ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Διπλωματική Εργασία

Κριτήρια Επιλογής Στατιστικών Μοντέλων

Λουκία Γεωργάτου-Πολίτου

Τριμελής Επιτροπή

Δημήτριος Φουσκάκης (Επιβλέπων), Αναπληρωτής Καθηγητής
Μιχαήλ Λουλάκης, Αναπληρωτής Καθηγητής
Βασίλης Παπανικολάου, Καθηγητής

Αθήνα, 14 Σεπτεμβρίου 2017

Περιεχόμενα

Ευχαριστίες

Καταρχάς, θα ήθελα να ευχαριστήσω θερμά τον Αναπληρωτή Καθηγητή και επιβλέποντα της παρούσας διπλωματικής εργασίας, κύριο Δημήτριο Φουσκάκη, για την πολύτιμη βοήθεια, τη μεγάλη υπομονή και κατανόησή του κατά την εκπόνησή της, αλλά ακόμα περισσότερο για το γεγονός πως μέσω των μαθημάτων και του τρόπου διδασκαλίας του συνέβαλε τα μέγιστα στην ανάπτυξη του ενδιαφέροντός μου για τη Στατιστική. Επίσης, θα ήθελα να ευχαριστήσω τον Αναπληρωτή Καθηγητή, κύριο Μιχαήλ Λουλάκη, και τον Καθηγητή, κύριο Βασίλη Παπανικολάου, για την τιμή που μου κάνουν να συμμετάσχουν στην επιτροπή εξέτασης της εργασίας. Ευχαριστώ, ακόμα, την επικεφαλής του τμήματος Βιοστατιστικής του Deutsches Krebsforschungszentrum, καθηγήτρια Annete Kopp-Schneider, για τη βοήθειά της στην εύρεση των πραγματικών δεδομένων που χρησιμοποιήθηκαν στην παρούσα εργασία.

Επίσης, ευχαριστώ πολύ την Ελενίτσα, που είναι πάντοτε δίπλα μου όσο μακρυά και αν βρίσκεται, την Εύα για την συμπαράσταση και τις συζητήσεις μας, αλλά και τους Πάνο και Εμμανουέλα για τη φιλοξενία και την απέραντη γενναιοδωρία τους, ιδιαιτέρως τη φετινή χρονιά.

Τέλος, το ευχαριστώ είναι πολύ λίγο για να εκφράσει την ευγνωμοσύνη μου προς την οικογένειά μου για την αγάπη και την αδιάλλειπτη στήριξή της σε όλα τα επίπεδα και υπό όλες τις συνθήκες, αλλά και για την πίστη των μελών της σε εμένα ακόμα κι όταν εγώ την χάνω.

Περίληψη

Η παρούσα διπλωματική εργασία αποτελείται από δύο ενότητες. Η πρώτη εστιάζεται στην παρουσίαση και ανάλυση των Κριτηρίων Πληροφορίας AIC (*Akaike Information Criterion*), BIC (*Bayesian Information Criterion*), EIC (*Extended Information Criterion*), FIC (*Focussed Information Criterion*) καθώς και ορισμένων επιλεγμένων, λιγότερο εκτενώς χρησιμοποιούμενων. Επίσης, γίνεται αναφορά στη Στάθμιση Μοντέλων και, τέλος, αξιολόγηση των Κριτηρίων AIC και BIC. Η δεύτερη ενότητα περιλαμβάνει την εφαρμογή μεθόδων επιλογής μεταβλητών για πραγματικά και προσομοιωμένα δεδομένα με χρήση κώδικα της R, ο οποίος παρατίθεται στο Παράρτημα.

Abstract

The present diploma thesis comprises two parts. The first one covers the description and analysis of the Information Criteria AIC (*Akaike Information Criterion*), BIC (*Bayesian Information Criterion*), EIC (*Extended Information Criterion*), FIC (*Focussed Information Criterion*) and of other selected, less extensively used Criteria. It also includes Model Averaging methodology and an evaluation of AIC and BIC. In the second part applications of model selection techniques using real as well as simulated data are presented. The relevant R code can be found in the Appendix.

Εισαγωγή

Στην Επιστήμη συχνά προσπαθούμε να κατανοήσουμε φαινόμενα με αβεβαιότητα, να προσδιορίσουμε τη δομή περίπλοκων συστημάτων και να προβούμε σε αξιόπιστες προβλέψεις σε σχέση με αυτά. Ένα πολύ βασικό και χρήσιμο εργαλείο για την επίτευξη των παραπάνω είναι τα στατιστικά μοντέλα, τα οποία κατέχουν κυρίαρχο ρόλο στην ανάλυση δεδομένων. Τα στατιστικά μοντέλα είναι κατανομές πιθανότητας, στις οποίες χρησιμοποιούνται τα δεδομένα που παρατηρήθηκαν με σκοπό να προσεγγιστούν οι πραγματικές κατανομές από τις οποίες αυτά προέρχονται. Επομένως, κάθε στατιστικός, δοθέντος ενός συνόλου δεδομένων, δύναται να επιλέξει ανάμεσα σε πληθώρα μοντέλων, πρέπει όμως να χρησιμοποιήσει το καταλληλότερο εξ αυτών, ώστε να περιγραφούν εύστοχα τα δεδομένα.
Πάνω σ΄αυτό το ζήτημα, υπάρχουν κάποια σημεία-κλειδιά που αφορούν γενικώς την επιλογή στατιστικών μοντέλων και οφείλουμε να συζητήσουμε:

- ***Τα μοντέλα είναι προσεγγιστικά***: Όταν κληθούμε είτε να κατασκευάσουμε είτε να επιλέξουμε ένα μοντέλο, πρέπει να γίνει αντιληπτό πως πιθανότατα δε θα μπορέσουμε να εικάσουμε το «σωστό» μοντέλο, το οποίο σχέδον πάντα είναι άγνωστο ή και περίπλοκο. Για πρακτικούς, λοιπόν, λόγους είναι προτιμότερο να δουλέψουμε με απλούστερα αλλά σχεδόν «σωστά» μοντέλα που αντικατοπτρίζουν μέρος μόνο του πραγματικού φαινομένου. Γι' αυτό το λόγο πολλές μέθοδοι επιλογής υιοθετούν και βασίζονται στη ρήση *«όλα τα μοντέλα είναι λάθος, αλλά κάποια απ' αυτά είναι χρήσιμα»* (G.E.P. Box).
- ***Συμβιβασμός μεροληψίας-διασποράς (The bias-variance trade-off)***: Στην προσαρμογή μοντέλων, γνωρίζουμε ότι λιγότερες παράμετροι προς εκτίμηση ναι μεν οδηγούν σε μικρότερη διασπορά, αλλά σχετίζονται με την αύξηση της μεροληψίας, ενώ για μεγάλο αριθμό παραμέτρων έχουμε μικρότερη μεροληψία, αυξημένη, όμως, διασπορά. Συνεπώς, κατά την επιλογή μοντέλου πρέπει να βρεθεί μία ισορροπία ανάμεσα σ' ένα μοντέλο υπεραπλουστευμένο (*underfitting*) που έχει υπερβολικά λίγες παραμέτρους και σ'ένα υπέρ το δέον περίπλοκο (*overfitting*) με υπερβολικά μεγάλο πλήθος παραμέτρων, ώστε να μην οδηγηθούμε σε παραπλανητικά και ανακριβή συμπεράσματα.

- ***Φειδώ(Parsimony)***: Η ιδέα της φειδούς βασίζεται στο «Νόμο της Συντομίας», ο οποίος αναφέρει πως δεν πρέπει να χρησιμοποιούνται περισσότερα πράγματα από αυτά που χρειάζεται. Έτσι, φειδωλά είναι τα μοντέλα εκείνα που έχουν την ικανότητα να εξηγούν επαρκώς τα δεδομένα κάνοντας χρήση μόνο όσων παραμέτρων ή επεξηγηματικών μεταβλητών είναι απαραίτητες γι' αυτό τον σκοπό. Επομένως, μόνο οι παράμετροι που πραγματικά έχουν σημασία οφείλουν να εισαχθούν σ' ένα μοντέλο, διότι διαφορετικά το μοντέλο περιπλέκεται, γεγονός που υπονομεύει την εξαγωγή χρήσιμων αποτελεσμάτων.
- ***Εστίαση(Focus)***: Καθότι κάποιες ποσότητες συχνά θεωρούνται σημαντικότερες από άλλες, είναι πιο αποτελεσματικό η επιλογή και η κατασκευή μοντέλων να πραγματοποιείται με στρατηγικές επιλογής που ευνοούν αυτές τις σημαντικές ποσότητες. Φυσικά, έτσι μπορεί να υπάρξουν διαφοροποιήσεις στις επιλογές μοντέλων, αφού οι σημαντικές σε κάθε περίπτωση ποσότητες δεν είναι απαραιτήτως ταυτόσημες για κάθε ερευνητή, κάτι που όμως είναι φυσιολογικό, αφού αντικατοπτρίζει διαφορετικές προτιμήσεις.
- ***Αντικρουόμενες προτάσεις(Conflicting recommendations)***: Πολλές φορές ανάλογα με τη μέθοδο επιλογής μοντέλου που ακολουθείται, προκύπτουν διαφορετικά αποτελέσματα και συμπεράσματα, όπως έχει ήδη γίνει κατανοητό. Αυτό το γεγονός δεν είναι οξύμωρο, απλώς δείχνει πόση έμφαση πρέπει να δοθεί στον τρόπο κατασκευής, τις ιδιότητες και τον σκοπό των μεθόδων επιλογής μοντέλων.
- ***Στάθμιση μοντέλων(Model averaging)***: Σε κάποιες περιπτώσεις, κατά την επιλογή στατιστικού μοντέλου παρατηρείται πως είναι περισσότερα του ενός αυτά τα οποία μπορούν να θεωρηθούν «κατάλληλα». Τότε, ο συνδυασμός τους μπορεί να προσφέρει καλύτερα και πιο ακριβή αποτελέσματα από το εάν επιλέγαμε μόνο ένα και μοναδικό ως «σωστό».

Είναι σαφές πως οι μέθοδοι επιλογής στατιστικών μοντέλων για να ανταποκριθούν στις ανάγκες, τους στόχους και τις εφαρμογές που εξυπηρετούν τα κάθε φορά επιλεγμένα μοντέλα, διαφέρουν μεταξύ τους. Όμως, οι πιο πολλές έχουν την ομοιότητα ότι ορίζονται με βάση το κατάλληλο Κριτήριο Πληροφορίας.

Ως Κριτήριο Πληροφορίας (*Information Criterion*) ορίζεται ένας μηχανισμός που χρησιμοποιεί τα εκάστοτε δεδομένα για να δώσει σε κάθε υποψήφιο μοντέλο ένα «σκορ». Έτσι, δημιουργείται μία λίστα από ταξινομημένα μοντέλα που ξεκινά από το καλύτερο και καταλήγει στο χειρότερο. Στην επιλογή κατάλληλων Κριτηρίων είναι σημαντικό να κατανοούμε όσο καλύτερα είναι δυνατό τα δεδομένα που έχουμε, τη φύση τους και το τι πληροφορίες μας «δίνουν», γιατί αυτά θα συμβάλλουν ώστε να αντιληφθούμε την

πολυπλοκότητα του μοντέλου και τους παράγοντες που είναι απαραίτητο να ενέχονται σε αυτό ή να το προσδιορίζουν. Σε ιδανικές συνθήκες, οπότε και θα μας ήταν γνωστά πολλά παραπάνω δεδομένα, θα λαμβάναμε υπ' όψιν περαιτέρω παράγοντες, όμως στην επιλογή μοντέλων μπορούμε να γνωρίζουμε μόνο τι συμπεράσματα προκύπτουν από τα δεδομένα μας και όχι σίγουρα για το ποια είναι η πλήρης πραγματικότητα.

Το ερώτημα είναι τι ακριβώς σημαίνει ο όρος «καταλληλότερο» ή «σωστό» μοντέλο; Η απάντηση δεν είναι απλή. Για να προχωρήσουμε στην επιλογή του κατάλληλου μοντέλου πρέπει σε κάθε περίπτωση να λάβουμε υπ' όψιν ποιος είναι ο σκοπός της στατιστικής μοντελοποίησης αλλά και πού πρόκειται να χρησιμοποιηθεί το μοντέλο. Κατά τον Akaike, ο σκοπός της μοντελοποίησης είναι να προβλέψει τα μελλοντικά δεδομένα με τη μεγαλύτερη δυνατή ακρίβεια (*predictive point of view*) και όχι να εντοπίσει την «πραγματική» κατανομή. Κι ενώ, εάν είχαμε στη διάθεσή μας άπειρα ή χωρίς «θόρυβο» δεδομένα, η ακριβής πρόβλεψη των μελλοντικών δεδομένων και η εύρεση της «πραγματικής» κατανομής τους θα αποτελούσαν δύο οπτικές με μικρές διαφορές, στην πραγματικότητα -οπότε και στις περισσότερες των περιπτώσεων- που τα δεδομένα υπό μελέτη είναι πεπερασμένου πλήθους, οι δύο αυτές οπτικές παρουσιάζουν σημαντικές διαφοροποιήσεις. Μάλιστα, όπως έχει διαπιστωθεί, συχνά απλούστερα -ακόμη και εκείνα στα οποία ενέχεται μεροληψία- μοντέλα δίνουν πολύ καλύτερες προβλέψεις από μοντέλα που είναι αποτέλεσμα της εκτίμησης της «πραγματικής» κατανομής. Ένας ακόμη παράγοντας που πρέπει να ληφθεί υπ' όψιν κατά την επιλογή μοντέλου είναι η οπτική του ερευνητή όσον αφορά την «πραγματική» κατανομή. Η παραδοσιακή οπτική υποθέτει ότι η «πραγματική» κατανομή είναι γνωστή ή τουλάχιστον υπάρχει. Μία πιο πρόσφατη δημοφιλής άποψη υποστηρίζει πως τα πραγματικά φαινόμενα είναι πολύ περίπλοκα και, άρα, δεν είναι εφικτό να περιγραφούν με ακρίβεια από μοντέλα. Δηλαδή, τα μοντέλα κατασκευάζονται βάσει της ήδη υπάρχουσας γνώσης του επιστήμονα όσον αφορά το αντικείμενο της μοντελοποίησης, όμως δεν υπάρχουν στην πραγματικότητα. Συνεπώς, σύμφωνα με την συγκεκριμένη οπτική, τα μοντέλα αποτελούν εργαλεία των οποίων η χρήση έχει ως αποτέλεσμα την εξαγωγή πληροφοριών από τα δεδομένα και, άρα, η επιλογή του «σωστού» έγκειται στο εάν αυτό είναι το κατάλληλο εργαλείο ώστε να εξαχθούν χρήσιμες πληροφορίες κι όχι το εάν προσεγγίζει με ακρίβεια την συμπεριφορά ενός φαινομένου.

Συνοψίζοντας, η επιλογή στατιστικού μοντέλου είναι κρίσιμο και αναπόσπαστο κομμάτι σχεδόν κάθε στατιστικής ανάλυσης και για να πραγματοποιηθεί είναι απαραίτητο να γίνει απολύτως σαφές τι θέλουμε να επιτύχουμε ή τι σκοπεύουμε να κάνουμε με τα παρεχόμενα δεδομένα, με λίγα λόγια ποιος είναι ο στόχος της ανάλυσης. Σύμφωνα με αυτόν και λαμβάνοντας υπ'όψιν τα προαναφερθέντα σημεία-κλειδιά, λαμβάνεται η τελική απόφαση με γνώμονα τα τελικά αποτελέσματα να προσφέρουν όσο το δυνατόν ακριβέστερες πληροφορίες και προβλέψεις.

1 **Akaike Κριτήριο Πληροφορίας (AIC)**

1.1 **Τύπος του AIC**

Το Akaike Κριτήριο Πληροφορίας (*Akaike Information Criterion*) προτάθηκε από τον Akaike (1974) και αποτελεί μία από τις πιο σημαντικές αλλά και δημοφιλείς μεθόδους επιλογής μοντέλων. Εφαρμογές του συναντώνται σε προβλήματα γραμμικών μοντέλων παλινδρόμησης και χρονοσειρών, στην Ανάλυση Επιβίωσης και γενικώς σε όλες τις περιπτώσεις σύγκρισης παραμετρικών μοντέλων. Είναι ενδεικτικό πως τα περισσότερα στατιστικά πακέτα για την επιλογή κατάλληλου μοντέλου χρησιμοποιούν το AIC.
Ο τύπος του είναι ο εξής:

$$AIC = -2\ \log L(\hat{\boldsymbol{\theta}}) + 2\dim(\boldsymbol{\theta}),$$

όπου $L(\hat{\boldsymbol{\theta}})$ είναι η μέγιστη πιθανοφάνεια, log ο νεπέριος λογάριθμος και $\dim(\boldsymbol{\theta})$ το μήκος του διανύσματος παραμέτρων $\boldsymbol{\theta}$.
Στην πράξη, αφού πρώτα υπολογιστεί η τιμή του AIC για κάθε υποψήφιο μοντέλο, διαλέγουμε αυτό με το μικρότερο «σκορ», αφού εκείνο είναι που βρίσκεται πιο κοντά στο πραγματικό μοντέλο. Αν κανένα από τα μοντέλα δεν είναι καλά προσαρμοσμένο, θα επιλεγεί το σχετικά καλύτερο από τα άλλα, όμως αυτό δε σημαίνει ότι είναι αντικειμενικά καλό. Γι' αυτό και οφείλουμε να κάνουμε ό,τι είναι δυνατό ώστε τα μοντέλα να στηρίζονται σε γερά «θεμέλια». Αυτό που μας αφορά κυρίως είναι η τιμή του AIC σε σχέση με τα υποψήφια μοντέλα και ακόμα περισσότερο οι διαφορές ανάμεσα στην τιμή του Κριτηρίου κάθε μοντέλου και όχι η απόλυτη τιμή του γενικά. Παρατηρώντας τη μορφή του Κριτηρίου αντιλαμβανόμαστε πως όσο προστίθενται παράμετροι στο μοντέλο, ο πρώτος όρος μειώνεται, ενώ ο δεύτερος, δηλαδή ο όρος ποινικοποίησης, αυξάνεται. Μπορεί να γίνει κατανοητό πως σε περίπτωση που συγκρίναμε τα μοντέλα μόνο μέσω του υπολογισμού της μέγιστης λογαριθμικής πιθανοφάνειας, δηλαδή χωρίς χρήση του δεύτερου όρου του AIC, το μοντέλο με τη μικρότερη τιμή θα ήταν αυτό με τις περισσότερες παραμέτρους, αφού η λογαριθμική πιθανοφάνεια αυξάνει με την προσθήκη περισσότερων παραμέτρων. Αντιθέτως, με την προσθήκη του δεύτερου όρου επιτυγχάνεται μία ισορροπία ανάμεσα στην καλή προσαρμογή ενός μοντέλου και την

απουσία υπερβολικά πολλών παραμέτρων, αφού κατά κάποιον τρόπο «τιμωρούνται» τα πιο σύνθετα μοντέλα-δηλαδή εκείνα με πολλές παραμέτρους. Έτσι επιτυγχάνεται ο *συμβιβασμός μεροληψίας- διασποράς*.

Παραδείγματα εφαρμογής του AIC αποτελούν τα παρακάτω:

Επιλογή κατάλληλης κατανομής

Έστω $\mathbf{Y} = (y_1, \ldots, y_n)$ παρατηρήσεις n πλήθους. Θέλουμε να ελέγξουμε εάν κάποιες υπολογιστικές διαδικασίες έχουν την ιδιότητα της έλλειψης μνήμης. Αν ναι, η κατάλληλη σ.π.π. για το ρυθμό αποτυχίας θα ήταν εκείνη της Εκθετικής Κατανομής, δηλαδή, η $f(y|\lambda) = \lambda\, e^{-\lambda y},\ y \geq 0$. Αν πάλι όχι, ο ρυθμός αποτυχίας μειώνεται όσο περνάει ο χρόνος και τότε θα χρησιμοποιούσαμε τη σ.π.π. της Weibull, δηλαδή την $f(y|\lambda, \alpha) = \alpha\lambda^{\alpha}\, y^{\alpha-1}\, e^{-(\lambda y)^{\alpha}},\ y \geq 0$.
Τα AIC για το πρώτο μοντέλο είναι:

$$AIC_{\varepsilon\kappa\theta} = -2 \sum_{i=1}^{n} \left[\log \hat{\lambda} - (\hat{\lambda}\, y_i) \right] + (2 \cdot 1),$$

όπου $\hat{\lambda}$ είναι η Ε.Μ.Π. της παραμέτρου λ της εκθετικής κατανομής, ενώ για το δεύτερο μοντέλο είναι:

$$AIC_{weib} = -2 \sum_{i=1}^{n} \left[\log \hat{\alpha} + \hat{\alpha}\ \log \hat{\lambda} + (\hat{\alpha} - 1)\ \log y_i - (\hat{\lambda}\, y_i)^{\hat{\alpha}} \right] + (2 \cdot 2),$$

όπου $\hat{\alpha}, \hat{\lambda}$ είναι οι Ε.Μ.Π. των παραμέτρων λ, α αντίστοιχα της Weibull κατανομής.

Μετά τον υπολογισμό των παραπάνω ποσοτήτων, επιλέγεται για την καταλληλότερη περιγραφή των δεδομένων η κατανομή με το μικρότερο AIC.

Επιλογή επεξηγηματικών μεταβλητών στη γραμμική παλινδρόμηση

Έστω πολλαπλό γραμμικό μοντέλο n παρατηρήσεων, όπου $\mathbf{Y} = (y_1, \ldots, y_n)$ είναι η μεταβλητή απόκρισης, $\mathbf{X} = (\mathrm{X}_1, \ldots, \mathrm{X}_p)$ οι επεξηγηματικές μεταβλητές, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ οι συντελεστές των επεξηγηματικών μεταβλητών και $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ τα τυχαία σφάλματα που έχουν διασπορά σ^2. Τότε, το μοντέλο με τη βοήθεια πινάκων γράφεται ως $\mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, όπου $\tilde{\mathrm{X}}$ είναι ο $n \times p + 1$ πίνακας σχεδιασμού. Είναι γνωστό πως

$\boldsymbol{\varepsilon} \overset{iid}{\sim} N_n(\mathbf{0}, \sigma^2\mathbf{I})$ και, άρα, $\mathbf{Y}|\mathbf{X} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.
Τότε η πιθανοφάνεια ως προς το διάνυσμα παραμέτρων $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ του μοντέλου ισούται με

$$L(\boldsymbol{\theta}) = \frac{1}{2\pi^{\frac{n}{2}}} \frac{1}{|\sigma^2\mathbf{I}|^{\frac{1}{2}}} e^{\left[-\frac{1}{2}\left(\mathbf{Y}-\tilde{\mathbf{X}}\boldsymbol{\beta}\right)^T(\sigma^2\mathbf{I})^{-1}\left(\mathbf{Y}-\tilde{\mathbf{X}}\boldsymbol{\beta}\right)\right]}$$

και, άρα, η λογαριθμική πιθανοφάνεια

$$\log L(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right)^T\left(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right).$$

Παραγωγίζουμε την l ως προς σ^2, $\boldsymbol{\beta}$ και έχουμε

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right)^T\left(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right) \ , \ \frac{\partial l}{\partial \boldsymbol{\beta}} = -\frac{1}{\sigma^2}\left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\boldsymbol{\beta} - \tilde{\mathbf{X}}^T\mathbf{Y}\right).$$

Θέτοντας τις παραγώγους ίσες με το μηδέν, τελικά, προκύπτουν οι Ε.Μ.Π. των σ^2, $\boldsymbol{\beta}$:

$$\hat{\sigma}^2 = \frac{\left(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}\right)^{\mathbf{T}}\left(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}\right)}{n} = \frac{SSE(\hat{\boldsymbol{\beta}})}{n} \ , \ \hat{\boldsymbol{\beta}} = \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\mathbf{Y}.$$

Έτσι,

$$l(\hat{\boldsymbol{\theta}}) = -\frac{n}{2} \ \log(2\pi) - n \ \log\hat{\sigma} - \frac{1}{2}n$$

και

$$AIC = n\log(2\pi) + 2n\log\hat{\sigma} + n + [2 \cdot (p+2)].$$

Τελικά, η επιλογή των κατάλληλων επεξηγηματικών μεταβλητών καθορίζεται από την ελαχιστοποίηση της ποσότητας $n\log\hat{\sigma}^2 + 2p$ σε όλα τα μοντέλα.

1.2 **Απόσταση Kullback-Leibler**

Αναμφισβήτητα, υπάρχουν κι άλλες στρατηγικές πέραν του AIC, όμως συγκεκριμένοι μαθηματικοί λόγοι κρύβονται πίσω από τη χρήση του AIC κι αυτοί σχετίζονται με τη συνάρτηση της απόστασης Kullback-Leibler. Τι είναι και πώς ορίζεται, λοιπόν, αυτή η συνάρτηση;

Έστω $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$ είναι τα n πλήθους δεδομένα μας. Τότε, για συνεχείς κατανομές, η απόσταση Kullback-Leibler (*Kullback-Leibler distance*) ορίζεται ως εξής:

$$I(g,f) = \int g(y)\ \log\left(\frac{g(y)}{f(y|\boldsymbol{\theta})}\right) dy,$$

όπου η g είναι η σ.π.π. του πραγματικού μοντέλου κατανομής των δεδομένων, ενώ η f η σ.π.π. του μοντέλου που χρησιμοποιείται για να προσεγγίσει την g. Με log συμβολίζεται ο νεπέριος λογάριθμος και με $\boldsymbol{\theta}$ το διάνυσμα παραμέτρων του προσεγγιστικού μοντέλου.

Είναι σημαντικό σ' αυτό το σημείο να αναφερθεί πως η g σε ρεαλιστικές συνθήκες αντικατοπτρίζει τη διαδικασία που παράγει τα δεδομένα κι έτσι σπανίως αποτελεί συγκεκριμένο μοντέλο. Είναι ενίοτε, λοιπόν, χρήσιμο να υποθέτουμε ότι η g έχει άπειρο πλήθος παραμέτρων κι έτσι οδηγεί σ' ένα σύνολο δεδομένων. Η υπόθεση αυτή είναι βεβαίως ανέφικτη, συμβάλλει ωστόσο στη διατήρηση της έννοιας της πραγματικότητας.
Η απόσταση Kullback-Leibler δηλώνει την απόσταση από την προσαρμοσμένη f στην πραγματική g, δηλαδή την πληροφορία που «χάνεται» όταν η f χρησιμοποιείται για να προσεγγίσει την g. Συνεπώς, όσο πιο μικρή η τιμή της απόστασης, τόσο πιο πολύ πλησιάζει η f την g. Βέβαια, αν θέλουμε να είμαστε ακριβείς η απόσταση αυτή δεν είναι απλώς απόσταση, καθότι το μέτρο της απόστασης της g από τη f δεν είναι ίσο με το μέτρο της απόστασης της f από την g. Είναι μία κατευθυνόμενη απόσταση, που δείχνει με πόση ακρίβεια προσεγγίζεται η g από τη f, δύο σ.π.π. που δε μπορεί η μία να μπει στη θέση της άλλης, δεν αντιμετατίθενται δηλαδή.

Ιδιότητες απόστασης Kullback-Leibler

- $I(g,f) \geq 0$

- $I(g,f) = 0 \Leftrightarrow g = f$ παντού.

Απόδειξη

Θεωρούμε τη συνάρτηση $K(t) = \log t - t + 1$, η οποία ορίζεται για $t > 0$.
Τότε η παράγωγός της, $K'(t) = \frac{1}{t} - 1$, για $t = 1$ ισούται με το μηδέν και η $K(t)$ για $t = 1$ παρουσιάζει μέγιστο που ισούται με μηδέν. Η ανισότητα $K(t) \leq 0$ ισχύει για κάθε $t \geq 0$ και η ισότητα μόνο για $t = 1$. Επομένως, $\log t \leq t - 1$ για κάθε $t \geq 0$.
Θέτουμε $t = \frac{f(y|\boldsymbol{\theta})}{g(y)}$ και η παραπάνω σχέση γίνεται:

$$\log\left(\frac{f(y|\boldsymbol{\theta})}{g(y)}\right) \leq \frac{f(y|\boldsymbol{\theta})}{g(y)} - 1.$$

Αν πολλαπλασιάσουμε και τα δύο μέλη με την $g(y)$ και ολοκληρώσουμε, τότε έχουμε:

$$\int \log\left(\frac{f(y|\boldsymbol{\theta})}{g(y)}\right) g(y)\, dy \leq \int \left(\frac{f(y|\boldsymbol{\theta})}{g(y)} - 1\right) g(y)\, dy.$$

Όμως,

$$\int \left(\frac{f(y|\boldsymbol{\theta})}{g(y)} - 1\right) g(y)\, dy = \int f(y|\boldsymbol{\theta})\, dy - \int g(y)\, dy = 0.$$

Άρα,

$$\int \log\left(\frac{g(y)}{f(y|\boldsymbol{\theta})}\right) g(y)\, dy \geq 0.$$

Είναι προφανές πως η ισότητα $\mathrm{I}(g,f) = 0$ ισχύει αν και μόνο αν $g = f$ παντού. ■

Άλλη μια σημαντική χρήση της απόστασης Kullback-Leibler είναι σε περιπτώσεις που έχουμε περισσότερα υποψήφια μοντέλα, έστω f_i $(i = 1, 2, \ldots)$, με παραμέτρους που ποικίλλουν και θέλουμε να εξακριβώσουμε ποιο από αυτά είναι πιο ακριβές, δηλαδή ποιες τιμές των παραμέτρων φέρνουν πιο κοντά κάθε f_i στο πραγματικό μοντέλο. Τότε, εάν παραπάνω από μία από τις προσεγγιστικές κατανομές πλησιάζουν αρκετά την πραγματική, ίσως γίνει δύσκολο να καταλήξουμε σε ένα ασφαλές συμπέρασμα μόνο ερευνώντας γραφήματα. Όταν χρησιμοποιούμε την απόσταση K-L πρέπει να αντιληφθούμε ότι αυτό που συγκρίνεται είναι οι κατανομές καθ' ολοκληρίαν και όχι μόνο με βάση το μέσο και τη διασπορά τους.

Κάποια άλλα μέτρα πέραν της απόστασης Kullback-Leibler για την εύρεση της απόστασης από τη g στην f είναι τα κάτωθι:

- **L_1-νόρμα:** $L_1(g,f) = \int |g(y) - f(y|\boldsymbol{\theta})|\, dy$
- **L_2-νόρμα:** $L_2(g,f) = \int |g(y) - f(y|\boldsymbol{\theta})|^2\, dy$
- **Απόσταση Hellinger:** $I_k(g,f) = \int \left(\sqrt{f(y|\boldsymbol{\theta})} - \sqrt{g(y)}\right)^2$

1.2.1 **Υπολογισμός της K-L**

Είναι προφανές πως στον υπολογισμό της Kullback-Leibler για να καταλήξουμε σ΄ ένα αποτέλεσμα πρέπει να είναι γνωστές και οι δύο σ.π.π.. Όμως, συνήθως η g είτε δεν είναι γνωστή είτε έχουμε ελλιπή στοιχεία γι' αυτή. Έτσι, η $I(g,f)$ γράφεται:

$$I(g,f) = \int g(y) \log g(y)\, dy - \int g(y) \log f(y|\hat{\boldsymbol{\theta}})\, dy \Rightarrow$$

$$I(g,f) = E_g[\log g(y)] - E_g[\log f(y|\hat{\boldsymbol{\theta}})].$$

αφού κάθε ολοκλήρωμα στη πρώτη ισότητα αποτελεί αναμενόμενη τιμή της g.

Η ποσότητα $H(g) = E_g[\log g(y)]$ ονομάζεται **αρνητική εντροπία κατά Shannon**, ενώ η $E_g[\log f(y|\hat{\boldsymbol{\theta}})]$ **αναμενόμενη λογαριθμική πιθανοφάνεια**. Η $H(g)$ αποτελεί μία (άγνωστη) σταθερά, έστω C, που εξαρτάται μόνο από την πραγματική σ.π.π. g. Έτσι, προκύπτει η εξής σχέση:

$$I(g,f) - C = -\int g(y) \log(f(y|\hat{\boldsymbol{\theta}})\, dy,$$

η οποία περιγράφει τη σχετική απόσταση της g με την f.

Επομένως, αν έχουμε δύο υποψήφια μοντέλα f_1 και f_2, το ποιο είναι πιο κοντά στο πραγματικό εξαρτάται από την αναμενόμενη λογαριθμική πιθανοφάνεια. Για παράδειγμα, αν το f_1 προσεγγίζει καλύτερα την g, η $I(g,f_1)$ είναι μικρότερη της $I(g,f_2)$ κι άρα $-\int g(y) \log(f_1(y|\hat{\boldsymbol{\theta}}))\, dy < -\int g(y) \log(f_2(y|\hat{\boldsymbol{\theta}}))\, dy$. Μάλιστα, παρόλο που δε μπορούμε

να βρούμε την ακριβή τιμή των $\mathrm{I}(g,f_1)$, $\mathrm{I}(g,f_2)$, είναι δυνατός ο υπολογισμός της διαφοράς τους $\mathrm{I}(g,f_2)-\mathrm{I}(g,f_1)$, δηλαδή μπορούμε να εξακριβώσουμε πόσο καλύτερη είναι η μία κατανομή έναντι της άλλης.

1.3 **Συλλογιστική πορεία του Akaike για το AIC**

Ο Akaike σκεπτόμενος πως για τον υπολογισμό της K-L χρειάζεται να είναι γνωστές και η g αλλά και οι παράμετροι $\boldsymbol{\theta}$ κάθε υποψήφιου μοντέλου, εισήγαγε ένα τρόπο εκτίμησης της αναμενόμενης σχετικής απόστασης Kullback-Leibler, δηλαδή της σχετικής απόστασης της g από την f, βασισμένο στη μέγιστη λογαριθμική πιθανοφάνεια διορθωμένη κατά μεροληψία. Η εκτίμηση αυτή είναι προσεγγιστική και, υπό κάποιες τεχνικές προϋποθέσεις, ασυμπτωτικά αμερόληπτη.
Δηλαδή, ουσιαστικά ο Akaike (1973) βρήκε ότι

$$C-\hat{E}_{\hat{\boldsymbol{\theta}}}[\mathrm{I}(g,f)]\approx \log L(\hat{\boldsymbol{\theta}})-\dim(\boldsymbol{\theta}),$$

όπου C είναι μία σταθερά και $\dim(\boldsymbol{\theta})$ το μήκος του διανύσματος παραμέτρων του προσεγγιστικού μοντέλου, που είναι και ο όρος διόρθωσης μεροληψίας.
Τέλος, ο Akaike κατέληξε στον ορισμό του τύπου του AIC πολλαπλασιάζοντας την ποσότητα $\log L(\hat{\boldsymbol{\theta}})-\dim(\boldsymbol{\theta})$ για ιστορικούς λόγους με τον αριθμό -2. Γενικά, ο αριθμός -2 συναντάται και σε άλλες συναφείς στατιστικές εκφράσεις κι επομένως έχει νόημα που και ο Akaike τον χρησιμοποίησε για να καταλήξει στην έκφραση του AIC.
Οφείλουμε να αναφέρουμε πως αυτή η σύνδεση της αναμενόμενης απόστασης Kullback-Leibler με τη μέγιστη λογαριθμική πιθανοφάνεια, που επιτεύχθηκε από τον Akaike, δρομολόγησε σημαντικές εξελίξεις στην ανάλυση πολύπλοκων δεδομένων και στον κλάδο της επιλογής μοντέλων (βλ. Stone (1982), Bozdogan (1987) και deLeeuw (1992)).

1.4 **Takeuchi Κριτήριο Πληροφορίας (TIC)**

Μία ειδική περίπτωση του Κριτηρίου Πληροφορίας Akaike είναι το Takeuchi Κριτήριο Πληροφορίας (*Takeuchi Information Criterion*). Αναπτύχθηκε από τον Takeuchi (1976). Η υπόθεση του Takeuchi ήταν ότι το πραγματικό μοντέλο βρίσκεται ανάμεσα στα υποψήφια. Το Κριτήριο αυτό χρησιμοποιείται σε περιπτώσεις που τα υποψήφια μοντέλα f δεν προσεγγίζουν καλά την g. Ο τύπος του είναι ο ακόλουθος:

$$TIC = -2\ \log L(\hat{\boldsymbol{\theta}}) + 2\ tr(I(\hat{\boldsymbol{\theta}})\ J(\hat{\boldsymbol{\theta}})^{-1}),$$

όπου ο όρος *tr* είναι το ίχνος, δηλαδή το άθροισμα των διαγωνίων στοιχείων του πίνακα $I(\hat{\boldsymbol{\theta}})\ J(\hat{\boldsymbol{\theta}})^{-1}$.

Οι παραπάνω πίνακες είναι διάστασης $\dim(\boldsymbol{\theta}) \times \dim(\boldsymbol{\theta})$ και, για n παρατηρήσεις $y_1, . . , y_n$, ισούνται με:

$$J(\hat{\boldsymbol{\theta}}) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(y_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\hat{\boldsymbol{\theta}}},$$

$$\mathrm{I}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\ \sum_{i=1}^{n} \frac{\partial \log f(y_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\ \frac{\partial \log f(y_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\bigg|_{\hat{\boldsymbol{\theta}}}.$$

Με βάση την περιγραφή του Κριτηρίου Πληροφορίας Takeuchi δημιουργείται η εύλογη ερώτηση: Γιατί να μη χρησιμοποιούμε το TIC αντί του AIC, απαλλασσόμενοι, έτσι, από την «έννοια» για το πόσο κοντά στην g είναι τα υποψήφια μοντέλα; Όμως, η ποιότητα των υποψήφιων μοντέλων είναι ένα κρίσιμο σημείο το οποίο πρέπει να ερευνάται πάντα στις στατιστικές αναλύσεις και στην επιλογή μοντέλων. Επίσης, η χρήση του TIC εμπεριέχει την εκτίμηση των στοιχείων των πινάκων $I(\hat{\boldsymbol{\theta}})$ και $J(\hat{\boldsymbol{\theta}})$, κάτι που οδηγεί σε αστάθεια στο αποτέλεσμα λόγω του σφάλματος από την προσέγγιση αυτών των στοιχείων.

Αποδεικνύεται πως μία απλή και φειδωλή εκτιμήτρια του όρου $tr(I(\hat{\boldsymbol{\theta}})\ J(\hat{\boldsymbol{\theta}})^{-1})$ είναι η ποσότητα $\dim(\boldsymbol{\theta})$, δηλαδή το μήκος του διανύσματος παραμέτρων $\boldsymbol{\theta}$. Το AIC είναι, έτσι, μία προσέγγιση του TIC με $tr(I(\hat{\boldsymbol{\theta}})\ J(\hat{\boldsymbol{\theta}})^{-1}) \approx \dim(\boldsymbol{\theta})$. Το θέμα είναι πως ο όρος $-2logL(\hat{\boldsymbol{\theta}})$ για μη καλές εκτιμήσεις της g κυριαρχεί γιατί η προσαρμογή είναι κακή, άρα η τιμή του θα είναι μεγαλύτερη σε σχέση με ακόμα και καλές εκτιμήσεις της g. Έτσι, αν τα υποψήφια μοντέλα είναι κοντά στην g, το TIC αποτελεί μία πολύ καλή προσέγγιση, όμως

στις περιπτώσεις που δεν ισχύει κάτι τέτοιο, μετατρέπεται σε κακή.
Το συγκεκριμένο Κριτήριο παρόλο που δε θεωρείται ασήμαντο, σπάνια εφαρμόζεται. Μάλιστα, κάποιοι επιστήμονες του κλάδου της Στατιστικής προτείνουν τη χρήση του μόνο σε περιπτώσεις που, πρώτον, το δείγμα είναι πραγματικά μεγάλο και, δεύτερον, περιμένουμε πολύ καλές προσεγγίσεις των πινάκων $I(\hat{\boldsymbol{\theta}})$, $J(\hat{\boldsymbol{\theta}})$.

1.5 **Διορθωμένο Κριτήριο Πληροφορίας Akaike (AIC_c)**

Όπως γνωρίζουμε, ενώ η πιθανοφάνεια αυξάνει όταν αυξάνεται το μέγεθος ενός δείγματος, ο όρος ποινικοποίησης εξαρτάται μόνο από το πλήθος των παραμέτρων. Κατά συνέπεια, σε περιπτώσεις που αυξάνεται πολύ το μέγεθος ενός δείγματος, ουσιαστικά ο όρος ποινικοποίησης του AIC γίνεται μικρός. Έτσι, η εφαρμογή του Κριτηρίου Πληροφορίας Akaike ενέχει τον κίνδυνο να επιλεγούν πολύπλοκα μοντέλα, παραβιάζοντας το σημείο-κλειδί της *φειδούς*. Με σκοπό την επίλυση αυτού του ζητήματος, δημιουργήθηκε από τους Hurvich & Tsai (1989) το διορθωμένο Κριτήριο Πληροφορίας Akaike (*corrected Akaike Information Criterion*), του οποίου ο τύπος είναι ο εξής:

$$AIC_c = -2\ \log L(\hat{\boldsymbol{\theta}}) + 2\dim(\boldsymbol{\theta})\left(\frac{n}{n-\dim(\boldsymbol{\theta})-1}\right) \quad \text{ή}$$

$$AIC_c = -2\ \log L(\hat{\boldsymbol{\theta}}) + 2\dim(\boldsymbol{\theta}) + \frac{2\ \dim(\boldsymbol{\theta})\ (\dim(\boldsymbol{\theta})+1)}{n-\dim(\boldsymbol{\theta})-1} \quad \text{ή}$$

$$AIC_c = AIC + \frac{2\ \dim(\boldsymbol{\theta})\ (\dim(\boldsymbol{\theta})+1)}{n-\dim(\boldsymbol{\theta})-1},$$

όπου n είναι το μέγεθος του δείγματος.

Ουσιαστικά, όπως φαίνεται στη πρώτη σχέση που αφορά το AIC_c, ο δεύτερος όρος που συναντάμε στο AIC πολλαπλασιάζεται με το κλάσμα $\frac{n}{n-\dim(\boldsymbol{\theta})-1}$. Έτσι, αν το μέγεθος του δείγματος είναι πολύ μεγάλο σε σχέση με το πλήθος των παραμέτρων του μοντέλου, τότε το κλάσμα αυτό είναι αμελητέο και εύκολα μπορούμε να χρησιμοποιήσουμε το AIC με καλά αποτελέσματα. Εάν βρεθούμε σε δίλημμα όσον αφορά τη χρήση του AICc, σημαντικό είναι να προσέξουμε πως $\dim(\boldsymbol{\theta})$ είναι το πλήθος των παραμέτρων του υποψήφιου μοντέλου με τις μεγαλύτερες διαστάσεις.

Σημειώνεται πως ένας πρακτικός κανόνας είναι χρήση του AIC_c να γίνεται σε περιπτώσεις που το πλήθος των παραμέτρων δεν είναι αρκετά μικρότερο από το πλήθος των παρατηρήσεων. Πάντως, είναι καλύτερο να μη συνδυάσουμε για το ίδιο σύνολο δεδομένων και τα δύο Κριτήρια, αλλά να περιοριστούμε στη χρήση είτε του ενός είτε του άλλου.

1.6 **Τροποποιημένα Κριτήρια Πληροφορίας Akaike QAIC και $QAIC_c$**

Υπάρχουν περιπτώσεις που η δειγματική διασπορά είναι μεγαλύτερη από τη θεωρητική διασπορά. Τότε λέμε ότι έχουμε **υπερ-διασπορά** *(overdispersion)* στα δεδομένα.
Οι Cox & Snell (1989) για την αντιμετώπιση αυτού του θέματος πρότειναν μία απλή και αποτελεσματική προσέγγιση για την εκτίμηση της υπερ-διασποράς με τη χρήση του χ^2-ελέγχου καλής προσαρμογής (*goodness-of-fit chi square statistic*):

$$\hat{c} = \frac{X^2}{df},$$

όπου df είναι οι βαθμοί ελευθερίας του εκάστοτε μοντέλου. Το $X^2 = \sum_{l=1}^{m} \frac{(O_l - E_l)^2}{E_l}$ είναι το στατιστικό του χ^2-ελέγχου, όπου m είναι το πλήθος των κλάσεων στις οποίες είναι ταξινομημένες οι παρατηρήσεις. Επίσης, O_l και E_l είναι αντίστοιχα ο αριθμός και ο αναμενόμενος αριθμός παρατηρήσεων στην l-κλάση. Ο E_l ισούται με το γινόμενο της πιθανότητας μιας τυχαίας παρατήρησης $y_i, i = 1, \ldots, n$ μίας τ.μ. Y να ανήκει στην κλάση l, κάτω από την προϋπόθεση ότι η συνάρτηση κατανομής της Y είναι η συγκεκριμένη που υποθέσαμε, με το πλήθος n των παρατηρήσεων.
Γενικά, πρέπει $1 \leq \hat{c} \leq 4$, διότι μεγαλύτερες τιμές συνήθως σημαίνουν πως η προσαρμογή του μοντέλου έχει κάποιο «ελάττωμα» και έτσι επηρεάζει και την τιμή του $\hat{c}$. Επομένως, σε μοντέλα με υπερ-διεσπαρμένα δεδομένα, η λογαριθμική πιθανοφάνεια γίνεται: $\frac{\log L(\hat{\boldsymbol{\theta}})}{\hat{c}}$. Σε περιπτώσεις που $\hat{c} < 1$, παίρνουμε $\hat{c} = 1$.
Κατ' επέκταση, το Κριτήριο Πληροφορίας Akaike (AIC) καθώς και το Διορθωμένο Κριτήριο Πληροφορίας Akaike (AICc) για υπερ-διεσπαρμένα δεδομένα μετασχηματίζονται ως εξής:

$$QAIC = -\frac{2\ \log L(\hat{\boldsymbol{\theta}})}{\hat{c}} + 2\dim(\boldsymbol{\theta})$$

$$QAIC_c = -\frac{2\ \log L(\hat{\boldsymbol{\theta}})}{\hat{c}} + 2\dim(\boldsymbol{\theta}) + \frac{2\dim(\boldsymbol{\theta})(\dim(\boldsymbol{\theta})+1)}{n-\dim(\boldsymbol{\theta})-1} \quad \text{ή}$$

$$QAIC_c = QAIC + \frac{2\dim(\boldsymbol{\theta})(\dim(\boldsymbol{\theta})+1)}{n-\dim(\boldsymbol{\theta})-1}$$

Είναι απαραίτητο να παρατηρήσουμε τα παρακάτω:
Πρώτον, ο συγκεκριμένος παράγοντας πρέπει να υπολογίζεται από το ολικό μοντέλο (*global model*) και όχι για κάθε υποψήφιο μοντέλο ξεχωριστά.
Δεύτερον, πλέον οι παράμετροι του μοντέλου είναι ίσες με το πλήθος των παραμέτρων αυξημένο κατά μία μονάδα λόγω του $\hat{c}$.
Η χρήση των συγκεκριμένων Κριτηρίων προτείνεται για διακριτά δεδομένα.

2 **Μπεϋζιανό Κριτήριο Πληροφορίας (BIC)**

2.1 **Βασικές Έννοιες Μπεϋζιανής Στατιστικής**

Το Μπεϋζιανό Κριτήριο Πληροφορίας (BIC) αποτελεί προϊόν Μπεϋζιανής Στατιστικής, συνεπώς θα ήταν χρήσιμο να αναφερθούν ορισμένες βασικές έννοιές της πριν την παρουσίαση του Κριτηρίου.

Κεντρική υπόθεση της Μπεϋζιανής ανάλυσης είναι ότι η αβεβαιότητα θα πρέπει να μοντελοποιείται με τη χρήση πιθανότητας και ότι τα συμπεράσματα στα οποία θα οδηγηθούμε θα πρέπει να είναι λογικά και να υπακούν στους νόμους των πιθανοτήτων. Η αρχική αβεβαιότητα για τις άγνωστες παραμέτρους διαμορφώνεται από τις εκ των προτέρων απόψεις του ερευνητή.
Βασικά χαρακτηριστικά αυτής της ανάλυσης είναι, πρώτον, ότι το άγνωστο διάνυσμα παραμέτρων **θ** ενός μοντέλου θεωρείται τυχαία μεταβλητή και, δεύτερον, ότι η συμπερασματολογία γίνεται με βάση το **θεώρημα Bayes**, που επιτρέπει τη χρήση δεσμευμένων πιθανοτήτων επί των παρατηρήσεων. Υπενθυμίζεται πως το θεώρημα Bayes για δύο ενδεχόμενα A, B ορίζεται ως εξής:

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

όπου $P(A)$, $P(B)$ είναι οι πιθανότητες των A, B, $P(A|B)$ η δεσμευμένη πιθανότητα του A δεδομένου του B κι αντίστοιχα $P(B|A)$ η δεσμευμένη πιθανότητα του B δεδομένου του A.
Η Μπεϋζιανή προσέγγιση αρχίζει με μία κατανομή, η οποία ονομάζεται **εκ των προτέρων κατανομή** (*prior distribution*), που εκφράζει τη γνώση μας για το άγνωστο **θ** πριν την συλλογή δεδομένων και στη συνέχεια με χρήση των δεδομένων και του θεωρήματος Bayes επανακαθορίζεται σε **εκ των υστέρων κατανομή** (*posterior distribution*).

Πιο συγκεκριμένα, στα πλαίσια του BIC, έστω ότι έχουμε:

- n πλήθους δεδομένα $y_1, \ldots, y_n$
- $\mathrm{M}_1, \mathrm{M}_2, \ldots, \mathrm{M}_r$ τα υποψήφια μοντέλα

- $f_i(y|\boldsymbol{\theta}_i)$ τη σ.π.π. κάθε υποψήφιου μοντέλου M_i με $\boldsymbol{\theta}_i$ το διάνυσμα παραμέτρων μήκους $\dim(\boldsymbol{\theta}_i)$
- $\pi_i(\boldsymbol{\theta}_i)$ την εκ των προτέρων (*prior*) κατανομή του $\boldsymbol{\theta}_i$ του μοντέλου M_i.

Τότε, η περιθώρια κατανομή των $y_1, \ldots, y_n$, η οποία συχνά αναφέρεται ως **περιθώρια πιθανοφάνεια** (*marginal likelihood*) για το *i*-οστό μοντέλο M_i είναι:

$$p_i(y_1, \ldots, y_n) = \int f(y_1, \ldots, y_n|\boldsymbol{\theta}_i)\, \pi_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \ .$$

Άρα, η εκ των υστέρων πιθανότητα για το *i*-οστό μοντέλο, σύμφωνα με το θεώρημα Bayes, είναι:

$$P(M_i|y_1, \ldots, y_n) = \frac{p_i(y_1, \ldots, y_n)\, P(M_i)}{\sum_{j=1}^{r}\ p_j(y_1, \ldots, y_n)\, P(M_j)},$$

όπου $P(M_i)$ είναι η εκ των προτέρων πιθανότητα του *i*-οστού μοντέλου με $i = 1, 2, \ldots, r$.

Η ποσότητα $P(M_i|y_1, \ldots, y_n)$ εκφράζει την ύστερη πιθανότητα του *i*-οστού μοντέλου, όταν οι παρατηρήσεις μας είναι $y_1, \ldots, y_n$. Εύλογα, λοιπόν, συμπεραίνουμε πως το υπόψηφιο μοντέλο με τη μεγαλύτερη εκ των υστέρων πιθανότητα είναι εκείνο που αναμένεται να επιλεγεί, δηλαδή επί της ουσίας εκείνο με το μεγαλύτερο αριθμητή καθώς όλα έχουν τον ίδιο παρονομαστή. Στην περίπτωση που και οι εκ των προτέρων πιθανότητες των μοντέλων είναι ίσες, επιλέγεται αυτό με τη μεγαλύτερη περιθώρια πιθανοφάνεια, ο υπολογισμός της οποίας επιτυγχάνεται συχνά με χρήση της προσέγγισης Laplace.

Μία, επίσης, σημαντική ποσότητα στη Μπεϋζιανή Στατιστική, η οποία χρησιμοποιείται για τον έλεγχο μίας υπόθεσης έναντι μίας εναλλακτικής, είναι ο **παράγοντας Bayes** (*Bayes factor*). Για λόγους απλότητας θεωρούμε δύο μοντέλα M_1 και M_2 με ύστερες πιθανότητες $P(M_1|y_1, \ldots, y_n)$ και $P(M_2|y_1, \ldots, y_n)$. Ο λόγος, τότε, των εκ των υστέρων συμπληρωματικών πιθανότητων (*posterior odds*) των δύο μοντέλων είναι

$$\frac{P(M_1|y_1, \ldots, y_n)}{P(M_2|y_1, \ldots, y_n)} = \frac{p_1(y_1, \ldots, y_n)}{p_2(y_1, \ldots, y_n)}\ \frac{P(M_1)}{P(M_2)}.$$

Έτσι, ο παράγοντας Bayes του μοντέλου M_1 έναντι του M_2 ορίζεται ως εξής:

$$B_{12} = \frac{p_1(y_1, \ldots, y_n)}{p_2(y_1, \ldots, y_n)} = \frac{\int f_1(y_1, \ldots, y_n|\boldsymbol{\theta}_1)\, \pi_1(\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1}{\int f_2(y_1, \ldots, y_n|\boldsymbol{\theta}_2)\, \pi_2(\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2} \ .$$

Ανάλογα με την τιμή του παράγοντα Bayes απορρίπτουμε ή δεχόμαστε την αρχική υπόθεση. Μεγάλες τιμές του παράγοντα Bayes (μεγαλύτερες της μονάδας), σημαίνουν ότι το M_1 στηρίζεται ισχυρότερα από το M_2 από τα εκάστοτε δεδομένα, κι έτσι αποτελούν ένδειξη υπέρ του μοντέλου της αρχικής υπόθεσης.

Παρατηρήσεις

- Μη-πληροφοριακή (*noninformative*) εκ των προτέρων κατανομή ονομάζεται μία κατανομή που εκφράζει ασαφείς ή γενικές πληροφορίες για μία μεταβλητή, αντίθετα με μία πληροφοριακή (*informative*), η οποία δίνει συγκεκριμένες και σαφείς πληροφορίες.
- Η παράμετρος της εκ των προτέρων κατανομής συχνά αναφέρεται ως υπερπαράμετρος (*hyperparameter*) $\boldsymbol{\lambda}$. Ο όρος αυτός χρησιμοποιείται για να μη συγχέεται με τις παραμέτρους του μοντέλου που αναλύεται κάθε φορά.

2.2 **Τύπος του BIC**

Το Μπεϋζιανό Κριτήριο Πληροφορίας (*Bayesian Information Criterion*) ή Κριτήριο Schwarz (*Schwartz Information Criterion*) αναπτύχθηκε από τον Schwarz (1978). Πρόκειται, επίσης, για ένα από τα πιο δημοφιλή και ευρέως χρησιμοποιούμενα Κριτήρια.

Ο τύπος του είναι ο εξής:

$$\mathrm{B}IC = -2\ \log L(\hat{\boldsymbol{\theta}}) + \log n\ \dim(\boldsymbol{\theta}),$$

όπου $L(\hat{\boldsymbol{\theta}})$ είναι η μέγιστη πιθανοφάνεια, log ο νεπέριος λογάριθμος, n το μέγεθος του δείγματος και $\dim(\boldsymbol{\theta})$ το μήκος του διανύσματος παραμέτρων $\boldsymbol{\theta}$.
Όπως και στην περίπτωση του AIC, επιλέγεται το μοντέλο με τη μικρότερη BIC-τιμή, αφού προσεγγίζει καλύτερα το πραγματικό.
Παρατηρώντας τη μορφή του BIC, είναι σαφές πως έχει πολλές ομοιότητες με εκείνη του AIC, καθώς κι εδώ γίνεται χρήση της μεγίστης πιθανοφάνειας αλλά και όρου ποινικοποίησης ($\log n\ \dim(\boldsymbol{\theta})$), ώστε με το συνδυασμό τους από τα υποψήφια μοντέλα να επιλεγούν εκείνα που είναι ακριβή αλλά και απλά. Η διαφορά είναι πως στην περίπτωση

του BIC ο όρος αυτός γίνεται πιο δριμύς, δηλαδή -τουλάχιστον για $n \geq 8$- η πολυπλοκότητα «τιμωρείται» αυστηρότερα. Επομένως, στην πράξη, τα μοντέλα που επιλέγονται με χρήση του BIC τείνουν να είναι πιο *φειδωλά*.

Παραδείγματα χρήσης του BIC είναι τα εξής:

Επιλογή κατάλληλης κατανομής

Όπως είδαμε και στο αντίστοιχο παράδειγμα για το AIC, αν θέλουμε να ελέγξουμε την ιδιότητα της έλλειψης μνήμης, για $\mathbf{Y} = (y_1, \ldots, y_n)$ παρατηρήσεις n πλήθους, κατάλληλες σ.π.π. αποτελούν είτε η $f(y|\lambda) = \lambda\, e^{-\lambda y},\ y \geq 0$ είτε η $f(y|\lambda, \alpha) = \alpha\lambda^{\alpha}\, y^{\alpha-1}\, e^{-(\lambda y)^{\alpha}},\ y \geq 0.$

Υπολογίζουμε τα αντίστοιχα BIC:

$$\mathrm{B}IC_{\varepsilon\kappa\theta} = -2 \sum_{i=1}^{n} \left[\log \hat{\lambda} - (\hat{\lambda}\, y_i) \right] + (1 \cdot \log n)$$

$$BIC_{weib} = -2 \sum_{i=1}^{n} \left[\log \hat{\alpha} + \hat{\alpha} \log \hat{\lambda} + (\hat{\alpha} - 1)\ \log y_i - (\hat{\lambda}\, y_i)^{\hat{\alpha}} \right] + (2 \cdot \log n),$$

Προφανώς, το καλύτερο μοντέλο είναι αυτό που έχει την ελάχιστη BIC-τιμή.

Επιλογή επεξηγηματικών μεταβλητών στη γραμμική παλινδρόμηση

Ομοίως με το αντίστοιχο παράδειγμα στο AIC, αν έχουμε γραμμικό μοντέλο $\mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ n παρατηρήσεων, τότε η μέγιστη λογαριθμική πιθανοφάνεια ως προς $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ είναι:

$$l(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - n \log \hat{\sigma} - \frac{1}{2}n\ ,$$

όπου

$$\hat{\sigma}^2 = \frac{\left(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}\right)^{\mathbf{T}} \left(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}\right)}{n} = \frac{SSE(\hat{\boldsymbol{\beta}})}{n}\ ,\ \ \hat{\boldsymbol{\beta}} = \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\mathbf{Y}$$

είναι οι Ε.Μ.Π. των σ^2, $\boldsymbol{\beta}$.

Άρα,

$$\mathrm{B}IC = n\ \log(2\pi) + 2n\ \log \hat{\sigma} + n + [(p+2) \cdot \log n].$$

Από την ελαχιστοποίηση της ποσότητας $n \log \hat{\sigma}^2 + 2p$ σε όλα τα μοντέλα, προκύπτει το βέλτιστο υποσύνολο επεξηγηματικών μεταβλητών που χρειαζόμαστε.

2.3 **Περιγραφή της κατασκευής του BIC**

Για να γίνει κατανοητή η διαδικασία κατασκευής του Μπεϋζιανού Κριτηρίου Πληροφορίας (BIC), είναι απαραίτητο να αναφερθούμε στην προσέγγιση Laplace για ολοκληρώματα.

Προσέγγιση Laplace για ολοκληρώματα

Έστω ένα απλό ολοκλήρωμα

$$\int e^{n\, q(\boldsymbol{\theta})}\, d\boldsymbol{\theta},$$

όπου n είναι το πλήθος των παρατηρήσεων και $q(\boldsymbol{\theta})$ μία πραγματική συνάρτηση ενός διανύσματος παραμέτρων $\boldsymbol{\theta}$.
Κατά την προσέγγιση Laplace εκμεταλλευόμαστε το γεγονός πως για μεγάλο n, η προς ολοκλήρωση συνάρτηση συγκεντρώνεται στην περιοχή της κορυφής, $\hat{\boldsymbol{\theta}}$, της $q(\boldsymbol{\theta})$ με αποτέλεσμα η τιμή του ολοκληρώματος να εξαρτάται μόνο από τη συμπεριφορά της συνάρτησης σ' αυτή την περιοχή.
Αφού $\frac{\partial q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0$, το ανάπτυγμα Taylor γύρω από το $\hat{\boldsymbol{\theta}}$ είναι:

$$q(\boldsymbol{\theta}) = q(\hat{\boldsymbol{\theta}}) - \frac{1}{2}\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\, J_q(\hat{\boldsymbol{\theta}})\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \ldots, \quad \text{όπου } J_q(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T}$$

Αντικαθιστούμε το ανάπτυγμα Taylor στο απλό ολοκλήρωμα που θεωρήσαμε κι έχουμε:

$$\int e^{n\,[\, q(\hat{\boldsymbol{\theta}}) - \frac{1}{2}\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\, J_q(\hat{\boldsymbol{\theta}})\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \ldots]}\, d\boldsymbol{\theta}$$

$$\approx\; e^{n\, q(\hat{\boldsymbol{\theta}})} \int e^{-\frac{n}{2}\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\, J_q(\hat{\boldsymbol{\theta}})\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}\, d\boldsymbol{\theta}.$$

Όμως, για το τυχαίο διάνυσμα $\boldsymbol{\theta}$ μήκους ισχύει ότι προσεγγιστικά $\boldsymbol{\theta} \sim N_{\dim(\boldsymbol{\theta})}(\hat{\boldsymbol{\theta}}, n^{-1}\, J_q(\hat{\boldsymbol{\theta}})^{-1})$. Έτσι:

$$\int e^{-\frac{n}{2}\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\, J_q(\hat{\boldsymbol{\theta}})\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}\, d\boldsymbol{\theta} = \frac{(2\pi)^{\frac{\dim(\boldsymbol{\theta})}{2}}}{n^{\frac{\dim(\boldsymbol{\theta})}{2}}\ |J_q(\hat{\boldsymbol{\theta}})|^{\frac{1}{2}}}.$$

Επομένως, τελικά, η προσέγγιση Laplace για ολοκληρώματα είναι η εξής:

$$\int e^{n\, q(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \approx \frac{(2\pi)^{\frac{\dim(\boldsymbol{\theta})}{2}}}{n^{\frac{\dim(\boldsymbol{\theta})}{2}}\ |J_q(\hat{\boldsymbol{\theta}})|^{\frac{1}{2}}}\ e^{n\, q(\hat{\boldsymbol{\theta}})}.$$

Μετά και την παραπάνω αναφορά, πλέον είναι εφικτή η περιγραφή της κατασκευής του Μπεϋζιανού Κριτηρίου Πληροφορίας:

Με χρήση της προσέγγισης Laplace και αφαιρώντας τη συμβολική εξάρτηση από κάθε μοντέλο M_i, αν $\boldsymbol{\theta}$ είναι το διάνυσμα παραμέτρων, η περιθώρια πιθανοφάνεια των δεδομένων $y_1, \ldots, y_n$ που ορίστηκε παραπάνω, μπορεί να γραφεί ως εξής:

$$p(y_1, \ldots, y_n) = \int f(y_1, \ldots, y_n|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta})\ d\boldsymbol{\theta} \Leftrightarrow p(y_1, \ldots, y_n) = \int e^{\log[f(y_1,\ldots,y_n|\boldsymbol{\theta})]}\ \pi(\boldsymbol{\theta})\ d\boldsymbol{\theta}.$$

Δεδομένου πως η $l(\boldsymbol{\theta}) = \log f(y_1, \ldots, y_n|\boldsymbol{\theta})$ είναι η συνάρτηση λογαριθμικής πιθανοφάνειας, τελικά:

$$p(y_1, \ldots, y_n) = \int e^{l(\boldsymbol{\theta})}\ \pi(\boldsymbol{\theta})\ d\boldsymbol{\theta}.$$

Όπως αναφέρθηκε παραπάνω, όταν το μέγεθος n είναι επαρκώς μεγάλο, η τιμή του ολοκληρώματος εξαρτάται μόνο από τη συμπεριφορά της προς ολοκλήρωση συνάρτησης γύρω από μία περιοχή της Ε.Μ.Π. της $l(\boldsymbol{\theta})$, δηλαδή του $\hat{\boldsymbol{\theta}}$, διότι εκεί συγκεντρώνεται η συνάρτηση αυτή. Επομένως, αφού για την εύρεση της Ε.Μ.Π. της $l(\boldsymbol{\theta})$, πρέπει να ισχύει $\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0$, τα αναπτύγματα Taylor της λογαριθμικής πιθανοφάνειας $l(\boldsymbol{\theta})$ και της εκ των προτέρων κατανομής $\pi(\boldsymbol{\theta})$ γύρω από την Ε.Μ.Π. $\hat{\boldsymbol{\theta}}$ αντίστοιχα είναι:

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

όπου $J(\hat{\boldsymbol{\theta}}) = -\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\ \partial \boldsymbol{\theta}^T}$.

$$\pi(\boldsymbol{\theta}) = \pi(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T\ \frac{\partial \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} + \ldots\ .$$

Αντικαθιστώντας τα αναπτύγματα Taylor και απλοποιώντας τα αποτελέσματα, η προσέγγιση της οριακής πιθανοφάνειας $p(y_1, . . , y_n)$ είναι

$$p(y_1, . . , y_n) = \int e^{\left[l(\hat{\boldsymbol{\theta}}) - \frac{n}{2}\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\, J(\hat{\boldsymbol{\theta}})\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \ldots\right]} \left[\pi(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \frac{\partial \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} + \ldots\right] d\boldsymbol{\theta}.$$

Επομένως,

$$p(y_1, . . , y_n) \approx e^{l(\hat{\boldsymbol{\theta}})}\, \pi(\hat{\boldsymbol{\theta}}) \int e^{-\frac{n}{2}\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\, J(\hat{\boldsymbol{\theta}})\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}\, d\boldsymbol{\theta}.$$

Για την εξαγωγή του παραπάνω αποτελέσματος αξιοποιήθηκε ότι το διάνυσμα $\hat{\boldsymbol{\theta}}$ συγκλίνει στο $\boldsymbol{\theta}$ με τάξη $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \mathcal{O}_p(n^{-\frac{1}{2}})$ αλλά και ότι $\int (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\, e^{-\frac{n}{2}\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\, J(\hat{\boldsymbol{\theta}})\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}\, d\boldsymbol{\theta} = 0$, αφού η πρώτη παράγωγος της πιθανοφάνειας για $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ είναι ίση με μηδέν.

Ολοκληρώνοντας ως προς $\boldsymbol{\theta}$, έχουμε:

$$e^{-\frac{n}{2}\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\, J(\hat{\boldsymbol{\theta}})\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}\, d\boldsymbol{\theta} = \frac{(2\pi)^{\frac{\dim(\boldsymbol{\theta})}{2}}}{n^{\frac{\dim(\boldsymbol{\theta})}{2}}\ |J(\hat{\boldsymbol{\theta}})|^{\frac{1}{2}}},$$

αφού η συνάρτηση που ολοκληρώθηκε είναι η συνάρτηση πυκνότητας της $N_{\dim(\boldsymbol{\theta})}(\hat{\boldsymbol{\theta}}, n^{-1}\, J(\hat{\boldsymbol{\theta}})^{-1})$. Επομένως, για μεγάλο n, όσον αφορά την περιθώρια πιθανοφάνεια ισχύει:

$$p(y_1, . . , y_n) \approx e^{l(\hat{\boldsymbol{\theta}})}\, \pi(\hat{\boldsymbol{\theta}})\, \frac{(2\pi)^{\frac{\dim(\boldsymbol{\theta})}{2}}}{n^{\frac{\dim(\boldsymbol{\theta})}{2}}\ |J(\hat{\boldsymbol{\theta}})|^{\frac{1}{2}}}\ .$$

Λογαριθμώντας και τα δύο μέλη της σχέσης $p(y_1, . . , y_n) = \int f(y_1, . . , y_n|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ και πολλαπλασιάζοντάς τα με τον αριθμό -2, προκύπτουν τα εξής:

$$-2\ \log p(y_1, . . , y_n) = -2\ \log\left[\int f(y_1, . . , y_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}\right]$$

$$\approx -2l(\hat{\boldsymbol{\theta}}) + \log n\ \dim(\boldsymbol{\theta}) + \log|J(\hat{\boldsymbol{\theta}})| - \dim(\boldsymbol{\theta}) \log(2\pi) - 2\ \log[\pi(\boldsymbol{\theta})].$$

Αγνοώντας τους όρους τάξης μικρότερης από $\mathcal{O}(1)$ ως προς το δείγμα n, τελικά, προκύπτει ο τύπος του Μπεϋζιανού Κριτηρίου Πληροφορίας (BIC).

2.4 **Deviance Κριτήριο Πληροφορίας (DIC)**

Το Deviance Κριτήριο Πληροφορίας (*Deviance Information Criterion*) αναπτύχθηκε από τους Spiegelhalter et al. (2002) και χρησιμοποιείται συχνά στη Μπεϋζιανή ανάλυση πολλών παραμέτρων σε πολύπλοκα ιεραρχικά μοντέλα.
Έστω $y_1, \ldots, y_n$ παρατηρήσεις και $\boldsymbol{\theta}$ το διάνυσμα παραμέτρων ενός μοντέλου μήκους $\dim(\boldsymbol{\theta})$. Τότε, δεδομένων του μοντέλου και των παρατηρήσεων, ορίζεται η **απόκλιση**:

$$D(\boldsymbol{\theta}) = -2 \log L(\boldsymbol{\theta}).$$

Οι αποκλίσεις χρησιμοποιούνται κυρίως στα μοντέλα παλινδρόμησης που υπάρχει ένα καλά ορισμένο «κορεσμένο» μοντέλο, δηλαδή ένα μοντέλο το οποίο έχει τόσες παραμέτρους όσες και παρατηρήσεις. Συνήθως, το ενδιαφέρον εστιάζεται κυρίως στις διαφορές τιμών που έχουν μεταξύ τους οι αποκλίσεις και όχι τόσο στις ίδιες τις τιμές της κάθε απόκλισης ξεχωριστά.
Έστω τώρα $\pi(\boldsymbol{\theta}|y_1, \ldots, y_n)$ η εκ των υστέρων κατανομή του διανύσματος παραμέτρων $\boldsymbol{\theta}$. Ο Spiegelhalter και οι συνεργάτες του όρισαν τον αποτελεσματικό αριθμό παραμέτρων (*effective number of parameters*) σε ένα μοντέλο ως εξής:

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}),$$

όπου $\bar{\boldsymbol{\theta}} = \mathrm{E}_{\pi(\boldsymbol{\theta}|y_1,\ldots,y_n)}[\boldsymbol{\theta}]$ είναι ο εκ των υστέρων μέσος του $\boldsymbol{\theta}$ και $\overline{D(\boldsymbol{\theta})} = \mathrm{E}_{\pi(\boldsymbol{\theta}|y_1,\ldots,y_n)}[D(\boldsymbol{\theta})]$ ο εκ των υστέρων μέσος της απόκλισης $D(\boldsymbol{\theta})$.
Δηλαδή, ουσιαστικά το πλήθος των αποτελεσματικών παραμέτρων ισούται με τη διαφορά του εκ των υστέρων μέσου της απόκλισης με την απόκλιση του εκ των υστέρων μέσου του διανύσματος παραμέτρων του μοντέλου.
Άρα, τελικά, ο τύπος του DIC είναι:

$$DIC = D(\bar{\boldsymbol{\theta}}) + 2\, p_D.$$

Το Deviance Κριτήριο Πληροφορίας θεωρείται Μπεϋζιανό μέτρο προσαρμογής, όπου ο όρος p_D λειτουργεί ως όρος ποινικοποίησης.
Στην πράξη, η εύρεση της DIC-τιμής είναι απλή σε περιπτώσεις προσομοίωσης δειγμάτων του $\boldsymbol{\theta}$ από την εκ των υστέρων κατανομή. Τότε, υπολογίζεται το $\bar{\boldsymbol{\theta}}$ ως η μέση τιμή των παρατηρήσεων μεγάλου πλήθους προσομοιωμένων τιμών του διανύσματος παραμέτρων και, στη συνέχεια, το p_D. Άρα, βρίσκεται η τιμή του Κριτηρίου και, τελικά, επιλέγεται το μοντέλο με τη μικρότερη τιμή DIC.

3 **Bootstrap Κριτήριο Πληροφορίας (EIC)**

3.1 **Μέθοδος Bootstrap**

Η μέθοδος Bootstrap αναπτύχθηκε από τον Efron (1979) ως μία -βελτιωμένη και πιο αποτελεσματική από την Jackknife- απάντηση στο ζήτημα της μη-παραμετρικής εκτίμησης της διακύμανσης, του τυπικού σφάλματος και της μεροληψίας.
Η διαδικασία που ακολουθείται κατά την Bootstrap είναι η εξής:
Έστω δεδομένα $\mathbf{Y} = y_1, \ldots, y_n$ μεγέθους n που προέρχονται από μία άγνωστη Κατανομή Πιθανότητας $G(y)$. Τότε, το (άγνωστο) διάνυσμα παραμέτρων $\boldsymbol{\theta}$ ως προς την G εκτιμάται από την $\hat{\boldsymbol{\theta}}$ και, στη συνέχεια, αξιολογείται η αξιοπιστία της εκτίμησης. Για την αξιολόγηση του σφάλματος στη συγκεκριμένη εκτίμηση, απαραίτητες είναι οι εξής ποσότητες:

$$b(G) = E_G[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}$$

$$\sigma^2(G) = E_G[\,(\hat{\boldsymbol{\theta}} - E_G[\hat{\boldsymbol{\theta}}])^2\,]\,,$$

όπου $b(G)$ και $\sigma^2(G)$ είναι η μεροληψία και η διακύμανση της εκτιμήτριας αντίστοιχα που εκφράζουν το σφάλμα της εκτιμήτριας και, όπως είναι εμφανές από τους τύπους τους, εξαρτώνται από την (άγνωστη) G. Σκοπός είναι να εκτιμηθούν από τα δεδομένα.
Για αυτό το σκοπό, μπορεί να ακολουθηθεί η μέθοδος Bootstrap, η περιγραφή της οποίας είναι η ακόλουθη:
Αρχικά, εκτιμάται η G από μία εμπειρική κατανομή $\hat{G}(y)$, η οποία δίνει ίση πιθανότητα $\frac{1}{n}$ σε καθεμία από τις n παρατηρήσεις $y_1, \ldots, y_n$ και 0 αλλού.
Στη συνέχεια, γίνεται λήψη τυχαίων δειγμάτων $\mathbf{Y}^*$ από την $\hat{G}(y)$, τα οποία ονομάζονται **bootstrap-δείγματα**. Η αντίστοιχη εκτιμήτρια είναι η $\hat{\boldsymbol{\theta}}^*$. Τότε, αν $E_{\hat{G}}$ είναι η αναμενόμενη τιμή ως προς την $\hat{G}$, οι bootstrap-εκτιμήσεις της μεροληψίας και της διακύμανσης είναι αντίστοιχα οι:

$$b(\hat{G}) = E_{\hat{G}}[\hat{\boldsymbol{\theta}}^*] - \hat{\boldsymbol{\theta}}$$

$$\sigma^2(\hat{G}) = E_{\hat{G}}[\,(\hat{\boldsymbol{\theta}}^*) - E_{\hat{G}}[(\hat{\boldsymbol{\theta}}^*)^2]\,.$$

Τέλος, αφού κάθε bootstrap-δείγμα $\mathbf{Y}^*(i) = (y_1^*(i), \ldots, y_n^*(i))$, $i = 1, \ldots, B$ προκύπτει από δειγματοληψία με επανάθεση των παρατηρήσεων, οι bootstrap-εκτιμήσεις προσεγγίζονται αριθμητικά με χρήση της μεθόδου Monte Carlo. Δηλαδή, τα bootstrap-δείγματα μεγέθους n εξάγονται B φορές κατ' επανάληψη $(\mathbf{Y}^*(i) : i = 1, .., B)$ και οι αντίστοιχες B εκτιμήτριες είναι οι $\hat{\boldsymbol{\theta}}^*(i)$.
Οι προσεγγίσεις,τότε, των $b(\hat{G})$, $\sigma^2(\hat{G})$ είναι οι εξής:

$$b(\hat{G}) \approx \frac{1}{B}\sum_{i=1}^{B} \hat{\boldsymbol{\theta}}^*(i) - \hat{\boldsymbol{\theta}}$$

$$\sigma^2(\hat{G}) \approx \frac{1}{B-1}\sum_{i=1}^{B}\left[\hat{\boldsymbol{\theta}}^*(i) - \frac{\sum_{i=1}^{B}\hat{\boldsymbol{\theta}}^*(i)}{B}\right]^2.$$

Δηλαδή, με τη Bootstrap εκτιμώνται οι ποσότητες που μας ενδιαφέρουν όχι αναλυτικά για κάθε εκτιμήτρια, αλλά αριθμητικά με τη χρήση του αλγόριθμου που αναφέρθηκε παραπάνω.
Θεωρητικά, αν το πλήθος B των Bootstrap επαναλήψεων γίνει απείρως μεγάλο, τα σφάλματα που προκύπτουν από την εφαρμογή της μεθόδου Monte Carlo μπορούν να αγνοηθούν. Στην πράξη, συνήθως επιλέγονται 50-200 επαναλήψεις όταν πρόκειται για εκτίμηση διακύμανσης, τυπικού σφάλματος και μεροληψίας και 1000-2000 για εκτίμηση ποσοστημορίων.

3.2 **Bootstrap-Εκτίμηση μεροληψίας**

Έστω n παρατηρήσεις $\mathbf{Y} = (y_1, .., y_n)$ που προέρχονται από την πραγματική κατανομή $G(y)$ (με σ.π. $g(y)$) και $f(y|\hat{\boldsymbol{\theta}})$ που χρησιμοποιείται για να προσεγγίσει την $g(y)$. Θέλουμε, στα πλαίσια της μεθόδου Bootstrap, να υπολογίσουμε τη μεροληψία της λογαριθμικής πιθανοφάνειας ως εκτιμήτριας της αναμενόμενης λογαριθμικής πιθανοφάνειας.
Όπως προαναφέρθηκε, κατά την εφαρμογή της Bootstrap, αρχικά, η $G(y)$ εκτιμάται από

την $\hat{G}(y)$. Βάσει ενός bootstrap-δείγματος $\mathbf{Y}^*$ που προέρχεται απ' την τελευταία, προκύπτει ένα μοντέλο $f(y|\hat{\boldsymbol{\theta}}^*)$. Με την $\hat{G}$ να θεωρείται η πραγματική κατανομή, η αναμενόμενη λογαριθμική πιθανοφάνεια του $f(y|\hat{\boldsymbol{\theta}}^*)$ ισούται με

$$E_{\hat{G}}\left[\log f(y|\hat{\boldsymbol{\theta}}^*)\right] = \frac{1}{n}\sum_{j=1}^{n}\log f(y_j|\hat{\boldsymbol{\theta}}^*)$$

$$\equiv \frac{1}{n}\, l(y_1,\ldots,y_n|\hat{\boldsymbol{\theta}}^*).$$

Έτσι, αφού η $\hat{G}$ θεωρείται η πραγματική κατανομή, η αναμενόμενη λογαριθμική πιθανοφάνεια είναι απλώς η λογαριθμική πιθανοφάνεια.

Αφού η λογαριθμική πιθανοφάνεια (εκτιμήτρια της αναμενόμενης λογαριθμικής πιθανοφάνειας) κατασκευάζεται με την επανάχρηση του bootstrap-δείγματος $\mathbf{Y}^*$, και αν $\hat{G}^*$ είναι η εμπειρική κατανομή που βασίζεται στο $\mathbf{Y}^*$, δύναται να εκφραστεί ως εξής:

$$E_{\hat{G}^*}\ \log\left[f(y|\hat{\boldsymbol{\theta}}^*)\right] = \frac{1}{n}\sum_{j=1}^{n}\log f(y_j^*|\hat{\boldsymbol{\theta}}^*)$$

$$\equiv \frac{1}{n}\, l(y_1^*,\ldots,y_n^*|\hat{\boldsymbol{\theta}}^*).$$

Έτσι, η bootstrap-εκτίμηση της $b(G)$, τελικά, είναι:

$$b^*(\hat{G}) = E_{\hat{G}}\left[l(y_1^*,\ldots,y_n^*|\hat{\boldsymbol{\theta}}^*) - l(y_1,\ldots,y_n|\hat{\boldsymbol{\theta}}^*)\right].$$

3.3 **Τύπος του EIC**

Το Boostrap ή Extended Κριτήριο Πληροφορίας (*Extended Information Criterion*) προέκυψε από την εφαρμογή της μεθόδου Bootstrap και αναπτύχθηκε από τους Efron (1983), Wong (1983), Konishi & Kitawaga (1996), Ishiguro et al. (1997), Cavannaugh & Shumway (1997).

Αντλούμε B bootstrap-δείγματα μεγέθους n και γράφουμε το i-οστό δείγμα ως $\mathbf{Y}_n^*(i), i = 1, \ldots, B$. Τότε, η διαφορά $D^*(i)$ της λογαριθμικής πιθανοφάνειας με την αναμενόμενη λογαριθμική πιθανοφάνεια ως προς το δείγμα $\mathbf{Y}^*(i)$ ορίζεται ως:

$$D^*(i) = l(y_1^*(i), \ldots, y_n^*(i)|\hat{\boldsymbol{\theta}}^*(i)) - l(y_1, \ldots, y_n|\hat{\boldsymbol{\theta}}^*(i)),$$

όπου $\hat{\boldsymbol{\theta}}^*(i)$ είναι η εκτίμηση του $\boldsymbol{\theta}$ που προέκυψε από το i-οστό bootstrap-δείγμα. Τότε η $b^*(\hat{G})$ μπορεί να προσεγγιστεί αριθμητικά ως εξής:

$$b^*(\hat{G}) \approx \frac{1}{B} \sum_{i=1}^{B} D^*(i) \equiv b_B(\hat{G}),$$

όπου B είναι τα bootstrap-δείγματα και $b_B(\hat{G})$ η bootstrap εκτίμηση της μεροληψίας $b(G)$ της λογαριθμικής πιθανοφάνειας.
Έτσι,τελικά, ο τύπος του Bootstrap (Extended) Κριτηρίου Πληροφορίας είναι ο εξής:

$$EIC = -2 \log L(\hat{\boldsymbol{\theta}}) + 2\, b_B(\hat{G}),$$

όπου $L(\hat{\boldsymbol{\theta}})$ είναι η μέγιστη πιθανοφάνεια και log ο νεπέριος λογάριθμος.
Επιλέγεται το μοντέλο με τη μικρότερη EIC-τιμή.

Σε μια εποχή που, λόγω της προόδου της Πληροφορικής, έχει γίνει εφικτή η χρήση των αριθμητικών αντί των αναλυτικών μεθόδων στη μοντελοποίηση, γενικότερα η μέθοδος Bootstrap αλλά και το EIC αποτελούν σημαντικές στατιστικές τεχνικές, αφού εφαρμόζονται σε ευρύ φάσμα εξαιρετικά περίπλοκων προβλημάτων ποικίλων μορφών.

4 **Focussed Κριτήριο Πληροφορίας (FIC)**

Η κεντρική ιδέα σε όλα τα Κριτήρια που περιγράφηκαν παραπάνω είναι η επιλογή ενός και μοναδικού κατάλληλου μοντέλου, του «καλύτερου» δηλαδή, ώστε να εξηγηθούν οι μηχανισμοί στους οποίους βασίζονται τα δεδομένα αλλά και για να προβλεφθούν τα μελλοντικά. Όμως, στην πραγματικότητα η επιλογή κατάλληλου μοντέλου πολύ συχνά είναι μόνο το πρώτο βήμα μίας στατιστικής ανάλυσης και διαφορετικές αναλύσεις μπορεί να στοχεύουν ή να επικεντρώνονται σε άλλα σημεία, κάτι που δεν λαμβάνεται υπ'όψιν κατά τις προαναφερθείσεις μεθόδους, αφού το «καλύτερο» μοντέλο επιλέγεται ανεξαρτήτως αυτού. Με βάση αυτό τον προβληματισμό οι Claeskens & Hjort (2003, 2008) εισήγαγαν το FIC (*Focussed Information Criterion*), απαλλασσόμενοι από την άποψη ότι μόνο ένα μοντέλο είναι το καταλληλότερο. Το FIC βασίζεται, λοιπόν, στην ιδέα ότι το «καλύτερο» μοντέλο εξαρτάται από την παράμετρο στην οποία κάθε ανάλυση εστιάζεται το ενδιάφερον (*focus parameter*), π.χ. τη μέση τιμή, το ποσοστημόριο κ.α. Έτσι, με το συγκεκριμένο Κριτήριο για διαφορετικές παραμέτρους εστίασης προτιμώνται διαφορετικά μοντέλα. Συνήθως, εφαρμογές του FIC συναντώνται κατά την επιλογή επεξηγηματικών μεταβλητών σε μοντέλα παλινδρόμησης.

4.1 **Όροι και συμβολισμοί στο FIC**

Πριν προχωρήσουμε στην παρουσίαση του FIC είναι απαραίτητο να ορισθούν και να επεξηγηθούν έννοιες και ποσότητες που χρησιμοποιούνται στην FIC-ανάλυση.

Κατά τη διαδικασία επιλογής μοντέλων υπάρχει μία λίστα υποψήφιων μοντέλων, ανάμεσά τους το απλούστερο αλλά και το πιο πολύπλοκο εξ αυτών. Ορίζουμε, λοιπόν, ως **περιορισμένο** (*narrow*) το πιο απλό μοντέλο που θα εξετασθεί το οποίο έχει ένα άγνωστο διάνυσμα παραμέτρων μήκους $\dim(\boldsymbol{\theta}) = p$ και γνωστό $\boldsymbol{\gamma_0}$, το οποίο συχνά είναι ένα μηδενικό διάνυσμα. Σ' αυτό το μοντέλο $\boldsymbol{\theta}_0$ είναι η πραγματική τιμή του $\boldsymbol{\theta}$. Επίσης, ως **ευρύ** (*wide*) ορίζεται το πιο πολύπλοκο μοντέλο το οποίο πέραν του $\boldsymbol{\theta}$ (που συναντάται σε όλα τα μοντέλα) έχει ένα επιπλέον άγνωστο διάνυσμα παραμέτρων μήκους $\dim(\boldsymbol{\gamma}) = q$.

Θεωρούμε πως στο ευρύ μοντέλο υπάρχει $\boldsymbol{\gamma_0}$, για το οποίο προκύπτει το περιορισμένο, δηλαδή το περιορισμένο αποτελεί ειδική περίπτωση του ευρέος. Στα υπόλοιπα «ενδιάμεσα» μοντέλα, που κατατάσσονται σε S υποσύνολα του $\{1, .., q\}$, το επιπλέον άγνωστο διάνυσμα παραμέτρων είναι το $\boldsymbol{\gamma}_j$ που περιέχει κάποιες μόνο από τις q τιμές του $\boldsymbol{\gamma}$. Δηλαδή, σε καθένα από αυτά υπάρχουν οι τιμές του $\boldsymbol{\gamma}_j$, για τις οποίες $j \in S$, ενώ για $S = \emptyset$ προκύπτει το περιορισμένο μοντέλο, αφού ο συγκεκριμένος συμβολισμός αντιστοιχεί στη μη ύπαρξη επιπλέον αγνώστων παραμέτρων.

Έστω $\mathbf{Y} = (y_1, .., y_n)$ είναι n παρατηρήσεις που ακολουθούν την κατανομή F. Τότε το μοντέλο S, αν S^c είναι το συμπληρωματικό του σύνολο, έχει συνάρτηση πυκνότητας $f(y|\boldsymbol{\theta}, \boldsymbol{\gamma_S}, \boldsymbol{\gamma_{0,S^c}})$. Ο συμβολισμός $\boldsymbol{\gamma_S}$ αναφέρεται στις τιμές του $\boldsymbol{\gamma}$ που ενέχονται στο S και το $\boldsymbol{\gamma_{0,S^c}}$ στις υπόλοιπες τιμές του $\boldsymbol{\gamma}$ που είναι ίσες με τις αντίστοιχες του $\boldsymbol{\gamma_0}$, δηλαδή αυτές για τις οποίες $j \notin S$. Επίσης, οι εκτιμήτριες των παραμέτρων του συγκεκριμένου μοντέλου συμβολίζονται ως $(\hat{\boldsymbol{\theta}}_S, \hat{\boldsymbol{\gamma}}_S)$ και μέσω αυτών καταλήγουμε στην Ε.Μ.Π. της παραμέτρου εστίασης $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}, \boldsymbol{\gamma})$, δηλαδή την $\hat{\boldsymbol{\mu}}_S = \mu(\hat{\boldsymbol{\theta}}_S, \hat{\boldsymbol{\gamma}}_S, \boldsymbol{\gamma_{0,S^c}})$.

Ακόμη, είναι απαραίτητο να αναφερθούν οι ποσότητες $U(x)$, $V(x)$ και J_S των οποίων θα γίνει χρήση παρακάτω. Στο $(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)$, λοιπόν, έχουμε:

$$\begin{bmatrix} U(y) \\ V(y) \end{bmatrix} = \begin{bmatrix} \frac{\partial \log f(y|\boldsymbol{\theta_0}, \boldsymbol{\gamma_0})}{\partial \boldsymbol{\theta}} \\ \frac{\partial \log f(y|\boldsymbol{\theta_0}, \boldsymbol{\gamma_0})}{\partial \boldsymbol{\gamma}} \end{bmatrix},$$

όπου οι διαστάσεις των U, V είναι p, q αντίστοιχα. Τότε ο πίνακας διασποράς J_{wide} στο $(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)$ του πιο πολύπλοκου μοντέλου έχει διαστάσεις $(p+q) \times (p+q)$ και ισούται με:

$$J_{wide} = Var_0 \begin{bmatrix} U(y) \\ V(y) \end{bmatrix} = \int \log f(y|\boldsymbol{\theta_0}, \boldsymbol{\gamma_0}) \begin{bmatrix} U(y) \\ V(y) \end{bmatrix} \begin{bmatrix} U(y) \\ V(y) \end{bmatrix}^T dy \quad \text{ή}$$

$$J_{wide} = \begin{bmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{bmatrix} \qquad J_{wide}^{-1} = \begin{bmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{bmatrix},$$

όπου J_{00} είναι ο πίνακας πληροφορίας του περιορισμένου μοντέλου και J_{11} είναι η διασπορά του περιορισμένου μοντέλου, επίσης.

Πλέον, μπορούν να οριστούν και οι ποσότητες $\boldsymbol{\omega}$, τ_0^2:

$$\boldsymbol{\omega} = J_{10} {J_{00}}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} - \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\gamma}} \qquad \tau_0^2 = \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} \right)^{\mathrm{T}} {J_{00}}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}},$$

όπου $\boldsymbol{\omega}$ είναι ένα διάνυσμα μήκους $\dim(\boldsymbol{\omega}) = q$ που καθορίζεται από τα ιδιαίτερα χαρακτηριστικά της $\boldsymbol{\mu}$ και τ_0 η τυπική απόκλιση της εκτιμήτριας για το απλούστερο μοντέλο. Επισημαίνεται πως οι $\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}, \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\gamma}}$ υπολογίζονται για το απλούστερο μοντέλο.

Αν v_S είναι ένα υποσύνολο v_j στοιχείων με $j \in S$ και $\mathbf{v}$ ένα διάνυσμα, το v_S μπορεί να

εκφραστεί και ως $v_S = \pi_S \, \mathbf{v}$ με π_S τον πίνακα προβολής του που έχει διαστάσεις $|S| \times q$, όπου $|S|$ είναι η πληθικότητα του S. Έτσι, ορίζεται ο $(|S| + p) \times (|S| + p)$ υποπίνακας του J_{wide}, ονόματι J_S, στον οποίο κρατάμε τις πρώτες p και τις τελευταίες q γραμμές και στήλες που αντιστοιχούν στα $j \in S$. Άρα, προκύπτει το εξής αποτέλεσμα:

$$J_S = \begin{bmatrix} J_{00} & J_{01,S} \\ J_{10,S} & J_{11,S} \end{bmatrix} \qquad\qquad J_S^{-1} = \begin{bmatrix} J^{00} & J^{01,S} \\ J^{10,S} & J^{11,S} \end{bmatrix}.$$

Αν $Q = J^{11}$, τότε, ορίζουμε τον $Q_S = J^{11,S} = \left(\pi_S \, Q^{-1} \, \pi_S^T\right)^{-1}$ για ένα μοντέλο S. Έστω, επίσης, $Q_S^0 = \pi_S^T \, Q_S \, \pi_S$ να είναι ένας $q \times q$ πίνακας του οποία τα στοιχεία ισούνται με αυτά του Q_S εκτός από τις γραμμές και τις στήλες του με δείκτη S^C και όπου τα στοιχεία του Q_S^0 είναι μηδενικά. Επίσης, ορίζεται ο πίνακας $G_S = Q_S^0 \, Q^{-1} = \pi_S^T \, Q_S \, \pi_S \, Q^{-1}$ διάστασης $q \times q$.

Τέλος, αφού δε γνωρίζουμε το πραγματικό μοντέλο και τόσο το περιορισμένο όσο και το ευρύ αποτελούν «ακραίες» περιπτώσεις (το μεν έχει μικρή διασπορά αλλά μεγάλη μεροληψία, ενώ το δε το αντίθετο), για να μην κυριαρχήσει η μεροληψία, στα πλαίσια του τοπικού εσφαλμένου προσδιορισμού (*local misspecification*) μοντέλου που εργαζόμαστε, υποθέτουμε ότι το πραγματικό είναι κάπου ενδιάμεσα και η κατανομή του f εκφράζεται και ως $f(y|\boldsymbol{\theta_0}, \boldsymbol{\gamma_0} + \frac{\boldsymbol{\delta}}{\sqrt{n}})$, όπου το $\boldsymbol{\delta}$ είναι ένα $q \times 1$ διάνυσμα που εκφράζει την «απόσταση» των υπόλοιπων μοντέλων από το περιορισμένο ή αλλιώς τους βαθμούς αναχωρήσεων (*degrees of departures*) των υπόλοιπων μοντέλων σε κατευθύνσεις $1, \ldots, q$. Αυτή η «απόσταση» του $\boldsymbol{\gamma}_S$ του εκάστοτε μοντέλου από το $\boldsymbol{\gamma}_0$ επιλέγεται να είναι τάξης $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ ώστε να προκύψουν πιο αποτελεσματικές προσεγγίσεις για μεγάλα δείγματα, στις οποίες το τετράγωνο της μεροληψίας και η διακύμανση θα είναι τάξης $\mathcal{O}\left(\frac{1}{n}\right)$. Έτσι, ορίζεται η εκτιμήτρια του $\boldsymbol{\delta}_{\mathbf{wide}}$, δηλαδή του $\boldsymbol{\delta}$ στο μεγαλύτερο μοντέλο, η οποία είναι $\mathbf{D}_n = \hat{\boldsymbol{\delta}}_{wide} = \sqrt{n}\,(\hat{\boldsymbol{\gamma}}_{wide} - \boldsymbol{\gamma}_0) \xrightarrow{d} \mathbf{D} \sim N_q(\boldsymbol{\delta}, Q)$.

4.2 **Τύπος του FIC**

Έστω ότι η παράμετρος προς εκτίμηση είναι της μορφής $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}, \boldsymbol{\gamma})$. Επομένως, μία εκτιμήτριά της είναι η $\boldsymbol{\mu}_S = \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_S, \hat{\boldsymbol{\gamma}}_S, \boldsymbol{\gamma}_{\mathbf{0,S^c}})$, αφού δε γνωρίζουμε πόσα από τα q στοιχεία του $\boldsymbol{\gamma}$ θα συμπεριληφθούν, αντίθετα με αυτά του διανύσματος $\boldsymbol{\theta}$ που θα συμπεριληφθούν όλα.

Έστω, επίσης, $MSE(S) = \sqrt{n}(\hat{\boldsymbol{\mu}}_S - \boldsymbol{\mu}_{true})$. Το FIC στηρίζεται σε μία εκτιμήτρια του MSE(S), δηλαδή $\mathbf{FIC = \hat{bias}^2(S) + \hat{Var}(S)}$. Επομένως, η ιδέα είναι να εκτιμηθεί το MSE(S) κάθε μοντέλου και, τελικά, να επιλεγεί ως καταλληλότερο εκείνο για το οποίο ελαχιστοποείται. Κάνουμε τη σύμβαση πως το $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$ εκφράζει την Ε.Μ.Π. του μοντέλου με τις $p + q$ παραμέτρους. Τότε, ο τύπος για το FIC κάθε μοντέλου S δίνεται από την εξής σχέση:

$$FIC = \hat{\boldsymbol{\omega}}^{\mathrm{T}} \, (\mathrm{I}_q - \hat{G}_S) \, \mathbf{D}_n \mathbf{D}_n^T \, (\mathrm{I}_q - \hat{G}_S)^T \, \hat{\boldsymbol{\omega}} + 2 \, \hat{\boldsymbol{\omega}}^{\mathrm{T}} \, \hat{Q}_S^0 \, \hat{\boldsymbol{\omega}} \quad \text{ή}$$

$$FIC = n \, \hat{\boldsymbol{\omega}}^{\mathrm{T}} \, (\mathrm{I}_q - \hat{G}_S) \, (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \, (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^T \, (\mathrm{I}_q - \hat{G}_S)^T \, \hat{\boldsymbol{\omega}} + 2 \, \hat{\boldsymbol{\omega}}^{\mathrm{T}} \, \hat{Q}_S^0 \, \hat{\boldsymbol{\omega}} \quad \text{ή}$$

$$FIC = (\hat{\boldsymbol{\psi}}_{wide} - \hat{\boldsymbol{\psi}}_S)^2 + 2 \, \hat{\boldsymbol{\omega}}_S^{\mathrm{T}} \, \hat{Q}_S^0 \, \hat{\boldsymbol{\omega}}_S,$$

όπου $\hat{\boldsymbol{\psi}}_{wide} = \hat{\boldsymbol{\omega}}^{\mathrm{T}} \mathbf{D}_n$ και $\hat{\boldsymbol{\psi}}_S = \hat{\boldsymbol{\omega}}^{\mathrm{T}} \hat{G}_S \, \mathbf{D}_n$ είναι οι εκτιμήσεις για το ευρύ και το S μοντέλο αντιστοίχως της ποσότητας $\boldsymbol{\psi} = \boldsymbol{\omega}^{\mathrm{T}} \, \boldsymbol{\delta}$. Επισημαίνεται, ακόμη, πως $\hat{\boldsymbol{\omega}}^{\mathrm{T}} \, \hat{Q}_S^0 \, \hat{\boldsymbol{\omega}} = \hat{\boldsymbol{\omega}}^{\mathrm{T}} \, \hat{Q}_S \, \hat{\boldsymbol{\omega}}_S$. Όπως αντιλαμβανόμαστε από τον τύπο, όσο πιο μεγάλο είναι το S τόσο πιο μικρός ο πρώτος όρος του δεξιού μέλους και τόσο μεγαλύτερος ο δεύτερος. Στην πράξη, βέβαια, συνήθως δεν υπολογίζουμε την τιμή του Κριτηρίου για όλα τα 2^q μοντέλα, αλλά αυτά που ανάλογα με τα δεδομένα και τη στατιστική ανάλυση, έχουμε θεωρήσει ως πιο πιθανά να προσεγγίζουν την πραγματικότητα. Σαφώς, τελικά, προτιμώνται τα μοντέλα με τις μικρότερες τιμές για το FIC.

Αν παρατηρήσουμε τις προηγούμενες σχέσεις, αντιλαμβανόμαστε πως για ένα απλούστερο μοντέλο S, ο όρος $(\hat{\boldsymbol{\psi}}_{wide} - \hat{\boldsymbol{\psi}}_S)^2$ αυξάνεται, ενώ ο $2 \, \hat{\boldsymbol{\omega}}_S^{\mathrm{T}} \, \hat{Q}_S^0 \, \hat{\boldsymbol{\omega}}_S$ είναι μειωμένος. Δηλαδή, η «ποινή» λόγω μεροληψίας είναι μεγάλη, ενώ η «αμοιβή» λόγω διασποράς μικρότερη. Προφανώς, για ένα περίπλοκο μοντέλο, συμβαίνει το αντίθετο. Έτσι, οδηγούμαστε στο συμπέρασμα ότι στο συγκεκριμένο Κριτήριο επιτυγχάνεται *συμβιβασμός μεροληψίας-διασποράς*. Μάλιστα, αν προσέξουμε τη σχέση $FIC = n \, \hat{\boldsymbol{\omega}}^{\mathrm{T}} \, (\mathrm{I}_q - \hat{G}_S) \, (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \, (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^T \, (\mathrm{I}_q - \hat{G}_S)^T \, \hat{\boldsymbol{\omega}} + 2 \, \hat{\boldsymbol{\omega}}^{\mathrm{T}} \, \hat{Q}_S^0 \, \hat{\boldsymbol{\omega}}$, γίνεται αντιληπτό πως αν το μοντέλο S δεν ταυτίζεται με το ευρύ (για το οποίο $\hat{G}_S = I$), όσο μεγαλώνει το μέγεθος του δείγματος, τόσο θα κυριαρχεί ο όρος $n \, \hat{\boldsymbol{\omega}}^{\mathrm{T}} \, (\mathrm{I}_q - \hat{G}_S) \, (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \, (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^T \, (\mathrm{I}_q - \hat{G}_S)^T \, \hat{\boldsymbol{\omega}}$ με αποτέλεσμα αυξημένες τιμές για το FIC. Ουσιαστικά, δηλαδή, για μεγάλα n θα επιλέγονται πιο ευρέα μοντέλα.

Ο υπολογισμός του FIC γίνεται πολύ απλούστερος αν ο $\hat{Q}$ είναι διαγώνιος πίνακας, διότι τότε ο G_S γίνεται, επίσης, διαγώνιος με μονάδα στις θέσεις j για τις οποίες $j \in S$ και 0 στις υπόλοιπες κι έτσι:

$$FIC = \left(\sum_{j \notin S} \hat{\boldsymbol{\omega}}_j \, D_{n,j} \right)^2 + 2 \sum_{j \in S} \hat{\boldsymbol{\omega}}_j^2 \, \hat{\kappa}_j^2, \quad \text{με} \quad \hat{Q} = diag(\hat{\kappa}_1^2, \ldots, \hat{\kappa}_q^2).$$

Στην πιο πάνω σχέση ο πρώτος όρος του δεξιού μέλους εκφράζει το τετράγωνο της μεροληψίας των παραμέτρων που δεν είναι στο εκάστοτε μοντέλο και ο δεύτερος είναι η διπλασιασμένη διασπορά των εκτιμητριών των παραμέτρων που περιέχονται στο μοντέλο. Από αυτήν την μορφή, γίνεται ακόμη πιο ξεκάθαρο πως όσο περισσότερες παράμετροι συμπεριλαμβάνονται σ' ένα μοντέλο, τόσο μικρότερη είναι η μεροληψία, σε αντίθεση με τη διασπορά που είναι αυξημένη (*συμβιβασμός μεροληψίας-διασποράς*).
Επίσης, υπάρχουν περιπτώσεις που η εκτιμήτρια του τετραγώνου της μεροληψίας (που αποτελεί όρο του FIC) είναι αρνητική. Για να αντιμετωπιστούν τέτοιες περιπτώσεις, ορίζεται το μεροληπτικά τροποποιημένο FIC* :

$$FIC^* = \begin{cases} FIC, \text{ αν δεν υπάρχει } N_n(S) \\ \hat{\boldsymbol{\omega}}^{\mathrm{T}} \, (\mathrm{I}_q + \hat{G}_S) \, \hat{Q} \, \hat{\boldsymbol{\omega}}, \text{ αν υπάρχει } N_n(S), \end{cases}$$

όπου N_n είναι η παρουσία αμελητέας μεροληψίας, δηλαδή

$$\left[\hat{\boldsymbol{\omega}}^{\mathrm{T}} \, (\mathrm{I}_q - \hat{G}_S) \, \hat{\boldsymbol{\delta}}_{wide} \right]^2 = n \left[\hat{\boldsymbol{\omega}}^{\mathrm{T}} \, (\mathrm{I}_q - \hat{G}_S) \, \hat{\boldsymbol{\gamma}}_{wide} \right]^2 < \hat{\boldsymbol{\omega}}^{\mathrm{T}} \, (\hat{Q} - \hat{Q}_S^0) \, \hat{\boldsymbol{\omega}}.$$

Στην πράξη, τα βήματα που ακολουθούνται για την εύρεση του καταλληλότερου μοντέλου κατά το FIC, συνοπτικά είναι:

- ***Καθορισμός της παραμέτρου εστίασης*** $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}, \boldsymbol{\gamma})$: Η σημασία αυτού το βήματος είναι ιδιαίτερη γιατί ουσιαστικά πρέπει να γίνει αντιληπτός ο λόγος της επιλογής μοντέλου.
- ***Καθορισμός της λίστας των υποψήφιων μοντέλων***: Θα ήταν πιο αποτελεσματικό εξαρχής η λίστα να αποτελείται μόνο από τα μοντέλα που θεωρούμε ότι πιθανώς θα ήταν κατάλληλα.
- ***Εκτίμηση του*** J_{wide}: Ένας τρόπος εκτίμησης του $J(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)$ είναι με χρήση ενός εμπειρικού Εσσιανού πίνακα. Δηλαδή, μπορεί να χρησιμοποιηθεί ο $J_{n,wide}(\hat{\boldsymbol{\theta}}_{narrow}, \boldsymbol{\gamma}_0)$ του $J_{n,wide}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$ με $J_{n,wide}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\zeta} \partial \boldsymbol{\zeta}^T} \log f(y_i | \boldsymbol{\theta}, \boldsymbol{\gamma})$ και $\boldsymbol{\zeta} = (\boldsymbol{\theta}, \boldsymbol{\gamma})$, κάτι που πολλές φορές προκύπτει ως αποτέλεσμα της εφαρμογής π.χ. της Newton-Raphson ή άλλων αριθμητικών μεθόδων. Επίσης, η εκτίμηση του $J(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)$ μπορεί να γίνει μέσω υπολογισμού, για παράδειγμα, 10000 τιμών του

$\begin{bmatrix} U(y) \\ V(y) \end{bmatrix}$ για την εκτίμηση του περιορισμένου μοντέλου, καθότι συχνά είναι πιο εύκολος ο υπολογισμός απ' ότι στο ευρύ. Τέλος, σε κάποιες περιπτώσεις μπορούμε να βρούμε συγκεκριμένο τύπο για την εκτίμηση $J(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)$. Με την εύρεση αυτής της εκτίμησης, γίνεται εφικτός και ο υπολογισμός των ποσοτήτων $\hat{Q},\ \hat{Q}_S,\ \hat{G}_S$.

- ***Εκτίμηση του γ για το ευρύ μοντέλο και υπολογισμός της εκτιμήτριας*** $\mathbf{D}_n$: Σ' αυτό το βήμα είναι δυνατή η χρήση είτε του $(\hat{\boldsymbol{\theta}}_{narrow}, \boldsymbol{\gamma}_0)$ είτε του $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$, όμως είναι προτιμητέο το τελευταίο διότι είναι συνεπές ακόμα κι όταν δεν ισχύουν οι υποθέσεις για το περιορισμένο μοντέλο και, έτσι, δεν εξαρτάται η ανάλυση από την «απόσταση» $\boldsymbol{\gamma}_0$ και $\boldsymbol{\gamma}$ και το μοντέλο είναι πιο ευσταθές.
- ***Εκτίμηση του*** $\boldsymbol{\omega}$: Η εύρεση της εκτιμήτριας των $\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\gamma}}$, $\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}$ γίνεται είτε με χρήση της $\hat{\boldsymbol{\theta}}$, η τιμή της οποίας προκύπτει είτε από αριθμητικές μεθόδους είτε από αυστηρό τύπο, είτε μέσω υπολογισμού της ποσότητας $\frac{\left[\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}, \boldsymbol{\gamma}_0 + \boldsymbol{\eta}\, e_i) - \boldsymbol{\mu}(\hat{\boldsymbol{\theta}} + \boldsymbol{\gamma}_0)\right]}{\boldsymbol{\eta}}$ για το $\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\gamma}}$ και της $\frac{\left[\boldsymbol{\mu}(\hat{\boldsymbol{\theta}} + \boldsymbol{\eta}\, e_i, \boldsymbol{\gamma}_0) - \boldsymbol{\mu}(\hat{\boldsymbol{\theta}} + \boldsymbol{\gamma}_0)\right]}{\boldsymbol{\eta}}$ για το $\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}$, για μικρό $\boldsymbol{\eta}$ (*linear predictor*), όπου $\boldsymbol{\eta} = \tilde{\mathbf{X}}\boldsymbol{\beta}$ με $\mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ ένα γραμμικό μοντέλο, και e_i η i-οστή τιμή του μοναδιαίου διανύσματος έχει τιμή 1 στη θέση i και 0 στις υπόλοιπες.

4.3 **Περιγραφή της κατασκευής του FIC**

Έστω $\mathbf{Y} = (y_1, \ldots, y_n)$ n-πλήθους δεδομένα. Υποθέτουμε ότι P_n, το n-οστό μοντέλο, είναι το πραγματικό μοντέλο με συνάρτηση πυκνότητας $f_n(y) = f(y|\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0 + \frac{\boldsymbol{\delta}}{\sqrt{n}})$.

Η κεντρική ιδέα της κατασκευής του FIC είναι η εκτίμηση του μέσου τετραγωνικού σφάλματος MSE του $\hat{\boldsymbol{\mu}}_S$. Για την περαιτέρω περιγραφή αυτής της διαδικασίας, κρίνεται απαραίτητη η παρουσίαση ενός θεωρήματος το οποίο απέδειξαν οι Hort & Claeskens (2003a) και αφορά την κατανομή που ακολουθεί το $\hat{\boldsymbol{\mu}}_S$. Για να γίνει αυτό θα χρειαστούν οι ποσότητες $\boldsymbol{\omega}$, τ_0^2 καθώς και οι ανεξάρτητες τυχαίες μεταβλητές $\mathbf{D} \sim N_q(\boldsymbol{\delta}, Q)$ και $\boldsymbol{\Lambda}_0 \sim N(0, \tau_0^2)$:

Θεώρημα

Για την εκτιμήτρια του $\boldsymbol{\delta}$ *στο ευρύ μοντέλο ισχύει*:

$$\mathbf{D}_n = \hat{\boldsymbol{\delta}}_{wide} = \sqrt{n}\,(\hat{\boldsymbol{\gamma}}_{wide} - \boldsymbol{\gamma}_0) \xrightarrow{d} \mathbf{D} \sim N_q(\boldsymbol{\delta}, Q),$$

ενώ για την Ε.Μ.Π. $\hat{\boldsymbol{\mu}}_S$ *της* $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}, \boldsymbol{\gamma})$:

$$\sqrt{n}(\hat{\boldsymbol{\mu}}_S - \boldsymbol{\mu}_{true}) \xrightarrow{d} \boldsymbol{\Lambda}_S = \boldsymbol{\Lambda}_0 + \boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{\delta} - G_S\, D).$$

Απόδειξη

Είναι γνωστό πως οι $U(y)$, $V(y)$ είναι οι παράγωγοι της $\log f(y|\boldsymbol{\theta}, \boldsymbol{\gamma})$ στο $(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)$ και $\bar{U}_n = \frac{1}{n}\sum_{i=1}^{n} U(y_i)$, $\bar{V}_n = \frac{1}{n}\sum_{i=1}^{n} V(y_i)$ οι μέσοι αυτών. Τότε οι Ε.Μ.Π. είναι πρώτης ασυμπτωτικής τάξης γραμμικές συναρτήσεις των $\bar{U}_n$, $\bar{V}_n$. Ο συνδυασμός ορισμάτων του αναπτύγματος Taylor με το Κ.Ο.Θ. κατά Lindeberg οδηγούν στο εξής:

$$\begin{bmatrix} \sqrt{n}\,(\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_0) \\ V\sqrt{n}\,(\hat{\boldsymbol{\gamma}}_S - \boldsymbol{\gamma}_{0,S}) \end{bmatrix} \doteq_d \begin{bmatrix} \sqrt{n}\,\bar{U}_n \\ \sqrt{n}\,\bar{V}_{n,S} \end{bmatrix} \xrightarrow{d} \begin{bmatrix} C_S \\ D_S \end{bmatrix} = J_S^{-1} \begin{bmatrix} J_{01}\,\boldsymbol{\delta} + U' \\ \pi_S\, J_{11}\,\boldsymbol{\delta} + V'_S \end{bmatrix},$$

όπου το σύμβολο $\doteq_d$ εκφράζει το γεγονός ότι η διαφορά συγκλίνει κατά πιθανότητα στο μηδέν και $(U'_S, V'_S) \sim N_{p+q}(0, J)$.
Έτσι, με πρόσθετα ορίσματα του αναπτύγματος Taylor

$$\sqrt{n}(\hat{\boldsymbol{\mu}}_S - \boldsymbol{\mu}_{true}) \xrightarrow{d} \boldsymbol{\Lambda}_S = \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}} C_S + \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\gamma}_S}\right)^{\mathrm{T}} D_S - \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\gamma}}\right)^{\mathrm{T}} \boldsymbol{\delta}.$$

Η υπόλοιπη απόδειξη περιγράφει τη διαδικασία ώστε να εκφραστεί κι αυτό το όριο σε μορφή πίνακα.
Οι δύο μεταβλητές που ορίστηκαν πριν την παρουσίαση του θεωρήματος ισούνται με $\mathbf{D} = \boldsymbol{\delta} + Q(V' - J_{10}J_{00}^{-1}\, U')$ $\boldsymbol{\Lambda}_0 = \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}} J_{00}^{-1}\, U'$.
Τέλος, μέσω αλγεβρικών πράξεων καταλήγουμε πως η $\boldsymbol{\Lambda}_S$ είναι ίδια με την $\boldsymbol{\Lambda}_0 + \boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{\delta} - G_S\, D)$.■
Σημειώνεται πως εδώ παρουσιάστηκαν τα βασικά βήματα της απόδειξης. Πιο πολλές πληροφορίες παρέχονται από τους Hjort&Claeskens (2003a, ενότητα 3 και παράρτημα).

Παρατηρείται πως οι οριακές μεταβλητές $\boldsymbol{\Lambda}_S \sim \mathrm{N}(\boldsymbol{\omega}^{\mathrm{T}}(\mathrm{I}_q - G_S)\boldsymbol{\delta}, \tau_S^2)$ με $\tau_S^2 = \tau_0^2 + \boldsymbol{\omega}^{\mathrm{T}}\, G_S\, Q\, G_S^T\, \boldsymbol{\omega} = \tau_0^2 + \boldsymbol{\omega}^{\mathrm{T}}\, Q_S^0\, \boldsymbol{\omega} = \tau_0^2 + \boldsymbol{\omega}^{\mathrm{T}}\, Q_S\, \boldsymbol{\omega}_S$.
Συνάγεται πως το απλούστερο μοντέλο έχει τη μικρότερη διασπορά, ίση με τ_0^2, και τη μεγαλύτερη μεροληψία, που είναι $\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\delta}$. Αντιθέτως, το πιο περίπλοκο μοντέλο έχει διασπορά $\tau_0^2 + \boldsymbol{\omega}^{\mathrm{T}}\, Q\, \boldsymbol{\omega}$ και μηδενική μεροληψία.
Επομένως, υπολογίζεται το οριακό μέσο τετραγωνικό σφάλμα της $\sqrt{n}(\hat{\boldsymbol{\mu}}_S - \boldsymbol{\mu}_{true})$, δηλαδή:

$$MSE(S, \boldsymbol{\delta}) = Var(\boldsymbol{\Lambda}_S) + bias^2(\boldsymbol{\Lambda}_S) \quad \text{ή}$$

$$MSE(S, \boldsymbol{\delta}) = \tau_0^2 + \boldsymbol{\omega}^{\mathrm{T}}\, Q_S^0\, \boldsymbol{\omega} + \boldsymbol{\omega}^{\mathrm{T}}\, (\mathrm{I}_q - G_S)\, \boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}}\, (\mathrm{I}_q - G_S)^{\mathrm{T}}\, \boldsymbol{\omega}.$$

Για να εκτιμηθεί η παραπάνω ποσότητα πρέπει να εκτιμηθούν οι $\tau_0, \boldsymbol{\omega}, Q_S, G_S$ και $\boldsymbol{\delta}$. Για τις πρώτες, η εύρεση εκτιμητριών είναι εύκολη, γιατί -σε αντίθεση με την τελευταία της οποίας η καλύτερη δυνατή είναι η $\mathbf{D}_n$-, έχουν συνεπείς εκτιμήτριες. Αποδεικνύεται, όμως, ότι ο μέσος της ποσότητας $\mathbf{D}\mathbf{D}^{\mathrm{T}}$ είναι $\boldsymbol{\delta}\, \boldsymbol{\delta}^{\mathrm{T}} + Q$ κι, έτσι, αντί να εκτιμηθεί μεμονωμένα η $\boldsymbol{\delta}$, εκτιμάται η $\boldsymbol{\delta}\, \boldsymbol{\delta}^{\mathrm{T}}$ με εκτιμήτρια την $\mathbf{D}_n \mathbf{D}_n^{\mathrm{T}} - \hat{Q}$. Επομένως, τελικά,

$$\hat{MSE}(S) = \hat{\tau_0}^2 + \hat{\boldsymbol{\omega}}^{\mathrm{T}}\, \hat{Q}_S^0\, \hat{\boldsymbol{\omega}} + \hat{\boldsymbol{\omega}}^{\mathrm{T}}\, (\mathrm{I}_q - \hat{G_S})\, (\mathbf{D}_n \mathbf{D}_n^{\mathrm{T}} - \hat{Q})\, (\mathrm{I}_q - \hat{G}_S)^{\mathrm{T}}\, \hat{\boldsymbol{\omega}} \quad \text{ή}$$

$$\hat{MSE}(S) = \hat{\tau_0}^2 - \hat{\boldsymbol{\omega}}^{\mathrm{T}}\, \hat{Q}\, \hat{\boldsymbol{\omega}} + \hat{\boldsymbol{\omega}}^{\mathrm{T}}\, (\mathrm{I}_q - \hat{G_S})\, \mathbf{D}_n \mathbf{D}_n^{\mathrm{T}}\, (\mathrm{I}_q - \hat{G}_S)^{\mathrm{T}}\, \hat{\boldsymbol{\omega}} + 2\, \hat{\boldsymbol{\omega}}^{\mathrm{T}}\, \hat{Q}_S^0\, \hat{\boldsymbol{\omega}},$$

όπου χρησιμοποιήθηκε ότι $Q_S^0\, Q^{-1}\, Q_S^0 = Q_S^0$.
Αφαιρώντας τους δύο πρώτους όρους, αφού δεν εξαρτώνται από το μοντέλο, καταλήγουμε στον τύπο του FIC.

Παρατήρηση
Έστω δύο υποψήφια μοντέλα S, S' και $\hat{\boldsymbol{\mu}}_S, \hat{\boldsymbol{\mu}}_{S'}$ οι αντίστοιχες εκτιμήτριες. Τότε η συσχέτισή τους συγκλίνει σε εκείνη των οριακών τους μεταβλητών, η οποία είναι:

$$corr(\boldsymbol{\Lambda}_S, \boldsymbol{\Lambda}_{S'}) = \frac{\tau_0^2 + \boldsymbol{\omega}^{\mathrm{T}}\, G_S\, Q\, G_{S'}^T\, \boldsymbol{\omega}}{(\tau_0^2 + \boldsymbol{\omega}^{\mathrm{T}}\, G_S\, Q\, G_S^T\, \boldsymbol{\omega})^{\frac{1}{2}}\, (\tau_0^2 + \boldsymbol{\omega}^{\mathrm{T}}\, G_{S'}\, Q\, G_{S'}^T\, \boldsymbol{\omega})^{\frac{1}{2}}}\ .$$

5 Άλλα Κριτήρια

5.1 Mallows C_p

Το Mallows C_p προτάθηκε από τον Mallows (1973) και χρησιμοποιείται συχνά ως μέτρο αξιολόγησης της προσαρμογής μοντέλων παλινδρόμησης.
Έστω p ένα υποσύνολο του συνόλου K των συντελεστών παλινδρόμησης ενός μοντέλου. Τότε η μορφή του συγκεκριμένου Κριτηρίου είναι η εξής:

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - n + 2\,p,$$

όπου:

- n είναι το μέγεθος του δείγματος
- $SSE_p = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ είναι το άθροισμα των τετραγώνων των υπολοίπων του γραμμικού μοντέλου με τους p συντελεστές παλινδρόμησης, όπου y_i και $\hat{y}_i$ ($i = 1, . . , n$) είναι αντίστοιχα οι παρατηρηθείσες και οι προβλεπόμενες τιμές της μεταβλητής απόκρισης του μοντέλου
- $\hat{\sigma}^2$ η εκτίμηση της διασποράς του μεγαλύτερου μοντέλου, δηλαδή εκείνου με τους K συντελεστές παλινδρόμησης.

Γενικά, σε μοντέλα p επεξηγηματικών μεταβλητών με μεγάλη μεροληψία το C_p αναμένεται να ισούται με p συν ένα θετικό όρο λόγω μεροληψίας (*bias term*), δηλαδή $C_p > p$. Στην αντίθετη περίπτωση, η τιμή του C_p είναι σχεδόν ίση με p.
Τελικά, ως καταλληλότερα θεωρούνται τα μοντέλα με C_p-τιμές μικρότερες από ή περίπου ίσες με το p. Μάλιστα, συχνά για να υπάρχει πιο σαφής εικόνα για το ποια είναι τα καλύτερα μοντέλα, συνιστάται η κατασκευή ενός γραφήματος του C_p σε συνάρτηση με τους p συντελεστές παλινδρόμησης.

5.2 **Hannan-Quinn Κριτήριο Πληροφορίας (HQIC)**

Το Hannan-Quinn Criterion Κριτήριο Πληροφορίας (*Hannan-Quinn Information Criterion*) αναπτύχθηκε από τους Hannan & Quinn (1979) αρχικά με σκοπό τον καθορισμό της τάξης των χρονοσειρών αυτοπαλινδρόμησης.
Ο τύπος του είναι ο εξής:

$$HQIC = -2\ \log L(\hat{\boldsymbol{\theta}}) + \log(\log n)\dim(\boldsymbol{\theta}),$$

όπου $L(\hat{\boldsymbol{\theta}})$ είναι η μέγιστη πιθανοφάνεια, log ο νεπέριος λογάριθμος, n το μέγεθος του δείγματος και $\dim(\boldsymbol{\theta})$ το μήκος του διανύσματος παραμέτρων $\boldsymbol{\theta}$.
Επιλέγεται ως καταλληλότερο το μοντέλο με τη μικρότερη HQIC-τιμή.
Η μορφή του HQIC παρουσιάζει ομοιότητες με εκείνη των AIC και BIC, μόνο που στο παρόν Κριτήριο ο όρος ποινικοποίησης ($\log(\log n)\dim(\boldsymbol{\theta})$) είναι εμφανώς λιγότερο δριμύς.
Γενικά, αν και βρίσκουμε συχνές αναφορές για το Hannan-Quinn Κριτήριο στη βιβλιογραφία, στην πράξη δε χρησιμοποείται με την ίδια συχνότητα.

5.3 **Final Prediction Error (FPE)**

Το Final Prediction Error αναπτύχθηκε από τον Akaike (1969, 1970) αρχικά στα πλαίσια της ανάλυσης χρονοσειρών με στόχο την επιλογή της τάξης των μοντέλων αυτοπαλινδρόμησης.
Ο τύπος του για περιπτώσεις γραμμικής παλινδρόμησης είναι ο εξής:

$$FPE = \frac{n+p}{n-p}\ \hat{\sigma}^2,$$

όπου n είναι το μέγεθος του δείγματος και $\hat{\sigma}^2 = \frac{SSE}{p}$ η εκτίμηση της διασποράς του μοντέλου με p συντελεστές παλινδρόμησης.
Προτιμάται το υποψήφιο μοντέλο με το μικρότερο FPE.
Άξια αναφοράς είναι η σχέση FPE και AIC: Αποδεικνύεται πως η ελαχιστοποίηση του Akaike Κριτηρίου Πληροφορίας είναι περίπου ισοδύναμη της ελαχιστοποίησης του

Τελικού Σφάλματος Πρόβλεψης. Έτσι, όσον αφορά τα μοντέλα αυτοπαλινδρόμησης, για ελάχιστη τιμή του AIC προκύπτει ένα μοντέλο που κατά προσέγγιση ελαχιστοποιεί και το FPE.
Στην πράξη, προτείνεται η χρήση του αντί του AIC σε περιπτώσεις δειγμάτων πολύ μικρού μεγέθους.

5.4 **ICOMP**

Οι Bozdogan (1988, 1990) και Bozdogan & Haughton (1998) πρότειναν το ICOMP (*I:information, COMP:complexity*) ως ένα μέτρο πληροφορίας ως προς την πολυπλοκότητα μοντέλων.
Ο τύπος του είναι ο εξής:

$$ICOMP = -2 \ \log L(\hat{\boldsymbol{\theta}}) + 2\, C,$$

όπου $L(\hat{\boldsymbol{\theta}})$ είναι η μέγιστη πιθανοφάνεια, log ο νεπέριος λογάριθμος, C ένα μέτρο πολυπλοκότητας.
Ουσιαστικά, ως όρος ποινικοποίησης αντί του πλήθους των παραμέτρων του μοντέλου, χρησιμοποιείται το C και ,τελικά, επιλέγεται εκείνο με το μικρότερο ICOMP.

Για γραμμικά μοντέλα $\mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ με $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ και $\boldsymbol{\varepsilon} \overset{iid}{\sim} N(\mathbf{0}, \sigma^2\mathbf{I})$, ο όρος C ισούται με:

$$C = \frac{p+1}{2} \ \log\left[\frac{tr(\hat{\sigma}^2(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}) + \frac{2\,\hat{\sigma^4}}{n}}{p+1} \right] - \frac{1}{2} \ \log|\hat{\sigma}^2(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}| - \frac{1}{2} \ \log\frac{2\,\hat{\sigma^4}}{n},$$

όπου $\hat{\sigma}^2$ είναι η εκτίμηση της διασποράς του μοντέλου.

Γενικά, το συγκεκριμένο Κριτήριο τόσο βιβλιογραφικά όσο και στην πράξη δεν συναντάται συχνά.

6 **Στάθμιση Μοντέλων**

Τις τελευταίες δεκαετίες έχει αναπτυχθεί αλλά και βελτιωθεί πληθώρα Κριτηρίων Επιλογής, άλλα στοχευμένα προς συγκεκριμένες κατηγορίες μοντέλων και άλλα πιο γενικά. Μάλιστα, κάποια από αυτά έχουν ενταχθεί σε στατιστικά πακέτα και χρησιμοποιούνται και από μη ειδικούς. Στην πράξη, συνήθως ένα ή περισσότερα Κριτήρια εφαρμόζονται, εντοπίζεται το καλύτερο μοντέλο που είναι επαρκές για την περιγραφή των δεδομένων και έπειτα η ανάλυση προχωρά σα να ήταν προαποφασισμένο το ποιο είναι το καταλληλότερο, χωρίς να δίνεται έμφαση στις συνέπειες της επιλογής αυτής. Όμως, είναι σαφές πως μία τέτοια τακτική εμπεριέχει μια κάποια αβεβαιότητα. Κι αυτό γιατί, πρώτον, οι εκτιμήτριες που διαμορφώνονται μετά την επιλογή μοντέλου (*post-selection estimators*) είναι ουσιαστικά συνδυασμοί των εκτιμητριών που θα είχαν προκύψει από μία διαφορετική επιλογή κατάλληλου μοντέλου από το Κριτήριο που εφαρμόστηκε και, δεύτερον, ότι συχνά η περαιτέρω ανάλυση ωφελείται περισσότερο από την επιλογή περισσότερων του ενός μοντέλων παρά από την αποκλειστική χρήση ενός και μοναδικού. Επομένως, μία διαφορετική οπτική είναι η στάθμιση μοντέλων (*model averaging*), δηλαδή η εκτίμηση της παραμέτρου που μας ενδιαφέρει μέσω όχι ενός, αλλά μεγαλύτερου πλήθους μοντέλων και η δημιουργία σταθμισμένου μέσου όρου (*weighted average*) των εκτιμητριών που προέκυψαν.

6.1 **Σταθμισμένες Εκτιμήτριες**

Έστω A ένα σύνολο που περιέχει S υποψήφια μοντέλα. Για το μοντέλο S, έστω $\hat{\boldsymbol{\mu}}_S$, το διάνυσμα της εκτιμήτριας του διανύσματος παραμέτρων $\boldsymbol{\mu}$. Υποθέτουμε ότι το καλύτερο μοντέλο επιλέχθηκε από το Akaike Κριτήριο Πληροφορίας, το συμβολίζουμε ως S_{AIC} και την αντίστοιχη εκτιμήτρια που θα προκύψει από αυτό $\hat{\boldsymbol{\mu}}_{S_{AIC}}$. Τότε η εκτιμήτρια που έπεται της επιλογής (*post-selection estimation*) είναι ίση με

$$\hat{\boldsymbol{\mu}}_{\mathbf{S_{AIC}}} = \sum_{S \in A} I\{S = S_{AIC}\}\, \hat{\boldsymbol{\mu}}_S,$$

δηλαδή ένα σταθμισμένο άθροισμα όλων των υποψήφιων εκτιμητριών στο οποίο τα βάρη είναι ίσα με 1 αν το μοντέλο είναι το επιλεγμένο από το AIC και 0 διαφορετικά. Τα συγκεκριμένα βάρη ονομάζονται indicator AIC-βάρη (*AIC indicator weights*). Αντίστοιχα, υπάρχουν και τα indicator BIC-βάρη (*BIC indicator weights*), τα indicator FIC-βάρη (*FIC indicator weights*) κ.τ.λ.
Γενικότερα, υπάρχουν εκτιμήτριες της μορφής

$$\hat{\boldsymbol{\mu}}_S = \sum_{S \in A} c(S)\, \hat{\boldsymbol{\mu}}_S,$$

όπου τα βάρη $\{c(S) : S \in A\}$ μπορούν να είναι οποιεσδήποτε τιμές (αν και συνήθως είναι μη αρνητικές), οι οποίες αθροίζουν στο 1. Οι εκτιμήτριες αυτές ονομάζονται **σταθμισμένες εκτιμήτριες** (*model average estimators*). Επί της ουσίας, η διαφορά των δύο αυτών τύπων εκτιμητριών είναι ότι στην πρώτη περίπτωση «επαναπαυόμαστε» στο μοντέλο που επιλέχθηκε, ενώ στη δεύτερη αναζητείται μία συμβιβαστική λύση ανάμεσα σε ένα πλήθος μοντέλων τα οποία θεωρούνται κατάλληλα.

Γνωστές περιπτώσεις σταθμισμένων εκτιμητριών είναι τα AIC-βάρη (*AIC weights*) και τα BIC-βάρη (*BIC weights*), τα οποία ορίζονται ως εξής:

$$c_{AIC}(S) = \frac{e^{\left(-\frac{1}{2}\Delta_{AIC,S}\right)}}{\sum_{S' \in A}\ e^{\left(-\frac{1}{2}\Delta_{AIC,S'}\right)}}\ ,$$

$$c_{BIC}(S) = \frac{e^{\left(-\frac{1}{2}\Delta_{BIC,S}\right)}}{\sum_{S' \in A}\ e^{\left(-\frac{1}{2}\Delta_{BIC,S'}\right)}}\ .$$

Στους παραπάνω τύπους Δ_{AIC} είναι οι λεγόμενες Akaike διαφορές (*Akaike differences*). Οι τιμές τους προκύπτουν από τη σχέση $\Delta_{AIC,S} = AIC_S - min_{S'} AIC_{S'}$, όπου $min_{S'} AIC_{S'}$ είναι η ελάχιστη τιμή του AIC που συναντάται στη λίστα με τα υποψήφια μοντέλα. Δηλαδή, υπολογίζουμε αρχικά την AIC-τιμή για κάθε μοντέλο, βρίσκουμε τη μικρότερη και στη συνέχεια για καθένα από αυτά βρίσκουμε τη διαφορά Akaike. Το άθροισμα στα c_{AIC} εκτείνεται σ΄όλα τα υποψήφια μοντέλα.

Ο λόγος που οι διαφορές αυτές είναι σημαντικές είναι ότι, όπως έχει ήδη αναφερθεί (βλ. Ενότητα 1.1), πρέπει να δίνεται περισσότερη έμφαση στις σχετικές διαφορές των τιμών του AIC παρά στην απόλυτη τιμή του. Εμπειρικά, έχει προκύψει πως εάν $\Delta_{AIC,S} \leq 2$, το μοντέλο είναι πολύ πιθανό να είναι ένα από τα καταλληλότερα.
Τα βάρη μπορούν να ερμηνευθούν ως η πιθανότητα το εκάστοτε μοντέλο να είναι το βέλτιστο, δεδομένων των παρατηρήσεων και της λίστας των υποψηφίων. Επομένως, για να ελέγξουμε το πόσο καλύτερο είναι το ένα μοντέλο έναντι κάποιου άλλου, διαιρούμε τα βάρη τους. Καθώς τα βάρη εξαρτώνται από το σύνολο των μοντέλων, είναι άμεσο πως εάν κάποιο από αυτά για οποιουσδήποτε λόγους αφαιρεθεί σε κάποιο σημείο της ανάλυσης, πρέπει να επαναϋπολογιστούν και τα βάρη.
Προφανώς, τα αντίστοιχα ισχύουν για τις διαφορές Δ_{BIC} και τα BIC-βάρη.

Επίσης, εφαρμογές σταθμισμένων εκτιμητριών συναντώνται στο γραμμικό μοντέλο παλινδρόμησης.
Δηλαδή, έστω γραμμικό μοντέλο

$$\mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{Z}}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

όπου:

- $i = 1, \ldots, n$
- $\mathbf{Y} = (y_1, \ldots, y_n)$ είναι οι τιμές της μεταβλητής απόκρισης
- $\boldsymbol{\varepsilon} \overset{iid}{\sim} N(0, \sigma^2)$
- $\tilde{\mathbf{X}}$ ο πίνακας σχεδιασμού για τις $\mathbf{X} = (X_1, \ldots, X_p)$ επεξηγηματικές μεταβλητές
- $\tilde{\mathbf{Z}}$ ο πίνακας σχεδιασμού για τις επιπλέον μεταβλητές $\mathbf{Z} = (Z_1, \ldots, Z_q)$ οι επεξηγηματικές που υπάρχει περίπτωση να υπάρχουν

Για $\boldsymbol{\mu} = \mathrm{E}(\mathbf{Y}|\mathbf{X}, \mathbf{Z})$ οι εκτιμήτριες είναι ίσες με:

$$\hat{\boldsymbol{\mu}}(\mathbf{X}, \mathbf{Z}) = \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}^* + \tilde{Z}\hat{\boldsymbol{\gamma}}^*,$$

όπου $\hat{\boldsymbol{\beta}}^* = \sum_{S \in A} c(S|\mathbf{D}_n)\, \hat{\boldsymbol{\beta}}_S$ και $\hat{\boldsymbol{\gamma}}^* = \sum_{S \in A} c(S|\mathbf{D}_n)\, \hat{\boldsymbol{\gamma}}_S$ είναι οι εκτιμήσεις των συντελεστών του μοντέλου.
Στην εκτίμηση της $\boldsymbol{\mu}$ στη γραμμική παλινδρόμηση, η στάθμιση του μοντέλου είναι ισοδύναμη με το να υπολογιστεί ο μέσος των συντελεστών παλινδρόμησης που προκύπτουν από την προσαρμογή των διαφορετικών μοντέλων παλινδρόμησης, κάτι που παύει να ισχύει σε πιο περίπλοκες περιπτώσεις εκτίμησης.

6.2 **Μπεϋζιανή Στάθμιση Μοντέλων (BMA)**

Η Μπεϋζιανή Στάθμιση Μοντέλων (*Bayesian Model Averaging*) χρησιμοποιεί την εκ των υστέρων σ.π.π. της παραμέτρου εστίασης $\boldsymbol{\mu}$ δεδομένων των παρατηρήσεων $y_1, \ldots, y_n$, δηλαδή την $\pi(\boldsymbol{\mu}|y_1, \ldots, y_n)$, ως ένα σταθμισμένο μέσο όρο των δεσμευμένων $\pi(\boldsymbol{\mu}|\mathrm{M}_i, y_1, \ldots, y_n)$, όπου $M_1, \ldots, M_r$ είναι τα υποψήφια μοντέλα, οι οποίες αποτελούν τα βάρη. Η $\pi(\boldsymbol{\mu}|y_1, \ldots, y_n)$ είναι απαλλαγμένη από την εξάρτηση από οποιοδήποτε μοντέλο κι, επομένως, βασικό χαρακτηριστικό κι αυτής της προσέγγισης είναι ότι λαμβάνεται υπόψη η αβεβαιότητα στα μοντέλα, κάτι που, όπως προαναφέρθηκε, αγνοείται σε άλλες τεχνικές.

Συγκεκριμένα, κατά τη Μπεϋζιανή Στάθμιση, αρχικά, προσδιορίζονται οι εκ των προτέρων πιθανότητες, $P(M_i)$, για κάθε μοντέλο από τη λίστα των υποψηφίων $M_1, \ldots, M_r$ και, στη συνέχεια, οι εκ των προτέρων σ.π.π. των παραμέτρων $\boldsymbol{\theta_i}$ κάθε μοντέλου M_i, δηλαδή οι $\pi(\boldsymbol{\theta}_i)$. Ύστερα, υπολογίζεται η εκ των υστέρων πιθανότητα του i-οστού μοντέλου $P(M_i|y_1, \ldots, y_n)$ (για ορισμούς και λεπτομερή περιγραφή βλ. Ενότητα 2.1). Θεωρώντας το M_i ως το πραγματικό μοντέλο, προκύπτει η εκ των υστέρων σ.π.π. της $\boldsymbol{\mu}$ για κάθε μοντέλο, η οποία συμβολίζεται ως $\pi(\boldsymbol{\mu}|\mathrm{M}_i, y_1, \ldots, y_n)$. Έτσι, καταλήγουμε στην εκ των υστέρων πυκνότητα της $\boldsymbol{\mu}$. Δηλαδή,

$$\pi(\boldsymbol{\mu}|y_1, \ldots, y_n) = \sum_{i=1}^{r} P(M_i|y_1, \ldots, y_n)\, \pi(\boldsymbol{\mu}|\mathrm{M}_i, y_1, \ldots, y_n).$$

Ανάλογα, προκύπτουν ο εκ των υστέρων μέσος της $\boldsymbol{\mu}$ που είναι ο σταθμισμένος μέσος των εκ των υστέρων μέσων κάθε μοντέλου καθώς κι η αντίστοιχη διακύμανση. Δηλαδή,

$$\mathrm{E}[\boldsymbol{\mu}|y_1, \ldots, y_n] = \sum_{i=1}^{r} P(M_i|y_1, \ldots, y_n)\, \mathrm{E}[\boldsymbol{\mu}|\mathrm{M}_i, y_1, \ldots, y_n]$$

$$Var[\boldsymbol{\mu}|y_1, \ldots, y_n] = \sum_{i=1}^{r} \Big[Var[\boldsymbol{\mu}|\mathrm{M}_i, y_1, \ldots, y_n] + \mathrm{E}[\boldsymbol{\mu}|\mathrm{M}_i, y_1, \ldots, y_n]\Big] \times$$

$$\times P(M_i|y_1, \ldots, y_n) - \mathrm{E}[\boldsymbol{\mu}|y_1, \ldots, y_n]^2.$$

Στην πράξη, η σωστή επιλογή εκ των προτέρων κατανομών, δεν είναι απλή. Κατά τους Claeskens & Hjort, εμφανίζονται προβλήματα στις περιπτώσεις που χρησιμοποιούνται διαφορετικές εκ των προτέρων κατανομές για κάθε μοντέλο και, ως αποτέλεσμα, προκύπτουν αντιφατικές εκ των προτέρων κατανομές της $\boldsymbol{\mu}$. Ο κάθε ερευνητής, δηλαδή, μπορεί να εκφράσει την εκ των προτέρων γνώμη του για την παράμετρο εστίασης, με έστω *r*, πλήθος τρόπων, όσα δηλαδή και τα μοντέλα, γεγονός που δεν είναι σίγουρο πως είναι πάντα σωστό. Συνιστάται, όταν δεν υπάρχουν αρκετές πληροφορίες όσον αφορά τα υποψήφια μοντέλα, να γίνεται χρήση μη-πληροφοριακών εκ των προτέρων κατανομών. Έτσι, συνήθως γίνεται χρήση της πιο απλής και παλαιάς μη-πληροφοριακής κατανομής, δηλαδή εκείνης που δίνει ίσες πιθανότητες σε όλα τα ενδεχόμενα, παρότι σε κάποιες περιπτώσεις αμφισβητείται εάν είναι όντως μη-πληροφοριακή. Γενικά, περαιτέρω μελέτη επί του θέματος όπως και προτάσεις μη-πληροφοριακών κατανομών για διάφορες περιπτώσεις μοντέλων έχουν προταθεί από τους Kass & Wasserman (1996), Jeffreys (1961), Rissanen (1983) και Berger & Pericchi (1996). Όσον αφορά την εκ των υστέρων κατανομή συνήθως υπολογίζεται μέσω μεθόδων Markov Chain Monte Carlo (MCMC) ή αριθμητικής ολοκλήρωσης.

Ένα ακόμη ζήτημα που μπορούμε να θίξουμε αφορά την περίπτωση πολύ μεγάλου πλήθους υποψηφίων μοντέλων. Γενικότερα, κατά την επιλογή κατάλληλου μοντέλου, είναι πιο φρόνιμο, όπως έχει προαναφερθεί, να εξετάζουμε- ανάλογα με τα δεδομένα μας και τη στατιστική ανάλυση- όχι το σύνολο των υποψηφίων μοντέλων, αλλά μόνο εκείνα που θεωρούμε πιθανότερο να προσεγγίζουν την πραγματικότητα. Έτσι, και κατά τη Μπεϋζιανή προσέγγιση, υπάρχουν περιπτώσεις που τα μοντέλα προς εξέταση είναι πάρα πολλά. Τότε, μπορεί να εφαρμοστεί η αρχή **Occam's Razor**, σύμφωνα με την οποία, όταν υπάρχουν πολλές εξηγήσεις που είναι συμβατές με ένα σύνολο δεδομένων, προτιμάται η απλούστερη. Δηλαδή, στη συγκεκριμένη περίπτωση, όταν υπάρχουν πολλά μοντέλα συμβατά με τα δεδομένα, προτιμώνται τα απλούστερα. Αυτό μπορεί να επιτευχθεί με τον παράγοντα Bayes (Jefferys & Berger (1992), MacKay (2003), Myung & Pitt (1997)), μέσω του οποίου επιτυγχάνεται ένα «αυτόματο» Occam's Razor. Κι αυτό διότι, χρησιμοποιεί την περιθώρια πιθανοφάνεια $p_i(y_1, \ldots, y_n)$, η οποία είναι η πιθανότητα οι παράμετροι που ελήφθησαν τυχαία από την εκ των προτέρων κατανομή, να παράξουν τα δεδομένα $y_1, \ldots, y_n$. Από τον τύπο της $p_i(y_1, \ldots, y_n) = \int f(y_1, \ldots, y_n|\boldsymbol{\theta}_i)\ \pi_i(\boldsymbol{\theta_i})\ d\boldsymbol{\theta}_i$ για ένα μοντέλο M_i, αντιλαμβανόμαστε πως μετρά τη μέση προσαρμογή ενός μοντέλου στα δεδομένα, αφού λειτουργεί ως σταθμισμένος μέσος της πιθανοφάνειας για όλες τις δυνατές τιμές του $\boldsymbol{\theta}$, δηλαδή για όλο τον παραμετρικό χώρο, με τις $\pi_i(\boldsymbol{\theta_i})d\theta_{\mathbf{i}}$ να αποτελούν τα σταθμισμένα βάρη. Σε περίπλοκα μοντέλα, οι τιμές των παραμέτρων είναι σε θέση να παράξουν μεγαλύτερο εύρος πιθανών δεδομένων. Όμως, η πιθανότητα ως προς στα δεδομένα πρέπει να ολοκληρώνει στο 1, κι έτσι η «εξάπλωση» της πυκνότητας, που επιτρέπει τα πιο περίπλοκα σετ δεδομένων, έχει ως αποτέλεσμα, τελικά, τα

απλούστερα δεδομένα να έχουν πιο μικρή πυκνότητα. Επομένως, η περιθώρια πιθανοφάνεια είναι μεγαλύτερη για μοντέλα τα οποία είναι απλούστερα, δηλαδή με λιγότερες παραμέτρους. Άρα, τελικά, στην αναζήτηση για το καταλληλότερο μοντέλο μειώνεται το σύνολο των προς εξέταση υποψηφίων μοντέλων, από το οποίο πλέον έχουν αφαιρεθεί εκείνα που δεν προσαρμόζονται καλά στα δεδομένα.

7 **Αξιολόγηση Κριτηρίων Επιλογής**

Σκοπός του συγκεκριμένου Κεφαλαίου είναι να παρουσιαστούν και να επεξηγηθούν οι έννοιες της συνέπειας και της αποδοτικότητας στα πλαίσια της σύγκρισης και της αξιολόγησης των Κριτηρίων Πληροφορίας, αλλά και η εξέταση των AIC και BIC ως προς αυτές τις έννοιες.

7.1 **Συνέπεια**

Η συνέπεια ενός Κριτηρίου μπορεί να εξεταστεί μέσω της ασθενούς (*weak consistency*) και της ισχυρής (*strong consistency*), έννοιες, δηλαδή, που σχετίζονται με την ασθενή (κατά πιθανότητα) και την ισχυρή (σχεδόν βέβαια) σύγκλιση, αλλά και της συνέπειας με την έννοια της φειδούς.

Ασθενής και Ισχυρή Συνέπεια

Υποθέτουμε ότι το πραγματικό μοντέλο, από το οποίο προέρχονται τα δεδομένα, υπάρχει και εμπεριέχεται στη λίστα με τα υποψήφια. Τότε, ένα Κριτήριο καλείται **ασθενώς συνεπές**, με πιθανότητα να τείνει στη μονάδα, είναι ικανό να επιλέξει το πραγματικό μοντέλο, καθώς το $n \to \infty$. Υπό τις ίδιες προϋποθέσεις ένα Κριτήριο έχει την ιδιότητα της **ισχυρής συνέπειας** αν η επιλογή του πραγματικού μοντέλου γίνεται σχεδόν βεβαίως.
Στην περίπτωση που δεν ισχύει η προϋπόθεση της ύπαρξης πραγματικού μοντέλου, τότε **ασθενώς συνεπές** είναι το Κριτήριο που, με πιθανότητα τείνουσα στο 1, επιλέγει το μοντέλο που βρίσκεται πιο «κοντά» στο πραγματικό, σύμφωνα με την απόσταση Kullback-Leibler, καθώς το $n \to \infty$. Αντίστοιχα, ορίζεται και η **ισχυρή συνέπεια**.

Προτού προχωρήσουμε στην παρουσίαση θεωρημάτων που αφορούν τις παραπάνω έννοιες, είναι απαραίτητος ο ορισμός μίας βοηθητικής ποσότητας:
Έστω, λοιπόν, δεδομένα $\mathbf{Y} = (y_1, .., y_n)$ και $r = 1, .., R$ το πλήθος των υποψηφίων μοντέλων. Τότε, $\boldsymbol{\theta}_r$ είναι οι παράμετροι του r μοντέλου και $f_{r,i}$ η σ.π.π. του r μοντέλου υπολογισμένη στην i-οστή παρατήρηση. Η γενική μορφή ενός Κριτηρίου, τότε, ορίζεται ως:

$$IC(M_r) = -2 \sum_{i=1}^{n} \log f_{r,i}(y_i|\hat{\boldsymbol{\theta}}_r) + c_{n,r},$$

όπου $\hat{\boldsymbol{\theta}}_r$ είναι οι Ε.Μ.Π. του $\boldsymbol{\theta}_r$ και $c_{n,r} > 0$ είναι ο όρος ποινικοποίησης για το εκάστοτε μοντέλο M_r.

Πλέον είναι εφικτή η παρουσίαση των σχετικών θεωρημάτων:

Θεώρημα Ασθενούς Συνέπειας

Υποθέτουμε ότι ανάμεσα στα υποψήφια μοντέλα που επιλέχθηκαν υπάρχει ακριβώς ένα μοντέλο έστω M_{r_0}*, για το οποίο η απόσταση Kullback-Leibler ως προς το πραγματικό μοντέλο ελαχιστοποιείται, τέτοιο ώστε να ισχύει ότι*

$$\liminf_{n\to\infty} \min_{r\neq r_0} \frac{1}{n} \sum_{i=1}^{n} \left[KL(g, f_{r,i}) - KL(g, f_{r_0,i})\right] > 0,$$

όπου g είναι η σ.π.π. του πραγματικού μοντέλου.

Έστω ότι ο αυστηρά θετικός όρος ποινικοποίησης είναι τέτοιος ώστε $c_{n,r} = o_p(n)$*. Τότε, με πιθανότητα τείνουσα στη μονάδα, το Κριτήριο Πληροφορίας επιλέγει το* M_{r_0} *ως το καλύτερο μοντέλο, καθώς το* $n \to \infty$.

Με άλλα λόγια, θα πρέπει $\lim_{n\to\infty} \frac{c_{n,r}}{n} \to 0$ για να είναι ένα Κριτήριο ασθενώς συνεπές.

Θεώρημα Ισχυρής Συνέπειας

Υποθέτουμε ότι ανάμεσα στα υποψήφια μοντέλα που επιλέχθηκαν υπάρχει ακριβώς ένα μοντέλο, έστω M_{r_0}*, για το οποίο η απόσταση Kullback-Leibler ως προς το πραγματικό μοντέλο ελαχιστοποιείται, τέτοιο ώστε να ισχύει ότι*

$$\liminf_{n\to\infty} \min_{r\neq r_0} \frac{1}{n} \sum_{i=1}^{n} \left[KL(g, f_{r,i}) - KL(g, f_{r_0,i})\right] > 0,$$

όπου g είναι η σ.π.π. του πραγματικού μοντέλου. Έστω ότι ο αυστηρά θετικός όρος ποινικοποίησης είναι τέτοιος ώστε $c_{n,r} = o(n)$ σχεδόν βεβαίως. Τότε,

$$P\left\{ \min_{l\neq r_0} \left(IC(M_l) - IC(M_{r_0}) \right) > 0, \ \text{για σχεδόν όλα τα } n \right\} = 1.$$

Συνέπεια

Στις προαναφερθείσες περιπτώσεις, βασική προϋπόθεση ήταν ότι το πραγματικό μοντέλο είναι ένα. Όμως, υπάρχουν στατιστικές αναλύσεις στις οποίες, τα μοντέλα που η K-L απόσταση ως προς το πραγματικό μοντέλο ελαχιστοποιείται, είναι περισσότερα του ενός. Τότε, επιλέγεται το μοντέλο με τις λιγότερες παραμέτρους, δηλαδή το πιο φειδωλό (*parsimonious*). Συχνά, η ιδιότητα αυτή αναφέρεται ως συνέπεια.
Με βάση το παρακάτω θεώρημα μπορεί να ελεγχθεί η συνέπεια ενός Κριτηρίου Πληροφορίας.

Θεώρημα Συνέπειας
Έστω J το σύνολο των δεικτών των μοντέλων στα οποία ελαχιστοποιείται η K-L απόσταση ως προς το πραγματικό μοντέλο g και J_0 το υποσύνολο του J που περιέχει τα πιο φειδωλά μοντέλα. Τότε:

Συνθήκη 1

Υποθέτουμε πως για όλα τα $r_0 \in J,\ l_0 \in J$ με $r_0 \neq l_0$:

- $\limsup_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[KL(g, f_{r_0,i}) - KL(g, f_{l_0,i})\right] < \infty.$
- *για κάθε δείκτη $j_0 \in J_0$ και $l \in J \setminus J_0$, ο όρος ποινικοποίησης έστω ότι είναι τέτοιος ώστε $P\left\{ \frac{c_{n,l} - c_{n,j_0}}{\sqrt{n}} \to \infty \right\} = 1.$*

Συνθήκη 2

Υποθέτουμε πως για όλα τα $r_0 \in J,\ l_0 \in J$ με $r_0 \neq l_0$:

- *ο λόγος λογαριθμικής πιθανοφάνειας* $\sum_{i=1}^{n} \log \frac{f_{r_0,i}(y_i|\boldsymbol{\theta}^*_{r_0})}{f_{l_0,i}(y_i|\boldsymbol{\theta}^*_{l_0})} = \mathcal{O}_p(1)$ *με* $\boldsymbol{\theta}^*_{r_0}$, $\boldsymbol{\theta}^*_{l_0}$ *να είναι τα διανύσματα παραμέτρων στα μοντέλα* M_{r_0}, M_{l_0} *αντίστοιχα.*
- *για κάθε δείκτη* $j_0 \in J_0$ *και* $l \in J \setminus J_0$ $P\Big\{(c_{n,l} - c_{n,j_0}) \to \infty\Big\} = 1.$

Αν ισχύει μία εκ των δύο συνθηκών, τότε με πιθανότητα να τείνει στη μονάδα, το Κριτήριο Πληροφορίας θα επιλέξει το μικρότερο μοντέλο για το οποίο ισχύει:

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

δηλαδή το πιο φειδωλό μοντέλο στο οποίο ελαχιστοποιείται η K-L ως προς το πραγματικό μοντέλο.

Σημειώνεται πως για να ικανοποιείται η δεύτερη συνθήκη, πρέπει η κατανομή που ακολουθεί το στατιστικό του λόγου λογαριθμικής πιθανοφάνειας να είναι φραγμένη.

7.2 **Αποδοτικότητα**

Η αποδοτικότητα (*efficiency*) ενός Κριτηρίου αφορά την ικανότητά του να αποδίδει σχεδόν το ίδιο καλά με το θεωρητικά βέλτιστο μοντέλο όσον αφορά το μέσο τετραγωνικό σφάλμα ή το αναμενόμενο τετραγωνικό σφάλμα πρόβλεψης. Σημειώνεται πως εδώ δε χρειάζεται το πραγματικό μοντέλο να είναι ένα από τα υποψήφια.

Έστω δεδομένα $y_1, \ldots, y_n$ και γραμμικό μοντέλο με μεταβλητή απόκρισης $\mathbf{Y} = (y_1, \ldots, y_n)$ και επεξηγηματικές μεταβλητές $\mathbf{X} = (X_1, \ldots, X_p)$, το οποίο γραφεται και ως

$$y_i = \beta_0 + \beta_1 \mathbf{x}_{i1} + \ldots + \beta_p \mathbf{x}_{ip} + \varepsilon_i,$$

όπου $i = 1, \ldots, n$, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ και $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip})$.

Στόχος είναι η πρόβλεψη της τιμής μίας νέας μεταβλητής απόκρισης, $\hat{\mathbf{Y}} = (\hat{y}_1, \ldots, \hat{y}_n)$. Έτσι, αν υπόθεσουμε ότι M είναι το σύνολο των επεξηγηματικών μεταβλητών, αναζητείται το υποσύνολο $\mathbf{x}_j$, $j \in M$, για το οποίο ελαχιστοποιείται το αναμενόμενο

σφάλμα πρόβλεψης. αναμενόμενο σφάλμα πρόβλεψης μη-δεσμευμένο (δηλαδή ανεξάρτητο από τα δεδομένα $y_1, \ldots, y_n$)

$$PE_n(M) = \sum_{i=1}^{n} E\left[(\hat{y}_{i,M} - y_{i,true})^2\right],$$

όπου οι $y_{i,true}$ είναι ανεξάρτητες των $y_1, \ldots, y_n$, αλλά προέρχονται από την ίδια κατανομή. Το $PE_n(M)$, ισοδύναμα, γίνεται

$$PE_n(M) = E\left[(\hat{\boldsymbol{\beta}}_{\mathbf{M}} - \boldsymbol{\beta}_{true})^{\mathrm{T}}\, \tilde{\mathrm{X}}^{\mathrm{T}} \tilde{\mathrm{X}}\, (\hat{\boldsymbol{\beta}}_{M} - \boldsymbol{\beta}_{true})\right] + n\sigma^2,$$

όπου είναι:

- $\hat{y}_{i,M}$ οι προβλεπόμενες τιμές, για την πρόβλεψη των οποίων χρησιμοποιείται το μοντέλο στο οποίο συμμετέχουν μόνο οι $\mathbf{x}_j,\ j \in M$, επεξηγηματικές μεταβλητές
- $\hat{\boldsymbol{\beta}}_M$ το διάνυσμα των παραμέτρων του μοντέλου που χρησιμοποιήθηκε στην πρόβλεψη αλλά και με μηδενικά στις θέσεις των επεξηγηματικών μεταβλητών που δε συμμετέχουν στο μοντέλο (ώστε να συμφωνούν οι διαστάσεις του με εκείνες του $\boldsymbol{\beta}_{true}$)
- $\tilde{\mathrm{X}}$ ο πίνακας σχεδιασμού των επεξηγηματικών μεταβλητών $\mathbf{X}$
- $Var[\varepsilon_i] = \sigma^2$.

Ένα Κριτήριο Πληροφορίας, τότε, είναι **αποδοτικό** αν ο λόγος της αναμενόμενης συνάρτησης απώλειας (π.χ. το αναμενόμενο σφάλμα πρόβλεψης) του επιλεγμένου μοντέλου με την αντίστοιχη θεωρητικά ελάχιστη αναμενόμενη συνάρτηση απώλειας που προκύπτει από τα υποψήφια μοντέλα, συγκλίνει κατά πιθανότητα στη μονάδα.
Δηλαδή, για να έχει το Κριτήριο την ιδιότητα της αποδοτικότητας πρέπει

$$\frac{PE_n(\hat{M})}{PE_n(M^*)} \xrightarrow{p} 1, \quad \text{καθώς } n \to \infty,$$

όπου $\hat{\mathrm{M}}$ είναι το σύνολο των επεξηγηματικών μεταβλητών του μοντέλου που επιλέχθηκε από το Κριτήριο Πληροφορίας, M^* το σύνολο των επεξηγηματικών μεταβλητών του μοντέλου με ελάχιστο αναμενόμενο σφάλμα πρόβλεψης.
Δηλαδή, γενικά, ένα μοντέλο καλείται **αποδοτικό** έαν επιλέγει το μοντέλο, που ο λόγος της αναμενόμενης συνάρτησης απώλειας του επιλεγμένου μοντέλου προς την ελάχιστη θεωρητικά αναμενόμενη συνάρτηση απώλειας, που προκύπτει από τα υποψήφια μοντέλα, τείνει κατά πιθανότητα στο 1, καθώς $n \to \infty$.

Στα πλαίσια της γραμμικής παλινδρόμησης ακολουθείται συχνά η προσέγγιση των Hurvich & Tsai (1995b) κατά την οποία διερευνάται η ασυμπτωτική αποδοτικότητα ενός Κριτηρίου Πληροφορίας. Αυτό επιτυγχάνεται μέσω του επόμενου θεωρήματος:

Θεώρημα Ασυμπτωτικής Αποδοτικότητας

Αν

- *η μέγιστη τάξη του πίνακα σχεδιασμού* $\tilde{\mathbf{X}}_M$*, που περιλαμβάνει στις στήλες του τις επεξηγηματικές μεταβλητές* $\mathbf{x}_j$*,* $j \in M$*, είναι ίση το πλήθος των επεξηγηματικών μεταβλητών στο σύνολο* M*, το οποίο συμβολίζεται ως* $|M|$*, και* $\max_M |M| = o(n^{\alpha})$ *για σταθερά* $\alpha \in (0, 1]$
- *υπάρχει σταθερά* $b \in [0, 0.5)$ *τέτοια ώστε για κάθε* $c > 0$

$$\sum_M e^{-cn^{-2b}\, PE_n(M)} \to 0, \quad \mu\varepsilon \;\; n \to \infty$$

- *Η μορφή του Κριτηρίου Πληροφορίας, όπου στο μοντέλο* M *η διασπορά εκτιμάται από την ποσότητα* $\frac{1}{n}\,(\mathbf{Y} - \tilde{\mathbf{X}}_M\hat{\boldsymbol{\beta}}_M)^T\,(\mathbf{Y} - \tilde{\mathbf{X}}_M\hat{\boldsymbol{\beta}}_M) = \frac{SSE_{\mathrm{M}}}{n}$ *είναι η εξής:*

$$IC(M) = (n + 2\,|M|)\hat{\sigma}^2_M,$$

τότε

$$\frac{PE_n(\hat{M}_{IC})}{PE_n(M^*)} - 1 = o_p(n^{-c}),$$

όπου $\hat{M}_{IC}$ *και* M^* *είναι τα σύνολα των επεξηγηματικών μεταβλητών του μοντέλου για το οποίο το* $IC(M)$ *λαμβάνει ελάχιστη τιμή και του μοντέλου με το ελάχιστο αναμενόμενο σφάλμα πρόβλεψης αντίστοιχα.* Επίσης, $c = \min\left\{\frac{1-a}{2}, b\right\}$.

Σημειώνεται πως από τη δεύτερη υπόθεση συνάγεται πως για όλα τα M πρέπει $n^{-2b}\, PE_n(M) \to \infty$, καθώς $n \to \infty$. Η συγκεκριμένη υπόθεση ικανοποιείται αν το πλήθος των επεξηγηματικών μεταβλητών του πραγματικού μοντέλου είναι άπειρο ή αν το πλήθος των επεξηγηματικών μεταβλητών αυξάνεται, όσο αυξάνεται και το μέγεθος του δείγματος, π.χ. σε εμφωλευμένα μοντέλα αν $M_1 = \{1\} \subset \ldots \subset M_q = \{1, \ldots, q\} \subset \ldots$, όπου $|M_q| = q$.

7.3 **Αξιολόγηση των AIC και BIC**

Εφόσον, πλέον, έχουν επεξηγηθεί οι ιδιότητες της συνέπειας και της αποδοτικότητας, μπορούμε να προχωρήσουμε στην αξιολόγηση των δημοφιλέστερων Κριτηρίων AIC, BIC.

Ασθενής Συνέπεια

Είναι προφανές πως και το AIC και το BIC γράφονται με τη γενική μορφή $IC(M_r) = -2\sum_{i=1}^{n} \log f_{r,i}(y_i|\hat{\boldsymbol{\theta}}_r) + c_{n,r}$ που ορίστηκε παραπάνω με όρους ποινικοποίησης $c_{n,r}^{AIC} = 2\dim(\boldsymbol{\theta})$ και $c_{n,r}^{BIC} = \dim(\boldsymbol{\theta})\log n$ αντίστοιχα, με $\dim(\boldsymbol{\theta})$ να είναι το μήκος του διανύσματος παραμέτρων $\boldsymbol{\theta}$.
Επομένως, σύμφωνα με το Θεώρημα Ασθενούς Συνέπειας προκύπτουν τα εξής:

$$\lim_{n\to\infty} \frac{c_{n,r}^{AIC}}{n} = \lim_{n\to\infty} \frac{2\ \dim(\boldsymbol{\theta})}{n} \to 0$$

$$\lim_{n\to\infty} \frac{c_{n,r}^{BIC}}{n} = \lim_{n\to\infty} \frac{\dim(\boldsymbol{\theta})\log n}{n} \to 0.$$

Άρα, και τα δύο Κριτήρια είναι ασθενώς συνεπή.

Ισχυρή Συνέπεια

Είναι άμεσο πως οι υποθέσεις του Θεωρήματος Ισχυρής Συνέπειας τόσο για το μοντέλο στο οποίο ελαχιστοποιείται η K-L απόσταση, όσο και για τον όρο ποινικοποίησης, ισχύουν και για το AIC και για το BIC.
Επομένως, αμφότερα τα Κριτήρια έχουν την ιδιότητα της ισχυρής συνέπειας.

Συνέπεια

Για να εξεταστεί η συνέπεια με βάση το αντίστοιχο θεώρημα, πρέπει να ελεγχθεί, αρχικά, μία από τις δύο συνθήκες του θεωρήματος. Συγκεκριμένα, ελέγχεται η δεύτερη.
Στα εμφωλευμένα μοντέλα, ο λόγος των διπλάσιων των μεγίστων λογαριθμικών πιθανοφανειών ακολουθεί ασυμπτωτικά μία κατανομή χ^2 με βαθμούς ελευθερίας το πλήθος των παραμέτρων στις οποίες διαφέρουν τα μοντέλα, όταν το πιο φειδωλό εξ

αυτών είναι πραγματικό. Η χ^2 είναι φραγμένη κατά πιθανότητα, δηλαδή $\mathcal{O}_p(1)$. Συνεπώς, αρκεί να ελεγχθεί αν για κάθε δείκτη $j_0 \in J_0$ και $l \in J \setminus J_0$ $P\Big\{(c_{n,l} - c_{n,j_0}) \to \infty\Big\} = 1$, ώστε να συμπεράνουμε αν τα Κριτήρια είναι συνεπή ή όχι.

Επομένως, για το AIC έχουμε:

$$P\Big\{ \left(c_{n,l}^{AIC} - c_{n,j_0}\right) \to \infty \Big\} = P\Big\{ 2 \left(\dim(\boldsymbol{\theta}_l) - \dim(\boldsymbol{\theta}_{j_0} \right) \to \infty \Big\} \neq 1,$$

αφού η ποσότητα $2 \left(\dim(\boldsymbol{\theta}_l) - \dim(\boldsymbol{\theta}_{j_0})\right)$ ισούται με μία σταθερά.

Ανάλογα, για το BIC προκύπτει:

$$P\Big\{ \left(c_{n,l}^{BIC} - c_{n,j_0}^{AIC}\right) \to \infty \Big\} = P\Big\{ \Big[\left(\dim(\boldsymbol{\theta}_l) \log n \right) - \left(\dim(\boldsymbol{\theta}_{j_0}) \log n \right) \Big] \to \infty \Big\} = 1,$$

διότι:

- $\dim(\boldsymbol{\theta}_l) > \dim(\boldsymbol{\theta}_{j_0})$, επειδή το J_0 είναι το σύνολο των μοντέλων με τις λιγότερες παραμέτρους και, άρα, η ποσότητα $\dim(\boldsymbol{\theta}_l) - \dim(\boldsymbol{\theta}_{j_0})$ είναι μία θετική σταθερά
- $\lim_{n\to\infty} \log n = \infty$.

Άρα, το BIC είναι συνεπές, ενώ το AIC όχι.
Σημειώνεται πως ακόμα ένα συνεπές Κριτήριο είναι το HQIC.

Αποδοτικότητα

Απόρροια του Θεωρήματος Ασυμπτωτικής Αποδοτικότητας, που αναφέρθηκε παράπανω, αποτελεί το κάτωθι:

Πόρισμα

- *Τα Κριτήρια AIC, FPE,* $AIC_c = AIC + \frac{2(|M|+1)(|M|+2)}{n-|M|+2}$ *και Mallows* C_p*, υπό τις υποθέσεις του Θεωρήματος Ασυμπτωτικής Αποδοτικότητας, είναι ασυμπτωτικά αποδοτικά*
- *Τα Κριτήρια BIC και* H*QIC δεν είναι ασυμπτωτικά αποδοτικά.*

Το συγκεκριμένο Πόρισμα αποδεικνύεται από τον Shibata (1980, θεώρημα 4.2).

Συμπεραίνουμε, λοιπόν, πως τα AIC και BIC, ικανοποιώντας τις ιδιότητες της ασθενούς και ισχυρής συνέπειας, επιλέγουν το μοντέλο με τη μικρότερη K-L-τιμή ως προς το πραγματικό, υπό την προϋπόθεση πως είναι ένα και μόνο ανάμεσα στα υποψήφια. Αντιθέτως, εάν δεν ισχύει η συγκεκριμένη προϋπόθεση και έχουμε περισσότερα του ενός μοντέλα με ελάχιστη K-L απόσταση ως προς το πραγματικό, τότε το BIC είναι το μόνο

που επιτυγχάνει την συνέπεια με την έννοια της φειδούς. Επίσης, το AIC έχει την ικανότητα ως βέλτιστο μοντέλο να επιλέξει εκείνο που ελαχιστοποιεί το αναμενόμενο τετραγωνικό σφάλμα πρόβλεψης, κάτι που δεν επιτυγχάνει το BIC. Είναι ενδιαφέρον να παρατηρήσουμε πως κανένα από τα δύο Κριτήρια δεν είναι ταυτοχρόνως συνεπές (με την έννοια της φειδούς) και αποδοτικό. Αυτό δεν είναι τυχαίο, καθότι, όπως έχει αποδειχθεί από τον Yang (2005), είναι αδύνατο οποιόδηποτε Κριτήριο να είναι και συνεπές και αποδοτικό.

Αν θέλουμε να προχωρήσουμε σε μία σύγκριση των δύο αυτών Κριτηρίων, πέραν της συνέπειας και της αποδοτικότητας, μπορούμε να σημειώσουμε πως, καταρχάς, είναι εμφανές ότι ο όρος ποινικοποίησης του BIC είναι «αυστηρότερος» του αντίστοιχου στο AIC κι, επομένως, έχει την τάση να επιλέγει πιο φειδωλά μοντέλα. Αυτός είναι ο λόγος για τον οποίο, παρά την ευρύτατη χρήση του, το AIC από κάποιους χαρακτηρίζεται ως υπερβολικά «γενναιόδωρο», αφού τείνει να επιλέγει πιο περίπλοκα μοντέλα από το BIC. Επίσης, αν και με τη πρώτη ματιά οι τύποι τους παρουσιάζουν ομοιότητες, πρέπει να γίνει κατανοητό πως προέρχονται από τελείως διαφορετικά πλαίσια. Το BIC, υπό την προϋπόθεση ότι ένα από τα υποψήφια μοντέλα είναι το πραγματικό, προσπαθεί να εντοπίσει εκείνα με τη μεγαλύτερη πιθανότητα να είναι τα πραγματικά. Αντιθέτως, κατά το AIC δεν υφίσταται αυτή η προϋπόθεση, αλλά γίνεται προσπάθεια, μέσω της χρήσης της απόστασης Kullback-Leibler, να επιλεχθεί το μοντέλο που περιγράφει με τον πιο επαρκή τρόπο την άγνωστη πραγματικότητα. Επομένως, είναι εμφανής η διαφορά στην ερώτηση που θέτει το κάθε Κριτήριο, αφού στην περίπτωση του BIC είναι η εύρεση του πραγματικού μοντέλου, ενώ στο AIC η πρόβλεψη μελλοντικών δεδομένων. Σύμφωνα με τους Wagenmakers & Farrell (2004), οι περισσότερες προσομοιώσεις που δείχνουν υπεροχή του AIC έναντι του BIC υποθέτουν πως η πραγματικότητα είναι απειροδιάσταση, άρα το πραγματικό μοντέλο δεν μπορεί να είναι ανάμεσα στα υποψήφια, ενώ στις προσομοιώσεις που καταδεικνύουν καλύτερη επίδοση του BIC έναντι του AIC, γίνεται η υπόθεση ότι και είναι κάποιο από τα υποψήφια και, μάλιστα, σχετικά μικρής διάστασης. Λόγω, λοιπόν, αυτής της διαφοράς πλαισίου των δύο Κριτηρίων, είναι πολύ δύσκολο να συγκριθούν τυπικά σε επίπεδο επιδόσεων.

8 **Εφαρμογή Κριτηρίων Επιλογής στην R**

Σ’ αυτό το κεφάλαιο παρουσιάζεται η εφαρμογή στην R για πραγματικά, αλλά και για προσομοιωμένα δεδομένα μεθόδων επιλογής μεταβλητών που χρησιμοποιούνται συνήθως σε γραμμικά μοντέλα παλινδρόμησης.

8.1 **Πραγματικά δεδομένα**

Τα πραγματικά δεδομένα που χρησιμοποιήθηκαν αφορούν μια έρευνα σε 97 άνδρες που πάσχουν από καρκίνο του προστάτη και πρόκειται να υποβληθούν σε ριζική προστατεκτομή.
Τα δεδομένα είναι αποθηκευμένα σ’ένα πλαίσιο 97 γραμμών και 9 στηλών, ονόματι **prostate**, η μορφή του οποίου είναι η εξής:

lcavol	**lweight**	**age**	**lbph**	**svi**	**lcp**	**gleason**	**pgg45**	**lpsa**
-0.58	2.77	50	-1.39	0	-1.39	6	0	-0.43
-0.99	3.32	58	-1.39	0	-1.39	6	0	-0.16
-0.51	2.69	74	-1.39	0	-1.39	7	20	-0.16
-1.20	3.28	58	-1.39	0	-1.39	6	0	-0.16
0.75	3.43	62	-1.39	0	-1.39	6	0	0.37
-1.05	3.23	50	-1.39	0	-1.39	6	0	0.77

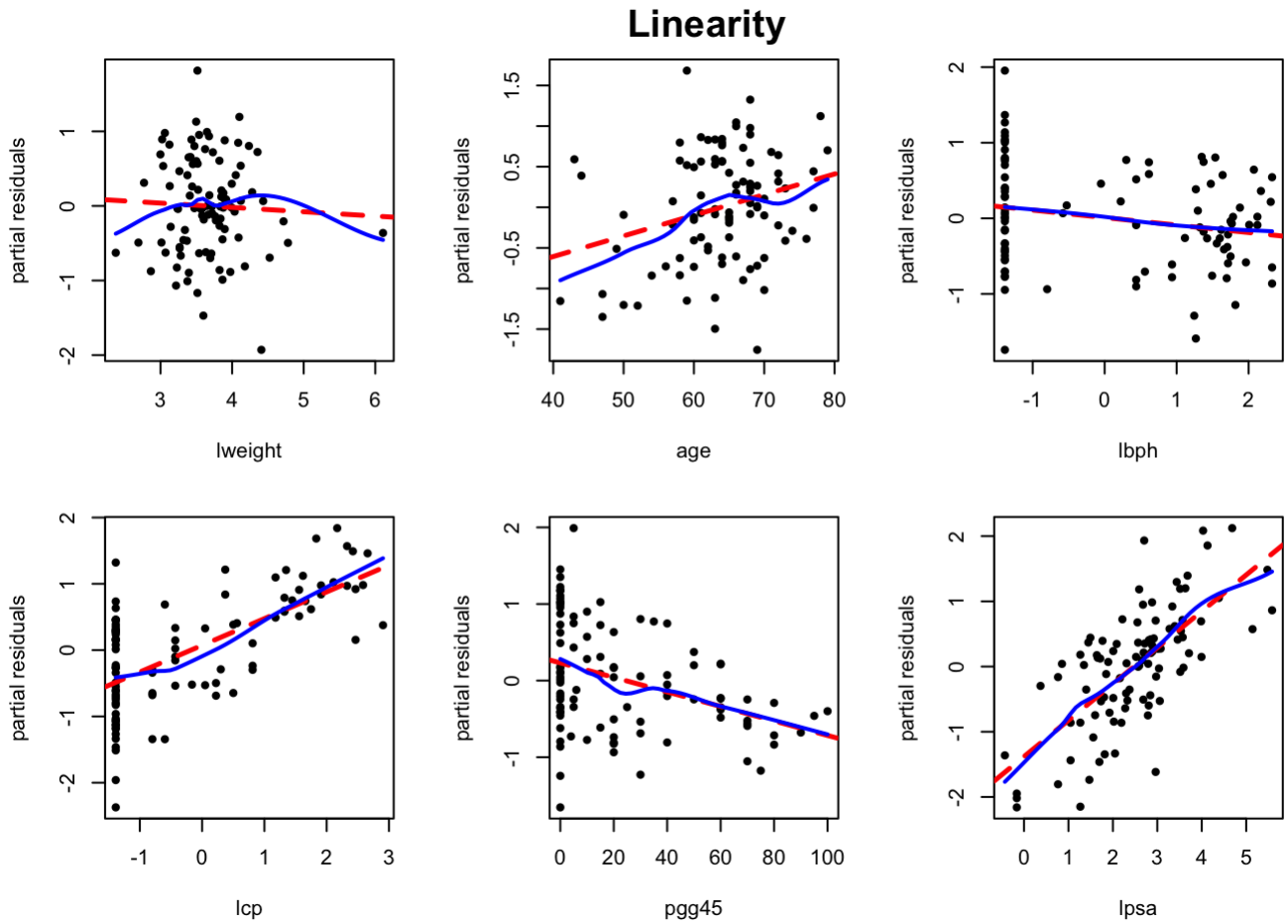
Στο παραπάνω πλαίσιο δεδομένων είναι:

- lcavol ο καρκινικός όγκος (σε λογαριθμική κλίμακα)
- lweight το βάρος προστάτη (σε λογαριθμική κλίμακα)
- age η ηλικία του ασθενή
- lbph το ποσόν καλοήθους υπερπλασίας του προστάτη (σε λογαριθμική κλίμακα)
- svi η διήθηση της σπερματοδόχου κύστης (0 ή 1)
- lcp η διήθηση της κάψας (σε λογαριθμική κλίμακα)
- gleason το Gleason score (σύστημα κατηγοριοποίησης του καρκίνου του προστάτη)
- pgg45 το ποσοστό Gleason score που είναι ίσο με 4 ή 5
- lpsa το προστατικό αντιγόνο (σε λογαριθμική κλίμακα)

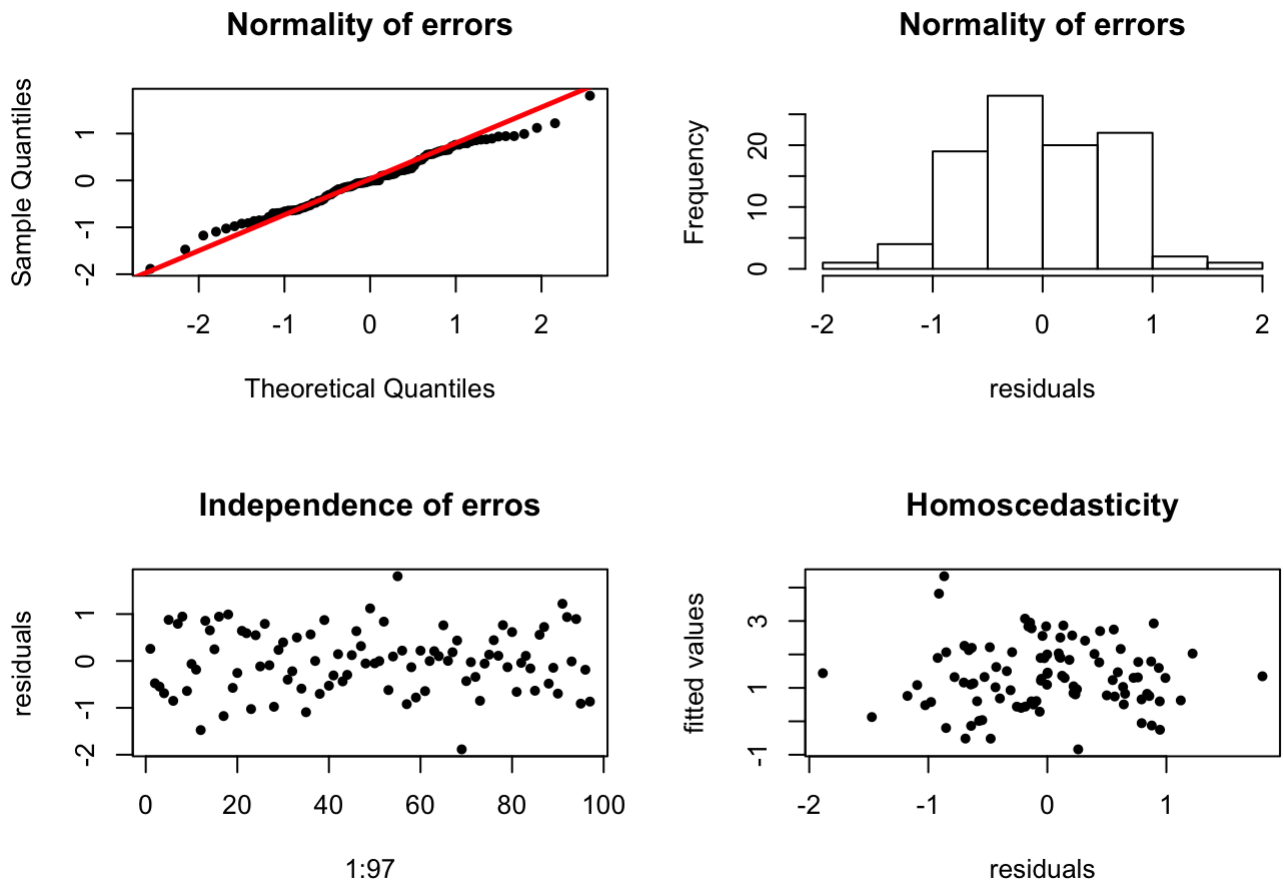
Στο γραμμικό μοντέλο που θα μελετηθεί, το ενδιαφέρον μας εστιάζεται στην πρόβλεψη του καρκινικού όγκου. Επομένως, η μεταβλητή απόκρισης είναι η lcavol και οι υπόλοιπες 8 επεξηγηματικές.

Οφείλουμε, όμως, προτού προχωρήσουμε στην περαιτέρω ανάλυση του μοντέλου να ελέγξουμε τις προϋποθέσεις του μοντέλου:

- ***Γραμμικότητα*** (*Linearity*): Εξετάζεται αν συνδέεται γραμμικά η κάθε επεξηγηματική μεταβλητή με τη δεσμευμένη μέση τιμή της μεταβλητής απόκρισης, υπό την προϋπόθεση πως όλες οι υπόλοιπες συνδέονται γραμμικά με τη δεσμευμένη μέση τιμή της μεταβλητής απόκρισης. Αυτό γίνεται εφικτό με την κατασκευή διαγράμματος διασποράς j-μερικών υπολοίπων $P_{ij} = \hat{\beta}_j \, x_{ij} + \varepsilon_i$, $(i = 1, \ldots, n \text{ και } \; j = 1, \ldots, p)$ για κάθε j ποσοτική επεξηγηματική μεταβλητή.
- ***Κανονικότητα σφαλμάτων*** (*Normality of errors*): Με τη χρήση qqplot και ιστογράμματος για τα υπόλοιπα, ελέγχεται εάν ακολουθείται η Κανονική Κατανομή.
- ***Ανεξαρτησία σφαλμάτων*** (*Independence of errors*): Ελέγχεται η τυχαία ή μη συμπεριφορά των υπολοίπων με ένα διάγραμμα υπολοίπων σε σχέση με την σειρά των δεδομένων.
- ***Ομοσκεδαστικότητα*** (*Homoscedascity*): Κατασκευάζεται γράφημα των υπολοίπων συναρτήσει των προβλεπόμενων τιμών της μεταβλητής απόκρισης, μέσω του οποίου διαπιστώνεται εάν η διασπορά των σφαλμάτων είναι ίδια για κάθε τιμή των επεξηγηματικών μεταβλητών. Δηλαδή, αν τα ζεύγη αυτών των τιμών δεν ακολουθούν συστηματικό τρόπο συμπεριφοράς, τότε πληρούται η συγκεκριμένη προϋπόθεση.



Εικόνα 8.1 *Έλεγχος Γραμμικότητας*



Εικόνα 8.2 *Έλεγχος Κανονικότητας σφαλμάτων, Ανεξαρτησίας σφαλμάτων και Ομοσκεδαστικότητας*

Συνεπώς, όπως παρατηρείται στην Εικόνα 8.1, αν εξαιρέσουμε την lweight, η υπόθεση της γραμμικότητας ικανοποιείται για όλες τις άλλες μεταβλητές. Επίσης, πληρούται η προϋπόθεση της κανονικότητας των σφαλμάτων, αφού στο qqplot της Εικόνας 8.2 πολλά σημεία συμπίπτουν με τη θεωρητική ευθεία και στο ιστόγραμμα η απεικόνισή τους προσεγγίζει την Κανονική Κατανομή. Τέλος, φαίνεται να μην ακολουθείται κάποια συστηματική συμπεριφορά στα δύο τελευταία διαγράμματα της Εικόνας 8.2, δηλαδή ικανοποιούνται οι υποθέσεις της ανεξαρτησίας και της ομοσκεδαστικότητας των σφαλμάτων.
Εφόσον οι προϋποθέσεις πληρούνται σε ικανοποιητικό βαθμό, προχωράμε στην προσαρμογή του μοντέλου:

```
Call:
lm(formula = lcavol ~ ., data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.88578 -0.48342 -0.01038  0.54901  1.80694

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.286986   0.825417  -1.559   0.1226
lweight     -0.057978   0.174868  -0.332   0.7410
age          0.025409   0.011137   2.281   0.0250 *
lbph        -0.098748   0.058312  -1.693   0.0940 .
svi1        -0.220671   0.256096  -0.862   0.3913
lcp          0.403017   0.084978   4.743 8.30e-06 ***
gleason7     0.300578   0.216390   1.389   0.1684
gleason8    -0.718327   0.758456  -0.947   0.3462
gleason9     0.773619   0.487484   1.587   0.1162
pgg45       -0.009469   0.004536  -2.087   0.0398 *
lpsa         0.557157   0.088071   6.326 1.09e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6969 on 86 degrees of freedom
Multiple R-squared:  0.6868,    Adjusted R-squared:  0.6504
F-statistic: 18.86 on 10 and 86 DF,  p-value: < 2.2e-16
```

Από τα αποτελέσματα της προσαρμογής συνάγεται ότι οι p-τιμές της ηλικίας (age), της διήθησης κάψας (lcp), του ποσοστού Gleason score 4 ή 5 (pgg45) και, τέλος, του προστατικού αντιγόνου (lpsa) είναι μικρότερες του 0.05. Αυτό σημαίνει πως οι συγκεκριμένες μεταβλητές είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας $\alpha = 5\%$, δηλαδή αν καθεμιά από αυτές αυξηθεί κατά μια μονάδα, ενώ οι υπόλοιπες επεξηγηματικές μεταβλητές παραμένουν σταθερές, προκαλείται μεταβολή στην αναμενόμενη τιμή του καρκινικού όγκου (lcavol). Επίσης, παρατηρείται πως η τιμή του πολλαπλού συντελεστή προσδιορισμού, R^2, ισούται με περίπου 0.69. Ο R^2, που παίρνει τιμές στο $[0, 1]$, εκφράζει το ποσοστό διασποράς της μεταβλητής απόκρισης που

εξηγείται με βάση το μοντέλο παλινδρόμησης, και, γενικά, όσο μεγαλύτερη είναι η τιμή του, υπό την προϋπόθεση ότι το μοντέλο είναι το κατάλληλο, τόσο πιο ισχυρή είναι γραμμική σχέση εξάρτησης της μεταβλητής απόκρισης με τις επεξηγηματικές. Κάθε φορά που προσθέτουμε μία επεξηγηματική μεταβλητή ο R^2 αυξάνεται. Για να μην υπάρχει, λοιπόν, αυτή η εξάρτηση υπολογίζεται ο διορθωμένος συντελεστής προσδιορισμού, $\tilde{R}^2$, του οποίου η τιμή στο συγκεκριμένο μοντέλο είναι ίση με 0.65. Κι οι δύο αυτές τιμές θεωρούνται ικανοποιητικές. Τέλος, το τυπικό σφάλμα, s, είναι περίπου 0.7, δε θεωρείται, δηλαδή, μεγάλο. Γενικά, όσο πιο μικρή τιμή έχει το s, τόσο καλύτερη είναι η προσαρμογή του μοντέλου.

Ένα τελευταίο ζήτημα που πρέπει να ελεγχθεί είναι η πολυσυγγραμμικότητα. Πολυσυγγραμμικότητα συναντάται σε ένα μοντέλο όταν υπάρχουν υψηλές συσχετίσεις μεταξύ των επεξηγηματικών μεταβλητών, δηλαδή όταν μία επεξηγηματική μεταβλητή μπορεί να προβλεφθεί γραμμικά με ικανοποιητική ακρίβεια από τις άλλες. Λόγω αυτού, οι συντελεστές παλινδρόμησης είναι εξαιρετικά ασταθείς και υφίστανται δραματικές αλλαγές τιμών σε περίπτωση που υπάρξουν μικρές μεταβολές στα δεδομένα του προβλήματος. Έτσι, τα αποτελέσματα που προκύπτουν δεν είναι έγκυρα, αφού δεν είναι δυνατή η αξιολόγηση της ουσιαστικής προσφοράς μιας συγκεκριμένης ανεξάρτητης μεταβλητής.
Ένας τρόπος να ελεγχθεί η ύπαρξη πολυσυγγραμμικότητας είναι μέσω του παράγοντα διόγκωσης διασποράς (*variance inflation factor*), του οποίου η τετραγωνική ρίζα μετρά τη «διόγκωση» του τυπικού σφάλματος σε σύγκριση με την τιμή του εάν η επεξηγηματική μεταβλητή δεν ήταν γραμμικά εξαρτημένη από τις άλλες επεξηγηματικές μεταβλητές. Για την $\mathbf{X}_j$ $(j = 1, . . , p)$ ο VIF ορίζεται ως

$$VIF_j = \frac{1}{1 - R_j^2},$$

όπου R_j^2 είναι ο συντελεστής προσδιορισμού στο μοντέλο όπου η $\mathbf{X}_j$ αποτελεί τη μεταβλητή απόκρισης και οι υπόλοιπες $p - 1$ του αρχικού μοντέλου αποτελούν τις επεξηγηματικές.
Γενικά, αν $VIF_j > 10$, τότε υπάρχει ένδειξη για έντονη πολυσυγγραμμικότητα.
Ο υπολογισμός του VIF για κάθε επεξηγηματική μεταβλητή στο μοντέλο που εξετάζεται μας δίνει τα εξής αποτελέσματα:

```
lweight      age     lbph     svi1      lcp gleason7 gleason8 gleason9
   1.49     1.36     1.41     2.22     2.79     2.28     1.17     2.32
  pgg45     lpsa
   3.24     2.04
```

Βλέπουμε, λοιπόν, πως δεν υπάρχει ένδειξη για έντονη πολυσυγγραμμικότητα στο μοντέλο, άρα μπορούμε να προχωρήσουμε στο επόμενο βήμα της ανάλυσης.

Ακολουθούν τα αποτελέσματα της χρήσης μεθόδων επιλογής μεταβλητών για το μοντέλο:

AIC

```
$min.AIC.value
[1] -64.76317

$best.model
(Intercept)        lpsa         lcp         age        lbph
-1.21022019  0.54272841  0.30907478  0.02003480 -0.08934519
```

Από την εφαρμογή του AIC παρατηρείται πως το καλύτερο μοντέλο περιλαμβάνει τις επεξηγηματικές μεταβλητές age, lbph, lcp και lpsa και έχει $AIC \approx$ -64.76.

BIC

```
$min.BIC.value
[1] -55.98627

$best.model
(Intercept)        lpsa         lcp
 0.09134534  0.53162111  0.32837535
```

Σύμφωνα με το BIC, το βέλτιστο μοντέλο, για το οποίο προέκυψε $BIC \approx$ -55.99 περιλαμβάνει μόνο τις lcp και lpsa.

EIC

```
$min.EIC.value
[1] 34.81095

$best.model.variables
[1] lcp,  lpsa
```

Το EIC, επίσης, επιλέγει τις ίδιες επεξηγηματικές μεταβλητές με το BIC. Στο επιλεγμένο μοντέλο το $EIC \approx 34.81$.

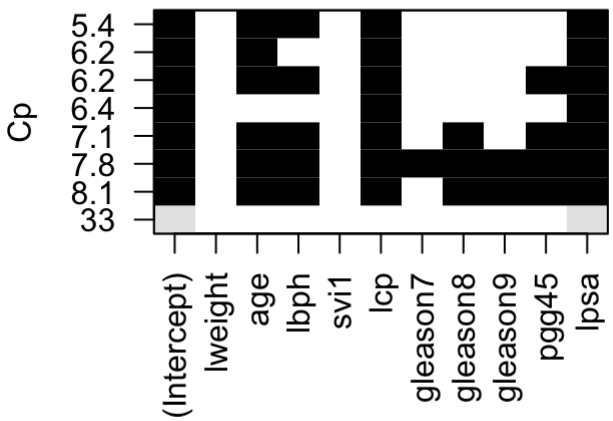
FIC

```
$min.FIC
[1] 20.60186

$best.model
(Intercept)     lweight         age        lbph         svi         lcp
-1.06283931 -0.07365342  0.02142020 -0.09034149 -0.20151683  0.33596961
       lpsa
 0.57570274
```

Το FIC επιλέγει ένα μοντέλο με $FIC \approx 20.6$, στο οποίο οι μόνες επεξηγηματικές μεταβλητές που απουσιάζουν είναι οι gleason και pgg45.
Σημειώνεται πως το παρόν Κριτήριο κατασκευάστηκε με στόχο στο καλύτερο μοντέλο να ελαχιστοποιείται το μέσο τετραγωνικό σφάλμα (MSE).

Mallows Cp

```
$min.Mallows_Cp.value
[1] 5.417451

$best.model
(Intercept)         age        lbph         lcp        lpsa
-1.21022019  0.02003480 -0.08934519  0.30907478  0.54272841
```



Εικόνα 8.3 *Κατάταξη υποψήφιων μοντέλων σύμφωνα με το Mallows* C_p

Στην Εικόνα 8.3 παρουσιάζεται μία κατάταξη των μοντέλων, σύμφωνα με το Mallows C_p, όπου τα σκούρα χρώματα δείχνουν ποιες μεταβλητές μετέχουν σε κάθε μοντέλο. Το συμπέρασμα και από το σχήμα και από τα αποτελέσματα είναι πως το Mallows C_p του βέλτιστου μοντέλου είναι περίπου 5.42. Επομένως, είναι πολύ κοντά στην τιμή 5, δηλαδή το πλήθος των παραμέτρων που επιλέχθηκαν, οι οποίες είναι η σταθερά, η age, η lbph, η lcp και η lpsa.

HQIC

```
$min.HQIC.value
[1] -67.16045

$best.model
(Intercept)        lpsa         lcp         age        lbph
-1.21022019  0.54272841  0.30907478  0.02003480 -0.08934519
```

Από το HQIC επιλέγεται ακριβώς το ίδιο μοντέλο με αυτό του Mallows Cp. Η ελάχιστη τιμή του Κριτηρίου είναι $HQIC \approx$ -67.16.

Στάθμιση με Akaike-βάρη

```
$models.weights
$models.weights$candidate.models
           delta     weight
1345   0.0000000 0.18500984
13457  0.7665242 0.12610904
145    0.8946208 0.11828520
45     1.0527658 0.10929232
13458  1.3193216 0.09565502
1457   1.6917601 0.07940243
13456  1.8038545 0.07507456
345    1.8848925 0.07209342
1456   1.9431633 0.07002325
123457 1.9710141 0.06905491

$models.weights$coefficient.codes
    age gleason    lbph     lcp    lpsa lweight   pgg45     svi
      1       2       3       4       5       6       7       8


$best.model
(Intercept)         age        lpsa         lcp        lbph
-0.85948970  0.01593413  0.54051844  0.33081639 -0.05374771
```

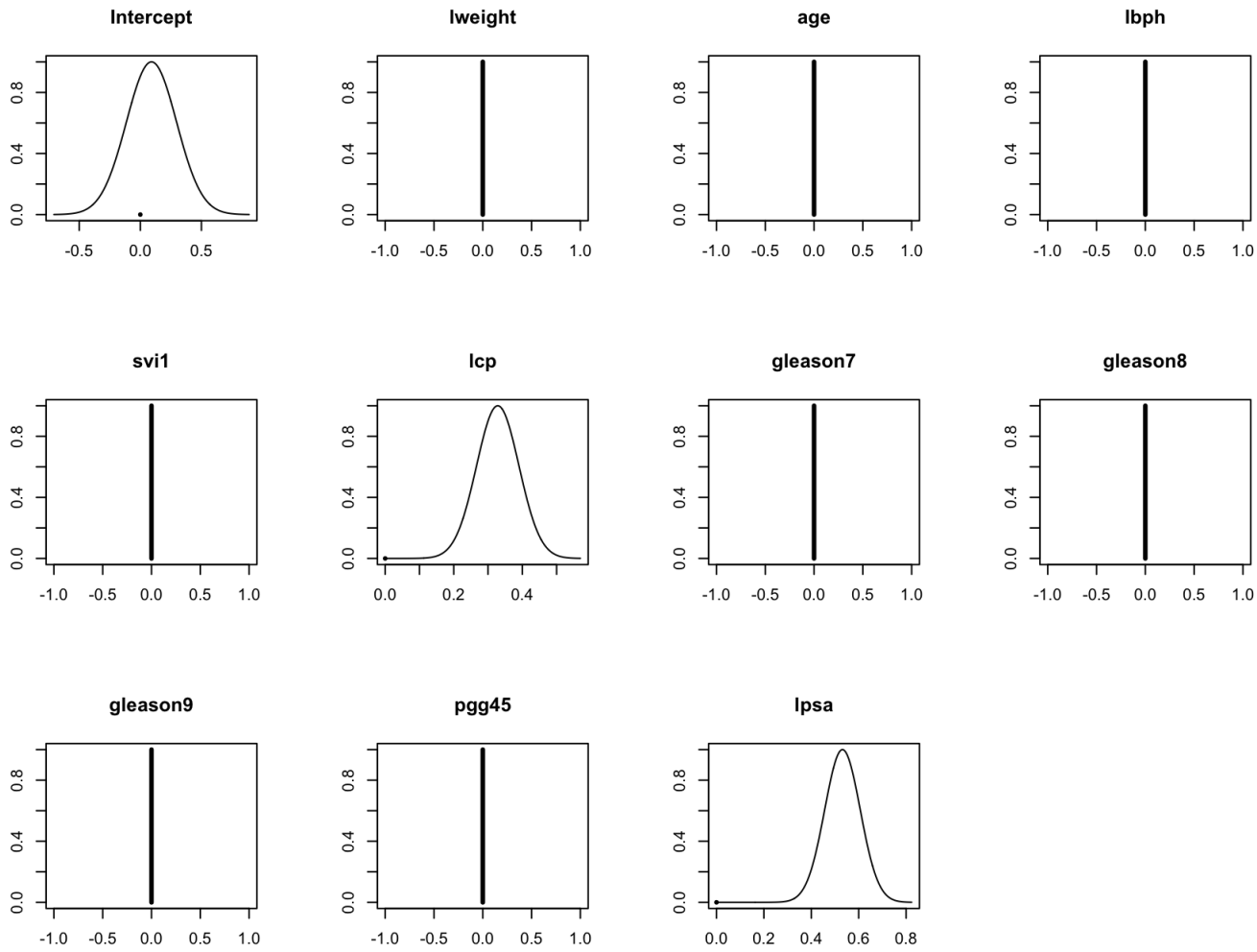
Στα αποτελέσματα της Στάθμισης με Akaike-βάρη βλέπουμε τη λίστα των υποψηφίων μοντέλων με Akaike διαφορές ≤ 2 (delta), δηλαδή τα μοντέλα που είναι τα καταλληλότερα, καθώς και τα βάρη τους (weight). Το καλύτερο έχει βάρος 0.19, δηλαδή

επί της ουσίας είναι περίπου $\frac{0.19}{0.13} = 1.46$ φορές πιο πιθανό να είναι καταλληλότερο, στα πλαίσια της απόστασης K-L, από το δεύτερο καλύτερο που έχει βάρος 0.13. Περιλαμβάνει τις μεταβλητές age, lbph, lcp και lpsa.

Δεν προκαλεί έκπληξη το γεγονός ότι καταλήγουμε εν τέλει στο ίδιο μοντέλο με το AIC, απλώς με τα Akaike-βάρη έχουμε στη διάθεσή μας και ένα ποσοτικό μέτρο ισχύος της επιλογής του συγκεκριμένου μοντέλου ως καλύτερου.

Μπεϋζιανή Στάθμιση

```
$posterior.probability.of.best.models
[1] 1

$best.models
(Intercept)         lcp        lpsa
 0.09134534  0.32837535  0.53162111
```



Εικόνα 8.4 *Εκ των υστέρων κατανομές των συντελεστών παλινδρόμησης του βέλτιστου μοντέλου*

Σύμφωνα με τη Μπεϋζιανή Στάθμιση, το βέλτιστο μοντέλο έχει ως επεξηγηματικές μεταβλητές μόνο τις lcp και lpsa. Μάλιστα, η εκ των υστέρων πιθανότητά του ισούται με 1.

Στην Εικόνα 8.4 απεικονίζονται οι εκ των υστέρων κατανομές των συντελεστών παλινδρόμησης. Όπως διακρίνεται, η εκ των υστέρων πιθανότητα ένας συντελεστής να είναι μηδενικός αναπαρίσταται από μία κάθετη ευθεία που περνά από το μηδέν και έχει ύψος ίσο με την πιθανότητα (βλ. π.χ. την lweight). Επομένως, παρατηρώντας και το συγκεκριμένο σχήμα, προκύπτουν τα ίδια συμπεράσματα για τις μεταβλητές του καταλληλότερου μοντέλου.

Τα αποτελέσματα της εφαρμογής των προαναφερθεισών μεθόδων παρουσιάζονται συνοπτικά στον Πίνακα 8.1:

Επεξηγηματικές μεταβλητές	**Μέθοδοι**							
	AIC	BIC	EIC	FIC	Mallows Cp	HQIC	MA	BMA
age	×			×	×	×	×	
gleason								
lbph	×			×	×	×	×	
lcp	×	×	×	×	×	×	×	×
lpsa	×	×	×	×	×	×	×	×
lweight				×				
pgg45								
svi				×				

Πίνακας 8.1 *Αποτελέσματα εφαρμογής μεθόδων επιλογής μεταβλητών*

Όσον αφορά τα ευρήματα, μπορούμε να σχολιάσουμε τα εξής:

Τα ίδια πιο φειδωλά μοντέλα, που έχουν μόνο δύο επεξηγηματικές μεταβλητές, επιλέχθηκαν από το BIC, τη Μπεϋζιανή Στάθμιση και το EIC. Στον αντίποδα, το καταλληλότερο, σύμφωνα με το FIC, μοντέλο αποτελεί και το πιο περίπλοκο, αφού έχει έξι μεταβλητές. Οι υπόλοιπες τέσσερις μέθοδοι (AIC, Mallows Cp, HQIC, Στάθμιση με Akaike-βάρη) επέλεξαν ως βέλτιστα ταυτόσημα «ενδιάμεσα» μοντέλα, που περιλαμβάνουν τέσσερις επεξηγηματικές μεταβλητές. Συνεπώς, σύμφωνα με την ανάλυση που προηγήθηκε, στην πρόβλεψη του καρκινικού όγκου απαραίτητες φαίνεται να είναι οι μεταβλητές για τη διήθηση κάψας (lcp) και το προστατικό αντιγόνο (lpsa), αφού επιλέγονται και από το σύνολο των μεθόδων επιλογής αλλά και ως στατιστικά σημαντικές λόγω p-τιμών. Ακριβώς το αντίθετο συμβαίνει με τη μεταβλητή του Gleason score, η

οποία μάλλον είναι περιττή. Πιθανότατα το ίδιο ισχύει και για την pgg45, διότι δεν εμπεριέχεται στο βέλτιστο μοντέλο καμιάς μεθόδου, αν και έχει p-τιμή ίση με 0.0398. Ακόμα, καθώς η age (p-τιμή= 0.025) περιλαμβάνεται στα καταλληλότερα μοντέλα 5 εκ των 8 μεθόδων, πολύ πιθανόν να είναι απαραίτητη για την πρόβλεψη. Κάτι ανάλογο ισχύει και για τη επεξηγηματική μεταβλητή που αφορά το ποσόν καλοήθους υπερπλασίας (lbph), της οποίας η p-τιμή είναι 0.094, δηλαδή σε επίπεδο σημαντικότητας $\alpha = 10\%$ είναι οριακά σημαντική. Τέλος, το βάρος του προστάτη (lweight) η διήθηση της σπερματοδόχου κύστης (svi) επιλέγονται μόνο από το FIC και, επίσης, είναι στατιστικά μη-σημαντικές μεταβλητές από άποψη p-τιμής. Επομένως, πολύ πιθανό να μην είναι χρήσιμες για την πρόβλεψη του καρκινικού όγκου.

8.2 **Προσομοιωμένα δεδομένα**

Λόγω της μεγάλης έκτασής τους, τα αποτελέσματα των προσομοιώσεων δεν παρατίθενται. Εν τούτοις, παρουσιάζονται τα συμπεράσματα που προέκυψαν. Σημειώνεται πως ο αντίστοιχος κώδικας βρίσκεται στο Παράρτημα.

Προσομοίωση για

$\mathbf{Y}|\mathbf{X} \sim N(3 + 2\,\mathbf{X}_{i1} + 0.8\,\mathbf{X}_{i2} - 1.2\,\mathbf{X}_{i15}, \sigma^2)$

Στη συγκεκριμένη προσομοίωση παράχθηκαν από μία Κανονική κατανομή $\mathrm{N}(0, 1)$ 20 τυχαίες μεταβλητές με πλήθος παρατηρήσεων ίσο με 100 η καθεμία, οι $\mathbf{X}_{ij}$ με $i = 1, \ldots, 100$ και $j = 1, \ldots, 20$, τις οποίες συμβολίζουμε ως $\mathbf{X}$. Επίσης, προσομοιώθηκαν 100 τιμές από μια τυχαία μεταβλητή, έστω $\mathbf{Y}$, για την οποία ισχύει ότι $\mathbf{Y}|\mathbf{X} \sim N(3 + 2\,\mathbf{X}_{i1} + 0.8\,\mathbf{X}_{i2} - 1.2\,\mathbf{X}_{i15}, \sigma^2)$. Οι $\mathbf{X}$ αποτέλεσαν τις επεξηγηματικές μεταβλητές του γραμμικού μοντέλου με μεταβλητή απόκρισης την $\mathbf{Y}$.
Δηλαδή, έχουμε

$$Y_i = 3 + 2\,\mathbf{X}_{i1} + 0.8\,\mathbf{X}_{i2} - 1.2\,\mathbf{X}_{i15} + \varepsilon_i,$$

όπου $i = 1, \ldots, 100$ και $\varepsilon_i \overset{iid}{\sim} (0, \sigma^2)$.
Η διαδικασία επαναλήφθηκε 100 φορές για $\sigma^2 = 1$ και $\sigma^2 = 2.5^2$.

Στόχος είναι να εξεταστεί εάν οι μέθοδοι επιλογής μεταβλητών διαλέγουν ως βέλτιστο το μοντέλο που περιλαμβάνει τις μεταβλητές από τις οποίες εξαρτάται η $\mathbf{Y}$, δηλαδή το πραγματικό μοντέλο, που σ' αυτή τη περίπτωση είναι γνωστό.
Τα αποτελέσματα που αφορούν τον εντοπισμό της πραγματικής κατανομής ακολουθούν στον Πίνακα 8.2:

(%) Εύρεση πραγματικού μοντέλου	Μέθοδοι								
	AIC	BIC	EIC	FIC	Mallows Cp	HQIC	MA	BMA	Καμία
$\sigma^2=1$	6	75	8	1	10	1	6	96	4
$\sigma^2=2.5^2$	4	41	10	0	6	1	4	78	21

Πίνακας 8.2 *Ποσοστό εύρεσης πραγματικής κατανομής*

Παρατηρούμε πως για $\sigma^2 = 1$ υπήρξαν 4 φορές που καμία μέθοδος δεν κατάφερε να εντοπίσει το πραγματικό μοντέλο. Για τις υπόλοιπες 96 η Μπεϋζιανή Στάθμιση επέλεξε το σωστό, ενώ το BIC για 75. Βέβαια, οφείλουμε να αναφέρουμε πως κατά την Μπεϋζιανή Στάθμιση, κάποιες φορές επιλέγονται πάνω από 1 κατάλληλα μοντέλα, αλλά το πραγματικό ήταν πάντοτε ανάμεσά τους. Το Mallows Cp και το EIC εντόπισαν επιτυχώς το βέλτιστο μοντέλο σε 10 και 8 επαναλήψεις αντιστοίχως. Ακόμα, το AIC επέλεξε το πραγματικό μοντέλο μόνο σε 6 φορές όπως και η Στάθμιση με Akaike-βάρη. Τέλος, τα χειρότερα αποτελέσματα ήταν αυτά των HQIC, FIC, τα οποία είχαν επιτυχία μόνο σε 1 περίπτωση.

Για $\sigma^2 = 2.5^2$ τα αποτελέσματα, όπως βλέπουμε στον Πίνακα 8.2, ήταν κάπως διαφορετικά. Υπήρξαν αρκετά περισσότερες επαναλήψεις που όλες οι μέθοδοι απέτυχαν να εντοπίσουν το πραγματικό μοντέλο, για την ακρίβεια 21. Το FIC απέτυχε πλήρως και το HQIC βρήκε μόνο μία φορά τη σωστή κατανομή. Το AIC και η Στάθμιση με Akaike-βάρη είχαν επιτυχία σε 4 περιπτώσεις. Επίσης, το EIC επέλεξε ορθά 10 φορές και το Mallows Cp 6. Τέλος, το BIC και η Μπεϋζιανή Στάθμιση εντόπισαν το πραγματικό μοντέλο σε 41 και 78 επαναλήψεις αντίστοιχα.

Επομένως, για αμφότερες τις τιμές της διασποράς, καταλήγουμε πως το BIC και η Μπεϋζιανή στάθμιση έχουν την καλύτερη επίδοση και, μάλιστα, με διαφορά από τα επόμενα καλύτερα που είναι τα EIC και Mallows Cp.

Προσομοίωση για

$$\mathbf{Y}|\mathbf{X} \sim N(4 + 2\,\mathbf{X}_{i1} - \mathbf{X}_{i5} + 1.5\,\mathbf{X}_{i7} + \mathbf{X}_{i11} + 0.5\,\mathbf{X}_{i13}, \sigma^2)$$

Με ανάλογο σκεπτικό δουλεύουμε και στη δεύτερη προσομοίωση.
Εδώ, προσομοιώθηκαν 10 μεταβλητές από μία Κανονική κατανομή N(0, 1), οι $\mathbf{X}_{im}^a$ με $i = 1, \ldots, 100$ και $m = 1, \ldots, 10$ και άλλες 5 από μία Κανονική κατανομή με μέση τιμή $0.3\,\mathbf{X}_{i1}^a + 0.5\,\mathbf{X}_{i2}^a + 0.7\,\mathbf{X}_{i3}^a + 0.9\,\mathbf{X}_{i4}^a + 1.1\,\mathbf{X}_{i5}^a$ και διασπορά ίση με 1, δηλαδή οι $\mathbf{X}_{il}^b$ με $i = 1, \ldots, 100$ και $l = 11, \ldots, 15$. Καθεμία από τις μεταβλητές έχει 100 παρατηρήσεις. Έτσι, από τις $\mathbf{X}_{im}^a$ και $\mathbf{X}_{il}^b$ προέκυψαν οι $\mathbf{X}_{ij}$ $i = 1, \ldots, 100$ και $j = 1, \ldots, 15$, τις οποίες συμβολίζουμε ως $\mathbf{X}$, και αποτέλεσαν τις επεξηγηματικές μεταβλητές του μοντέλου. Τέλος, προσομοιώθηκαν 100 τιμές από μια τυχαία μεταβλητή, την $\mathbf{Y}$, όπου $\mathbf{Y}|\mathbf{X} \sim N(4 + 2\,\mathbf{X}_{i1}^a - \mathbf{X}_{i5}^a + 1.5\,\mathbf{X}_{i7}^a + \mathbf{X}_{i11}^b + 0.5\,\mathbf{X}_{i13}^b, \sigma^2)$. Η $\mathbf{Y}$ αποτέλεσε τη μεταβλητή απόκρισης. Δηλαδή, το γραμμικό μοντέλο είναι

$$Y_i = 4 + 2\,\mathbf{X}_{i1} - \mathbf{X}_{i5} + 1.5\,\mathbf{X}_{i7} + \mathbf{X}_{i11} + 0.5\,\mathbf{X}_{i13}, + \varepsilon_i,$$

όπου $i = 1, \ldots, 100$ και $\varepsilon_i \overset{iid}{\sim} (0, \sigma^2)$.
Όπως φαίνεται από τη διαδικασία προσομοίωσης, λόγω της σχέσης των $\mathbf{X}_{il}^b$, $i = 1, \ldots, 100$, $l = 11, \ldots, 15$ με τις $\mathbf{X}_{im}^a$, $i = 1, \ldots, 100$, $m = 1, \ldots, 10$, είναι λογικό να παρουσιάζεται πολυσυγγραμμικότητα στο παραπάνω γραμμικό μοντέλο.

Και αυτή η προσομοίωση επαναλήφθηκε 100 φορές για $\sigma^2 = 1$ και $\sigma^2 = 2.5^2$.

(%) Εύρεση πραγματικού μοντέλου	**Μέθοδοι**								
	AIC	BIC	EIC	FIC	Mallows Cp	HQIC	MA	BMA	Καμία
σ^2=1	12	52	31	0	14	7	12	99	1
σ^2=2,5^2	14	30	22	0	17	8	14	51	30

Πίνακας 8.3 *Ποσοστό εύρεσης πραγματικής κατανομής*

Σύμφωνα με τον Πίνακα 8.3, για τιμή διασποράς $\sigma^2 = 1$, υπήρξε μόνο μία φορά που όλες οι μέθοδοι απέτυχαν. Όσον αφορά τις υπόλοιπες, και πάλι το FIC δεν κατάφερε να εντοπίσει το πραγματικό βέλτιστο μοντέλο ποτέ, ενώ το HQIC μόνο σε 7 επαναλήψεις.

Αντίθετα, η Μπεϋζιανή Στάθμιση ήταν επιτυχής σε 99 περιπτώσεις και το BIC σε 52. Επίσης, το σωστό μοντέλο βρέθηκε 31 φορές από το EIC και 14 από το Mallows Cp. Τέλος, το AIC και η Στάθμιση με Akaike-βάρη έδωσαν σωστά αποτελέσματα 12 φορές.

Και πάλι, για $\sigma^2 = 2.5^2$, παρατηρείται διαφοροποίηση στα αποτελέσματα, καθώς υπήρξαν αρκετές επαναλήψεις που όλες οι μέθοδοι απέτυχαν να εντοπίσουν το πραγματικό μοντέλο και συγκεκριμένα 30, όπως παρατηρείται στον Πίνακα 8.3. Το FIC απέτυχε και πάλι πλήρως, ενώ το HQIC βρήκε τη σωστή κατανομή σε 8 μόνο περιπτώσεις. Ακόμα, το πραγματικό μοντέλο επιλέχθηκε από το AIC και τη Στάθμιση με Akaike-βάρη 14 φορές, ενώ από το EIC σε 22 επαναλήψεις και από το Mallows Cp 17. Και πάλι το BIC και η Μπεϋζιανή Στάθμιση εντόπισαν το πραγματικό μοντέλο τις περισσότερες φορές, δηλαδή 30 και 51 αντίστοιχα. Ένα φαινόμενο που παρατηρήθηκε στη συγκεκριμένη προσομοίωση είναι πως τα μοντέλα που επέλεγαν το BIC και η Μπεϋζιανή Στάθμιση, όταν δεν επέλεγαν το σωστό, ήταν στις περισσότερες περιπτώσεις υπερβολικά φειδωλά, δηλαδή περιελάμβαναν $3 - 4$ επεξηγηματικές μεταβλητές.

Συνεπώς και για τη δεύτερη προσομοίωση, καταλήγουμε πως το BIC και η Μπεϋζιανή Στάθμιση έχουν την καλύτερη επίδοση, ιδιαίτερως η τελευταία. Βέβαια, για διασπορά $\sigma^2 = 2.5^2$ και στις δύο προσομοιώσεις αυξήθηκε αρκετά το πλήθος των επαναλήψεων στις οποίες αποτύγχανε το σύνολο των μεθόδων να εντοπίσει την πραγματική κατανομή.

Καταλήγοντας, είναι ξεκάθαρο ότι στις δύο περιπτώσεις προσομοιώσεων το BIC και η Μπεϋζιανή Στάθμιση υπερέχουν έναντι των υπολοίπων μεθόδων. Σημειώνεται, πάντως, πως και τα AIC, EIC, Mallows Cp και η Στάθμιση με Akaike-βάρη επέλεξαν τις περισσότερες φορές μοντέλα τεσσάρων, στην πρώτη προσομοίωση, και έξι, στη δεύτερη, μεταβλητών που περιείχαν και τις μεταβλητές του πραγματικού μοντέλου. Αντιθέτως, μοντέλα πολλών μεταβλητών επελέγησαν από το HQIC και το FIC, τα οποία υπήρξαν τελείως αδύναμα στον εντοπισμό του πραγματικού μοντέλου. Συμπεραίνουμε, λοιπόν, πως η φειδώ και, γενικότερα, η προσέγγιση των BIC και της Μπεϋζιανής Στάθμισης είχαν ως αποτέλεσμα την επιτυχή εύρεση της πραγματικής κατανομής. Ειδικότερα, η Μπεϋζιανή Στάθμιση είχε την καλύτερη επίδοση, γεγονός που φανερώνει το πλεονέκτημα της Στάθμισης. Βέβαια, πρέπει να αναφερθεί ότι στις συγκεκριμένες προσομοιώσεις υιοθετήθηκε η υπόθεση ότι υπάρχει πραγματικό μοντέλο και μάλιστα μικρών διαστάσεων σε σχέση με το πλήρες, κάτι που, όπως αναφέρθηκε και στην Ενότητα 7.3, ίσως και να «προωθεί» την καλύτερη επίδοση των συγκεκριμένων μεθόδων.

Παράρτημα

Στο Παράρτημα παρουσιάζεται ο κώδικας για την επιλογή μεταβλητών με τις απαραίτητες διευκρινίσεις.
Σημειώνεται πως ακολουθήθηκαν στην R οι τεχνικές που παρουσιάζονται από την Claeskens κατά την επιλογή μεταβλητών στα γραμμικά μοντέλα (βλ. Claeskens, G. (March 2011): *Short course: Model selection and Model Averaging*, University of Groningen).

Πραγματικά δεδομένα

Προσαρμογή μοντέλου

```
install.packages("faraway")
library(faraway,quietly=TRUE)
data(prostate)
prostate$gleason<-factor(prostate$gleason)
prostate$svi<-factor(prostate$svi)
results<-lm(lcavol ~., data=prostate)
```

AIC

```
install.packages("MASS")
library(MASS,quietly=TRUE)
fit0<-lm(lcavol~1, data=prostate)
results.AIC<-stepAIC(fit0,  k=2,direction="forward",trace=FALSE,scope=list(lower=~1,up
per=results)) #AIC
print(list(min.AIC.value=min(results.AIC$anova$AIC),
best.model=results.AIC$coefficients))
```

Έγινε χρήση της συνάρτησης *stepAIC* του πακέτου *MASS*. Ως αρχικό μοντέλο τέθηκε αυτό που περιλαμβάνει μόνο τη σταθερά (*fit0*), επιλέχθηκε η διαδοχική πρόσθεση (*direction="forward"*) ως τεχνική επιλογής και εύρος μοντέλων προς εξέταση από το αρχικό μέχρι αυτό που περιλαμβάνει όλες τις επεξηγηματικές μεταβλητές (*scope=list(lower=~1, upper=results)*). Η επιλογή μεταβλητών βασίστηκε στο AIC (*k=2*)

και η εμφάνιση πληροφοριών κατά το τρέξιμο της συνάρτησης δεν ήταν επιθυμητή (*trace=FALSE*). Ως αποτέλεσμα, επιστράφηκε η ελάχιστη AIC-τιμή αλλά και το μοντέλο στο οποίο συναντάται.

BIC

```
results.BIC<-stepAIC(fit0,k=log(nrow(prostate)),trace=FALSE,direction="forward",
scope=list(lower=~1,upper=results)) #BIC
print(list(min.BIC.value=min(results.BIC$anova$AIC),
best.model=results.BIC$coefficients))
```

Παρά το «παραπλανητικό» όνομά της, η *stepAIC* χρησιμοποιείται και για το BIC, αρκεί να αλλάξει η τιμή του *k*, που είναι η ποσότητα που πολλαπλασιάζεται με το *p* στον όρο ποινικοποίησης. Επομένως, εδώ θέσαμε *k=log(nrow(prostate))*.

EIC

```
install.packages("reams")
library(reams,quietly=TRUE)
results.EIC<-eic(y=prostate$lcavol,X=prostate[, -1],
nboot=100)
min.EIC<-min(results.EIC$eic)
min.EIC.model<-t(as.matrix(results.EIC$best))
colnames(min.EIC.model)<-colnames(prostate[,-1])
min.EIC.model.vars<-noquote(paste(names(which(min.EIC.model[1,]=="TRUE")),collapse=
",  "))
print(list(min.EIC.value=min.EIC,best.model.variables=min.
EIC.model.vars))
```

Εφαρμόστηκε η συνάρτηση *eic* του πακέτου *reams*. Τα ορίσματα που χρησιμοποιήθηκαν ήταν η μεταβλητή απόκρισης (*y=prostate$lcavol*), ο πίνακας των τιμών των επεξηγηματικών μεταβλητών (*X=prostate[, -1]*) και το πλήθος των bootstrap-δειγμάτων (*nboot=100*). Έτσι, επιστράφηκαν οι επεξηγηματικές μεταβλητές του βέλτιστου μοντέλου καθώς και η EIC-τιμή του.

FIC

```
source("http://feb.kuleuven.be/public/u0043181/modelselection/programs/FICforwardsear
ch.txt")
FIC<-function(data, var.fixed = NULL, FIC.type = "MSE") {
result <-NULL
responses<-data[,1]
var.names<-names(data[,-1])
data.predict<-data[,-1]
n.predict<-length(responses)
p<-length(var.fixed)
q<-length(var.names)-p
form.null<-paste(colnames(data)[1],"~ 1")
if (p>0) {
  var.float<-var.names[-match(var.fixed,var.names)]
  form.null<-paste(form.null,paste(var.fixed,collapse="+"),
sep=" + ")
} else {
  var.float<-var.names
}
form.full<-paste(form.null,paste(var.float,collapse="+"),sep="+")
model.full<-lm(as.formula(form.full),data=data)
Jn.full<-summary(model.full)$cov.unscaled
Jn.full<-solve(Jn.full) / n.predict
coefs.full<-model.full$coefficients
n.dummies<-pmax(1,unlist(lapply(data.predict,count.levels))-1)
names(n.dummies)<-var.names
p.dummies<-sum(n.dummies[var.fixed])
q.dummies<-sum(n.dummies)-p.dummies
d.float<-d.fixed<-NULL
var.factor<-unlist(lapply(data.predict,is.factor))
data.predict.rearranged<-NULL
  i<-0
for (v in c(var.fixed,var.float)) {
  i<-i+1
  t.data<-data.predict[,v]
  t.vars<-v
  if (var.factor[v]) {
    t.vars<-paste(v,levels(t.data)[-1],sep="")
    t.data<-model.matrix(~ t.data)[,-1]
  }
  if (i > p) {
    d.float<-c(d.float,t.vars)
  } else {
    d.fixed<-c(d.fixed,t.vars)
  }
data.predict.rearranged<-cbind(data.predict.
rearranged,t.data)
}
data.predict.rearranged<-as.matrix(cbind(1,data.predict.rearranged))
d.fixed<-c("(Intercept)",d.fixed)
rownames(data.predict.rearranged)<-rownames(data.predict)
colnames(data.predict.rearranged)<-c(d.fixed, d.float)
Jn.00<-as.matrix(Jn.full[d.fixed,d.fixed])
Jn.10<-as.matrix(Jn.full[d.float,d.fixed])
Jn.11<-as.matrix(Jn.full[d.float,d.float])
colnames(Jn.10)<-colnames(Jn.00)<-rownames(Jn.00)<-d.fixed
colnames(Jn.11)<-rownames(Jn.11)<-rownames(Jn.10)<-d.float
delta<-sqrt(n.predict)*coefs.full[d.float]
```

```
Qn.inv<-as.matrix(Jn.11-Jn.10%*%solve(Jn.00) %*% t(Jn.10))
result$variables<-matrix(NA, nrow=n.predict,ncol=q)
result$FIC<-matrix(NA,nrow=n.predict,ncol=q+1)
result$prediction<-rep(NA,n.predict)
result$focus<-rep(NA,n.predict)
rownames(result$variables)<-rownames(result$FIC)<-
names(result$prediction)
<-names(result$focus)<-
rownames(data.predict)
colnames(result$variables)<-var.float
colnames(result$FIC)<-paste("FIC",0:q,sep="-")
for (i in 1:n.predict) {
x.zero<-data.predict.rearranged[i, ]
if (FIC.type=="MSE") {
prediction <- FIC.MSE.search(x.zero, delta, var.float, d.fixed, d.float, n.dummies[va
r.float], Qn.inv, Jn.00, Jn.10, q)
  }
if (q > 0) {
result$variables[i, ]<-prediction$order
}
result$FIC[i, ]<-prediction$FIC.values
var.selected<-colnames(result$variables)[prediction$order>0]
n.selected<-sum(prediction$order > 0)
form<-form.null
if (n.selected > 0) {
form<-paste(form,paste(var.selected,collapse="+"),sep="+")
  }
model.fit<-lm(as.formula(form), data=data)
coefs<-rep(0,sum(n.dummies)+1)
names(coefs)<-c(d.fixed,d.float)
coefs[names(model.fit$coefficients)]<-model.fit$coefficients
result$focus[i]<-c(t(x.zero) %*% coefs)
}
result$prediction<-result$focus>0
return(list(min.FIC=min(result$FIC[n.predict,]),
best.model=model.fit$coefficients))
}

results.FIC<-FIC(data=prostate) #FIC
print(results.FIC)
```

Ο κώδικας για το FIC αποτέλεσε τροποποιημένη μορφή του προγράμματος στην ιστοσελίδα της Claeskens.

Πρόκειται για μία συνάρτηση που βρίσκει το βέλτιστο μοντέλο με βάση την ελαχιστοποίηση του MSE, χρησιμοποιώντας τη διαδοχική πρόσθεση μεταβλητών, και στην αρχική της μορφή γράφηκε από τον Van Kerckhoven.

Η συνάρτηση *FIC* χρησιμοποιεί τις συναρτήσεις *count.levels*, η οποία επιστρέφει το πλήθος των κατηγοριών μίας κατηγορικής μεταβλητής, την *FIC.MSE*, που υπολογίζει το MSE, και την *FIC.MSE.search*, με τη χρήση της οποίας πραγματοποιείται η διαδοχική πρόσθεση μεταβλητών. Ο κώδικας των τριών τελευταίων συναρτήσεων παρατίθεται στον σύνδεσμο που προαναφέρθηκε. Όσον αφορά το κυρίως πρόγραμμα της *FIC* διευκρινίζεται ότι με *p* συμβολίζονται οι μεταβλητές που θέλουμε να είναι στο επιλεγμένο

μοντέλο (αν υπάρχουν), *q* εκείνες που εξετάζεται αν θα συμπεριληφθούν, ενώ με *form.null* και *form.full* το περιορισμένο και το ευρύ μοντέλο αντίστοιχα. Μετά τον καθορισμό αυτών των ποσοτήτων, υπολογίστηκε ο πίνακας J_{wide}^{-1} (*Jn.full*) και κατασκευάστηκε ένα πλαίσιο δεδομένων, το *data.predict.rearranged*, με στοιχεία τους συντελεστές παλινδρόμησης κάθε επεξηγηματικής μεταβλητής αλλά και τη σταθερά για *n* μοντέλα παλινδρόμησης, όπου *n* είναι το μέγεθος του δείγματος. Ακόμα, βρέθηκε η $\hat{\boldsymbol{\delta}}_{wide}$ (*delta*). Στη συνέχεια, για να γίνει δυνατός ο υπολογισμός του πίνακα $Q_{wide} = (J_{11} - J_{10}J_{00}{}^{-1}J_{01})^{-1}$, υπολογίστηκαν τα στοιχεία του J_{wide}^{-1}, δηλαδή τα J^{00}, J^{10}, J^{11}, τα οποία στον κώδικα συμβολίζονται ως *Jn.00*, *Jn.10* και *Jn.11* αντίστοιχα και βρέθηκε ο πίνακας Q_{wide}^{-1} (*Qn.inv*). Έτσι, κατέστη δυνατή η εφαρμογή της *FIC.MSE.search* για κάθε μοντέλο (*x.zero*). Επίσης, βρέθηκε το βέλτιστο μοντέλο (*model.fit*) και, τελικά, επιστράφηκε η λίστα *result* με στοιχεία *variables*, *FIC*, *prediction* και *focus*. Το πρώτο στοιχείο της αναφέρεται στις μεταβλητές που επιλέχθηκαν και τη σειρά με την οποία επιλέχθηκαν στο κάθε μοντέλο, π.χ. αν κάποια μεταβλητή έχει τον αριθμό 2, σημαίνει ότι επιλέχθηκε δεύτερη και το αντίστοιχο μοντέλο περιέχει τη σταθερά, τη μεταβλητή που επιλέχθηκε πρώτη και εκείνη που επιλέχθηκε δεύτερη. Το δεύτερο στοιχείο αναφέρεται στις τιμές του FIC για όλα τα *n* μοντέλα και για εύρος μοντέλων από το περιορισμένο μέχρι το ευρύ. Για παράδειγμα, *FIC-2* κάθε μοντέλου αναφέρεται στο μοντέλο που προαναφέρθηκε. Το στοιχείο *prediction* είναι *TRUE* αν η παράμετρος ενδιαφέροντος είναι θετική και *FALSE* διαφορετικά. Το τελευταίο στοιχείο της λίστας περιέχει τις τιμές της παραμέτρου ενδιαφέροντος. Τέλος, επιλέχθηκε να επιστραφούν η ελάχιστη FIC-τιμή (*min.FIC*) και το βέλτιστο μοντέλο (*best.model*).

Mallows Cp

```
install.packages("leaps")
library(leaps,quietly=TRUE)
results.cp<-regsubsets(x=lcavol~.,data=prostate, intercept=TRUE, method="exhaustive")
 #Mallows' Cp
min.cp.value<-min(summary(results.cp)$cp)
min.cp.model<-coef(results.cp, which.min(summary(results.cp)$cp))
print(list(min.Mallows_Cp.value=min.cp.value,
best.model=min.cp.model))
plot(results.cp,scale="Cp", main="Mallows Cp plot")
```

Έγινε χρήση της *regsubsets* του πακέτου *leaps* με ορίσματα το μοντέλο (*x=lcavol~.*), το πλαίσιο δεδομένων (*data=prostate*), την προσθήκη της σταθεράς (*intercept=TRUE*) και τεχνική τη διαδοχική προσθήκη μεταβλητών (*method="forward"*). Επιστράφηκαν το βέλτιστο μοντέλο και η τιμή του Mallows Cp αυτού του μοντέλου. Επίσης, κατασκευάστηκε ένα γράφημα κατάταξης των μοντέλων (*plot(results.cp,scale="Cp", main="Mallows Cp plot")*).

HQIC

```
results.HQIC<-
stepAIC(fit0,k=log(log(nrow(prostate))),trace=FALSE,direction="forward",scope=list(lo
wer=~1,upper=results)) #HQIC
print(list(min.HQIC.value=min(results.HQIC$anova$AIC),best.model=results.HQIC$coeffic
ients))
```

Εφαρμόστηκε ξανά η *stepAIC* με *k=log(log(nrow(prostate)))*, καθότι αυτή η ποσότητα χρησιμοποιείται στον όρο ποινικοποίησης του HQIC.

Στάθμιση με Akaike-βάρη

```
install.packages("MuMIn")
library(MuMIn,quietly=TRUE)
results.MA<-dredge(results, rank=AIC) #MA (Akaike weights)
results.MA<-model.avg(results.MA,
subset=delta <= 2, fit=TRUE)
cand.models<-results.MA$msTable[,-c(1:3)]
cand.models.codes<-attr(results.MA$msTable, "term.codes")
cand.models.list<-list(candidate.models=cand.models,  coefficient.codes=cand.models.co
des)
best.model<-results.MA$coefficients["full",c("
(Intercept)","age","lpsa","lcp","lbph")]
print(list(models.weights=cand.models.list,best.model=best.model))
```

Έγινε χρήση των συναρτήσεων *dredge*, *model.avg* του πακέτου *MuMIn*. Επιστράφηκε το βέλτιστο μοντέλο, (*rank=AIC*), ως πρώτο στοιχείο μίας λίστας υποψηφίων μοντέλων που είναι πολύ πιθανό να είναι, επίσης, κατάλληλα (*subset=delta <= 2*). Το *fit=TRUE* τέθηκε για να προσαρμοστούν τα μοντέλα της λίστας.

Μπεϋζιανή Στάθμιση

```
install.packages("BMA")
library(BMA,quietly=TRUE)
results.BMA<-bicreg(x=prostate[,-1], y=prostate[,1], strict=TRUE) #BMA
best.postprob<-results.BMA$postprob
best.models<-results.BMA$postmean[results.BMA$postmean!=0]
print(list(posterior.probability.of.best.models=best.postprob,
best.models=best.models))
plot(results.BMA)
```

Χρησιμοποιήθηκε το πακέτο *BMA* και, συγκεκριμένα, η συνάρτηση *bicreg* με ορίσματα τις επεξηγηματικές (*x=prostate[,-1]*) και τη μεταβλητή απόκρισης (*y=prostate[,1]*). Έτσι, προέκυψαν το βέλτιστο ή τα βέλτιστα μοντέλα καθώς και οι εκ των υστέρων πιθανότητές τους. Θέσαμε *strict=TRUE* ώστε κάθε μοντέλο του οποίου κάποιο υπό-μοντέλο είναι

καταλληλότερο, να αφαιρείται από τη λίστα και, έτσι, να επιστραφούν πιο φειδωλά μοντέλα. Επίσης, κατασκευάστηκε γράφημα στο οποίο παρουσιάζονται οι εκ των υστέρων κατανομές των συντελεστών παλινδρόμησης.

Προσομοιωμένα δεδομένα

Προσομοίωση για

$$\mathbf{Y}|\mathbf{X} \sim N(3 + 2\,\mathbf{X}_{i1} + 0.8\,\mathbf{X}_{i2} - 1.2\,\mathbf{X}_{i15}, \sigma^2)$$

```
library(MASS,quietly=TRUE)
library(leaps,quietly=TRUE)
library(reams,quietly=TRUE)
library(MuMIn,quietly=TRUE)
library(BMA, quietly=TRUE)
source("FIC.R")
set.seed(1234)
sim1 <- function(sd, n, p) {
X <- matrix(rnorm((n*p),mean=0,sd=1), n, p)
mu <- 3 + 2*X[,1] + 0.8*X[,2] - 1.2*X[,15]
Y <- rnorm(n, mean=mu, sd=sd)
lm<-data.frame(Y,X)
lmf<- lm(Y~., data=lm)
lm0<-lm(Y~1, data=lm)
sim.AIC<-stepAIC(lm0,k=2,direction="forward",trace=FALSE,scope=list(lower=~1,upper=lm
f)) #AIC
sim.BIC<-stepAIC(lm0,k=log(n),direction="forward",  trace=FALSE,  scope=list(lower=~1,
upper=lmf)) #BIC
sim.HQIC<-stepAIC(lm0,k=log(log(n)),direction="forward",trace=FALSE,
scope=list(lower=~1, upper=lmf)) #HQIC
sim.cp<-regsubsets(Y~.,data=lm,intercept = TRUE,method = "forward")#Mallows
min.sim.cp<-coef(sim.cp, which.min(summary(sim.cp)$cp))
sim.EIC<-eic(Y,lm[,-1],nboot = 100) #EIC
min.sim.EIC<-sim.EIC$best
min.sim.EIC<-t(as.matrix(sim.EIC$best))
colnames(min.sim.EIC)<-colnames(lm[,-1])
min.sim.FIC<-FIC(data=lm)$best.model #FIC
sim.MA<-dredge(lmf, rank = AIC, m.lim=c(1,4))
sim.MA<-model.avg(sim.MA, fit = TRUE) #Model Averaging (Akaike weights)
min.sim.MA<-list(best.model=sim.MA$msTable[1,-c(1:3)],coefficient.codes=attr(sim.MA$m
sTable, "term.codes"))
sim.bma<-bicreg(X,Y, strict = TRUE) #Bayesian Model Averaging
min.sim.BMA<-sim.bma$which
results.list<-list(AIC=sim.AIC$coefficients,BIC=sim.BIC$coefficients,  HQIC=sim.HQIC$c
oefficients,  Mallows.Cp=min.sim.cp,  EIC=min.sim.EIC,FIC=min.sim.FIC,   MA=min.sim.MA,
BMA=min.sim.BMA)
return(results.list)
}
sim1a<-replicate(100, sim1(1,100,20),simplify = FALSE)
print(sim1a)
sim1b<-replicate(100, sim1(2.5,100,20),simplify = FALSE)
print(sim1b)
```

Προσομοίωση για

$$\mathbf{Y}|\mathbf{X} \sim N(4 + 2\,\mathbf{X}_{i1} - \mathbf{X}_{i5} + 1.5\,\mathbf{X}_{i7} + \mathbf{X}_{i11} + 0.5\,\mathbf{X}_{i13}, \sigma^2)$$

```
library(MASS,quietly=TRUE)
library(leaps,quietly=TRUE)
library(reams,quietly=TRUE)
library(MuMIn,quietly=TRUE)
library(BMA, quietly=TRUE)
source("FIC.R")
set.seed(1234)
sim2 <- function(sd, n, p, k) {
Xa <- matrix(rnorm((n*k), mean=0, sd=1), n, k)
Xb<-  matrix(rnorm((n*(p-k)),
mean= 0.3*Xa[,1]+0.5*Xa[,2]+0.7*Xa[,3]+0.9*Xa[,4]+1.1*Xa[,5],sd=1), n, p-k)
X <- cbind(Xa,Xb)
mu <- 4 + 2*X[,1] -X[,5] + 1.5*X[,7] + X[,11] +0.5*X[,13]
Y <- rnorm(n, mean=mu, sd=sd)
lm<-data.frame(Y,X)
lmf<- lm(Y~., data=lm)
lm0<-lm(Y~1, data=lm)
sim.AIC<-stepAIC(lm0,k=2,direction="forward",    trace=FALSE,    scope=list(lower=~1,
upper=lmf)) #AIC
sim.BIC<-stepAIC(lm0,k=log(n),direction="forward",  trace=FALSE,  scope=list(lower=~1,
upper=lmf)) #BIC}
sim.HQIC<-stepAIC(lm0,k=log(log(n)),direction="forward",trace=FALSE,
scope=list(lower=~1, upper=lmf)) #HQIC
sim.cp<-regsubsets(Y~.,data=lm,intercept = TRUE,method = "forward") #Mallows
min.sim.cp<-coef(sim.cp, which.min(summary(sim.cp)$cp))
sim.EIC<-eic(Y,lm[,-1],nboot = 100) #EIC
min.sim.EIC<-sim.EIC$best
min.sim.EIC<-t(as.matrix(sim.EIC$best))
colnames(min.sim.EIC)<-colnames(lm[,-1])
min.sim.FIC<-FIC(data=lm)$best.model #FIC
sim.MA<-dredge(lmf, rank = AIC, m.lim=c(1,6))
sim.MA<-model.avg(sim.MA, fit = TRUE) #Model Averaging (Akaike weights)
min.sim.MA<-list(best.model=sim.MA$msTable[1,-c(1:3)],coefficient.codes=attr(sim.MA$m
sTable, "term.codes"))
sim.bma<-bicreg(X,Y, strict = TRUE)
min.sim.BMA<-sim.bma$which
results.list<-list(AIC=sim.AIC$coefficients,BIC=sim.BIC$coefficients,  HQIC=sim.HQIC$c
oefficients,  Mallows.Cp=min.sim.cp,   EIC=min.sim.EIC,FIC=min.sim.FIC,  MA=min.sim.MA,
BMA=min.sim.BMA)
return(results.list)
}
sim2a<-replicate(100, sim2(1,100,15,10),simplify = FALSE)
print(sim2a)
sim2b<-replicate(100, sim2(2.5,100,15,10),simplify = FALSE)
print(sim2b)
```

Κατά τις προσομοιώσεις χρησιμοποιήθηκαν οι ίδιες τεχνικές και συναρτήσεις με εκείνες για τα πραγματικά δεδομένα. Μόνο μία διαφοροποίηση υπήρξε όσον αφορά τη Στάθμιση με Akaike-βάρη. Εδώ ήταν γνωστά τα πραγματικά μοντέλα και, άρα, το πλήθος των

επεξηγηματικών μεταβλητών σ’ αυτά. Έτσι, στη *dredge* το υποσύνολο υποψήφιων μοντέλων καθορίστηκε με χρήση του ορίσματος *m.lim*, με το οποίο ορίζεται ελάχιστο και μέγιστο πλήθος επεξηγηματικών μεταβλητών στα μοντέλα προς εξέταση.

Πηγές

Διεθνείς Πηγές

Behl, P., Dette, H., Frondel, M., Tauchmann H. (2012): Choice is suffering: A Focused Information Criterion for model selection, *Economic Modelling* **29**: 817-822

Burnham, K.P., Anderson, D.R. (2002): *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (Second edition)*, Springer

Burnham, K.P., Anderson, D.R. (2004): Multimodel Inference: Understanding AIC and BIC in Model Selection, *Sociological Methods & Research* **33**: 261-304

Busemeyer, J.R., Wang, Z., Townsend, J.T., Eidels, A. (eds) (2015): *The Oxford Handbook of Computational and Mathematical Psychology*, Oxford University Press

Claeskens, G., Hjort, N.L. (2008): *Model Selection and Model Averaging*, Cambridge University Press

Claeskens, G., Hjort, N.L. (2003): The Focused Information Criterion, *Journal of the American Statistical Association* **98**: 900–916

Hjort, N.L., Claeskens, G. (2003): Frequentist Model Average Estimators, *Journal of the American Statistical Association* **98**: 879-899

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999): Bayesian Model Averaging: A Tutorial, *Statistical Science* **14**: 382-417

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013): *An Introduction to Statistical Learning-with Applications in R*, Springer

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

Kuha, J. (2004): AIC and BIC: Comparisons of Assumptions and Performance, *Sociological Methods & Research* **33**: 188-229

MacKay, D.J.C. (2003): *Information Theory, Inference and Learning Algorithms*, Cambridge University Press

Miller, A.J. (2002): *Subset Selection in Regression (Second Edition)*, Chapman and Hall/CRC

Tichy, D. (2016): Multiple linear regression and variable selection, *Lecture Series: Biostatistical Case Studies using R and Bioconductor*, German Cancer Research Center (DKFZ)

Wagenmakers, E.J., Farrell (2004): AIC model selection using Akaike weights, *Psychonomic Bulletin & Review* **11**: 192–196

Ελληνικές Πηγές

Κοκολάκης, Γ., Φουσκάκης, Δ. (2009): *Στατιστική Θεωρία & Εφαρμογές*, Εκδόσεις Συμεών

Οικονόμου, Π., Καρώνη, Χ. (2010): *Στατιστικά Μοντέλα Παλινδρόμησης*, Εκδόσεις Συμεών

Φουσκάκης, Δ.(2013): *Ανάλυση Δεδομένων με χρήση της R*, Εκδόσεις Τσότρας

Ηλεκτρονικές Πηγές

www.math.rug.nl/stat/models/files/claeskens.pdf: Claeskens, G. (2011): *Short course: Model selection and Model Averaging,* University of Groningen

www.feb.kuleuven.be/public/u0043181/modelselection/programs/FICforwardsearch.txt: Claeskens, G., Van Kerckhoven, J. (2006): *Program: FIC for normal regression, with forward search strategy*

www.math.ntua.gr/~fouskakis/Intro_Bayesian_Analysis.pdf: Φουσκάκης, Δ. (2017): Εισαγωγή στην Μπευζιανή Στατιστική, *Μπευζιανή Στατιστική και MCMC*, Δ.Π.Μ.Σ. Εφαρμοσμένων Μαθηματικών Επιστημών, Σ.Ε.Μ.Φ.Ε., Ε.Μ.Π.