοντέλων βρίσκεται αρκετά κοντά στην κορυφή και η επιλογή ενός και μοναδικού μοντέλου είναι δύσκολη; Σε τέτοιες περιπτώσεις μέθοδοι πάνω στη στάθμιση μοντέλων μπορούν να προσφέρουν καταλληλότερες λύσεις.

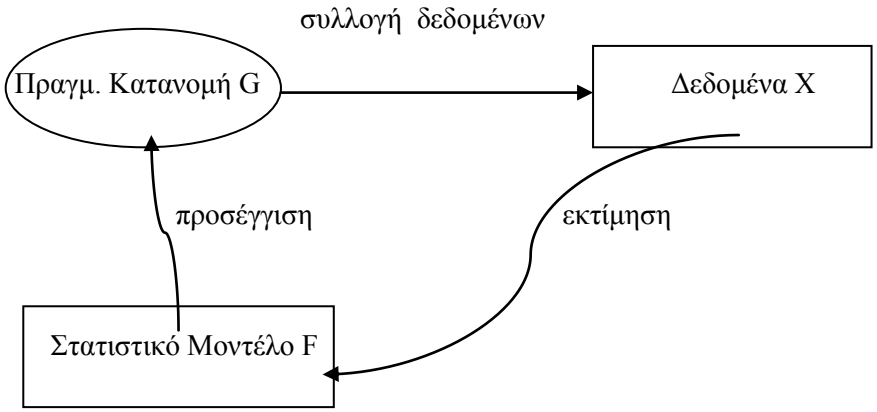
1.3. **Στατιστικό Μοντέλο**

Ένα στατιστικό μοντέλο μπορεί να θεωρηθεί μια κατανομή πιθανότητας που χρησιμοποιεί τα δεδομένα που παρατηρήθηκαν (δείγμα) έτσι ώστε να προσεγγίζει την πραγματική κατανομή του πληθυσμού από την οποία προέρχεται το δείγμα. Για παράδειγμα παίρνουμε το ύψος 20 φοιτητών συγκεκριμένου έτους και εκτιμάμε την κατανομή που μπορεί να ακολουθεί το ύψος των φοιτητών αυτού του έτους ή προσπαθούμε να προβλέψουμε τι ύψος θα έχουν οι φοιτητές που θα έρθουν το επόμενο έτος. Άρα ο σκοπός του στατιστικού μοντέλου είναι να περιγράψει προσεγγιστικά όσο καλύτερα γίνετε τη δομή των δεδομένων ή να προβλέψει μελλοντικά δεδομένα του ίδιου πληθυσμού παίρνοντας πληροφορίες από ένα συγκεκριμένο δείγμα.

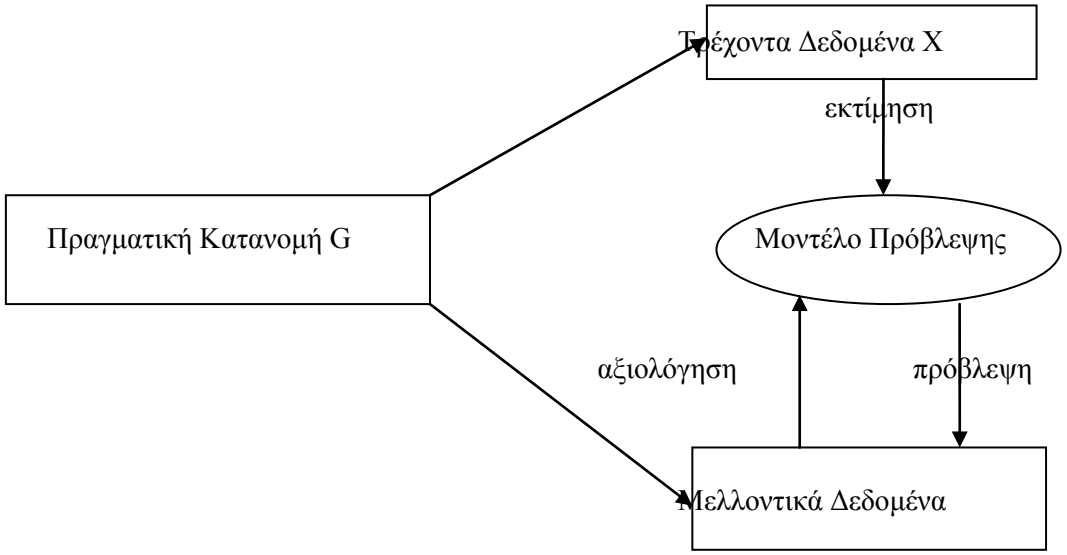
Η εύρεση του κατάλληλου μοντέλου αποτελεί το κύριο ζητούμενο της ανάλυσης δεδομένων. Για παράδειγμα, εφαρμόζοντας ένα γραμμικό μοντέλο παλινδρόμησης, η βασική ερώτηση είναι ποιό υποσύνολο των επεξηγηματικών μεταβλητών θα χρησιμοποιηθεί, ενώ εφαρμόζοντας ένα πολυωνυμικό μοντέλο παλινδρόμησης, η βασική ερώτηση είναι ποιά θα είναι η επιλογή της σειράς των επεξηγηματικών μεταβλητών που θα χρησιμοποιηθούν.

Στο Διάγραμμα 1.1 περιγράφεται η διαδικασία δημιουργίας του στατιστικού μοντέλου. Αρχικά συλλέγουμε τις παρατηρήσεις μας οι οποίες αποτελούν τα δεδομένα μας. Στη συνέχεια προσαρμόζουμε στα δεδομένα ένα κατάλληλο μοντέλο και υπολογίζουμε τους συντελεστές του με κάποια μέθοδο, όπως η μέθοδος ελαχίστων τετραγώνων. Το μοντέλο

που προσαρμόσαμε μαζί με τους συντελεστές που προσδιορίσαμε είναι το στατιστικό μοντέλο, το οποίο αποτελεί μια προσέγγιση του πραγματικού μοντέλου. Στο Διάγραμμα 1.2 βλέπουμε πώς δημιουργείται ένα στατιστικό μοντέλο πρόβλεψης. Αυτό που αξίζει να σχολιάσουμε είναι η αξιολόγηση του παραγόμενου μοντέλου πρόβλεψης μέσα από τα ίδια τα παράγωγά του. Αυτό επιτυγχάνεται με το να εξετάσουμε σε ποιο βαθμό τα μελλοντικά δεδομένα που παράγει το μοντέλο μας προέρχονται από την πραγματική κατανομή G.



Διάγραμμα 1.1: Διαδικασία Μοντελοποίησης



Διάγραμμα 1.2: Δημιουργία στατιστικού μοντέλου πρόβλεψης.

1.4. **Δεσμευμένα Μοντέλα Κατανομής.**

Αν η κατανομή μιας τυχαίας μεταβλητής (τ.μ.) Y καθορίζεται με τέτοιο τρόπο έτσι ώστε να εξαρτάται από μια διάστασης p μεταβλητή $\vec{X} = (X_1, X_2, ..., X_p)^T$, τότε η κατανομή της τ.μ. Y εκφράζεται ως $F(Y \mid \vec{X})$ και αυτό καλείται *μοντέλο δεσμευμένης* ή *υπό συνθήκη κατανομή*. Υπάρχουν διάφοροι τρόποι με τους οποίους μια τυχαία μεταβλητή εξαρτάται από άλλες μεταβλητές. Ένας από αυτούς τους τρόπους είναι και η γραμμική παλινδρόμηση.

Κεφάλαιο ΙΙ

Βασικές Έννοιες

2.1. Μοντέλο παλινδρόμησης

Το μοντέλο παλινδρόμησης χρησιμοποιείται έτσι ώστε να μοντελοποιηθεί η σχέση μεταξύ μιας μεταβλητής απόκρισης Y και διαφόρων επεξηγηματικών μεταβλητών $\vec{X} = (X_1, X_2, ..., X_p)^T$. Αυτό είναι ισοδύναμο με την παραδοχή ότι η κατανομή πιθανότητας της μεταβλητής απόκρισης Y εξαρτάται από τις επεξηγηματικές μεταβλητές $\vec{X}$. Τότε η δεσμευμένη κατανομή δίνεται υπό την μορφή $F(Y \mid \vec{X})$.

Έστω $\{(y_i, \vec{x}_i);\ i = 1,2,...,n\}$ να είναι n ζεύγη από παρατηρούμενα δεδομένα της μεταβλητής απόκρισης Y και της διαστατής p μεταβλητής $\vec{X}$. Τότε το μοντέλο

$$Y_i = u(\vec{X}_i) + \varepsilon_i \text{ με } i = 1,2,...,n \tag{1.1}$$

των παρατηρούμενων δεδομένων καλείται *μοντέλο παλινδρόμησης* όπου

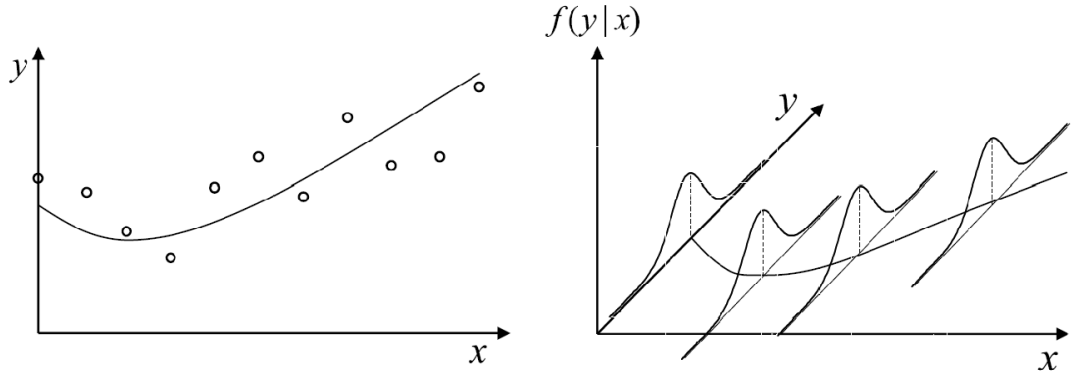
1. $u(\vec{X}_i)$ είναι μια συνάρτηση των επεξηγηματικών μεταβλητών $\vec{X}$.
2. ε_i είναι ο όρος του σφάλματος ή «θόρυβος», ο οποίος υποθέτουμε ότι είναι ανεξάρτητα κατανεμημένος με $E[\varepsilon_i] = 0$ και διασπορά $V[\varepsilon_i] = \sigma^2$. Συχνά υποθέτουμε ότι το ε_i είναι κανονικά κατανεμημένο, δηλαδή ότι $\varepsilon_i \sim N(0, \sigma^2)$.

Έστω ότι $\varepsilon_i \sim N(0, \sigma^2)$. Τότε $Y_i \mid \vec{X}_i = \vec{x}_i \sim N(u(\vec{X}_i), \sigma^2)$. Δηλαδή το $Y_i \mid \vec{X}_i = \vec{x}_i$ ακολουθεί κανονική κατανομή με μέση τιμή $u(\vec{X}_i)$ και διασπορά σ^2. Η συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) δίνεται από τον τύπο

$$f(y_i \mid \vec{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{\left(y_i - u(\vec{x}_i)\right)^2}{2\sigma^2} \right\} \text{ με } i = 1,2,...,n. \qquad (1.2)$$

Η κατανομή της Εξίσωσης 1.2 είναι μια δεσμευμένη κατανομή της οποίας ο μέσος είναι του τύπου $E(Y \mid \vec{X} = \vec{x}) = u(\vec{x})$ καθώς η Y εξαρτάται από τις τιμές των επεξηγηματικών μεταβλητών.

Παρακάτω παραθέτουμε το Διάγραμμα 2.1 όπου μας βοηθάει να κατανοήσουμε τις δύο μορφές της παλινδρόμησης.



Διάγραμμα 2.1: Μοντέλο παλινδρόμησης (αριστερά) και μοντέλο δεσμευμένης κατανομής (δεξιά) στο οποίο ο μέσος της μεταβλητής απόκρισης είναι συναρτήσει της επεξηγηματικής μεταβλητής X.

Η αριστερή γραφική παράσταση στο Διάγραμμα 2.1 δείχνει 11 παρατηρήσεις και την μέση συνάρτηση $u(X)$ της μονοδιάστατης μεταβλητής X με την μεταβλητή απόκρισης Y. Η τιμή y_i της μεταβλητής απόκρισης Y στο σημείο x_i παρατηρείται ως

$$y_i = \mu_i + \varepsilon_i \qquad \text{με} \qquad i = 1,2,...,n,$$

με μέση τιμή $E[Y_i \mid X_i] = \mu_i$ και θόρυβο ε_i. Η ποσότητα $u(X)$ αντιπροσωπεύει την μέση δομή του φαινομένου, ενώ το ε_i είναι ο θόρυβος που προκαλεί η διακύμανση των y_i δεδομένων.

Η δεξιά γραφική παράσταση στο Διάγραμμα 2.1 δείχνει μια δεσμευμένη κατανομή που εκτιμάται χρησιμοποιώντας ένα μοντέλο παλινδρόμησης. Για δεδομένη τιμή μιας επεξηγηματικής μεταβλητής X, η κατανομή πιθανότητας είναι $f(Y \mid X)$ για την οποία ο μέσος είναι $u(X)$. Γι' αυτό το μοντέλο παλινδρόμησης την Εξίσωση 1.2 ορίζεται ως μια κλάση κατανομών σύμφωνα με την τιμή της X.

2.2. **Γραμμικό μοντέλο παλινδρόμησης**

Αν η συνάρτηση παλινδρόμησης ή η συνάρτηση μέσου $u(\vec{X})$ μπορεί να προσεγγιστεί από μια γραμμική συνάρτηση των $\vec{X}$, τότε το μοντέλο μπορεί να εκφραστεί ως

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \varepsilon_i \quad \text{για} \quad i = 1,2,...,n.$$

Γράφοντας την παραπάνω εξίσωση υπό μορφή πινάκων έχουμε

$$Y_i = \vec{\beta}\vec{X}_i + \varepsilon_i \text{ για } i = 1,2,...,n,$$

όπου $\vec{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^{\mathrm{T}}$ και $\vec{X}_i = (1, X_{i1}, X_{i2}, ..., X_{ip})$.

Το παραπάνω μοντέλο ονομάζεται *γραμμικό μοντέλο* και είναι το μοντέλο παλινδρόμησης που χρησιμοποιείται περισσότερο.

Ένα γραμμικό μοντέλο παλινδρόμησης με ε_i κανονικά κατανεμημένο ($\varepsilon_i \sim N(0,\sigma^2)$) έχει σ.π.π.

$$f(y_i \mid x_i;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{\left(y_i - \vec{\beta}\vec{x}_i\right)^2}{2\sigma^2} \right\} \text{ με } i = 1,2,...,n, \qquad (1.3)$$

όπου οι άγνωστοι παράμετροι του μοντέλου είναι $\vec{\theta} = (\vec{\beta}, \sigma^2)$.

Ένα γραμμικό μοντέλο ονομάζεται *απλό γραμμικό μοντέλο* όταν έχει μια μόνο επεξηγηματική μεταβλητή, ενώ όταν έχει περισσότερες ονομάζεται *γενικό ή πολλαπλό γραμμικό μοντέλο*.

Ένα από τα κύρια ζητήματα στο γραμμικό μοντέλο είναι να καθορίσουμε το σύνολο των επεξηγηματικών μεταβλητών που θα εισαχθούν στο μοντέλο έτσι ώστε να εξηγήσουμε καλύτερα τις αλλαγές στην κατανομή της μεταβλητής απόκρισης. Το πρόβλημα αυτό αναφέρεται στη βιβλιογραφία ως *πρόβλημα επιλογής μεταβλητών* (variable selection problem). Όπως θα δούμε και παρακάτω, ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι χρησιμοποιώντας τα κριτήρια πληροφορίας.

Εκτός από το απλό και το γενικό γραμμικό μοντέλο υπάρχουν και άλλα είδη μοντέλων που μπορούμε να χρησιμοποιήσουμε, όπως το πολυωνυμικό μοντέλο παλινδρόμησης, το μη γραμμικό μοντέλο παλινδρόμησης κ.α.

2.3. **Μέθοδος Ελαχίστων Τετραγώνων**

Αφού έχουμε επιλέξει το γραμμικό μοντέλο ως κατάλληλο μοντέλο, αυτό που καλούμαστε τώρα να κάνουμε είναι να εκτιμήσουμε τις παραμέτρους του έτσι ώστε να προσαρμόζεται κατάλληλα τα δεδομένα μας. Πρακτικά καλούμαστε να προσαρμόσουμε μια ευθεία στα δεδομένα μας εκτιμώντας τους συντελεστές αυτής της ευθείας μέσω των δεδομένων μας. Αυτοί οι συντελεστές λέγονται *συντελεστές παλινδρόμησης* και μια από τις μεθόδους υπολογισμού τους είναι η μέθοδος ελαχίστων τετραγώνων.

Η συγκεκριμένη μέθοδος έγκειται στην ελαχιστοποίηση της παράστασης

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

ως προς β_0, β_1 με (x_i, y_i) $i = 1,2,...,n$ ζεύγη παρατηρήσεων.

Παραγωγίζοντας την παραπάνω εξίσωση ως προς β_0, β_1 και εξισώνοντας τις δύο εξισώσεις με το μηδέν έχουμε ένα ζεύγος εξισώσεων, τις *κανονικές εξισώσεις* από τις οποίες λαμβάνουμε τις εκτιμήσεις $\hat{\beta}_0, \hat{\beta}_1$ των β_0, β_1.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{X})^2} \qquad \text{κ΄} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

όπου $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ και $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

Από τις δεύτερες παράγωγους του S^2 επιβεβαιώνεται ότι πράγματι στα σημεία που προκύπτουν από τις κανονικές εξισώσεις το S^2 παρουσιάζει ελάχιστο.

Παρακάτω παραθέτουμε κάποιες συναρτήσεις που θα μας βοηθήσουν στον ορισμό των μέτρων καταλληλότητας.

$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$, που είναι το άθροισμα των τετραγώνων παλινδρόμησης. (1.4)

$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$, που είναι το άθροισμα τετραγώνων των υπολοίπων. (1.5)

$SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2$, που είναι το ολικό άθροισμα τετραγώνων. (1.6)

Η σχέση που συνδέει τις τρεις αυτές ποσότητες εύκολα αποδεικνύεται ότι είναι $SST = SSR + SSE$.

2.4. Μέτρα καταλληλότητας.

2.4.1. Συντελεστής προσδιορισμού R^2

Ένας τρόπος για να κρίνουμε την προσαρμογή ή καταλληλότητα του απλού γραμμικού μοντέλου είναι η ποσότητα

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

που λέγετε *συντελεστής προσδιορισμού* και εκφράζει το ποσοστό της μεταβλητότητας της τ.μ. Y που εξηγείται από την $\vec{X}$. Το πεδίο τιμών του είναι $[0,1]$ ή $[0,100]$ όταν παρουσιάζεται ως ποσοστό, καθώς $SST \geq SSR$ όπως φαίνεται και από τους παραπάνω τύπους (1.4, 1.5, 1.6). Γενικά όσο πιο κοντά στο 1 και αντίστοιχα στο 100, είναι η τιμή του συντελεστή προσδιορισμού τόσο ισχυρότερη είναι η γραμμική σχέση εξάρτησης των Y και X.

Στο γενικό γραμμικό μοντέλο ο συντελεστής προσδιορισμού μπορεί να χρησιμεύσει και για την εισαγωγή ή την εξαγωγή μιας μεταβλητής στο μοντέλο. Αν για παράδειγμα, με την εισαγωγή μιας μεταβλητής ο συντελεστής αυξάνει σημαντικά, αυτό μπορεί να εκληφθεί θετικά ως προς την εισαγωγή αυτής της μεταβλητής στο μοντέλο μας. Εκτός του συντελεστή προσδιορισμού, στο πολλαπλό γραμμικό μοντέλο χρησιμοποιείται και ο *προσαρμοσμένος συντελεστής προσδιορισμού* όπου δεν είναι τίποτα άλλο από τον συντελεστή προσδιορισμού με την διαφορά ότι κάθε άθροισμα είναι διαιρεμένο με τους βαθμούς ελευθερίας του. Οι βαθμοί ελευθερίας ενός αθροίσματος ισοδυναμούν με το πλήθος των ανεξάρτητων εξισώσεων που χρειάζονται για τον υπολογισμό του εν λόγω αθροίσματος. Έτσι έχουμε

$$R_{adj}^2 = 1 - \frac{MSE}{MST}, \text{ όπου } MSE = \frac{1}{n - p - 1} SSE \text{ και } MST = \frac{1}{n-1} SST.$$

Ο προσαρμοσμένος μέσος χρησιμοποιείται περισσότερο όταν ο αριθμός των παρατηρήσεων είναι πολύ κοντά με τον αριθμό των μεταβλητών και έτσι ο απλός συντελεστής προσαρμογής δεν είναι αξιόπιστος. Το πλεονέκτημα του προσαρμοσμένου

συντελεστή προσαρμογής είναι ότι δεν εξαρτάται από τον αριθμό των επεξηγηματικών μεταβλητών.

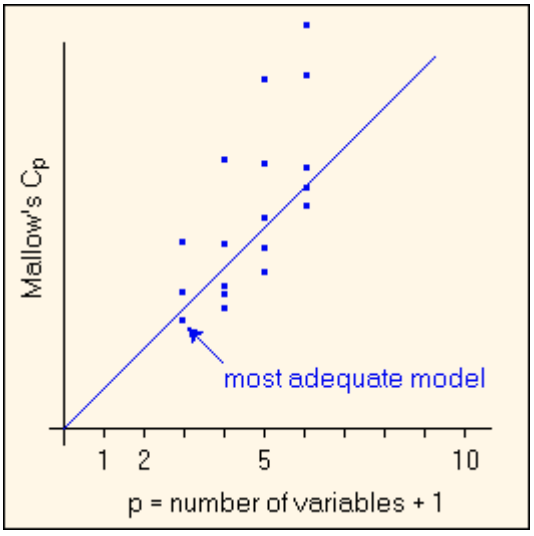
2.4.2. **Στατιστική Συνάρτηση** $C_p - Mallows$

Ένα άλλο μέτρο για την αξιολόγηση της καταλληλότητας του μοντέλου είναι η στατιστική συνάρτηση $C_p - Mallows$

$$C_p = \frac{SSE_\kappa}{S^2} + 2p - n,$$

όπου p είναι ο αριθμός των επεξηγηματικών μεταβλητών στο μοντέλο και n ο αριθμός των παρατηρήσεων. Ο όρος SSE_κ συμβολίζει το άθροισμα τετραγώνων των υπολοίπων για το μοντέλο που περιλαμβάνει κ μεταβλητές από τις p (κ≤p) που έχουμε στη διάθεση μας, ενώ ο S^2 είναι το μέσο τετραγωνικό υπόλοιπο όταν χρησιμοποιούνται και οι p υποψήφιες μεταβλητές.

Ο τρόπος για να επιλέξουμε το καταλληλότερο μοντέλο χρησιμοποιώντας τη στατιστική συνάρτηση $C_p - Mallows$ είναι να υπολογιστεί το C_p για όλα τα μοντέλα με τους δυνατούς συνδυασμούς των υποψήφιων μεταβλητών και να κάνουμε την γραφική παράσταση του αριθμού των επεξηγηματικών μεταβλητών προς την τιμή C_p. Το σημείο με την μικρότερη τιμή του C_p (όπως φαίνεται και στο Σχήμα 2.2) είναι το καταλληλότερο μοντέλο. Στην περίπτωση όπου υπάρχει ισοβαθμία μεταξύ κάποιων τιμών επιλέγουμε αυτή που είναι πιο κοντά στην ευθεία $y = x$.



Σχήμα 2.2: Επιλογή καταλληλότερου μοντέλου με βάση το μέτρο καταλληλότητας $C_p - Mallows$.

Κεφάλαιο III

Βασικά Χαρακτηριστικά Akaike Κριτήριου Πληροφορίας (AIC)

3.1. Εισαγωγή

Για την επιλογή ενός μοντέλου μεταξύ υποψήφιων μοντέλων, το AIC (Akaike Information Criterion) αποτελεί μια από τις πιο αξιόπιστες, δημοφιλείς και εύκολα εφαρμόσιμες στρατηγικές επιλογής μοντέλου (model selection). Η κύρια ιδέα του έγγυτε στην κατάλληλη ποινικοποίηση της μέγιστης πιθανοφάνειας κάθε μοντέλου με σκοπό την επιλογή του καταλληλότερου. Το AIC είναι ένα κριτήριο πληροφορίας. Με τον όρο κριτήριο πληροφορίας εννοούμε έναν μηχανισμό που χρησιμοποιεί τα δεδομένα ενός προβλήματος δίνοντας σε κάθε υποψήφιο μοντέλο μια τιμή που το χαρακτηρίζει.

3.2. Γενική Μορφή του AIC (Akaike Information Criterion)

Η γενική μορφή του Akaike κριτηρίου πληροφορίας είναι

$$AIC(M) = 2\log.likelihood_{\max}(M) - 2\dim(M) \qquad (3.1)$$

για κάθε υποψήφιο μοντέλο M, όπου dim(M) είναι η διάσταση του παραμετρικού χώρου του μοντέλου M.

Σχολιάζοντας αρχικά την μορφή του AIC, θα λέγαμε ότι προσπαθεί να «παντρέψει» την καλή προσαρμογή του μοντέλου (goodness of fit), μέσα από τον υπολογισμό της μέγιστης πιθανοφάνειας του $\left(2\log.likelihood_{\max}(M)\right)$, με την όσο το δυνατό μικρότερη πολυπλοκότητά του (least complexity), μέσα από τον όρο ποινικοποίησης $2\dim(M)$.

Η ανάγκη ποινικοποίησης της μέγιστης πιθανοφάνειας εξυπηρετεί την αρχή της φειδωλότητας που πρέπει να διέπει μια στρατηγική επιλογής μοντέλου όπως αναφέραμε και στο Κεφάλαιο Ι. Εάν χρησιμοποιούσαμε σαν μέτρο σύγκρισης των μοντέλων μόνο τη μέγιστη πιθανοφάνεια, μοιραία θα επιλέγαμε το πολυπλοκότερο μοντέλο ως καταλληλότερο, το οποίο συχνά είναι δύσκολο τόσο να προσαρμοστεί στα δεδομένα, όσο και να εξαχθούν από αυτό χρήσιμα συμπεράσματα.

Το AIC αποτελεί μια από τις πιο γενικές στρατηγικές ποινικοποίησης έχοντας εφαρμογές σε οποιαδήποτε τομέα της στατιστικής χρειάζεται σύγκριση μοντέλων. Πολλά από τα σύγχρονα προγράμματα στατιστικής χρησιμοποιούν το AIC για την επιλογή του κατάλληλου μοντέλου, τόσο σε απλά προβλήματα ανεξάρτητων και ισόνομων κατανεμημένων δεδομένων, όσο και σε πιο πολύπλοκα μοντέλα, όπως αυτά με χρονοσειρές ή τα παραμετρικά μοντέλα διακινδύνευσης που χρησιμοποιούνται στην ανάλυση επιβίωσης.

Τα προτερήματα αλλά και τα μειονεκτήματα του AIC, καθώς και όλων των κριτηρίων πληροφορίας που χρησιμοποιούν τη μέγιστη πιθανοφάνεια, απορρέουν από την συσχέτιση του AIC με τις *εκτιμήτριες μέγιστης πιθανοφάνειας* και την *ποσότητα πληροφορίας Kullback-Leibler*.

3.3. **Εκτιμήτριες μέγιστης πιθανοφάνειας**

Οι εκτιμήτριες μέγιστης πιθανοφάνειας είναι εκείνες οι δειγματοσυναρτήσεις που μεγιστοποιούν την πιθανοφάνεια ενός μοντέλου σύμφωνα με τα δεδομένα μας. Τον τρόπο εύρεσης των εκτιμητριών μέγιστης πιθανοφάνειας θα προσπαθήσουμε να δούμε μέσα από δύο παραδείγματα.

Παράδειγμα 3.1: Εύρεση εκτιμητριών μέγιστης πιθανοφάνειας στο πολλαπλό γραμμικό μοντέλο παλινδρόμησης.

Έστω $\{y_i, x_{i1}, x_{i2}, ..., x_{ip}\}$, $i = 1,2,3...,n$ n σύνολα δεδομένων τα οποία έχουν παρατηρηθεί για την μεταβλητή απόκρισης Y και τις p επεξηγηματικές μεταβλητές $\{X_1, X_2, ..., X_p\}$. Για να περιγράψουμε τη σχέση μεταξύ των μεταβλητών μας υποθέτουμε το παρακάτω γενικό γραμμικό μοντέλο παλινδρόμησης,

$$Y_i = \vec{X}_i^t \vec{\beta} + \varepsilon_i,\ \varepsilon_i \sim \mathrm{N}(0, \sigma^2),\ i = 1,2,3,...,n,$$

όπου $\vec{X}_i = (1, X_{i1}, X_{i2}, ..., X_{ip})$ και $\vec{\beta} = (\beta_0, \beta_1, ..., \beta_p)^t$.

Η δεσμευμένη σ.π.π. των y_i είναι $f(y_i \mid x_i;\vec{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - x_i^t\vec{\beta})^2\right\}$.

Η αντίστοιχη λογαριθμική πιθανοφάνεια είναι ίση με

$$\begin{aligned} l(\vec{\theta}) = \log L(\vec{\theta}) &= \sum_{i=1}^{n} \log\ f(y_i \mid x_i;\vec{\theta}) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^t\vec{\beta})^2 \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - x\vec{\beta})^t(y - x\vec{\beta}), \end{aligned}$$

όπου $y = (y_1, y_2, \ldots y_n)^t$ και $x = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n)^t$. Εξισώνοντας τις μερικές παραγώγους του $l(\vec{\theta})$ ως προς τις παραμέτρους $\vec{\theta} = (\vec{\beta}, \sigma^2)$ με το 0, έχουμε τις κανονικές εξισώσεις,

$$\frac{\partial l(\vec{\theta})}{\partial \vec{\beta}} = \frac{1}{2\sigma^2}(-2x^t y + 2x^t x\vec{\beta}) = 0,$$

$$\frac{\partial l(\vec{\theta})}{\partial \sigma^2} = \frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - x\vec{\beta})^t(y - x\vec{\beta}) = 0.$$

Λύνοντας το σύστημα των παραπάνω εξισώσεων οι εκτιμήτριες μέγιστης πιθανοφάνειας $\hat{\vec{\theta}}_{ML} = (\hat{\vec{\beta}}, \hat{\sigma}^2)$ του $\vec{\theta}_{ML} = (\vec{\beta}, \sigma^2)$ θα δίνονται από τις δειγματοσυναρτήσεις,

$$\hat{\vec{\beta}} = (x^t x)^{-1} x^t \mathrm{y} \qquad \hat{\sigma}^2 = \frac{1}{n}(y - x\hat{\beta})^t(y - x\hat{\beta}).$$

Παράδειγμα 3.2: Εύρεση εκτιμητριών μέγιστης πιθανοφάνειας στο λογιστικό μοντέλο παλινδρόμησης.

Έστω $\{y_i, x_{i1}, x_{i2}, \ldots, x_{ip}\}$, $(i = 1,2,3,\ldots,n)$ n σύνολα δεδομένων τα οποία έχουν παρατηρηθεί για την μεταβλητή απόκρισης Y και τις p επεξηγηματικές μεταβλητές $\{X_1, X_2, \ldots, X_p\}$. Η μεταβλητή απόκρισης Y ακολουθεί την κατανομή Bernoulli με σ.π.π. $f(y; p) = p^y(1-p)^{1-y}$ για $y \in \{0,1\}$, όπου p η πιθανότητα να συμβεί ένα γεγονός. Για κάθε μια από τις παρατηρήσεις μεγέθους n, $Y_i = 1$ σημαίνει ότι το γεγονός συμβαίνει για την παρατήρηση i, αλλιώς $Y_i = 0$. Για να περιγράψουμε τη σχέση μεταξύ των μεταβλητών μας υποθέτουμε το παρακάτω λογιστικό μοντέλο παλινδρόμησης (logistic model) που έχει την μορφή,

$$\log it(p_i) = \beta_0 + \beta_1 \vec{X}_{i1} + \beta_2 \vec{X}_{i2} + ... + \beta_{ip} = \vec{X}_i^t \vec{\beta}$$

με $\vec{X}_i = (1, X_{i1}, X_{i2}, ..., X_{ip})$, $\vec{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ και $\log it(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$.

Συνεπώς,

$$E[Y_i] = P(Y_i = 1 \mid X_i) = p_i = \frac{\exp(X_i^t \vec{\beta})}{1 + \exp(X_i^t \vec{\beta})} \text{ για } i = 1, 2, 3, ..., n,$$

με $\vec{\beta}$ το διάνυσμα παραμέτρων διάστασης p. Η λογαριθμική πιθανοφάνεια του μοντέλου θα είναι

$$l(\vec{\beta}) = \log\{L(\vec{\beta})\} = \sum_{i=1}^{n} \log f(y_i, \vec{\beta}) = \sum_{i=1}^{n} \{y_i \log p_i + (1 - y_i) \log(1 - p_i)\} \Rightarrow$$

$$l(\vec{\theta}) = \sum_{i=1}^{n} \left[y_i x_i^t \vec{\beta} - \log\{1 + \exp(x_i^t \vec{\beta})\} \right].$$

Εξισώνοντας τις μερικές παραγώγους του $l(\vec{\beta})$ ως προς $\vec{\beta}$, με το 0, έχουμε τις κανονικές εξισώσεις. Το σύστημα των $p+1$ κανονικών εξισώσεων των $\vec{\beta}$ που προκύπτει δεν επιλύεται αναλυτικά. Έτσι, χρησιμοποιούμε την αριθμητική μέθοδο προσέγγισης Newton-Raphson για να πάρουμε τις εκτιμήτριες μέγιστης πιθανοφάνειας $\hat{\vec{\beta}}$ των $\vec{\beta}$.

Η αριθμητική προσεγγιστική μέθοδος Newton-Raphson εύρεσης των εκτιμητριών μέγιστης πιθανοφάνειας των συντελεστών λογιστικής παλινδρόμησης αποτελείται από τρία βήματα:

1. Επιλέγουμε μια αρχική εκτίμηση του συντελεστή παλινδρόμησης, όπως $\beta_0 = 0$.
2. Για κάθε δείκτη t, βρίσκουμε τον συντελεστή:

$$\beta_t = \beta_{t-1} + \left(\vec{X}_i^t \vec{V}_{t-1} \vec{X}_i \right)^{-1} X_i^t \left(\vec{Y} - \vec{p}_{t-1} \right),$$

όπου,

- $\vec{Y}$ είναι το διάνυσμα των μεταβλητών απόκρισης που παίρνε τιμές (0,1),
- p_{t-1} είναι το διάνυσμα των προσαρμοσμένων πιθανοτήτων απόκρισης (fitted response probabilities) για τον προηγούμενο δείκτη, του οποίου η i συντεταγμένη είναι $p_{i,t-1} = \frac{1}{1 + \exp\left(-X_i^t \beta_{t-1}\right)}$,

- V_{t-1} είναι ο διαγώνιος πίνακας με στοιχεία τα $p_{i,t-1}\left(1-p_{i,t-1}\right)$.

3. Το 2ο Βήμα επαναλαμβάνεται μέχρι το β_t να είναι αρκετά κοντά στο β_{t-1}. Ο ασυμπωτικά εκτιμώμενος πίνακας διασποράς των συντελεστών δίνεται από τον τύπο $\left(X_i^t V X_i\right)^{-1}$.

3.4. **Απόσταση κατά** K-L (Kullback-Leibler distance)

Σε οποιαδήποτε στατιστική ανάλυση μας δίνεται ένα σύνολο παρατηρήσεων (δείγμα). Αυτό το σύνολο παρατηρήσεων είναι τιμές τυχαίων μεταβλητών που συνήθως δε γνωρίζουμε τις κατανομές που ακολουθούν ή έχουμε ελλιπή γνώση για αυτές. Έτσι, προσπαθούμε μέσα από πληροφορίες που αντλούμε από τις παρατηρήσεις μας, να εξαγάγουμε χρήσιμα συμπεράσματα ως προς την κατανομή που ακολουθεί το δείγμα μας. Έχοντας εκτιμήσει την κατανομή ή κάποια στοιχεία αυτής όπως οι παράμετροι της, μπορούμε να έχουμε μια εικόνα για το προς τα πού θα κινηθεί το δείγμα μας σε μια ενδεχόμενη προσθήκη νέων παρατηρήσεων, έχοντας υιοθετήσει το μοντέλο πρόβλεψης.

Για παράδειγμα, έστω ένα δείγμα από παρατηρήσεις όπου διάφορα στοιχεία συνηγορούν στο ότι οι τυχαίες μεταβλητές μας ακολουθούν την κανονική κατανομή με μέση τιμή μ και διασπορά σ^2. Τότε γνωρίζουμε ότι προσεγγιστικά με πιθανότητα περίπου 95%, παρατηρήσεις που προέρχονται από τον ίδιο πληθυσμό θα ανήκουν στο διάστημα $(\bar{x}-2s, \bar{x}+2s)$ με $\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$ και $s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)$.

Στη διαδικασία μοντελοποίησης προσπαθούμε να κατασκευάσουμε ένα μοντέλο, σύμφωνα με την κατανομή πιθανότητας που υποθέτουμε, *προσαρμοσμένη κατανομή*, έτσι ώστε εφαρμόζοντας το μοντέλο στα δεδομένα, να εκτιμήσουμε την *πραγματική κατανομή* από την οποία προέρχονται.

Στην διαδικασία επιλογής μοντέλου με την βοήθεια του κριτηρίου πληροφορίας AIC, προσπαθούμε να βρούμε ποιό από τα υποψήφια μοντέλα μας προσεγγίζει καλύτερα την πραγματική κατανομή των δεδομένων ή αλλιώς ποιό μοντέλο ελαχιστοποιεί την απόσταση μεταξύ προσαρμοσμένης και πραγματικής κατανομής. Ένα εργαλείο μέτρησης αυτής της απόστασης είναι και η *ποσότητα πληροφορίας κατά Kullback-Leibler* ή αλλιώς *απόκλιση κατά K-L*.

Ορισμός Ποσότητας Πληροφορίας κατά K-L

Έστω $\vec{X} = (X_1, X_2, ..., X_p)$ διάνυσμα συνεχών τ.μ. το οποίο χαρακτηρίζεται από την σ.π.π. $f(\vec{x}|\vec{\theta})$ η οποία είναι μια παραμετρική συνάρτηση με άγνωστες παραμέτρους τα ορίσματα του διανύσματος $\vec{\theta} \in R^{\kappa}$, διάστασης κ. Έστω επίσης ότι υπάρχει ένα πραγματικό διάνυσμα παραμέτρων $\vec{\theta}^* \in R^{\kappa}$ τέτοιο ώστε το $\vec{\theta}$ να πλησιάζει όσο το δυνατό πιο κοντά το $\vec{\theta}^*$, ή αλλιώς τέτοιο ώστε η προσαρμοσμένη κατανομή $f(\vec{x}|\vec{\theta})$ να αποκλίνει όσο το δυνατό λιγότερο από την πραγματική κατανομή $f(\vec{x}|\vec{\theta}^*)$.

Το πόσο κοντά είναι αυτά τα δύο μοντέλα, δηλαδή η προσαρμοστικότητα (goodness of fit) του $f(\vec{x}|\vec{\theta})$ ως προς το $f(\vec{x}|\vec{\theta}^*)$, μετράται με την ποσότητα πληροφορίας κατά K-L.

Η ποσότητα πληροφορίας κατά K-L ορίζεται ως

$$I(\vec{\theta}^*;\vec{\theta}) = \mathrm{H}(\vec{\theta}^*;\vec{\theta}^*) - \mathrm{H}(\vec{\theta}^*;\vec{\theta}), \qquad (3.2)$$

όπου $\mathrm{H}(\vec{\theta}^*;\vec{\theta}^*) = E[\log f(\vec{\mathrm{X}}|\vec{\theta})] = \int f(\vec{x}|\vec{\theta}^*)\log f(\vec{x}|\vec{\theta}^*)d\vec{x}$ ονομάζεται η κατά Shannon αρνητική εντροπία, η οποία είναι σταθερή για δεδομένο $f(\vec{x}|\vec{\theta}^*)$, συμβολίζοντας με log τον νεπέριο λογάριθμο.

Η ποσότητα $\mathrm{H}(\vec{\theta}^*;\vec{\theta}) = \int f(\vec{x}|\vec{\theta}^*)\log f(\vec{x}|\vec{\theta})d\vec{x}$ ονομάζεται αναμενόμενη ως προς $f(\vec{X}|\vec{\theta}^*)$ λογαριθμική πιθανοφάνεια ή αλλιώς *αναμενόμενη λογαριθμική πιθανοφάνεια* και μετράει την προσαρμογή της $f(\vec{x}|\vec{\theta})$ στην $f(\vec{x}|\vec{\theta}^*)$.

Η ποσότητα πληροφορίας κατά Kullback-Leibler μετράει, κατά κάποιο τρόπο, την απόκλιση μεταξύ δύο συναρτήσεων γι' αυτό και αναφέρετε συχνά ως *απόκλιση κατά K-L.*

Μια πιο γενική μορφή της απόκλισης μεταξύ δύο συναρτήσεων είναι η παρακάτω

$$D(g; f) = \int u\left(\frac{g(x)}{f(x)}\right)g(x)dx,$$

όπου για να πάρουμε την απόκλιση K-L αρκεί να χρησιμοποιήσουμε $u = \log$.

Πολλές φορές επηρεασμένοι από τον παραπάνω συμβολισμό της απόκλισης μεταξύ δύο συναρτήσεων την απόκλιση κατά K-L την συμβολίζουμε ως $KL(g; f)$ όπου f είναι η προσαρμοσμένη σ.π.π. και g η πραγματική σ.π.π. των παρατηρήσεων μας.

Εκτός από την απόκλιση κατά K-L υπάρχουν και άλλα μέτρα της απόκλισης μεταξύ της πραγματική και της προσαρμοσμένης κατανομής των δεδομένων. Μερικά από αυτά παρουσιάζονται στον παρακάτω Πίνακα 3.1.

$\chi^2(g; f) = \sum_{i=1}^{k} \frac{g_i^2}{f_i} - 1 = \sum_{i=1}^{k} \frac{(f_i - g_i)^2}{f_i}$	χ^2 -statistic,
$I_K(g; f) = \int \left\{ \sqrt{f(x)} - \sqrt{g(x)} \right\}^2 dx$	Hellinger distance,
$I_\lambda(g; f) = \frac{1}{\lambda} \int \left\{ \left(\frac{g(x)}{f(x)} \right) \right\} dx$	Generalized information,
$D(g; f) = \int u\left(\frac{g(x)}{f(x)} \right) g(x)\, dx$	Divergence
$L_1(g; f) = \int \lvert g(x) - f(x) \rvert\, dx$	L^1 - norm

Πίνακας 3.1: Διαφορετικά μέτρα της απόστασης δύο κατανομών.

Για την καλύτερη κατανόηση της ποσότητας πληροφορίας K-L ή απόκλισης K-L μεταξύ της πραγματικής και της προσαρμοσμένης κατανομής παραθέτουμε το παρακάτω παράδειγμα.

Παράδειγμα 3.1: Kullback-Leibler απόστασης.

Έστω F η οικογένεια κανονικών κατανομών $\left\{ N(\mu, \sigma^2) : -\infty < \mu < +\infty, < \sigma^2 < +\infty \right\}$. Έστω επίσης $x | \theta^* \sim N(\mu^*, \sigma^{*2})$ να είναι η πραγματική κατανομή g με διάνυσμα παραμέτρων $\vec{\theta}^* = (\mu^*, \sigma^{*2})$, και $x | \vec{\theta} \sim N(\xi, \sigma^2)$ να είναι το προσαρμοσμένο μοντέλο f με διάνυσμα παραμέτρων $\vec{\theta} = (\xi, \sigma^2)$.

Τότε έχουμε ότι

$$KL(g; f) = \frac{1}{2} \log\left(\frac{\sigma^{*2}}{\sigma^2} \right) + \frac{1}{2} \left[\frac{\sigma^2}{\sigma^{*2}} + \frac{(\mu^* - \xi)^2}{\sigma^{*2}} \right] - \frac{1}{2}. \qquad (3.3)$$

Παρατηρούμε ότι η συμμετοχή των μέσων στην Εξίσωση 3.3 είναι της μορφής $(\mu^* - \xi)^2$ ενώ οι διασπορές συμμετέχουν με τον λόγο σ^{*2} / σ^2.

Στην Εξίσωση 3.3 φαίνεται ξεκάθαρα η συμμετοχή των παραμέτρων της πραγματικής κατανομής $\vec{\theta}^* = (\mu^*, \sigma^{*2})$ στον υπολογισμό της τελικής απόκλισης. Αυτό καθιστά τον

υπολογισμό αδύνατο καθώς οι ποσότητες αυτές αποτελούν θεωρητικές τιμές που μόνο να εκτιμηθούν μπορούν.

Έτσι λοιπόν, το κρίσιμο ερώτημα που καλούμαστε να απαντήσουμε είναι πώς θα μετασχηματίσουμε την Kullback-Leibler πληροφορία έτσι ώστε να μπορεί να χρησιμοποιηθεί, έστω και προσεγγιστικά, μέσω της χρήσης των δεδομένων, για την εκτίμηση της απόκλισης μεταξύ προσαρμοσμένου και πραγματικού μοντέλου;

Μέση λογαριθμική πιθανοφάνεια ως εκτιμήτρια της απόστασης K-L.

Συμβολίζοντας με $l(\vec{\theta})$ την λογαριθμική πιθανοφάνεια ενός δείγματος $(x_1, x_2, ..., x_n)$ προερχόμενο από την $f(X;\vec{\theta})$ με $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_\kappa)$ τότε μπορούμε να ορίσουμε την *μέση λογαριθμική πιθανοφάνεια του δείγματος* ως

$$\frac{1}{n} l(\vec{\theta}) = \frac{1}{n} \log L(\vec{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log f(x_i, \vec{\theta}) = l_n(\vec{\theta}),$$

η οποία αποτελεί μια εκτιμήτρια της απόστασης μεταξύ της πραγματικής σ.π.π. $f(\vec{x} | \vec{\theta}^*)$ και της προσαρμοσμένης σ.π.π. $f(\vec{x} | \vec{\theta})$.

Καθώς η ποσότητα πληροφορίας $KL(\vec{\theta}^*; \vec{\theta})$ είναι μη υπολογίσιμη, πρέπει να εκτιμηθεί από μια ποσότητα $\tilde{KL}(\vec{\theta}^*; \vec{\theta})$. Για την εύρεση της εν λόγω εκτιμήτρια χρησιμοποιώντας την Εξίσωση 3.2 έχουμε:

$$\tilde{KL}(\vec{\theta}^*; \vec{\theta}) = \tilde{H}(\vec{\theta}^*; \vec{\theta}^*) - \tilde{H}(\vec{\theta}^*; \vec{\theta}) \Rightarrow$$

$$\tilde{H}(\vec{\theta}^*; \vec{\theta}) = -\tilde{KL}(\vec{\theta}^*; \vec{\theta}) + \tilde{H}(\vec{\theta}^*; \vec{\theta}^*).$$

Καθώς η ποσότητα $\tilde{H}(\vec{\theta}^*; \vec{\theta}^*) = \tilde{H}(\vec{\theta}^*)$ παραμένει σταθερή για δεδομένη $f(\vec{x} | \vec{\theta}^*)$, η μεγιστοποίηση της αναμενόμενης λογαριθμικής πιθανοφάνειας $\tilde{H}(\vec{\theta}^*; \vec{\theta})$, ισούται ασυμπωτικά με την ελαχιστοποίηση της πληροφορίας κατά K-L.

Χρησιμοποιώντας δεδομένα για τις $X_1, X_2, ..., X_n$ τυχαίες μεταβλητές παίρνουμε τις εκτιμήτριες μέγιστης πιθανοφάνειας $\hat{\vec{\theta}}$ για το $\vec{\theta}$.

Έτσι,

$$\tilde{KL}\left(\vec{\theta}^*;\hat{\vec{\theta}}\right)=\tilde{\mathrm{H}}\left(\vec{\theta}^*;\vec{\theta}^*\right)-\frac{1}{n}\sum_{i=1}^{n}\log f\left(x_i,\hat{\vec{\theta}}\right)=$$
$$=\tilde{\mathrm{H}}\left(\vec{\theta}^*;\vec{\theta}^*\right)-I_n\left(\hat{\vec{\theta}}\right).$$

Καθώς,

$$I_n\left(\hat{\vec{\theta}}\right)=\frac{1}{n}I\left(\hat{\vec{\theta}}\right)\to\int f\left(\vec{x}\,|\,\theta^*\right)\log f\left(\vec{x}\,|\,\theta\right)d\vec{x}=E\left[\log f\left(\vec{X}\,|\,\theta\right)\right]=\tilde{\mathrm{H}}\left(\vec{\theta}^*;\vec{\theta}\right)=\tilde{H}\left(\vec{\theta}^*\right),$$

όπου η σύγκλιση είναι σχεδόν παντού, με πιθανότητα 1.

Συνεπώς έχουμε ότι

$$I_n\left(\hat{\vec{\theta}}\right)=\tilde{KL}\left(\vec{\theta}^*;\hat{\theta}\right)+\tilde{\mathrm{H}}\left(\vec{\theta}^*\right).$$

Λόγω της παραπάνω σύγκλισης οι εκτιμήτριες μέγιστης πιθανοφάνειας $\hat{\theta}$ που μεγιστοποιούν την $I(\theta)$ τείνουν σχεδόν παντού στο θ_0, το οποίο αποτελεί το *σημείο ελάχιστης πληροφορία κατά K-L*. Συνεπώς οι εκτιμήτριες μέγιστης πιθανοφάνειας παρέχουν την καλύτερη παραμετρική προσέγγιση της πραγματικής σ.π.π. $f\left(\vec{x}\,|\,\theta^*\right)$ των δεδομένων κάτω από την υπόθεση του προσαρμοσμένου μοντέλου $f\left(\vec{x}\,|\,\theta\right)$.

Καθώς η εκτιμήτρια της K-L πληροφορίας βασίζεται πάνω στη μέση λογαριθμική πιθανοφάνεια, η οποία είναι εκτιμήτρια και της αναμενόμενης λογαριθμικής πιθανοφάνειας, και οι εκτιμήτριες μέγιστης πιθανοφάνειας είναι μεροληπτικές ως προς το δείγμα μας, είναι αναπόφευκτη η εισαγωγή ενός σφάλματος εκτίμησης της K-L ποσότητας πληροφορίας. Για την ελαχιστοποίηση αυτού του σφάλματος εισάγουμε κάποιον όρο ποινικοποίησης στα κριτήρια πληροφορίας.

Αυτός είναι ο λόγος που για συγκεκριμένου μεγέθους n παρατηρήσεων, θέλοντας να έχουμε τη λογαριθμική πιθανοφάνεια κοντά στο θεωρητικό επίπεδο, στα διάφορα κριτήρια που χρησιμοποιούν την Kullback-Leibler απόκλιση, διορθώνουμε την μεροληψία που εισάγεται λόγω της χρήσης των εκτιμητριών μέγιστης πιθανοφάνειας με κάποιο όρο ποινικοποίησης.

Συμπερασματικά

Σύμφωνα με τα παραπάνω θα λέγαμε ότι η εκτίμηση της ποσότητα πληροφορίας κατά Kullback-Leibler $\tilde{KL}\left(\vec{\theta}^*;\vec{\theta}\right)$ είναι ίση με ένα σταθερό μέρος $\mathrm{H}\left(\vec{\theta}^*;\vec{\theta}^*\right)$, συν ένα μεταβλητό μέρος $\mathrm{H}\left(\vec{\theta}^*;\vec{\theta}\right)$, κάτω από κάποια συγκεκριμένη πραγματική σ.π.π. $f\left(x|\theta^*\right)$ των δεδομένων. Καθώς εμείς ενδιαφερόμαστε για σύγκριση μεταξύ διαφορετικών μοντέλων, ο σταθερός όρος αποτελεί έναν πρόσθετο όρο που μπορούμε να τον παραλείψουμε στις συγκρίσεις μας. Έτσι, ο μόνος όρος προς εκτίμηση είναι ο μεταβλητός όρος $\mathrm{H}\left(\vec{\theta}^*;\vec{\theta}\right)$ όπου τον εκτιμάμε με την βοήθεια της μέσης λογαριθμικής πιθανότητας σε συνδυασμό με τις εκτιμήτριες μέγιστης πιθανοφάνειας. Αυτή η μεθοδολογία εκτίμησης όμως εισάγει μια μεροληψία στο μοντέλο μας που προσπαθούμε να διορθώσουμε με κάποιο όρο ποινικοποίησης της λογαριθμικής πιθανοφάνειας μας.

Την παραπάνω συλλογιστική πορεία είχε και ο Akaike (1974), όταν στο AIC ενσωμάτωνε τον όρο ποινικοποίησης $2p$ χρησιμοποιώντας εκτιμήτριες μέγιστης πιθανοφάνειας για την εκτίμηση της αναμενόμενης λογαριθμικής πιθανοφάνειας από την μέση λογαριθμική πιθανοφάνεια.

Κεφάλαιο IV

Akaike **Κριτήριο Πληροφορίας και οι Επεκτάσεις του**: TIC, AICc, WIeAIC

4.1. **Εισαγωγή**

Έχοντας αναφερθεί στις εκτιμήτριες μέγιστης πιθανοφάνειας τόσο σαν εργαλείο εκτίμησης των άγνωστων παραμέτρων ενός μοντέλου, όσο και σαν την καλύτερη προσεγγιστικά μέθοδο ελαχιστοποίησης της απόστασης μεταξύ προσαρμοσμένης και πραγματικής κατανομής όταν χρησιμοποιούμε τη μέση λογαριθμική πιθανοφάνεια, είμαστε έτοιμοι να παρουσιάσουμε ένα σχεδιάγραμμα της απόδειξης του AIC για να καταλάβουμε πώς δουλεύει. Στο σημείο αυτό πρέπει να εισάγουμε δύο καινούργιες έννοιες οι οποίες θα μας βοηθήσουν στο έργο μας.

4.2. **Διάνυσμα Επίδοσης και Συνάρτηση Πίνακα Πληροφορίας**

Έστω $\log f(Y,\vec{\theta})$ η λογαριθμική πιθανοφάνεια ενός μοντέλου. Τότε ορίζουμε τις ποσότητες

$$\vec{u}(Y,\vec{\theta}) = \frac{\partial \log f(Y,\vec{\theta})}{\partial \vec{\theta}} \quad \text{και} \quad \mathrm{I}(\mathrm{Y},\vec{\theta}) = \frac{\partial^2 \log f(Y,\vec{\theta})}{\partial \vec{\theta} \partial \vec{\theta}^t},$$

όπου η πρώτη είναι μια διάστασης p συνάρτηση που καλείται *διάνυσμα επίδοσης* (score vector) του μοντέλου με ορίσματα $\partial \log f(Y,\theta)/\partial \theta_j$ για j=1,2,...,p , και η δεύτερη είναι ένας $p \times p$ πίνακας όπου καλείται *πίνακας πληροφορίας* του μοντέλου. Με p συμβολίζουμε τη διάσταση του παραμετρικού χώρου. Έχοντας ορίσει τις παραπάνω ποσότητες, θέτουμε

$$J = -E_g I(Y,\theta_0) \quad \text{και} \quad \mathrm{K} = Var_g u(Y,\theta_0), \tag{4.1}$$

όπου θ_0 το σημείο στο οποίο είναι ελάχιστη η πληροφορία κατά K-L. Οι παραπάνω $p\times p$ πίνακες είναι ίσοι όταν η $g(Y)$ πραγματική κατανομή, συμπίπτει με την $f(Y,\theta_0)$ προσαρμοσμένη κατανομή, με θ_0 τις παραμέτρους που ελαχιστοποιούν της K-L απόκλιση, για κάθε Y. Στην περίπτωση αυτή ο πίνακας

$J(\theta_0)=-\int f(y,\theta_0)\mathrm{I}(y,\theta_0)\,dy$ καλείται *πίνακας πληροφορίας Fisher* του μοντέλου.

Αντίστοιχα, στην περίπτωση της παλινδρόμησης για πραγματική κατανομή $g(Y\mid X)$ ορίζουμε τους πίνακες $U(Y\mid X,\vec{\theta})=\dfrac{\partial \log f(Y\mid X,\vec{\theta})}{\partial\vec{\theta}}$, $\mathrm{I}(\mathrm{Y}|\mathrm{X},\vec{\theta})=\dfrac{\partial^2 \log f(Y\mid X,\vec{\theta})}{\partial\vec{\theta}\partial\vec{\theta}^t}$,

$$J_n=-n^{-1}\sum_{i=1}^{n}\int g(Y\mid \vec{X}_i)I(Y\mid X_i,\theta_{0,n})\mathrm{dy} \quad \text{και} \quad \mathrm{K}_n=n^{-1}\sum_{i=1}^{n}Var_g u(Y\mid X_i,\theta_{0,n}),$$

όπου το $\theta_{0,n}$ αναφέρεται στην ελάχιστη απόσταση κατά K-L στην περίπτωση της παλινδρόμησης. Οι εκτιμήσεις των J_n, K_n τότε είναι ίσες με

$$\hat{J}_n=-n^{-1}\partial^2 l(\hat{\vec{\theta}})/\partial\vec{\theta}\partial\vec{\theta}^2=-n^{-1}\sum_{i-1}^{n} l(y_i\mid X_i,\hat{\vec{\theta}}), \tag{4.2a}$$

$$\hat{\mathrm{K}}_n=n^{-1}\sum_{i=1}^{n}u(y_i\mid X_i,\hat{\vec{\theta}})u(y_i\mid x_i,\hat{\vec{\theta}})^t. \tag{4.2b}$$

Οι εκτιμήσεις $\hat{J}_n,\hat{\mathrm{K}}_n$ των J_n,K_n, όπως θα δούμε παρακάτω, χρησιμοποιούνται για επεκτάσεις του AIC με διορθωμένη μεροληψία που εισάγεται λόγω των εκτιμητριών μέγιστης πιθανοφάνειας.

4.3. Σχεδιάγραμμα Απόδειξης AIC

Όπως έχουμε αναφέρει το AIC στηρίζεται πάνω στη ποσότητα πληροφορίας κατά K-L. Έστω ότι έχουμε προσαρμόσει στα δεδομένα μας διάφορα μοντέλα με σ.π.π. $f(\cdot,\hat{\theta})$, με $\hat{\theta}$ η εκτιμήτρια μέγιστης πιθανοφάνειας, ενώ η πραγματική σ.π.π. των δεδομένων μας είναι $g(\cdot)$.

Τότε η ποσότητα πληροφορίας κατά K-L θα είναι

$$KL(g,\ f(\cdot,\hat{\theta}))=\int g(y)\{\log g(y)-\log f(y,\hat{\theta})\}dy=\int g(y)\log g(y)\ dy-R_n,$$

όπου ο πρώτος όρος είναι σταθερός ανάμεσα στα διάφορα προσαρμοσμένα μοντέλα, καθώς εξαρτάται μόνο από την πραγματική κατανομή, και ο δεύτερος όρος R_n είναι

τυχαία μεταβλητή και εξαρτάται από τα δεδομένα μας μέσω των εκτιμητριών μέγιστης πιθανοφάνειας.

Έτσι, η αναμενόμενη τιμή του R_n θα είναι

$$Q_n = E_g[R_n] = E_g\left[\int g(y)\log f(y,\hat{\theta})\,\mathrm{dy}\right].$$

Αυτό που προσπαθεί να κάνει το AIC είναι να εκτιμήσει την αναμενόμενη τιμή του R_n για κάθε μοντέλο και να διαλέξει το μοντέλο με την καλύτερη εκτίμηση, πετυχαίνοντας έτσι έμμεσα να διαλέξει το μοντέλο με την μικρότερη K-L απόκλιση από την πραγματική κατανομή.

Μια μέθοδος εκτίμησης του Q_n είναι να εκτιμήσουμε την αναμενόμενη τιμή του R_n από τον αντίστοιχο δειγματικό μέσο.

Δηλαδή $\hat{Q}_n = n^{-1}\sum_{i=1}^{n}\log f(Y_i,\hat{\theta}) = n^{-1}l(\hat{\theta})$ όπου $n^{-1}l(\hat{\theta}) = l_n(\hat{\theta})$ είναι η μέση λογαριθμική πιθανοφάνεια.

Έχοντας ορίσει τις ποσότητες J, K στη Σχέση 4.1, προχωράμε στην εκτίμηση της μεροληψίας της $\hat{Q}_n$ με την βοήθεια των ποσοτήτων αυτών.

Η εκτιμήτρια $\hat{Q}_n$ της R_n, τείνει να υπερεκτιμήσει την υπό εκτίμηση ποσότητα Q_n

καθώς $E(\hat{Q}_n - Q_n) = \frac{p^*}{n}$, $\mathrm{p}^* = Tr(J^{-1}K)$, όπου το $\frac{p^*}{n}$ ονομάζεται *γενικευμένη διάσταση* του μοντέλου.

Άρα μία διορθωμένη εκτιμήτρια της Q_n ως προς την μεροληψία είναι η ποσότητα

$$\hat{Q}_n' = \hat{Q}_n - \frac{p^*}{n} = n^{-1}\{l(\hat{\theta}) - p^*\}.$$

θέτοντας $p^* = p$ όπου p η διάσταση του παραμετρικού χώρου του μοντέλου, υποθέτοντας έτσι ότι το υποψήφιο μοντέλο είναι το πραγματικό, οδηγούμαστε στον τύπο

$$AIC(M) = 2l_n(\hat{\theta}_{ML}) - 2p, \tag{4.2}$$

όπου αποτελεί και τον γενικό τύπο του Akaike κριτήριου πληροφορίας.

Παράδειγμα 4.1: Επιλογή μεταβλητών σε ένα γραμμικό μοντέλο παλινδρόμησης.

Έστω ότι έχουμε μια μεταβλητή απόκρισης Y και p επεξηγηματικές μεταβλητές $X_1, X_2, ..., X_p$. Το γραμμικό μοντέλο παλινδρόμησης τότε θα είναι

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon, \quad \text{όπου} \quad \varepsilon \sim \mathrm{N}(0, \sigma^2).$$

Η δεσμευμένη κατανομή της μεταβλητής απόκρισης Y δεδομένου των επεξηγηματικών μεταβλητών είναι

$$P\left(Y \mid X_1, X_2, ..., X_p\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}\left(Y - \sum_{i=0}^{p} \beta_i X_i \right)^2 \right\} \quad \text{με } \mathrm{X}_0 = 1.$$

Συνεπώς, για n σύνολα ανεξάρτητων παρατηρήσεων $\{y_i, x_{i1}, x_{i2}, ..., x_{ip}$ όπου $i = 1, 2, ..., n\}$ η πιθανοφάνεια του μοντέλου παλινδρόμησης θα είναι

$$L\left\{\beta_0, \beta_1, \beta_2, ..., \beta_p, \sigma^2\right\} = \prod_{i=1}^{n} p(Y_i \mid X_{i1}, X_{i2}, ..., X_{ip}).$$

Άρα η λογαριθμική πιθανοφάνεια θα δίνεται από τον τύπο

$$l\left(\beta_0, \beta_1, \beta_2, ..., \beta_p, \sigma^2\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \sum_{i=0}^{p} a_i X_i \right)^2 \quad \text{με} \quad \mathrm{X}_0 = 1. \tag{4.3}$$

Αντίστοιχα, οι εκτιμήτριες μέγιστης πιθανοφάνειας των συντελεστών παλινδρόμησης $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$ θα προέρχονται ως λύση του συστήματος των γραμμικών εξισώσεων

$$X^t X \vec{\beta} = X^t Y,$$

όπου $\vec{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^t$, $\vec{X}$ ο πίνακας διάστασης $n \times (p+1)$ και $\vec{Y}$ το διάνυσμα διάστασης n.

$$\vec{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \quad \text{και} \quad \vec{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_{n-1} \\ Y_n \end{bmatrix}$$

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

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \left(\sum_{j=0}^{p} \hat{\beta}_j x_{ij} \right) \right]^2 \quad \text{με} \quad x_{i0} = 1. \tag{4.4}$$

Αντικαθιστώντας τις εκτιμήτριες μας στην Σχέση 4.3 έχουμε

$$l(\beta_0, \beta_1, \beta_2, ..., \beta_p, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\hat{\sigma}^2 - \frac{n}{2}.$$

Καθώς ο αριθμός των ελεύθερων παραμέτρων που περιέχονται στο πολυπαραμετρικό μοντέλο παλινδρόμησης μας είναι $p+2$ ($p+1$ συντελεστές παλινδρόμησης και το σ^2) το AIC για αυτό το μοντέλο θα είναι $AIC = n(\log 2\pi + 1) + n\log\hat{\sigma} + 2(p+2)$.

Στην πολλαπλή ανάλυση παλινδρόμησης πολλές φορές κάποιες από τις αρχικές επεξηγηματικές μεταβλητές δεν χρησιμοποιούνται στο τελικό μοντέλο πρόβλεψης της μεταβλητής απόκρισης, καθώς κάνουν το μοντέλο ασταθές στις προβλέψεις του. Επιλέγοντας το μοντέλο που έχει το μικρότερο AIC για διαφορετικούς συνδυασμούς επεξηγηματικών μεταβλητών πετυχαίνουμε ένα κατάλληλο και φειδωλό μοντέλο πρόβλεψης.

Ως εφαρμογή των παραπάνω, θεωρούμε τον παρακάτω πίνακα στον οποίο αποτυπώνεται η μικρότερη ημερήσια θερμοκρασία τον μήνα Ιανουάριο κατά μέσο όρο, Y, από το 1971 μέχρι το 2000, το γεωγραφικό πλάτος, X_1, το γεωγραφικό μήκος, X_2 και το υψόμετρο, X_3, σε 25 πόλεις της Ιαπωνίας. Σκοπός μας είναι να προβλέψουμε την κατά μέσο όρο ελάχιστη ημερήσια θερμοκρασία το μήνα Ιανουάριο, λαμβάνοντας υπ' όψην τις επεξηγηματικές μεταβλητές X_1, X_2, X_3. Διάφορα στοιχεία συνηγορούν ότι το καλύτερο μοντέλο πρόβλεψης για αυτό το σκοπό είναι το γραμμικό μοντέλο παλινδρόμησης, με σφάλμα που ακολουθεί την κανονική κατανομή με μέση τιμή 0 και διασπορά σ^2,

$$Y_i = a_0 + a_1 X_{i1} + a_2 X_{i2} + a_3 X_{i3} + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2).$$

n	Cities	Temp. (y)	Latitude (x_1)	Longitude (x_2)	Altitude (x_3)
1	Wakkanai	−7.6	45.413	141.683	2.8
2	Sapporo	−7.7	43.057	141.332	17.2
3	Kushiro	−11.4	42.983	144.380	4.5
3	Nemuro	−7.4	43.328	145.590	25.2
4	Akita	−2.7	39.715	140.103	6.3
5	Morioka	−5.9	39.695	141.168	155.2
6	Yamagata	−3.6	38.253	140.348	152.5
7	Wajima	0.1	37.390	136.898	5.2
8	Toyama	−0.4	36.707	137.205	8.6
9	Nagano	−4.3	36.660	138.195	418.2
10	Mito	−2.5	36.377	140.470	29.3
11	Karuizawa	−9.0	36.338	138.548	999.1
12	Fukui	0.3	36.053	136.227	8.8
13	Tokyo	2.1	35.687	139.763	6.1
14	Kofu	−2.7	35.663	138.557	272.8
15	Tottori	0.7	35.485	134.240	7.1
16	Nagoya	0.5	35.165	136.968	51.1
17	Kyoto	1.1	35.012	135.735	41.4
18	Shizuoka	1.6	34.972	138.407	14.1
19	Hiroshima	1.7	34.395	132.465	3.6
20	Fukuoka	3.2	33.580	130.377	2.5
21	Kochi	1.3	33.565	133.552	0.5
22	Shionomisaki	4.7	33.448	135.763	73.0
23	Nagasaki	3.6	32.730	129.870	26.9
24	Kagoshima	4.1	31.552	130.552	3.9
25	Naha	14.3	26.203	127.688	28.1

Πίνακας 4.1: Ελάχιστες ημερήσιες θερμοκρασίες σε πόλεις της Ιαπωνίας κατά μέσο όρο στο διάστημα 1971-2000.

Δεδομένου του Πίνακα 4.1 με τιμές $\{(y_i, x_{i1}, x_{i2}, x_{i3})$ για $i = 1,2,...,n\}$ η πιθανοφάνεια του γενικού μοντέλου παλινδρόμησης με τρείς επεξηγηματικές μεταβλητές θα είναι

$$L(\beta_0,\beta_1,\beta_2,\beta_3,\sigma^2) = \prod_{i=1}^{n} p(Y \mid X_{i1}, X_{i2}, X_{i3}) =$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{3} a_j x_j\right)^2\right\} \text{ με } \quad x_0 = 1,$$

ενώ η λογαριθμική πιθανοφάνεια του θα είναι

$$l(\beta_0,\beta_1,\beta_2,\beta_3,\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{3}\beta_j x_j\right)^2 \quad \text{με} \quad x_0 = 1\,. \tag{4.5}$$

Οι εκτιμήτριες $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ των συντελεστών παλινδρόμησης που προκύπτουν με την μέθοδο μέγιστης πιθανοφάνειας καθώς και η εκτιμήτρια μέγιστης πιθανοφάνειας της διασποράς των υπολοίπων, $\hat{\sigma}^2$ εισάγονται στην Εξίσωση 4.1 και έχουμε

$$l(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\hat{\sigma}^2 - \frac{n}{2}.$$

Έχοντας υπολογίσει την γενική λογαριθμική πιθανοφάνεια για το γραμμικό μοντέλο παλινδρόμησης που προσαρμόσαμε στα δεδομένα μας με τέσσερις επεξηγηματικές μεταβλητές, εφαρμόζουμε την ίδια ακριβώς μεθοδολογία και για τα υπόλοιπα υποψήφια μοντέλα. Υποψήφια μοντέλα θεωρούνται όλοι οι δυνατοί συνδυασμοί των επεξηγηματικών μεταβλητών που συμμετέχουν στο πρόβλημα συμπεριλαμβάνοντας κάθε φορά και τον σταθερό όρο X_0. Το σύνολο δηλαδή των υποψήφιων μοντέλων είναι 8.

Άρα το AIC για κάθε μοντέλο θα προκύπτει από τον τύπο

$$AIC = n(\log 2\pi + 1) + n\log\hat{\sigma}^2 + 2(\kappa + 2),$$

όπου κ ο αριθμός των επεξηγηματικών μεταβλητών που συμπεριλαμβάνονται σε κάθε μοντέλο.

No	**Επεξηγηματικές Μεταβλητές**	**Διασπορά υπολοίπων**	k	AIC	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
1	X_0, X_1	4,411	1	154,8727	29,8483	-0,8207	-	-
2	X_0, X_2	4,621	1	157,2947	94,2642	-	-0,6881	-
3	X_0, X_3	5,066	1	162,0716	0,887343	-	-	-0,011071
4	X_0, X_1, X_2	4,496	2	156.7647	43,0150	-0,6974	-0,1288	-
5	X_0, X_1, X_3	3,799	2	148.004	30,910146	-0,822073	-	-0,011111
6	X_0, X_2, X_3	4,271	2	154.0933	87,299331	-	-0,631197	-0,009299
7	X_0, X_1, X_2, X_3	3,863	3	149.7176	12,294185	-0,996884	-0,182614	-0,011632
8	X_0	5,477	0	165.1907	-0,1192	-	-	-

Πίνακας 4.2: Αποτελέσματα εφαρμογής AIC στην R, στα υποψήφια μοντέλα του παραδείγματος 4.2.

Σχολιάζοντας τον Πίνακα 4.2 θα λέγαμε ότι το $5^{ο}$ μοντέλο (με επεξηγηματικές μεταβλητές X_0, X_1, X_3) είναι το μοντέλο που επιλέγεται ως καταλληλότερο από το AIC. Το μοντέλο που περιλαμβάνει το σύνολο των μεταβλητών ($7^{ο}$ μοντέλο) και θα περιμέναμε να είναι το επιλεγόμενο, βλέπουμε ότι έρχεται δεύτερο. Αυτό είναι αποτέλεσμα της ύπαρξης στον τύπο του AIC του όρου ποινικοποίησης, καθώς η αύξηση

της πιθανοφάνειας του μοντέλου με την προσθήκη της X_2, λόγω της αύξησης της διασποράς, είναι 0,064 ενώ του όρου ποινικοποίησης είναι κατά 1,00.

Το καταλληλότερο μοντέλο που προτείνει το AIC δεν είναι απαραίτητα και το επιλεγόμενο μοντέλο από τον ερευνητή κάθε φορά. Συγκεκριμένα, στο δικό μας παράδειγμα κάποιος ερευνητής θα μπορούσε να διαλέξει το $1^{ο}$ μοντέλο ως το καταλληλότερο μεταξύ αυτών που έχουν μια επεξηγηματική μεταβλητή. Κάποιος από τους λόγους που θα τον ωθούσαν στην απόφαση αυτή είναι πιθανώς η αδυναμία συλλογής δεδομένων πολλών μεταβλητών. Συνεπώς, το Akaike κριτήριο πληροφορίας προτείνει καταλληλότερο μοντέλο και ο ερευνητής με βάση και άλλα δεδομένα επιλέγει ποιο θα προσαρμόσει στα δεδομένα του.

4.4. **Βελτίωση του** AIC **από το** Takeuchi **Κριτήριο πληροφορίας** (TIC)

Όπως είδαμε και παραπάνω το κλειδί στην γενική μορφή του AIC είναι ότι η μεροληψία της εκτιμήτριας $\hat{Q}_n$ της Q_n μπορεί να προσεγγιστεί από τη γενικευμένη διάσταση του μοντέλου $\frac{p^*}{n}$. Διαφορετικές προσεγγίσεις της μεροληψίας του $\hat{Q}_n$ μπορούμε να επιτύχουμε από διαφορετικές εκτιμήσεις του p^* όπου οδηγούν σε ένα διορθωτικό παράγοντα μεροληψίας $n^{-1}l_{n,\max} - \hat{p}^*$ για την εκτιμήτρια $\hat{Q}_n$.

Συγκεκριμένα, στην έκφραση του AIC έχουμε θέσει το p^* ίσο με το $p = \dim(M)$, υποθέτοντας έτσι ότι το πραγματικό μοντέλο συμπίπτει με το προσαρμοσμένο μοντέλο. Αυτήν την αυθαίρετη υπόθεση προσπαθούμε να αντικαταστήσουμε με μια άλλη εκτίμηση του p^* με το Takeuchi κριτήριο πληροφορίας (TIC).

Ο Takeuchi (1976) πρότεινε ένα κριτήριο, που προκύπτει με τη βοήθεια των ποσοτήτων $J, K,$ της εξίσωσης 4.1 και έχει την μορφή

$$TIC = 2l_n(\hat{\hat{\theta}}_{ML}) - 2\hat{p}^* \quad \text{με} \quad \hat{p}^* = Tr(\hat{J}_n^{-1}\hat{K}_n),$$

όπου οι ποσότητες $\hat{J}, \hat{K}$ παράγονται από τις εξισώσεις 4.2a, 4.2b αντίστοιχα.

Αν υπολογίζοντας τα $\hat{J}_n, \hat{K}_n$ και το $\mathrm{Tr}(\hat{J}_n^{-1}\hat{K}_n)$ αντιμετωπίσουμε δυσκολία υπολογισμού λόγω μεγάλου μεγέθους των πινάκων ή λόγω πολυπλοκότητας των ορισμάτων τους, τότε μπορούμε να εισάγουμε μια μεταβλητή d τέτοια ώστε $K_n = (1+d)J_n$, όπου την

ονομάζουμε υπερδιάστατη (overdispersion) μεταβλητή. Τότε, το $\mathrm{Tr}(\hat{J}_n^{-1}\hat{K}_n)$ είναι ίσο με το $(1+d)p$ έχοντας να εκτιμήσουμε μόνο το d.

Το TIC δεν αποτελεί μια διόρθωση του AIC ως προς τα κύρια χαρακτηριστικά του AIC, όπως είναι οι εκτιμήτριες μέγιστης πιθανοφάνειας ή η απόκλιση κατά K-L, αλλά αποτελεί μια διόρθωση ως προς την γενική υπόθεση ότι το πραγματικό μοντέλο είναι ανάμεσα στα υποψήφια.

4.5. **Διορθωμένο** Akaike **κριτήριο πληροφορίας** (AICc)

Προσπαθώντας να διορθώσουμε κάποιες από τις αδυναμίες του AIC ως προς τα κύρια χαρακτηριστικά του, παρατηρούμε ότι όσο αυξάνει το μέγεθος n του δείγματος, τόσο το AIC τείνει να επιλέγει πιο πολύπλοκα μοντέλα. Αυτό οφείλεται στο γεγονός ότι η μέγιστη πιθανοφάνεια ενός μοντέλου αυξάνει γραμμικά με την αύξηση του μεγέθους n του δείγματος, ενώ ο παράγοντας ποινικοποίησης της εξαρτάται μόνο από το μέγεθος του παραμετρικού χώρου του μοντέλου. Κύριος στόχος του διορθωμένου Akaike κριτηρίου πληροφορίας (AICc) είναι να διορθώσει αυτή την αδυναμία του AIC.

Θα προσπαθήσουμε να δούμε την παραγωγή αυτής της διόρθωσης μέσα από τα γενικά γραμμικά μοντέλα.

Ξέρουμε ότι

$$AIC_{GML} = -2n\log(\hat{\sigma}) - 2(p+1) - n - n\log(2\pi) \text{ με } \hat{\sigma}^2 = \frac{\|ress\|^2}{n} \text{ και ress=Y-X}\hat{\beta}.$$

Άρα το AIC μας συμβουλεύει να επιλέξουμε εκείνο το μοντέλο που ελαχιστοποιεί το $n\log\hat{\sigma} + p$ (καθώς οι υπόλοιποι όροι παραμένουν σταθεροί ανάμεσα στα υποψήφια μοντέλα).

Γενικά το AIC προσπαθεί να εκτιμήσει την αναμενόμενη K-L ποσότητα πληροφορίας του υποψήφιου μοντέλου $f(Y \mid x, \hat{\theta})$ από τον μηχανισμό γέννησης των δεδομένων $\left(\textit{πραγματική κατανομή } g(Y \mid x)\right)$. Στην δική μας περίπτωση, του γενικού γραμμικού μοντέλου, $\vec{\theta} = (\beta_1, \beta_2, ..., \beta_p, \sigma^2)$.

Έστω ότι η άγνωστη παράμετρος και η τυπική απόκλιση του $g(y \mid x)$ είναι $\xi(x)$, σ_0 αντίστοιχα. Υποθέτοντας ότι το υποψήφιο μοντέλο είναι ίσο με το πραγματικό έχουμε ότι $\xi_i = \xi(x_i) = x_i^t\beta$.

Αν αντί της εκτιμήτριας μέγιστης πιθανοφάνειας $\hat{\sigma}^2_{ML}$ χρησιμοποιήσουμε μια αμερόληπτη εκτιμήτρια, όπου προκύπτει εύκολα από το γεγονός ότι $\| ress \|^2 \sim \sigma^2 X^2_{n-p}$, έχουμε ότι

$$\hat{\sigma}^2_{u.e} = \frac{\| ress \|^2}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x^t_{ij}\hat{\beta}_j \right)^2 ,$$

όπου με $\hat{\sigma}^2_{u.e}$ συμβολίζουμε τις αμερόληπτες εκτιμήτριες.

Έτσι, οι εκτιμήτριες για το σ^2 είναι

$$\hat{\sigma}^2 = \frac{1}{n-a}\| ress \|^2 \quad \begin{matrix} για\ a=0 \ \text{εκτιμήτρια μέγιστης πιθανοφάνειας (ML)} \\ για\ a=p \ \text{αμερόληπτη εκτιμήτρια(U.E)} \end{matrix} .$$

Ακολουθώντας την ίδια συλλογιστική πορεία όπως στο σχεδιάγραμμα της απόδειξης του AIC αλλά για το γενικό γραμμικό μοντέλο, έχουμε ότι η μεροληψία της εκτιμήτρια $\hat{Q}_n$ της ποσότητας R_n είναι $E_g\left(\hat{Q}_n - R_n \right) = \frac{p+1}{n} + \frac{n-a}{n-p-2}$.

Σύμφωνα με τα παραπάνω οδηγούμαστε σε δύο στρατηγικές διορθώσεις της μορφής του AIC.

$1^{η}$ στρατηγική διόρθωσης $AIC\ \left(AIC^1_c \right)$.

Αφήνουμε την μέγιστη πιθανοφάνεια όπως στο AIC εκτιμώντας τις παραμέτρους μας με εκτιμήτριες μέγιστης πιθανοφάνειας, ενώ τροποποιούμε τον όρο ποινικοποίησης έτσι ώστε να διορθώσουμε την μεροληψία που εισάγεται λόγο των εκτιμητριών μέγιστης πιθανοφάνειας.

Έτσι έχουμε $AIC^1_c = 2l_n(\hat{\theta}_{ML}) - 2(p+1)\frac{n}{n-p-2}$.

$2^{η}$ στρατηγική διόρθωσης $AIC\ \left(AIC^2_c \right)$.

Αφήνουμε τον όρο ποινικοποίησης ως έχει και στην μέγιστη πιθανοφάνεια χρησιμοποιούμε αμερόληπτη εκτιμήτρια διασποράς και όχι μέγιστης πιθανοφάνειας.

Έτσι έχουμε,

$$AICc^2 = 2l(\hat{\theta}_{u.e}) - 2(p+1) \quad \text{με} \quad \hat{\sigma}^2_{u.e} = \frac{\| ress \|^2}{n-p-2},$$

όπου το $\hat{\theta}_{u.e}$ περιλαμβάνει τις εκτιμήτριες μέγιστης πιθανοφάνειας των συντελεστών παλινδρόμησης και την αμερόληπτη εκτιμήτρια $\hat{\sigma}^2_{u.e}$.

Οι δύο αυτές στρατηγικές είναι προϊόν της ίδιας ανάλυσης και δεν μπορούν να συγκριθούν μεταξύ τους έτσι ώστε να πούμε ποιά είναι καλύτερη. Από την άλλη πλευρά, μόνο η 1^{η} στρατηγική μπορεί να γενικευτεί και να χρησιμοποιηθεί και σε άλλα παραμετρικά μοντέλα παλινδρόμησης εκτός του γραμμικού με τον τύπο

$$AICc_g^2 = 2l(\hat{\theta}) - 2\dim(\theta)\frac{n}{n-\dim(\theta)-1},$$

όπου g η πραγματική κατανομή των δεδομένων.

Σχολιάζοντας τις διορθωμένες μορφές του AIC θα λέγαμε ότι η πρώτη προσπαθεί να διορθώσει την μεροληψία που εισάγεται λόγο της χρήσης εκτιμητριών μέγιστης πιθανοφάνειας κάνοντας τον όρο ποινικοποίησης να εξαρτάται από το μέγεθος του δείγματος, ενώ η δεύτερη προσπαθεί να διορθώσει την μεροληψία αυτή χρησιμοποιώντας για την εκτίμηση της διασποράς μια αμερόληπτη εκτιμήτρια.

Καμιά από τις δύο στρατηγικές δεν αντιμετωπίζει το πρόβλημα επηρεασμού των εκτιμητριών από ακραίες τιμές, πρόβλημα που συναντάται συχνά σε στατιστικές αναλύσεις. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι η σταθμισμένη επέκταση του AIC.

4.6. **Σταθμισμένα** Akaike **κριτήρια πληροφορίας** (WAIC)

Όπως είδαμε στις προηγούμενες παραγράφους το Akaike κριτήριο πληροφορίας χρησιμοποιεί εκτιμήτριες μέγιστης πιθανοφάνειας ή αμερόληπτες εκτιμήτριες έτσι ώστε να υπολογίσει της πιθανοφάνεια ενός μοντέλου και στη συνέχεια να ποινικοποιήσει αυτή την πιθανοφάνεια με ένα παράγοντα διαφορετικό ανάλογα σε ποιό κριτήριο αναφερόμαστε (AIC, AICc, TIC). Οι εκτιμήτριες μέγιστης πιθανοφάνειας αλλά και οι αμερόληπτες εκτιμήτριες, αποτελούν δειγματοσυναρτήσεις που αντιμετωπίζουν την κάθε παρατήρηση ισότιμα. Έτσι, ακραίες παρατηρήσεις μπορεί να επηρεάσουν πάρα πολύ τις εκτιμήτριες μας με ανάλογα αποτελέσματα και στην τιμή του AIC. Τη λύση σε αυτό το πρόβλημα έρχεται να δώσει η θεωρία των σταθμισμένων συναρτήσεων πιθανοφάνειας.

Ακραία ή *μεμονωμένη παρατήρηση* ονομάζεται μια παρατήρηση που είναι μακριά από το κύριο σώμα των παρατηρήσεων μας. Ο κύριος τρόπος εύρεσης αυτών των παρατηρήσεων είναι μέσω του διαγράμματος διασποράς των παρατηρήσεων μας.

Επίσης, *σταθμισμένη συνάρτηση πιθανοφάνειας* ονομάζεται η συνάρτηση πιθανοφάνειας που έχει ενσωματωμένη μια συνάρτηση βάρους των παραμέτρων και των μεταβλητών του προβλήματος.

Δηλαδή, έστω η συνάρτηση μέγιστης πιθανοφάνειας ενός μοντέλου

$$\sum_{i=1}^{n} \log f(y_i \mid x_i; \hat{\beta}, \hat{\sigma}) .$$

Η αντίστοιχη σταθμισμένη συνάρτηση μέγιστης πιθανοφάνειας θα είναι

$$\sum_{i=1}^{n} w(y_i - \hat{\beta} x_i, \hat{\sigma}) \log f(y_i \mid x_i; \hat{\beta}, \hat{\sigma}).$$

Μέθοδος Επιλογής Μοντέλου μέσω της Σταθμισμένης K-L Απόστασης

Στην προηγούμενη ενότητα εισαγάγαμε μια σταθμισμένη σ.π.π. και βρήκαμε τις εκτιμήτριες μας μέσω των σταθμισμένων εξισώσεων επίδοσης. Σε αυτή την ενότητα θα δείξουμε ότι ενθυλακώνοντας μια συνάρτηση $w(\cdot)$ στην σ.π.π., για να είμαστε συνεπής ως προς τον στόχο του AIC για επιλογή του μοντέλου που ελαχιστοποιεί την απόκλιση μεταξύ του προσαρμοσμένου μοντέλου με σ.π.π. f_θ και της πραγματικής κατανομής με σ.π.π. g, θα πρέπει να εισάγουμε έναν επιπλέον παράγοντα που εξαρτάται από τις παραμέτρους του προβλήματος μας.

Ορισμός σταθμισμένης Kullback-Leibler απόστασης.

Έστω g σ.π.π. της πραγματική κατανομή των δεδομένων, f_θ σ.π.π. της προσαρμοσμένη κατανομή στα δεδομένα και $w(\cdot)$ μία μη αρνητική συνάρτηση βάρους. Ορίζουμε ως

$$d_w(g, f_\theta) = \int w\big(\log(g \mid f_\theta) - (g - f_\theta)\big) dy, \qquad (4.8)$$

την σταθμισμένη K-L απόκλισης της προσαρμοσμένης f από την πραγματική κατανομή g των δεδομένων. Όταν η $w(\cdot)$ είναι σταθερή καταλήγουμε στον κλασικό ορισμό της K-L απόστασης.

Βασιζόμενοι στην φιλοσοφία του AIC, ζητούμενο είναι η ελαχιστοποίηση της ποσότητας $d_w(g, f_\theta)$. Η ελαχιστοποίηση της $d_w(g, f_\theta)$ είναι ισοδύναμη με την μεγιστοποίηση της ποσότητας

$$H(\theta) = \int w(g \log f_\theta - f_\theta) dy.$$

Έτσι, ορίζουμε ως σταθμισμένες εκτιμήτριες μέγιστης πιθανοφάνεια $\hat{\theta}_w$ τις τιμές που μεγιστοποιούν την Εξίσωση 4.9.

$$H_n(\theta) = n^{-1}\sum_{i=1}^{n} w(y_i)\log f(y_i,\theta) - \int w f_\theta dy. \qquad (4.9)$$

Η Εξίσωση 4.9 αποτελείται από την σταθμισμένες εξισώσεις επίδοσης $n^{-1}\sum_{i=1}^{n} w(y_i)\log f(y_i,\theta)$ αλλά και από τον επιπλέον παράγοντα $\int w f_\theta dy$.

Σημειώνουμε εδώ ότι μεγιστοποιώντας απλά την σταθμισμένη συνάρτηση πυκνότητας πιθανότητας $\sum_{i=1}^{n} w(y_i)\log f(y_i,\theta)$ χωρίς τον διορθωτικό όρο $-\int w f_\theta dy$ θα καταλήγαμε σε μη επαρκή εκτιμήτριες.

Αφού το $\hat{\theta}_w$ είναι σημείο μεγιστοποίησης της Εξίσωσης 4.9, θα είναι λύση και της εξίσωσης $H_n^1(\theta) = 0$ όπου $H_n^1(\theta) = n^{-1}\sum_{i=1}^{n} w(y_i)u - \xi(\theta)$ είναι η παράγωγος του $H_n(\theta)$ με $\xi = \int w u_\theta f_\theta dy$ την παράγωγο του $\int w f_\theta dy$, ως προς θ.

Έτσι η εκτιμήτρια $\hat{\theta}_w$ είναι συνεπής ως προς το σημείο $\theta_{w,0}$ σημείο ελάχιστης σταθμισμένης απόσταση $d_w(g, f_\theta)$.

Χρησιμοποιώντας την σύγκλιση κατά πιθανότητα p έχουμε ότι

$$H_n^1(\theta) \to_p \int w(g - f_\theta)u_\theta dy. \qquad (4.10)$$

Άρα, έχουμε ότι $\int w g u_\theta dy = \int w f_\theta u_\theta dy$ στο σημείο $\theta_{w,0}$, όπου το σημείο $\theta_{w,0}$ είναι το σημείο ελάχιστης απόκλισης $d_w(g, f_\theta)$ μεταξύ της προσαρμοσμένης κατανομής με σ.π.π. f_θ και της πραγματικής κατανομής με σ.π.π. g.

Συνεπώς, για να συγκρίνουμε την ικανότητα πρόβλεψης διαφόρων μοντέλων με την σταθμισμένη K-L απόσταση πρέπει να υπολογίσουμε την αναμενόμενη K-L απόσταση $d_w\left(g(\cdot), f_\theta(\cdot,\hat{\theta})\right)$. Η ποσότητας αυτή διαφέρει κατά μια σταθερά από την ποσότητα $Q_n = E_g R_n$ όπου $R_n = \int w\left(g\log f\left(\cdot,\hat{\theta}_W\right) - f\left(\cdot,\hat{\theta}_W\right)\right)dy$, g η πραγματική κατανομή των δεδομένων και $\hat{\theta}_W$ οι σταθμισμένες εκτιμήτριες μέγιστης πιθανοφάνειας.

Μετά από υπολογισμούς, που ξεφεύγουν από τον σκοπό αυτής της διπλωματικής, καταλήγουμε στον τύπο του WAIC

$$WAIC = 2nH_n(\hat{\theta}_w) - 2\tilde{p}^*, \text{ με } \tilde{p}^* = Tr(\tilde{J}_n^{-1}\tilde{K}_n).$$

Η συνάρτηση του κριτηρίου που μεγιστοποιείται είναι η

$$H_n(\theta) = n^{-1}\sum_{i=1}^{n} w(x_i, y_i)\log f(y_i \mid x_i, \theta) - n^{-1}\sum_{i=1}^{n}\int w(x_i, y)\, f(y \mid x_i, \theta)\, dx,$$

και οι ποσότητες που συμμετέχουνε στο κριτήριο ισούνται με

$\tilde{J}_n^{-1} = -H_n^2(\hat{\theta}_w)$, $\tilde{K}_n = n^{-1}\sum_{i=1}^{n} u_i u_i^t$, με $u_i = w(x_i, y_i)u\left(y_i \mid x_i, \hat{\theta}_w\right) - \xi(\hat{\theta}_w \mid x_i)$ και $\xi(\hat{\theta}_w \mid x_i) = \int w(x_i, y) f\left(y \mid x_i, \hat{\theta}_w\right)$.

Το AIC αποτελεί πάλι ειδική λύση του WAIC αν χρησιμοποιήσουμε ως $w(\cdot) = 1$.

Σχολιάζοντας θα λέγαμε ότι το WAIC μπορεί να αποτελέσει αξιόπιστο κριτήριο πληροφορίας με οποιαδήποτε μη αρνητική συνάρτηση $w(\cdot)$. Συνήθως όμως, χρησιμοποιούμαι συναρτήσεις τέτοιες ώστε να «κανονικοποιούνται» τα κομμάτια εκείνα του δείγματος που αποκλίνουν από το κύριο σώμα, αποφεύγοντας έτσι συνέπειες που θα είχαμε από αυτά στις εκτιμήτριες μας.

Ως προς τον όρο του WAIC, $H_n\left(\hat{\theta}_w\right)$, έχουμε να παρατηρήσουμε ότι αποτελείται από δύο επιμέρους όρους. Ο πρώτος είναι η σταθμισμένη μέγιστη πιθανοφάνεια του μοντέλου μας, και ο δεύτερος είναι ο επιπρόσθετος όρος στον οποίο αναφερθήκαμε στην εισαγωγή της παραγράφου. Ως προς τον όρο $2\tilde{p}^*$ που χρησιμοποιείται σαν όρος ποινικοποίησης του $2nH_n(\hat{\theta}_w)$, η μορφή του θυμίζει την μορφή του όρου ποινικοποίησης του TIC. Πράγματι, είναι στην ίδια λογική αρκεί να επιλέξουμε στην θέση της πιθανοφάνειας την σταθμισμένη πιθανοφάνεια $H_n(\theta)$. Άρα το WAIC, εκτός από μια γενική μορφή του AIC που αντιμετωπίζει τις ακραίες παρατηρήσεις ενός δείγματος, αποτελεί και μια βελτιωμένη μορφή ως προς την υπόθεση ότι το πραγματικό μοντέλο είναι το υποψήφιο μας.

Σταθμισμένο Akaike **Κριτήριο πληροφορίας** Agostinelli (wleAIC).

Ας υποθέσουμε ότι έχουμε ένα γραμμικό μοντέλο παλινδρόμησης της μορφής $Y = X\beta + \varepsilon$ όπου ο παραμετρικός χώρος είναι διάστασης p και τα σφάλματα ε έχουν κοινή διασπορά σ^2 που είναι άγνωστη. Σκοπός μας είναι να εκτιμήσουμε τους

συντελεστές $\vec{\beta}$. Κάνοντας το διάγραμμα διασποράς έχουμε ενδείξεις ότι υπάρχουν ακραίες παρατηρήσεις στο δείγμα μας και έτσι οι μέθοδοι εκτιμητριών μέγιστης πιθανοφάνειας και ελαχίστων τετραγώνων που συμπίπτουν στο απλό γραμμικό μοντέλο, δεν συνίστανται σαν λύσεις. Έχοντας αυτό υπόψη του Ο Agostinelli (2002) πρότεινε τις *σταθμισμένες εκτιμήτριες μέγιστης πιθανοφάνειας (W-ε.μ.π)* έτσι ώστε να κατασκευαστεί ένα *σταθμισμένο κριτήριο επιλογής μοντέλων*.

Έτσι, ενώ τις εκτιμήτριες μέγιστης πιθανοφάνειας τις βρίσκουμε λύνοντας τις εξισώσεις επιδόσεων της μορφής

$$\sum_{i=1}^{n} u(Y_i \mid x_i, \beta, \sigma) = 0,$$

οι σταθμισμένες εκτιμήτριες μέγιστης πιθανοφάνειας $\hat{\theta}_w = (\hat{\beta}_w, \hat{\sigma}_w)$ βρίσκονται λύνοντας τις εξισώσεων της μορφής

$$\sum_{i=1}^{n} w(y_i - x_i^t\beta, \sigma)\, u(Y_i \mid x_i, \beta, \sigma) = 0. \qquad (4.6)$$

Με $u(Y \mid x, \theta) = \partial \log f(Y \mid x, \theta)/\partial\theta$ και στις δύο παραπάνω περιπτώσεις. Η επιπλέον συνάρτηση βάρους που χρησιμοποιείτε ορίζεται ως

$$w(y_i - x_i^t\beta, \sigma) = \min\{1, \max(0, r(\hat{e}_i) + 1/(\hat{e}_i + 1))\},$$

όπου με $\hat{e}_i$ συμβολίζουμε τα *υπόλοιπα εξομάλυνσης κατά Pearson*, και για $r(a) = a$ παίρνουμε τις εξισώσεις επίδοσης των εκτιμητριών μέγιστης πιθανοφάνειας καθώς τότε όλα τα βάρη είναι ίσα με 1. Το κύριο χαρακτηριστικό της σταθμισμένης μεθόδου Agostinelli είναι η χρησιμοποίηση υπολοίπων εξομάλυνσης κατά Pearson για την κατασκευή της συνάρτησης βάρους αντί των κανονικών υπολοίπων $e_i = y_i - x_i^t\hat{\beta}$.

Τα *υπόλοιπα εξομάλυνσης κατά Pearson* ορίζονται ως

$$e_i = \left(\hat{f}_e(e_i) / \hat{m}(e_i) - 1 \right),$$

όπου $\hat{f}_e(t) = n^{-1}\sum_{i=1}^{n} h^{-1} K(h^{-1}(e_i - t))$ είναι η εκτιμήτρια της σ.π.π. των υπολοίπων με την μέθοδο των πυρήνων (kernels) και $\hat{m}(t) = \int h^{-1} K\left(h^{-1}(t - s)\right)\Phi(s, \sigma^2)\, ds$.

Έχοντας ορίσει τις σταθμισμένες εξισώσεις επιδόσεων όπως παραπάνω, ο Agostinelli πρότεινε τον μετασχηματισμό του AIC από την μορφή της Εξίσωσης 4.2. στην μορφή

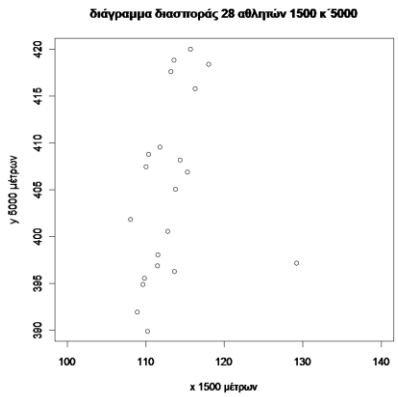
$$wleAIC = 2\sum_{i=1}^{n} w(y_i - x_i^t\hat{\beta},\hat{\sigma}) \log f(y_i \mid x_i,\hat{\beta},\hat{\sigma}) - 2(p+1), \qquad (4.7)$$

όπου οι εκτιμήτριες, $\hat{\theta} = (\hat{\beta},\hat{\sigma})$ παράγονται από την Εξίσωση 4.6, κατασκευάζοντας έτσι μια *σταθμισμένη στρατηγική επιλογής μοντέλου*.

Παράδειγμα 4.3: Εφαρμογή wleAIC Agostinelli.

Στο ευρωπαϊκό πρωτάθλημα αγώνων πατινάζ του 2004 που έγινε στην Heereveen της Ολλανδίας, ο Νορβηγός Eskin Ervik είχε μια πτώση στον τρίτο αγώνα δρόμου των 1500 μέτρων. Ο τελικός του χρόνος, 2:09:20, αποτελούσε μια μεμονωμένη ή ακραία παρατήρηση στη τελική λίστα αποτελεσμάτων των 28 αθλητών που συμμετείχαν, όπως φαίνεται εμφανώς και στο διάγραμμα 4.5.1. Θέλοντας να δείξουμε πόσο επηρεάζει μια ακραία παρατήρηση μια ανάλυση δεδομένων, εφαρμόζουμε ένα πολυωνυμικό μοντέλο παλινδρόμησης συσχετίζοντας τους χρόνους των αθλητών αυτών στους αγώνες των 1500 και 5000 μέτρων. Το πολυωνυμικό μοντέλο που εφαρμόζουμε είναι το

$Y = \beta_0 + \beta_1 X + ... + \beta_4 X^4 + \varepsilon$ με $\varepsilon \sim N(0,\sigma^2)$ όπου X, Y είναι οι χρόνοι σε δευτερόλεπτα των 1500 και 5000 μέτρων των 28 αθλητών αντίστοιχα.



Διάγραμμα 4.3:Διάγραμμα διασποράς 28 αθλητών στους αγώνες των 1500κ΄5000 μέτρων.

Από το διάγραμμα διασποράς 4.3 είναι εμφανές ότι το μοντέλο παλινδρόμησης που ταιριάζει καλύτερα στα δεδομένα μας είναι το απλό γραμμικό μοντέλο ($M1$: $Y=\beta_0 + \beta_1 X$), με την μεμονωμένη παρατήρηση όμως να αποτελεί ένα «αγκάθι» στην προσαρμογή αυτού του μοντέλου. Εφαρμόζοντας και τα τέσσερα πολυωνυμικά μοντέλα από την παραπάνω οικογένεια πολυωνυμικών μοντέλων θα προσπαθήσουμε να διαλέξουμε το καταλληλότερο με βάση το Akaike κριτήριο πληροφορία και το wleAIC σταθμισμένο κριτήριο πληροφορίας του Agostinelli. Εφαρμόζοντας και τα δύο κριτήρια στο δείγμα των 28 αθλητών αλλά και στο δείγμα των 27 αθλητών (αφαιρώντας την μεμονωμένη παρατήρηση της πτώσης του Ervik), σχολιάζουμε τα αποτελέσματα.

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα εφαρμογής των M1, M2, M3, M4 πολυωνυμικών μοντέλων παλινδρόμησης πάνω στα δεδομένα μας με 27 και 28 αθλητές με τις αντίστοιχες τιμές των AIC, WleAIC.

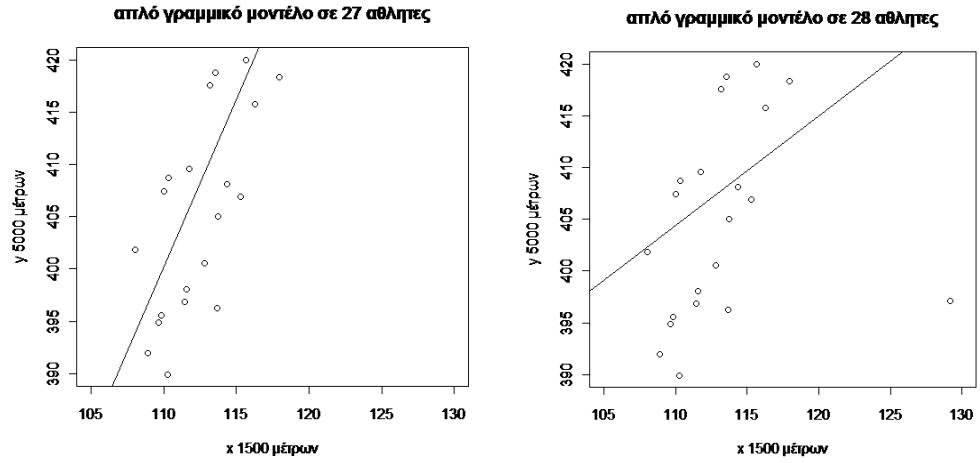
#	Αθλητές που συμμετέχουν σε αγώνες 1500 κ΄5000 μέτρων								
#	28 Αθλητές					27 Αθλητές			
#	Επεξ. μεταβλητές	AIC		wleAIC		AIC		WleAIC	
#	###	rank	score	rank	score	rank	score	rank	score
M1	x	2	227,1026	1	184,6	1	206,1201	1	183,6
M2	x, x^2	1	213,8950	2	201,3	2	207,4309	3	190,8
M3	x, x^2, x^3	3	215,8177	3	201,6	3	209,4139	2	193,6
M4	x, x^2, x^3, x^4	4	217,8161	4	204,3	4	210,5252	4	202,5

Πίνακας 4.4: Στοιχεία εφαρμογής AIC, WleAIC στα πολυωνυμικά μοντέλα παλινδρόμησης πάνω στα δείγματα των 28 και 27 αθλητών στους αγώνες των 1500, 5000 μέτρων.

Σχολιάζοντας τον παραπάνω Πίνακα 4.4, παρατηρούμε ότι το AIC στο δείγμα των 28 αθλητών προτείνει ως καταλληλότερο μοντέλο το M2 (AIC=213,8950) όπου $M2:\ Y=\beta_0+\beta_1 X+\beta_2 X^2$, με το απλό γραμμικό μοντέλο M1 να έρχεται δεύτερο στην κατάταξη και να ακολουθούν το M3 και το M4. Από την άλλη πλευρά, στο δείγμα των 27 αθλητών το AIC βλέπουμε ότι επιβεβαιώνει τον ισχυρισμό που διατυπώσαμε παραπάνω βλέποντας το διάγραμμα διασποράς των παρατηρήσεων μας, προτείνοντας το M1 (206,1201) ως καταλληλότερο μοντέλο με τα M2, M3, M4 να ακολουθούν. Συνεπώς, είναι εμφανής ο επηρεασμός της στρατηγικής AIC για την επιλογή του κατάλληλου μοντέλου από την μεμονωμένη παρατήρηση. Στατιστικά σωστή επιλογή μοντέλου θεωρείται αυτή που είναι καλύτερη για παρατηρήσεις απαλλαγμένες από τις ακραίες τιμές.

Θέλοντας να διορθώσουμε αυτή την αδυναμία του AIC εφαρμόζουμε το ενβαρές κριτήριο πληροφορίας του Agostinelli, WleAIC. Παρατηρούμε ότι το WleAIC πετυχαίνει να επιλέξει το απλό γραμμικό μοντέλο, M1 (WleAIC=184,6), ανάμεσα από όλα τα υποψήφια τα μοντέλα προτείνοντάς το ως καταλληλότερο, με τα M2, M3, M4 να ακολουθούν. Συνεπώς, συμπεραίνουμε ότι το WleAIC καταφέρνει να εξαλείψει το πρόβλημα που εισήγαγε η πτώση του Ervik στην ανάλυσή μας.

Παρουσιάζοντας το Διάγραμμα 4.5 και τον Πίνακα 4.6, γίνεται ακόμα πιο εμφανές το πόσο πολύ επηρεάζεται το μοντέλο που προσαρμόζουμε σε ένα σύνολο παρατηρήσεων όταν υπάρχουν μεμονωμένες παρατηρήσεις.



Διάγραμμα 4.5: Προσαρμογή του απλού γραμμικού μοντέλου στο δείγμα των 27 παρατηρήσεων (απαλλαγμένο από την ακραία παρατήρηση του Ervik) και στο δείγμα των 28 παρατηρήσεων.

#	Αθλητές που συμμετέχουν σε αγώνες 1500 κ΄5000 μέτρων			
#	28 Αθλητές		27 Αθλητές	
#	Intercept	slope	intercept	slope
AIC	288.22	1.06	46.03	3.22
WleAIC	46.87	3.21	46.64	3.21

Πίνακας 4.6: Πίνακας συντελεστών παλινδρόμησης εφαρμόζοντας τις στρατηγικές AIC, WleAIC στα δείγματα των 27 και 28 αθλητών.

Πολλές φορές η αναγνώριση των μεμονωμένων παρατηρήσεων δεν είναι τόσο εύκολη υπόθεση όπως στο συγκεκριμένο παράδειγμα. Έτσι, εκτός από το διάγραμμα διασποράς μπορούμε να χρησιμοποιήσουμε και τα διαγράμματα των μεταβλητών ξεχωριστά αλλά και την εμπειρία μας, έτσι ώστε να αποφασίσουμε να χρησιμοποιήσουμε μια ενβαρή μέθοδο. Από την άλλη πλευρά, η χρήση ενβαρών μεθόδων πρέπει να γίνεται με μεγάλη προσοχή, καθώς η χρήση τους σε παραδείγματα που δεν ενδείκνυται μπορεί να έχει ως αποτέλεσμα να χαθούν χρήσιμες πληροφορίες από το δείγμα μας.

Για την εξαγωγή των αποτελεσμάτων χρησιμοποιήθηκε η έκδοση της R 2.12.0 και η συνάρτηση wle.aic για το WleAIC που έχει κατασκευαστεί από τον ίδιο τον Agostinelli από τα πακέτα circular κ΄ wle ενώ για το AIC η συνάρτηση AIC.

Κεφάλαιο V

Μπεϋζιανό κριτήριο πληροφορίας (BIC) και σύγκρισή του με το (AIC)

5.1. Εισαγωγή

Στο 3^{ο} κεφάλαιο είδαμε ότι μια αξιόπιστη στρατηγική επιλογής μοντέλου μεταξύ διάφορων υποψήφιων μοντέλων είναι το Akaike κριτήριο πληροφορίας (AIC). Μια άλλη εξίσου αξιόπιστη και δημοφιλή στρατηγική αποτελεί και το Μπεϋζιανό κριτήριο πληροφορίας (BIC). Το εν λόγω κριτήριο αποτελεί προϊόν της Μπεΰζιανής στατιστικής, ποινικοποιώντας, όπως και το AIC, την μέγιστη πιθανοφάνεια ενός μοντέλου με σκοπό την προσέγγιση του πραγματικού μοντέλου. Το Μπεϋζιανό κριτήριο πληροφορίας παρουσιάστηκε από τον Akaike (1977) σαν μια προσπάθεια βελτίωσης του Akaike κριτηρίου πληροφορίας.

5.2. Γενική Μορφή του BIC (Bayesian information criterion)

Η γενική μορφή του Μπεϋζιανού κριτήριου πληροφορίας για ένα υποψήφιο μοντέλο M, με μέγεθος δείγματος n και διάσταση παραμετρικού χώρου $\dim(M)$ είναι

$$BIC(M) = 2l_{\max}(M) - (\log n)\dim(M). \qquad (5.1)$$

Είναι εμφανές ότι ο τύπος του BIC μοιάζει πάρα πολύ με αυτόν του AIC. Οι δύο αυτές στρατηγικές επιλογής μοντέλου χρησιμοποιούν σαν κύριο εργαλείο τους τις μέγιστες λογαριθμικές πιθανοφάνειες των υποψήφιων μοντέλων και προσπαθούν με ένα κατάλληλο όρο ποινικοποίησης, τον $2\dim(M)$ για το AIC και τον $(\log n)\dim(M)$ για το BIC, να διαλέξουν το μοντέλο που προσαρμόζεται καλύτερα στα δεδομένα.

Μια πρώτη σύγκριση των δύο όρων ποινικοποίησης δείχνει ότι και οι δύο εξαρτώνται γραμμικά από τον αριθμό των παραμέτρων που συμμετέχουν στο υπό εξέταση μοντέλο, ενώ στο BIC ο όρος ποινικοποίησης εξαρτάται και από τον λογάριθμο του αριθμού των παρατηρήσεων.

Το μοντέλο που έχει την μεγαλύτερη τιμή από όλα τα υποψήφια μοντέλα, σύμφωνα με τον τύπο 5.1 του Μπεϋζιανού κριτήριου πληροφορίας, αποτελεί και το προτεινόμενο μοντέλο από αυτή την στρατηγική. Συνήθως η τιμή του $BIC(M)$ είναι αρνητική. Έτσι το προτεινόμενο μοντέλο είναι αυτό με την μικρότερη απόλυτη τιμή.

Παραθέτουμε παρακάτω το παράδειγμα 4.2 υπολογίζοντας αυτή τη φορά και το BIC για κάθε υποψήφιο μοντέλο, παρουσιάζοντας έτσι μια εφαρμογή του BIC αλλά και μια σύγκριση του με το AIC.

Παράδειγμα 5.1.: Εφαρμογή BIC σε πολυπαραμετρικό γραμμικό μοντέλο παλινδρόμησης στα δεδομένα του Παραδείγματος 4.2 του 4^{ου} κεφαλαίου σελ 29-31.

Όπως μπορούμε να δούμε και στη σελίδα 29 στο συγκεκριμένο παράδειγμα, σκοπός μας είναι να προβλέψουμε την κατά μέσο όρο ελάχιστη ημερήσια θερμοκρασία το μήνα Ιανουάριο. Τα δεδομένα μας δίνονται στον Πίνακα 4.1 Ο πίνακας περιέχει τέσσερις μεταβλητές, μία μεταβλητή απόκρισης και τρείς επεξηγηματικές, ενώ τα υποψήφια μοντέλα είναι οκτώ.

Αριθμός μοντέλου	**Επεξηγηματικές μεταβλητές**	**Διασπορά υπολοίπων**	**K Αριθμός παραμέτρων**	**A**kaike information criterion (AIC)	Bayesian information criterion (BIC)
1	X_0, X_1	4,411	1	154,8727(4)	158,5293(3)
2	X_0, X_2	4,621	1	157,2947(6)	160,7282(5)
3	X_0, X_3	5,066	1	162,0716(7)	165,7282(7)
4	X_0, X_1, X_2	4,496	2	156,7647(5)	161,6402(6)
5	X_0, X_1, X_3	3,799	2	148,0040(1)	152,8759(1)
6	X_0, X_2, X_3	4,271	2	154,0933(3)	158,9689(4)
7	X_0, X_1, X_2, X_3	3,863	3	149,7176(2)	155,8120(2)
8	X_0	5,477	0	165,1907(8)	167,6284(8)

Πίνακας 5.1.: Αποτελέσματα εφαρμογής AIC, BIC στην R, στα αντίστοιχα υποψήφια μοντέλα στο Παράδειγμα 5.1 Στη στήλη με την τιμή του κάθε κριτηρίου μέσα στην παρένθεση, υπάρχει και η αντίστοιχη κατάταξη του κάθε μοντέλου.

Τόσο το AIC όσο και το BIC συμφωνούν απόλυτα στο μοντέλο που προτείνουν ως καταλληλότερο ($5^{ο}$ μοντέλο με τις επεξηγηματικές μεταβλητές X_0, X_1, X_3) καθώς και στο μοντέλο που θεωρούν πιο ακατάλληλο ($8^{ο}$ μοντέλο μόνο με την σταθερά X_0) στην αντίστοιχη κατάταξη τους. Στις ενδιάμεσες θέσεις της κάθε κατάταξης, εκεί που η διαφορά μεταξύ της βαθμολογίας των μοντέλων είναι μικρή, παρουσιάζονται οι διαφοροποιήσεις μεταξύ των δύο κριτηρίων. Βλέπουμε για παράδειγμα το $6^{ο}$ και το $1^{ο}$ μοντέλο να εναλλάσσουν τις θέσεις τους από το ένα κριτήριο στο άλλο. Αυτό οφείλεται στο γεγονός ότι ο όρος ποινικοποίησης του BIC είναι πιο «αυστηρός» στις μεταβολές του μεγέθους του παραμετρικού χώρου ενός μοντέλου. Έτσι η αύξηση των μεταβλητών κατά μία (στο $1^{ο}$ μοντέλο συμμετέχει μόνο η X_1 μεταβλητή, ενώ στο $6^{ο}$ μοντέλο συμμετέχουν οι X_2, X_3) υπερκαλύπτει την αύξηση της διασποράς κατά 0.14 (4,271-4,411). Μια γενική παρατήρηση επίσης είναι ότι το BIC χρησιμοποιεί ένα μικρότερο εύρος βαθμολογίας των υποψήφιων μοντέλων (152,8759-167,6284) σε σχέση με το AIC (148,004-165,1907). Αυτό οφείλεται στον μεγαλύτερο όρο ποινικοποίησης του BIC όταν το μέγεθος του δείγματος είναι $n > 8$.

Η απόδειξη του Μπεϋζιανού κριτήριου πληροφορίας γίνεται μέσω της Μπεΰζιανής στατιστικής που δεν αποτελεί αντικείμενο της συγκεκριμένης διπλωματικής. Μια λεπτομερή ανάλυση του BIC μπορεί κανείς να βρει στις εργασίες του Akaike (1978a, 1978b). Εμείς θα προσπαθήσουμε παρακάτω να εξετάσουμε το BIC μέσα από τη σύγκριση του με το AIC ως προς την συνέπεια (consistency) και την αποδοτικότητα (efficiency) τους .

5.3. Τρόποι Αξιολόγησης Κριτηρίων Πληροφορίας

5.3.1. Συνέπεια

Ένας τρόπος αξιολόγησης των κριτηρίων πληροφορίας είναι μέσω της συνέπειάς τους (consistency). Η συνέπεια ενός κριτηρίου πληροφορίας αξιολογείται με δύο τρόπους. Ο πρώτος τρόπος είναι η *ασθενής* και η *ισχυρή συνέπεια (weak and strong consistency)*, όπου αξιολογείται η ικανότητα του κριτηρίου πληροφορίας κατά πιθανότητα και σχεδόν παντού αντίστοιχα, να προτείνει ως καταλληλότερο το μοντέλο που ελαχιστοποιεί την απόκλιση K-L, καθώς το $n \to \infty$. Ο δεύτερος τρόπος είναι αν ένα κριτήριο πληροφορίας είναι συνεπές ή όχι. Η συνέπεια εδώ αναφέρεται στην ικανότητα του κριτηρίου πληροφορίας να τείνει να προτείνει σαν καταλληλότερο μοντέλο αυτό που όχι μόνο ελαχιστοποιεί την απόκλιση K-L, αλλά είναι και το πιο φειδωλό ως προς τον αριθμό των μεταβλητών του (parsimonious model). Παρακάτω θα προσπαθήσουμε, ορίζοντας ακριβέστερα τις παραπάνω έννοιες και χρησιμοποιώντας τρία Θεωρήματα, να

αξιολογήσουμε τα κριτήρια πληροφορίας AIC, BIC ως προς την ασθενή συνέπεια, την ισχυρή συνέπεια και συνέπεια τους.

Ασθενής και Ισχυρή Συνέπεια

Υποθέτοντας ότι υπάρχει ένα πραγματικό μοντέλο από το οποίο προέρχονται τα δεδομένα μας και ότι αυτό το μοντέλο είναι ένα από τα υποψήφια μοντέλα, από την μέθοδο επιλογής ζητάμε να αναγνωρίσει και να μας προτείνει το συγκεκριμένο μοντέλο ως καταλληλότερο για την ανάλυση των δεδομένων μάς. Η επιτυχία ή όχι της μεθόδου πάνω σε αυτό, σχετίζεται με την συνέπεια της συγκεκριμένης μεθόδου και κατά επέκταση και του κριτηρίου πληροφορίας που χρησιμοποιεί.

Συγκεκριμένα, μια μέθοδος επιλογής μοντέλων είναι *ασθενώς συνεπής* αν, με πιθανότητα να τείνει στο 1, είναι ικανή να επιλέξει το πραγματικό μοντέλο μεταξύ όλων των υποψήφιων μοντέλων. Από την άλλη πλευρά, *ισχυρώς συνεπής* είναι μια μέθοδος επιλογής αν σχεδόν παντού, επιτυγχάνει να επιλέξει το πραγματικό μοντέλο μεταξύ όλων των υποψήφιων μοντέλων.

Συχνά, δεν θέλουμε να κάνουμε την υπόθεση ότι το πραγματικό μοντέλο είναι ανάμεσα στα υποψήφια μοντέλα. Τότε μπορούμε να μιλήσουμε για ασθενή και ισχυρή συνέπεια όταν μια μέθοδος επιλογής τείνει να επιλέξει το μοντέλο, μεταξύ των υποψήφιων μοντέλων, που βρίσκεται πιο κοντά στο πραγματικό κατά την απόκλιση K-L. Η σύγκλιση που εννοείται στους παραπάνω ορισμούς, είναι για $n \to \infty$.

Πολλά κριτήρια πληροφορίας μπορούν να γραφούν με μια κοινή μορφή. Το AIC καθώς και το BIC όπως φαίνεται και από τις εξισώσεις 5.2, 5.3 είναι κάποια από αυτά.

$$AIC(M) = 2l_n(M) - 2\dim(M), \tag{5.2}$$

$$BIC(M) = 2l_n(M) - (\log n)\dim(M). \tag{5.3}$$

Τα δύο αυτά κριτήρια είναι κατασκευασμένα από το διπλάσιο της τιμής της μέγιστης λογαριθμικής πιθανοφάνειας $2l_n(M)$, μείον έναν όρο ποινικοποίησης που εξαρτάται από την πολυπλοκότητα του μοντέλου. Ο όρος ποινικοποίησης του BIC, $(\log n)\dim(M)$, είναι μεγαλύτερος και άρα αυστηρότερος ως προς την εισαγωγή μεταβλητών, από αυτόν του AIC, $2\dim(M)$ για $n > 8$. Έτσι, σαν συμπέρασμα θα λέγαμε ότι το BIC αποθαρρύνει την επιλογή μοντέλων με πολλές μεταβλητές σε σχέση με το AIC. Για να μπορέσουμε να εισάγουμε τα Θεωρήματα συνέπειας θα πρέπει να ορίσουμε μια συγκεκριμένη μορφή κριτηρίων στα οποία θα αναφερόμαστε.

Για τον ορισμό αυτής της γενικής μορφής των κριτηρίων πληροφορίας θα χρησιμοποιήσουμε κάποιους συμβολισμούς. Έστω $\kappa = 1,...,K$ τα υποψήφια μοντέλα και $i = 1,...,n$ οι παρατηρήσεις που έχουμε στην διάθεση μας. Θα συμβολίζουμε τις

παραμέτρους του κ μοντέλου ως $\vec{\theta}_\kappa$ και την συνάρτηση πιθανοφάνειας της i παρατήρησης στο μοντέλο αυτό ως $f_{\kappa,i}\left(y_i;\vec{\theta}_\kappa\right)$. Αντίστοιχα, στην περίπτωση της παλινδρόμησης ως $f_{\kappa,i}\left(y_i, x_i;\vec{\theta}_\kappa\right)$.

Τότε, η γενική μορφή των κριτηρίων που αναφερόμαστε είναι

$$IC(M_\kappa) = 2\sum_{i=1}^{n}\log f_{\kappa,i}\left(y_i, x_i;\hat{\vec{\theta}}_\kappa\right) - C_{n,\kappa}, \tag{5.4}$$

όπου $\hat{\vec{\theta}}_\kappa$ είναι οι εκτιμήτριες μέγιστης πιθανοφάνειας των $\vec{\theta}_\kappa$ και $C_{n,\kappa} = C(n)\,C(\kappa) > 0$ είναι ο όρος ποινικοποίησης του κριτηρίου για το μοντέλο M_κ με δείγμα μεγέθους n. Για παράδειγμα, στο AIC είναι $C_{n,\kappa} = 2\dim(\theta)$ ενώ στο BIC είναι $C_{n,\kappa} = \log n \dim(\theta)$. Ο πολλαπλασιαστής στον όρο της μέγιστης πιθανοφάνειας μπορεί να παραληφθεί από τα κριτήρια αυτού του τύπου, καθώς υπάρχει για καθαρά ιστορικούς λόγους. Άλλα παραδείγματα κριτηρίων αυτής της μορφής είναι τα AIC_c, TIC.

Τώρα είμαστε έτοιμοι να αναφερθούμε στα αντίστοιχα Θεωρήματα ασθενής και ισχυρής συνέπειας.

Θεώρημα 5.1: *Ασθενής Συνέπεια*. Έστω ότι ανάμεσα στα υποψήφια μοντέλα μας υπάρχει ένα και μόνο ένα μοντέλο M_{κ_0} τέτοιο ώστε να ελαχιστοποιείται η K-L απόκλιση. Έστω επίσης ότι ισχύουν οι παρακάτω προϋποθέσεις:

(a) $\liminf_{n\to\infty} \min_{\kappa\neq\kappa_0} n^{-1}\sum_{i=1}^{n}\left\{ KL\left(g; f_{\kappa,i}\right) - KL\left(g; f_{\kappa_0,i}\right)\right\} > 0$,

(b) Το αυστηρά θετικό κομμάτι του όρου ποινικοποίησης είναι της τάξης του $O_p(n)$.

Τότε, με πιθανότητα που τείνει στο 1, το κριτήριο πληροφορίας επιλέγει το μοντέλο M_{κ_0} ως το καταλληλότερο.

Συνέπεια Θεωρήματος 5.1: Για να είναι ένα κριτήριο ασθενώς συνεπές, θα πρέπει ο όρος ποινικοποίησης του, διαιρεμένος με το n, να τείνει στο 0 για $n\to\infty$.

Σημείωση: Η ασθενής συνέπεια ισχύει και στην οριακή περίπτωση όπου ένας από τους παράγοντες του $C_{n,\kappa}$ είναι 0, αρκεί τα άλλα κομμάτια να είναι αυστηρά θετικά.

Εφαρμόζοντας την παραπάνω συνέπεια του Θεωρήματος 5.1 για τα AIC, BIC όπου $C_{n,\kappa}$ είναι $2\dim(\theta)$ και $\log n \dim(\theta)$ αντίστοιχα, έχουμε ότι:

Για το AIC:

$$\frac{C_{n,\kappa}}{n} = \frac{2}{n}\dim(\theta_\kappa) \to_{n\to\infty} 0,$$

ενώ η οριακή περίπτωση όπου ένας παράγοντας του όρου ποινικοποίησης είναι ίσος με μηδέν δεν υφίσταται καθώς $\frac{2}{n} > 0$ και $\dim(\theta_\kappa) > 0$.

Για το BIC:

$$\frac{C_{n,\kappa}}{n} = \frac{\log n}{n}\dim(\theta_\kappa) \to_{n\to\infty} 0,$$

ενώ για την οριακή περίπτωση όπου $\log n = 0 \Rightarrow n = 1$, το $\dim(\theta_\kappa) > 0$.

Άρα το AIC αλλά και το BIC είναι ασθενώς συνεπή.

Συνεπώς, τόσο το AIC όσο και το BIC τείνουν να επιλέξουν μεταξύ των υποψήφιων μοντέλων σαν καταλληλότερο, το μοντέλο που ελαχιστοποιεί την απόκλιση K-L από το πραγματικό μοντέλο καθώς το $n \to \infty$.

Θεώρημα 5.2: *Ισχυρή Συνέπεια*. Έστω ότι ανάμεσα στα υποψήφια μοντέλα υπάρχει ένα και μόνο ένα μοντέλο που ελαχιστοποιεί την απόκλιση K-L από το πραγματικό μοντέλο. Έστω επίσης ότι ισχύουν οι παρακάτω υποθέσεις:

(a) $\liminf_{n\to\infty} \min_{\kappa\neq\kappa_0} n^{-1}\sum_{i=1}^{n}\left\{ KL(g; f_{\kappa,i}) - KL(g; f_{\kappa_0,i})\right\} > 0$,

(b) Το αυστηρά θετικό κομμάτι του όρου ποινικοποίησης να είναι της τάξης του $O(n)$ σχεδόν παντού στο n,

Τότε $P\left\{\min_{l\neq\kappa_0}\left(IC(M_{\kappa_0}) - IC(M_l)\right) > 0, \text{ για σχεδόν όλα τα n}\right\} = 1$.

Η υπόθεση για τον όρο ποινικοποίησης είναι προφανές ότι ικανοποιείται από τους στοχαστικούς όρους των AIC και BIC. Συνεπώς τα AIC και BIC είναι ισχυρώς συνεπή καθώς τείνουν να επιλέγουν το μοντέλο που ελαχιστοποιεί την απόσταση K-L από το πραγματικό μοντέλο για σχεδόν όλα τα n.

Συνέπεια

Στα παραπάνω δύο Θεωρήματα μιλήσαμε για μοναδικότητα μοντέλων που προσεγγίζουν την ελάχιστη απόκλιση κατά K-L από το πραγματικό μοντέλο, καθώς το $n \to \infty$. Τι γίνεται όταν δύο ή και περισσότερα μοντέλα προσεγγίζουν εξίσου καλά αυτή την ελάχιστη απόκλιση; Αυτό το φαινόμενο δεν είναι και τόσο σπάνιο στις στατιστικές αναλύσεις. Σε αυτή την ερώτηση έρχεται να δώσει απάντηση το τρίτο Θεώρημα. Το Θεώρημα συνέπειας αξιολογεί όχι μόνο εάν το μοντέλο που επιλέγεται από ένα κριτήριο πληροφορίας, τείνει να ελαχιστοποιήσει την απόκλιση K-L από το πραγματικό μοντέλο, αλλά και αν αυτό το μοντέλο είναι και το πιο φειδωλό από αυτά που τείνουν να ελαχιστοποιούν την απόσταση αυτή, καθώς το $n \to \infty$.

Έστω J το σύνολο των δεικτών των μοντέλων που τείνουν να ελαχιστοποιήσουν την απόσταση K-L, και $J_0 \subset J$ οι δείκτες των μοντέλων με τις λιγότερες παραμέτρους.

Θεώρημα 5.3: *Συνέπεια.* Έστω ότι ικανοποιείται ένα από τα παρακάτω σύνολά υποθέσεων:

i) Για $\kappa_0 \neq l_0 \in J$ ισχύει ότι,

(a) $\lim Sup_{n\to\infty} n^{-1} \sum_{i=1}^{n} \left\{ KL\left(g;\, f_{\kappa_0,i}\right) - KL\left(g;\, f_{l_0,i}\right) \right\} < \infty$,

(b) ενώ για τους όρους ποινικοποίησης ισχύει ότι $\forall j_0 \in J_0$ και $\forall l \in J \setminus J_0$,
$P\left\{ \left(\left(C_{n,l} - C_{n,j_0} \right) / \sqrt{n} \right) \to \infty \right\} = 1$.

ii) Για $\kappa_0 \neq l_0 \in J$ ισχύει ότι,

(a) ο λόγος λογαριθμικής πιθανοφάνειας $\sum_{i=1}^{n} \log \frac{f_{\kappa_0,i}\left(y_i;\theta^*_{\kappa_0}\right)}{f_{l_0,i}\left(y_i;\theta^*_{l_0}\right)} = O_p(1)$,

(b) ενώ για τους όρους ποινικοποίησης ισχύει ότι $\forall j_0 \in J_0$ και $\forall l \in J \setminus J_0$,
$P\left\{ \left(C_{n,l} - C_{n,j_0} \right) \to \infty \right\} = 1$.

Τότε, με πιθανότητα που τείνει στο 1, το κριτήριο πληροφορίας θα διαλέξει το μοντέλο που και θα ελαχιστοποιεί την απόσταση K-L από το πραγματικό μοντέλο και θα έχει και τις λιγότερες μεταβλητές (parsimonious model),

$$\lim_{n\to\infty} \left\{ \min_{l \in J \setminus J_0} \left(IC\left(M_{j_0}\right) - IC\left(M_l\right) \right) > 0 \right\} = 1.$$

Στο δεύτερο σύνολο υποθέσεων απαιτείται η κατανομή του στατιστικού λόγου λογαριθμικής πιθανοφάνειας να είναι φραγμένη. Οι ασυμπωτικές κατανομές ενός τέτοιου στατιστικού λόγου περιγράφονται επαρκώς από τον Vuong (1989).

Εφαρμόζοντας στα AIC, BIC την υπόθεση ii.b. έχουμε:

Για το BIC:

$$C_{n,l} = \log n \dim(\theta_l) \text{ και } C_{n,j_0=} \log n \dim\left(\theta_{j_o}\right).$$

$$\text{Επειδή } \dim(\theta_l) > \dim\left(\theta_{j_o}\right) \text{ και για n} \to \infty \text{ έχουμε ότι } \log n \to \infty,$$

$$P\left\{\left(C_{n,l} - C_{n,j_0}\right) \to \infty\right\} = 1.$$

Άρα ισχύει η υπόθεση μας. Συνεπώς το BIC είναι συνεπές.

Για το AIC:

$$C_{n,l} = 2\dim(\theta_l) \; και \; C_{n,j_0=} 2\dim\left(\theta_{j_o}\right).$$

$$\text{Επίσης } 2\dim(\theta_l) > 2\dim\left(\theta_{j_o}\right).$$

$$\text{Επείδη για n} \to \infty \text{ έχουμε ότι } 2\left(\dim(\theta_l) - \dim\left(\theta_{j_o}\right)\right) = σταθερο,$$

$$P\left\{\left(C_{n,l} - C_{n,j_0}\right) \to \infty\right\} \neq 1.$$

Άρα δεν ισχύει η υπόθεση μας. Συνεπώς το AIC δεν είναι συνεπές.

Συνοψίζοντας τα παραπάνω περί συνέπειας, θα λέγαμε ότι η ασθενής και η ισχυρή συνέπεια έχουν να κάνουν περισσότερο με την ασθενή και την ισχυρή σύγκλιση όταν το $\text{n} \to \infty$, καθώς αναφέρονται στην σύγκλιση κατά πιθανότητα και σχεδόν παντού αντίστοιχα. Ως προς την ασθενή και την ισχυρή συνέπεια δεν έχουμε διαφορές ανάμεσα στα δύο κριτήρια πληροφορίας που εξετάζουμε AIC, BIC. Από την άλλη πλευρά, το BIC είναι συνεπές ενώ το AIC δεν είναι, σύμφωνα με το Θεώρημα 5.3. Συνεπώς τόσο το AIC όσο και το BIC τείνουν να επιλέξουν ανάμεσα στα υποψήφια μοντέλα αυτό που ελαχιστοποιεί την K-L απόκλιση από το πραγματικό μοντέλο, όταν αυτό το μοντέλο είναι μοναδικό, καθώς το $\text{n} \to \infty$ (ασθενής και ισχυρή συνέπεια). Όταν όμως υπάρχουν πολλά ανταγωνιστικά μοντέλα που ελαχιστοποιούν εξίσου την απόκλιση K-L από το πραγματικό μοντέλο, τότε μόνο το BIC τείνει να επιλέξει το πιο φειδωλό από αυτά (συνέπεια).

5.3.2. Αποδοτικότητα

Ένας άλλος τρόπος αξιολόγησης ενός κριτηρίου πληροφορίας είναι κατά πόσο συμπεριφέρεται «σχεδόν εξίσου καλά» με την επιλογή του καλύτερου θεωρητικά μοντέλου με βάση το τετραγωνικό σφάλμα απώλειας. Για παράδειγμα, όταν έχουμε ένα μοντέλο πρόβλεψης μπορούμε να αξιολογήσουμε ένα κριτήριο πληροφορίας με το αν συμπεριφέρεται «σχεδόν εξίσου καλά» με την επιλογή του θεωρητικά καλύτερου μοντέλου που προτείνει το αναμενόμενο τετραγωνικό σφάλμα πρόβλεψης. Όταν ένα κριτήριο πληροφορίας ικανοποιεί την παραπάνω ιδιότητα αποκαλείται αποδοτικό (efficient).

Παρακάτω θα προσπαθήσουμε να δούμε πώς εξετάζουμε αυτή την «σχεδόν εξίσου καλή» συμπεριφορά ενός κριτηρίου πληροφορίας με το τετραγωνικό σφάλμα απώλειας μέσω ενός μοντέλου πρόβλεψης.

Έστω ότι έχουμε να διαλέξουμε το καλύτερο σύνολο μεταβλητών στο παρακάτω μοντέλο παλινδρόμησης

$$\vec{Y}_i = \beta_0 + \beta_1 \vec{X}_{1i} + \beta_2 \vec{X}_{2i} + ... + \beta_k \vec{X}_{ki} + \vec{\varepsilon}_i \text{ για } i = 1,..., n,$$

όπου $Var(\vec{\varepsilon}_i) = \sigma^2$ και $\vec{X}_i = \vec{X}_{1,i}, \vec{X}_{2,i},..., \vec{X}_{k,i}$ επεξηγηματικές μεταβλητές. Σκοπός μας είναι η πρόβλεψη μιας καινούργιας ανεξάρτητης μεταβλητής απόκρισης $\hat{Y}_i$. Έστω $S = P(X)$ το δυναμοσύνολο του συνόλου $X = \{X_1, X_2,..., X_k\}$. Εμείς θέλουμε να διαλέξουμε αυτό το υποσύνολο επεξηγηματικών μεταβλητών του S που ελαχιστοποιεί το αναμενόμενο σφάλμα πρόβλεψης υπό το παρατηρούμενο σύνολο δεδομένων $y_1, y_2,..., y_n$. Άρα προσπαθούμε να ελαχιστοποιήσουμε την ποσότητα

$$\sum_{i=1}^{n} E\left[\left(\hat{\vec{Y}}_{s,i} - \vec{Y}_{true,i}\right)^2 \mid y_1, y_2,..., y_n\right], \qquad (5.5)$$

όπου $\vec{Y}_{true,i}$ είναι ανεξάρτητες των $Y_1, Y_2,..., Y_n$ αλλά προέρχονται από την ίδια κατανομή. Οι προβλεπόμενες τιμές $\hat{\vec{Y}}_{s,i}$ είναι για τα διάφορα υποσύνολα $s \subset S = P(X)$ και εξαρτώνται από τα $y_1, y_2,..., y_n$.

Χρησιμοποιώντας την ανεξαρτησία μεταξύ των νέων παρατηρήσεων από τις $y_1, y_2,..., y_n$, μπορούμε να γράψουμε ότι

Για διαφορετικού τύπου μοντέλα, όπως για παράδειγμα τα αυτοσυσχετιζόμενα μοντέλα για τετραγωνικό σφάλμα απώλειας χρησιμοποιούμε το μέσο τετραγωνικό σφάλμα.

$$\sum_{i=1}^{n} E\left[\left(\hat{\vec{Y}}_{s,i} - \vec{Y}_{true,i}\right)^2 \mid y_1, y_2, ..., y_n\right] = \sum_{i=1}^{n} E\left[\left(\hat{\vec{Y}}_{s,i} - E\left[\vec{Y}_{true,i}\right]\right)^2 \mid y_1, y_2, ..., y_n\right] + n\sigma^2$$

$$\sum_{i=1}^{n}\left(\hat{\vec{\beta}}_s - \vec{\beta}_{true}\right)^t \vec{X}_i^t \vec{X}_i \left(\hat{\vec{\beta}}_s - \vec{\beta}_{true}\right) + n\sigma^2,$$

όπου το διάνυσμα $\hat{\vec{\beta}}_s$, που περιέχει τους συντελεστές παλινδρόμησης των μεταβλητών που περιέχονται στο σύνολο $s \subset S$ και παίρνουν μέρος στο μοντέλο πρόβλεψης των $\hat{\vec{Y}}_{s,i}$, έχει μηδενικά για όσες επεξηγηματικές μεταβλητές δεν ανήκουν στο σύνολο s. Το διάνυσμα $\hat{\vec{\beta}}_{true}$ αποτελείται από τους συντελεστές παλινδρόμησης του μοντέλου που συμμετέχουν όλες οι μεταβλητές $\vec{X}_i$ για την παραγωγή των $\vec{Y}_{true,i}$.

Αν ένα κριτήριο πληροφορίας επιλέγει εκείνο το μοντέλο πρόβλεψης των $\hat{\vec{Y}}_{\hat{s}_0,i}$ που πιάνει το κατώτερο όριο της ποσότητας 5.5, καθώς το $n \to \infty$, τότε λέμε *ότι το* κριτήριο πληροφορίας είναι αποδοτικό υπό το συγκεκριμένο σύνολο δεδομένων $y_1, y_2, ..., y_n$.

Εκτός από την δεσμευμένη αναμενόμενη ποσότητα 5.5 μπορούμε, θεωρητικά, να δουλέψουμε και με το μη δεσμευμένο σφάλμα πρόβλεψης

$$L_n(s) = \sum_{i=1}^{n} E\left[\left(\hat{\vec{Y}}_{s,i} - \vec{Y}_{true,i}\right)^2\right] \tag{5.6}$$

όπου για την συγκεκριμένη περίπτωση της γραμμικής παλινδρόμησης γίνεται

$$L_n(s) = E\left[\sum_{i=1}^{n}\left(\hat{\vec{\beta}}_s - \vec{\beta}_{true}\right)^t \vec{X}_i^t \vec{X}_i \left(\hat{\vec{\beta}}_s - \vec{\beta}_{true}\right)\right] + n\sigma^2 .$$

Συμβολίζοντας με $\hat{s}_0^* \subset S$ το σύνολο των επεξηγηματικών μεταβλητών του μοντέλου όπου ελαχιστοποιείται το αναμενόμενο σφάλμα πρόβλεψης και με $\hat{s}_0 \subset S$ το σύνολο των μεταβλητών του μοντέλου που επιλέχτηκε από το κριτήριο πληροφορίας, έχουμε ότι ένα κριτήριο πληροφορίας είναι αποδοτικό όταν

$$\frac{\sum_{i=1}^{n} E_{\hat{s}_0}\left[\left(\hat{\vec{Y}}_{\hat{s}_0,i} - \vec{Y}_{true,i}\right)^2\right]}{\sum_{i=1}^{n} E\left[\left(\hat{\vec{Y}}_{s_0^*,i} - \vec{Y}_{true,i}\right)^2\right]} = \frac{L_n(\hat{s}_0)}{L_n(s_0^*)} \to_p 1 \text{ για } n \to \infty,$$

όπου το E_{s_0} συμβολίζει ότι η αναμενόμενη τιμή υπολογίζεται ως προς όλα τα υποσύνολα του S εκτός από το $\hat{s}_0$.

Παρομοίως, ένα κριτήριο πληροφορίας καλείται αποδοτικό αν επιλέγει εκείνο το μοντέλο όπου ο λόγος της αναμενόμενης συνάρτησης απώλειας του μοντέλου αυτού, προς την ελάχιστη αναμενόμενη, θεωρητικά, απώλεια που μπορεί να προκύψει από τα προτεινόμενα μοντέλα , τείνει στο 1 κατά πιθανότητα.

Το παρακάτω Θεώρημα αποδεικνύει ότι τα κριτήρια πληροφορίας AIC, AIC_c, είναι ασυμπωτικά αποδοτικά ενώ το BIC δεν είναι.

Υποθέσεις Θεωρήματος 5.4.

1η υπόθεση: Έστω ότι για όλα τα διαφορετικά υποσύνολα επεξηγηματικών μεταβλητών s του δυναμοσυνόλου S, ο αντίστοιχος πίνακας σχεδιασμού X_s έχει μέγιστο βαθμό $|s|$, όπου $|s|$ συμβολίζουμε τον αριθμό των μεταβλητών του συνόλου s, και $\max_s |s| = O(n^a)$ για σταθερά $a \in (0,1]$.

2η υπόθεση: Μία άλλη τεχνική υπόθεση που απαιτείται για την ύπαρξη της μοναδικότητας είναι ότι για $b \in [0, 0.5)$ και για κάθε $c > 0$,

$$\sum_s \exp\{-cn^{-2b} L_n(s)\} \to 0, \text{ για } n \to \infty.$$

Εφαρμόζοντας τα παραπάνω για όλα τα $s \subset S$ έχουμε ότι $n^{-2b} L_n(s) \to \infty$.

Σημείωση: Αυτή η υπόθεση δεν χρειάζεται όταν ο αριθμός των μεταβλητών παλινδρόμησης στο πραγματικό μοντέλο είναι άπειρος ή όταν ο αριθμός των μεταβλητών αυξάνεται με την αύξηση του μεγέθους του δείγματος. Οι Hurvich και Tsai (1995) αναλύουν διεξοδικά τις τεχνικές υποθέσεις του παρακάτω Θεωρήματος.

3η υπόθεση: Τα κριτήρια πληροφορίας για τα οποία εφαρμόζεται το παρακάτω Θεώρημα έχουν την μορφή

$$IC(s) = (n + 2|s|)\hat{\sigma}_s^2, \qquad (5.7)$$

όπου η διασπορά του μοντέλου s εκτιμάται από το

$$n^{-1}(Y - X_s\beta_s)^t (Y - X_s\beta_s) = SSE_S / n.$$

Το AIC και το BIC προφανώς και μπορούν να γραφθούν σε αυτή την μορφή.

Θεώρημα 5.4: Έστω $\hat{s}_0$ το σύνολο των επεξηγηματικών μεταβλητών που επιλέχθηκαν για την ελαχιστοποίηση του κριτηρίου της μορφής 5.7. Αν ισχύουν οι παραπάνω τρείς υποθέσεις τότε για s_0^* το σύνολο των επεξηγηματικών μεταβλητών που ελαχιστοποιούν την $L_n(s)$ στην εξίσωση 5.6., και θέτοντας $c = \min\{(1-a)/2, b\}$ έχουμε ότι

$$\frac{L_n(\hat{s}_0)}{L_n(s_0^*)} - 1 = O_p(n^{-c}).$$

Για την απόδειξη του παραπάνω θεωρήματος παραπέμπουμε στους Hurvich και Tsai (1995).

Συνέπειες Θεωρήματος 5.4:

(a) Τα κριτήρια πληροφορίας $AIC, AIC_c = AIC(s) + 2(|s|+1)(|s|+2)/(n-|s|+2)$ είναι ασυμπωτικά αποδοτικά κάτω από τις υποθέσεις του Θεωρήματος 5.1.,

(b) Το κριτήριο πληροφορίας BIC δεν είναι ασυμπωτικά αποδοτικό.

Για την απόδειξη των παραπάνω συνεπειών παραπέμπουμε στον Shibata (1980, Θεώρημα 2).

Συνοψίζοντας θα λέγαμε ότι το AIC και το $AICc$ είναι αποδοτικά, σύμφωνα με την συνέπεια του Θεωρήματος 5.3., ενώ το BIC δεν είναι αποδοτικό. Συνεπώς, τα AIC, AIC_c τείνουν να επιλέξουν σαν καταλληλότερο από τα υποψήφια μοντέλα το ίδιο μοντέλο που επιλέγεται και σαν καταλληλότερο από την ελαχιστοποίηση του τετραγωνικού σφάλματος απώλειας (στην περίπτωση του μοντέλου πρόβλεψης, το αναμενόμενο τετραγωνικό σφάλμα πρόβλεψης), καθώς το $n \to \infty$.

Συμπερασματικά

Ανακεφαλαιώνοντας, τόσο το AIC όσο και το BIC επιτυγχάνουν να προτείνουν σαν καταλληλότερο μοντέλο το μοντέλο που ελαχιστοποιεί την απόσταση K-L από το πραγματικό μοντέλο (ασθενής και ισχυρή συνέπεια). Από την άλλη πλευρά, μόνο το BIC επιτυγχάνει να προτείνει το πιο φειδωλό μοντέλο που ελαχιστοποιεί την K-L απόσταση από την πραγματική κατανομή (συνέπεια) ως καταλληλότερο. Επίσης, το AIC επιτυγχάνει να επιλέγει σαν καταλληλότερο μοντέλο αυτό που επιλέγεται και με την ελαχιστοποίηση του τετραγωνικού σφάλματος, ενώ το BIC όχι (αποδοτικότητα).

Επίλογος

Στην εν λόγω διπλωματική εξετάσαμε πώς τα κριτήρια πληροφορίας, AIC, TIC, AICc και WAIC, βοηθάνε στην επιλογή των μεταβλητών που θα εισάγουμε στο τελικό μοντέλο παλινδρόμησης μιας ανάλυσης δεδομένων. Τόσο το AIC, όσο και οι επεκτάσεις TIC, AICc και WAIC, αποτελούν κριτήρια πληροφορίας ποινικοποίησης της μέγιστης πιθανοφάνειας κάθε μοντέλου. Τα βασικά δομικά υλικά του AIC είναι οι εκτιμήτριες μέγιστης πιθανοφάνειας και η απόκλιση K-L. Οι εκτιμήτριες μέγιστης πιθανοφάνειας βοηθάνε στην εύρεση της μέγιστης πιθανοφάνειας κάθε μοντέλου, ενώ η απόκλιση K-L βοηθάει στην ελαχιστοποίηση της απόκλισης μεταξύ του επιλεγόμενου μοντέλου-κατανομής (προσαρμοσμένη κατανομή) και της κατανομής που προέρχονται τα δεδομένα (πραγματική κατανομή).

Το TIC αποτελεί μια διόρθωση του AIC ως προς την υπόθεση ότι η πραγματική κατανομή είναι ανάμεσα στα προτεινόμενα μοντέλα. Όσον αφορά το AICc, αποτελεί μια διόρθωση του AIC χρησιμοποιώντας αμερόληπτες εκτιμήτριες αντί για εκτιμήτριες μέγιστης πιθανοφάνειας, διορθώνοντας έτσι την μεροληψία του AIC ως προς το μέγεθος του δείγματος n. Επίσης παρουσιάζεται η σταθμισμένη διόρθωση του AIC, η WAIC όπως ονομάζεται από τον Agostinelli, η οποία προσπαθεί να αντιμετωπίσει την ύπαρξη ακραίων παρατηρήσεων στα δεδομένα εισάγοντας μια συνάρτηση βάρους στην πιθανοφάνεια κάθε μοντέλου χρησιμοποιώντας τα υπόλοιπα εξομάλυνσης Pearson.

Τέλος παρουσιάζεται μια σύγκριση του AIC με το BIC, ως προς την ασθενή συνέπεια, την ισχυρή συνέπεια, την συνέπεια και την αποδοτικότητα. Η σύγκριση αυτή μέσα από συγκεκριμένα Θεωρήματα εξάγει δύο βασικά συμπεράσματα. Πρώτον, και τα δύο κριτήρια, AIC και BIC, καταφέρνουν να επιλέξουν το μοντέλο που είναι πιο κοντά στην πραγματική κατανομή των δεδομένων σύμφωνα με την απόσταση K-L, καθώς το $n \to \infty$. Δεύτερον, το μεν AIC πετυχαίνει να επιλέγει το μοντέλο που επιλέγεται και με την ελαχιστοποίηση του τετραγωνικού σφάλματος, καθώς το $n \to \infty$ (αποδοτικότητα) το δε BIC πετυχαίνει να επιλέγει το πιο φειδωλό μοντέλο από αυτά που ελαχιστοποιούν την K-L απόσταση, μεταξύ προσαρμοσμένης και πραγματικής κατανομής των δεδομένων (συνέπεια).

Ένα ερώτημα που γεννάται άμεσα είναι πώς θα μπορούσαμε να συνδυάσουμε την αποδοτικότητα του AIC με την συνέπεια του BIC; Πάνω σε αυτό το ερώτημα ο Yang (2005) απέδειξε ότι δεν είναι δυνατόν μέσα από την Μπεϋζιανή στάθμιση μοντέλων να συνδυαστούν τα θετικά των δύο αυτών κριτηρίων. Υπάρχει κάποιος τρόπος μέσα από την στάθμιση μοντέλων ή οποιαδήποτε άλλη στρατηγική που να μπορεί να δώσει απάντηση σε αυτό το ερώτημα; Αυτό μένει να αποδειχθεί από την επιστημονική κοινότητα στο μέλλον.

Βιβλιογραφία

Α) Αγγλική

Akaike, H. (1969). Fitting autoregressive model for prediction. *Ann. Inst. Statist. Math.*. 21. 243–247.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csaki, F. (editors), *Second International Symposium on Information Theory*. Akademiai Kiado. Budapest. 267–281.

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

Akaike, H. (1977). On entropy maximization principle. *In Applications of Statistics (Proceedings of Symposium, Wright State University, Dayton, Ohio, 1976)*. North-Holland, Amsterdam. 27–41

Akaike, H. (1978a). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math*. 30. 9–14.

Akaike, H. (1978b). A new look at the Bayes procedure. *Biometrika*. 65. 53–59.

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika*. 66. 237-242.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrica.* 52. *Ann. Inst. Statist. Math.* 317-332.

Agostinelli, C. (2002). Robust model selection in regression via weighted likelihood methodology. *Statistics & Probability Letters*. 56. 289–300.

Bozdogan, H. *(*1987*).* Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika.* 52. 345-370.

Burnhmam, K. (2002*). Model selction and multimodel inference. 2*nd edition. Springer New York.

Burnhmam, K. and Anderson. R. (2004). Understanding AIC and BIC in Model Selection. Socialogical Methods & Research 2004. *33. 261-304.*

Claeskens, G. (2008). *Model selection and model averaging. 2nd edition*. Cambridge university press.

Czepiel, S. (2008) Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation. *logistic-regression-notes.*

Dalgaard, P. (2008). *Introductory Statistic with R. 2nd Edition*. Springer New York.
Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku* (*Mathematic Sciences*). 153. 12–18.

Terasvirta, T. (1986). Model selection criteria and model selection tests in regression models. 13. 159-171.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? *Biometrika*, 92. 937–950.

Β) Ελληνική

Καρώνη Χ. *Ανάλυση Παλινδρόμησης* (2007). Εκδόσεις Ε.Μ.Π. Αθήνα.

Κοκολάκης Γ. και Σπηλιώτης, Ι. (2002). *Εισαγωγή στις Πιθανότητες*. Εκδόσεις Συμεών. Αθήνα.

Κοκολάκης, Γ. και Φουσκάκης, Δ. (2005). *Σημειώσεις Στατιστικής*. Εκδόσεις Ε.Μ.Π. Αθήνα.

Draper, N.R. and Smith, H (1997). Εφαρμοσμένη ανάλυση παλινδρόμησης. 2η αγγλική έκδοση. Μετάφραση Χατζηκωνσταντινίδης. Ε. και Καλαματιανού. Α. Εκδόσεις Παπαζήση Αθήνα.