Συμπερασματολογία και επιλογή μεταβλητών για μοντέλα ποσοστημοριακής παλινδρόμησης

Πανεπιστήμιο Αθηνών
Τμήμα Μαθηματικών
Μεταπτυχιακό Στατιστικής και Επιχειρησιακής Έρευνας

Διπλωματική Εργασία

Χαράλαμπος Χανιαλίδης
Οκτώβριος 2010

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Στη μνήμη του πατέρα μου Παναγιώτη
και της γιαγιάς μου Ουρανίας.

Περίληψη

Σκοπός της εργασίας αυτής είναι η παρουσίαση της μεθόδου ποσοστημοριακής παλινδρόμησης. Η ποσοστημοριακή παλινδρόμηση είναι στη πραγματικότητα επέκταση της «κλασικής» γραμμικής παλινδρόμησης, πιο ανθεκτική σε ακραίες τιμές και μας επιτρέπει να ασχολούμαστε με όποιο ποσοστημόριο της δεσμευμένης κατανομής επιθυμούμε. Επίσης, η χρησιμοποίηση της ενδείκνυται σε περιπτώσεις όπου παραβιάζονται οι υποθέσεις του γραμμικού μοντέλου (ανεξαρτησία, κανονικότητα, ομοσκεδαστικότητα των τυχαίων όρων).

Στην παρούσα εργασία ασχολούμαστε με δύο βασικά θέματα στατιστικής συμπερασματολογίας για μοντέλα ποσοστημοριακής παλινδρόμησης. Πρώτον, αναφερόμαστε στην εκτίμηση των αγνώστων παραμέτρων του μοντέλου (και στην αξιολόγηση της αβεβαιότητας των εκτιμήσεων αυτών) χρησιμοποιώντας δύο βασικές μεθοδολογίες στατιστικής συμπερασματολογίας που στηρίζονται στη συνάρτηση πιθανοφάνειας: την εκτίμηση μέγιστης πιθανοφάνειας (κλασική προσέγγισή της) και τη μπεϋζιανή συμπερασματολογία. Δεύτερον, μας απασχολεί η επιλογή επεξηγηματικών μεταβλητών για μοντέλα ποσοστημοριακής παλινδρόμησης. Συγκεκριμένα, εντοπίζουμε ποιες, από ένα σύνολο διαθέσιμων επεξηγηματικών μεταβλητών, είναι οι καταλληλότερες να περιληφθούν στο μοντέλο.

Τέλος, εφαρμόζουμε τη μεθοδολογία εκτίμησης και επιλογής μοντέλου που αναπτύχθηκε σε χρηματοοικονομικά δεδομένα. Συγκεκριμένα, μέσω των κριτηρίων AIC και BIC, βρίσκουμε τα μοντέλα που εξηγούν καλύτερα τη μεταβλητότητα των αποδόσεων τεσσάρων αμοιβαίων κεφαλαίων υψηλού κινδύνου μέσω γραμμικής και ποσοστημοριακής παλινδρόμησης.

Ευχαριστίες

Αρχικά θέλω να ευχαριστήσω την οικογένεια μου και πιο συγκεκριμένα τα αδέρφια μου, Μάκη και Σπύρο Χανιαλίδη που άθελά τους έβαλαν τον πήχη αρκετά ψηλά κάτι το οποίο είχε ως αποτέλεσμα να προσπαθώ από μικρή ηλικία να τους φτάσω. Όσον αφορά τη μητέρα μου, Αγάπη Χανιαλίδη, δεν μπορώ να πω τίποτα άλλο πέρα από το ότι ήταν είναι και θα παραμείνει ο πιο σημαντικός άνθρωπος στη ζωή μου. Επίσης θα ήθελα να ευχαριστήσω τη Γραμματεία Κότσιαλου η οποία συνεχίζει να ανέχεται τις παραξενιές μου (και είναι πολλές) και είναι δίπλα μου όποτε και αν χρειάζομαι τη βοήθεια της. Αισθάνομαι υπερβολικά τυχερός που έχω ένα φίλο σαν τον Θάνο Τσουάνα η ανιδιοτέλεια του οποίου μερικές φορές είναι υπερβολική.

Ακόμα, η παρούσα διπλωματική εργασία δεν θα είχε πραγματοποιηθεί χωρίς τη βοήθεια της Λουκίας Μελιγκοτσίδου η οποία ακόμα και τώρα μου φαίνεται παράξενο για το πόσο υπομονετική ήταν μαζί μου. Η μεταδοτικότητα της, η σωστή καθοδήγηση της σε όλη τη διάρκεια της εργασίας και η ικανότητα της να μετατρέπει πολύπλοκα ζητήματα σε απλά ήταν απαραίτητα στοιχεία για την ολοκλήρωση της διπλωματικής εργασίας. Την ευχαριστώ πολύ και της ζητάω συγγνώμη για τις απίστευτες απορίες και ερωτήσεις που τις έστελνα σε email σχεδόν καθημερινά. Ακόμα θέλω να ευχαριστήσω τον Αποστόλη Μπουρνέτα για την βοήθειά του σε όλη τη διάρκεια των μεταπτυχιακών μου σπουδών. Ιδιαίτερη μνεία αξίζει στον Αντώνη Οικονόμου χωρίς τα μαθήματα του οποίου τα προπτυχιακά μου χρόνια στο Μαθηματικό Αθήνας θα ήταν τουλάχιστον βαρετά. Η αμεσότητα του και ο τρόπος διδασκαλίας του είναι παραδείγματα προς μίμηση.

Τέλος, θα ήθελα να πω ένα μεγάλο ευχαριστώ στους συμφοιτητές μου Χρήστο Γραμματικό, Νίκο Δικαιοσυνόπουλο και Βασίλη Σιτοκωνσταντίνου οι οποίοι εκτός από τη βοήθεια που μου έδωσαν σε αρκετά μαθήματα, ήταν πάντα προσιτοί και έτοιμοι να συζητήσουν ό,τι είχα στο μυαλό μου.

Κεφάλαιο 1

Γραμμική παλινδρόμηση

> There are three kinds of lies:
> lies, damned lies, and statistics.
>
> Benjamin Disraeli

Η ανάλυση παλινδρόμησης (regression analysis) περιλαμβάνει τεχνικές για ανάλυση συγκεκριμένων μεταβλητών και μοντελοποίηση αυτών όταν αυτό που μας ενδιαφέρει είναι η σχέση μεταξύ μίας μεταβλητής (Y) και μίας άλλης (X) ή πολλών άλλων μεταβλητών (X_i $i = 1, \ldots, n$). Η Y ονομάζεται εξαρτημένη μεταβλητή (response variable) και οι X_i ανεξάρτητες ή επεξηγηματικές (covariates).

Με την ανάλυση παλινδρόμησης έχουμε τη δυνατότητα να απαντήσουμε σε μια πληθώρα ερωτημάτων που αφορούν τη σχέση μεταξύ εξαρτημένης μεταβλητής και επεξηγηματικών μεταβλητών όπως «τι συμβαίνει στην εξαρτημένη μεταβλητή κατά μέση τιμή όταν κάποια επεξηγηματική αλλάζει, ενώ όλες οι υπόλοιπες παραμένουν σταθερές;» Όταν υπάρχει γραμμική σχέση μεταξύ της αναμενόμενης τιμής της Y και των X_i τότε αυτή μπορεί να περιγραφεί από ένα μοντέλο γραμμικής παλινδρόμησης (linear regression). Εδώ να τονίσουμε ότι είναι σπάνιο η πραγματική σχέση μεταξύ δύο ή περισσοτερών μεταβλητών να είναι τελείως γραμμική αλλά υποθέτουμε ότι το μοντέλο αποτελεί μια περιγραφή της σχέσης μεταξύ των μεταβλητών και δε συμμορφώνεται αναγκαστικά με τη πραγματική τους σχέση. Τη προσεγγίζει όμως τόσο ώστε να δικαιολογείται η χρήση του μοντέλου. Αποκλίσεις των δεδομένων από το εφαρμοζόμενο μαθηματικό μοντέλο μπορεί να απεικονίζουν την ασυμφωνία μεταξύ του μοντέλου και της φύσης, καθώς επίσης και πηγές ανακρίβειας στη συλλογή των πληροφοριών. Γενικά επιλέγουμε το μοντέλο που ταιριάζει όσο το δυνατόν περισσότερο με τα παρατηρούμενα δεδομένα έτσι ώστε το μεγαλύτερο μέρος της απόκλισης των δεδομένων από το μοντέλο να οφείλεται στην ανακρίβεια που προέρχεται

από τα ίδια τα δεδομένα και όχι στην ακαταλληλότητα του μοντέλου.

Στη περίπτωση της μίας επεξηγηματικής μεταβλητής έχουμε απλή γραμμική παλινδρόμηση (simple linear regression) ενώ όταν έχουμε παραπάνω, ονομάζεται πολλαπλή γραμμική παλινδρόμηση (multiple linear regression). Στις επόμενες παραγράφους θα δούμε λίγο πιο αναλυτικά το μοντέλο πολλαπλής γραμμικής παλινδρόμησης.

1.1 Το μοντέλο γραμμικής παλινδρόμησης

Έστω ότι από ένα πληθυσμό λαμβάνουμε ένα δείγμα μεγέθους n και για κάθε παρατήρηση καταγράφουμε τις τιμές της εξαρτημένης μεταβλητής και των ανεξάρτητων. Με βάση αυτά τα δεδομένα θα διερευνήσουμε τη σχέση μεταξύ των μεταβλητών. Έστω ότι έχουμε ένα δείγμα $\{Y_i, X_{i1}, \ldots, X_{ik}\},\ i = 1, \ldots, n$, από n παρατηρήσεις με k επεξηγηματικές μεταβλητές. Το μοντέλο της πολλαπλής παλινδρόμησης υποθέτει ότι η εξαρτημένη μεταβλητή συνδέεται με τις ανεξάρτητες με τη σχέση

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n,$$

όπου β_0 είναι ο σταθερός όρος της παλινδρόμησης και δείχνει τι τιμή έχει η εξαρτημένη μεταβλητή όταν όλες οι επεξηγηματικές εχουν την τιμή 0 και $\beta_1, \ldots, \beta_n$ είναι οι παράμετροι κλίσης και δείχνουν τι μεταβολή θα υποστεί κατά μέση τιμή η Y όταν υπάρξει μεταβολή μίας συγκεκριμένης επεξηγηματικής μεταβλητής κατά μία μονάδα, ενώ όλες οι υπόλοιπες παραμείνουν σταθερές (π.χ. η β_4 δείχνει την επίπτωση που θα έχει η Y όταν η X_4 μεταβληθεί κατά μία μονάδα ενώ όλες οι υπόλοιπες παραμείνουν σταθερές). Στο υπόδειγμα περιλαμβάνεται και ο στοχαστικός όρος ϵ_i (σφάλμα της παλινδρόμησης) ο οποίος είναι μία τυχαία μεταβλητή με τιμές μη παρατηρήσιμες. Αυτές οφείλονται σε τυχαίους μη σταθερούς παράγοντες που επηρεάζουν τις μεταβολές στις τιμές των ανεξάρτητων μεταβλητών. Για αυτό το λόγο οι τιμές του θεωρούνται ότι προέρχονται από κατανομές με $E(\epsilon_i) = 0$, είναι ανεξάρτητες μεταξύ τους καθώς και με τις τιμές των ανεξάρτητων μεταβλητών. Υποθέτουμε επίσης την ύπαρξη ομοσκεδαστικότητας (όλοι οι όροι ϵ_i έχουν την ίδια διακύμανση $\forall i = 1, \ldots, n$). Δηλαδή έχουμε ότι

$$E(\epsilon_i \epsilon_j) = \begin{cases} \sigma^2 & \forall i = j \\ 0 & \forall i \neq j \end{cases}$$

Κάτω από τις παραπάνω υποθέσεις για το στοχαστικό όρο του γραμμικού μοντέλου έχουμε ότι

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k.$$

Οι συντελεστές της πολλαπλής παλινδρόμησης $\beta_0, \beta_1, \ldots, \beta_k$, καθώς και η διασπορά του στοχαστικού όρου είναι άγνωστες παράμετροι και στόχος μας είναι να εκτιμήσουμε όλες τις παραπάνω παραμέτρους όσο το δυνατόν καλύτερα (δηλ. πιο κοντά στις πραγματικές τους τιμές).

Εδώ θα χρειαστεί να τονίσουμε δύο πράγματα. Αρχικά, ότι η γραμμικότητα του μοντέλου αναφέρεται στις παραμέτρους του και όχι στις επεξηγηματικές μεταβλητές του (π.χ. $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \ldots + \beta_n x_{ik} + \epsilon_i$ είναι ένα γραμμικό μοντέλο αντιθέτως με το $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2^2 x_{i2} + \ldots + \beta_n x_{ik} + \epsilon_i$). Επίσης, κάθε παράμετρος επηρεάζεται από όλες τις επεξηγηματικές μεταβλητές που περιλαμβάνονται στο μοντέλο. Έτσι, για παράδειγμα, η παράμετρος β_1 δεν εξηγεί την συνολική επίδραση της X_1 πάνω στην Y αλλά την «επιπλέον» επίδραση όταν προσθέσουμε την συγκεκριμένη μεταβλητή στο μοντέλο, αν όλες οι άλλες παραμείνουν ίδιες. Άρα ως αποτέλεσμα των παραπάνω έχουμε ότι οι εκτιμήσεις των παραμέτρων μπορεί να αλλάζουν κάθε φορά που είτε προσθέτουμε καινούργιες μεταβλητές στο μοντέλο είτε αφαιρούμε ήδη υπάρχουσες.

Το μοντέλο πολλαπλής παλινδρόμησης

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i \quad i = 1, \ldots, n.$$

μπορεί να γραφτεί σε μορφή πινάκων ως εξής. Αν θέσουμε

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

θα έχουμε ότι

$$Y = X\beta + \epsilon,$$

με τις υποθέσεις ότι $E(\epsilon) = 0$ και $V(\epsilon) = \sigma^2 I$ όπου σ^2 η διακύμανση του διαταρακτικού όρου και I ο μοναδιαίος πίνακας $n \times n$.

1.2 Εκτίμηση των συντελεστών παλινδρόμησης

Όπως είπαμε και πριν, σκοπός μας είναι να εκτιμήσουμε όσο το δυνατόν καλύτερα τις παραμέτρους του μοντέλου μας. Μία από τις πιο διαδεδομένες μεθόδους εκτίμησης είναι η μέθοδος των ελαχίστων τετραγώνων (ordinary least squares) η οποία είναι εννοιολογικά απλή και υπολογιστικά ξεκάθαρη.[1] Ένα άλλο αρκετά

[1] Στη περίπτωση ετεροσκεδαστικότητας η μέθοδος των σταθμισμένων ελαχίστων τετραγώνων(weighted least squares) δίνει καλύτερες εκτιμήσεις.

σημαντικό πλεονέκτημα αυτής της μεθόδου είναι ότι μπορεί να μας δώσει αρκετά ρεαλιστικά αποτελέσματα έχοντας ενα σχετικά μικρό σύνολο παρατηρήσεων. Σε αυτή τη μέθοδο η εκτίμηση των συντελεστών του γραμμικού υποδείγματος $\beta_0, \beta_1, \ldots, \beta_k$ στηρίζεται στην αρχή ότι οι εκτιμήσεις της εξαρτημένης μεταβλητής Y (που θα προκύψουν από την εκτίμηση των συντελεστών) θα πρέπει να έχουν την ελάχιστη δυνατή απόσταση από τις παρατηρούμενες τιμές τους. Αυτό θα συμβεί αν ελαχιστοποιήσουμε το άθροισμα των τετραγώνων των καταλοίπων $\hat{\epsilon_i}$ όπου ως κατάλοιπα ορίζουμε τη διαφορά μεταξύ των παρατηρούμενων τιμών των δεδομένων και των προβλεπόμενων ή προσαρμοσμένων τιμών που προκύπτουν από το γραμμικό μοντέλο. Άρα, αν θέσουμε

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{pmatrix}$$

το διάνυσμα των εκτιμώμενων παραμέτρων, η παλινδρόμηση εκτιμάται από την ευθεία

$$\hat{Y} = X\hat{\beta}$$

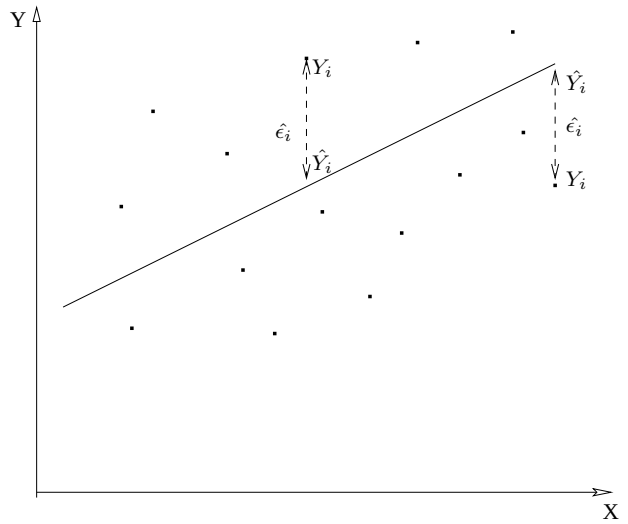
και τα κατάλοιπα ορίζονται ως

$$\hat{\epsilon_i} = Y - X\hat{\beta},$$

όπου η τιμή $\hat{Y_i}$ καλείται προσαρμοσμένη τιμή (fitted value) ενώ η Y_i παρατηρούμενη τιμή (observed value).

Στο σχήμα 1.1 ο οριζόντιος άξονας παριστάνει τις τιμές μιας ανεξάρτητης μεταβλητής και ο κάθετος τις τιμές της εξαρτημένης μεταβλητής Y. Οι κουκκίδες του σχήματος αντιστοιχούν σε ζεύγη παρατηρήσεων (Y_i, X_i) και η ευθεία που βρίσκεται στο σχήμα είναι η ευθεία της παλινδρόμησης η οποία όπως είπαμε και πριν, εκτιμάται από την ευθεία $\hat{Y} = X\hat{\beta}$ (δηλ. οι προσαρμοσμένες τιμές $\hat{Y_i}$ είναι σημεία της). Τα κατάλοιπα τα οποία βρίσκονται είναι οι αποκλίσεις των προσαρμοσμένων τιμών από τις παρατηρούμενες, δηλαδή $\hat{\epsilon_i} = Y_i - \hat{Y_i}$. Εδώ να τονίσουμε ότι η μέθοδος ελαχίστων τετραγώνων δεν ελαχιστοποιεί το άθροισμα των καταλοίπων, το οποίο είναι σταθερά 0 (γιατί οι αρνητικές τιμές των καταλοίπων απαλείφονται με τις θετικές τιμές τους), αλλά το άθροισμα των τετραγώνων τους.

Θέλουμε λοιπόν να ελαχιστοποιήσουμε το άθροισμα



Σχήμα 1.1: Γραμμή παλινδρόμησης και κατάλοιπα

$$\begin{aligned}\sum \hat{\epsilon_i}^2 = \hat{\epsilon}'\hat{\epsilon} &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}.\end{aligned}$$

Για να ελαχιστοποιήσουμε ως προς $\hat{\beta}$ πρέπει

$$\frac{\partial(Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta})}{\partial\hat{\beta}} = 0.$$

Χρησιμοποιώντας διαφορικό λογισμό έχουμε ότι το παραπάνω ισχύει όταν

$$X'X\hat{\beta} = X'Y$$

Αν ο πίνακας $X'X$ είναι αντιστρέψιμος, ο εκτιμητής ελαχίστων τετραγώνων που προκύπτει δίνεται από τον τύπο

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

Μπορεί να αποδειχτεί ότι $E(\hat{\beta}) = \beta$ και $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$ δηλ. ότι ο εκτιμητής $\hat{\beta}$ είναι αμερόληπτος και αποτελεσματικός.

Εδώ πρέπει να υπενθυμίσουμε ότι για να είναι ο πίνακας $X'X$ αντιστρέψιμος, πρέπει οι στήλες του παραπάνω πίνακα να αποτελούν ανεξάρτητα διανύσματα. Αυτό στη πραγματικότητα σημαίνει ότι οι ανεξάρτητες μεταβλητές του μοντέλου αποτελούν διαφορετικές πηγές ερμηνευτικότητας της εξαρτημένης μεταβλητής Y. Αν δε συμβαίνει αυτό (άρα $X'X$ μη αντιστρέψιμος δηλαδή ισχύει $|(X'X)| = 0$) τότε κάποιες από τις ανεξάρτητες μεταβλητές είναι τέλεια εξαρτημένες μεταξύ τους ή έχουν υψηλό βαθμό εξάρτησης. Αυτό το πρόβλημα μπορούμε να το λύσουμε με τον «αποκλεισμό» από το μοντέλο αυτών των μεταβλητών ή με αύξηση του αριθμού των παρατηρήσεων, ελπίζοντας ότι αυτό θα έχει ως αποτέλεσμα τη μείωση της συσχέτισης μεταξύ των ανεξάρτητων μεταβλητών, οι οποίες προηγουμένως παρουσίαζαν υψηλή συσχέτιση.

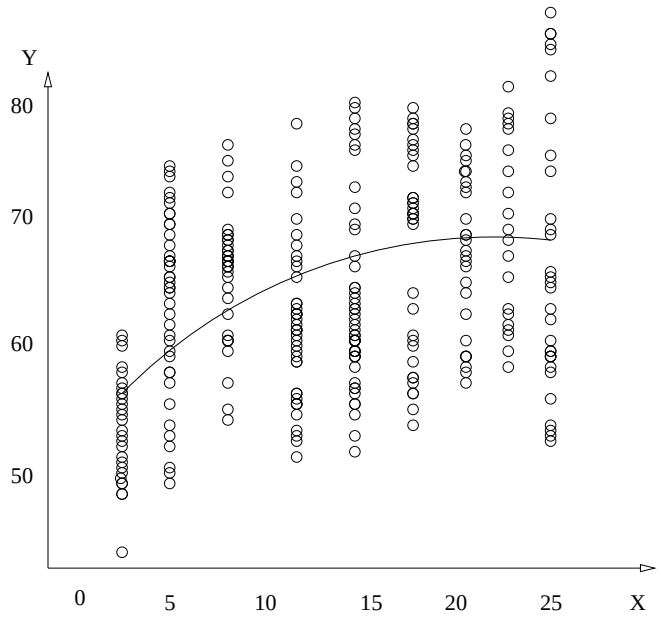
Ένας διαφορετικός τρόπος να προκύψει ο εκτιμητής ελαχίστων τετραγώνων (ε.ε.τ.) είναι να δούμε ότι η εκτίμηση του $\hat{\beta}$ ουσιαστικά γίνεται μέσω της ελαχιστοποίησης της τετραγωνικής συνάρτησης απώλειας (quadrative loss function) $r(u) = u^2$, όπου ελαχιστοποιείται το άθροισμα

$$\sum_{i=1}^{n} r(Y - X\hat{\beta}) = \sum_{i=1}^{n} (Y - X\hat{\beta})^2.$$

Η παραπάνω προσέγγιση στοχεύει στην εκτίμηση της δεσμευμένης μέσης τιμής $E[Y|X = x]$, η οποία προκύπτει ως η τιμή $\hat{Y}$ που ελαχιστοποιεί την δεσμευμένη τετραγωνική συνάρτηση απώλειας $E[(Y - \hat{Y})^2|X = x]$.

1.3 Μειονεκτήματα και ελλείψεις της γραμμικής παλινδρόμησης

Σε ορισμένες εφαρμογές, το μοντέλο της γραμμικής παλινδρόμησης παρουσιάζει μειονεκτήματα και ελλείψεις οι οποίες έχουν ως αποτέλεσμα να μην μπορούμε να βγάλουμε συμπεράσματα ή όταν μπορούμε αυτά να μην είναι σωστά. Ορισμένα συνήθη προβλήματα της γραμμικής παλινδρόμησης προκύπτουν όταν παραβιάζονται οι υποθέσεις του μοντέλου, δηλαδή η ανεξαρτησία, η κανονικότητα ή η ομοσκεδαστικότητα των τυχαίων όρων. Γενικότερα ωστόσο, η γραμμική παλινδρόμηση μοντελοποιεί μόνο τη δεσμευμένη μέση τιμή της εξαρτημένης μεταβλητής και όχι ολόκληρη τη κατανομή της, δηλαδή κάθε σημείο της ευθείας της παλινδρόμησης μας δείχνει ποια θα είναι η μέση τιμή της Y δεδομένου ότι οι ανεξάρτητες μεταβλητές παίρνουν κάποια συγκεκριμένη τιμή. Στη πραγματικότητα υποθέτει πως η επίδραση που έχει η κάθε μία ανεξάρτητη μεταβλητή πάνω στην εξαρτημένη Y είναι σταθερή σε όλα τα ποσοστημόρια της κατανομής της, άρα και στις ουρές της κάτι το οποίο μπορεί να μην είναι τόσο ρεαλιστική παραδοχή στη πράξη.



Σχήμα 1.2: Μισθοί 459 καθηγητών (σε χιλιάδες δολλάρια) ως συνάρτηση της διδακτικής τους εμπειρίας

Ας δούμε το παρακάτω παράδειγμα, το οποίο συσχετίζει το μισθό 459 καθηγητών στατιστικής στην Αμερική με τη διδακτική τους εμπειρία [;]. Στο σχήμα 1.2 ο οριζόντιος άξονας παριστάνει τα χρόνια διδακτικής εμπειρίας των καθηγητών και ο κάθετος τον ετήσιο μισθό τους (σε χιλιάδες δολλάρια). Η καμπύλη παλινδρόμησης του σχήματος είναι αυτή που συνδέει όσο το δυνατόν καλύτερα την ανεξάρτητη μεταβλητή X (διδακτική εμπειρία) με την εξαρτημένη Y (μισθός). Για τη μοντελοποίηση αυτής της σχέσης έχει χρησιμοποιηθεί καμπύλη παλινδρόμησης και όχι ευθεία γιατί χρησιμοποιήσαμε το μοντέλο $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$.

Θα πρέπει εδώ να αναφέρουμε ένα μεγάλο μειονέκτημα της μεθόδου των ελαχίστων τετραγώνων (μ.ε.τ.). Λόγω του ότι ασχολούμαστε με τη (δεσμευμένη) μέση τιμή της εξαρτημένης μεταβλητής Y γνωρίζουμε ότι πιθανές ακραίες τιμές (outliers) θα επηρεάσουν αρκετά τα αποτελεσματά μας, κάτι που μας απομακρύνει από το να βγάλουμε ρεαλιστικά συμπεράσματα. Πιο συγκεκριμένα, από την καμπύλη της παλινδρόμησης παρατηρούμε ότι $E(Y|X = 25) \approx 65$ δηλαδή ότι ένας καθηγητής που δουλεύει 25 χρόνια αμοίβεται με 65000 δολλάρια το χρόνο κατά μέσο όρο. Ενώ η προηγούμενη πρόταση μας δίνει κάποια πληροφορία για το συγκεκριμένο υπόδειγμα, η αλήθεια είναι πως δεν είναι αρκετή για να βγάλουμε ασφαλή συμπεράσματα για όλους τους καθηγητές με διδακτική εμπειρία 25 χρόνων. Δε μπορεί κάποιος να ισχυριστεί ότι οι καθηγητές οι οποί-

οι έχουν 25 χρόνια διδακτική εμπειρία αμοίβονται με 65000 δολλάρια. Πρέπει αρχικά να προσθέσει, στη προηγούμενη πρόταση, τις πολύ σημαντικές λέξεις «κατά μέσο όρο» αλλά ακόμα και τότε δεν είναι τόσο χρήσιμη η πληροφορία όσο ακούγεται αρχικά. Θα ήταν αρκετά πιο ενδιαφέρον να βρούμε το άνω όριο του μισθού των παραπάνω καθηγητών με πιθανότητα 0.5 δηλ. αφότου έχουμε διατάξει με αύξουσα σειρά (ως προς τους μισθούς τους) όλους τους καθηγητές με διδακτική εμπειρία 25 ετών να βρούμε την τιμή του μισθού (έστω y) τέτοια ώστε $P(Y < y|X = 25) = 0.5$.

Ερωτήσεις όπως « Τι μισθό παίρνει το 80% των καθηγητών οι οποίοι έχουν 25 χρόνια διδακτική εμπειρία;» ή «Ποιά είναι η διαφορά σε μισθούς του 90% των καθηγητών με 15 χρόνια διδακτικης εμπειρίας με το αντίστοιχο ποσοστο αυτών με 2 χρόνια διδ. εμπειρία;» αλλά και πολλές άλλες δεν μπορούν να απαντηθούν με τη γραμμική παλινδρόμηση στο μέσο.

Λογικό επακόλουθο των προηγουμένων ήταν να γεννηθεί η ανάγκη για ανθεκτικότερες (robust) εναλλακτικές μεθόδους που δε θα παρουσιάζαν τις παραπάνω αδυναμίες.

Κεφάλαιο 2

Ποσοστημοριακή παλινδρόμηση

Is it safe? Is it safe?

Marathon Man
Dr Christian Szell

2.1 Ποσοστημόρια κατανομών τυχαίων μεταβλητών

Όπως είδαμε στο προηγούμενο κεφάλαιο η γραμμική παλινδρόμηση παρουσιάζει κάποια σημαντικά προβλήματα τα οποία δε μας επιτρέπουν να έχουμε απαραίτητες πληροφορίες για τα δεδομένα μας, το οποίο έχει ως αποτέλεσμα να μη βγάζουμε ρεαλιστικά συμπεράσματα.

Συνοπτικά, στη γραμμική παλινδρόμηση, χρησιμοποιώντας τη μ.ε.τ. εκτιμούσαμε τη δεσμευμένη μέση τιμή $E[Y|X = x]$ και από αυτή, εξάγαμε τα συμπεράσματά μας. Η γραμμή (ή καμπύλη) της παλινδρόμησης μας έδινε τη δεσμευμένη μέση τιμή της εξαρτημένης μεταβλητής Y δεδομένου των ανεξάρτητων μεταβλητών X_i. Έχοντας μόνο αυτή τη πληροφορία είδαμε ότι δε γνωρίζουμε τι πραγματικά συμβαίνει στη δεσμευμένη κατανομή της εξαρτημένης μεταβλητής. Είναι κάτι αντίστοιχο με το να προσπαθούμε να αντλήσουμε πληροφορίες για μία κατανομή γνωρίζοντας μόνο τη μέση τιμή της. Τα συμπεράσματα που θα βγάζαμε είναι πολύ πιθανό να ήταν λανθασμένα και παραπλανητικά. Για το λόγο αυτό, προκειμένου να μπορέσουμε να εκτιμήσουμε τα ποσοστημόρια της εξαρτημένης μεταβλητής Y ως προς X_i καταφεύγουμε στη λεγόμενη ποσοστημοριακή παλινδρόμηση (quantile regression).

Αρχικά ας θυμηθούμε πως βρίσκουμε τα τεταρτημόρια ενός δείγματος και με αντίστοιχο τρόπο μπορούμε να βρούμε όποιο δειγματικό ποσοστημόριο θέλουμε. Έστω πως έχουμε το διατεταγμένο δείγμα

$$\{3, 6, 7, 8, 8, 10, 13, 15, 16, 20\}$$

το όποιο περιέχει άρτιο πλήθος στοιχείων (10 αριθμούς). Το πρώτο τεταρτημόριο είναι η τιμή που χωρίζει τα δεδομένα σε αναλογία 1 προς 3 όπου το $\frac{1}{4}$ των στοιχείων του δείγματος είναι μικρότερα της τιμής αυτής και τα $\frac{3}{4}$ είναι μεγαλύτερα. Εδώ βλέπουμε ότι το πρώτο τεταρτημόριο είναι η τιμή 7. Το δεύτερο τεταρτημόριο (διάμεσος) είναι η τιμή που χωρίζει τα δεδομένα σε αναλογία 1 προς 1 όπου τα μισά στοιχεία είναι μικρότερα της τιμής αυτής και τα άλλα μισά μεγαλύτερα. Τη τιμή αυτή τη βρίσκουμε από το ημιάθροισμα $\frac{8+10}{2} = 9$ αν και οποιοσδήποτε αριθμός ανάμεσα στο 8 και 10 ικανοποιεί τον προηγούμενο ορισμό της διαμέσου[1]. Εδω να τονίσουμε ότι η διάμεσος δεν είναι ίδια με τη μέση τιμή (10.6) του δείγματος κάτι που ισχύει γενικά και για τις κατανομές τυχαίων μεταβλητών εκτός αν η συγκεκριμένη κατανομή είναι συμμετρική. Αντίστοιχα, μπορούμε να πούμε πως το p-ποσοστημόριο είναι η τιμή που είναι μεγαλύτερη από το $\frac{p}{100}$ των τιμών και μικρότερη από το $\frac{1-p}{100}$ όπου $p \in (0, 100)$.

Λέμε επίσης ότι η επίδοση ενός μαθητή σε ένα διαγώνισμα βρίσκεται στο 90ο ποσοστημόριο αν ο μαθητής έχει αποδόσει καλύτερα από το 90% των μαθητών και χειρότερα από το 10%. Έτσι, μισοί μαθητές έχουν αποδόσει καλύτερα από τον μαθητή με επίδοση στο 50ο ποσοστημόριο και οι άλλοι μισοί χειρότερα.

2.2 Ποσοστημοριακές συναρτήσεις

Η ποσοστημοριακή παλινδρόμηση είναι στη πραγματικότητα επέκταση της «κλασικής» γραμμικής παλινδρόμησης. Παρουσιάστηκε το 1978 από τους Koenker και Bassett[**?**] ως μία πιο ανθεκτική μέθοδος σε ακραίες τιμές από αυτή των ελαχίστων τετραγώνων και σήμερα έχει αρκετές εφαρμογές, μεταξύ άλλων στην οικονομετρία και στην οικολογία. Γενικά, μπορεί να χρησιμοποιηθεί σε περιπτώσεις όπου το γραμμικό μοντέλο παρουσιάζει μία «αδύναμη» σχέση μεταξύ της μέσης τιμής της εξαρτημένης μεταβλητής και των τιμών των ανεξάρτητων, υπό την έννοια ότι ενώ μπορεί να υπάρχουν μεταβολές στις τιμές των ανεξάρτητων μεταβλητών, αυτές είναι τέτοιες ώστε να μην αντικατοπτρίζονται στη δεσμευμένη μέση τιμή της Y.

Η ιδέα της ποσοστημοριακής παλινδρόμησης είναι να εκτιμηθούν οι δεσμευμένες ποσοστημοριακές συναρτήσεις, μοντέλα στα οποία ποσοστημόρια της

[1] Αν το πλήθος των στοιχείων ήταν περιττός αριθμός π.χ. $2k+1$ τότε η διάμεσος θα ήταν η τιμή που θα ήταν μεγαλύτερη από k αριθμούς του συνόλου και μικρότερη από τους άλλους k

δεσμευμένης κατανομής της εξαρτημένης μεταβλητής εκφράζονται ως συναρτήσεις των ανεξάρτητων μεταβλητών. Ως ποσοστημοριακή συνάρτηση μιας κατανομής πιθανότητας F ορίζεται η αντίστροφη F^{-1} της αθροιστικής συνάρτησης κατανομής της.

Η ποσοστημοριακή συνάρτηση επιστρέφει τη τιμή, κάτω από την οποία θα πέσουν τυχαίες εκλογές από τη κατανομή, $p \times 100$ τοις εκατό των φορών, όπου $p \in (0, 1)$. Συγκεκριμένα, επιστρέφει τη τιμή x τέτοια ώστε

$$P(X \leq x) = p.$$

Σύμφωνα με τα προηγούμενα όταν έχουμε $p = 0.5$ οδηγούμαστε στη λεγόμενη συνάρτηση διαμέσου (median regression). Ας δούμε για παράδειγμα την Εκθετική(λ) κατανομή η οποία έχει αθροιστική συνάρτηση κατανομής

$$F(x) = 1 - e^{-\lambda x}.$$

Η αντίστροφη της είναι η

$$F^{-1}(p, \lambda) = -\frac{\ln\{1-p\}}{\lambda} \quad \text{για } 0 \leq p \leq 1.$$

Πιο συγκεκριμένα ακόμα, τα τεταρτημόρια της είναι

$$F^{-1}(0.25, \lambda) = \frac{ln\{4/3\}}{\lambda},$$
$$F^{-1}(0.5, \lambda) = \frac{ln\{2\}}{\lambda},$$
$$F^{-1}(0.75, \lambda) = \frac{ln\{4\}}{\lambda}.$$

Η τιμή που έχουμε βρει πιο πάνω για το πρώτο τεταρτημόριο μας δείχνει ότι το 25% των φορών που λαμβάνουμε τιμές από την κατανομή θα είναι μικρότερες της τιμής αυτής. Αντίστοιχα για το δεύτερο τεταρτημόριο έχουμε ότι

$$P(X \leq \frac{ln\{2\}}{\lambda}) = 0.5 \ .$$

Στον πίνακα 2.1 δίνονται οι ποσοστημοριακές συναρτήσεις ορισμένων γνωστών κατανομών.

2.3 Ποσοστημοριακή παλινδρόμηση

Για διευκόλυνση θα θέσουμε τη ποσοστημοριακή συνάρτηση

$$F_y^{-1}(p|X) = Q_y(p|X).$$

Πίνακας 2.1: Αθροιστικές συναρτήσεις γνωστών κατανομών και οι ποσοστημοριακές συναρτήσεις τους.

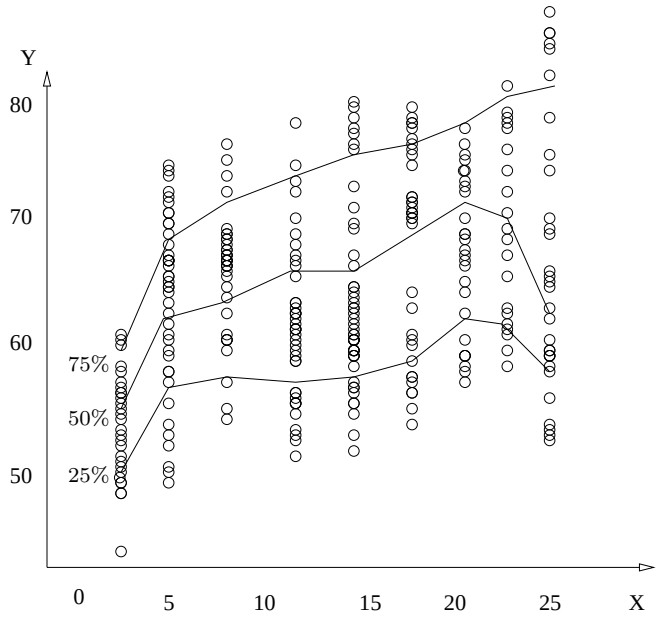
Κατανομή	$F_Y(y)$	$Q_Y(p)$
Εκθετική (λ)	$1-e^{-\lambda x}$	$-\frac{\ln\{1-p\}}{\lambda}$
Παρέτο (α,β)	$1-(\frac{\alpha}{y})^{\beta}$	$\alpha(1-p)^{-\frac{1}{\beta}}$
Ομοιόμορφη (α,β)	$\frac{y-\alpha}{\beta-\alpha}$	$p(\beta-\alpha)+\alpha$
Λογιστική (μ,s)	$\frac{1}{1+e^{-(y-\mu)/s}}$	$\mu-s\ln\frac{1-p}{p}$

Για παράδειγμα, αν $p=0.9$, $Q_y(0.9|X)$ δείχνει το 90ο ποσοστημόριο της δεσμευμένης κατανομής της Y ως προς X_i ή με άλλα λόγια 90% των τιμών της Y είναι μικρότερες ή ίσες από τη συγκεκριμένη συνάρτηση των X_i δηλαδή $P(Y\leq y|X)=0.9$.

Βρίσκοντας με αντίστοιχο τρόπο όποια ποσοστημόρια θέλουμε για τη συγκεκριμένη κατανομή που έχουμε, μέσω της ποσοστημοριακής παλινδρόμησης έχουμε ως αποτέλεσμα να εκτιμούμε καλύτερα τη δεσμευμένη κατανομή της Y και να έχουμε μία πιο πλήρη εικόνα για τις σχέσεις μεταξύ των μεταβλητών συγκριτικά με άλλες μεθόδους παλινδρόμησης.

Η δυνατότητα που έχουμε να ασχολούμαστε με όποιο ποσοστημόριο της κατανομής επιθυμούμε, είναι σημαντική σε πολλές εφαρμογές. Για παράδειγμα, όταν θέλουμε να βρούμε το ποσοστό ενός πληθυσμού το οποίο ζει κάτω από συνθήκες φτώχειας ή το ύψος των φόρων που πληρώνουν στο κράτος άνθρωποι με αρκετά υψηλά εισοδήματα. Γίνεται κατανοητό ότι στη πρώτη περίπτωση αυτό που μας ενδιαφέρει είναι η συμπεριφορά στη κάτω ουρά (low tail) της κατανομής, ενώ στη δεύτερη περίπτωση στην άνω (upper tail). Πρέπει να τονίσουμε ότι ενώ οι εκτιμήσεις των παραμέτρων στη γραμμική ποσοστημοριακή παλινδρόμηση έχουν την ίδια έννοια με αυτές σε οποιοδήποτε άλλο γραμμικό μοντέλο, εδώ οι εκτιμήσεις είναι διαφορετικές για κάθε ποσοστημόριο που εξετάζουμε. Άρα ένα ακόμα πλεονέκτημα της ποσοστημοριακής παλινδρόμησης (π.π.) είναι ότι αλλαγές σε διαφορετικά μέρη της κατανομής μπορούν να δικαιολογηθούν από την ύπαρξη διαφορετικών ανεξάρτητων μεταβλητών. Επίσης, ανεξάρτητες μεταβλητές που δε θα είχαν θεωρηθεί στατιστικά σημαντικές με τη μ.ε.τ., εδώ μπορούν να είναι. Ως αποτέλεσμα των παραπάνω, ανεξάρτητες μεταβλητές που είναι στατιστικά σημαντικές για κάποιο ποσοστημόριο δεν είναι απαραίτητο ότι θα είναι και για άλλα, κάτι που κάνει τα αποτελέσματα μας για τη δεσμευμένη κατανομή της Y πιο αξιόπιστα.

Η μ.ε.τ. παρουσιάζει όπως είδαμε πρόβλημα όταν υπάρχει ετεροσκεδαστικότητα (οι όροι ϵ_i δεν έχουν την ίδια διακύμανση) με αποτέλεσμα αλλαγές στη μέση τιμή να υποτιμούν, υπερτιμούν ή ακομά και να αποτυγχάνουν να παρατη-



Σχήμα 2.1: Οι καμπύλες ποσοστημοριακής παλινδρόμησης για τα τρία τεταρτημόρια

ρήσουν, σημαντικές αλλαγές στη δεσμευμένη κατανομή της εξαρτημένης Y.

Το μοντέλο της πολλαπλής γραμμικής ποσοστημοριακής παλινδρόμησης είναι

$$y_i = \beta_{0,p} + \beta_{1,p}x_{i1} + \ldots + \beta_{k,p}x_{ik} + \epsilon_{i,p} \quad i = 1, \ldots, n.$$

και ισχύει

$$Q_y(p|X) = \beta_{0,p} + \beta_{1,p}X_1 + \ldots + \beta_{k,p}X_k.$$

δηλαδή υποθέτουμε ότι

$$Q_y(\epsilon_{i,p}|X) = 0$$

όπου αντίστοιχα στη μ.ε.τ. είχαμε $E(\epsilon_i|X) = 0$. Από τη προηγούμενη σχέση καταλαβαίνουμε ότι για κάθε ποσοστημόριο που επιλέγουμε θα έχουμε και μια ξεχωριστή γραμμή παλινδρόμησης.

Χρησιμοποιώντας τώρα τις καινούργιες γνώσεις μας, στα δεδομένα του σχήματος 1.2, όπου φαίνεται η σύνδεση του μισθού 459 καθηγητών με τη διδακτική τους εμπειρία, μπορούμε να βρούμε γραμμές παλινδρόμησης για τα τεταρτημόρια της δεσμευμένης κατανομής Y (έτσι ώστε να έχουμε μία πιο ολοκληρωμένη εικόνα για τη κατανομή του μισθού τους).

Στο σχήμα 2.1 παριστάνονται πάλι τα ίδια δεδομένα με το σχήμα 1.2 αλλά εδώ δίνονται οι καμπύλες παλινδρόμησης για τα τρία τεταρτημόρια $(25\%, 50\%, 75\%)$

της Y. Όπως φαίνεται από το σχήμα, τώρα έχουμε μία πιο ξεκάθαρη περιγραφή των δεδομένων από αυτή που μας έδινε η γραμμική παλινδρόμηση για το συγκεκριμένο παράδειγμα. Εδώ φαίνεται ότι

$$P(Y \leq 60|X = 25) = 0.5$$

το οποίο σημαίνει ότι υπάρχει πιθανότητα 0.5 ένας καθηγητής με 25 χρόνια διδακτικής εμπειρίας να πληρώνεται το πολύ 60000 δολλάρια. Μία πληροφορία η οποία είναι αρκετά πιο σημαντική από το να γνωρίζουμε κατά μέση τιμή ποιος θα ήταν ο μισθός του (πληροφορία που παίρναμε χρησιμοποιώντας τη γραμμική παλινδρόμηση στο μέσο).

Ακόμα μπορούμε να δούμε από το σχήμα 2.1 ότι υπάρχει ένα 25% των καθηγητών του δείγματος με διδακτική εμπειρία 25 ετών το οποίο πληρώνεται με μισθούς άνω των 79000 δολλαρίων. Την παραπάνω πληροφορία την αντλούμε από την ευθεία παλινδρομήσης του 75ου ποσοστημόριου όπου ισχύει

$$P(Y \leq 79|X = 25) = 0.75$$

άρα

$$P(Y > 79|X = 25) = 0.25 \ .$$

Το ποσοστημόριο με το οποίο ασχολούμαστε πιο πολύ στις εφαρμογές αυτού του είδους παλινδρόμησης είναι το 50ο ή αλλιώς η διάμεσος. Στη ποσοστημοριακή παλινδρόμηση της διαμέσου εκτιμάται η δεσμευμένη διάμεσος της εξαρτημένης μεταβλητής Y δεδομένου των ανεξάρτητων X_i.

2.4 Εκτίμηση των παραμέτρων της ποσοστημοριακής παλινδρόμησης

Όπως στο προηγούμενο κεφάλαιο είδαμε ότι η δεσμευμένη μέση τιμή είναι η λύση σε ένα πρόβλημα ελαχιστοποίησης (του αθροίσματος των τετραγώνων των καταλοίπων) αντίστοιχα η δεσμευμένη διάμεσος της εξαρτημένης μεταβλητής Y ως προς τις ανεξάρτητες μεταβλητές X_i είναι επίσης η λύση σε ένα καινούργιο πρόβλημα ελαχιστοποίησης, αυτό της ελαχιστοποίησης του αθροίσματος των απολύτων τιμών των καταλοίπων. Επίσης είδαμε ότι για να προκύψει ο εκτιμητής ελαχίστων τετραγώνων, πρέπει να ελαχιστοποιήσουμε ως προς $\hat{\beta}$ το άθροισμα $\sum_{i=1}^{n} r(Y - X\hat{\beta})$ όπου $r(u) = u^2$ η τετραγωνική συνάρτηση απώλειας.

Στη ποσοστημοριακή παλινδρόμηση διαμέσου υποθέτουμε πάλι ότι ισχύει η σχέση

$$y_i = \beta_{0,0.5} + \beta_{1,0.5}x_{i1} + \ldots + \beta_{k,0.5}x_{ik} + \epsilon_{i,p} \quad i = 1, \ldots, n.$$

μεταξύ της εξαρτημένης μεταβλητής Y και των ανεξάρτητων X_i. Αν θέσουμε ότι

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \beta_{0.5} = \begin{pmatrix} \beta_0(0.5) \\ \beta_1(0.5) \\ \vdots \\ \beta_k(0.5) \end{pmatrix}, \epsilon_{0.5} = \begin{pmatrix} \epsilon_{1,0.5} \\ \epsilon_{2,0.5} \\ \vdots \\ \epsilon_{n,0.5} \end{pmatrix}$$

το μοντέλο ποσοστημοριακής παλινδρόμησης σε μορφή πινάκων γράφεται

$$Y = X\beta_{0.5} + \epsilon_{0.5}.$$

Για να προκύψει ο εκτιμητής του διανύσματος $\beta_{0.5}$ (για τη διάμεσο) ζητάμε να ελαχιστοποιηθεί ως προς $\beta_{0.5}$ το άθροισμα

$$\sum_{i=1}^{n} \rho_{0.5}(Y - X\beta_{0.5})$$

όπου $\rho_{0.5}(u) = 0.5|u|$ ονομάζεται απόλυτη συνάρτηση απώλειας(absolute loss function). Έχουμε επίσης ότι

$$\begin{aligned} \rho_{0.5}(u) &= 0.5|u| \\ &= 0.5uI_{[0,\infty)}(u) - (1-0.5)uI_{(-\infty,0)}(u). \end{aligned}$$

όπου

$$I_A(u) = \begin{cases} 1 & u \in A \\ 0 & u \notin A \end{cases}$$

δείκτρια συνάρτηση ενός συνόλου A.

Ο προηγούμενος ορισμός γενικεύεται για κάθε ποσοστημόριο p, όμως τώρα ζητάμε να ελαχιστοποιηθεί ως προς β_p το άθροισμα

$$\sum_{i=1}^{n} \rho_p(Y - X\beta_p)$$

όπου $\rho_p(u) = p|u|$ είναι η απόλυτη συνάρτηση απώλειας. Με αντίστοιχο τρόπο έχουμε

$$\begin{aligned} \rho_p(u) &= p|u| \\ &= puI_{[0,\infty)}(u) - (1-p)uI_{(-\infty,0)}(u). \end{aligned}$$

η οποία είναι γνωστή ως συνάρτηση ελέγχου (check function).

Επομένως έχουμε ότι ο εκτιμητής $\hat{\beta}_p$ για το συγκεκριμένο ποσοστημόριο p είναι

$$\hat{\beta}_p = \underset{\beta_p}{\operatorname{argmin}}\{\sum_{i=1}^{n} \rho_p(Y - X\beta_p)\}$$

Αντίστοιχα ο εκτιμητής διαμέσου είναι

$$\hat{\beta_{0.5}} = \underset{\beta_{0.5}}{\operatorname{argmin}}\{\sum_{i=1}^{n} \rho_{0.5}(Y - X\beta_p)\}.$$

Σε αντίθεση με τη μέθοδο ελαχίστων τετραγώνων, η εύρεση αυτού του εκτιμητή δεν είναι τόσο εύκολη. Μετατρέποντας όμως τα παραπάνω ως ένα πρόβλημα γραμμικού προγραμματισμού η εύρεση του μπορεί να επιτευχθεί. Μπορούμε να δούμε το y_i ως συνάρτηση θετικών στοιχείων, έτσι έχουμε

$$\begin{aligned} y_i &= \sum_{k=1}^{K} \beta_{k,p} x_{ik} + \epsilon_{i,p} \\ &= \sum_{k=1}^{K} (\beta^1_{k,p} - \beta^2_{k,p}) x_{ik} + (\mu_{i,p} - \nu_{i,p}). \end{aligned}$$

όπου $\beta^1_{k,p} \geq 0, \beta^2_{k,p} \geq 0 \ \forall k = 1, \ldots, K$ και $\mu_{i,p} \geq 0, \nu_{i,p} \geq 0 \ \forall i = 1, \ldots, n$.

Μετά τις παραπάνω μετατροπές, αρκει να βρούμε τη λύση στο ακόλουθο πρόβλημα

$$\min_{\beta^1_{k,p}, \beta^2_{k,p}, \mu_{i,p}, \nu_{i,p}} \sum_{i=1}^{n} p\mu_{i,p} + (1-p)\nu_{i,p}$$

όπου

$$y_i = \sum_{k=1}^{K} (\beta^1_{k,p} - \beta^2_{k,p}) x_{ik} + (\mu_{i,p} - \nu_{i,p})$$

και

$$\beta^1_{k,p}, \beta^1_{k,p}, \mu_{i,p}, \nu_{i,p} \geq 0 \quad \forall (i,k).$$

Αν θέσουμε

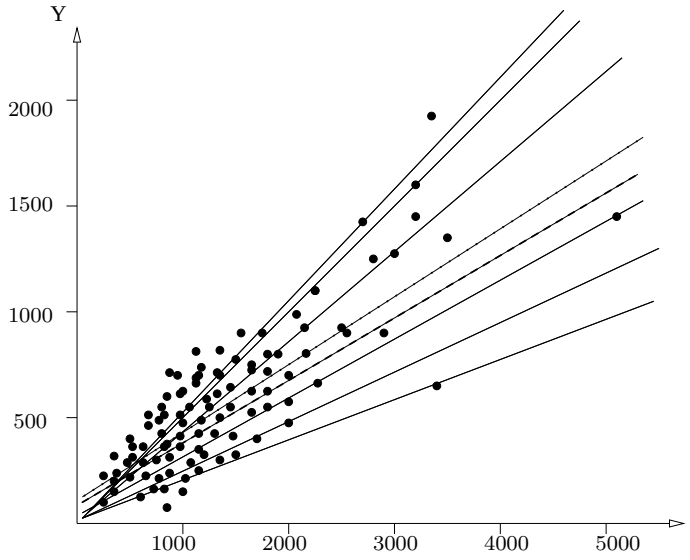
$$A = (X, -X, I, -I), \ z = (\beta^{1'}_{k,p}, \beta^{2'}_{k,p}, \mu^{'}_{i,p}, \nu^{'}_{i,p})$$

και

$$c = (0, 0, pi', (1-p)i')'$$

έχουμε το πρόβλημα γραμμικού προγραμματισμού

$$\min_{z} c'z$$



Σχήμα 2.2: Καμπύλες Engel για διαφορετικά ποσοστημόρια, σε 235 νοικοκυριά της Ευρώπης

$$Az = y \quad (z \geq 0)$$

και το αντίστοιχο δυαδικό πρόβλημα του

$$\max_{w} w'y$$

$$w'A \leq c'$$

όπου $w \in (p-1, p)^n$ [;]. Εδώ πρέπει να πούμε ότι αν ο πίνακας X είναι πλήρους τάξης τότε και το αρχικό αλλα και το δυϊκό πρόβλημα έχουν εφικτή λύση με ίσες βέλτιστες τιμές ($\min c'z = \max w'y$).

Ας δούμε ένα ακόμα παράδειγμα εφαρμογής της ποσοστημοριακής παλινδρόμησης. Μία καμπύλη Engel περιγράφει πως η ζητούμενη ποσότητα ενός προϊόντος αλλάζει, σε σχέση με αλλαγές στο εισόδημα του καταναλωτή. Γραφικά, η καμπύλη Engel παριστάνεται στο πρώτο τεταρτημόριο του καρτεσιανού συστήματος όπου στον άξονα των X βρίσκεται το εισόδημα και στον άξονα των Y βρίσκεται η κατανάλωση του συγκεκριμένου προϊόντος. Συγκεκριμένα στο σχήμα 2.2 παρουσιάζονται στοιχεία από έρευνα του 1857 σε 235 νοικοκυριά από όλη την Ευρώπη όπου στον άξονα των X βρίσκεται το εισόδημα των νοικοκυριών και στον άξονα των Y βρίσκονται τα χρήματα που σπαταλούν τα νοικοκύρια σε φαγητό[;]. Στο σχήμα επίσης υπάρχουν οι ευθείες παλινδρόμησης για τα ποσοστημόρια $p \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$ και η ευθεία

για τη δεσμευμένη μέση τιμή η οποία παριστάνεται με διακεκκομένες (ενώ της διαμέσου με κουκκίδες).

Από το γράφημα φαίνεται μία τάση αύξηση της διασποράς των χρημάτων που σπαταλούνται για φαγητό όσο αυξάνεται το εισόδημα των νοικοκυριών. Επίσης από την απόσταση που έχουν τα υψηλά ποσοστημόρια βλέπουμε ότι η δεσμευμένη κατανομή των χρημάτων που σπαταλούν τα νοικοκυριά για φαγητό είναι ασύμμετρη αριστερά δηλαδή ότι οι μεγάλες συχνότητες συγκεντρώνονται στο αριστερό άκρο της κατανομής. Οι ευθείες της δεσμευμένης μέσης τιμής και διαμέσου είναι αρκετά διαφορετικές κάτι που μπορεί να εξηγηθεί και από την ασυμμετρία της δεσμευμένης κατανομής αλλά και από τις δύο ακραιές τιμές στις οποίες έχουμε νοικοκυριά με μεγάλο εισόδημα να σπαταλούν λίγα (συγκριτικά με τα άλλα νοικοκυριά) χρήματα για φαγητό.

Εδώ, πρέπει να πούμε ότι υπάρχει και η λανθασμένη εντύπωση ότι η ποσοστημοριακή παλινδρόμηση θα μπορούσε να επιτευχθεί με την κατηγοριοποίηση της εξαρτημένης μεταβλητής Y σε υποσύνολα σύμφωνα με τη κατανομή της και μετά να εφαρμόσουμε τη μέθοδο των ελαχίστων τετραγώνων σε αυτά τα υποσύνολα. Να τονίσουμε ότι ακόμα και στα ακραία ποσοστημόρια λαμβάνονται υπ'οψιν όλες οι παρατηρήσεις του δείγματος κατά την εφαρμογή της ποσοστημοριακής παλινδρόμησης. Κάθε ευθεία που υπάρχει στο σχήμα 2.2 αν και σε τελική ανάλυση καθορίζεται από ένα ζευγάρι σημείων, χρειάζεται όλες τις παρατηρήσεις έτσι ωστε να επιλέξουμε το συγκεκριμένο ζευγάρι. Αντίστοιχα, όταν έχουμε k παραμέτρους να εκτιμήσουμε, χρειαζόμαστε k σημεία για την ευθεία αλλά το ποια θα είναι αυτά βασίζεται σε όλο το δείγμα. Αντιθέτως, το να χωρίσουμε το δείγμα μας σε υποσύνολα σύμφωνα με τις ανεξάρτητες μεταβλητές είναι μία έγκυρη επιλογή και συνήθως ακολουθείται στη μη παραμετρική ποσοστημοριακή παλινδρόμηση, όπου για κάθε υποσύνολο υπολογίζεται μια μεταβλητή ως ποσοστημόριο.

2.5 Εφαρμογές ποσοστημοριακής παλινδρόμησης

Τα τελευταία χρόνια η μέθοδος της ποσοστημοριακής παλινδρόμησης χρησιμοποιείται ολοένα και περισσότερο σε αρκετούς κλάδους της οικονομίας όπως για παράδειγμα τον κλάδο που αφορά την εργασία και τους μισθούς εργατών και υπαλλήλων (labour economics). Ο Chamberlain[**?**] βρήκε ότι όσο αφορά τους εργαζόμενους σε βιομηχανίες, η αύξηση που έχουν στο μισθό τους λόγω του ότι ανήκουν σε κάποιο συνδικάτο, στο 10ο ποσοστημόριο είναι 28% και μειώνεται όσο ανεβαίνουμε και φτάνει στο 90ο ποσοστημόριο στο 0.3%. Χρησιμοποιώντας τη μέθοδο ελαχίστων τετραγώνων θα είχαμε ότι η δεσμευμένη

μέση τιμή της αύξησης αυτής είναι 15.8% κάτι που σύμφωνα με τα προηγούμενα καταλαβαίνουμε ότι είναι παραπλανητικό και έχει «επηρεαστεί» από τη κάτω ουρά της δεσμευμένης κατανομής.

Μια άλλη εφαρμογή της μεθόδου βρίσκεται στην ιατρική και συγκεκριμένα στα διαγράμματα αναφοράς (reference charts) τα οποία χρησιμοποιούνται για τη προκαταρκτική διάγνωση του ατόμου έτσι ώστε να εντοπιστεί οποιαδήποτε μέτρηση η οποία βρίσκεται στην πάνω ή κάτω ουρά της κατανομής, μια πληροφορία η οποία συνήθως είναι απαραίτητη για τη περαιτέρω διάγνωση του ατόμου, λόγω του ότι επισημαίνει στον ιατρό μία μεγάλη απόκλιση από τα «αναμενόμενα».

Επίσης, στην ανάλυση επιβίωσης (survival analysis) οι εφαρμογές της ποσοστημοριακής παλινδρόμησης περιλαμβάνουν μελέτες συσχέτισης συγκεκριμένων μεταβλητών (π.χ. πόσα χρόνια καπνίζει ένα άτομο) με τους χρόνους επιβίωσης των ατόμων και αυτό γίνεται γιατί κάθε μεταβλητή έχει εντελώς διαφορετικές επιδράσεις σε χαμηλού, μεσαίου και υψηλού ρίσκου άτομα.

Η μέθοδος ποσοστημοριακής παλινδρόμησης χρησιμοποιείται αρκετά και στην υδρολογία, η οποία γενικά ασχολείται με τη κίνηση, κατανομή και ποιότητα του νερού στη γη αλλά και πιο συγκεκριμένα μοντελοποιεί βροχοπτώσεις και ρεύματα ποταμών. Για να κάνουμε μια σωστή πρόβλεψη για τη ποσότητα γλυκού νερού που χρειαζόμαστε για τα επόμενα 50 χρόνια χρειάζεται να ξέρουμε ποια είναι η πιθανότητα ξηρασίας, αντίστοιχα για το σχεδιασμό υπονόμων χρειάζεται να γνωρίζουμε τη πιθανότητα βροχόπτωσης. Αν υποθέσουμε ότι $q_p(x)$ είναι η ποσοστημοριακή συνάρτηση μιας μεταβλητής όπως το μέγιστο ετήσιο ύψος πλημμύρας, για δεδομένο p_0 τότε υπάρχει πιθανότητα $100(1-p_0)\%$ να υπερβεί το $q_{p_0}(x)$, κάτι που είναι πολυ σημαντικό να το γνωρίζουμε έτσι ωστε να λάβουμε προληπτικά μέτρα.

Μία ακόμα εφαρμογή της ποσοστημοριακής παλινδρόμησης είναι σε θέματα εκπαίδευσης και εκπαιδευτικής μεταρρύθμισης όπως στο αν υπάρχουν καλύτερα αποτελέσματα σε μία τάξη με λιγότερους μαθητές από μία άλλη, και εν τέλει αν είναι υπέρ του μαθητή να βρίσκεται σε μία τέτοια τάξη. Λόγω των πλεονεκτημάτων της μεθόδου, που ήδη έχουμε αναλύσει, τα αποτελέσματα που θα έχουμε θα αναφέρονται όχι μόνο στον «μέσο» μαθητή αλλά και στον «κακό» και σε αυτόν με άριστες επιδόσεις. Ο Levin[?] σε έρευνα του πάνω σε Ολλανδούς μαθητές, χρησιμοποιώντας τη ποσοστημοριακή παλινδρόμηση έβγαλε ως συμπέρασμα ότι το να μειωθεί το μέγεθος μιας τάξης δεν έχει ως άμεσο αποτέλεσμα τη βελτίωση των αποδόσεων των μαθητών, αντιθέτως μάλιστα όταν οι μαθητές που έφευγαν από την τάξη έτσι ώστε να μειωθεί το μεγεθός της ήταν μαθητές με κακές επιδόσεις, τότε αυτό είχε ως αποτέλεσμα οι εναπομείναντες «κακοί» μαθητές της τάξης να έχουν χειρότερα αποτελέσματα από πριν. Ο Tian[?] μέσω ποσοστημοριακής παλινδρόμησης συσχέτισε το οικογενειακό περιβάλλον (αριθμός γονιών, αριθμός αδερφών, φύλλο μαθητή, χώρα καταγωγής, κοινωνικοοικονομική κατάσταση γονιών, κ.α.) ενός μεγάλου δείγματος

μαθητών στο Καναδά με τα αποτελέσματα τους στα μαθηματικά και είδε ότι ο αριθμός των γονιών είναι ένας σημαντικός παράγοντας για την επίδοση των μαθητών από το 5ο ποσοστημόριο έως το 50ο. Αυτό το αποτέλεσμα επίσης δείχνει ότι ο αυξανόμενος αριθμός διαζυγίων των τελευταίων χρόνων επηρεάζει αρνητικά τις επιδόσεις των μαθητών. Επίσης ο αριθμός των αδερφών έχει αρνητική επίδραση στις αποδόσεις των μαθητών που βρίσκονται στα χαμήλα και μεσαία ποσοστημόρια. Τέλος, από την παραπάνω έρευνα βγαίνει το συμπέρασμα ότι η επίδραση που έχει η κοινωνικοοικονομική κατάσταση του πατέρα είναι πιο σημαντική από αυτή της μητέρας.

Κεφάλαιο 3

Συμπερασματολογία με βάση τη πιθανοφάνεια

> You come at the king, you best not miss.
>
> *The Wire*
> Omar Devone Little

Στα πρώτα 2 κεφάλαια είδαμε αρχικά πως μπορούμε να περιγράψουμε μαθηματικά σχέσεις μεταξύ μιας μεταβλητής και άλλων μεταβλητών. Η μοντελοποίηση της σχέσης μεταξύ των μεταβλητών αυτών είναι μία από τις σημαντικές δραστηριότητες με τις οποίες ασχολείται η στατιστική. Μία άλλη είναι η συμπερασματολογία, η οποία αναφέρεται στην εκτίμηση των αγνώστων παραμέτρων του μοντέλου αλλά και στην αξιολόγηση της αβεβαιότητας των εκτιμήσεων. Με το πρώτο μέρος της στατιστικής συμπερασματολογίας (εκτίμηση παραμέτρων) ασχοληθήκαμε στα προηγούμενα κεφάλαια. Είναι φανερό ότι υπάρχει ένας σημαντικός δεσμός μεταξύ της στατιστικής μοντελοποίησης και της συμπερασματολογίας, εννοώντας ότι αν το εφαρμοζόμενο μαθηματικό μοντέλο που έχουμε επιλέξει δεν είναι το κατάλληλο τότε δεν έχουν καμία «δύναμη» οι εκτιμήσεις των παραμέτρων που έχουμε βρει και θα είναι παραπλανητικές. Στο κεφάλαιο αυτό, αφού πρώτα ορίσουμε τη συνάρτηση πιθανοφάνειας θα αναφερθούμε σε 2 βασικές μεθοδολογίες στατιστικής συμπερασματολογίας που στηρίζονται σε αυτή: την εκτίμηση μέγιστης πιθανοφάνειας (κλασική προσέγγισή της) και τη μπεϋζιανή συμπερασματολογία. Στο επόμενο κεφάλαιο θα δούμε πως μπορούμε να πραγματοποιήσουμε στατιστική συμπερασματολογία μέσω της συνάρτησης πιθανοφάνειας για τα μοντέλα γραμμικής και ποσοστημοριακής παλινδρόμησης.

3.1 Συνάρτηση πιθανοφάνειας

Συνάρτηση πιθανοφάνειας (likelihood function) των διακριτών ή συνεχών τυχαίων μεταβλητών $X_1, X_2, \ldots, X_n$ είναι η από κοινού συνάρτηση πιθανότητας

$$L(\theta) = L(\theta; x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n; \theta)$$

όπου θ είναι η άγνωστη παράμετρος που θέλουμε να εκτιμήσουμε. Η συνάρτηση πιθανοφάνειας δίνει τη πιθανότητα να παρατηρήσουμε τις συγκεκριμένες τιμές του δείγματος δεδομένου ότι η άγνωστη παράμετρος παίρνει την τιμή θ. Ειδικά αν $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από συνάρτηση πιθανότητας $f(x; \theta)$ τότε

$$L(\theta) = L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta).$$

Έστω ότι θέλουμε να βρούμε τη συνάρτηση πιθανοφάνειας από ένα τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από την κατανομή Bernoulli(θ). Γνωρίζουμε ότι

$$f(x; \theta) = \theta^x (1-\theta)^{1-x} \quad x = 0, 1 \quad 0 < \theta < 1.$$

Άρα έχουμε

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(x_i; \theta) \\ &= \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \\ &= \theta^{\sum x_i} (1-\theta)^{n-\sum x_i}. \end{aligned}$$

Έστω τώρα ότι έχουμε ένα τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από την κανονική κατανομή $N(\mu, \sigma^2)$ όπου μ είναι η μέση τιμή και η σ^2 η διασπορά της κανονικής κατανομής). Άρα $\theta = (\mu, \sigma^2)$ και γνωρίζουμε ότι

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} \quad \mu \in \mathbb{R}, \sigma^2 \geq 0$$

Τώρα θα έχουμε

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\} \\ &= (2\pi\sigma^2)^{-n/2} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\}. \end{aligned}$$

Μπορούμε να παρατηρήσουμε ότι η συνάρτηση πιθανοφάνειας είναι συνάρτηση του τυχαίου δείγματος $X_i,\ i = 1, \ldots, n$ και της άγνωστης παραμέτρου θ αλλά επειδή οι τιμές του δείγματος είναι ήδη γνωστές και σταθερές, εμείς ενδιαφερόμαστε μόνο για το πως μεταβάλλεται η πιθανοφάνεια σε σχέση με την άγνωστη παράμετρο θ.

3.2 Εκτίμηση μέγιστης πιθανοφάνειας

Από τον ορισμό της συνάρτησης πιθανοφάνειας, είναι λογικό μια καλή εκτίμηση της παραμέτρου θ να είναι η τιμή $\hat{\theta}$ η οποία ικανοποιεί τη σχέση

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta).$$

Η τιμή $\hat{\theta}$ είναι αυτή η οποία μεγιστοποιεί τη συνάρτηση πιθανοφάνειας, δηλαδή είναι η πιο πιθανή τιμή της παραμέτρου για να έχουμε πάρει το συγκεκριμένο δείγμα. Αυτή η τιμή καλείται εκτιμήτρια μέγιστης πιθανοφάνειας. Μια σημαντική παρατήρηση που μπορούμε να κάνουμε εδώ και η οποία χρησιμοποιείται αρκετά για την εύρεση της εκτιμήτριας μέγιστης πιθανοφάνειας είναι ότι η τιμή $\hat{\theta}$ μεγιστοποιεί επίσης και την συνάρτηση

$$l(\theta) = \log L(\theta)$$

η οποία ονομάζεται λογαριθμική πιθανοφάνεια.

Ας βρούμε αρχικά την εκτιμήτρια μέγιστης πιθανοφάνειας για το πρώτο παράδειγμά μας στο οποίο το τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ ακολουθούσε την κατανομή Bernoulli(θ).

Έχουμε ήδη βρει ότι $L(\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$. Για να βρούμε τωρα για ποια τιμή του θ μεγιστοποιείται η $L(\theta)$, σύμφωνα με τα προηγούμενα αρκεί να βρούμε που μεγιστοποιείται η $l(\theta) = \log L(\theta)$.

Άρα έχουμε

$$\begin{aligned} l(\theta) &= \log L(\theta) \\ &= \sum x_i \ln \theta + (n - \sum x_i) \ln(1-\theta). \end{aligned}$$

Για να βρούμε αρχικά το ακρότατο της γνωρίζουμε ότι αυτό θα είναι το σημείο στο οποίο ισχύει η σχέση

$$\frac{\partial l(\theta)}{\partial \theta} = 0.$$

Άρα

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{\sum x_i}{\theta} - \frac{n - \sum x_i}{1-\theta} = 0$$

$$\sum x_i - \theta \sum x_i = n\theta - \theta \sum x_i$$
$$n\theta = \sum x_i \Rightarrow \hat{\theta} = \frac{\sum x_i}{n}.$$

Για να είναι τώρα η $\hat{\theta} = \bar{x}$ η τιμή που μεγιστοποιεί την $l(\theta)$ αρκεί να ισχύει

$$\frac{\partial^2 l(\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}} < 0$$

$$\begin{aligned}
\frac{\partial^2 l(\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}} &= -\frac{\sum x_i}{\hat{\theta}^2} - \frac{n - \sum x_i}{(1-\hat{\theta})^2} \\
&= -\frac{\sum x_i}{\bar{x}^2} - \frac{n - \sum x_i}{(1-\bar{x})^2} \\
&= -\frac{\sum x_i}{(\frac{\sum x_i}{n})^2} - \frac{n - \sum x_i}{(1-(\frac{\sum x_i}{n})^2} \\
&= \dots \\
&= -n^2(\frac{1}{\sum x_i} + \frac{1}{n - \sum x_i}) < 0.
\end{aligned}$$

Άρα η εκτιμήτρια μέγιστης πιθανοφάνειας (ε.μ.π.) τυχαίου δείγματος $X_1, \dots, X_n$ της κατανομής Bernoulli(θ) είναι η $\hat{\theta} = \bar{x}$.

Πριν ασχοληθούμε με το δεύτερο παράδειγμά μας θα δώσουμε μερικούς ορισμούς οι οποίοι θα είναι σημαντικοί για τη συμπερασματολογία μέσω πιθανοφάνειας. Εδώ πρέπει να τονίσουμε ότι η εκτιμήτρια μέγιστης πιθανοφάνειας δεν είναι αρκετή για να εξάγουμε συμπεράσματα για την άγνωστη παράμετρο αλλά χρειάζεται ολόκληρη η συνάρτηση πιθανοφάνειας για τη στατιστική συμπερασματολογία.

Κανονικά προβλήματα ονομάζονται εκείνα στα οποία η $l(\theta)$ γύρω από την εκτιμήτρια μέγιστης πιθανοφάνειας μπορεί να εκτιμηθεί από μία τετραγωνική συνάρτηση, έχοντας ως αποτέλεσμα να αναπαρίσταται από την $\hat{\theta}$ και τη καμπυλότητα της στη $\hat{\theta}$. Σε τέτοιες περιπτώσεις η λογαριθμική πιθανοφάνεια $l(\theta)$ ονομάζεται κανονική. Η καμπυλότητα της εκτιμήτριας μέγιστης πιθανοφάνειας, $\hat{\theta}$, είναι η $I(\hat{\theta})$ όπου

$$I(\theta) = -\frac{\partial^2}{\partial \theta^2} l(\theta)$$

ονομάζεται παρατηρούμενη πληροφορία Fisher για τη θ. Η αναμενόμενη πληροφορία Fisher βρίσκεται από τον τύπο

$$\mathbf{I}(\theta) = -E\Big[\frac{\partial^2}{\partial \theta^2} l(\theta)\Big].$$

Μία μεγάλη καμπυλότητα $I(\hat{\theta})$, σημαίνει ότι υπάρχει μία στενή κορυφή στη λογαριθμική πιθανοφάνεια κάτι το οποίο μας κάνει να είμαστε πιο σίγουροι για την εκτίμηση μας.

Ως τώρα έχουμε ασχοληθεί με παραδείγματα τα οποία είχαν μία άγνωστη παράμετρο θ και στα οποία κάτι που είναι αρκετά σημαντικό να γνωρίζουμε είναι το τυπικό σφάλμα της εκτίμησής μας $\hat{\theta}$. Για κανονικές $l(\theta)$, η εύρεση του τυπικού σφάλματος της εκτίμησής μας στηρίζεται στο τύπο

$$se(\hat{\theta}) = \mathbf{I}^{-1/2}(\hat{\theta}),$$

όπου θ είναι μονόμετρο μέγεθος.

Στο δεύτερο παράδειγμα τώρα, είχαμε τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από την κανονική κατανομή $N(\mu, \sigma^2)$ και ας υποθέσουμε ότι η διασπορά της κανονικής κατανομής είναι γνωστή και η άγνωστη παράμετρος είναι η μέση τιμή,δηλ. $\theta = \mu$. Είχαμε βρει ότι

$$L(\theta) = (2\pi\sigma^2)^{-n/2} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\}$$

Η διαδικασία που ακολουθούμε για να βρούμε την εκτιμήτρια $\hat{\theta}$ για τη μέση τιμή θ η οποία να μεγιστοποιεί την πιθανοφάνεια, είναι ανάλογη με πριν απλά τώρα θέλουμε να ισχύει

$$\frac{\partial l(\theta)}{\partial \theta} = 0,$$

έτσι ώστε να βρούμε που παρουσιάζει ακρότατο η συνάρτηση πιθανοφάνειας. Έπειτα για να εξετάσουμε αν αυτό μεγιστοποιεί τη πιθανοφάνεια, θέλουμε να δείξουμε ότι

$$\frac{\partial^2 l(\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}} < 0$$

Με την ίδια λογική όπως πριν έχουμε ότι

$$\hat{\theta} = \bar{x}$$

και

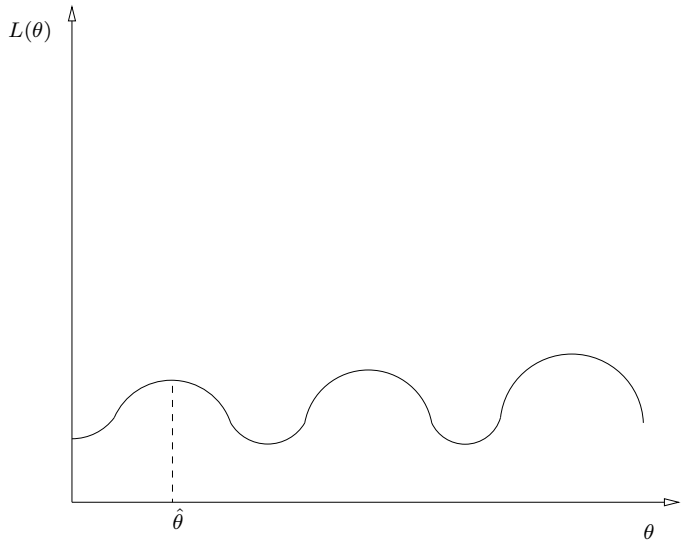
$$\frac{\partial^2 l(\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}} = -\frac{n}{\sigma^2} < 0.$$

Ακόμα, βρίσκουμε ότι

$$I(\hat{\theta}) = \frac{n}{\sigma^2}$$

και ότι το τυπικό σφάλμα της $\hat{\theta}$ είναι

$$se(\hat{\theta}) = \frac{\sigma}{\sqrt{n}}.$$



Σχήμα 3.1: Διάγραμμα πιθανοφάνειας και άγνωστης παραμέτρου

Πριν συνεχίσουμε, πρέπει να τονίσουμε ότι σε μία οποιαδήποτε συνάρτηση η ικανοποίηση της ανίσωσης $\frac{\partial^2 l(\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}} < 0$ δεν προυποθέτει ότι στο σημείο $\hat{\theta}$ βρίσκεται το ολικό μέγιστο της συνάρτησης αυτής, αλλά υπάρχει και η περίπτωση να δώσει τοπικό μέγιστο όπως στη συνάρτηση που παρουσιάζεται στο σχήμα 3.1. Όμως επειδή τα παραπάνω αφορούν κανονικές πιθανοφάνειες δεν πρόκειται να συναντήσουμε ποτέ συναρτήσεις τέτοιας μορφής.

3.3 Ιδιότητες της εκτιμήτριας μέγιστης πιθανοφάνειας

Ο λόγος που ασχολούμαστε με τις εκτιμήτριες μέγιστης πιθανοφάνειας είναι ότι αυτές έχουν αρκετές επιθυμητές ιδιότητες οι οποίες έχουν ως αποτέλεσμα οι παραγόμενες εκτιμήσεις να είναι κοντά στην πραγματικότητα και ρεαλιστικές κάτι το οποίο είναι ο στόχος οποιασδήποτε μεθόδου στατιστικής συμπερασματολογίας.

I Οι εκτιμήτριες μέγιστης πιθανοφάνειας έχουν την ιδιότητα του Αναλλοίωτου (Invariance). Αν $\hat{\theta}$ είναι η εκτιμήτρια μέγιστης πιθανοφάνειας της παραμέτρου της θ και $u(\theta)$ μία συνάρτηση της θ με μονότιμη αντίστροφη συνάρτηση, τότε η εκτιμήτρια μέγιστης πιθανοφάνειας της $u(\theta)$ είναι η $u(\hat{\theta})$.

II Η ε.μ.π. $\hat{\theta}$ είναι ασυμπτωτικά συνεπής εκτιμητής της παράμετρου θ, δηλαδή ισχύει $\hat{\theta} \to \theta$ για $n \to \infty$.

III Η ε.μ.π. $\hat{\theta}$ ακολουθεί ασυμπτωτικά τη κανονική κατανομή με μέση τιμή θ και διασπορά $I^{-1}(\theta)$.

IV Η ε.μ.π. $\hat{\theta}$ είναι από όλες τις αμερόληπτες εκτιμήτριες της θ αυτή με την ελάχιστη διασπορά

3.4 Διαστήματα εμπιστοσύνης βασισμένα στη πιθανοφάνεια

Η γνώση της εκτιμήτριας μέγιστης πιθανοφάνειας και του τυπικού της σφάλματος δεν είναι αρκετή για τη στατιστική συμπερασματολογία. Ο Fisher ήταν ο πρώτος που σκέφτηκε από τη παρατηρούμενη συνάρτηση πιθανοφάνειας να βγάζουμε συμπεράσματα για την αβεβαιοτητά μας για την άγνωστη παράμετρο θ μέσω διαστημάτων εμπιστοσύνης. Ένα διάστημα βασισμένο στη πιθανοφάνεια ορίζεται ως το σύνολο των τιμών που μπορεί να πάρει η παράμετρος θ οι οποίες έχουν πιθανοφάνεια $L(\theta)$ τέτοια ώστε $\frac{L(\theta)}{L(\hat{\theta})} > C$ για κάποιο σημείο αποκοπής (cutoff point) C. Τώρα, αν ορίσουμε ως $D(\theta)$ την τ.μ.

$$D(\theta) = 2(l(\hat{\theta}) - l(\theta))$$

έχουμε ότι η $D(\theta)$ ακολουθεί τη κατανομή $\mathbf{X}^2$ με ένα βαθμό ελευθερίας δηλαδή

$$D(\theta) \sim \mathbf{X}_1^2.$$

Για μία άγνωστη παράμετρο θ, έχουμε ότι

$$\begin{aligned}
\mathbf{Pr}(\frac{L(\theta)}{L(\hat{\theta})} > C) &= \mathbf{Pr}(\log \frac{L(\hat{\theta})}{L(\theta)} > logC) \\
&= \mathbf{Pr}(\log \frac{L(\hat{\theta})}{L(\theta)} < -logC) \\
&= \mathbf{Pr}(2 \log \frac{L(\hat{\theta})}{L(\theta)} < -2logC) \\
&= \mathbf{Pr}(D(\theta) < 2c)
\end{aligned}$$

όπου $c = -\log(C)$.

Άρα για $0 < \alpha < 1$ αν διαλέξουμε $c = -\frac{1}{2}\mathbf{X}^2_{1,(1-\alpha)}$ όπου $\mathbf{X}^2_{1,(1-\alpha)}$ είναι το $100(1-\alpha)$ εκατοστημόριο της $\mathbf{X}^2_1$ κατανομής, έχουμε

$$\mathbf{Pr}(\frac{L(\theta)}{L(\hat{\theta})} > C) = \mathbf{Pr}(\mathbf{X}^2_1 < \mathbf{X}^2_{1,(1-\alpha)}) = 1 - \alpha$$

Αυτό σημαίνει ότι αν επιλέξουμε το προαναφερθέν c τότε το βασισμένο στη πιθανοφάνεια διάστημα

$$\{\theta, \frac{L(\theta)}{L(\hat{\theta})} > C\}$$

είναι ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την άγνωστη παράμετρο θ.

Επειδή η επιλογή του σημείου αποκοπής c είναι ένα ζήτημα που παραμένει ακόμα ανοιχτό αν και ο Fisher θεωρούσε ότι τιμές της παραμέτρου με λιγότερο από 6.7% πιθανοφάνεια είναι λογικό να θεωρούνται «ύποπτες», μπορούμε να πούμε ότι για $\alpha = 0.05$ και 0.01 η τιμή του c θα είναι 0.15 και 0.04 αντίστοιχα. Άρα αν θελουμε να βρούμε ένα 95% ή 99% διάστημα εμπιστοσύνης για τη μέση τιμή μιας κανονικής κατανομής επιλέγουμε σημείο αποκοπής 15% ή 4% αντίστοιχα.

Έστω ότι οι τ.μ. $X_1, X_2, \ldots, X_{10}$ παριστάνουν τα αποτελέσματα 10 ανεξάρτητων ρίψεων νομίσματος, όπου αν $X_i = 1$ το αποτέλεσμα της i ρίψης ήταν κορώνα ενώ αν $X_i = 0$ ήταν γράμματα. Μας ενδιαφέρει, δεδομένου ότι παρατηρήσαμε $X = \sum_{i=1}^{10} X_i = 8$, να εκτιμήσουμε την άγνωστη παράμετρο θ, η οποία δείχνει την πιθανότητα να έρθει κορώνα, και να βρούμε διαστήματα στα οποία εκτιμούμε ότι θα βρίσκεται η θ βασισμένα στη πιθανοφάνεια. Η τ.μ. X ακολουθεί τη Διωνυμική $B(n, \theta)$ κατανομή για $n = 10$, άρα έχουμε

$$f(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x = 1, 2, \ldots, 10, \quad 0 < \theta < 1$$

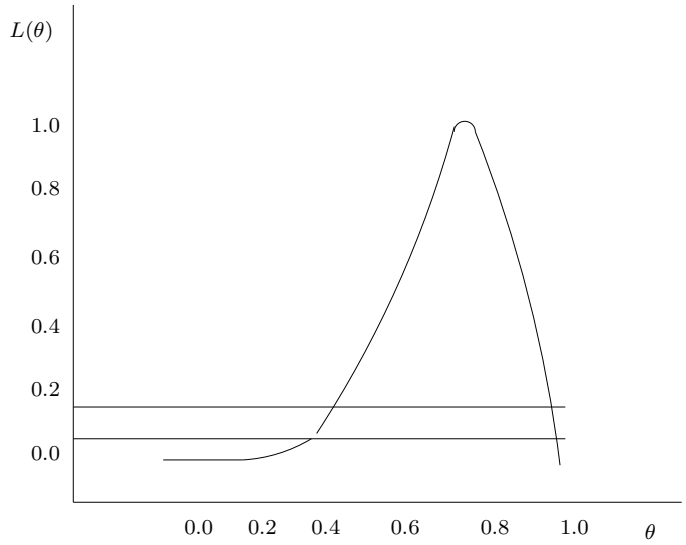
όπου $x = 8$ οι ρίψεις που το αποτέλεσμα τους ήταν γράμματα. Μπορούμε να δούμε ότι

$$l(\theta) = x \log\theta + (n-x)\log(1-\theta)$$

και

$$\hat{\theta} = \frac{x}{n} = 0.8 \ .$$

Στο σχήμα 3.2 βλέπουμε πως μεταβάλλεται η πιθανοφάνεια ανάλογα με τη παράμετρο θ. Οι δύο ευθείες παριστάνουν τα σημεια αποκοπής $C = 0.04$ και $C = 0.15$. Παρατηρούμε ότι η συνάρτηση πιθανοφάνειας μεγιστοποιείται για $\theta = 0.8$ και ότι το διάστημα που είναι βασισμένο στη πιθανοφάνεια για $C = 0.04$ είναι το $(0.41, 0.98)$ ενώ αν $c = 0.15$ τότε το διάστημα είναι $(0.50, 0.96)$. Το πρώτο είναι το 99% διάστημα εμπιστοσύνης για την θ και το δεύτερο το 95%.



Σχήμα 3.2: Διαστήματα βασισμένα στη πιθανοφάνεια για 15% και 4% σημεία αποκοπής

Εδώ σημαντικό είναι να τονίσουμε ένα λάθος που γίνεται συνήθως όταν αναφερόμαστε σε διαστήματα εμπιστοσύνης. Όταν μιλάμε για ένα 95% διάστημα εμπιστοσύνης μιας άγνωστης παραμέτρου δεν εννοούμε ότι υπάρχει πιθανότητα 95% το διάστημα αυτό να περιέχει την άγνωστη παράμετρο αλλά ότι αν επαναλάβουμε τη διαδικασία αρκετές φορές, η παράμετρος θα ανήκει στο 95% των διαστημάτων που θα βρούμε.

3.5 Διανυσματικές παράμετροι και προφίλ πιθανοφάνεια

Ως τώρα σε όλα τα παραδείγματα που είδαμε και βρίσκαμε αρχικά τη πιθανοφάνεια, η άγνωστη παράμετρος ήταν μία ή αλλιώς η άγνωστη παράμετρος θ ήταν μονόμετρο μέγεθος. Τώρα θα δούμε τι συμβαίνει όταν έχουμε περισσότερες από μία άγνωστες παραμέτρους ή αλλιώς όταν η θ είναι διανυσματικό μέγεθος δηλαδή όταν $\theta = (\theta_1, \theta_2, \ldots, \theta_s)$ $i \geq 2$. Τα προβλήματα αυτά είναι συνήθως πιο πολύπλοκα από πριν. Έστω $\theta = (\theta_1, \theta_2, \ldots, \theta_s)$ είναι η s-διάστατη άγνωστη παράμετρος.

Η πληροφορία Fisher εδώ είναι ο πίνακας που περιέχει τις δεύτερες μερικές παραγώγους της λογαριθμικής πιθανοφάνειας υπολογισμένος στη τιμή της

εκτιμήτριας μέγιστης πιθανοφάνειας. Τα στοιχεία του δίνονται από τη σχέση

$$I_{ij}(\hat{\theta}) = -\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\theta)|_{\theta=\hat{\theta}} \quad i,j = 1,2,\ldots,s.$$

Αντίστοιχα ο πίνακας της αναμενόμενης πληροφορίας Fisher βρίσκεται από τον τύπο

$$\mathbf{I}_{ij}(\hat{\theta}) = -E\Big[\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\theta)\Big] \quad i,j = 1,2,\ldots,s.$$

Σε κανονικά προβλήματα, η $l(\theta)$ μπορεί να αναπαρίσταται από την ε.μ.π. θ και τον πίνακα $I(\hat{\theta})$.

Έστω ότι έχουμε ένα τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από την κανονική κατανομή $N(\mu, \sigma^2)$ όπου τώρα έχουμε δύο άγνωστους παραμέτρους (τη μέση τιμή και τη διασπορά της κανονικής κατανομής). Άρα $\theta = (\mu, \sigma^2)$ και γνωρίζουμε ότι

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}, \quad \mu \in \mathbb{R}, \sigma^2 \geq 0.$$

Έχουμε ότι

$$L(\theta) = L(\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\}.$$

Η διαδικασία που ακολουθούμε για να βρούμε εκτιμήσεις $\hat{\theta_1}, \hat{\theta_2}$ για τη μέση τιμή $\mu = \theta_1$ και τη διασπορά $\sigma^2 = \theta_2$ οι οποίες να μεγιστοποιούν την πιθανοφάνεια είναι ανάλογη με πριν απλά τώρα θέλουμε να ισχύουν οι σχέσεις

$$\frac{\partial l(\theta_1,\theta_2)}{\partial\theta_1} = 0$$

και

$$\frac{\partial l(\theta_1,\theta_2)}{\partial\theta_2} = 0$$

έτσι ώστε να βρούμε που παρουσιάζει ακρότατα η συνάρτηση πιθανοφάνειας και ύστερα, για να εξετάσουμε αν αυτά μεγιστοποιούν τη πιθανοφάνεια θέλουμε να ελέγξουμε αν

$$\frac{\partial^2 l(\theta_1,\theta_2)}{\partial\theta_i^2}|_{\theta_i=\hat{\theta_i}} < 0 \quad i = 1,2.$$

Με την ίδια λογική όπως πριν έχουμε ότι

$$\hat{\theta_1} = \bar{x}, \quad \hat{\theta_2} = \frac{1}{\sum(x_i-\bar{x})^2}.$$

Οι εκτιμήτριες μεγίστης πιθανοφάνειας λοιπόν για τη μέση τιμή μ και τη διασπορά σ^2 ενός τυχαίου δείγματος που ακολουθεί κανονική κατανομή είναι απλά η δειγματική μέση τιμή και η μεροληπτική δειγματική διασπορά αντίστοιχα, η οποία μπορούμε να προσθέσουμε ότι είναι ασυμπτωτικά αμερόληπτη.

Ο πίνακας πληροφορίας Fisher είναι

$$I(\hat{\theta}) = \begin{pmatrix} n/\hat{\sigma}^2 & 0 \\ 0 & n/(2\hat{\sigma}^4) \end{pmatrix}$$

Σε προβλήματα που έχουμε διανυσματική άγνωστη παράμετρο θ η προφίλ πιθανοφάνεια κάθε μίας παραμέτρου ξεχωριστά είναι ίση με τη μέγιστη τιμή της συνάρτησης πιθανοφάνειας, κρατώντας τη συγκεκριμένη παράμετρο σταθερή αλλά βάζοντας ως τιμές όλων των υπόλοιπων παραμέτρων τις εκτιμήτριες μέγιστης πιθανοφάνειας τους. Μαθηματικά, η προφίλ πιθανοφάνεια της παραμέτρου $PL(\theta_i)$, όταν η άγνωστη παράμετρός μας είναι η $\theta = (\theta_i, \xi)$, ορίζεται ως

$$PL(\theta_i) = \max_{\xi} L(\theta_i, \xi) = L(\theta_i, \hat{\xi})$$

Στο προηγούμενο παράδειγμα μας

$$PL(\sigma^2) = (2\pi\sigma^2)^{-n/2} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\}$$

και

$$PL(\mu) = \Big(\frac{2\pi}{n}\sum_{i=1}^{n}(x_i - \mu)^2\Big)^{-n/2} exp\{-\frac{n}{2}\}$$

οι προφίλ πιθανοφάνειες των άγνωστων παραμέτρων.

3.6 Η μέθοδος Bootstrap για την εύρεση τυπικών σφαλμάτων

Η μέθοδος Bootstrap παρουσιάστηκε από τον Effron (1979) ως μία μέθοδος η οποία βοηθάει στην εύρεση των κατανομών των εκτιμητριών των αγνώστων παραμέτρων και επείδη είναι μία μέθοδος η οποία στηρίζεται στις υπολογιστικές δυνάμεις του υπολογιστή, χρόνο με το χρόνο γινόταν όλο και πιο διαδεδομένη. Επίσης η μέθοδος Bootstrap όπως θα δούμε παρακάτω μας βοηθάει στην εύρεση τυπικών σφαλμάτων για εκτιμήτριες που δεν ξέρουμε τι κατανομή ακολουθούν.

Ας υποθέσουμε ότι έχουμε $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από κατανομή F και έχουμε παρατηρήσει $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. Έστω $\hat{\theta}$ η εκτιμήτρια της μέσης τιμής του παραπάνω δείγματος.

Ας ασχοληθούμε αρχικά με την (πιο πιθανή) περίπτωση να μην γνωρίζουμε την κατανομή F. Η ιδέα της μη-παραμετρικής μεθόδου Bootstrap βασίζεται στο ότι αρχικά θα προσομοιώσουμε μέσω υπολογιστή ένα μεγάλο πλήθος, έστω B, δειγμάτων μεγέθους n από μία εκτιμήτρια της F, την εμπειρική συνάρτηση κατανομής F_n του δείγματος που έχουμε παρατηρήσει. Η εμπειρική συνάρτηση κατανομής F_n ενός δείγματος $(x_1, x_2, \ldots, x_n)$ είναι μία διακριτή συνάρτηση κατανομής που δίνει πιθανότητα $1/n$ σε κάθε παρατηρούμενη τιμή $x_1, x_2, \ldots, x_n$ και ορίζεται ως

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

όπου $I(A)$ είναι η δείκτρια συνάρτηση. Ένα δείγμα μεγέθους n από την F_n είναι ένα δείγμα με επανατοποθέτηση από το αρχικό τυχαίο δείγμα $(x_1, x_2, \ldots, x_n)$. Από το κάθε ένα από τα B δείγματα που έχουν προσομοιωθεί μπορούμε να βρούμε τη μέση τιμή του $\hat{\theta_i}^*$, $i = 1, 2, \ldots, B$. Τώρα έχουμε B τιμές οι οποίες, αν το B είναι αρκετά μεγάλο, είναι αρκετά καλές προσεγγίσεις της κατανομής της $\hat{\theta}$. Το τυπικό σφάλμα της $\hat{\theta}$ μπορούμε να το βρούμε από τον τύπο

$$s_{\hat{\theta}} = \sqrt{\frac{1}{B}\sum_{i=1}^{B}(\hat{\theta_i}^* - \bar{\theta}^*)^2}$$

όπου

$$\bar{\theta}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta_i}^*.$$

Στη περίπτωση που η κατανομή της F είναι γνωστή τότε κάνουμε προσομοίωση απευθείας από την κατανομή αυτή και συνεχίζουμε οπως προηγουμένως.

Ας δούμε και ένα παράδειγμα όπου $n = 2$ και $X_1 = c < X_2 = d$ τυχαίο δείγμα και θέλουμε να εκτιμήσουμε τη μέση τιμή. Εδώ έχουμε ότι X_1^*, X_2^* πιθανά αποτελέσματα με

$$\mathbf{Pr}(X_i^* = c) = \mathbf{Pr}(X_i^* = d) = \frac{1}{2}, \quad i = 1, 2.$$

Για $n = 2$ έχουμε 2^2 πιθανά ζευγάρια

$$(c, c), (c, d), (d, c), (d, d)$$

κάθε ένα με πιθανότητα $\frac{1}{4}$. Άρα έχουμε ότι

$$\hat{\theta_i}^* = \frac{X_1^* + X_2^*}{2} = \begin{cases} c & p = \frac{1}{4} \\ \frac{c+d}{2} & p = \frac{1}{2} \\ d & p = \frac{1}{4} \end{cases} \qquad i = 1, 2, \ldots, B$$

όπου $\hat{\theta_i}^*$ είναι η μέση τιμή του i δείγματος που έχουμε από τη προσομοίωση στον υπολογιστή και B είναι το πλήθος των δειγμάτων μεγέθους 2 που θα προσομοιωθούν. Στη συνέχεια μπορούμε να βρούμε και το $\bar{\theta}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta_i}^*$ το οποίο είναι απαραίτητο για την εύρεση του τυπικού σφάλματος της εκτιμήτριας της μέσης τιμής.

Με παρόμοιο τρόπο όπως αυτός των τυπικών σφαλμάτων μπορούμε να βρούμε τη εκτιμήσεις Bootstrap για τη συνδιακύμανση δύο άγνωστων παραμέτρων. Αν υποθέσουμε ότι $\hat{\theta_1}^*$ και $\hat{\theta_2}^*$ είναι δύο άγνωστες παράμετροι (θα μπορούσαν να είναι η μέση τιμή και η διασπορά της κανονικής κατανομής) έχουμε ότι

$$Cov_B(\hat{\theta_1}^*, \hat{\theta_2}^*) = \frac{1}{B}\sum_{i=1}^{B}(\hat{\theta_{1i}}^* - \hat{\theta_1}^*)(\hat{\theta_{i2}}^* - \hat{\theta_2}^*)$$

όπου $(\hat{\theta_{1i}}^*, \hat{\theta_{i2}}^*)$ είναι οι τιμές Bootstrap των άγνωστων παραμέτρων από το i δείγμα. Στην ουσία η μέθοδος Bootstrap μπορεί να χρησιμοποιηθεί για την εκτίμηση του πίνακα συνδιακύμανσης διανύσματος παραμέτρων. Συνοψίζοντας, μπορούμε να δούμε ότι η μέθοδος Bootstrap για να εφαρμοστεί απλά υποθέτει ότι το δείγμα που έχουμε είναι αντιπροσωπευτικό του πληθυσμού. Ισχύει επίσης ότι είναι υπολογιστικά απαιτητική (για δείγματα μεγέθους n έχουμε n^n πιθανά αποτελέσματα) αλλά με τη συνεχή πρόοδο των υπολογιστών μπορεί να χρησιμοποιείται άφοβα από τον κάθε ένα.

Για περισσότερες πληροφοριές γύρω από την μέθοδο Bootstrap μπορεί κάποιος να διαβάσει τις σημειώσεις των Καρλή[;], Μελιγκοτσίδου[;], Hung Chen[?].

3.7 Μπεϋζιανή συμπερασματολογία

Στη μπεϋζιανή στατιστική κάθε άγνωστη παράμετρος θεωρείται τυχαία μεταβλητή αντιθέτως με αυτά που έχουμε δει ως τώρα όπου όλες οι άγνωστες παράμετροι του μοντέλου μας ήταν σταθερές. Αυτή η καινούργια αντίληψη των πραγμάτων έχει ως αποτέλεσμα κάθε άγνωστη παράμετρος $\theta \in \Theta$, όπου Θ το πεδίο τιμών της, να έχει μία πυκνότητα πιθανότητας $\pi(\theta)$ η οποία έχει τις παρακάτω ιδιότητες

- $\pi(\theta) \geq 0, \forall\theta \in \Theta$.
- $\int_\Theta \pi(\theta)d\theta = 1$ ή $\sum_\Theta \pi(\theta) = 1$.

Η συνάρτηση $\pi(\theta)$ ονομάζεται εκ των προτέρων κατανομή (prior distribution) της θ και εκφράζει την προσωπική μας αντίληψη για την άγνωστη παράμετρο θ ή συνοψίζει κάποιες εκ των προτέρων (πριν τη συλλογή των δεδομένων)

πληροφορίες για την θ. Τώρα, αφότου έχουμε παρατηρήσει τα δεδομένα $x = (x_1, x_2, \ldots, x_n)$ η γνώση μας για την παράμετρο θ θα «ανανεωθεί» μέσω της συνάρτησης πιθανοφάνειας $L(x|\theta) = L(\theta)$. Τελικά, σύμφωνα με τη μπεϋζιανή προσέγγιση οποιαδήποτε συμπερασματολογία για την άγνωστη παράμετρο θ γίνεται μέσω της εκ των υστέρων κατανομής της (posterior distribution) $p(\theta|x)$ η οποία δίνεται από το Bayes Θεώρημα μέσω του τύπου

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int_\Theta L(x|\theta)\pi(\theta)d\theta}, \quad \theta \in \Theta.$$

Ως αποτέλεσμα του θεωρήματος του Bayes έχουμε ότι η εκ των υστέρων κατανομή της παραμέτρου θ είναι ανάλογη του γινομένου της πιθανοφάνειας και της εκ των προτέρων κατανομής της θ άρα έχουμε

$$p(\theta|x) \propto L(x|\theta)\pi(\theta).$$

Ας υποθέσουμε ότι έχουμε ένα τυχαίο δείγμα $x = (x_1, x_2, \ldots, x_n)$ μιας τυχαίας μεταβλητής X που ακολουθεί την Poisson(θ) κατανομή, έτσι ώστε

$$f(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}, \quad \theta \geq 0.$$

Γνωρίζουμε ήδη ότι $E(X) = \theta$ και $V(X) = \theta$.

Η πιθανοφάνεια είναι

$$\begin{aligned} L(x|\theta) &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \\ &\propto e^{-n\theta}\theta^{\sum x_i} \end{aligned}$$

Αν υποθέσουμε ότι $\theta \sim Gamma(r, q)$, έχουμε

$$\pi(\theta) = \frac{q^r}{\Gamma(r)}\theta^{r-1}e\{-q\theta\}, \quad \theta > 0,$$

$$E(\theta) = \frac{r}{q} \quad V(\theta) = \frac{r}{q^2}.$$

Σύμφωνα με το θεώρημα Bayes θα έχουμε

$$\begin{aligned} p(\theta|x) &\propto \theta^{r-1}exp\{-q\theta\} \times exp\{-n\theta\}\theta^{\sum x_i} \\ &= \theta^{(r+\sum x_i-1)}exp\{-(q+n)\theta\} \\ &= \theta^{R-1}exp\{-Q\theta\} \end{aligned}$$

όπου $R = r + \sum x_i$ και $Q = q + n$. Υπάρχει μόνο μία συνάρτηση πυκνότητας η οποία είναι ανάλογη σε αυτό που έχουμε βρει, έτσι έχουμε

$$\theta|x \sim Gamma(R, Q)$$

Άρα από εκεί που είχαμε μια αρχική εκτίμηση για τη θ, με τη βοήθεια της πιθανοφάνειας φτάσαμε στο σημείο να γνωρίζουμε την εκ των υστέρων κατανομή της θ. Από την εκ των υστέρων κατανομή $p(\theta|x)$ μπορούμε επίσης να πάρουμε μια «βέλτιστη» σημειακή εκτίμηση της θ που να συνοψίζει τη πληροφορία της εκ των υστέρων κατανομής όπως τη μέση τιμή

$$E_p(\theta|x) = \frac{\int_\Theta \theta L(x|\theta)\pi(\theta)d\theta}{\int_\Theta L(x|\theta)\pi(\theta)d\theta}$$

ή ακόμα και τη κορυφή της εκ των υστέρων κατανομής.

3.8 Διαστήματα αξιοπιστίας

Τα διαστήματα αξιοπιστίας (Credibility intervals) είναι το αντίστοιχο των διαστημάτων εμπιστοσύνης της κλασικής στατιστικής. Όμως όπως τονίσαμε και πριν, στη κλασική στατιστική τα διαστήματα εμπιστοσύνης δεν είναι πιθανοθεωρητικά ενώ εδώ που η άγνωστη παράμετρος είναι τυχαία μεταβλητή το διάστημα $C_\alpha(x)$ είναι ένα $100(1-\alpha)\%$ διάστημα αξιοπιστίας αν

$$\int_{C_\alpha(x)} p(\theta|x)d\theta = 1 - \alpha.$$

Η παραπάνω ισότητα μας λέει ότι με εκ των υστέρων πιθανότητα $1-\alpha$ η παράμετρος θ περιέχεται μέσα στο $C_\alpha(x)$. Όταν η θ είναι διανυσματικό μέγεθος τότε έχουμε περιοχή αξιοπιστίας. Τα διαστήματα αξιοπιστίας (όπως και τα διαστήματα εμπιστοσύνης) δεν είναι μοναδικά. Οποιαδήποτε περιοχή του παραμετρικού χώρου με εκ των υστέρων πιθανότητα $1-\alpha$ είναι $100(1-\alpha)\%$ περιοχή αξιοπιστίας. Αυτό που θα θέλαμε είναι η περιοχή αξιοπιστίας να περιέχει τις πιο πιθανές τιμές της παραμέτρου και να είναι η πιο μικρή περιοχή με πιθανότητα $1-\alpha$.

Τα παραπάνω παριστάνονται από ένα διάστημα (ή περιοχή) της μορφής

$$C_\alpha(x) = \{\theta : p(\theta|x) \geq \gamma\}$$

όπου γ είναι τέτοιο ώστε να ικανοποιεί

$$\int_{C_\alpha(x)} p(\theta|x)d\theta = 1 - \alpha.$$

Τέτοια διαστήματα (ή περιοχές) ονομάζονται υψηλότερης πυκνότητας πιθανότητας (highest posterior density intervals/regions). Αν η εκ των υστέρων κατανομή είναι μονοκόρυφη τότε η περιοχή υψηλότερης πυκνότητας πιθανότητας θα είναι ένα διάστημα της μορφής (a, b). Μικρές τιμές του α δίνουν μεγάλα διαστήματα αξιοπιστίας ενώ μεγάλες τιμές του α δίνουν μικρά διαστήματα.

3.9 Επιλογή της εκ των προτέρων κατανομής

Όπως ήδη έχουμε δει, η μπεϋζιανή συμπερασματολογία βασίζεται πάνω στη γνώση της συνάρτησης πιθανοφάνειας $L(x|\theta)$ αλλά και της εκ των προτέρων κατανομής $\pi(\theta)$. Οι «ενστάσεις» που έχουν οι κλασικοί στατιστικοί βασίζονται συνήθως στο ότι τα συμπεράσματα που βγάζουμε εξαρτώνται από την επιλογή της εκ των προτέρων κατανομής της άγνωστης παραμέτρου θ. Έτσι, θεωρούν, ότι χάνουμε την αντικειμενικότητα μας και επιλέγουμε την εκ των προτέρων κατανομή που θα μας οδηγήσει στα επιθυμητά συμπεράσματα.

Όμως τις περισσότερες φορές έχουμε μία εκ των προτέρων πεποίθηση για την θ την οποία αντικατοπτρίζουμε μέσω της εκ των προτέρων κατανομής. Ακόμα και όταν δεν έχουμε καμία γνώση ή πληροφορία για την θ, μπορούμε να κατασκευάσουμε μία εκ των προτέρων κατανομή που να δείχνει αυτή την έλλειψη πληροφορίας. Χρησιμοποιώντας αυτή τη μη πληροφοριακή εκ των προτέρων κατανομή (uninformative prior) και δεδομένου ότι έχουμε αρκετά δεδομένα, η εκ των υστέρων κατανομή της θ θα μοιάζει στη πιθανοφάνεια και έτσι η συγκεκριμένη prior κατανομή που χρησιμοποιήσαμε δεν έχει σημαντική επίδραση στις εκτιμήσεις που θα κάνουμε στη συνέχεια. Συνοψίζοντας, παρακάτω υπάρχουν καποιές σημαντικές παρατηρήσεις που αφορούν την εκ των προτέρων κατανομή.

I Υποκειμενική ανάλυση λόγω προσωπικής επιλογής prior κατανομής.

II Η επιλογή της prior επηρεάζει όλο και λιγότερο τα αποτελέσματα όσο αυξάνονται τα δεδομένα (αν έχουμε λίγα δεδομένα η posterior κατανομή θα είναι σταθμισμένη προς την prior).

III Μπορεί να έχουμε μια ιδέα για την prior κατανομή (π.χ. τη μέση τιμή και τη διασπορά της), και τίποτα παραπάνω. Τότε επιλέγουμε μια συναρτησιακή μορφή της $\pi(\theta)$ που να μας διευκολύνει υπολογιστικά.

IV Όταν δεν έχουμε καμία εκ των προτέρων πληροφορία για τη θ, μπορούμε να χρησιμοποιήσουμε μια μη πληροφοριακή prior η οποία θα αντικατοπτρίζει την άγνοια μας.

3.10 Συζυγείς εκ των προτέρων κατανομές

Όπως έχουμε ήδη αναφέρει, η επιλογή της εκ των προτέρων κατανομής $\pi(\theta)$ είναι σημαντική και δεν πρέπει να γίνεται τυχαία. Οι υπολογιστικές δυσκολίες αρχίζουν και εμφανίζονται όταν στο θεώρημα Bayes θέλουμε να υπολογίσουμε την σταθερά κανονικοποιήσης

$$\int_{\theta\in\Theta} L(x|\theta)\pi(\theta)d\theta$$

στον παρονομαστή.

Στο τελευταίο παράδειγμά μας είχαμε ότι $x = (x_1, x_2, \ldots, x_n)$ είναι τυχαίο δείγμα μιας τυχαίας μεταβλητής X που ακολουθεί την Poisson(θ) κατανομή. Αν οι γνώσεις μας για την παράμετρο θ είναι τέτοιες ώστε να είμαστε σίγουροι ότι ανήκει στο διάστημα $[0, 1]$, με όλες τις ενδιάμεσες τιμές ισοπίθανες, μπορούμε να πούμε ότι $\pi(\theta) = 1, 0 \leq \theta \leq 1$ και έχουμε ως αποτέλεσμα ότι

$$L(\theta|x) \propto exp(-n\theta)\theta^{\sum x_i}.$$

Τότε η σταθερά κανονικοποίησης είναι

$$\int_0^1 exp(-n\theta)\theta^{\sum x_i}d\theta.$$

Το παραπάνω ολοκλήρωμα, μπορεί να υπολογιστεί μόνο αριθμητικά. Το αποτέλεσμα της επιλογής που κάναμε για την εκ των προτέρων κατανομή ήταν να βρέθουμε σε αδιέξοδο. Στο αρχικό όμως παράδειγμά μας όπου είχαμε υποθέσει $\theta \sim Gamma(r, q)$, δεν είχαμε αυτό το πρόβλημα και μάλιστα είχαμε φτάσει στο συμπέρασμα $\theta|x \sim Gamma(R, Q)$.

Όταν επιλέγουμε μία εκ των προτέρων κατανομή $\pi(\theta)$ για τη θ και έχουμε ως αποτέλεσμα η εκ των υστέρων κατανομή $p(\theta|x)$ της θ να ανήκει στη ίδια οικογένεια κατανομών με την $\pi(\theta)$, τότε η $\pi(\theta)$ ονομάζεται συζυγής εκ των προτέρων κατανομή (conjugate prior).

Εδώ πρέπει να τονίσουμε ότι αν και οι συζυγείς εκ των προτέρων κατανομές διευκολύνουν τα μαθηματικά, πρέπει να τις χρησιμοποιούμε μόνο όταν είναι κατάλληλες δηλαδή μόνο όταν είναι κοντά στις πραγματικές πεποιθήσεις μας για την άγνωστη παράμετρο θ.

Η μοναδική περίπτωση να «παίρνουμε» συζυγείς εκ των προτέρων κατανομές είναι για κατανομές που ανήκουν στην εκθετική οικογένεια κατανομών (exponential family). Σε αυτές ισχύει ότι

$$L(x|\theta) = h(x)g(\theta)exp\{t(x)c(\theta)\}$$

Πίνακας 3.1: Συζυγείς εκ των προτέρων κατανομές για τις παραμέτρους γνωστών κατανομών.

Πιθανοφάνεια	$\pi(\theta)$	$p(\theta)$
$x \sim$ Διωνυμική(n, θ)	$Beta(r, q)$	$Beta(r + x, q + n - x)$
$x_1, \ldots, x_n \sim$ Γεωμετρική(θ)	$Beta(r, q)$	$Beta(r + n, \sum x_i - n)$
$x_1, \ldots, x_n \sim Poisson(\theta)$	$Gamma(r, q)$	$Gamma(r + \sum x_i, q + n)$
$x_1, \ldots, x_n \sim Normal(\theta, \tau^{-1})$, ($\tau$ γνωστό)	$Normal(b, c^{-1})$	$Normal(\frac{cb+n\tau\bar{x}}{c+n\tau}, \frac{1}{c+n\tau})$

όπου οι συναρτήσεις h, g, t, c είναι τέτοιες ώστε

$$\int L(x|\theta)dx = g(\theta) \int h(x)exp\{t(x)c(\theta)\}dx = 1.$$

Ενώ αυτό μπορεί αρχικά να φαίνεται περιοριστικό δεν είναι τόσο γιατί στην εκθετική οικογένεια κατανομών ανήκουν πάρα πολλές γνωστές κατανομές όπως η εκθετική κατανομή, η Poisson , η Gamma, η διωνυμική και τέλος η κανονική κατανομή με γνωστή διασπορά.

Στο πίνακα 3.1 υπάρχουν κάποιες γνωστές συναρτήσεις κατανομών στη πρώτη στήλη ενώ στη δεύτερη στήλη βρίσκονται οι εκ των προτέρων κατανομές που επιλέγουμε εμείς για τις άγνωστες παραμέτρους. Τέλος στη τρίτη στήλη που έχουμε τις εκ των υστέρων κατανομές της παραμέτρου θ παρατηρούμε ότι ανήκουν στην ίδια οικογένεια κατανομών με τις $\pi(\theta)$ που επιλέξαμε πριν άρα οι εκ των προτέρων κατανομές είναι συζυγείς.

3.11 Δεσμευμένες και περιθώριες εκ των υστέρων κατανομές

Στα πιο πολλά στατιστικά προβλήματα οι άγνωστες παράμετροι που θα έχουμε θα είναι αρκετές και είναι σύνηθες κάποιες από αυτές ή ακόμα και μόνο μία από αυτές να είναι «πιο σημαντική» η εκτίμηση της από τις άλλες. Η μέθοδος ανάλυσης προβλημάτων με αρκετές παραμέτρους στη Μπεϋζιανή στατιστική δε χρειάζεται καμία επιπλέον πληροφορία από αυτές που έχουμε ήδη αναφέρει.

Έστω ότι έχουμε μία διανυσματική άγνωστη παράμετρο $\theta = (\theta_1, \theta_2, \ldots, \theta_s)$ η οποία αποτελείται από s άγνωστες παραμέτρους τις οποίες θέλουμε να εκτιμήσουμε. Αφού ορίσουμε μια εκ των προτέρων κατανομή $\pi(\theta)$ για την θ, αυτή όταν συνδυαστεί με την πιθανοφάνεια $L(x|\theta)$ θα μας δώσει μέσω του Θεωρήματος

του Bayes την εκ των υστέρων κατανομή

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta)\pi(\theta)d\theta}$$

όπως και πριν απλά τώρα η σταθερά κανονικοποίησης είναι ένα πολυδιάστατο ολοκλήρωμα. Επίσης η εκ των υστέρων κατανομή θα είναι πολυδιάστατη.

Η δεσμευμένη εκ των υστέρων κατανομή (conditional posterior distribution) $p_i(\theta_i|y, \theta_{-i})$ μίας συγκεκριμένης συνιστώσας θ_i της θ δεδομένου των τιμών των υπολοίπων συνιστωσών θ_{-i} ορίζεται ως

$$p_i(\theta_i|x, \theta_{-i}) \propto p(\theta|x)$$

όπου οι τιμές των θ_{-i} θεωρούνται σταθερές. Στη πραγματικότητα όταν βρίσκουμε την από κοινού εκ των υστέρων κατανομή $p(\theta|x)$ τη θεωρούμε ως συνάρτηση μόνο της παραμέτρου θ_i, με σταθερές τις τιμές των άλλων παραμέτρων.

Η μπεϋζιανή συμπερασματολογία στηρίζεται στην εύρεση της περιθώρια εκ των υστέρων κατανομή (marginal posterior distribution) $p(\theta_i|x)$ μίας παραμέτρου θ_i όπου αυτή ορίζεται ως

$$p(\theta_i|x) = \int p(\theta|x)d\theta_{-i},$$

άρα βρίσκεται αφού πρώτα ολοκληρώσουμε την από κοινού εκ των υστέρων κατανομή $p(\theta|x)$ ως προς τις υπόλοιπες παραμέτρους θ_{-i}.

3.12 Μέθοδοι Markov chain Monte Carlo

Ένας αρκετά σημαντικός λόγος που η μπεϋζιανή στατιστική έχει ραγδαία ανάπτυξη τα τελευταία χρόνια είναι οι αλγόριθμοι προσομοίωσης Markov chain Monte Carlo. Αυτοί μας επιτρέπουν να ξεπεράσουμε προβλήματα μεγάλης υπολογιστικής δυσκολίας στα οποία μέχρι τότε δεν είχαμε απάντηση. Επίσης μας επιτρέπουν να είμαστε πιο ρεαλιστικοί στη μοντελοποίηση μας, αφού πλέον είναι εφικτή η συμπερασματολογία για πιο πολύπλοκα μοντέλα.

Συγκεκριμένα οι μέθοδοι Markov chain Monte Carlo (MCMC) είναι αλγόριθμοι προσομοίωσης οι οποίοι μας δίνουν δείγματα από την από κοινού εκ των υστέρων κατανομή των παραμέτρων βασισμένες, όπως προδίδει και το όνομα τους, στη δημιουργία μιας Μαρκοβιανής αλυσίδας η οποία μετά από ένα μεγάλο αριθμό βημάτων συγκλίνει στη κατανομή αυτή. Στη συνέχεια θα αναλύσουμε δύο MCMC μεθόδους, τον αλγόριθμο Gibbs και τον αλγόριθμο Metropolis-Hastings.

3.12.1 Αλγόριθμος Gibbs

Ας δούμε αρχικά ένα παράδειγμα όπου $X_i|\mu,\omega \sim N(\mu,\frac{1}{\omega})$ ανεξάρτητα για $i = 1,2,\ldots,n$ όπου οι άγνωστες παράμετροι μ,ω είναι η μέση τιμή και η ακρίβεια της κανονικής κατανομής αντίστοιχα. Άρα έχουμε ότι

$$f(x|\mu,\omega) = \prod_{i=1}^{n}(\frac{\omega}{2\pi})^{1/2}exp\{-\frac{\omega}{2}(x_i-\mu)^2\} \propto \omega^{n/2}exp\{-\frac{\omega}{2}\sum_{i=1}^{n}(x_i-\mu)^2\}$$

Ας υποθέσουμε ότι οι εκ των προτέρων κατανομές των μ,ω είναι

$$\mu \sim N(\mu_0,\frac{1}{\kappa_0}) \quad , \ \omega \sim Gamma(\alpha_0,\lambda_0)$$

όπου μ,ω είναι ανεξάρτητες μεταξύ τους. Η από κοινού εκ των υστέρων κατανομή των μ,ω είναι

$$f(\mu,\omega|x) \propto e^{-\frac{\omega}{2}\sum_{i=1}^{n}(x_i-\mu)^2}e^{-\frac{\kappa_0}{2}(\mu-\mu_0)^2}\omega^{\frac{n}{2}+\alpha_0-1}e^{-\lambda_0\omega}$$

Εδώ μπορούμε να παρατηρήσουμε ότι αν και οι εκ των προτέρων κατανομές των παραμέτρων είναι γνωστές κατανομές, η από κοινού εκ των υστέρων κατανομή είναι μία πολύπλοκη διδιάστατη κατανομή. Όμως οι δεσμευμένες εκ των υστέρων κατανομές των μ,ω ανήκουν στην ίδια οικογένεια κατανομών με τις εκ των προτέρων κατανομές τους, και πιο συγκεκριμένα

$$(\mu|\omega,x) \sim N(\frac{\kappa_0\mu_0+\omega\sum_{i=1}^{n}x_i}{\kappa_0+n\omega},\frac{1}{\kappa_0+n\omega}),$$

$$(\omega|\mu,x) \sim Gamma(\alpha_0+\frac{n}{2},\lambda_0+\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}).$$

Το φαινόμενο αυτό είναι γνωστό ως δεσμευμένη συζυγία (conditional conjugacy) και είναι αυτό που μας επιτρέπει να χρησιμοποιήσουμε τον αλγόριθμο Gibbs ο οποίος «απαιτεί» ως απαραίτητη πληροφορία οι δεσμευμένες εκ των υστέρων κατανομές να ανήκουν σε γνωστές κατανομές γιατί από αυτές θα κάνει την προσομοίωση. Έστω ότι έχουμε d άγνωστες παραμέτρους $(\theta_1,\theta_2,\ldots,\theta_d)$ και θέλουμε να πάρουμε ένα δείγμα από την από κοινού εκ των υστέρων κατανομή $f(\theta_1,\theta_2,\ldots,\theta_d|x)$.

Ο αλγόριθμος Gibbs έχει τα εξής βήματα

I Αρχικά έχουμε $\theta = (\theta_1^{(0)},\theta_2^{(0)},\ldots,\theta_d^{(0)})$.

II Προσομοιώνουμε $\theta_1^{(1)}$ από τη δεσμευμένη κατανομή $f(\theta_1|x,\theta_2^{(0)},\theta_3^{(0)},\ldots,\theta_d^{(0)})$.

III Προσομοιώνουμε $\theta_2^{(1)}$ από τη δεσμευμένη κατανομή $f(\theta_2|x,\theta_1^{(1)},\theta_3^{(0)},\ldots,\theta_d^{(0)})$.

IV ...

V Προσομοιώνουμε $\theta_d^{(1)}$ από τη δεσμευμένη κατανομή $f(\theta_d|x, \theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{d-1}^{(1)})$.

VI Επαναλαμβάνουμε τα βήματα $II - V$.

Η σύγκλιση της παραπάνω Μαρκοβιανής αλυσίδας στην από κοινού εκ των υστέρων κατανομή $f(\theta_1, \theta_2, \ldots, \theta_d|q)$ είναι εγγυημένη. Η παραπάνω διαδικασία συμπληρώνεται μετά από ένα μεγάλο αριθμό επαναλήψεων, αφού όμως πρώτα «αφαιρέσουμε» τα πρώτα δείγματα που θα πάρουμε από τις αρχικές επαναλήψεις (burn-in period) τα οποία δεν αποτελούν «ρεαλιστικά» δείγματα της από κοινού εκ των υστέρων κατανομής.

Όπως έχουμε ήδη τονίσει, τη παραπάνω διαδικασία την ακολουθούμε όταν χρειαζόμαστε δείγματα από την από κοινού εκ των υστέρων κατανομή. Αρκετά προβλήματα στα οποία η εύρεση του ολοκληρώματος στο κλάσμα

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

δεν είναι εφικτή και ως αποτέλεσμα δεν είναι εφικτή και η εύρεση της εκ των υστέρων κατανομής, ο αλγόριθμος Gibbs μας δίνει τη λύση, δίνοντας μας δείγματα από την από κοινού εκ των υστέρων κατανομή των άγνωστων παραμέτρων χωρίς να χρειαστεί να υπολογίσουμε το προηγούμενο ολοκλήρωμα.

3.12.2 Αλγόριθμος Metropolis-Hastings

Εκτός από τον αλγόριθμο Gibbs υπάρχουν και αρκετοί άλλοι MCMC αλγόριθμοι. Όπως έχουμε ήδη αναφέρει, ο αλγόριθμος Gibbs μπορεί να χρησιμοποιηθεί μόνο όταν μπορούμε να προσομοιώσουμε από τις δεσμευμένες εκ των υστέρων κατανομές των άγνωστων παραμέτρων του προβλήματος (δηλαδή μόνο οταν οι δεσμευμένες εκ των υστέρων κατανομές είναι γνωστές κατανομές). Είναι αρκετά πιθανό όμως να μην είναι γνωστές οι κατανομές κάτι που θα έχει ως αποτέλεσμα να μην μπορούμε να προσομοιώσουμε από αυτές. Σε αυτές τις περιπτώσεις ένας MCMC αλγόριθμος που μπορούμε να χρησιμοποιήσουμε είναι ο Metropolis-Hastings.

Ας υποθέσουμε ότι $X_i|\mu, \omega \sim Cauchy(\mu, \frac{1}{\omega})$ ανεξάρτητα για $i = 1, 2, \ldots, n$. Άρα έχουμε ότι

$$f(x|\mu, \omega) = \prod_{i=1}^{n} f(x_i|\mu, \omega) = \prod_{i=1}^{n} \frac{\omega^{1/2}}{\pi} \frac{1}{1 + \omega(x_i - \mu)^2}.$$

Ας υποθέσουμε επίσης, όπως και στο προηγούμενο παράδειγμα, ότι οι εκ των προτέρων κατανομές των μ, ω είναι

$$\mu \sim N(\mu_0, \frac{1}{\kappa_0}) \quad , \ \omega \sim Gamma(\alpha_0, \lambda_0)$$

όπου μ, ω είναι ανεξάρτητες μεταξύ τους. Η από κοινού εκ των υστέρων κατανομή των μ, ω είναι

$$f(\mu, \omega|x) \propto \{\prod_{i=1}^{n} \frac{1}{1+\omega(x_i-\mu)^2}\} e^{-\frac{\kappa_0}{2}(\mu-\mu_0)^2} \omega^{\frac{n}{2}+\alpha_0-1} e^{-\lambda_0\omega} I[\omega > 0].$$

Οι δεσμευμένες εκ των υστέρων κατανομές των μ, ω είναι

$$f(\mu|\omega, x) \propto \{\prod_{i=1}^{n} \frac{1}{1+\omega(x_i-\mu)^2}\} e^{-\frac{\kappa_0}{2}(\mu-\mu_0)^2}$$

και

$$f(\omega|\mu, x) \propto \{\prod_{i=1}^{n} \frac{1}{1+\omega(x_i-\mu)^2}\} \omega^{\frac{n}{2}+\alpha_0-1} e^{-\lambda_0\omega} I[\omega > 0].$$

Μπορούμε να παρατηρήσουμε ότι καμία από τις παραπάνω κατανομές δεν είναι γνωστή κατανομή και άρα δε θα μπορούσαμε να συνεχίσουμε με τον αλγόριθμο Gibbs. Το προηγούμενο παράδειγμα τονίζει την ανάγκη για πιο γενικευμένους αλγόριθμους από τον Gibbs , όπως είναι ο Metropolis-Hastings αλλά πριν τον δούμε πιο αναλυτικά είναι σημαντικό να δούμε τα βήματα που ακολουθούν οι γενικευμένοι αλγόριθμοι.

I Χωρίζουμε τις άγνωστες παραμέτρους σε d σύνολα $\theta_1, \ldots, \theta_d$ όπου το κάθε ένα έχει διάσταση ≥ 1.

II Αρχίζουμε με $(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_d^{(0)})$.

III Εκσυγχρονίζουμε το $\theta_1^{(0)}$ σε $\theta_1^{(1)}$ σύμφωνα με τη δεσμευμένη κατανομή $f(\theta_1|x, \theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_d^{(0)})$.

IV Εκσυγχρονίζουμε το $\theta_2^{(0)}$ σε $\theta_2^{(1)}$ σύμφωνα με τη δεσμευμένη κατανομή $f(\theta_2|x, \theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_d^{(0)})$.

V ...

VI Εκσυγχρονίζουμε το $\theta_d^{(0)}$ σε $\theta_d^{(1)}$ σύμφωνα με τη δεσμευμένη κατανομή $f(\theta_d|x, \theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{d-1}^{(1)})$.

VII Επαναλαμβάνουμε τα βήματα $III - VI$.

Όπως και πριν, έτσι και εδώ, διώχνουμε τα αρχικά δείγματα που θα πάρουμε και τα υπόλοιπα μπορούμε να θεωρήσουμε ότι είναι από τη ζητούμενη κατανομή. Πριν δούμε τον μηχανισμό εκσυγχρόνισης των τιμών του αλγόριθμου Metropolis-Hastings είναι σημαντικό να αναφέρουμε ότι οι λέξεις «σύμφωνα με» των προηγούμενων προτάσεων δεν σημαίνουν ότι προσομοιώνουμε από τη δεσμευμένη κατανομή των θ_i (δεν μπορούμε να το κάνουμε αυτό γιατι δεν τη γνωρίζουμε).

Ας υποθέσουμε τώρα ότι όπως τρέχουμε τον γενικευμένο MCMC αλγόριθμο έχουμε φτάσει στη $j-$οστή επανάληψη έχοντας τις τιμές $\theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_d^{(j)}$ και θέλουμε να προσομοιώσουμε την τιμή $\theta_1^{(j+1)}$, την επόμενη τιμή της θ_1. Ο μηχανισμός εκσυγχρόνισης των τιμών του αλγόριθμου Metropolis-Hastings είναι ο ακόλουθος.

1. Προτείνουμε μια υποψήφια τιμή θ_1^{can} η οποία είναι από μία τυχαία κατανομή με συνάρτηση πυκνότητας $q(\theta_1^{\text{can}} | \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_d^{(j)})$.

2. Επιλέγουμε ως επόμενη τιμή της θ_1 στην μαρκοβιανή αλυσίδα την $\theta_1^{(j+1)}$ όπου

$$\theta_1^{(j+1)} = \begin{cases} \theta_1^{\text{can}} & p \\ \theta_1^{(j)} & 1-p \end{cases}$$

όπου

$$p = \min\{1, \frac{f(\theta_1^{\text{can}} | x, \theta_2^{(j)}, \ldots, \theta_d^{(j)})}{f(\theta_1^{j} | x, \theta_2^{(j)}, \ldots, \theta_d^{(j)})} \frac{q(\theta_1^{j} | \theta_1^{\text{can}}, \theta_2^{(j)}, \ldots, \theta_d^{(j)})}{q(\theta_1^{\text{can}} | \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_d^{(j)})}\}$$

και $f(\theta_1^{\text{can}} | x, \theta_2^{(j)}, \ldots, \theta_d^{(j)})$ είναι η συνάρτηση πυκνότητας της δεσμευμένης κατανομής της θ_1 υπολογισμένη στη τιμή $\theta_1 = \theta_1^{\text{can}}$ και αντίστοιχα για την $f(\theta_1^{j} | x, \theta_2^{(j)}, \ldots, \theta_d^{(j)})$.

Τέλος, μπορούμε να προσθέσουμε ότι η επιλογή της τυχαίας συνάρτησης $q(\theta_1^{\text{can}} | \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_d^{(j)})$ αρκετές φορές μπορεί να είναι προφανής αλλά γενικά δε επιλέγουμε πάντα κάποια συγκεκριμένη. Ο αλγόριθμος Metropolis-Hastings έχει το πολύ σημαντικό πλεονέκτημα ότι δεν είναι απαραίτητο οι δεσμευμένες εκ των υστέρων κατανομές να είναι γνωστές όπως στον Gibbs. Ακόμα βλέπουμε ότι ο αλγόριθμος Gibbs είναι μία ειδική περίπτωση του Metropolis-Hastings όπου η τυχαία συνάρτηση πυκνότητας $q(\theta_1^{\text{can}} | \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_d^{(j)}) = f(\theta_1^{\text{can}} | \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_d^{(j)})$ και εδώ έχουμε ότι η πιθανότητα $\theta_1^{(j+1)} = \theta_1^{\text{can}}$ είναι πάντα 1[;].

Κεφάλαιο 4

Συμπερασματολογία με βάση τη πιθανοφάνεια για το γραμμικό και ποσοστημοριακό μοντέλο παλινδρόμησης

> And I'd join the movement
> If there was one
> I could believe in
> Yeah I'd break bread and wine
> If there was a church I could
> receive in.
>
> *Acrobat*
> U2

Έχοντας αναλύσει στο προηγούμενο κεφάλαιο όλες τις απαραίτητες έννοιες που πρέπει να γνωρίζουμε έτσι ώστε να είμαστε σε θέση να κατανοήσουμε τις δύο μεθοδολογίες στατιστικής συμπερασματολογίας που στηρίζονται στη συνάρτηση πιθανοφάνειας, σε αυτό το κεφάλαιο θα δούμε πως πραγματοποιούμε στατιστική συμπερασματολογία μέσω της συνάρτησης πιθανοφάνειας για τα μοντέλα γραμμικής και ποσοστημοριακής παλινδρόμησης (με την κλασική αλλά και την μπεϋζιανή προσέγγιση).

4.1 Εκτίμηση μέγιστης πιθανοφάνειας για το μοντέλο γραμμικής παλινδρόμησης

Ας δούμε αρχικά την περίπτωση της απλής γραμμικής παλινδρόμησης όπου έχουμε το μοντέλο

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, 2, \ldots, n$$

όπου $\epsilon_i \sim N(0, \sigma^2)$. Η άγνωστη παράμετρος μας εδώ είναι

$$\theta = \left[\beta_0, \beta_1, \sigma^2\right]$$

Η συνάρτηση πιθανοφάνειας μπορεί να γραφτεί ως

$$L(\beta_0, \beta_1, \sigma^2|y) \propto (\sigma^2)^{-\frac{n}{2}} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\}$$

και η αντίστοιχη λογαριθμική πιθανοφάνεια

$$l(\beta_0, \beta_1, \sigma^2|y) = -\frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Τώρα έχουμε ότι

$$\begin{bmatrix} \frac{\partial l}{\partial \beta_0} \\ \frac{\partial l}{\partial \beta_1} \\ \frac{\partial l}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \\ \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \end{bmatrix}$$

Αν θέσουμε τα στοιχεία του παραπάνω πίνακα ίσα με 0 έτσι ώστε να βρούμε τις εκτιμήτριες μέγιστης πιθανοφάνειας της παραμέτρου $\theta = \left[\beta_0, \beta_1, \sigma^2\right]$ έχουμε ότι

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n}$$

και $\hat{\beta}_0$ και $\hat{\beta}_1$ είναι οι λύσεις στο σύστημα πινάκων

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}.$$

Αύτες συγκεκριμένα είναι

$$\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

και

$$\beta_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2}$$

Εδώ μπορούμε να παρατηρήσουμε ότι οι εκτιμήσεις που παίρνουμε είναι ίδιες με αυτές της μεθόδου των ελαχίστων τετραγώνων.

Τέλος, ο πίνακας της αναμενόμενης πληροφορίας Fisher δίνεται ως

$$\mathbf{I}(\theta) = \sigma^{-2}\begin{bmatrix} n & \sum_{i=1}^n x_i & 0 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & 0 \\ 0 & 0 & (2\sigma^2)^{-1}n \end{bmatrix}$$

Τώρα θα ασχοληθούμε με το μοντέλο πολλαπλής γραμμικής παλινδρόμησης

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n.$$

όπου $\epsilon_i \sim N(0, \sigma^2 I)$. Έχουμε ήδη δει ότι αν θέσουμε

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

θα έχουμε ότι

$$Y = X\beta + \epsilon,$$

με τις υποθέσεις ότι $E(\epsilon) = 0$ και $V(\epsilon) = \sigma^2 I$ όπου σ^2 η διακύμανση του διαταρακτικού όρου και I ο μοναδιαίος πίνακας $n \times n$. Εδώ η άγνωστη παράμετρος είναι

$$\theta = \left[\beta, \sigma^2\right].$$

Μπορούμε να δούμε ότι $y_i \sim N(\mu_i, \sigma^2)$ όπου $\mu_i = x_i'\beta$ και $x_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ik} \end{pmatrix}$ για $i = 1, \ldots, n$.

Η συνάρτηση πιθανοφάνειας είναι

$$L(\theta) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - x_i'\beta)^2\}$$

Με αντίστοιχο τρόπο όπως πριν έχουμε ότι

$$\begin{aligned} l(\beta, \sigma^2|y) &= -\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \\ &= -\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}[(y - X\beta)'(y - X\beta)] \\ &= -\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}[y'y - 2\beta' X'y + \beta' X'X\beta] \end{aligned}$$

και

$$\begin{aligned} \frac{\partial l(\beta, \sigma^2|y)}{\partial \beta} &= -\frac{1}{2\sigma^2}[\frac{\partial[y'y - 2\beta' X'y + \beta' X'X\beta]}{\partial \beta}] \\ &= -\frac{1}{2\sigma^2}[-2X'y + 2X'X\beta] \\ &= \frac{1}{2\sigma^2}[X'y - X'X\beta]. \end{aligned}$$

Τέλος, έχουμε ότι

$$\begin{aligned} \frac{\partial l(\beta, \sigma^2|y)}{\partial \beta} &= 0 \\ \frac{1}{2\sigma^2}[X'y - X'X\beta] &= 0 \\ X'X\beta &= X'y \\ \beta &= (X'X)^{-1}X'y. \end{aligned}$$

Για τη διασπορά σ^2 έχουμε ότι

$$\frac{\partial l(\beta, \sigma^2|y)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}[y - X\beta)'(y - X\beta)]$$

Άρα

$$\begin{aligned} \frac{\partial l(\beta, \sigma^2|y)}{\partial \sigma^2} &= 0 \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}[y - X\beta)'(y - X\beta)] &= 0 \\ \frac{1}{2\sigma^4}[y - X\beta)'(y - X\beta)] &= \frac{n}{2\sigma^2} \\ \frac{1}{\sigma^2}[y - X\beta)'(y - X\beta)] &= n. \end{aligned}$$

Στη προηγούμενη ισότητα αντικαθιστούμε το β με την εκτιμήτρια του $\hat{\beta}$ που έχουμε ήδη βρει και βρίσκουμε

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y - X'\hat{\beta})^2.$$

Συνοψίζοντας έχουμε ότι οι εκτιμήτριες μέγιστης πιθανοφάνειας[;] και ο πίνακας πληροφορίας Fisher είναι

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y,\\ \hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2\\ I(\hat{\beta}) &= \hat{\sigma}^{-2}(X'X).\end{aligned}$$

Τα τυπικά σφάλματα των εκτιμήσεων των παραμέτρων της παλινδρόμησης $\hat{\beta}$ είναι η τετραγωνική ρίζα των διαγώνιων στοιχείων του πίνακα

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

Για να διαβάσει κάποιος πιο αναλυτικά την μέθοδο εύρεσης ε.μ.π. για το μοντέλο γραμμικής παλινδρόμησης μπορεί να κοιτάξει τα άρθρα των Franklin[?] Gonzalez[?].

4.2 Μπεϋζιανή συμπερασματολογία για το μοντέλο γραμμικής παλινδρόμησης

Θα ασχοληθούμε με το μοντέλο πολλαπλής γραμμικής παλινδρόμησης και την εκτίμηση των συντελεστών παλινδρόμησης μέσω μπεϋζιανής στατιστικής. Το μοντέλο όπως θυμόμαστε είναι

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n.$$

όπου $\epsilon_i \sim N(0, \sigma^2 I)$ και το οποίο μπορεί να γραφτεί ως

$$Y = X\beta + \epsilon,$$

με τις υποθέσεις ότι $E(\epsilon) = 0$ και $V(\epsilon) = \sigma^2 I$ όπου σ^2 η διακύμανση του διαταρακτικού όρου και I ο μοναδιαίος πίνακας $n \times n$. Εδώ, όπως και πριν, η άγνωστη παράμετρος είναι

$$\theta = \left[\beta, \sigma^2\right]$$

Η συνάρτηση πιθανοφάνειας τώρα είναι

$$L(\beta, \sigma^2|y) \propto (\sigma^2)^{-n/2} exp[-\frac{(y - X\beta)'(y - X\beta)}{2\sigma^2}].$$

Επίσης έχουμε

$$\begin{aligned}(y - X\beta)'(y - X\beta) &= (y - X\hat{\beta} - X\beta + X\hat{\beta})'(y - X\hat{\beta} - X\beta + X\hat{\beta}) \\ &= [y - X\hat{\beta} - X(\beta - \hat{\beta})][y - X\hat{\beta} - X(\beta - \hat{\beta})] \\ &= (y - X\hat{\beta})'(y - X\hat{\beta}) + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \\ &= S + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}).\end{aligned}$$

Όπου

$$S = (y - X\hat{\beta})'(y - X\hat{\beta})$$

είναι το άθροισμα των τετραγώνων των καταλοίπων και

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

είναι η ε.μ.π. για το β. Άρα τώρα έχουμε

$$L(\beta, \sigma^2|y) \propto (\sigma^2)^{-n/2} exp[-\frac{S + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})}{2\sigma^2}].$$

Ας υποθέσουμε τώρα ως εκ των προτέρων κατανομή της $\theta = [\beta, \sigma^2]$ την αντίστροφη κανονική Γάμμα κατανομή ($NIG(a, d, m, V)$) (Normal Inverse Gamma) με υπερπαραμέτρους a, d, m, V

$$\pi(\theta) = \frac{(a/2)^{d/2}}{(2\pi)^{p/2}|V|^{1/2}\Gamma(d/2)}(\sigma^2)^{-(d+p+2)/2} exp[-\{(\beta - m)'V^{-1}(\beta - m) + a\}/(2\sigma^2)],$$

άρα

$$\pi(\theta) \propto (\sigma^2)^{-(d+p+2)/2} exp[-\{(\beta - m)'V^{-1}(\beta - m) + a\}/(2\sigma^2)]$$

και γνωρίζοντας ότι

$$p(\theta|y) \propto \pi(\theta)L(\theta|y)$$

έχουμε

$$p(\theta|y) \propto (\sigma^2)^{-(d+n+p+2)/2} exp[-Q/(2\sigma^2)]$$

όπου

$$\begin{aligned}Q &= (y - X\beta)'(y - X\beta) + (\beta - m)'V^{-1}(\beta - m) + a \\ &= \beta'(V^{-1} + X'X)\beta - \beta'(V^{-1}m + X'y) - (m'V^{-1} + y'X)\beta + (m'V^{-1}m + y'y + a) \\ &= (\beta - m^*)'(V^*)^{-1}(\beta - m^*) + a^*,\end{aligned}$$

με

$$V^* = (V^{-1} + X'X)^{-1},$$

$$m^* = (V^{-1} + X'X)^{-1}(V^{-1}m + X'y),$$

$$a^* = a + m'V^{-1}m + y'y - (m^*)'(V^*)^{-1}m^*.$$

Τελικά, αν θέσουμε $d^* = d + n$ έχουμε ότι η από κοινού εκ των υστέρων κατανομή του θ

$$p(\theta|y) = p(\beta, \sigma^2|y) \sim NIG(a^*, d^*, m^*, V^*)$$

Για να βρούμε τώρα τις περιθώριες εκ των υστέρων κατανομές των β, σ^2 θα ολοκληρώσουμε την από κοινού εκ των υστέρων κατανομή $p(\beta, \sigma^2|y)$ ως προς σ^2 και β αντίστοιχα.

$$\begin{aligned} p(\beta|y) &= \int p(\beta, \sigma^2|y)d\sigma^2 \\ &\propto \int (\sigma^2)^{-(d+n+p+2)/2} exp[-Q/(2\sigma^2)]d\sigma^2 \\ &\propto \{1 + (\beta - m^*)'(a^*V^*)^{-1}(\beta - m^*)\}^{-(d^*+p)/2}, \end{aligned}$$

και έχουμε ότι η περιθώρια εκ των υστέρων κατανομή της β ακολουθεί την πολυδιάστατη t κατανομή με d^* βαθμούς ελευθερίας και υπερπαράμετρους m^* και a^*V^*.

Αντίστοιχα για την περιθώρια εκ των υστέρων κατανομή της άγνωστης παραμέτρου σ^2 έχουμε

$$\begin{aligned} p(\sigma^2|y) &= \int p(\beta, \sigma^2|y)d\beta \\ &\propto \int (\sigma^2)^{-(d+n+p+2)/2} exp[-Q/(2\sigma^2)]d\beta \\ &\propto (\sigma^2)^{-(d^*+2)/2} exp[-a^*/(2\sigma^2)], \end{aligned}$$

και έχουμε ότι η περιθώρια εκ των υστέρων κατανομή της σ^2 ακολουθεί την αντίστροφη Γάμμα κατανομή ($IG(a^*, d^*)$) (Inverse Gamma) με υπερπαραμέτρους $a*$ και d^*.

Περισσότερες πληροφορίες για μπεϋζιανή συμπερασματολογία στο μοντέλο γραμμικής παλινδρόμησης μπορούν να βρεθούν στα βιβλία των Sorensen, Gianola[**?**]και του O'Hagan[**?**].

4.3 Εκτίμηση μέγιστης πιθανοφάνειας για το μοντέλο ποσοστημοριακής παλινδρόμησης

Στα προηγούμενα υποκεφάλαια ασχοληθήκαμε με τη στατιστική συμπερασματολογία στο απλό γραμμικό μοντέλο μέσω της συνάρτησης πιθανοφάνειας σύμφωνα με τη κλασική αλλά και τη μπεϋζιανή στατιστική. Αυτό που είναι πολύ σημαντικό για τη συμπερασματολογία σε οποιοδήποτε μοντέλο παλινδρόμησης είναι η γνώση της κατανομής που ακολουθεί ο στοχαστικός όρος ϵ_i. Στο γραμμικό μοντέλο των προηγούμενων υποκεφαλαίων είχαμε ότι

$$\epsilon_i \sim N(0, \sigma^2).$$

Στο μοντέλο ποσοστημοριακής παλινδρόμησης η κατανομή που χρησιμοποιείται συνήθως είναι η ασσύμετρη κατανομή Laplace (assymetric Laplace). Μια τυχαία μεταβλητή U λέμε ότι ακολουθεί την ασσύμετρη Laplace κατανομή όταν η συνάρτηση πυκνότητας πιθανότητας της δίνεται από τον τύπο

$$g_p(u) = p(1-p)exp\{-\rho_p(u)\}, \quad 0 < p < 1, \quad u \in [0, \infty)$$

όπου

$$\rho_p(u) = puI_{[0,\infty)}(u) - (1-p)uI_{(-\infty,0)}(u)$$

η συνάρτηση ελέγχου (check function). Όταν $p = 1/2$ έχουμε ότι $g_p(u) = \frac{1}{4}exp\{-\frac{|u|}{2}\}$ η οποία είναι η συνάρτηση πιθανότητας της συμμετρικής Laplace κατανομής. Για όλες τις άλλες τιμές του p, η κατανομή είναι ασσύμετρη.

Η μέση τιμή της κατανομής είναι $\frac{1-2p}{p(1-p)}$ και είναι αρνητική για $p > 1/2$, θετική για $p < 1/2$ και 0 για $p = 1/2$ ενώ η διασπορά της είναι $\frac{1-2p+2p^2}{p^2(1-p)^2}$ η οποία αυξάνεται ταχύτατα καθώς το p πλησιάζει τη μονάδα.

Στη ποσοστημοριακή παλινδρόμηση έχουμε ήδη αναφέρει τη σχέση

$$y_i = \beta_{0,p} + \beta_{1,p}x_{i1} + \ldots + \beta_{k,p}x_{ik} + \epsilon_{i,p} \quad i = 1, \ldots, n.$$

η οποια συνδέει την εξαρτημένη μεταβλητή Y με τις ανεξάρτητες X_i.

Είχαμε θέσει

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}, \beta_p = \begin{pmatrix} \beta_0(p) \\ \beta_1(p) \\ \vdots \\ \beta_k(p) \end{pmatrix}, \epsilon_p = \begin{pmatrix} \epsilon_{1,p} \\ \epsilon_{2,p} \\ \vdots \\ \epsilon_{n,p} \end{pmatrix}$$

και ως αποτέλεσμα έχουμε ότι το μοντέλο ποσοστημοριακής παλινδρόμησης γράφεται σε μορφή πινάκων ως

$$Y = X\beta_p + \epsilon_p.$$

Είχαμε επίσης δει ότι για να βρούμε τις εκτιμήσεις ελαχίστων τετραγώνων, $\hat{\beta}_p$, ζητάμε να ελαχιστοποιηθεί ως προς β_p το άθροισμα

$$\sum_{i=1}^{n} \rho_p(Y - X\beta_p)$$

όπου $\rho_p(u) = p|u|$ είναι η απόλυτη συνάρτηση απώλειας. Αυτό είναι ισοδύναμο με το να μεγιστοποιήσουμε ως προς β_p τη πιθανοφάνεια

$$L(y_i|\beta_p, \sigma_p, x_1, \ldots, x_n) = (\frac{p(1-p)}{\sigma_p})^n exp\{-\frac{1}{\sigma_p}\sum_{i=1}^{n} \rho_p(y_i - \beta_p x_i)\}$$

όπου $0 < p < 1,\ \sigma_p > 0$.

Η πιθανοφάνεια στη παραπάνω εξίσωση δεν είναι απαραίτητο ότι αντιστοιχεί στη κατανομή της Y, όμως έχει το πλεονέκτημα ότι οι ε.μ.π. είναι ίδιες με τις εκτιμήσεις που παίρνουμε αν ελαχιστοποιήσουμε το προηγούμενο άθροισμα και ως αποτέλεσμα έχουν τις ίδιες ιδιότητες ασυμπτωτικά.

Έτσι έχουμε ως αποτέλεσμα οι εκτιμήτριες μέγιστης πιθανοφάνειας [;] να είναι

$$\hat{\beta}_p = \underset{\beta_p}{\operatorname{argmin}}[\sum_{i=1}^{n} \rho_p(y_i - \beta_p x_i)]$$

και

$$\hat{\sigma}_p = \frac{1}{n}\sum_{i=1}^{n} \rho_p(y_i - \beta_p x_i).$$

Για την εύρεση των τυπικών σφαλμάτων των εκτιμήσεων μπορεί να χρησιμοποιηθεί η Bootstrap καθώς αυτά δεν μπορούν να υπολογιστούν αναλυτικά[;]. Τέλος, εκτός από την ασσύμετρη Laplace κατανομή, ως κατανομή του στοχαστικού όρου μπορεί να χρησιμοποιηθεί κάποια που να ανήκει στην tick-exponential οικογένεια κατανομών. Για περισσότερες πληροφορίες μπορεί κάποιος να διαβάσει το άρθρο της Komunjer [;].

4.4 Μπεϋζιανή συμπερασματολογία για το μοντέλο ποσοστημοριακής παλινδρόμησης

Αν και δεν υπάρχει μια τυπική συζυγής εκ των προτέρων κατανομή που να χρησιμοποιούμε πάντα για τη ποσοστημοριακή παλινδρόμηση, λόγω των MCMC

(Markov chain Monte Carlo) μεθόδων μπορούμε να βρίσκουμε τις εκ των υστέρων κατανομές των άγνωστων παραμέτρων. Η διευκόλυνση που μας παρέχουν οι μέθοδοι MCMC έχει ως αποτέλεσμα να επιλέγουμε σχεδόν όποια εκ των προτέρων κατανομή επιθυμούμε. Επίσης, εκτός του ότι μπορούμε να πάρουμε τις δεσμευμένες και τις περιθώριες εκ των υστέρων κατανομές των άγνωστων παραμέτρων, κάνοντας μπεϋζιανή συμπερασματολογία μπορούμε να ενσωματώσουμε στις παραμέτρους οποιαδήποτε αβεβαιότητα έχουμε για αυτές.

Έχοντας ένα δείγμα $y = (y_1, y_2, \ldots, y_n)$, η από κοινού εκ των υστέρων κατανομή των άγνωστων παραμέτρων β_p, σ_p δίνεται από τον τύπο

$$\pi(\beta_p, \sigma_p|y) \propto L(y|\beta_p, \sigma_p)p(\beta_p)p(\sigma_p),$$

όπου $p(\beta_p), p(\sigma_p)$ είναι οι εκ των προτέρων κατανομές των β_p, σ_p και $L(y|\beta_p, \sigma_p)$ είναι η συνάρτηση πιθανοφάνειας η οποία μπορεί να γραφτεί ως

$$L(y_i|\beta_p, \sigma_p, x_1, \ldots, x_n) = (\frac{p(1-p)}{\sigma_p})^n exp\{-\frac{1}{\sigma_p}\sum_{i=1}^{n}\rho_p(y_i - \beta_p x_i)\}$$

όπου $0 < p < 1,\ \sigma_p > 0$.

Ενώ όπως είπαμε και πριν μπορούμε να χρησιμοποιήσουμε οποιαδήποτε εκ των προτέρων κατανομή για τη β_p οι Yu και Moyeed[**?**] έδειξαν ότι ακόμα και μία ακατάλληλη (improper) ομοιόμορφη κατανομή να δεχτούμε ως εκ των προτέρων κατανομή της β_p θα έχουμε μια από κοινού εκ των υστέρων κατανομή η οποία εκτός από το ότι θα είναι ανάλογη της πιθανοφάνειας, θα είναι και κατάλληλη (proper).[1]

Μπορούμε να χρησιμοποιήσουμε ως εκ των προτέρων κατανομή για τη β_p την κανονική κατανομή με μέση τιμή 0 και διασπορά σ_p^2 και για τη σ_p την ομοιόμορφη κατανομή (η οποία όπως είπαμε είναι ακατάλληλη) και άρα έχουμε ότι

$$p(\beta_p) \propto exp\{-\frac{\beta_p^2}{2\sigma_p^2}\}$$

και

$$p(\sigma_p) \propto 1.$$

Άρα έχουμε

$$\pi(\beta_p, \sigma_p|y) \propto (\frac{p(1-p)}{\sigma_p})^n exp\{-\frac{1}{\sigma_p}\sum_{i=1}^{n}\rho_p(y_i - \beta_p x_i)\}exp\{-\frac{\beta_p^2}{2\sigma_p^2}\}$$

[1] Ως ακατάλληλη εκ των προτέρων κατανομή θεωρείται μία εκ των προτέρων κατανομή $p(\beta)$ η οποία δεν έχει την ιδιότητα

$$0 < \int p(\beta)d\beta < \infty$$

Η δεσμευμένη εκ των υστέρων κατανομή της β_p είναι

$$\pi(\beta_p|y,\sigma_p) \propto exp\{-\frac{1}{\sigma_p}\sum_{i=1}^{n}\rho_p(y_i-\beta_p x_i)\}exp\{-\frac{\beta_p^2}{2\sigma_p^2}\}$$

και της σ_p

$$\pi(\sigma_p|y,\beta_p) \propto \sigma_p^{-n}exp\{-\frac{1}{\sigma_p}\sum_{i=1}^{n}\rho_p(y_i-\beta_p x_i)\}exp\{-\frac{\beta_p^2}{2\sigma_p^2}\}.$$

Επειδή οι δεσμευμένες εκ των υστέρων κατανομές δεν είναι γνωστές κατανομές, δεν μπορούμε να προσομοιώσουμε από αυτές και συνεχίζουμε, χρησιμοποιώντας τον αλγόριθμο Metropolis-Hastings. Στη περίπτωση που ανήκαν σε γνωστές κατανομές θα χρησιμοποιούσαμε, όπως έχουμε ήδη αναφέρει στο τρίτο κεφάλαιο, τον αλγόριθμο Gibbs.

Κεφάλαιο 5

Κριτήρια επιλογής στατιστικού μοντέλου βασισμένα στη πιθανοφάνεια

Rhaegar fought valiantly,
Rhaegar fought nobly,
Rhaegar fought honorably.
And Rhaegar died.

A Storm Of Swords
Ser Jorah Mormont

Στα προηγούμενα κεφάλαια παρουσιάσαμε τις σχέσεις μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών για τα μοντέλα της πολλαπλής γραμμικής (linear) και ποσοστημοριακής (quantile) παλινδρόμησης. Όπως έχουμε ήδη τονίσει, σκοπός μας όταν υποθέτουμε κάποιο στατιστικό μοντέλο είναι να περιγράψουμε τη σχέση μεταξύ των μεταβλητών όσο μπορούμε πιο κοντά στη πραγματικότητα. Αν μπορούμε να το κάνουμε αυτό χρησιμοποιώντας όσο το δυνατόν λιγότερες μεταβλητές η δουλειά μας γίνεται υπολογιστικά πιο εύκολη. Επίσης σε αρκετά μοντέλα παλινδρόμησης μας ενδιαφέρει να επιλέξουμε ποιές, από ένα σύνολο διαθέσιμων επεξηγηματικών μεταβλητών, είναι οι καταλληλότερες να περιληφθούν στο μοντέλο. Θέλουμε όσο το δυνατό καλύτερη προσαρμογή με όσο το δυνατό λιγότερες παραμέτρους. Επιλογή μεταξύ των «υποψήφιων» μοντέλων από τα οποία θα επιλέξουμε αυτό πάνω στο οποίο θα βασίσουμε όλη τη συμπερασματολογία μας γίνεται επίσης όταν υπάρχει αβεβαιότητα για το πλήθος των ανεξάρτητων μεταβλητών που θέλουμε να συμπεριληφθούν στο αρχικό μοντέλο γραμμικής ή ποσοστημοριακής παλινδρόμησης. Παρακάτω θα δούμε κάποια βασικά κριτήρια επιλογής μοντέλου τα οποία

βασίζονται στη πιθανοφάνεια.

5.1 Akaike Information Criterion

Το Akaike Information Criterion (AIC) παρουσιάστηκε από τον Hirotsugu Akaike το 1974 ως ένα μέτρο καλής προσαρμογής (goodness of fit) ενός στατιστικού μοντέλου. Τα στατιστικά μοντέλα κατατάσσονται βάσει του AIC τους, με καλύτερο αυτό που έχει το μικρότερο AIC. Η τιμή του AIC για κάθε μοντέλο δίνεται από τον τύπο

$$\mathrm{AIC} = -2\log L + 2k,$$

όπου k είναι ο αριθμός των παραμέτρων του μοντέλου και L είναι η μέγιστη τιμή της συνάρτησης πιθανοφάνειας του συγκεκριμένου μοντέλου. Μπορούμε ακόμα να ερμηνεύσουμε τον πρώτο όρο του AIC ως ένα μέτρο καλής προσαρμογής και τον δεύτερο ως μια ποινή (penalty) που έχουμε επειδή στο μοντέλο μας έχουμε k παραμέτρους για εκτίμηση. Μπορούμε να παρατηρήσουμε ότι η συνάρτηση AIC είναι αύξουσα ως προς των αριθμό των εκτιμώμενων παραμέτρων. Αυτό φυσικά μας αποτρέπει από το να προσθέτουμε παραμέτρους προς εκτίμηση γιατί έτσι η ποινή που θα έχει αυτό το μοντέλο θα είναι μεγαλύτερη. Είναι σημαντικό να πούμε ότι η τιμή AIC μας βοηθάει στο να κατατάσσουμε τα μοντέλα και να επιλέγουμε αυτό με τη μικρότερη τιμή αλλά από μόνη της δεν μας βοηθάει σε τίποτα και ούτε μπορούμε να βγαλουμε συμπεράσματα αν π.χ. ένα μοντέλο έχει $\mathrm{AIC} = 6$. Πρέπει πάντα να γίνεται σύγκριση με τις τιμές των AIC των άλλων μοντέλων.

Ας υποθέσουμε ότι έχουμε το μοντέλο πολλαπλής γραμμικής παλινδρόμησης

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i \quad i = 1, 2, \ldots, n$$

όπου

$$y_i \sim N(x_i^{'}\beta, \sigma^2) \quad i = 1, 2, \ldots, n$$

και

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

αποτελείται από $k+1$ παραμέτρους. Επίσης δεν γνωρίζουμε και τη διασπορά σ^2 του στοχαστικού όρου άρα έχουμε συνολικά $k+2$ άγνωστες παραμέτρους στο παραπάνω στατιστικό μοντέλο.

Όπως έχουμε ήδη δείξει

$$\hat{\beta} = (X'X)^{-1}X'Y,$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2$$

και η συνάρτηση πιθανοφάνειας είναι

$$L(\beta, \sigma^2) = \big(\frac{1}{2\pi\sigma^2}\big)^{n/2} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\}.$$

Άρα, με αντικατάσταση των εκτιμητριών μέγιστης πιθανοφάνειας στον παραπάνω τύπο έχουμε

$$\log L(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{n}{2}.$$

Η τιμή AIC του παραπάνω μοντέλου θα είναι

$$\text{AIC} = n\log(2\pi\hat{\sigma}^2) + n + 2(k+2).$$

Αν τώρα έχουμε το μοντέλο πολλαπλής γραμμικής παλινδρόμησης

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{k-1}x_{ik-1} + \epsilon_i \quad i = 1, 2, \ldots, n$$

δηλαδή το ίδιο μοντέλο με πριν αλλά χωρίς την παράμετρο β_k. Αντίστοιχα, αφότου βρούμε τις καινούργιες εκτιμήτριες μέγιστης πιθανοφάνειας, θα βρούμε τιμή AIC

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

και στο τέλος θα επιλέξουμε το μοντέλο με τη μικρότερη τιμή μεταξύ των δύο διαφορετικών τιμών AIC που βρήκαμε. Τέλος, φαίνεται εδώ ότι αν αυξάνουμε τις παραμέτρους η τιμή της εκτίμησης $\hat{\sigma}^2$ μειώνεται αλλά προσθέτουμε στο AIC την ποινή επειδή χρησιμοποιήσαμε παραπάνω παραμέτρους.

5.2 Bayesian Information Criterion

Το Bayesian Information Criterion (BIC) παρουσιάστηκε από τον Gideon E. Schwarz ως ένα πιο συντηρητικό ως προς τις παραμέτρους του μοντέλου κριτήριο επιλογής από το AIC. Το BIC ευνοεί μοντέλα με μικρότερο αριθμό παραμέτρων δίνοντας μεγαλύτερη ποινή όσο αυξάνουμε τον αριθμό των παραμέτρων. Η τιμή του BIC για κάθε μοντέλο δίνεται από τον τύπο

$$\text{BIC} = -2\log L + klogn.$$

όπου k είναι ο αριθμός των παραμέτρων του μοντέλου, n το πλήθος των παρατηρήσεων και L είναι η μέγιστη τιμή της συνάρτησης πιθανοφάνειας του συγκεκριμένου μοντέλου. Ισχύουν ακριβώς τα ίδια με το AIC όσον αφορά ποιο μοντέλο επιλέγουμε αλλά και του ότι η τιμή που βρίσκουμε δεν μας δίνει καμία πληροφορία από μόνη της.

Παρατηρούμε ότι το BIC είναι αύξουσα συνάρτηση ως προς k αλλά και $\hat{\sigma}^2$. Έτσι για να μειώσουμε την τιμή του BIC πρέπει είτε να μειώσουμε τις παραμέτρους, είτε τη διασπορά του στοχαστικού όρου ή και τα δύο.

Αν υποθέσουμε πάλι ότι έχουμε το μοντέλο πολλαπλής γραμμικής παλινδρόμησης

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i \quad i = 1, 2, \ldots, n$$

Θα βρούμε τιμή BIC

$$\text{BIC} = n \log(2\pi\hat{\sigma}^2) + n + (k+2) \log n$$

και στη συνέχεια, όπως και πριν, κάνουμε σύγκριση με τις τιμές BIC των άλλων υποψήφιων μοντέλων και επιλέγουμε αυτό με τη μικρότερη.

Μπορούμε συνοψίζοντας να δούμε κάποια χαρακτηριστικά του κριτηρίου BIC όπως ότι είναι ανεξάρτητο της εκ των προτέρων κατανομής, τιμωρεί την πολυπλοκότητα του μοντέλου όπου ως πολυπλοκότητα αναφέρουμε τον αριθμό των παραμέτρων που έχει το στατιστικό μοντέλο και τέλος ότι είναι άμεσα συνδεδεμένο με το AIC.

5.3 Likelihood ratio test

Το Likelihood ratio test είναι ένας έλεγχος που μας βοηθάει να επιλέξουμε μεταξύ δύο μοντέλων, ένα εκ των οποίων είναι υπερσύνολο του άλλου με την έννοια ότι περιέχει τις ίδιες παραμέτρους και έχει τουλάχιστον μία παράμετρο περισσότερη. Το Likelihood ratio συνήθως συμβολίζεται με το ελληνικό γράμμα Λ και είναι ίσο με το πηλίκο των πιθανοφανειών των δύο μοντέλων.

Αν έχουμε δύο μοντέλα στα οποία στο πρώτο υποθέτουμε ότι η άγνωστη παράμετρος $\theta = \theta_0$ ενώ στο δεύτερο $\theta = \theta_1$ αυτό γράφεται συνήθως ως

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta = \theta_1$$

όπου H_0 ονομάζεται μηδενική υπόθεση (null hypothesis) και η H_1 εναλλακτική υπόθεση (alternative hypothesis). Άρα εδώ έχουμε ότι

$$\Lambda(x) = \frac{L(\theta_0|x)}{L(\theta_1|x)}.$$

Ο αριθμητής του κλάσματος αντιστοιχεί στη μέγιστη πιθανότητα να παρατηρήσουμε ένα τέτοιο αποτέλεσμα αν ισχύει η μηδενική υπόθεση και ο παρονομαστής στη μέγιστη πιθανότητα να παρατηρήσουμε ένα τέτοιο αποτέλεσμα δεδομένου ότι ισχύει η εναλλακτική υπόθεση.

Χαμηλές τιμές του $\Lambda(x)$ δείχνουν ότι το αποτέλεσμα που έχουμε είναι λιγότερο πιθανό να έχει πραγματοποιηθεί λόγω της μηδενικής υπόθεσης συγκριτικά με την εναλλακτική. Άρα μπορούμε να απορρίψουμε την μηδενική υπόθεση. Μεγάλες τιμές του $\Lambda(x)$ δείχνουν ότι το αποτέλεσμα που έχουμε είναι πιο πιθανό να έχει συμβεί επειδή ισχύει η μηδενική υπόθεση παρά η εναλλακτική. Άρα εδώ δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση H_0.[1]

Πιο συγκεκριμένα

- Αν $\Lambda(x) > c$ τότε δεν απορρίπτουμε την H_0.
- Αν $\Lambda(x) < c$ τότε απορρίπτουμε την H_0.
- Αν $\Lambda(x) = c$ τότε απορρίπτουμε την H_0 με πιθανότητα q,

όπου c, q έχουν επιλεγεί ώστε

$$q\mathbf{Pr}(\Lambda(x) = c|H_0) + \mathbf{Pr}(\Lambda(x) < c|H_0) = \alpha$$

με α το επίπεδο σημαντικότητας.

[1]Μπορούμε να παρατηρήσουμε ότι ανεξαρτήτως του $\Lambda(x)$, δεν δεχόμαστε ποτέ την H_0. Αυτό οφείλεται στο ότι δεν γνωρίζουμε αν είναι αληθινή με το Likelihood ratio test απλά είτε έχουμε στοιχεία που να την απορρίπτουν ή δεν έχουμε και άρα δεν την απορρίπτουμε

Κεφάλαιο 6

Εφαρμογές

> There is no such thing as an impartial jury because there are no impartial people. There are people that argue on the web for hours about who their favorite character on "Friends" is.
>
> *The Daily Show*
> Jon Stewart

Τις τελευταίες δεκαετίες τα αμοιβαία κεφάλαια υψηλού κινδύνου (hedge funds) αρχίζουν και γίνονται όλο και πιο διαδεδομένα στην αγορά. Τα hedge funds χαρακτηρίζονται κυρίως από μεγάλη ευελιξία στην επενδυτική τους πολιτική και τις στρατηγικές που ακολουθούν. Κυκλοφορούν στις πλέον ώριμες χρηματοοικονομικές αγορές. Δεν είναι τυχαίο ότι κάποιες ημέρες τα αμοιβαία κεφάλαια υψηλού κινδύνου αντιπροσωπεύουν μέχρι και το 40% του καθημερινού όγκου συναλλαγών αγορών όπως το Euronext, το χρηματιστήριο του Λονδίνου και της Νέας Υόρκης. Λογικό επακόλουθο της πληθώρας στρατηγικών που έχουμε με τα αμοιβαία κεφάλαια υψηλού κινδύνου είναι η προσπάθεια εύρεσης στατιστικών/οικονομετρικών μοντέλων που να αντικατοπτρίζουν τις τιμές τους δηλαδή η εύρεση μιας σχέσης μεταξύ των τιμών των αμοιβαίων κεφαλαίων υψηλού κινδύνου ως προς κάποιες άλλες μεταβλητές.

Για να περιγραφεί η σχέση μεταξύ των τιμών των αμοιβαίων κεφαλαίων υψηλού κινδύνου και των άλλων μεταβλητών γνωστών και ως παράγοντες κινδύνου (risk factors) χρησιμοποιούνται συνήθως μοντέλα γραμμικής παλινδρόμησης. Αυτό που είναι ιδιαίτερα ενδιαφέρον είναι να βρεθούν οι «σημαντικοί» παράγοντες κινδύνου για κάθε hedge fund, δηλαδή να εντοπιστούν οι καταλληλότερες επεξηγηματικές μεταβλητές που εξηγούν καλύτερα την μεταβλητότητα των αποδόσεων τους. Αυτό είναι συνήθως δύσκολο εξαιτίας των πολύπλοκων και

διαφορετικών στρατηγικών που ακολουθούν οι διαχειριστές αυτών των κεφαλαίων άλλα επίσης και από την έλλειψη γνώσης των δραστηριοτήτων τους.

Έχουν χρησιμοποιηθεί αρκετές μέθοδοι επιλογής μοντέλων στη γραμμική παλινδρόμηση όπως η stepwise regression αλλά και τα κριτήρια AIC και BIC είναι αρκετά διαδεδομένα για την επιλογή των μεταβλητών κινδύνου που θα επιλεγούν να συμπεριληφθούν στο μοντέλο.

Μέσω της γραμμικής παλινδρόμησης βρίσκουμε τη σχέση της αναμενόμενης τιμής του αμοιβαίου κεφαλαίου υψηλού κινδύνου με τις μεταβλητές κινδύνου κάτι που δεν είναι αρκετό γιατί, λόγω της πολύπλοκης φύσης τους, οι αποδόσεις τους ενδέχεται να παρουσιάζουν υψηλό βαθμό μη κανονικότητας, παχιές ουρές, ασυμμετρία και κύρτωση. Η παρουσία έστω και ενός από τα προηγούμενα χαρακτηριστικά κάνει τη κλασική μέθοδο παλινδρόμησης μη αποτελέσματική και τις εκτιμήσεις της μη ρεαλιστικές.

Επειδή σκοπός μας είναι να βρούμε τις επιπτώσεις που έχουν οι μεταβλητές κινδύνου σε ολόκληρη τη δεσμευμένη κατανομή των αμοιβαίων κεφαλαίων υψηλού κινδύνου ενδείκνυται η ποσοστημοριακή παλινδρόμηση. Η προσέγγιση αυτή θα μας αποκαλύψει πιθανές διαφορές, στις επιδράσεις που έχουν οι μεταβλητές κινδύνου πάνω στις τιμές των αμοιβαίων κεφαλαίων, σε διαφορετικά ποσοστημόρια. Επίσης κάποιες μεταβλητές οι οποίες στη γραμμική παλινδρόμηση δε θα θεωρούνταν σημαντικές και ως αποτέλεσμα θα ήταν εκτός μοντέλου εδώ μπορεί να είναι σημαντικές για τα υψηλά ή χαμηλά δεσμευμένα ποσοστημόρια των τιμών των αμοιβαίων κεφαλαίων. Έτσι έχουμε ως αποτέλεσμα να μπορούμε να συσχετίσουμε οποιαδήποτε μελλοντική επένδυση σε αμοιβαία κεφάλαια με συγκεκριμένες μεταβλητές.

Στη συνέχεια θα ασχοληθούμε με 4 αμοιβαία κεφάλαια υψηλού κινδύνου και θα βρούμε από 14 επεξηγηματικές μεταβλητές (risk factors) τις πιο «σημαντικές» σύμφωνα με τη γραμμική παλινδρόμηση αλλά και τις πιο «σημαντικές» σύμφωνα με τη ποσοστημοριακή παλινδρόμηση για τα ποσοστημόρια $(0.1, 0.25, 0.5, 0.75, 0.9)$. Η επιλογή του «καλύτερου» μοντέλου θα γίνει μέσω των κριτηρίων AIC και BIC. Σκοπός του συγκεκριμένου κεφαλαίου είναι να δούμε τις διαφορές της ποσοστημοριακής παλινδρόμησης με τη κλασική παλινδρόμηση και για αυτό δε θα σταθούμε στην οικονομική ανάλυση των αποτελεσμάτων μας αλλά στη στατιστική. Η ανάλυση έγινε σε γλώσσα προγραμματισμού MATLAB και όλοι οι κώδικες που χρησιμοποιήθηκαν παρουσιάζονται στο παράρτημα.

6.1 Περιγραφή δεδομένων

Τα 4 αμοιβαία κεφάλαια υψηλού κινδύνου με τα οποία θα ασχοληθούμε είναι τα: Convertible Arbitrage (CA), Equity Non-Hedge (ENH), Distressed Secu-

rities (DS), Merger Arbitrage (MA). Η στρατηγική του CA περιλαμβάνει τη ταυτόχρονη αγορά μετατρέψιμων χρεογράφων και τον δανεισμό μετοχών (άμεση πώληση τους και αγορά των τίτλων σε επόμενη χρονική φάση, όταν θα έχει μειωθεί η τιμή) προκειμένου να διασφαλιστούν τα κέρδη (short selling). Το αμοιβαίο κεφάλαιο ΕΝΗ επιτρέπει το δανεισμό χρημάτων με σκοπό την επένδυση τους (leverage). Το αμοιβαίο κεφάλαιο DS περιέχει χρεόγραφα αναγκαστικής εκποίησης και τέλος στο αμοιβαίο κεφάλαιο MA υπάρχουν μετοχές από εταιρίες προς συγχώνευση οι οποίες αγοράζονται και πουλούνται ταυτόχρονα. Η περίοδος στην οποία αναφέρονται οι αποδόσεις των παραπάνω αμοιβαίων κεφαλαίων που θα αναλύσουμε είναι από τον Απρίλιο του 1990 έως τον Δεκέμβριο του 2005. Η περίοδος αυτή περιλαμβάνει κάποια γεγονότα τα οποία επηρέασαν τις τιμές των αμοιβαίων κεφαλαίων με αποτέλεσμα να υπάρχει μεγάλη μεταβλητότητα στις αποδόσεις τους.

Στον πίνακα 6.1 παρουσιάζουμε κάποια περιγραφικά στατιστικά στοιχεία για τις αποδόσεις των αμοιβαίων κεφαλαίων υψηλού κινδύνου για τη συγκεκριμένη περίοδο. Παρατηρούμε στον πίνακα 6.1 οτι τα δεδομένα είναι ανομοιογενή. Υπάρχουν αμοιβαία κεφάλαια τα οποία έχουν υψηλές τιμές αποδόσεων (ΕΝΗ) και άλλα με χαμηλές (CA, MA). Επίσης η τυπική απόκλιση δείχνει τη διαφορά που υπάρχει στη μεταβλητότητα των αμοιβαίων κεφαλαίων. Παρατηρούμε οτι σε αμοιβαία κεφάλαια όπως το ΕΝΗ η μεταβλητότητα των τιμών τους είναι πολύ μεγαλύτερη από άλλα όπως το CA και επίσης οτι και τα 4 έχουν θετικές μέσες αποδόσεις. Επίσης, και τα 4 αμοιβαία κεφάλαια παρουσιάζουν αρνητική ασσυμετρία κάτι που σημαίνει οτι η ακραία πτώση των τιμών τους είναι πιο πιθανή από την ακραία αύξηση τους. Τα αμοιβαία κεφάλαια DS, MA παρουσιάζουν μεγάλη κύρτωση κάτι που υποδεικνύει παχιές ουρές. Όλα τα παραπάνω μας οδηγούν στο συμπέρασμα οτι τα αμοιβαία κεφάλαια υψηλού κινδύνου παρουσιάζουν υψηλό βαθμό μη κανονικότητας.

Οι επεξηγηματικές μεταβλητές από τις οποίες θα επιλέξουμε τις πιο «σημαντικές» είναι

1.RUS Ο χρηματιστηριακός δείκτης μεταβλητού επιτοκίου Russell 3000 (Russell 3000 equity index excess return).

2.RUS(−1) Η χρονική υστέρηση πρώτης τάξης του δείκτη Russell 3000 (Russell 3000 equity index excess return lagged once).

3.MXUS Ο διεθνής (εκτός ΗΠΑ) δείκτης κεφαλαίων της Morgan Stanley (Morgan Stanley Capital Idternational world excluding USA index excess return).

4.MEM Ο διεθνής δείκτης κεφαλαίων αναπτυσσόμενων αγορών της Morgan Stanley (Morgan Stanley Capital Idternational emerging markets index excess return).

Πίνακας 6.1: Στατιστικά στοιχεία για τα αμοιβαία κεφάλαια υψηλού κινδύνου.

A.K.	Μέση τιμή	Τυπ. απόκλ.	Κύρτωση	Ασσυμ.	25ο Ποσ.	50ο Ποσ.	75ο Ποσ.
CA	0.48	0.99	5.37	−1.18	0.03	0.68	1.14
ENH	1.00	4.05	3.63	−0.53	−1.60	1.70	3.53
DS	0.84	1.75	8.50	−0.68	−0.05	0.81	1.72
MA	0.50	1.08	13.38	−2.39	0.08	0.64	1.12

5.SMB Ο δείκτης όγκου των Fama and French (Fama and French's 'size').

6.HML Ο δείκτης αξίας αγοράς των Fama and French(Fama and French's 'book-to-market').

7.MOM Ο δείκτης ορμής του Carhart (Carhart's 'momentum' factor).

8.SBGC Ο δείκτης κρατικών ομολόγων και συλλογικών εντόκων γραμματείων των Salomon Brothers (Salomon Brothers world government and corporate bond index excess return).

9.SBWG Ο δείκτης κρατικών ομολόγων των Salomon Brothers (Salomon Brothers world government bond index excess return).

10.LHY Ο δείκτης υψηλών αποδόσεων Lehman (Lehman high yield index excess return).

11.DEFSPR Η διαφορά αποδόσεων ανάμεσα στα συλλογικά γραμμάτεια BAA και στα 10ετή ομόλογα (Difference between the yield on the BAA-rated corporate bonds and the 10-year bonds).

12.FRBI Ο σταθμισμένος δείκτης ανταγωνιστικότητας της Federal Reserve Bank (Federal Reserve Bank competitiveness weighted dollar-index excess return).

13.GSCI Ο δείκτης εμπορευμάτων της Goldman Sachs (Goldman Sachs commodity index excess return).

14.VIX Η μεταβολή στο δείκτη μεταβλητότητας S&P 500 (Change in S&P 500 implied volatility index).

6.2 Μοντέλα γραμμικής παλινδρόμησης και επιλογή μεταβλητών

Τα αποτελέσματα για κάθε ένα αμοιβαίο κεφάλαιο υψηλού κινδύνου συμπεριλαμβάνονται στον πίνακα 6.2. Εκτός από τα «καλύτερα» μοντέλα σύμφωνα με κάθε κριτήριο, υπάρχουν και άλλες δύο σειρές οι οποίες δείχνους τις εκτιμήσεις των παραμέτρων τους (συμπεριλαμβάνοντας και την $\hat{\beta}_0$) σύμφωνα με τη μέθοδο μέγιστης πιθανοφάνειας και τα τυπικά σφάλματα των εκτιμητών (standard errors).

Αρχικά παρατηρούμε οτι το κάθε αμοιβαίο κεφάλαιο μπορεί να έχει εντελώς διαφορετικές «σημαντικές» επεξηγηματικές μεταβλητές από τα άλλα κάτι που είναι αναμενόμενο επειδή οι στρατηγικές του κάθε κεφαλαίου διαφέρουν από τα υπόλοιπα. Επίσης, όσον αφορά τα καλύτερα μοντέλα γραμμικής παλινδρόμησης βλέπουμε οτι τα μοντέλα που έχουν επιλεγεί με βάση το κριτήριο BIC περιέχουν πάντα μεταβλητές που υπήρχαν και στα μοντέλα που επιλέχτηκαν με βαση το κριτήριο AIC και επίπλεον στα καλύτερα μοντέλα που έχουν επιλεγεί με το κριτήριο BIC υπάρχουν πάντα λιγότερες επεξηγηματικές μεταβλητές.

6.3 Μοντέλα ποσοστημοριακής παλινδρόμησης και επιλογή μεταβλητών

Τα αποτελέσματα για κάθε ένα αμοιβαίο κεφάλαιο υψηλού κινδύνου συμπεριλαμβάνονται στους πίνακες 6.3, 6.4, 6.5, 6.6. Σε όλους αυτούς τους πίνακες εκτός από τα «καλύτερα» μοντέλα σύμφωνα με κάθε κριτήριο και για κάθε ποσοστημόριο, υπάρχουν και άλλες δύο σειρές οι οποίες δείχνουν τις εκτιμήσεις μέγιστης πιθανοφάνειας των παραμέτρων (συμπεριλαμβάνοντας και την $\hat{\beta}_0$) και τα τυπικά σφάλματα τους (standard errors) τα οποία έχουν εκτιμηθεί με τη μέθοδο Bootstrap. Όσον αφορά τα καλύτερα μοντέλα ποσοστημοριακής παλινδρόμησης μπορούμε εύκολα να δούμε οτι σε κάθε αμοιβαίο κεφάλαιο υψηλού κινδύνου υπάρχουν σημαντικές διαφορές στο ποιες μεταβλητές κινδύνου είναι σημαντικές ανάλογα το ποσοστημόριο παλινδρόμησης. Κάποιες μεταβλητές που είναι σημαντικές για συγκεκριμένα ποσοστημόρια δεν είναι σίγουρο οτι θα είναι και για άλλα. Επίσης η επίδραση μίας επεξηγηματικής μεταβλητής πάνω στο αμοιβαίο κεφάλαιο δεν παραμένει απαραίτητα θετική ή αρνητική σε όλη τη δεσμευμένη κατανομή του. Πιο συγκεκριμένα παρατηρούμε οτι στις ουρές των δεσμευμένων κατανομών των αμοιβαίων κεφαλαίων (10ο και 90ο ποσοστημόριο) τα καλύτερα μοντέλα περιέχουν συνήθως περισσότερες μεταβλητές κινδύνου από όλα τα άλλα ποσοστημόρια. Ως αποτέλεσμα έχουμε αρκετές μεταβλητές κινδύνου που είναι σημαντικές για να εξηγήσουν χαμηλά ή υψηλά ποσοστημόρια να μην είναι

σημαντικές στα αντίστοιχα μοντέλα γραμμικής παλινδρόμησης. Χαρακτηριστικό παράδειγμα για τα προηγούμενα είναι το αμοιβαίο κεφάλαιο MA όπου στο 90ο ποσοστημόριο το καλύτερο μοντέλο χρειάζεται 10 από τις συνολικά 14 μεταβλητές κινδύνου ενώ στο αντίστοιχο μοντέλο γραμμικής παλινδρόμησης και στο 50ο ποσοστημόριο χρειάζεται μόνο 3. Μπορούμε επίσης να δούμε οτι οι εκτιμήσεις των παραμέτρων β_7 και β_{13} στο 10ο ποσοστημόριο είναι θετικές ενώ στο 90ο αρνητικές, κάτι που έχει ως αποτέλεσμα να «αλληλοεξουδετερώνεται» η επίδραση των αντίστοιχων επεξηγηματικών μεταβλητών πάνω στο αμοιβαίο κεφάλαιο MA κατά μέσο όρο (ή στη διάμεσο). Το ίδιο ακριβώς συμβαίνει με τις εκτιμήσεις των παραμέτρων β_9 και β_{11} στο αμοιβαίο κεφάλαιο CA (κριτήριο AIC) και την εκτίμηση της β_{11} στο αμοιβαίο κεφάλαιο ENH (κριτήριο AIC).

6.4 Ανάλυση και σύγκριση γραμμικών με ποσοστημοριακών μοντέλων παλινδρόμησης

Βλέποντας τους πίνακες 6.2 και 6.3, 6.4, 6.5, 6.6 παρατηρούμε οτι, αν και τα μοντέλα της γραμμικής παλινδρόμησης ποτέ δε συμπίπτουν με αυτά της ποσοστημοριακής παλινδρόμησης διαμέσου, έχουν πάντα αρκετές μεταβλητές κινδύνου ίδιες και κάποιες φορές διαφέρουν κατά μία μόνο μεταβλητή όπως στις περιπτώσεις των αμοιβαίων κεφαλαίων υψηλού κινδύνου DS, ENH σύμφωνα με το κριτήριο BIC. Επίσης οι εκτιμήσεις των παραμέτρων των μοντέλων είναι αρκετά διαφορετικές. Αυτό οφείλεται στην έλλειψη συμμετρίας και γενικά κανονικότητας των αποδόσεων των hedge funds.

Αν τώρα λάβουμε υπόψη όλη την ανάλυση ποσοστημοριακής παλινδρόμησης, δηλαδή τα μοντέλα για όλα τα ποσοστημόρια, η σύγκριση της με τη μέθοδο γραμμικής παλινδρόμησης δίνει πολύ ενδιαφέροντα συμπεράσματα. Τα αποτελέσματα που παρουσιάζονται στην παράγραφο 6.3 δίνουν σαφείς ενδείξεις ότι η μέθοδος ποσοστημοριακής παλινδρόμησης υπερτερεί για την ανάλυση των αποδόσεων των hedge funds. Φαίνεται ότι γενικά οι μεταβλητές που επηρεάζουν τις αποδόσεις είναι περισσότερες από αυτές που ανιχνεύει η γραμμική παλινδρόμηση, η οποία βρίσκει τις μεταβλητές κινδύνου που είναι στατιστικά σημαντικές κατά μέσο όρο. Αντίθετα, η ποσοστημοριακή παλινδρόμηση, μοντελοποιώντας όλη τη δεσμευμένη κατανομή των αποδόσεων, είναι σε θέση να εντοπίσει ενδεχόμενες επιδράσεις επεξηγηματικών μεταβλητών μόνο στις ουρές της κατανομής ή σε συγκεκριμένα ποσοστημόριά της. Τα πλεονεκτήματα της μεθόδου έναντι της γραμμικής παλινδρόμησης είναι πιο εμφανή σε περιπτώσεις όπου υπάρχουν αποκλίσεις των δεδομένων από την κανονικότητα, όπως στην εφαρμογή μας σε αποδόσεις αμοιβαίων κεφαλαίων υψηλού κινδύνου.

Παρατηρούμε επίσης οτι είτε αναφερόμαστε στα καλύτερα μοντέλα γραμμικής παλινδρόμησης είτε σε αυτά της ποσοστημοριακής, αυτά που έχουν τις λιγότερες μεταβλητές είναι τα μοντέλα που επιλέχτηκαν με βάση τη τιμή του κριτηρίου BIC. Αυτό δε γίνεται τυχαία αλλά επειδή όπως έχουμε ήδη τονίσει, το κριτήριο BIC τιμωρεί την πολυπλοκότητα του μοντέλου και άρα είναι λογικό να έχουν λιγότερες μεταβλητές αυτά τα μοντέλα.

Πίνακας 6.2: Μοντέλα γραμμικής παλινδρόμησης για τα αμοιβαία κεφάλαια

Α.Κ.	Κριτήριο					Επεξηγ. Μεταβλητές						
CA	AIC		2	4	5	8	10					
	$\hat{\beta}$	0.0042	0.0494	0.0347	0.0331	0.1370	0.0493					
	s.e.	0.0007	0.0160	0.0106	0.0213	0.0531	0.0226					
	BIC		2	4	8	10						
	$\hat{\beta}$	0.0043	0.0539	0.0387	0.1232	0.0575						
	s.e.	0.0007	0.0158	0.0103	0.0525	0.0220						
ENH	AIC		1	2	4	5	6	9	10	12	13	14
	$\hat{\beta}$	0.0063	0.6920	0.0448	0.0704	0.4370	-0.1045	0.0812	0.0640	0.1613	0.0271	0.0780
	s.e.	0.0008	0.0332	0.0212	0.0175	0.0294	0.0234	0.0494	0.0288	0.0796	0.0142	0.0316
	BIC		1	4	5	6	10	14				
	$\hat{\beta}$	0.0066	0.6969	0.0673	0.4548	-0.0950	0.0720	0.0948				
	s.e.	0.0008	0.0335	0.0174	0.0284	0.0230	0.0289	0.0307				
DS	AIC		2	4	5	8	10	11	12	14		
	$\hat{\beta}$	0.0074	0.0924	0.0898	0.0956	0.2301	0.0823	-2.7857	0.1513	-0.0556		
	s.e.	0.0009	0.0229	0.0168	0.0287	0.0748	0.0306	0.8219	0.0770	0.0292		
	BIC		2	4	5	8	10	11				
	$\hat{\beta}$	0.0073	0.0787	0.0988	0.0952	0.2161	0.0918	-3.1576				
	s.e.	0.0009	0.0225	0.0145	0.0292	0.0759	0.0309	0.8192				
MA	AIC		1	2	4	5	6	7	11			
	$\hat{\beta}$	0.0037	0.0996	0.0653	0.0405	0.0732	0.0516	0.0298	0.8227			
	s.e.	0.0006	0.0221	0.0164	0.0136	0.0233	0.0221	0.0135	0.5584			
	BIC		1	2	4							
	$\hat{\beta}$	0.0043	0.0787	0.0681	0.0460							
	s.e.	0.0006	0.0210	0.0152	0.0130							

Πίνακας 6.3: Μοντέλα ποσοστημοριακής παλινδρόμησης για τα αμοιβαία κεφάλαια:Μέρος Α

Α.Κ.	Ποσοστ.	Κριτήριο	Επεξηγ. Μεταβλητές											
CA	0.10	AIC		2	3	4	5	6	8	9	10	11		
		$\hat{\beta}$	-0.0083	0.0960	-0.0732	0.0981	-0.0703	-0.0609	0.2603	0.1015	0.0716	-2.0879		
		s.e.	0.0018	0.0480	0.0546	0.0382	0.0727	0.0682	0.1557	0.1181	0.0542	2.0486		
	0.25			1	2	3	4	8	10	11	13	14		
		$\hat{\beta}$	-0.0001	0.0673	0.0759	-0.0405	0.0268	0.1199	0.0881	1.1586	0.0220	0.0315		
		s.e.	0.0013	0.0384	0.0279	0.0406	0.0253	0.0848	0.0444	1.5808	0.0163	0.0388		
	0.50			1	2	5	10	11						
		$\hat{\beta}$	0.0059	0.0470	0.0439	0.0282	0.0972	2.3189						
		s.e.	0.0006	0.0130	0.0141	0.0182	0.0364	0.6483						
	0.75			2	4	5	6	7	8	11	13			
		$\hat{\beta}$	0.0099	0.0297	0.0244	0.0293	-0.0325	-0.0431	0.1298	1.7262	0.0147			
		s.e.	0.0006	0.0179	0.0100	0.0223	0.02221	0.0159	0.0399	0.5544	0.0130			
	0.90			2	4	5	6	7	8	9	10	11	12	13
		$\hat{\beta}$	0.0125	0.0470	0.0231	0.0257	-0.0184	-0.0169	0.1072	-0.0507	0.0144	1.3574	-0.0619	0.0224
		s.e.	0.0011	0.0306	0.0202	0.0374	0.0315	0.0157	0.0777	0.0612	0.0756	0.8819	0.0888	0.0207
	0.10	BIC		2	3	4	8	11						
		$\hat{\beta}$	-0.0090	0.0763	-0.0515	0.1004	0.3214	-1.5298						
		s.e.	0.0017	0.0502	0.0417	0.0272	0.1074	1.7396						
	0.25			2	8	10								
		$\hat{\beta}$	0.0001	0.0801	0.1398	0.1368								
		s.e.	0.0009	0.0237	0.0516	0.0453								
	0.50			1	2	5	10	11						
		$\hat{\beta}$	0.0059	0.0470	0.0439	0.0282	0.0972	2.3189						
		s.e.	0.0005	0.0119	0.0130	0.0174	0.0353	0.7283						
	0.75			2	4	5	6	7	8	11				
		$\hat{\beta}$	0.0095	0.0360	0.0246	0.0346	-0.0343	-0.1454	0.1354	1.7523				
		s.e.	0.0006	0.0165	0.0112	0.0191	0.0215	0.0171	0.0372	0.5899				
	0.90			2	4	5	10	11						
		$\hat{\beta}$	0.0135	0.0469	0.0254	0.0434	0.0225	1.4495						
		s.e.	0.0007	0.0281	0.0183	0.0185	0.0654	0.6323						

Πίνακας 6.4: Μοντέλα ποσοστημοριακής παλινδρόμησης για τα αμοιβαία κεφάλαια:Μέρος Β

Α.Κ.	Ποσοστ.	Κριτήριο	Επεξηγ. Μεταβλητές												
ΕΝΗ	0.10	AIC		1	2	4	5	6	7	9	10	11	13		
		β	-0.0065	0.6239	0.0668	0.0988	0.3765	-0.1062	0.0137	0.1657	0.1366	-1.0523	0.0291		
		s.e.	0.0014	0.0337	0.0260	0.0254	0.0357	0.0339	0.0183	0.0568	0.0544	1.1602	0.0143		
	0.25			1	2	4	5	6	9	10	13	14			
		$\hat{\beta}$	-0.0019	0.6191	0.0493	0.0896	0.3838	-0.0997	0.1176	0.0803	-0.2368	0.0257			
		s.e.	0.0012	0.0546	0.0348	0.0360	0.0553	0.0394	0.0809	0.0360	0.0177	0.0503			
	0.50			1	3	4	5	6	7	8	14				
		$\hat{\beta}$	0.0061	0.6358	0.0405	0.0670	0.4986	-0.1042	-0.0030	0.1146	0.0855				
		s.e.	0.0012	0.0530	0.0399	0.0272	0.0477	0.0376	0.0194	0.0982	0.0424				
	0.75			1	3	4	5	6	7	11	12	14			
		$\hat{\beta}$	0.0130	0.6677	0.0266	0.0841	0.4943	-0.1164	-0.0218	1.0484	0.2198	0.0768			
		s.e.	0.0011	0.0532	0.0395	0.0314	0.0401	0.0404	0.0271	1.0466	0.0943	0.0644			
	0.90			1	2	3	4	5	6	7	11	12	13	14	
		$\hat{\beta}$	0.0194	0.7972	0.0404	0.0289	0.0333	0.4596	-0.0906	-0.0313	0.9426	0.2875	0.0559	0.1717	
		s.e.	0.0019	0.1183	0.0507	0.0443	0.0358	0.0725	0.0616	0.0381	1.3739	0.1597	0.0375	0.1284	
	0.10	BIC		1	2	4	5	6	9	10	11	13			
		β	-0.0066	0.6241	0.0667	0.0988	0.3670	-0.1183	0.1828	0.1097	-1.4624	0.0337			
		s.e.	0.0014	0.0290	0.0268	0.0231	0.0350	0.0256	0.0597	0.0469	1.2791	0.0143			
	0.25			1	4	5	6	9	10	14					
		$\hat{\beta}$	-0.0012	0.6604	0.0849	0.4333	-0.0803	0.1154	0.0647	0.0622					
		s.e.	0.0012	0.0475	0.0330	0.0480	0.0405	0.0579	0.0392	0.0342					
	0.50			1	4	5	6	14							
		$\hat{\beta}$	0.0066	0.7335	0.0653	0.4626	-0.0779	0.1339							
		s.e.	0.0009	0.0460	0.0220	0.0497	0.0369	0.0388							
	0.75			1	4	5	6	14							
		$\hat{\beta}$	0.0126	0.7033	0.0744	0.5043	-0.1011	0.1060							
		s.e.	0.0010	0.0456	0.0191	0.0426	0.0380	0.0760							
	0.90			1	4	5	6	13	14						
		$\hat{\beta}$	0.0191	0.8209	0.0194	0.5412	-0.0399	0.0511	0.1668						
		s.e.	0.0014	0.0852	0.0360	0.0562	0.0626	0.0346	0.1172						

Πίνακας 6.5: Μοντέλα ποσοστημοριακής παλινδρόμησης για τα αμοιβαία κεφάλαια:Μέρος Γ

Α.Κ.	Ποσοστ.	Κριτήριο	Επεξηγ. Μεταβλητές											
DS	0.10	AIC		1	2	4	7	8	10	11	12	13		
		$\hat{\beta}$	-0.0049	0.0565	0.1012	0.0865	0.0284	0.1460	0.1738	-2.4498	0.1634	0.0548		
		s.e.	0.0019	0.0410	0.0386	0.0466	0.0331	0.1148	0.0737	1.3073	0.1219	0.0291		
	0.25		1	2	4	5	7	8	10	11	12			
		$\hat{\beta}$	0.0013	0.0517	0.0885	0.0422	0.0742	0.0172	0.1879	0.1361	-4.1552	0.0601		
		s.e.	0.0011	0.0298	0.0269	0.0255	0.0331	0.0186	0.1023	0.0681	1.1223	0.0923		
	0.50		2	3	4	5	8	10	11	13				
		$\hat{\beta}$	0.0066	0.0671	0.0392	0.0725	0.0459	0.2820	0.0763	-4.0166	-0.0375			
		s.e.	0.0012	0.0356	0.0332	0.0261	0.0636	0.1316	0.0770	1.3440	0.0289			
	0.75		1	2	4	5	6	7	8	9	10	11	12	
		$\hat{\beta}$	0.0132	0.0580	0.0663	0.0907	0.2061	0.0664	0.0494	0.1487	0.1209	0.0567	-3.0243	0.2432
		s.e.	0.0014	0.0372	0.0414	0.0213	0.0524	0.0512	0.0369	0.1273	0.0858	0.0508	1.4736	0.1377
	0.90		2	3	4	5	6	8	9	10	11	12	14	
		$\hat{\beta}$	0.0209	0.0227	-0.0687	0.1291	0.2505	0.1051	0.2439	0.0795	0.0914	-2.0042	0.3356	-0.0300
		s.e.	0.0021	0.0575	0.0684	0.0374	0.0710	0.0450	0.1893	0.1125	0.0714	2.3445	0.1495	0.0563
	0.10	BIC		1	2	4	8	10	11	12	13			
		$\hat{\beta}$	-0.0045	0.0721	0.0912	0.0687	0.1782	0.1561	-2.5255	0.1923	0.0371			
		s.e.	0.0018	0.0384	0.0391	0.0456	0.0902	0.0641	1.3520	0.1249	0.0261			
	0.25			1	2	4	5	8	10	11				
		$\hat{\beta}$	0.0014	0.0336	0.0851	0.0479	0.0687	0.2319	0.1242	-4.4157				
		s.e.	0.0010	0.0388	0.0274	0.0222	0.0330	0.0902	0.0648	1.1975				
	0.50			2	4	8	10	11						
		$\hat{\beta}$	0.0059	0.0903	0.0847	0.2298	0.1082	-4.3613						
		s.e.	0.0014	0.0298	0.0174	0.1233	0.0753	1.3483						
	0.75			2	4	5	8	10	11					
		$\hat{\beta}$	0.0142	0.0539	0.0861	0.1690	0.2461	0.0731	-3.0941					
		s.e.	0.0011	0.0366	0.0156	0.0492	0.0891	0.0632	1.2087					
	0.90			4	5	6	8	10	12					
		$\hat{\beta}$	0.0202	0.1209	0.2548	0.1302	0.2649	0.0703	0.3399					
		s.e.	0.0021	0.0282	0.0413	0.0419	0.1564	0.0580	0.1753					

Πίνακας 6.6: Μοντέλα ποσοστημοριακής παλινδρόμησης για τα αμοιβαία κεφάλαια:Μέρος Δ

Α.Κ.	Ποσοστ.	Κριτήριο					Επεξηγ. Μεταβλητές						
MA	0.10	AIC		1	2	3	4	5	6	7	11	12	13
		$\hat{\beta}$	-0.0070	0.1146	0.0915	0.0426	0.0604	0.0574	0.0378	0.0344	2.1157	0.1076	0.0369
		s.e.	0.0016	0.0453	0.0276	0.0501	0.0382	0.0322	0.0481	0.0245	1.3575	0.0944	0.0289
	0.25			1	2	4	5	7	8	12	14		
		$\hat{\beta}$	-0.0007	0.0916	0.0453	0.0319	0.0648	0.0182	0.1471	0.1189	-0.0411		
		s.e.	0.0009	0.0306	0.0329	0.0222	0.0319	0.0148	0.0654	0.0892	0.0417		
	0.50			1	2	3	4	5	7	9	11	12	
		$\hat{\beta}$	0.0051	0.0761	0.0671	-0.0379	0.0491	0.0544	0.0222	0.1074	1.3755	0.1353	
		s.e.	0.0007	0.0320	0.0148	0.0231	0.0183	0.0206	0.0119	0.0452	1.0334	0.0539	
	0.75			1	2	4	5	6	9	11	12	13	14
		$\hat{\beta}$	0.0096	0.0562	0.0435	0.0214	0.0738	0.0182	0.0689	0.8799	0.1802	-0.0225	-0.0230
		s.e.	0.0007	0.0235	0.0223	0.0150	0.0257	0.0256	0.0383	0.7499	0.0621	0.0138	0.0286
	0.90			1	2	5	7	8	10	11	12	13	14
		$\hat{\beta}$	0.0128	0.0449	0.0327	0.0713	-0.0303	0.1028	-0.0539	1.0319	0.1585	-0.0290	-0.0396
		s.e.	0.0007	0.0177	0.0234	0.0272	0.0179	0.0634	0.0351	0.6407	0.0572	0.0132	0.0288
	0.10	BIC		1	2	4	7	13					
		$\hat{\beta}$	-0.0059	0.1187	0.0890	0.0595	0.0275	0.0365					
		s.e.	0.0019	0.0450	0.0369	0.0390	0.0238	0.0281					
	0.25			1	2	4							
		$\hat{\beta}$	-0.0004	0.0696	0.0499	0.0507							
		s.e.	0.0008	0.0276	0.0307	0.0198							
	0.50			1	2	5							
		$\hat{\beta}$	0.0054	0.0837	0.0533	0.0690							
		s.e.	0.0007	0.0246	0.0178	0.0139							
	0.75			1	2	5	12	13					
		$\hat{\beta}$	0.0100	0.0699	0.0254	0.0772	0.0790	-0.0231					
		s.e.	0.0006	0.0172	0.0156	0.0259	0.0334	0.0089					
	0.90			1	2	5	7	8	10	11	12	13	14
		$\hat{\beta}$	0.0128	0.0449	0.0327	0.0713	-0.0303	0.1028	-0.0539	1.0319	0.1585	-0.0290	-0.0396
		s.e.	0.0007	0.0199	0.0212	0.0242	0.0155	0.0549	0.0336	0.7342	0.0615	0.0145	0.0286

Παράρτημα Α΄

Κώδικας Matlab

Ignorance is bliss. Oedipus ruined a great sex life by asking too many questions.

The Colbert Report
Stephen Colbert

Α΄.1 Κώδικας Matlab για περιγραφικά στατιστικά στοιχεία

Έχοντας θέσει ως y τις τιμές του αμοιβαίου κεφαλαίου υψηλού κινδύνου αρκεί να πληκτρολογήσουμε τα παρακάτω έτσι ωστε να πάρουμε τα περιγραφικά στοιχεία του πίνακα 6.1

```
mean(y)
std(y)
prctile(y,[25 50 75])
skewness(y)
kurtosis(y)
```

Α΄.2 Κώδικας Matlab για γραμμική παλινδρόμηση

Σκοπός μας είναι να επίλεξουμε τα «καλύτερα» μοντέλα γραμμικής παλινδρόμησης μέσω των κριτηρίων AIC, BIC . Επειδή έχουμε 14 μεταβλητές κινδύνου

όλα τα δυνατά μοντέλα είναι $2^{14} = 16384$. Από αυτά θα επιλέξουμε 2 μοντέλα, ένα για κάθε κριτήριο. Αρχικά θα πρέπει να δημιουργήσουμε ένα πίνακα στη MATLAB μέσω του οποίου να επιλέγουμε κάθε δυνατό μοντέλο (από αυτό που θα περιέχει μόνο την πρώτη μεταβλητή κινδύνου εως και αυτό που θα τις περιέχει όλες).

```
N=14;
all_models=zeros(1,N);
for i=1:N
    models_i=nchoosek(1:14,i);
    [a,b]=size(models_i);
    models_i=[models_i zeros(a,N-b)];
    all_models=[all_models; models_i];
end

save all_models all_models
```

Ως αποτέλεσμα των παραπάνω είναι να δημιουργηθεί ο πίνακας

$$all\_models = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 2 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 14 & 0 & 0 & 0 & \dots & 0 \\ 1 & 2 & 0 & 0 & \dots & 0 \\ 1 & 3 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 13 & 14 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2 & 3 & 4 & \dots & 14 \end{pmatrix}$$

ο οποίος έχει 16384 γραμμές και 14 στήλες όπου κάθε μία στήλη αντιπροσωπεύει τις μεταβλητές κινδύνου και κάθε γραμμή το κάθε δυνατό μοντέλο.

Μέσω της MATLAB θα υπολογίσουμε τις τιμές των κριτηρίων AIC, BIC και έπειτα θα επιλέξουμε τα μοντέλα αυτά με τις ελάχιστες τιμές.

```
function [minaic,minbic,B]=minaibi(y,X)
load all_models
load datafactors_hfrci.txt % Matrix that contains the risk factors
for i=1:size(all_models14,1)
    model=all_models14(i,:);
    model(model==0)=[];
```

```
    X=datafactors_hfrci(:,model);
    [n,k]=size(X);
    s=regstats(y,X,'linear'); % Linear regression
    AIC=n*log(s.r'*s.r)+2*(k+1); % Computing AIC, BIC
    BIC=n*log(s.r'*s.r)+log(n)*(k+1);
    B(i,:)=[AIC BIC];
end
minaic=min(B(:,1)); % Computing  the min AIC, BIC
minbic=min(B(:,2));
end
```

Ο τύπος που χρησιμοποιούμε για την εύρεση του AIC βασίζεται στον υπολογισμό της λογαριθμικής πιθανοφάνειας ως μια σταθερά κοινή για όλα τα μοντέλα. Έχουμε

$$L(\theta) = \big(\frac{1}{2\pi\sigma^2}\big)^{n/2} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\}$$

και άρα ισχύει οτι

$$\begin{aligned} l(\theta) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma^2}) - \frac{1}{2\hat{\sigma^2}}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2\} \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma^2}) - \frac{1}{2\hat{\sigma^2}}n\hat{\sigma^2} \\ &= -\frac{n}{2}\log(\hat{\sigma^2}) + c. \end{aligned}$$

Τέλος, έχουμε οτι

$$\text{AIC} = n\log(\hat{\sigma^2}) + 2k.$$

Αντίστοιχα βρίσκουμε και τις τιμές του κριτηρίου BIC. Η προηγούμενη συνάρτηση μας επιστρέφει ένα πίνακα B ο οποίος έχει 16384 γραμμές και 2 στήλες όπου στη πρώτη στήλη αποθηκεύουμε τη τιμή του κριτηρίου AIC ενώ στη δεύτερη τη τιμή του κριτηρίου BIC για κάθε ένα μοντέλο ξεχωριστά. Επίσης η συνάρτηση επιστρέφει τις ελάχιστες τιμές των κριτηρίων. Τέλος, για να βρούμε σε ποια σειρά του πίνακα B, και ως αποτέλεσμα σε ποιο μοντέλο, αντιστοιχούν οι δύο τιμές που έχουμε ήδη βρει πληκτρολογούμε

```
[val1 ind1]=min(B(:,1)); % Value of minaic and index of that model
[val2 ind2]=min(B(:,2)); % Value of minbic and index of that model
allmodels(ind1,:) %Model with minimum AIC
allmodels(ind2,:) %Model with minimum BIC
```

Τέλος, οι εντολές που χρειάζονται για την εύρεση των εκτιμητών της γραμμικής παλινδρόμησης και των τυπικών σφαλμάτων τους, για το αμοιβαίο κεφάλαιο CA σύμφωνα με το κριτήριο AIC είναι

```
b1=[2 4 5 8 10]; %Best model factors
X=datafactors_hfrci(:,b1);
s1=regstats(y,X,'linear');
s1.beta %Regression coefficients
sqrt(diag(s1.covb)) %Standard errors
```

Α΄.3 Κώδικας Matlab για ποσοστημοριακή παλινδρόμηση

Σκοπός μας τώρα είναι να επίλεξουμε τα «καλύτερα» μοντέλα ποσοστημοριακής παλινδρόμησης μέσω των κριτηρίων AIC, BIC . Επειδή στη συνάρτηση πιθανοφάνειας υπάρχει η συνάρτηση ελέγχου (check function) πρέπει να δημιουργήσουμε μία συνάρτηση που να την υπολογίζει. Αυτή τη δουλειά τη κάνει η παρακάτω συνάρτηση

```
function sum = check(y,X,betaq,p)
sum=0;
qq=y-X*betaq;
for i=1:size(qq,1)
    if (qq(i)>=0)
       sum=sum+p*qq(i);
    elseif (qq(i)<0)
        sum=sum+(p-1)*qq(i);
    end
end
```

η οποία δέχεται ως ορίσματα εκτός από τις τιμές του αμοιβαίου κεφαλαίου (y) και τις τιμές των μεταβλητών ρίσκου (X), τους συντελεστές της ποσοστημοριακής παλινδρόμησης β_p και το ποσοστημόριο p.

Οι συντελεστές της ποσοστημοριακής παλινδρόμησης θα βρεθούν από τη συνάρτηση

```
function b = rq(X, y, p)
% Construct the dual problem of quantile regression
% Solve it with lp_fnm
%
%
```

```
[m n] = size(X);
u = ones(m, 1);
a = (1 - p) .* u;
b = -lp_fnm(X', -y', X' * a, u, a)';

function y = lp_fnm(A, c, b, u, x)
% Solve a linear program by the interior point method:
% min(c * u), s.t. A * x = b and 0 < x < u
% An initial feasible solution has to be provided as x
%
% Function lp_fnm of Daniel Morillo & Roger Koenker
% Translated from Ox to Matlab by Paul Eilers 1999
% Modified by Roger Koenker 2000--
% More changes by Paul Eilers 2004


% Set some constants
beta = 0.9995;
small = 1e-5;
max_it = 50;
[m n] = size(A);

% Generate inital feasible point
s = u - x;
y = (A' \  c')';
r = c - y * A;
r = r + 0.001 * (r == 0);    % PE 2004
z = r .* (r > 0);
w = z - r;
gap = c * x - y * b + w * u;

% Start iterations
it = 0;
while (gap) > small & it < max_it
    it = it + 1;

    %   Compute affine step
    q = 1 ./ (z' ./ x + w' ./ s);
    r = z - w;
    Q = spdiags(sqrt(q), 0, n, n);
    AQ = A * Q;          % PE 2004
```

```
rhs = Q * r';        % "
dy = (AQ' \ rhs)';   % "
dx = q .* (dy * A - r)';
ds = -dx;
dz = -z .* (1 + dx ./ x)';
dw = -w .* (1 + ds ./ s)';

% Compute maximum allowable step lengths
fx = bound(x, dx);
fs = bound(s, ds);
fw = bound(w, dw);
fz = bound(z, dz);
fp = min(fx, fs);
fd = min(fw, fz);
fp = min(min(beta * fp), 1);
fd = min(min(beta * fd), 1);

% If full step is feasible, take it. Otherwise modify it
if min(fp, fd) < 1

    % Update mu
    mu = z * x + w * s;
    g = (z + fd * dz) * (x + fp * dx) + (w + fd * dw) * (s + fp * ds);
    mu = mu * (g / mu) ^3 / ( 2* n);

    % Compute modified step
    dxdz = dx .* dz';
    dsdw = ds .* dw';
    xinv = 1 ./ x;
    sinv = 1 ./ s;
    xi = mu * (xinv - sinv);
    rhs = rhs + Q * (dxdz - dsdw - xi);
    dy = (AQ' \ rhs)';
    dx = q .* (A' * dy' + xi - r' -dxdz + dsdw);
    ds = -dx;
    dz = mu * xinv' - z - xinv' .* z .* dx' - dxdz';
    dw = mu * sinv' - w - sinv' .* w .* ds' - dsdw';

    % Compute maximum allowable step lengths
    fx = bound(x, dx);
    fs = bound(s, ds);
```

```
        fw = bound(w, dw);
        fz = bound(z, dz);
        fp = min(fx, fs);
        fd = min(fw, fz);
        fp = min(min(beta * fp), 1);
        fd = min(min(beta * fd), 1);

    end

    % Take the step
    x = x + fp * dx;
    s = s + fp * ds;
    y = y + fd * dy;
    w = w + fd * dw;
    z = z + fd * dz;
    gap = c * x - y * b + w * u;
    %disp(gap);
end

function b = bound(x, dx)
% Fill vector with allowed step lengths
% Support function for lp_fnm
b = 1e20 + 0 * x;
f = find(dx < 0);
b(f) = -x(f) ./ dx(f);
```

Για τον υπολογισμό των τιμών των κριτηρίων AIC, BIC και την επιλογή των μοντέλων με τις ελάχιστες τιμές θα χρησιμοποιήσουμε τη παρακάτω συνάρτηση.

```
function [minaicq,minbicq,Q]=minaibiq(y,X,p)
load all_models
load datafactors_hfrci.txt % Matrix that contains the risk factors
for i=1:16384
    model=all_models(i,:);
    model(model==0)=[];
    X=datafactors_hfrci(:,model);
    [n,k]=size(X);
    X=[ones(n,1) X];
    betaq=rq(X, y, p);    %Quantile regression function
    ch=check(y,X,betaq,p);
    AIC=2*n*log(ch)-2*n*log(p*(1-p))+2*n-2*n*log(n)+2*(k+1);
    BIC=2*n*log(ch)-2*n*log(p*(1-p))+2*n-2*n*log(n)+log(n)*(k+1);
```

```
    Q(i,:)=[AIC BIC];
end
minaicq=min(Q(:,1)); % Computing  the min AIC, BIC
minbicq=min(Q(:,2));
end
```

Ο τύπος που χρησιμοποιούμε για την εύρεση του AIC ισχύει γιατί

$$L(y_i|\beta_p, \sigma_p, x_1, \ldots, x_n) = (\frac{p(1-p)}{\sigma_p})^n exp\{-\frac{1}{\sigma_p}\sum_{i=1}^{n}\rho_p(y_i - \beta_p x_i)\}$$

και άρα ισχύει οτι

$$\begin{aligned} l(\theta) &= n\log(p(1-p)) - n\log(\hat{\sigma_p}) - \frac{1}{\hat{\sigma_p}}\sum_{i=1}^{n}\rho_p(y_i - \hat{\beta}_p x_i) \\ &= \ldots \\ &= n\log(p(1-p)) - n\log(\sum_{i=1}^{n}\rho_p(y_i - \hat{\beta}_p x_i)) + n\log(n) - n. \end{aligned}$$

Τέλος, έχουμε οτι

$$\text{AIC} = 2n\log(\sum_{i=1}^{n}\rho_p(y_i - \beta_p x_i)) - 2n\log(p(1-p)) + 2n - 2n\log(n) + 2k.$$

Αντίστοιχα βρίσκουμε και τις τιμές του κριτηρίου BIC. Η προηγούμενη συνάρτηση μας επιστρέφει ένα πίνακα Q ο οποίος έχει 16384 γραμμές και 2 στήλες όπου στη πρώτη στήλη αποθηκεύουμε τη τιμή του κριτηρίου AIC ενώ στη δεύτερη τη τιμή του κριτηρίου BIC για κάθε ένα μοντέλο ξεχωριστά. Επίσης η συνάρτηση επιστρέφει τις ελάχιστες τιμές των κριτηρίων. Τέλος, για να βρούμε σε ποια σειρά του πίνακα Q, και ως αποτέλεσμα σε ποιο μοντέλο, αντιστοιχούν οι δύο τιμές που έχουμε ήδη βρει πληκτρολογούμε

```
[val1 ind1]=min(Q(:,1)); % Value of minaic and index of that model
[val2 ind2]=min(Q(:,2)); % Value of minbic and index of that model
allmodels(ind1,:) %Model with minimum AIC
allmodels(ind2,:) %Model with minimum BIC
```

Τα τυπικά σφάλματα των εκτιμητών της ποσοστημοριακής παλινδρόμησης θα τα πάρουμε χρησιμοποιώντας τη μέθοδο Bootstrap με τη βοήθεια της παρακάτω συνάρτησης

```
function [bootestimates,covmat_boot]=b_bootstrap(F,data,p,b_lsq,B,m);

T=length(data);
bootestimates=zeros(length(b_lsq),B);
for bb=1:B
    bootsample=unidrnd(T,1,m);
    boot_factors=F(bootsample,:);
    boot_data=data(bootsample);
    b_boot=rq(boot_factors,boot_data,p);
    bootestimates(:,bb)=[b_boot];

end
mles=repmat([b_lsq],[1,B]);
covmat_boot=(m/T)*(bootestimates-mles)*(bootestimates'-mles')/B;
```

η οποία δέχεται ως ορίσματα τον πίνακα που περιέχει τις «καλύτερες» μεταβλητές του μοντέλου, τις τιμές y του αμοιβαίου κεφαλαίου, το ποσοστημόριο p, τα $\hat{\beta}_p$, τον αριθμό επαναλήψεων B και τέλος το πλήθος m των τιμών του αμοιβαίου κεφαλαίου.

Τέλος, οι εντολές που χρειάζονται για την εύρεση των εκτιμητών της γραμμικής παλινδρόμησης και των τυπικών σφαλμάτων τους, για το αμοιβαίο κεφάλαιο CA σύμφωνα με το κριτήριο AIC είναι

```
b1q=[2 3 4 5 6 8 9 10 11]
X=datafactors_hfrci(:,b1q);
[n k]=size(X);
X=[ones(n,1) X];
beta1q=rq(X,y,0.1);
[bootestimates,covmat_boot]=b_bootstrap(X,y,0.1,beta1q,100,189);
sqrt(diag(covmat_boot))
```