Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

Μικτά Μοντέλα και Παλινδρόμηση Ποσοστιαίων Σημείων

Ηλίας Κ. Γκανέτσος

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Μάιος 2007

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- ……. (Επιβλέπων)
- ……..
- ……..

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

Mixed Models and Quantile Regression

By

Ilias K. Gkanetsos

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
May 2007

Στον Αποστόλη

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά όλους εκείνους που με τον τρόπο τους ο καθένας με βοήθησαν να ολοκληρώσω την παρούσα διπλωματική εργασία, η οποία αποτελεί μέρος των σπουδών μου για την απόκτηση του Μεταπτυχιακού Διπλώματος Εφαρμοσμένης Στατιστικής. Τις μεγαλύτερες ευχαριστίες οφείλω στον Λέκτορα του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης, κ. Γεώργιο Πιτσέλη, που με τη διαρκή του στήριξη και καθοδήγηση συνέβαλε στο να ολοκληρωθεί επιτυχώς η συγγραφή της εργασίας. Επίσης θα ήθελα να ευχαριστήσω την Επίκουρη Καθηγήτρια κα Μαρία Κατέρη και τον Επίκουρο Καθηγητή κ. Κωνσταντίνο Πολίτη που συμμετείχαν στην τριμελή επιτροπή, αλλά και τον Καθηγητή του Πανεπιστημίου του Southampton κ. Νίκο Τζαβίδη για το πολύτιμο υλικό που μου παρείχε για την ολοκλήρωση ενός σημαντικού μέρους της διπλωματικής.

Επίσης επιθυμώ να ευχαριστήσω την οικογένεια μου για την εμπιστοσύνη και την υποστήριξη που μου παρείχε όλο αυτό το διάστημα των 2,5 ετών. Τέλος από όλους τους αναγνώστες ζητώ κατανόηση για τυχόν λάθη και ελλείψεις που υπάρχουν στο παρόν σύγγραμμα.

Ηλίας Γκανέτσος

Μάιος 2007

Περίληψη

Η ανάλυση παλινδρόμησης είναι η πλέον γνωστή τεχνική για να μελετήσουμε τη μέση συμπεριφορά μιας μεταβλητής, δοθέντων άλλων μεταβλητών που την επηρεάζουν. Ωστόσο, πολλές φορές έχει ενδιαφέρον να μελετήσουμε την επιρροή αυτή σε όλο το εύρος της κατανομής της εξαρτημένης μεταβλητής και όχι μόνο στη μέση της τιμή. Στην παρούσα διπλωματική εργασία παρουσιάζουμε μία ανθεκτική σε έκτροπες παρατηρήσεις τεχνική που δίνει την απάντηση στο παραπάνω ζήτημα, την Παλινδρόμηση Ποσοστιαίων Σημείων. Θα αναλυθεί το μοντέλο, θα ερμηνευτούν τα χαρακτηριστικά εκτίμησής του και θα συνοψιστούν τα πλεονεκτήματά του έναντι της Κλασικής Παλινδρόμησης.

Στη συνέχεια θα δούμε μια σειρά εφαρμογών της μεθόδου. Θα διαπιστώσουμε με τη βοήθεια ενός απλού σετ δεδομένων την ανθεκτικότητα της Παλινδρόμησης της Διαμέσου σε έκτροπες παρατηρήσεις σε σχέση με την Κλασική Παλινδρόμηση, θα δούμε μεγάλες έρευνες όπως την εφαρμογή της μεθόδου σε δεδομένα οικονομικά και ιατρικά και θα αναλύσουμε βάσει αυτής κάποια εκλογικά αποτελέσματα των τελευταίων Ελληνικών βουλευτικών εκλογών, καθώς και οικονομικά χαρακτηριστικά κάποιων ασφαλιστικών εταιριών μηχανοκίνητων οχημάτων που δραστηριοποιούνται στην Ελλάδα.

Τέλος, η μελέτη θα επεκταθεί στη σύνδεση της Παλινδρόμησης Ποσοστιαίων Σημείων με τα Γραμμικά Μικτά Μοντέλα. Θα παρουσιαστεί και θα αξιολογηθεί η μέθοδος της Μ-Παλινδρόμησης Ποσοστιαίων Σημείων, που αποτελεί μία εναλλακτική επιλογή για δεδομένα που αναλύονται με τα Μικτά Μοντέλα. Η νέα αυτή τεχνική, αν και δεν μπορεί να κριθεί πιο αξιόπιστη, ξεπερνάει κάποια προβλήματα που θα αντιμετωπίζαμε αν επιλέγαμε την ανάλυση μέσω των Γραμμικών Μικτών Μοντέλων.

Abstract

Regression analysis is the most prevalent technique for studying the mean behaviour of a variable, given other variables which affect it. That said, it is often interesting to investigate this correlation at the whole range of the dependent variable's distribution rather than just its mean. In this dissertation, a resistant to outlying observations technique, called Quantile Regression, will be presented and analyzed in an effort to provide an answer to the aforementioned issue. After we analyze the model, we will reach some conclusions in regards to the estimation characteristics and we will summarize its advantages to Classical Regression.

Moreover, a number of practical applications of this method will be portrayed. Through a simple set of variables, the robustness of the Median Regression in outliers will be shown and contrasted to that of the Classic Linear Regression. International studies will be reviewed, which contain economic and medical applications of the method. Furthermore, this method will be used to analyze and explain electoral results taken from the last Greek Parliamentary Elections, as well as to analyze some automobile insurance companies' financial characteristics as it concerns the case of Greece.

Finally, the study will consider the connection between Quantile Regression and Linear Mixed Models. The M-Quantile Regression will also be evaluated; this method provides an alternative way to deal with data which are analyzed through Mixed Models. This new technique, though it may not be considered more reliable, does solve some problems that we would have faced if we had chosen to analyze a set of variables through Linear Mixed Models.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κατάλογος Συντομογραφιών

ΠΠΣ	Παλινδρόμηση Ποσοστιαίων Σημείων
ΚΓΠ	Κλασική Γραμμική Παλινδρόμηση
ΠΔ	Παλινδρόμηση Διαμέσου
ΓΜΜ	Γραμμικό Μικτό Μοντέλο
Μ-ΠΠΣ	Μ-Παλινδρόμηση Ποσοστιαίων Σημείων

ΚΕΦΑΛΑΙΟ 1

Η ΠΑΛΙΝΔΡΟΜΗΣΗ ΠΟΣΟΣΤΙΑΙΩΝ ΣΗΜΕΙΩΝ

1.1 Εισαγωγή

Σε πολλές εφαρμοσμένες μελέτες παρατηρείται ότι μία εξαρτημένη μεταβλητή που μας ενδιαφέρει επηρεάζεται έμμεσα, μέσω κάποιων επεξηγηματικών μεταβλητών. Όπως είναι γνωστό, ένα συνηθισμένο μοντέλο παλινδρόμησης μοντελοποιεί τη σχέση μιας ή περισσοτέρων επεξηγηματικών μεταβλητών X με τη δεσμευμένη μέση τιμή μιας εξαρτημένης μεταβλητής Y, δοθέντος $X=x$. Τα κλασικά αυτά μοντέλα παλινδρόμησης δεν είναι πάντα κατάλληλα όσον αφορά τη φύση των δεδομένων και ποιες ακριβώς τιμές της εξαρτημένης μεταβλητής θέλουμε να εξηγήσουμε. Αυτό συμβαίνει σε πολλές περιπτώσεις, όπου ο υποπληθυσμός δεν έχει τόσο ενδιαφέρον να μελετηθεί στη μέση συμπεριφορά του και προκύπτει η ανάγκη να βρούμε τρόπους να μοντελοποιήσουμε ως προς τις ακραίες του τιμές. Αντί λοιπόν να εξηγούμε τη μέση συμπεριφορά της Y δοθέντος της X, θα μπορούσαμε να μελετήσουμε τη συμπεριφορά της Y δοθέντος της X σε όλο το εύρος της κατανομής της πρώτης.

Την απάντηση στα παραπάνω ζητήματα έρχεται να δώσει η Παλινδρόμηση Ποσοστιαίων Σημείων (ΠΠΣ), της οποίας η θεωρία αναπτύχθηκε από τους Roger Koenker και Gilbert Basset το 1978 στην εργασία τους με τίτλο «Regression Quantiles» που δημοσίευσε το επιστημονικό περιοδικό Econometrica. Η ΠΠΣ είναι μια επέκταση της κλασικής μεθόδου παλινδρόμησης που βασίζεται στη μέθοδο ελαχίστων τετραγώνων και έχει σκοπό την εκτίμηση και συμπερασματολογία γύρω από συναρτήσεις ποσοστιαίων σημείων της δεσμευμένης κατανομής της εξαρτημένης μεταβλητής δοθέντων άλλων ανεξάρτητων. Όπως η Κλασική Γραμμική Παλινδρόμηση (ΚΓΠ)

χρησιμοποιείται για να εκτιμούμε μοντέλα για δεσμευμένες συναρτήσεις του μέσου όρου, έτσι και η ΠΠΣ διαθέτει έναν άλλο μηχανισμό για εκτίμηση μοντέλων υπό συνθήκη συναρτήσεων της διαμέσου, καθώς και όλου του εύρους των ποσοστιαίων σημείων της κατανομής. Έτσι, η στατιστική αυτή τεχνική συμπληρώνει την ΚΓΠ, προσφέροντας μια πιο πλήρη στατιστική ανάλυση στοχαστικών σχέσεων μεταξύ τυχαίων μεταβλητών. Όπως γίνεται αντιληπτό η μέθοδος αυτή είναι πολύ πιο ανθεκτική στην ύπαρξη έκτροπων παρατηρήσεων στο υπό μελέτη δείγμα, ενώ έχει μια σειρά επιπλέον πλεονεκτημάτων που θα μελετήσουμε παρακάτω.

Η σπουδαιότητα της μεθόδου της ΠΠΣ φαίνεται και από τη θεωρητική ενασχόληση σπουδαίων στατιστικών με αυτήν, αλλά και από τις πολλές και ενδιαφέρουσες εφαρμογές της σε διάφορους τομείς. Τη θεωρία των Koenker και Basset (1978) ήρθαν να συμπληρώσουν μεταξύ άλλων οι Powell (1986), Koenker και Portnoy (1987, 1997), Portnoy (1991), Guttembrunner και Jureckova (1992), Hendricks και Koenker (1991), Buchinsky (1998), Chamberlain (1994), Chaudhuri, Doksum και Samarov (1997) και Knight (1998).

Στη συνέχεια θα επιχειρήσουμε να δώσουμε τα βασικά στοιχεία της θεωρίας που θεμελίωσαν οι παραπάνω στατιστικοί και στην οποία βασίζεται η μέθοδος της ΠΠΣ.

1.2 Το Μοντέλο Παλινδρόμησης Ποσοστιαίων Σημείων

Έστω τυχαία μεταβλητή X με συνάρτηση κατανομής:

$$F(x) = \Pr(X \le x). \qquad (1.2.1)$$

Το τ-οστό ποσοστιαίο σημείο, για $0 < t < 1$, ορίζεται ως:

$$Q(t) = \inf\{x : F(X) \ge t\}, \qquad (1.2.2)$$

όπου X είναι μια μεταβλητή με συνάρτηση κατανομής την (1.2.1).

Ο ορισμός του ποσοστιαίου σημείου λέει ότι μια παρατήρηση στο *τ*-οστό ποσοστημόριο είναι μεγαλύτερη από το *τ%* των παρατηρήσεων και μικρότερη από το *(1-τ)%* των παρατηρήσεων. Το *25%*-ποσοστιαίο σημείο είναι η τιμή $Q(1/4)$, η διάμεσος, που παίζει τον κεντρικό ρόλο, είναι η τιμή $Q(1/2)$, το *75%*-ποσοστιαίο σημείο είναι η τιμή $Q(3/4)$ κ.ο.κ. Σε συνδυασμό με τη συνάρτηση κατανομής, η συνάρτηση ποσοστιαίων σημείων, παρέχει μια πιο πλήρη εικόνα της τυχαίας μεταβλητής *X*.

Έστω τώρα (y_i, x_i), *i=1,2,…,n*, ένα δείγμα από κάποιον πληθυσμό, όπου y_i οι τιμές της εξαρτημένης μεταβλητής που μας ενδιαφέρει και x_i το διάνυσμα των επεξηγηματικών μεταβλητών. Το γενικό μοντέλο ΠΠΣ, όπως το περιγράφει ο Bunchinsky (1998), παίρνει τη γραμμική μορφή:

$$y_i = x_i^T b(t) + u_i(t), \qquad (1.2.3)$$

για $i=1,2,\ldots,n$, όπου $b(t)$ είναι ένα *kx1* διάνυσμα συντελεστών προς εκτίμηση, x_i είναι το διάνυσμα στήλη, που είναι η αντιμετάθεση της *i*-οστής γραμμής του *nxk* πίνακα *X* των επεξηγηματικών μεταβλητών, y_i είναι η *i*-οστή παρατήρηση της εξαρτημένης μεταβλητής και $u_i(t)$ είναι ένας άγνωστος όρος σφάλματος. Το *τ*-οστό δεσμευμένο ποσοστιαίο σημείο της *y* δοθέντος της *x* μπορεί να γραφεί ως εξής:

$$Q_t(y_i \mid x_i) = x_i^T b(t). \qquad (1.2.4)$$

Για παράδειγμα, έστω y_i η απόδοση του *i*-οστού μαθητή σε κάποιο σχολείο και έστω x_i οι τιμές του διανύσματος των επεξηγηματικών μεταβλητών για τον *i*-οστό μαθητή, που μπορεί να είναι το κοινωνικό και οικονομικό του επίπεδο, η μόρφωση των γονιών του κλπ.

Σύμφωνα με το κλασικό μοντέλο ΓΠ, οι αποδόσεις των μαθητών θα έχουν αναμενόμενη τιμή που θα είναι γραμμική συνάρτηση των επεξηγηματικών μεταβλητών που αναφέραμε παραπάνω και θα συνδέονται με τη σχέση:

$$E(y_i \mid x_i) = x_i^T b, \qquad (1.2.5)$$

ή αν θέλουμε να δηλώσουμε σαν *α* τον σταθερό όρο:

$$E(y_i \mid x_i) = a + x_i^T b. \qquad (1.2.6)$$

Το διάνυσμα των συντελεστών β, περιγράφει τον τρόπο με τον οποίο οι επεξηγηματικές μεταβλητές x_i επηρεάζουν την μέση απόδοση των μαθητών y_i.

Όσον αφορά το γραμμικό μοντέλο ΠΠΣ, η τιμή που εξαρτάται από τις επεξηγηματικές μεταβλητές που έχουμε χρησιμοποιήσει δεν είναι η μέση τιμή της δεσμευμένης κατανομής της μεταβλητής της απόδοσης των μαθητών αλλά το *τ*-οστό ποσοστιαίο σημείο της. Έτσι το μοντέλο για το *τ*-οστό ποσοστιαίο σημείο είναι το εξής:

$$Q_t(y_i \mid x_i) = x_i^T b(t), \qquad (1.2.7)$$

ή αν θέλουμε να δηλώσουμε σαν $a(t)$ τον σταθερό όρο για το *τ*-οστό ποσοστιαίο σημείο:

$$Q_t(y_i \mid x_i) = a(t) + x_i^T b(t). \qquad (1.2.8)$$

Αναφορικά με το παραπάνω παράδειγμα, ενώ η ΚΓΠ μελετά το πώς παράγοντες όπως κοινωνικό επίπεδο, οικονομικό επίπεδο, μόρφωση γονέων κλπ επηρεάζουνε την απόδοση μόνο του μέσου μαθητή, το μοντέλο της ΠΠΣ προχωράει παραπέρα και μελετά το πώς οι παραπάνω παράγοντες επηρεάζουνε την απόδοση ξεχωριστά των καλών, των μεσαίων και των κακών μαθητών ανάλογα με την επιλογή του *τ*.

Ας επικεντρωθούμε τέλος στο ρόλο που παίζουν τα κατάλοιπα Η σχέση (1.2.4) σε συνδυασμό με τη σχέση (1.2.3) μας λέει ότι $Q_t(u_i(t) \mid x_i) = 0$ για δοθέν διάνυσμα επεξηγηματικών x_i. Θεωρούμε το μοντέλο πρόβλεψης:

$$\hat{y}_i = x_i^T b(t), \qquad (1.2.9)$$

το οποίο προβλέπει την απόκριση της εξαρτημένης μεταβλητής y για συγκεκριμένες τιμές των επεξηγηματικών μεταβλητών. Τότε τα κατάλοιπα $u_i(t) = y_i - \hat{y}_i$ αποτελούν ένα μέτρο για το πόσο καλά σχετίζεται το μοντέλο με τις πραγματικές αποκρίσεις. Η διαφορά ανάμεσα στην ΠΠΣ και στην ΚΓΠ είναι ότι αντί να ισχύει $E(u_i) = 0$, έχουμε $Q_t(u_i(t)) = 0$ για την παλινδρόμηση του ποσοστιαίου σημείου με αντίστοιχη τιμή τ. Αυτό οδηγεί στη σχέση:

$$Q_t(u_i(t)) = Q_t(y_i - \hat{y}_i) = 0. \qquad (1.2.10)$$

Στο σημείο αυτό πρέπει να σημειωθεί ότι επειδή ο καθορισμός των ποσοστιαίων σημείων απαιτεί διάταξη των παρατηρήσεων, ο τελεστής Q_t δεν είναι γραμμικός, οπότε:

$$Q_t(y_i - \hat{y}_i) \neq Q_t(y_i) - Q_t(\hat{y}_i). \qquad (1.2.11)$$

1.3 Εκτίμηση παραμέτρων μέσω βελτιστοποίησης

Τα ποσοστιαία σημεία έχουν άμεση σχέση με τη διάταξη των δειγματικών παρατηρήσεων που χρησιμοποιούμε για να τα ορίσουμε. Έτσι μπορούμε να τα μελετήσουμε από μια άλλη οπτική γωνία, ως ένα πρόβλημα βελτιστοποίησης. Όπως ορίζουμε τον δειγματικό μέσο ως τη λύση ενός προβλήματος ελαχιστοποίησης του αθροίσματος των τετραγώνων των καταλοίπων, έτσι ορίζουμε την διάμεσο ως τη λύση

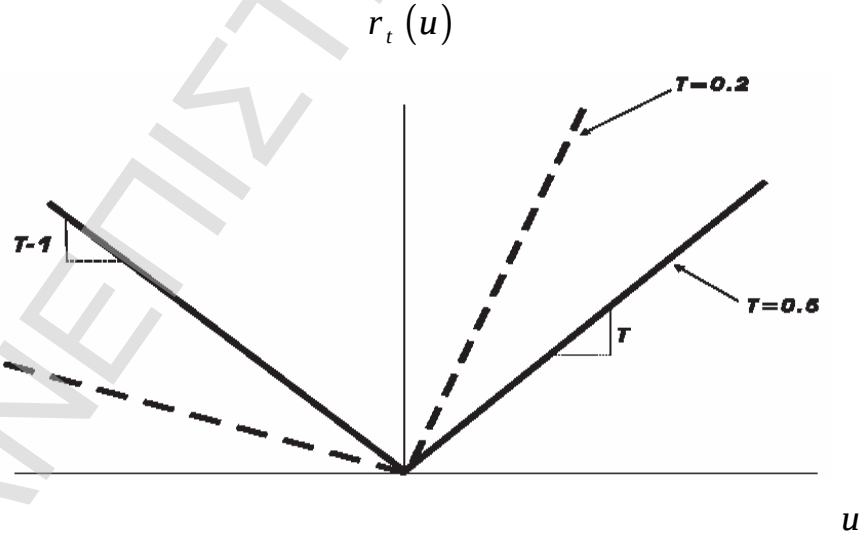
ενός προβλήματος ελαχιστοποίησης ενός αθροίσματος απόλυτων καταλοίπων. Τι γίνεται όμως με τα υπόλοιπα ποσοστιαία σημεία; Αν η συνάρτηση της συμμετρικής απόλυτης τιμής προσεγγίζει τη διάμεσο, τότε μπορούμε να χρησιμοποιήσουμε μία συνάρτηση μη συμμετρικής σταθμισμένης απόλυτης τιμής ώστε να προσεγγίσουμε και τα υπόλοιπα ποσοστιαία σημεία. Με βάση το παραπάνω σκεπτικό η προσέγγιση των ποσοστιαίων σημείων ανάγεται στη λύση του παρακάτω προβλήματος ελαχιστοποίησης:

$$\min_{x\in\Re} \sum r_t(y_i - x), \qquad (1.3.1)$$

όπου $r_t(\cdot)$ η «συνάρτηση ελέγχου», που ορίζεται ως εξής:

$$r_t(u) = u(t - I(u<0)) = t\cdot u^+ + (1-t)\cdot u^-, \qquad (1.3.2)$$

όπου τα $u^+ = u\cdot I[u>0]$ και $u^- = -u\cdot I[u<0]$ δηλώνουν αντίστοιχα τα θετικά και αρνητικά μέρη του $u\in\Re$. Στο παρακάτω σχήμα φαίνεται και η γραφική απεικόνιση της εν λόγω συνάρτησης για διάφορες τιμές του *τ*:



ΣΧΗΜΑ 1-1 : Συνάρτηση Ελέγχου για *τ = 0.2, 0.5, 0.8*.
(Πηγή: Koenker, Hallock, 2001)

Έχοντας επιτύχει τον ορισμό των μη δεσμευμένων ποσοστιαίων σημείων ως ένα πρόβλημα βελτιστοποίησης, είναι τώρα εύκολο να ορίσουμε με ανάλογο τρόπο και τα δεσμευμένα ποσοστιαία σημεία. Η μέθοδος ελαχίστων τετραγώνων της ΚΓΠ προσφέρει ένα μοντέλο για τον τρόπο που θα ακολουθήσουμε. Έστω ένα τυχαίο δείγμα $\{y_1, y_2, ..., y_n\}$. Αν λύσουμε το παρακάτω πρόβλημα ελαχιστοποίησης:

$$\min_{m\in\Re} \sum_{i=1}^{n} (y_i - m)^2, \qquad (1.3.3)$$

τότε παίρνουμε τον δειγματικό μέσο ως μία εκτίμηση του μη δεσμευμένου πληθυσμιακού μέσου *EY*.

Αν αντικαταστήσουμε την παράμετρο μ με την παραμετρική συνάρτηση $\mu(x_i, \beta)$ και λύσουμε το παρακάτω πρόβλημα ελαχιστοποίησης:

$$\min_{b\in\Re^p} \sum_{i=1}^{n} (y_i - m(x_i, b))^2, \qquad (1.3.4)$$

τότε θα πάρουμε μια εκτίμηση της δεσμευμένης μέσης τιμής $E(Y|x)$.

Στην ΠΠΣ ακολουθείται ο ίδιος ακριβώς τρόπος. Για να πάρουμε μια εκτίμηση της δεσμευμένης συνάρτησης της διαμέσου απλά αντικαθιστούμε στη σχέση (1.3.1) την παράμετρο ξ με την παραμετρική συνάρτηση $\xi(x_i, \beta)$ και θέτουμε $\tau=1/2$. Στην περίπτωση της διαμέσου η συνάρτηση ελέγχου είναι η συνάρτηση της απόλυτης τιμής, δηλαδή $r_t(\cdot) = |\cdot|$, επομένως αρκεί να βρούμε τη λύση ενός προβλήματος ελαχιστοποίησης του αθροίσματος των απόλυτων καταλοίπων. Για να βρούμε τώρα τις εκτιμήσεις των δεσμευμένων συναρτήσεων άλλων ποσοστιαίων σημείων απλά αντικαθιστούμε τις απόλυτες τιμές με την $r_t(\cdot)$, για το αντίστοιχο *τ*, και λύνουμε το εξής πρόβλημα ελαχιστοποίησης:

$$\min_{b\in\Re^p} \sum r_t(y_i - x(x_i, b)). \qquad (1.3.5)$$

Εναλλακτικά, χωρίς τη χρήση της συνάρτησης ελέγχου $r_t(\cdot)$, για να βρούμε τις εκτιμήσεις μπορούμε να λύσουμε το παρακάτω πρόβλημα ελαχιστοποίησης που πρότειναν οι Koenker και Basset το 1978 και είναι ισοδύναμο του (1.3.5):

$$\min_{b\in\Re^p}\left[\sum_{i\in\{i:y_i\geq x(x_i,b)\}} t\cdot|y_i-x(x_i,b)|+\sum_{i\in\{i:y_i<x(x_i,b)\}}(1-t)\cdot|y_i-x(x_i,b)|\right]. \qquad (1.3.6)$$

1.4 Υπολογιστικά θέματα

Ένα από τα πλεονεκτήματα της ΠΠΣ όπως είδαμε παραπάνω είναι ότι η εκτίμηση των παραμέτρων μπορεί να παρασταθεί σαν ένα πρόβλημα γραμμικού προγραμματισμού, κάτι που έχει σημαντική σημασία και από θεωρητική αλλά και από πρακτική σκοπιά.

Συγκεκριμένα έχουμε το πρόβλημα ελαχιστοποίησης (1.3.5). Το πρόβλημα αυτό, όταν η $x\left(x,\hat{b}(t)\right)$ είναι μια γραμμική συνάρτηση μπορεί να λυθεί πολύ εύκολα με μεθόδους Γραμμικού Προγραμματισμού. Τέτοιες είναι οι παρακάτω:

► Μέθοδος των Barrodale & Roberts (Barrodale & Roberts, 1973 και 1974) (για μικρό μέγεθος δείγματος).

► Μέθοδος Simplex (Koenker & D'Orey, 1987 και 1993) (για μέτριο μέγεθος δείγματος).

► Μέθοδος Εσωτερικού Σημείου (Portnoy & Koenker, 1997) (για αρκετά μεγάλο μέγεθος δείγματος).

► Μέθοδος Εσωτερικού Σημείου με preprocessing (Portnoy & Koenker, 1997) (για πολύ μεγάλο μέγεθος δείγματος $n>10^5$).

► Μέθοδος εξομάλυνσης (Chen, 2004).

Ο Buchinsky (1998), μελέτησε τα υπολογιστικά θέματα της μεθόδου αρκετά συνοπτικά αλλά και από μια «μη τεχνική» σκοπιά.

Από την έκφραση $y_i = x_i^T b(t) + u_i(t)$ μπορούμε να περάσουμε σε μία άλλη σχέση γράφοντας το y_i σαν μια συνάρτηση μόνο θετικών στοιχείων και εν συνεχεία να την μετατρέψουμε σε έναν πίνακα. Με τον τρόπο αυτό μετατρέπουμε τον αρχικό τύπο σε πρόβλημα γραμμικού προγραμματισμού. Δηλαδή μπορούμε να γράψουμε :

$$y_i = \sum_{j=1}^{K} x_{ij} b_{tj} + u_{tj} = \sum_{j=1}^{K} x_{ij} (b_{tj}^{1} - b_{tj}^{2}) + (e_{ti} - v_{ti}), \qquad (1.4.1)$$

όπου b_{tj}^{1} , b_{tj}^{2} , e_{ti} και v_{ti} μη αρνητικά με $j = 1,...,K$ και $i = 1,...,n$.

Η μορφή του αρχικού προβλήματος γραμμικού προγραμματισμού θα είναι τότε σε πίνακες :

$$\min_z c^T z$$

με περιορισμούς $Az = y,\ z \geq 0, \qquad (1.4.2)$

όπου $A = (X. - X, I_n, -I_n)$, $z = \left(b^{1^T}, b^{2^T}, u^T, v^T\right)^T$ και $c = \left(0^T, 0^T, t \cdot l^T, (1-t) \cdot l^T\right)^T$. Επιπλέον I_n είναι ο *n*-διάστατος ταυτοτικός πίνακας, 0^T είναι ένα *K x 1* διάνυσμα-στήλη με μηδενικά και *l* είναι ένα *n x 1* μοναδιαίο διάνυσμα.

Τέλος όσον αφορά δυική μορφή του παραπάνω προβλήματος γραμμικού προγραμματισμού, αυτή είναι εύκολο να γραφεί, με τη βοήθεια της σχέσης (1.4.2) :

$$\max_w w^T y$$

με περιορισμούς $w^T y \leq c^T$. $\qquad (1.4.3)$

Το θεώρημα της δυικότητας λέει πως υπάρχουν λύσεις και για τους δυο τύπους αν ο πίνακας X είναι πλήρους τάξεως. Επιπλέον το θεώρημα ισορροπίας του Γραμμικού Προγραμματισμού εγγυάται το βέλτιστο αυτών των λύσεων.

1.5 Ερμηνεία μοντέλων Παλινδρόμησης Ποσοστιαίων Σημείων

Θεωρούμε πάλι το μοντέλο (1.1.4) της ΠΠΣ. Όπως θα δούμε σε μία από τις παρακάτω παραγράφους, για κάθε μονότονο μετασχηματισμό $h(\cdot)$ ισχύει:

$$Q_{h(Y)}(t \mid X = x) = h(Q_Y(t \mid X = x)).$$

Από την παραπάνω σχέση προκύπτει άμεσα ότι αν $Q_{h(Y)}(t \mid X = x) = x^T b(t)$, τότε:

$$\frac{\partial Q_Y(t \mid X = x)}{\partial x_j} = \frac{\partial h^{-1}(x^T b)}{\partial x_j}. \qquad (1.5.1)$$

Επομένως, αν για παράδειγμα η h είναι η λογαριθμική συνάρτηση, τότε από τη σχέση $Q_{\log(Y)}(t \mid X = x) = x^T b(t)$, προκύπτει ότι $\frac{\partial Q_Y(t \mid X = x)}{\partial x_j} = \exp(x^T b) b_j$.

Άρα αρκεί να προσπαθούμε σε κάθε περίπτωση να δίνουμε την ερμηνεία της μερικής παραγώγου $\frac{\partial Q_Y(t \mid X = x)}{\partial x_j}$ κάτι που τις περισσότερες φορές χρειάζεται ιδιαίτερη προσοχή.

1.6 Ιδιότητες Παλινδρόμησης Ποσοστιαίων Σημείων

1.6.1 Equivariance

Ας υποθέσουμε ότι έχουμε ένα μοντέλο για τη θερμοκρασία ενός υγρού που εκφράζεται μέσω της τυχαίας μεταβλητής y και αποφασίζουμε να αλλάξουμε την κλίμακα των μετρήσεων από Κελσίου σε Φαρενάιτ, ή αποφασίζουμε να μελετήσουμε και την επίδραση του αθροίσματος ή της διαφοράς των δύο επεξηγηματικών μεταβλητών που διαθέτουμε. Οι αλλαγές αυτές δε θα έχουν βασική επίδραση στις εκτιμήσεις μας. Όταν τα δεδομένα αλλάζουν με παρόμοιο με τους παραπάνω προβλέψιμους τρόπους, τότε αναμένουμε και οι εκτιμήσεις της παλινδρόμησης να αλλάζουν και αυτές έτσι ώστε να αφήνουν την ερμηνεία των αποτελεσμάτων αμετάβλητη. Αρκετές τέτοιες ιδιότητες μπορούν να συνοψιστούν υπό τον τίτλο «equivariance» και συχνά προσφέρουν μια σημαντική βοήθεια στην προσεκτική ερμηνεία των αποτελεσμάτων μιας στατιστικής ανάλυσης.

Έστω ότι εκτελείται παλινδρόμηση του τ-οστού ποσοστιαίου σημείου σε ένα σετ δεδομένων (y,X) και προκύπτει ο εκτιμητής $\hat{b}(t; y, X)$. Τέσσερις βασικές equivariance ιδιότητες του $\hat{b}(t; y, X)$ συνοψίζονται στο παρακάτω θεώρημα:

Θεώρημα (Koenker & Basset, 1978): Έστω **A** ένας *pxp* ουδέτερος πίνακας , $\gamma \in \Re^p$ και $\alpha > 0$. Τότε, για κάθε $\tau \in [0,1]$:

(i) $\hat{b}(t; ay, X) = a\hat{b}(t; y, X)$

(ii) $\hat{b}(t; -ay, X) = -a\hat{b}(1-t; y, X)$

(iii) $\hat{b}(t; y + Xg, X) = \hat{b}(t; y, X) + g$

(iv) $\hat{b}(t; y, XA) = A^{-1}\hat{b}(t; y, X)$

Οι ιδιότητες *(i)* και *(ii)* υποδηλώνουν την equivariance της κλίμακας, η ιδιότητα *(iii)* συνήθως ονομάζεται equivariance της παλινδρόμησης ή της μετατόπισης (shift equivariance), ενώ η ιδιότητα *(iv)* έχει να κάνει με την αναπαραμετροποίηση (equivariance to reparameterization of design).

Τα ποσοστιαία σημεία όμως διαθέτουν μία ακόμη equivariance ιδιότητα. Αυτή η ιδιότητα ονομάζεται equivariance σε μονότονους μετασχηματισμούς. Έστω $h(\cdot)$ μία αύξουσα συνάρτηση του $\Re$. Τότε, για κάθε τυχαία μεταβλητή Y ισχύει:

$$Q_{h(Y)}(t) = h(Q_Y(t)). \qquad (1.6.1)$$

Αυτό σημαίνει ότι τα ποσοστιαία σημεία της μετασχηματισμένης τυχαίας μεταβλητής $h(Y)$ είναι απλώς τα μετασχηματισμένα ποσοστιαία σημεία της αρχικής τυχαίας μεταβλητής *Y*. Στο σημείο αυτό πρέπει να σημειωθεί ότι η μέση τιμή δεν έχει αυτή την ιδιότητα, δηλαδή:

$$E(h(Y)) \neq h(E(Y)). \qquad (1.6.2)$$

Η σχέση (1.6.1) προκύπτει άμεσα από το γεγονός ότι για κάθε μονότονο μετασχηματισμό h ισχύει:

$$P(Y \leq y) = P(h(Y) \leq h(y)). \qquad (1.6.3)$$

Η ιδιότητα (1.6.3) έχει αρκετά σημαντικά αποτελέσματα. Συχνά θέτουμε ως εξής το μοντέλο της ΚΓΠ:

$$h(y_i, l) = x_i^T b + u_i, \qquad (1.6.4)$$

όπου το $h(y, l)$ δηλώνει κάποιον μετασχηματισμό της αρχικής εξαρτημένης μεταβλητής y, έτσι ώστε η μέση τιμή $E(h(y_i, l)|x)$ να έχει γραμμική σχέση με τις επεξηγηματικές μεταβλητές x και η διακύμανση $V(h(y_i, l)|x)$ να είναι ανεξάρτητη των x (ομοσκεδαστικότητα).

1.6.2 Ανθεκτικότητα (Robustness)

Η ανθεκτικότητα του εκτιμητή της ΠΠΣ σε ακραίες τιμές της εξαρτημένης μεταβλητής *y* μπορεί να αξιολογηθεί με κλασικές τεχνικές όπως η καμπύλη ευαισθησίας, η συνάρτηση επίδρασης (influence function) ή το breakdown point. Μπορεί όμως να φανεί και εύκολα αν σκεφτούμε το εξής. Ας φανταστούμε ένα διάγραμμα διασποράς κάποιων δεδομένων με την προσαρμοσμένη γραμμή της παλινδρόμησης του *τ*-οστού ποσοστιαίου σημείου. Ας θεωρήσουμε ότι παίρνουμε ένα οποιοδήποτε σημείο, ας πούμε το y_i, το οποίο βρίσκεται πάνω από τη γραμμή, και το μετακινούμε αρκετά μακριά από τη γραμμή, προς την κατεύθυνση του άξονα *y*. Εφόσον δεν περάσαμε πάνω από την γραμμή προσαρμογής, θα παρατηρήσουμε ότι η νέα προσαρμογή της γραμμής παλινδρόμησης του *τ*-οστού ποσοστιαίου σημείου είναι ίδια με την παλιά, κάτι που δε θα συνέβαινε αν κάναμε το ίδιο πείραμα με τη γραμμή της μέσης απόκρισης της ΚΓΠ. Αυτό το χαρακτηριστικό της ανθεκτικότητας του εκτιμητή της ΠΠΣ μπορεί να διατυπωθεί μαθηματικά στο παρακάτω θεώρημα:

Θεώρημα (Koenker 2005): Έστω *D* ένας διαγώνιος πίνακας με μη αρνητικά στοιχεία d_i, με *i=1,…,n*. Τότε:

$$\hat{b}(t; y, X) = \hat{b}(t; X\hat{b}(t; y, X) + D\hat{u}, X), \tag{1.6.5}$$

όπου $\hat{u} = y - X\hat{b}(t; y, X)$.

Στο κεφάλαιο 3 θα ασχοληθούμε εκτενέστερα με αυτή την ιδιότητα της μεθόδου της ΠΠΣ, καθώς θα συνδέσουμε τη μέθοδο με κάποιους ιδιαίτερα ανθεκτικούς εκτιμητές.

1.6.3 Ασυμπτωτικές ιδιότητες

Εκτός από τις παραπάνω βασικές ιδιότητες της ΠΠΣ, η μέθοδος έχει και κάποιες ενδιαφέρουσες ασυμπτωτικές ιδιότητες.

Έστω ότι το τ-οστό δεσμευμένο ποσοστιαίο σημείο της y δοθέντος της x έχει την εξής παραμετρική μορφή:

$$Q_Y(t \mid X = x) = g(x, b_0(t)). \qquad (1.6.6)$$

Γεννάται λοιπόν το ερώτημα κάτω από ποιες συνθήκες ο εκτιμητής

$$\hat{b}_n(t) = \arg\min_{b \in \Re^p} \sum_{i=1}^{n} r_t(y_i - g(x_i, b)) \qquad (1.6.7)$$

συγκλίνει κατά πιθανότητα στο $b_0(t)$, ή αλλιώς πότε $\left\| \hat{b}_n(t) - b_0(t) \right\| \to 0$, όταν $n \to \infty$. Με λίγα λόγια θέλουμε να ελέγξουμε τη συνέπεια του εκτιμητή $\hat{b}_n(t)$.

Στην απλούστερη περίπτωση των δειγματικών ποσοστιαίων σημείων μιας μεταβλητής y με κατανομή F, έχουμε ότι:

$$\hat{x}_n(t) = \arg\min_{x \in \Re} \sum_{i=1}^{n} r_t(y_i - x) \qquad (1.6.8)$$

και θέλουμε να δούμε αν $\hat{x}_n(t) \to x_0(t)$ με την υπόθεση ότι η F έχει ένα μοναδικό τ-οστό ποσοστιαίο σημείο, το $x_0(t) = F^{-1}(t)$. Στην περίπτωση αυτή αποδεικνύεται ότι αν η F έχει μία συνεχή συνάρτηση πυκνότητας f, τότε:

$$\sqrt{n}\left(\hat{x}_n(t)-x_0(t)\right)\sim\ N\left(0,w^2\right), \tag{1.6.9}$$

όπου $w^2=\dfrac{t\left(1-t\right)}{f^2\left(F^{-1}(t)\right)}$.

Τα παραπάνω αποτελέσματα αφορούν όπως είπαμε την περίπτωση των δειγματικών ποσοστιαίων σημείων μιας μεταβλητής y. Η εισαγωγή τώρα των επεξηγηματικών μεταβλητών x περιπλέκει την κατάσταση ,αφού πλέον περνάμε στην ασυμπτωτική θεωρία της εκτίμησης συναρτήσεων δεσμευμένων ποσοστιαίων σημείων, την οποία θεμελίωσαν οι Koenker και Basset το 1978 και συμπλήρωσαν οι He και Shao το 1996.

Θεωρούμε μία γενική μορφή του γραμμικού μοντέλου ΠΠΣ. Έστω $Y_1,Y_2,...$ ανεξάρτητες τυχαίες μεταβλητές με αντίστοιχες συναρτήσεις κατανομής $F_1,F_2,....$ Έχουμε την παρακάτω γραμμική συνάρτηση ως προς x του τ-οστού δεσμευμένου ποσοστιαίου σημείου:

$$Q_{Y_i}(t\mid x)=x^T b(t). \tag{1.6.10}$$

Οι δεσμευμένες συναρτήσεις κατανομής των Y_i μπορούν να γραφούν $P(Y_i<y\mid x_i)=F_{Y_i}(y\mid x_i)=F_i(y),$ οπότε:

$$Q_{Y_i}(t\mid x_i)=F_{Y_i}^{-1}(t\mid x_i)\equiv x_i(t). \tag{1.6.11}$$

Πριν δούμε ποια ακριβώς είναι η ασυμπτωτική συμπεριφορά του εκτιμητή $\hat{b}_n(t)$ της ΠΠΣ πρέπει πρώτα να θέσουμε τις παρακάτω δύο συνθήκες:

1) Οι συναρτήσεις κατανομών $\{F_i\}$ είναι απόλυτα συνεχείς και έχουν συνεχείς συναρτήσεις πυκνότητας $f_i(x)$ ομοιόμορφα φραγμένες από 0 μέχρι ∞.

2) Υπάρχουν θετικά ορισμένοι πίνακες D_0 και $D_1(t)$ τέτοιοι ώστε:

α) $\lim_{n\to\infty} n^{-1}\sum x_i x_i^T = D_0$

β) $\lim_{n\to\infty} n^{-1}\sum f_i(x_i(t))x_i x_i^T = D_1(t)$

γ) $\frac{\max_{i=1,\ldots,n}\|x_i\|}{\sqrt{n}} \to 0$

Κάτω λοιπόν από τις παραπάνω συνθήκες, η ασυμπτωτική συμπεριφορά του εκτιμητή $\hat{b}_n(t)$ της ΠΠΣ είναι η εξής:

$$\sqrt{n}\left(\hat{b}_n(t) - b(t)\right) \sim N\left(0, t(1-t)D_1^{-1}D_0D_1^{-1}\right). \qquad (1.6.12)$$

1.7 Στατιστική Συμπερασματολογία της ΠΠΣ

Έχουν προταθεί κατά καιρούς πολλά τεστ για τον έλεγχο της μηδενικής υπόθεσης

$$H_0: Rb(t) = r(t) \qquad (1.7.1)$$

Ο Koenker προτείνει τον εξής έλεγχο της H_0 που είναι ουσιαστικά ένας έλεγχος λόγου πιθανοφανειών. Στην περίπτωση της Παλινδρόμησης της Διαμέσου ο έλεγχος της υπόθεσης H_0 για το μοντέλο (1.2.3) για τ=1/2 βασίζεται στο στατιστικό

$$L_n = 8\left(\tilde{V}(1/2) - \hat{V}(1/2)\right)/s(1/2), \qquad (1.7.2)$$

όπου $\tilde{V}(t) = \min_{\{b\in\Re^p\}}\sum r_t\left(y_i - x_i^T b\right)$, $\hat{V}(t) = \min_{\{b\in\Re^p | Rb=r\}}\sum r_t\left(y_i - x_i^T b\right)$ και $s(t) = \left[f\left(F^{-1}(t)\right)\right]^{-1}$.

Αποδεικνύεται ότι κάτω από τη μηδενική υπόθεση η ποσότητα L_n ακολουθεί ασυμπτωτικά την c_q^2 κατανομή, όπου q=rank*(R)*.

Η παραπάνω προσέγγιση μπορεί να επεκταθεί άμεσα και για άλλα ποσοστιαία σημεία εκτός από τη διάμεσο. Το αντίστοιχο στατιστικό που ακολουθεί ασυμπτωτικά την c_q^2 κατανομή στο οποίο θα βασίζεται ο έλεγχος της H_0 για το τ-ποσοστιαίο σημείο είναι το εξής:

$$L_n = \frac{2}{l^2(t)s(t)}\left[\tilde{V}_n(t) - \hat{V}_n(t)\right], \qquad (1.7.3)$$

όπου $l^2(t) = t(1-t)$ και $\tilde{V}(t)$, $\hat{V}(t)$ και $s(t)$ όπως έχουν οριστεί παραπάνω.

1.8 Επιλογή βέλτιστου μοντέλου

Ένα ερώτημα που τίθεται ύστερα από τη μελέτη όλων των βασικών συνιστωσών της μεθόδου της ΠΠΣ είναι η επιλογή του βέλτιστου μοντέλου ΠΠΣ με μεθόδους αντίστοιχες των stepwise αλγορίθμων της ΚΓΠ. Αν και σε αυτόν τον τομέα η αλήθεια είναι ότι δεν υπάρχει μία κοινά αποδεκτή μέθοδος η οποία να χρησιμοποιείται ευρέως, έχουν προταθεί 2 κριτήρια.

Ο Machado το 1993 πρότεινε το κριτήριο:

$$SIC(j) = \log(\hat{s}_j) + \frac{1}{2} p_j \log n, \qquad (1.8.1)$$

όπου $\hat{s}_j = n^{-1}\sum_{i=1}^{n} r_{1/2}\left(y_i - x_i^T b_n(1/2)\right)$ και p_j η διάσταση του j μοντέλου.

Εναλλακτικά θα μπορούσε κανείς να χρησιμοποιήσει το κριτήριο του Akaike:

$$AIC(j) = \log(\hat{s}_j) + p_j. \qquad (1.8.2)$$

1.9 Πλεονεκτήματα Παλινδρόμησης Ποσοστιαίων Σημείων

Όπως είδαμε παραπάνω, η μέθοδος της ΠΠΣ έχει μια σειρά από πλεονεκτήματα, τα οποία θα επιχειρήσουμε να συνοψίσουμε:

► Η ΠΠΣ μπορεί να χρησιμοποιηθεί για να μελετηθεί σε όλο το εύρος της η δεσμευμένη κατανομή της Y δοθέντος $X=x$ και όχι μόνο στη μέση της συμπεριφορά. Αυτό επιτυγχάνεται με τη δοκιμή της αντίστοιχης τιμής του $t \in (0,1)$ ανάλογα με το ποια σημεία της δεσμευμένης κατανομής της Y μας ενδιαφέρουν.

► Το πρόβλημα της εκτίμησης των παραμέτρων του μοντέλου της ΠΠΣ μπορεί να εκφραστεί μέσω Γραμμικού Προγραμματισμού, κάτι που κάνει πολύ εύκολη την εκτίμηση χρησιμοποιώντας κάποιον από τους αλγορίθμους που αναφέραμε στην παράγραφο 1.4.

► Τα ποσοστιαία σημεία έχουν την ιδιότητα της equivariance σε μονότονους μετασχηματισμούς. Για παράδειγμα σε κάποιο παράδειγμα που η εξαρτημένη μεταβλητή εκφράζει μισθούς η δεσμευμένη διάμεσος των λογαρίθμων των μισθών είναι ίση με το λογάριθμο της δεσμευμένης διαμέσου των μισθών.

► Η αντικειμενική συνάρτηση της ΠΠΣ είναι ένα σταθμισμένο άθροισμα απολύτων αποκλίσεων, που δίνει ανθεκτικούς εκτιμητές, κι έτσι το εκτιμώμενο διάνυσμα των παραμέτρων δεν είναι ευαίσθητο σε παρατηρούμενες έκτροπες παρατηρήσεις της εξαρτημένης μεταβλητής.

► Όταν τα σφάλματα δεν ακολουθούν την κανονική κατανομή, οι εκτιμητές της ΠΠΣ είναι πολύ πιο αξιόπιστοι από τους εκτιμητές ελαχίστων τετραγώνων.

► Μέσω της μεθόδου της ΠΠΣ μπορούμε να μελετήσουμε την ετεροσχεδαστικότητα των δεδομένων μας.

ΚΕΦΑΛΑΙΟ 2

ΕΦΑΡΜΟΓΕΣ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΠΟΣΟΣΤΙΑΙΩΝ ΣΗΜΕΙΩΝ

2.1 Εισαγωγή

Όπως είδαμε και στο Κεφάλαιο 1, η μέθοδος της Παλινδρόμησης Ποσοστιαίων Σημείων έχει μία σειρά από πλεονεκτήματα τα οποία παρέχουν στον στατιστικό αναλυτή τρόπους για την αντιμετώπιση προβλημάτων που η Κλασική Παλινδρόμηση δεν μπορεί να λύσει. Επομένως, όπως είναι φυσικό, η μέθοδος έχει βρει ευρεία εφαρμογή σε πάρα πολλούς τομείς.

Έχει χρησιμοποιηθεί για την εκτίμηση των ανώτερων και των χαμηλότερων καμπυλών που αντιστοιχούν σε ποσοστιαία σημεία και έχουν να κάνουν για παράδειγμα με την ηλικία, το φύλο και άλλες επεξηγηματικές μεταβλητές, χωρίς να τεθούν αυστηρές παραμετρικές υποθέσεις στις σχέσεις μεταξύ αυτών των καμπυλών. Ως παράδειγμα μπορούμε να αναφέρουμε ότι μέσω της μεθόδου έχει μοντελοποιηθεί η απόδοση των μαθητών στα σχολεία συναρτήσει κοινωνικοοικονομικών χαρακτηριστικών καθώς και άλλων μεταβλητών όπως σχολικές δαπάνες, προσόντα δασκάλων κλπ. Οι μεταβλητές αυτές δεν επηρεάζουν όλο τον πληθυσμό κατά τον ίδιο τρόπο, οπότε η ΠΠΣ μας βοηθάει να μελετήσουμε την επίδραση των παραπάνω μεταβλητών στην απόδοση των ισχυρότερων μαθητών και να δούμε αν είναι ίδιος ο τρόπος που οι αδύνατοι μαθητές επηρεάζονται. Μία τέτοια εφαρμογή έγινε από τους Eide και Showalter (1998).

Επίσης σημαντικές είναι οι εφαρμογές στη μελέτη της κατανομής των μισθών και στις τάσεις της εισοδηματικής ανισότητας. Χαρακτηριστικές είναι οι εργασίες των

Buchinsky και Hunt (1997) και Eide και Showalter (1999). Μία από τις πιο σημαντικές εφαρμογές της μεθόδου, στην οποία θα αναφερθούμε αναλυτικότερα και παρακάτω, ήταν η ανάλυση των παραγόντων που οδηγούν σε νεογέννητα πολύ χαμηλού βάρους από τον Abreveya (2001). Η έρευνα του Abreveya έδωσε κίνητρο για περαιτέρω ενασχόληση με τη μέθοδο σε τομείς όπως η οικονομετρία και η βιοστατιστική Η μέθοδος της ΠΠΣ έχει εφαρμοστεί και σε πολλές ακόμα περιπτώσεις όπως στην ανάλυση των υψηλότερων προσφορών σε δημοπρασίες από τους Donald και Paarsch (1993), στην εκτίμηση των παραγόντων υψηλού ρίσκου στα οικονομικά από τον Tsay (2002), στην ανάλυση επιβίωσης σε ακραίες διάρκειες από τους Koenker και Geling (2001) και στην ανάλυση των παραγόντων που επιδρούν στα προσεγγιστικά φράγματα βιολογικών διαδικασιών από τον Cade (2003).

Τέλος μεγάλη είναι η συνεισφορά της μεθόδου στη μελέτη των χρηματιστηρίων. Χαρακτηριστικές είναι οι εργασίες των Buchinsky (1994, 1997), που μελετά το αμερικάνικο χρηματιστήριο, Fitzenberger (1999), που μελετά το χρηματιστήριο της Γερμανίας, Machado και Mata (1999), που μελετούν το χρηματιστήριο της Πορτογαλίας, Garcia, Hernadez και Lopez-Nicolas (2001), που μελετούν το χρηματιστήριο της Ισπανίας και Schultz και Mwabu (1998), που μελετούν το χρηματιστήριο της Νοτίου Αφρικής.

Παρακάτω θα δώσουμε πέντε παραδείγματα εφαρμογής της μεθόδου. Στο πρώτο παράδειγμα χρησιμοποιούμε ένα μικρό υποθετικό σετ δεδομένων, στο δεύτερο εφαρμόζουμε τη μέθοδο στα δεδομένα του Engel, στο τρίτο αναλύουμε την έρευνα του Abreveya, στο τέταρτο χρησιμοποιούμε την ΠΠΣ για την ανάλυση κάποιων ελληνικών εκλογικών αποτελεσμάτων και τέλος στο πέμπτο αναλύουμε μέσω της μεθόδου οικονομικά χαρακτηριστικά κάποιων ασφαλιστικών εταιριών.

2.2 Σύγκριση Παλινδρόμησης Διαμέσου με Κλασική Παλινδρόμηση

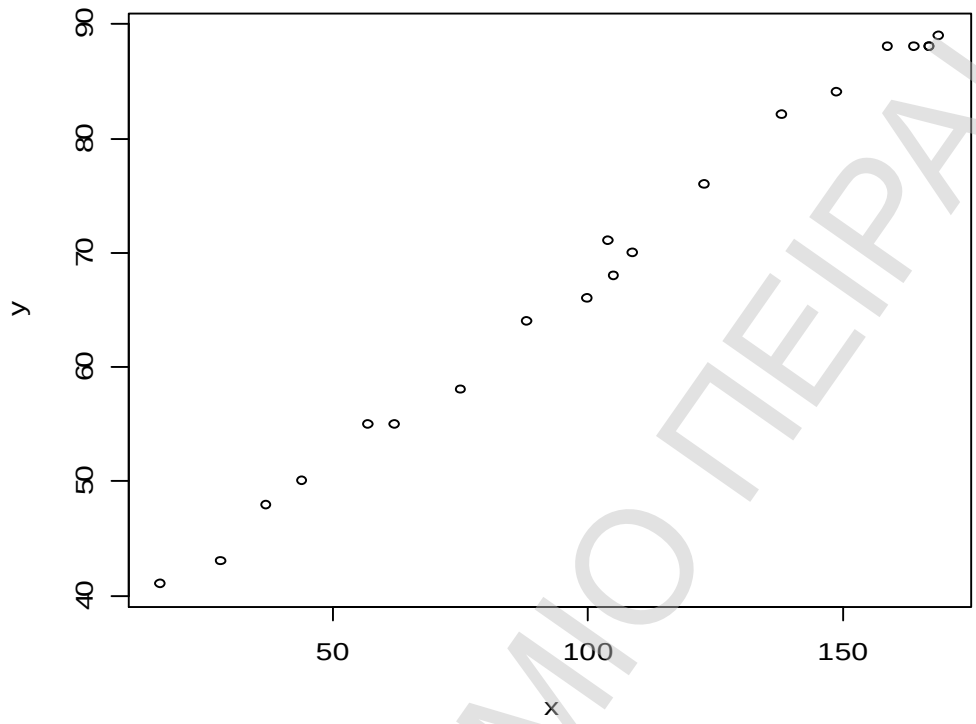
Με τη βοήθεια του στατιστικού πακέτου της γλώσσας **R** θα επιχειρήσουμε να συγκρίνουμε την ΠΠΣ με την ΚΓΠ σε δεδομένα που δεν υπάρχει έκτροπη παρατήρηση καθώς και σε δεδομένα που εμφανίζεται έκτροπη παρατήρηση.

Έστω ότι έχουμε το παρακάτω σετ δεδομένων που αποτελείται από 20 τιμές μιας εξαρτημένης μεταβλητής y και τις αντίστοιχες 20 μιας επεξηγηματικής x:

Παρατήρηση	x	y
1η	123	76
2η	109	70
3η	62	55
4η	104	71
5η	57	55
6η	37	48
7η	44	50
8η	100	66
9η	16	41
10η	28	43
11η	138	82
12η	105	68
13η	159	88
14η	75	58
15η	88	64
16η	164	88
17η	169	89
18η	167	88
19η	149	84
20η	167	88

ΠΙΝΑΚΑΣ 2-1 : Πίνακας δεδομένων (x, y).

Αν κάνουμε το διάγραμμα διασποράς των παραπάνω δεδομένων θα παρατηρήσουμε ότι η εξαρτημένη μεταβλητή y και η επεξηγηματική x διέπονται από μία γραμμική σχέση, χωρίς να υπάρχει κάποια έκτροπη παρατήρηση. Το σχήμα που προκύπτει είναι χαρακτηριστικό:



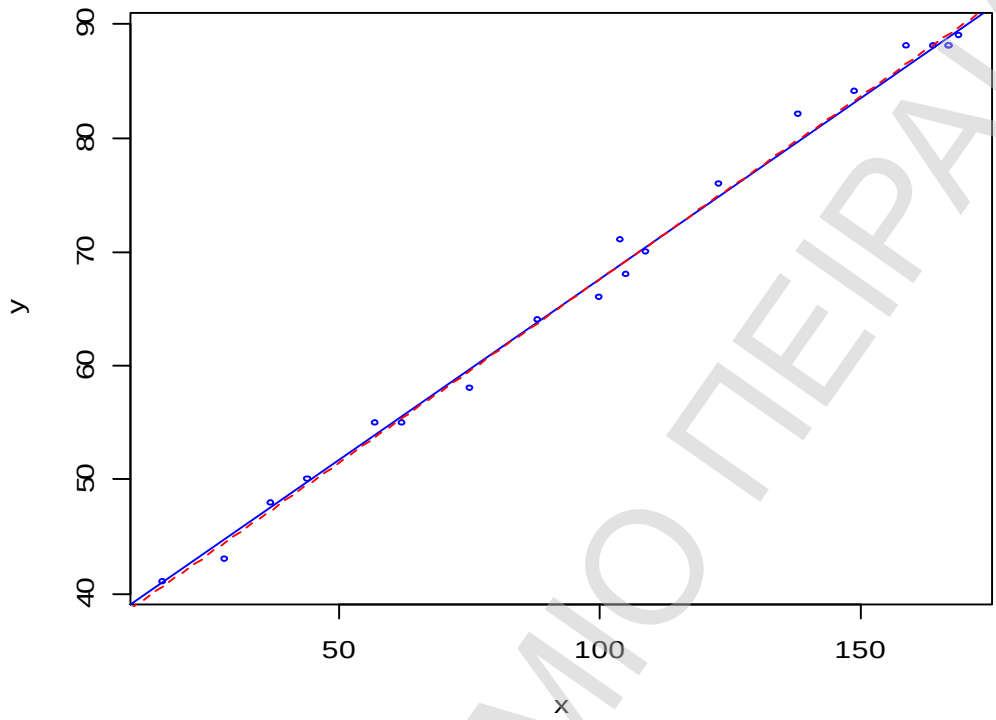
ΣΧΗΜΑ 2-1: Διάγραμμα διασποράς των (*x*, *y*) χωρίς έκτροπη παρατήρηση.

Στα παραπάνω δεδομένα με τη βοήθεια του στατιστικού πακέτου της **R** εφαρμόζουμε την ΠΠΣ συγκεκριμένα στο 0.5-ποσοστιαίο σημείο, δηλαδή στη διάμεσο. Επίσης εφαρμόζουμε και την ΚΓΠ που χρησιμοποιεί τη μέθοδο των ελαχίστων τετραγώνων. Τα συγκριτικά αποτελέσματα της μεθόδου της Παλινδρόμησης της Διαμέσου (ΠΔ) με τη μέθοδο της ΚΓΠ δίνονται στον παρακάτω πίνακα:

Μέθοδος Κλασικής Γραμμικής Παλινδρόμησης					
Μοντέλο: $E(y_i \mid x_i) = \hat{\beta}_0^{LS} + x_i^T \hat{\beta}_1^{LS}$, R^2 =99,47					
$\hat{\beta}_0^{LS}$	35,458	Std.Error	0,635	p-value	0
$\hat{\beta}_1^{LS}$	0,322	Std.Error	0,006	p-value	0
Μέθοδος Παλινδρόμησης Διαμέσου					
Μοντέλο: $Q_{1/2}(y_i \mid x_i) = \hat{\beta}_0^{QR} + x_i^T \hat{\beta}_1^{QR}$					
$\hat{\beta}_0^{QR}$	35,919	Std.Error	0,942	p-value	0
$\hat{\beta}_1^{QR}$	0,318	Std.Error	0,008	p-value	0

ΠΙΝΑΚΑΣ 2-2: Συγκριτικός πίνακας Κλασικής Γραμμικής Παλινδρόμησης και Παλινδρόμησης Διαμέσου για τα δεδομένα (x, y) χωρίς έκτροπη παρατήρηση.

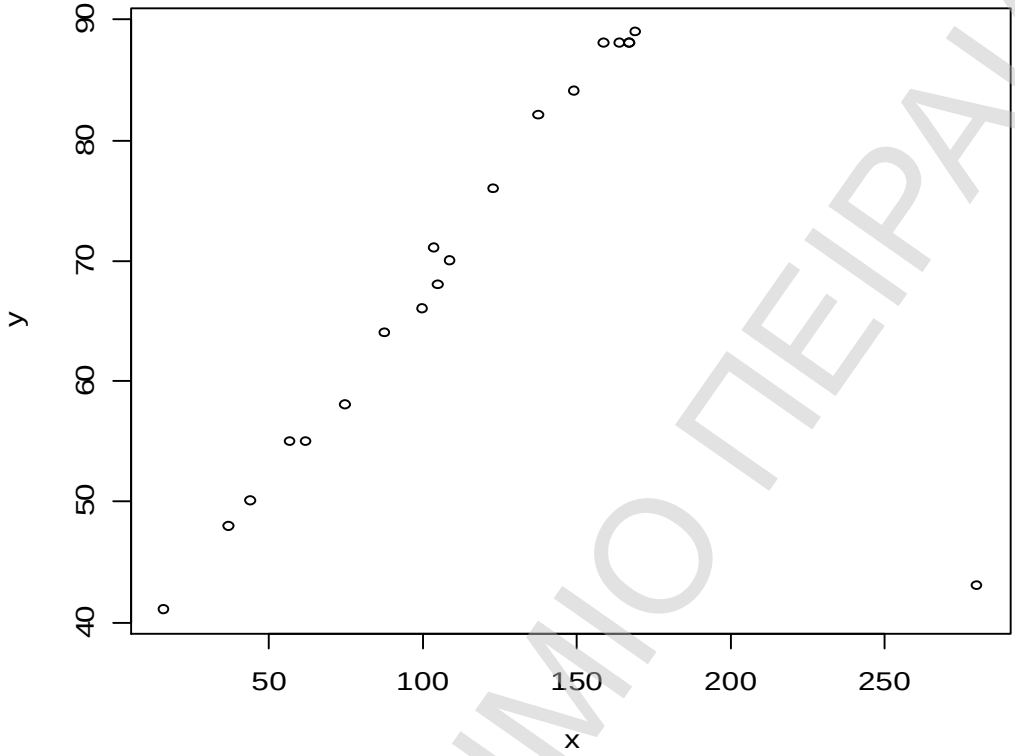
Είναι φανερό ότι η παλινδρόμηση με την κλασική μέθοδο των ελαχίστων τετραγώνων μας δίνει παρόμοια αποτελέσματα με την παλινδρόμηση της διαμέσου, αφού οι εκτιμήσεις για τους συντελεστές παλινδρόμησης είναι πολύ κοντά. Επίσης η προσαρμογή του μοντέλου της ΚΓΠ είναι σχεδόν άψογη καθώς ο συντελεστής παλινδρόμησης R^2 είναι 99.5%. Η πολύ καλή προσαρμογή του μοντέλου της ΚΓΠ, καθώς και αυτή του μοντέλου της ΠΔ φαίνονται ξεκάθαρα και στο παρακάτω διάγραμμα με τις προσαρμοσμένες πλέον ευθείες παλινδρόμησης. Η ευθεία της ΚΓΠ είναι η διακεκομμένη, ενώ αυτή της ΠΔ είναι η συνεχής ευθεία. Βλέπουμε ότι για τα δεδομένα χωρίς καμία έκτροπη παρατήρηση οι δύο ευθείες σχεδόν ταυτίζονται και εξηγούν σχεδόν 100% τη γραμμική σχέση των δύο μεταβλητών:



ΣΧΗΜΑ 2-2: Διάγραμμα διασποράς των (*x*, *y*) χωρίς έκτροπη παρατήρηση με τις προσαρμοσμένες ευθείες παλινδρόμησης.

Έστω τώρα ότι έχουμε πάλι τα δεδομένα του πίνακα 2-1 με τις 20 τιμές της εξαρτημένης μεταβλητής y και τις αντίστοιχες 20 της επεξηγηματικής x με μία όμως αλλαγή: αντί για την τιμή *28* στη $10^{η}$ παρατήρηση της μεταβλητής x έχουμε την τιμή *280*.

Αν κάνουμε το διάγραμμα διασποράς των νέων δεδομένων θα παρατηρήσουμε ότι η εξαρτημένη μεταβλητή y και η επεξηγηματική x διέπονται και πάλι από μία γραμμική σχέση, αλλά υπάρχει πλέον μία έκτροπη παρατήρηση. Το σχήμα που προκύπτει είναι χαρακτηριστικό:



ΣΧΗΜΑ 2-3: Διάγραμμα διασποράς των (*x*, *y*) με έκτροπη παρατήρηση.

Στα νέα δεδομένα με την παρουσία της έκτροπης παρατήρησης εφαρμόζουμε και πάλι την ΠΠΣ στη διάμεσο καθώς και την ΚΓΠ. Τα συγκριτικά αποτελέσματα των δύο μεθόδων δίνονται στον παρακάτω πίνακα:

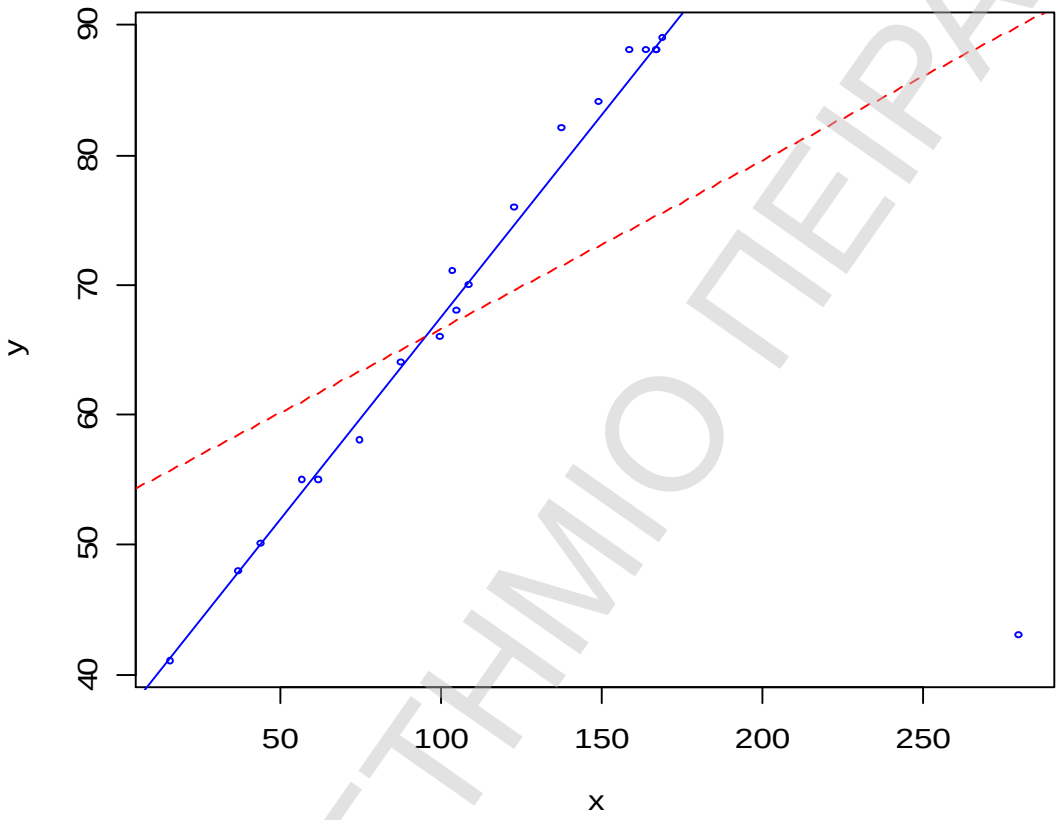
Μέθοδος Κλασικής Γραμμικής Παλινδρόμησης					
Μοντέλο: $E(y_i \mid x_i) = \hat{\beta}_0^{LS} + x_i^T \hat{\beta}_1^{LS}$, R^2 =23,80					
$\hat{\beta}_0^{LS}$	53,536	Std.Error	7,151	p-value	0
$\hat{\beta}_1^{LS}$	0,13	Std.Error	0,055	p-value	0,029
Μέθοδος Παλινδρόμησης Διαμέσου					
Μοντέλο: $Q_{1/2}(y_i \mid x_i) = \hat{\beta}_0^{QR} + x_i^T \hat{\beta}_1^{QR}$					
$\hat{\beta}_0^{QR}$	36,272	Std.Error	1,01	p-value	0
$\hat{\beta}_1^{QR}$	0,312	Std.Error	0,016	p-value	0

ΠΙΝΑΚΑΣ 2-3: Συγκριτικός πίνακας Κλασικής Γραμμικής Παλινδρόμησης και Παλινδρόμησης Διαμέσου για τα δεδομένα (x, y) με έκτροπη παρατήρηση.

Βλέπουμε στον παραπάνω πίνακα ότι τα αποτελέσματα της ΠΔ έχουν παραμείνει σχεδόν αναλλοίωτα. Η εκτίμηση του σταθερού όρου από *35.92* που ήταν στα δεδομένα χωρίς έκτροπη παρατήρηση έγινε τώρα *36.27* και είναι και πάλι στατιστικά σημαντική. Επίσης είναι σημαντικός και ο συντελεστής της επεξηγηματικής μεταβλητής x, ο οποίος από *0.32* που εκτιμήθηκε να είναι στα δεδομένα χωρίς έκτροπη παρατήρηση τώρα έγινε *0.31*.

Αντίθετα, η έκτροπη παρατήρηση φαίνεται να επηρέασε την εφαρμογή της ΚΓΠ αφού ο πίνακας 2-3 μας οδηγεί σε εντελώς διαφορετικά συμπεράσματα από τον πίνακα 2-2 με τα αντίστοιχα αποτελέσματα της παλινδρόμησης με την κλασική μέθοδο των ελαχίστων τετραγώνων για τα δεδομένα χωρίς την έκτροπη παρατήρηση. Ο σταθερός όρος από *35.46* που είχε εκτιμηθεί, εκτιμάται πλέον *53.54*, ενώ η εκτίμηση του συντελεστή της επεξηγηματικής μεταβλητής x από *0.32* έχει γίνει *0.13*. Επίσης ενώ η προσαρμογή του μοντέλου στα δεδομένα χωρίς την έκτροπη παρατήρηση ήταν απόλυτα ικανοποιητική, στα νέα δεδομένα με την έκτροπη παρατήρηση δεν είναι καθόλου καλή. Ο συντελεστής παλινδρόμησης R^2 από *99.5%* είναι πλέον *23.8%*. Η πολύ κακή προσαρμογή του μοντέλου της ΚΓΠ, σε συνδυασμό με την σταθερά ικανοποιητική

προσαρμογή του μοντέλου της ΠΔ φαίνονται ξεκάθαρα και στο παρακάτω διάγραμμα με τις προσαρμοσμένες ευθείες παλινδρόμησης:



ΣΧΗΜΑ 2-4: Διάγραμμα διασποράς των (*x*, *y*) με έκτροπη παρατήρηση με τις προσαρμοσμένες ευθείες παλινδρόμησης.

Είναι ξεκάθαρο και από το παραπάνω σχήμα ότι η ευθεία της ΚΓΠ έχει επηρεαστεί πολύ από την παρουσία της έκτροπης παρατήρησης και δεν μπορεί να εξηγήσει επαρκώς τη γραμμική σχέση που διέπει τα δεδομένα. Αντίθετα, βλέπουμε ότι η ευθεία της ΠΔ δεν έχει επηρεαστεί σχεδόν καθόλου από την έκτροπη παρατήρηση. Το παραπάνω παράδειγμα αναδεικνύει το πλεονέκτημα της μεθόδου της ΠΠΣ έναντι της κλασικής μεθόδου των ελαχίστων τετραγώνων όσον αφορά την ανθεκτικότητα σε ακραίες τιμές.

2.3 Δεδομένα του Engel (1857)

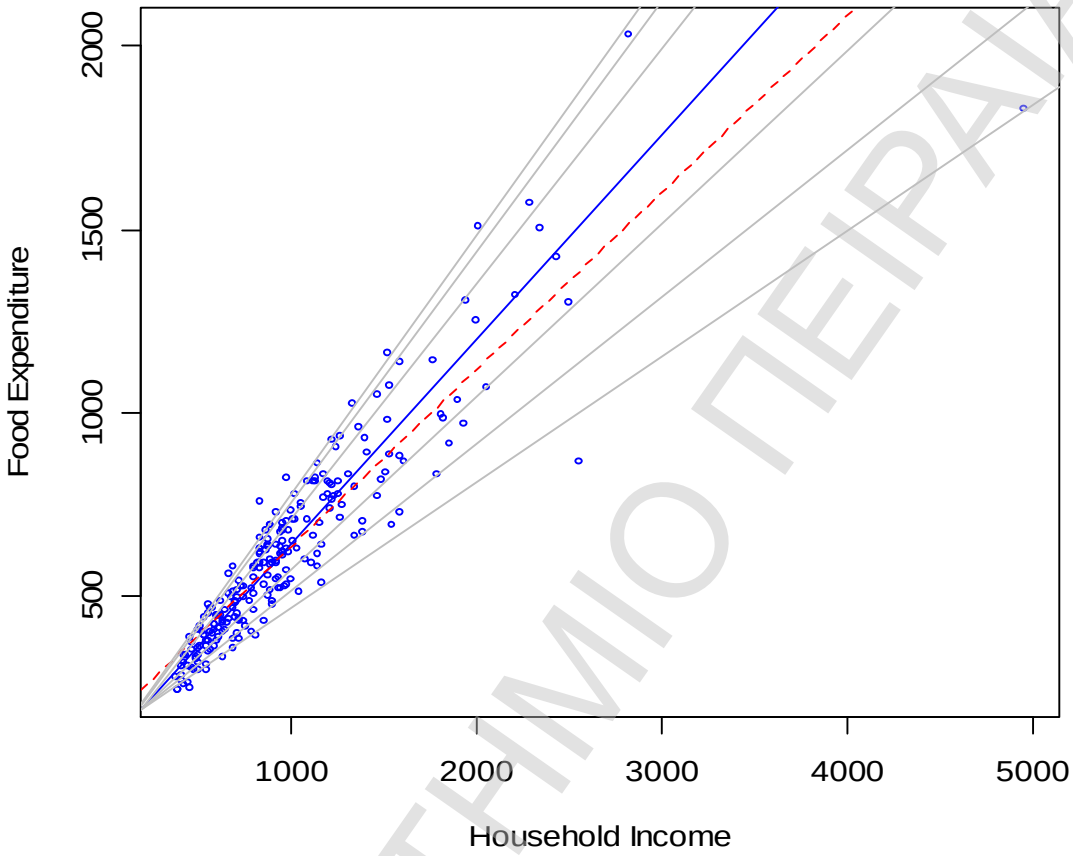
Το 1857 στο Βέλγιο ο Engel έκανε μία έρευνα για να αναλύσει τη σχέση μεταξύ των εξόδων ενός νοικοκυριού για τρόφιμα και του εισοδήματός του. Χρησιμοποίησε δεδομένα από 235 νοικοκυριά της εργατικής τάξης και θεώρησε ότι το εισόδημα ενός νοικοκυριού αποτελεί επεξηγηματική μεταβλητή για τα έξοδα για τρόφιμα, που είναι η εξαρτημένη μεταβλητή.

Τα δεδομένα του Engel υπάρχουν στη βιβλιοθήκη της **R** στο πακέτο "quantreg". Μπορεί κανένας να τα δει απλά πλήκτρολογώντας τις εντολές:

```
> data(engel)
> engel
```

εφόσον έχει εγκαταστήσει και ενεργοποιήσει το πακέτο "quantreg" (βλ. Παράρτημα Α).

Στα δεδομένα αυτά μπορούμε να εφαρμόσουμε τη μέθοδο της ΠΠΣ καθώς και την κλασική μέθοδο των ελαχίστων τετραγώνων και να βγάλουμε χρήσιμα συμπεράσματα. Παρακάτω βλέπουμε οπτικοποιημένα τα αποτελέσματα της εφαρμογής της ΠΠΣ και της ΚΓΠ, αφού έχουμε τις αντίστοιχες προσαρμοσμένες ευθείες στο διάγραμμα διασποράς με τα δεδομένα του Engel. Η ευθεία της ΚΓΠ είναι η διακεκομμένη, ενώ η ευθεία της ΠΔ είναι η συνεχής. Οι υπόλοιπες έξι ευθείες χρώματος γκρι είναι οι προσαρμοσμένες ευθείες παλινδρόμησης για τα ποσοστιαία σημεία με $\tau = 0.05, 0.1, 0.25, 0.75, 0.9, 0.95$ αντίστοιχα από την ευθεία με τη μικρότερη κλίση μέχρι την ευθεία με τη μεγαλύτερη κλίση:

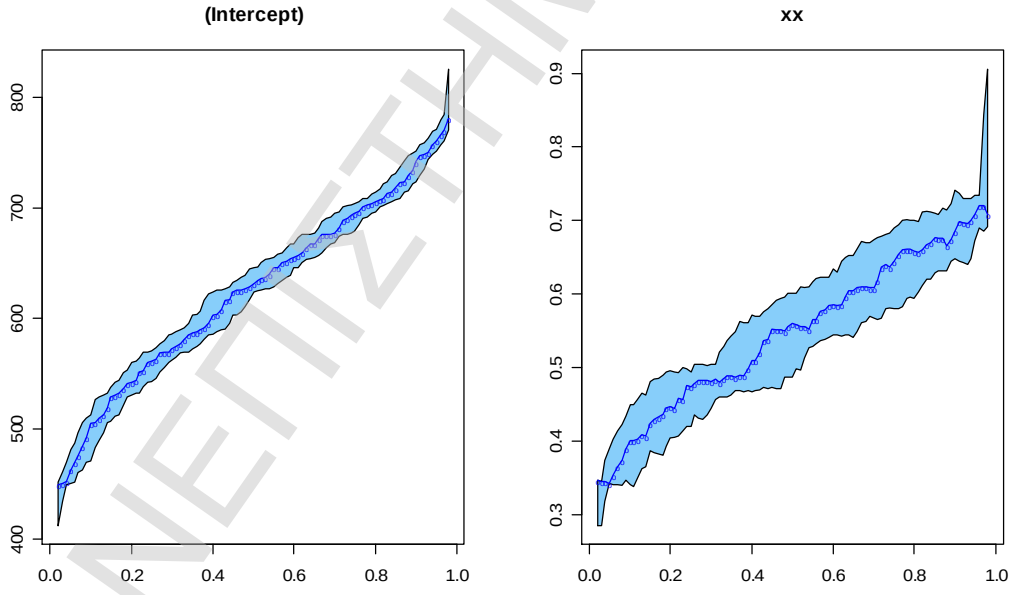


ΣΧΗΜΑ 2-5: Διάγραμμα διασποράς των δεδομένων του Engel με τις προσαρμοσμένες ευθείες παλινδρόμησης.

Από το διάγραμμα αυτό φαίνεται καθαρά η αυξητική τάση της διασποράς των εξόδων για τρόφιμα καθώς αυξάνεται το εισόδημα του νοικοκυριού (ετεροσχεδαστικότητα). Τα διαστήματα μεταξύ των προσαρμοσμένων ευθειών της ΠΠΣ δείχνουν ότι η δεσμευμένη κατανομή της μεταβλητής των εξόδων για τρόφιμα είναι λοξή προς τα αριστερά: τα στενότερα διαστήματα των ευθειών των ανώτερων ποσοστιαίων σημείων δηλώνουν μεγάλη πυκνότητα στην πάνω «ουρά» της κατανομής, ενώ τα πλατύτερα διαστήματα των ευθειών των κατώτερων ποσοστιαίων σημείων δηλώνουν μικρή πυκνότητα στην κάτω «ουρά» της κατανομής. Όσον αφορά τις προσαρμογές των

ευθειών της διαμέσου και της ΚΓΠ, βλέπουμε ότι σε αυτό το παράδειγμα αυτές είναι αρκετά διαφορετικές. Αυτό εξηγείται από την ασυμμετρία της δεσμευμένης πυκνότητας που περιγράψαμε παραπάνω, καθώς και από τις δύο ακραίες τιμές που παρατηρούνται στο δείγμα και φαίνονται στο παραπάνω διάγραμμα. Είχαμε δηλαδή δύο νοικοκυριά με σχετικά χαμηλά έξοδα για τρόφιμα και υψηλό εισόδημα, που επηρέασαν την ευθεία της ΚΓΠ «τραβώντας» την προς τα κάτω αλλά άφησαν σχεδόν ανεπηρέαστες τις ευθείες της ΠΠΣ. Αυτό το χαρακτηριστικό της έλλειψης ανθεκτικότητας της ΚΓΠ έχει ως συνέπεια η συγκεκριμένη μέθοδος να μας δίνει μία φτωχή εκτίμηση της δεσμευμένης μέσης τιμής για τα φτωχότερα νοικοκυριά του δείγματος, μιας και η αντίστοιχη ευθεία περνάει πάνω από τη συντριπτική πλειοψηφεία των παρατηρήσεων με πολύ χαμηλή τιμή στη μεταβλητή του εισοδήματος.

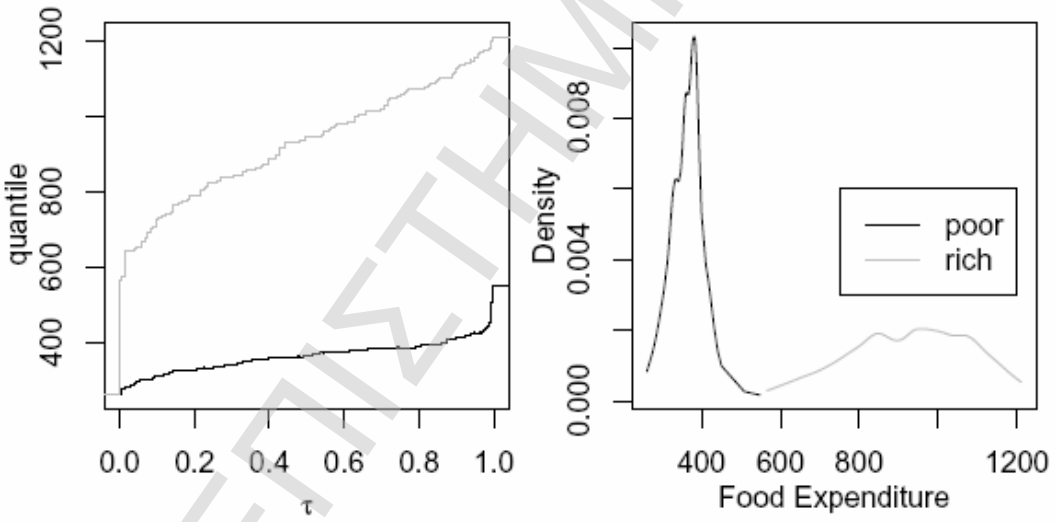
Παρακάτω παριστάνονται οι εκτιμήσεις της ΠΠΣ με ένα 90% διάστημα εμπιστοσύνης για τον σταθερό όρο και το συντελεστή της επεξηγηματικής μεταβλητής, δηλαδή του εισοδήματος του νοικοκυριού:



ΣΧΗΜΑ 2-6: Παλινδρόμηση Ποσοστιαίων Σημείων για τα δεδομένα του Engel.

Στα παραπάνω διαγράμματα η μεταβλητή του εισοδήματος του νοικοκυριού είναι κεντραρισμένη στη μέση τιμή της, επομένως στο διάγραμμα με το σταθερό όρο παριστάνονται οι εκτιμήσεις της ΠΠΣ για τα έξοδα για τρόφιμα ενός νοικοκυριού με μέσο εισόδημα. Στο δεύτερο διάγραμμα με τις εκτιμήσεις για το συντελεστή της επεξηγηματικής μεταβλητής φαίνεται ότι το εισόδημα του νοικοκυριού επηρεάζει περισσότερο τα έξοδα για τρόφιμα στα ανώτερα ποσοστιαία σημεία και λιγότερο στα χαμηλότερα.

Τέλος ενδιαφέρον παρουσιάζουν και τα παρακάτω δύο διαγράμματα, όπου παριστάνονται η εκτιμώμενη συνάρτηση παλινδρόμησης ποσοστιαίων σημείων και η αντίστοιχη εκτιμώμενη εμπειρική συνάρτηση πυκνότητας για τα σχετικά φτωχά νοικοκυριά με εισόδημα *504.5* Βελγικά φράγκα (0.1-ποσοστιαίο σημείο) και τα σχετικά πλούσια με εισόδημα *1538.99* Βελγικά φράγκα (0.9-ποσοστιαίο σημείο). Τα φτωχά νοικοκυριά παριστάνονται με τη μαύρη γραμμή και τα πλούσια με τη γκρι:



ΣΧΉΜΑ 2-7: Συνάρτηση παλινδρόμησης ποσοστιαίων σημείων και εμπειρική συνάρτηση πυκνότητας για τα *0.1* και *0.9* ποσοστιαία σημεία των δεδομένων του Engel.

2.4 Βάρος νεογέννητων (Abreveya 2001)

Στο παράδειγμα αυτό θα δούμε μία επέκταση της έρευνας του Abreveya για την επίδραση διάφορων δημογραφικών και κοινωνικών χαρακτηριστικών της μητέρας στα βάρη των νεογέννητων στις Η.Π.Α.. Το χαμηλό βάρος γέννας είναι συνδεδεμένο με μία σειρά άλλων προβλημάτων υγείας που ακολουθούν. Για το λόγο αυτόν έχει δοθεί ιδιαίτερο ενδιαφέρον στους παράγοντες που επηρεάζουν το βάρος γέννας με σκοπό να μειωθεί η συχνότητα των νεογέννητων χαμηλού βάρους.

Στο μεγαλύτερο μέρος της ανάλυσης το προβλήματος έχει χρησιμοποιηθεί η ΚΓΠ, με τη χρήση της μεθόδου ελαχίστων τετραγώνων. Παρόλα αυτά έχει διαπιστωθεί ότι οι εκτιμήσεις των διαφόρων επιδράσεων στη δεσμευμένη μέση τιμή της εξαρτημένης μεταβλητής, που στην περίπτωσή μας είναι το βάρος γέννας, δεν ήταν πάντα ενδεικτικές του μεγέθους και της φύσης των επιδράσεων αυτών στα χαμηλότερα ποσοστιαία σημεία της εν λόγω μεταβλητής. Στην προσπάθεια να επικεντρωθεί η ανάλυση στη συγκεκριμένη περιοχή της κατανομής χρήσιμη ήταν η συμβολή της μεθόδου της ΠΠΣ. Έχοντας εκτιμήσει, λοιπόν, μία οικογένεια δεσμευμένων συναρτήσεων ποσοστιαίων σημείων, δόθηκε μία πιο πλήρη εικόνα των επιδράσεων των επεξηγηματικών μεταβλητών.

Το δείγμα της μελέτης προήλθε από μητέρες εγγεγραμμένες ως λευκές ή έγχρωμες, μεταξύ 18 και 45 ετών που κατοικούσαν στις Η.Π.Α.. Το μέγεθος του δείγματος ήταν 198377 νεογέννητα. Όπως είπαμε η εξαρτημένη μεταβλητή ήταν το *βάρος γέννας* ενώ οι επεξηγηματικές μεταβλητές ήταν οι παρακάτω:

► Η *εκπαίδευση* της μητέρας (κατηγορική), που χωρίστηκε σε τέσσερις κατηγορίες:

- *«κάτω από λύκειο»*
- *«λύκειο»*
- *«πάνω από λύκειο» (“some college”)*
- *«απόφοιτος πανεπιστημίου»*

Η κατηγορία που παραλήφθηκε ήταν η «*κάτω από λύκειο*», γι' αυτό και οι συντελεστές ερμηνεύτηκαν σε σχέση με αυτήν την κατηγορία.

► Η *ιατρική περίθαλψη* της μητέρας πριν από τη γέννα (κατηγορική), που χωρίστηκε σε τέσσερις κατηγορίες:

- «*καμία επίσκεψη πριν τη γέννα*»
- «*η πρώτη επίσκεψη πριν τη γέννα στο πρώτο τρίμηνο της εγκυμοσύνης*»
- «*η πρώτη επίσκεψη πριν τη γέννα στο δεύτερο τρίμηνο της εγκυμοσύνης*»
- «*η πρώτη επίσκεψη πριν τη γέννα στο τελευταίο τρίμηνο της εγκυμοσύνης*»

Η κατηγορία που παραλήφθηκε ήταν η «*η πρώτη επίσκεψη πριν τη γέννα στο πρώτο τρίμηνο της εγκυμοσύνης*».

► Το *φύλο του νεογέννητου* (κατηγορική)*:*

- «*αγόρι*»
- «*κορίτσι*»

► Η *οικογενειακή κατάσταση της μητέρας* (κατηγορική).

- «*παντρεμένη*»
- «*ανύπαντρη*»

► Η *ηλικία της μητέρας* (συνεχής).

► Το *τετράγωνο της ηλικίας της μητέρας* (συνεχής).

► Αν η μητέρα ήταν *καπνίστρια* ή όχι (κατηγορική).

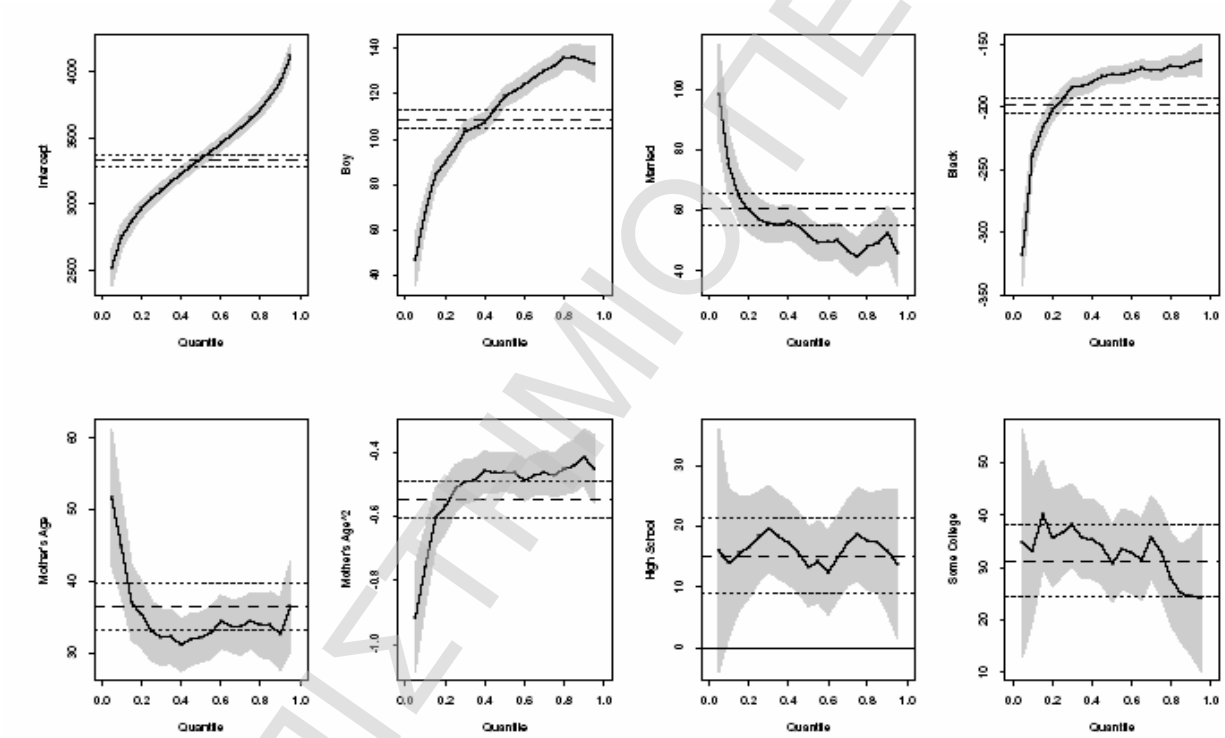
- «*καπνίστρια*»
- «*μη καπνίστρια*»

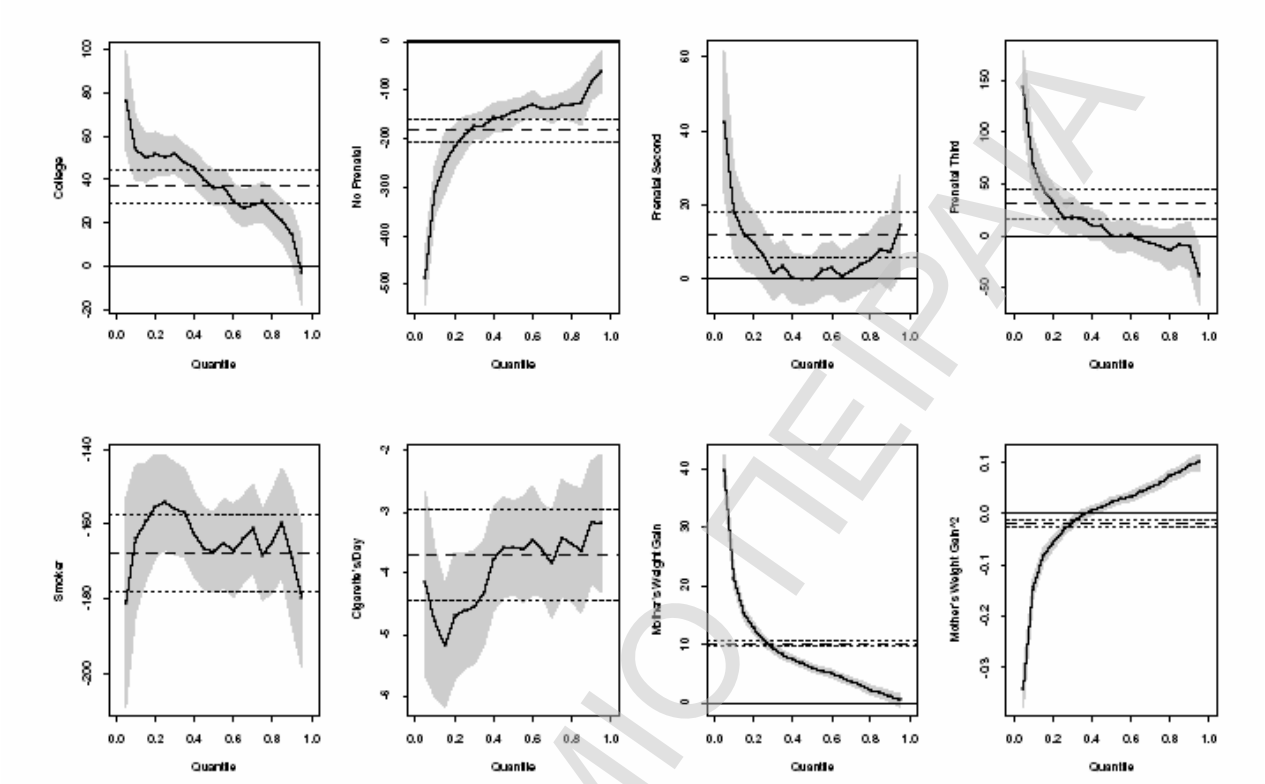
► Ο *αριθμός των τσιγάρων που κάπνιζε ανά ημέρα* (συνεχής).

► Τα *κιλά που πήρε κατά την εγκυμοσύνη* (συνεχής).

► Το *τετράγωνο των κιλών που πήρε κατά την εγκυμοσύνη* (συνεχής).

Τα παρακάτω σχήμα μας δίνει οπτικοποιημένα όλα τα αποτελέσματα της εφαρμογής της μεθόδου της ΠΠΣ στα δεδομένα:





ΣΧΉΜΑ 2-8: Παλινδρόμηση Ποσοστιαίων Σημείων για το βάρος γέννας. (Πηγή: Koenker, Hallock, 2001)

Σε κάθε ένα από τα παραπάνω διαγράμματα παριστάνεται ο συντελεστής ΠΠΣ για την αντίστοιχη επεξηγηματική μεταβλητή. Η έντονη γραμμή είναι η γραμμή των εκτιμήσεων $\hat{b}_j(t)$ με j=1,2,…,16 για τα διάφορα ποσοστιαία σημεία *τ*, όπως φαίνονται στον οριζόντιο άξονα, ενώ η γκρίζα περιοχή πάνω και κάτω από τη γραμμή αυτή παριστάνει ένα 90% διάστημα εμπιστοσύνης για τις σημειακές αυτές εκτιμήσεις. Η διακεκομμένη γραμμή σε κάθε διάγραμμα μας δείχνει την εκτίμηση της μέσης επίδρασης της επεξηγηματικής μεταβλητής με την κλασική μέθοδο των ελαχίστων τετραγώνων ενώ οι δύο παράλληλες γραμμές που βρίσκονται πάνω και κάτω από τη γραμμή της μέσης επίδρασης μας δίνουν αντίστοιχα το πάνω και το κάτω άκρο ενός 90% διαστήματος εμπιστοσύνης για την εκτίμηση αυτή. Τέλος, σε κάποια διαγράμματα υπάρχει μία

οριζόντια ευθεία γραμμή που μας δείχνει σε ποιο σημείο η γραμμή των εκτιμήσεων περνάει από το 0.

Ο συντελεστής παλινδρόμησης σε ένα δοθέν ποσοστιαίο σημείο δείχνει την επίδραση που έχει στο *βάρος γέννας* η αλλαγή μίας μονάδας της αντίστοιχης επεξηγηματικής μεταβλητής, θεωρώντας ότι οι υπόλοιπες παραμένουν σταθερές.

Στο πρώτο διάγραμμα του σχήματος ο σταθερός όρος του μοντέλου (intercept) μπορεί να ερμηνευτεί ως το εκτιμώμενο δεσμευμένο ποσοστιαίο σημείο της κατανομής της μεταβλητής του βάρους γέννας ενός κοριτσιού που γεννήθηκε από ανύπαντρη, λευκή μητέρα, με μόρφωση μικρότερη από λύκειο, είναι 27 ετών, κατά την εγκυμοσύνη της πήρε 13 κιλά, δεν κάπνιζε και η πρώτη της επίσκεψη στο γιατρό ήταν στο πρώτο τρίμηνο της εγκυμοσύνης. Η ηλικία της μητέρας και τα κιλά που πήρε κατά τη διάρκεια της εγκυμοσύνης της έχουν επιλεγεί να παίρνουν τη μέση τιμή των αντίστοιχων μεταβλητών στο δείγμα.

Από το δεύτερο διάγραμμα βλέπουμε ότι σύμφωνα με την εκτίμηση της μέσης επίδρασης με τη μέθοδο των ελαχίστων τετραγώνων τα αγόρια που γεννιούνται είναι περίπου 110 γραμμάρια βαρύτερα από τα κορίτσια. Είναι όμως φανερό από τα αποτελέσματα της ΠΠΣ ότι η διαφορά αυτή είναι πολύ μικρότερη στα χαμηλότερα ποσοστιαία σημεία της κατανομής και κάπως μεγαλύτερη από 110 γραμμάρια στα ανώτερα ποσοστιαία σημεία της κατανομής. Για παράδειγμα τα νεογέννητα αγόρια είναι 45 γραμμάρια βαρύτερα από τα κορίτσια στο 0.05-ποσοστιαίο σημείο, ενώ είναι περίπου 130 γραμμάρια βαρύτερα στο 0.95-ποσοστιαίο σημείο. Γενικότερα σε κάθε ποσοστιαίο σημείο φαίνεται από το διάγραμμα ποια είναι η διαφορά μεταξύ αγοριών και κοριτσιών όσον αφορά το βάρος τους, δεδομένου ότι οι υπόλοιπες επεξηγηματικές μεταβλητές παραμένουν σταθερές. Είναι επίσης φανερό ότι το διάστημα εμπιστοσύνης για τη μέση επίδραση δεν αναπαριστά το πραγματικό εύρος αυτών των διαφορών.

Στο τρίτο διάγραμμα βλέπουμε ότι αν η μητέρα είναι παντρεμένη, φαίνεται να υπάρχει μία θετική επίδραση στο βάρος του νεογέννητού της κυρίως στα χαμηλά ποσοστιαία σημεία της κατανομής. Και σε αυτή την περίπτωση το διάστημα εμπιστοσύνης για τη μέση επίδραση δεν δείχνει την πραγματική εικόνα όλου του εύρους της κατανομής.

Στο τέταρτο διάγραμμα παρατηρούμε ότι τα βάρη των νεογέννητων από μαύρες μητέρες είναι πολύ διαφορετικά από τα αντίστοιχα από λευκές μητέρες, κυρίως στα χαμηλά ποσοστιαία σημεία της κατανομής του βάρους της γέννας. Ενδεικτικά, η διαφορά στα βάρη γέννας μεταξύ ενός νεογέννητου από μαύρη μητέρα και ενός άλλου από λευκή μητέρα στο 0.05-ποσοστιαίο σημείο είναι περίπου το 1/3 του κιλού υπέρ του νεογέννητου από λευκή μητέρα. Και πάλι το διάστημα εμπιστοσύνης για τη μέση επίδραση δεν είναι αντιπροσωπευτικό.

Στα επόμενα δύο διαγράμματα έχουμε δύο συνεχείς μεταβλητές, την ηλικία της μητέρας και το τετράγωνό της. Αυτό που έχουμε να παρατηρήσουμε είναι ότι στα πολύ χαμηλά ποσοστιαία σημεία η ηλικία της μητέρας έχει μεγαλύτερη επίδραση στο βάρος του νεογέννητου σε σχέση με τα υπόλοιπα ποσοστιαία σημεία.

Τα τρία διαγράμματα που ακολουθούν μας δείχνουν τις επιδράσεις του παράγοντα της εκπαίδευσης της μητέρας. Η εκπαίδευση που φτάνει μέχρι την αποφοίτηση από το λύκειο έχει μία αρκετά ομοιόμορφη επίδραση σε όλο το εύρος της κατανομής γύρω στα +15 γραμμάρια. Δηλαδή οι μητέρες που έχουν τελειώσει το λύκειο γεννούν περίπου 15 γραμμάρια βαρύτερα νεογέννητα σε σχέση με αυτές που δεν έχουν τελειώσει το λύκειο. Στη συνέχεια βλέπουμε ότι οι μητέρες που έχουν κάποια εκπαίδευση άνω του λυκείου έχουν θετική επίδραση στο βάρος γέννας κυρίως στα χαμηλότερα ποσοστιαία σημεία της κατανομής. Η επίδραση αυτής της κατηγορίας ποικίλει από 35 γραμμάρια στα χαμηλά ποσοστιαία σημεία σε 25 γραμμάρια στα ανώτερα ποσοστιαία σημεία. Στα δύο τελευταία αυτά διαγράμματα τα διαστήματα εμπιστοσύνης για τις μέσες επιδράσεις είναι αντιπροσωπευτικά για το εύρος της κατανομής, κάτι που δεν ισχύει στο διάγραμμα της επίδρασης της πανεπιστημιακής εκπαίδευσης της μητέρας. Το πτυχίο πανεπιστημίου έχει μία πολύ θετική επίδραση στα χαμηλά ποσοστιαία σημεία, άλλα όσο προχωράμε στα ανώτερα ποσοστιαία σημεία της κατανομής η επίδραση αυτή όλο και μειώνεται.

Παρακάτω έχουμε την επίδραση της ιατρικής περίθαλψης της μητέρας κατά τη διάρκεια της εγκυμοσύνης. Τα νεογέννητα από μητέρες που δεν έκαναν καμία επίσκεψη στο γιατρό πριν τη γέννα βρέθηκαν να είναι περίπου 150 γραμμάρια ελαφρύτερα από εκείνα που γεννήθηκαν από μητέρες που έκαναν την πρώτη τους επίσκεψη στο πρώτο τρίμηνο της εγκυμοσύνης. Στα χαμηλότερα ποσοστιαία σημεία της κατανομής η

επίδραση αυτή είναι αρκετά μεγαλύτερη. Μάλιστα στο 0.05-ποσοστιαίο σημείο τα νεογέννητα αυτά είναι περίπου μισό κιλό ελαφρύτερα. Αντίθετα, οι μητέρες που ακύρωσαν τις επισκέψεις μέχρι το δεύτερο ή το τρίτο τρίμηνο στα χαμηλά ποσοστιαία σημεία της κατανομής γεννούν βαρύτερα μωρά από τις μητέρες που έκαναν την πρώτη τους επίσκεψη στο πρώτο τρίμηνο. Το τελευταίο αυτό αποτέλεσμα είναι σίγουρα αρκετά απρόσμενο, αλλά ίσως ερμηνεύεται από το γεγονός η επιλογή της επίσκεψης ανήκει στην ίδια τη μητέρα και οι μητέρες αυτές αισθάνονταν σίγουρες για το αποτέλεσμα της γέννας τους, ακυρώνοντας τις πρώτες επισκέψεις. Πάντως, αξίζει να σημειωθεί ότι στα πάνω 3/4 της κατανομής οι αντίστοιχες επιδράσεις δεν είναι σημαντικές.

Όσον αφορά το κάπνισμα από τα αντίστοιχα διαγράμματα βλέπουμε ότι η επίδρασή του στο βάρος των νεογέννητων είναι σταθερά αρνητική σε όλο το εύρος της κατανομής. Οι μητέρες που καπνίζουν γεννούν ελαφρύτερα νεογέννητα κατά μέσο όρο 175 γραμμάρια. Στη συνέχεια βλέπουμε την επίδραση των τσιγάρων που καπνίζει η μητέρα την ημέρα. Εδώ υπάρχει μία αρνητική επίδραση γύρω στα 4 με 5 γραμμάρια ανά τσιγάρο την ημέρα σε όλο το εύρος της κατανομής. Δηλαδή μία μητέρα που καπνίζει ένα πακέτο την ημέρα εμφανίζεται να έχει ως αποτέλεσμα ένα νεογέννητο περίπου 80 με 100 γραμμάρια ελαφρύτερο σε σχέση με μία μη καπνίστρια. Τα παραπάνω συμπεράσματα ισχύουν για όλο το εύρος της κατανομής, κάτι που φαίνεται και από το γεγονός ότι στα αντίστοιχα διαγράμματα για το κάπνισμα η οριζόντια γραμμή της μέσης επίδρασης που προκύπτει από την ΚΓΠ είναι σχεδόν όλη καλυμμένη από το 90% σκιασμένο διάστημα εμπιστοσύνης της ΠΠΣ.

Τέλος μελετάται η επίδραση στο βάρος του νεογέννητου της συνεχούς μεταβλητής του βάρους που πήρε η μητέρα κατά την εγκυμοσύνη της. Όπως θα περίμενε κανείς, ο παράγοντας αυτός βρέθηκε να έχει πολύ ισχυρή ανάλογη επίδραση στην εξαρτημένη μεταβλητή, γεγονός που φαίνεται και από τα πολύ στενά διαστήματα εμπιστοσύνης τόσο της ΚΓΠ όσο και της ΠΠΣ. Και πάλι στα χαμηλότερα ποσοστιαία σημεία έχουμε μεγαλύτερη επίδραση του παράγοντα σε σχέση με τα ανώτερα ποσοστιαία σημεία.

Από το τελευταίο διάγραμμα του τετραγώνου της παραπάνω μεταβλητής φαίνεται πόσο παραπλανητικά μπορούν να είναι τα αποτελέσματα της ΚΓΠ. Η εκτίμηση της ΚΓΠ δείχνει ότι η επίδραση του τετραγώνου του βάρους που πήρε η μητέρα κατά την

εγκυμοσύνη δεν είναι σημαντική, δηλαδή υπάρχει μόνο γραμμική επίδραση. Αντίθετα οι εκτιμήσεις της ΠΠΣ δίνουν μία εντελώς διαφορετική εικόνα, όπου η τετραγωνική επίδραση είναι αρκετά σημαντική σε όλο το εύρος της κατανομής εκτός από το σημείο που η αντίστοιχη γραμμή των εκτιμήσεων διέρχεται από το 0, δηλαδή το ποσοστιαίο σημείο τ=0.33.

2.5 Βουλευτικές εκλογές (Ελλάδα, Μάρτιος 2004)

Στην εφαρμογή αυτή χρησιμοποιούμε ένα σετ δεδομένων με τις ψήφους που πήρε κάθε κόμμα σε 47 εκλογικά τμήματα του δήμου Αγίων Αναργύρων στις τελευταίες Ελληνικές βουλευτικές εκλογές της 7^{ης} Μαρτίου του 2004 (βλ. Παράρτημα).

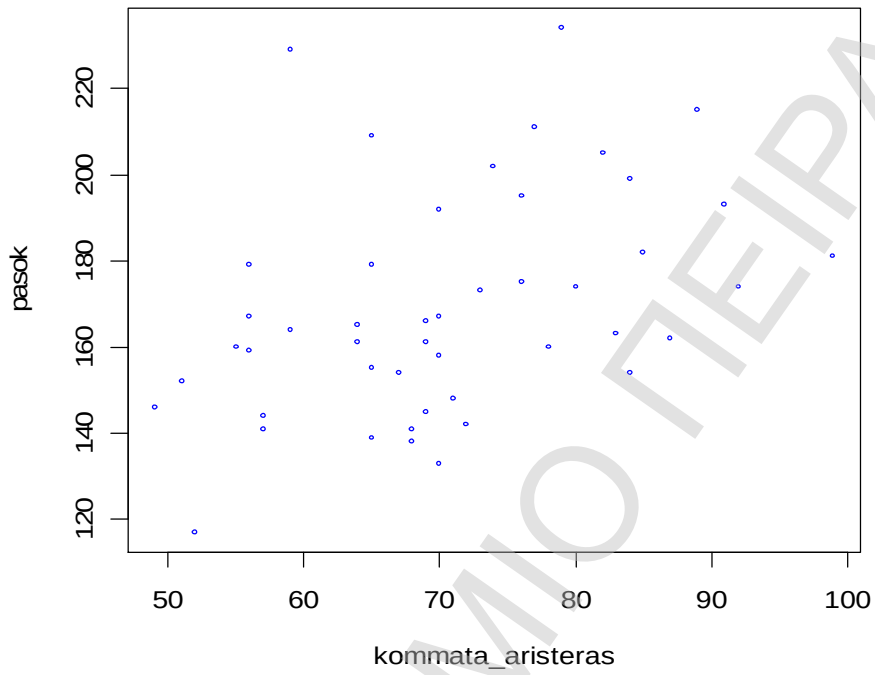
Επιλέξαμε να εξετάσουμε τη σχέση που υπάρχει μεταξύ των ψήφων των κομμάτων που χαρακτηρίζονται ως «αριστερά» και των ψήφων του ΠΑ.ΣΟ.Κ. Συγκεκριμένα μέσω της μεθόδου της ΠΠΣ, όπως θα δούμε παρακάτω μελετήσαμε τον τρόπο που το σύνολο των ψήφων της αριστεράς επηρεάζει όχι μόνο τη μέση τιμή, αλλά όλο το φάσμα της κατανομής των ψήφων που παίρνει το ΠΑ.ΣΟ.Κ. Έτσι ως εξαρτημένη μεταβλητή έχουμε θεωρήσει τις ψήφους του ΠΑ.ΣΟ.Κ. σε κάθε εκλογικό τμήμα και ως μοναδική επεξηγηματική μεταβλητή έχουμε θεωρήσει το σύνολο των ψήφων του Συνασπισμού, του Κ.Κ.Ε. και του ΔΗ.Κ.ΚΙ. σε κάθε εκλογικό τμήμα.

Αρχικά ας δούμε τον πίνακα με τα περιγραφικά στατιστικά των δύο μεταβλητών:

Descriptive Statistics		
	pasok	aristera
Mean	169.43	70.54
Median	164	70
Standard Deviation	25.92	11,83
Sample Variance	671.64	139.95
Range	117	50
Minimum	117	49
Maximum	234	99

ΠΙΝΑΚΑΣ 2-4: Πίνακας περιγραφικών στατιστικών για τα δεδομένα των βουλευτικών εκλογών.

Το διάγραμμα διασποράς των δύο μεταβλητών είναι το παρακάτω:



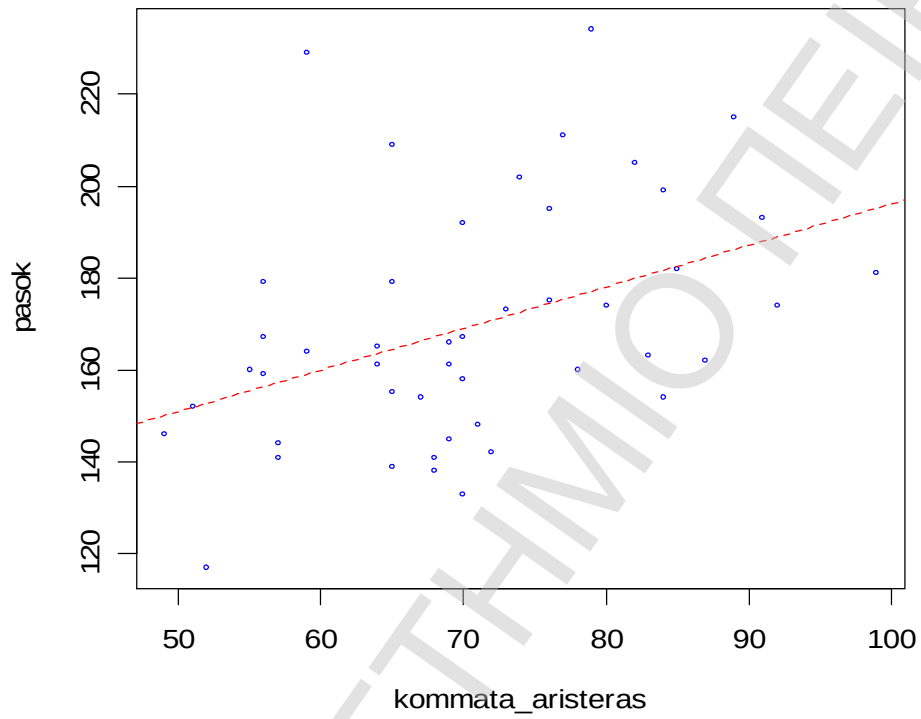
ΣΧΗΜΑ 2-9: Διάγραμμα διασποράς των ψήφων ΠΑ.ΣΟ.Κ. με τις ψήφους της αριστεράς.

Το παραπάνω σχήμα δε μας δίνει πολύ σαφή εικόνα, αν και φαίνεται ότι η εφαρμογή της ΚΓΠ δε θα μας δώσει αξιόπιστα αποτελέσματα λόγω της ετεροσχεδαστικότητας που παρατηρείται. Αντίθετα, όπως έχουμε πει, η ΠΠΣ σε τέτοια δεδομένα είναι πιο χρήσιμη. Πράγματι, η εφαρμογή της ΚΓΠ μας δίνει τα παρακάτω αποτελέσματα:

Μέθοδος Κλασικής Γραμμικής Παλινδρόμησης					
Μοντέλο: $E(y_i \mid x_i) = \hat{\beta}_0^{LS} + x_i^T \hat{\beta}_1^{LS}$, $R^2 = 16{,}98\%$					
$\hat{\beta}_0^{LS}$	105.71	Std.Error	21.28	p-value	0
$\hat{\beta}_1^{LS}$	0.90	Std.Error	0.29	p-value	0,004

ΠΙΝΑΚΑΣ 2-5: Πίνακας Κλασικής Γραμμικής Παλινδρόμησης για τα δεδομένα των βουλευτικών εκλογών.

Από τον παρακάτω πίνακα βλέπουμε ότι η προσαρμογή του μοντέλου της ΚΓΠ είναι πολύ κακή αφού ο συντελεστής παλινδρόμησης R^2 είναι μόλις *16.98%*. Αυτό εξάλλου φαίνεται και από την προσαρμογή της ευθείας παλινδρόμησης, που όπως αναμενόταν, λόγω της διασποράς των δεδομένων, δεν είναι καθόλου καλή:

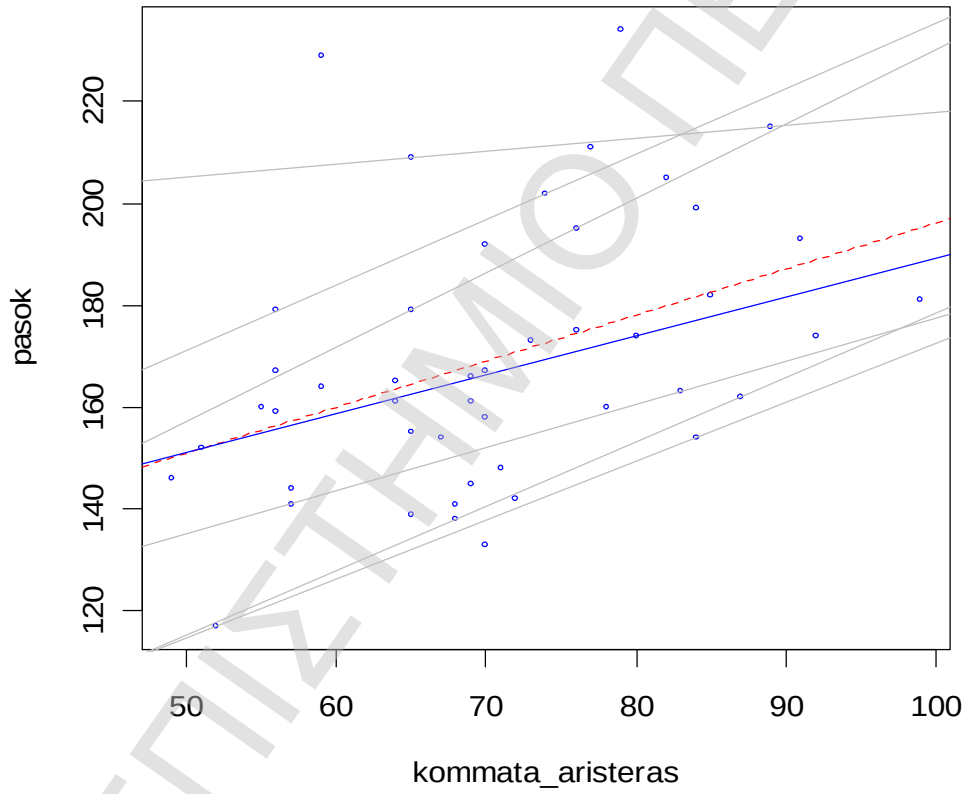


ΣΧΗΜΑ 2-10: Διάγραμμα διασποράς των ψήφων ΠΑ.ΣΟ.Κ. με τις ψήφους της αριστεράς με την προσαρμοσμένη ευθεία ΚΓΠ.

Είναι φανερή και από το παραπάνω σχήμα η ασυμμετρία της δεσμευμένης πυκνότητας της εξαρτημένης μεταβλητής, που είναι οι ψήφοι του ΠΑ.ΣΟ.Κ. Φαίνεται ότι υπάρχει μία αυξητική τάση της διασποράς των ψήφων του ΠΑ.ΣΟ.Κ. καθώς αυξάνονται οι ψήφοι της αριστεράς. Την κατάσταση αυτή μπορεί να περιγράψει καλύτερα, όπως είπαμε, η εφαρμογή της μεθόδου της ΠΠΣ.

Με τη βοήθεια του στατιστικού πακέτου της **R**, εφαρμόζουμε την ΠΠΣ για τα ποσοστιαία σημεία με $\tau = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$ και παίρνουμε το νέο

διάγραμμα διασποράς με τις αντίστοιχες ευθείες παλινδρόμησης για τα παραπάνω ποσοστιαία σημεία, καθώς και τον πίνακα με αποτελέσματα των παλινδρομήσεων αυτών. Στο παρακάτω σχήμα η ευθεία της ΚΓΠ είναι η διακεκομμένη, ενώ η ευθεία της ΠΔ είναι η συνεχής. Οι υπόλοιπες έξι ευθείες χρώματος γκρι είναι οι προσαρμοσμένες ευθείες παλινδρόμησης για τα ποσοστιαία σημεία με $\tau = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$ αντίστοιχα από κάτω προς τα πάνω:



ΣΧΗΜΑ 2-11: Διάγραμμα διασποράς των ψήφων ΠΑ.ΣΟ.Κ. με τις ψήφους της αριστεράς με τις προσαρμοσμένες ευθείες ΚΓΠ και ΠΠΣ για $\tau = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$.

Μέθοδος Παλινδρόμησης Ποσοστιαίων σημείων			
Μοντέλο: $Q_\tau(y_i \mid x_i) = \hat{\beta}_0^{QR} + x_i^T \hat{\beta}_1^{QR}$			
τ=0.05			
$\hat{\beta}_0^{QR}$	56.88	95% δ.ε.	[19.88 , 150.44]
$\hat{\beta}_1^{QR}$	1.16	95% δ.ε.	[-0.46 , 1.41]
τ=0.10			
$\hat{\beta}_0^{QR}$	52.11	95% δ.ε.	[42.55 , 114.75]
$\hat{\beta}_1^{QR}$	1.26	95% δ.ε.	[-0.34 , 1.35]
τ=0.25			
$\hat{\beta}_0^{QR}$	92.77	95% δ.ε.	[57.41 , 125.62]
$\hat{\beta}_1^{QR}$	0.85	95% δ.ε.	[0.43 , 1.25]
τ=0.50			
$\hat{\beta}_0^{QR}$	113.31	95% δ.ε.	[63.77 , 134.09]
$\hat{\beta}_1^{QR}$	0.76	95% δ.ε.	[0.43 , 1.43]
τ=0.75			
$\hat{\beta}_0^{QR}$	84.45	95% δ.ε.	[64.78 , 152.72]
$\hat{\beta}_1^{QR}$	1.45	95% δ.ε.	[0.41 , 1.72]
τ=0,90			
$\hat{\beta}_0^{QR}$	107.44	95% δ.ε.	[43.83 , 276.38]
$\hat{\beta}_1^{QR}$	1.28	95% δ.ε.	[-1.06 , 2.91]
τ=0,95			
$\hat{\beta}_0^{QR}$	192.75	95% δ.ε.	[25.19 , 394.14]
$\hat{\beta}_1^{QR}$	0.25	95% δ.ε.	[-1.03 , 3.47]

ΠΙΝΑΚΑΣ 2-6: Πίνακας Παλινδρόμησης Ποσοστιαίων Σημείων για τ = *0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95* για τα δεδομένα των βουλευτικών εκλογών.

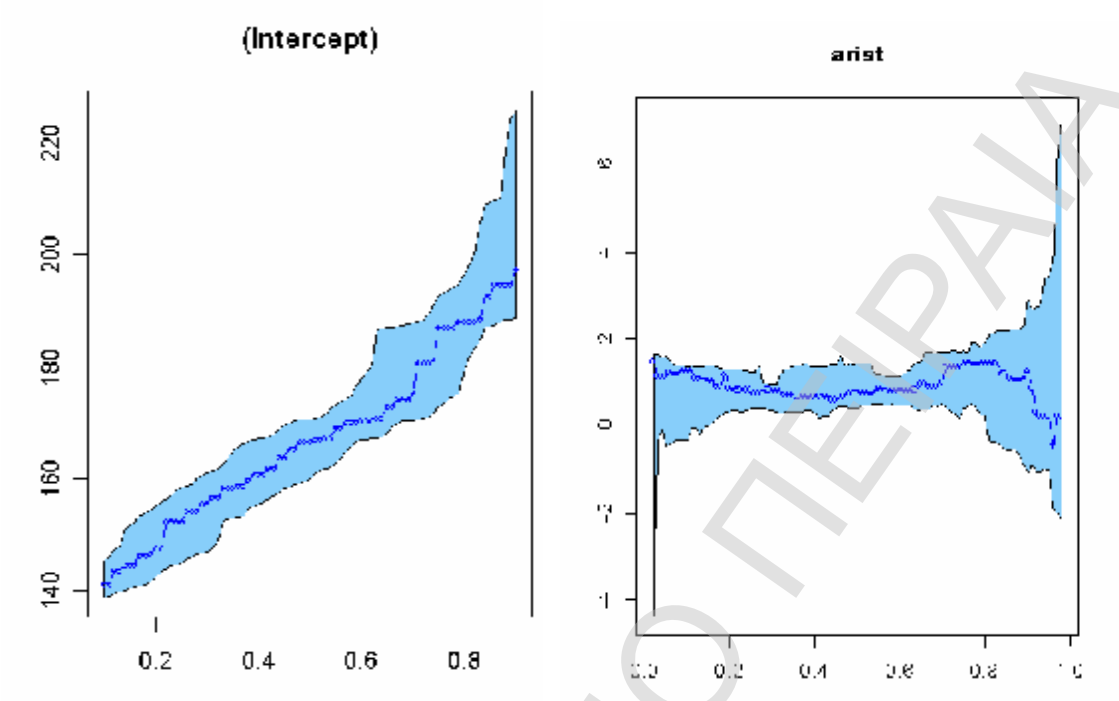
Από τον παραπάνω πίνακα μπορούμε πλέον να βγάλουμε κάποια χρήσιμα συμπεράσματα για την αποτελεσματικότητα της χρήσης της ΠΠΣ στα δεδομένα μας.

Ο εκτιμώμενος σταθερός όρος του μοντέλου $\hat{b}_0^{QR}$, αν και στατιστικά σημαντικός στα ποσοστιαία σημεία που έχουμε επιλέξει, στο παράδειγμά μας δεν έχει κάποια ποιοτική σημασία αφού δηλώνει το πως συμπεριφέρεται το εκάστοτε ποσοστιαίο σημείο της κατανομής των ψήφων του ΠΑ.ΣΟ.Κ. όταν δεν υπάρχει κανένας ψηφοφόρος των τριών κομμάτων της αριστεράς, ενδεχόμενο το οποίο είναι μη ρεαλιστικό.

Επομένως θα επικεντρωθούμε στην ερμηνεία του εκτιμώμενου συντελεστή $\hat{b}_1^{QR}$ που δείχνει την επιρροή του αριθμού των ψήφων της αριστεράς στον αριθμό των ψήφων

του ΠΑ.ΣΟ.Κ. στα διάφορα ποσοστιαία σημεία της κατανομής των τελευταίων. Για τα ποσοστιαία σημεία με $\tau = 0.05, 0.1, 0.9, 0.95$ ο $\hat{b}_1^{QR}$ είναι στατιστικά μη σημαντικός αφού το 0 περιέχεται μέσα στα διαστήματα εμπιστοσύνης για τις αντίστοιχες εκτιμήσεις. Άρα μπορούμε να πούμε ότι τα συγκεκριμένα ποσοστιαία σημεία της κατανομής των ψήφων του ΠΑ.ΣΟ.Κ. δεν επηρεάζονται από τον αριθμό των ψήφων της αριστεράς. Αντίθετα η Παλινδρόμηση των ποσοστιαίων σημείων με $\tau = 0.25$ και 0.75, καθώς και της διαμέσου, αν και οριακά, μας δίνει στατιστικά σημαντικούς συντελεστές. Φαίνεται δηλαδή πως για αυτά τα ποσοστιαία σημεία, υπάρχει μια μικρή θετική επιρροή των ψήφων της αριστεράς στις ψήφους του ΠΑ.ΣΟ.Κ.

Τα παρακάτω διάγραμματα συνοψίζουν τα παραπάνω συμπεράσματα και ταυτόχρονα μας δείχνουν πως συμπεριφέρεται όλο το φάσμα των ποσοστιαίων σημείων της κατανομής του αριθμού των ψήφων του ΠΑ.ΣΟ.Κ. σε σχέση με τον αριθμό των ψήφων των κομμάτων της αριστεράς. Η μεταβλητή των ψήφων της αριστεράς είναι κεντραρισμένη στη μέση τιμή της, επομένως στο διάγραμμα με το σταθερό όρο παριστάνονται οι εκτιμήσεις της ΠΠΣ για τις ψήφους του ΠΑ.ΣΟ.Κ. όταν οι ψήφοι της αριστεράς κυμαίνονται στη μέση τους τιμή. Στο δεύτερο διάγραμμα παριστάνονται οι εκτιμήσεις της ΠΠΣ για το συντελεστή $\hat{b}_1^{QR}$ με ένα 95% διάστημα εμπιστοσύνης. Το ότι όλες οι στατιστικά σημαντικές εκτιμήσεις βρίσκονται πάνω από το 0 δηλώνει τη θετική επιρροή του αριθμού των ψήφων της αριστεράς στον αριθμό των ψήφων του ΠΑ.ΣΟ.Κ, σχεδόν σε όλο το φάσμα της κατανομής. Τέλος, όπου περιέχεται στο διάστημα εμπιστοσύνης το 0, πράγμα που συμβαίνει μόνο στα πολύ χαμηλά και στα πολύ υψηλά ποσοστιαία σημεία της κατανομής, έχουμε μη στατιστική σημαντικότητα για αυτά τα ποσοστιαία σημεία, δηλαδή μηδενική επιρροή του αριθμού των ψήφων της αριστεράς.



ΣΧΗΜΑ 2-12: Παλινδρόμηση Ποσοστιαίων Σημείων για τα δεδομένα των βουλευτικών εκλογών.

2.6 Ασφαλιστικές εταιρίες μηχανοκίνητων οχημάτων (1996-2002)

Η εφαρμογή αυτή είναι ουσιαστικά μία επέκταση της έρευνας του Pitselis (2006,2007) που στις αντίστοιχες εργασίες μελετά μέσω κατάλληλων μοντέλων τα οικονομικά χαρακτηριστικά που πρέπει να έχει μία ασφαλιστική εταιρία ώστε να περιορίσει τα ελλείμματα και να αποφύγει το ενδεχόμενο της πτώχευσης. Τα δεδομένα (Διεύθυνση Ασφαλιστικών Επιχειρήσεων και Αναλογιστικής της Γενικής Διεύθυνσης Εσωτερικού Εμπορίου της Γενικής Γραμματείας Εμπορίου του Υπουργείου Ανάπτυξης) τα πήραμε από τις συγκεκριμένες εργασίες και αφορούν τον ισολογισμό κάποιων ασφαλιστικών εταιριών που δραστηριοποιούνται στην Ελλάδα στον τομέα των μηχανοκίνητων οχημάτων. Συγκεκριμένα, η ανάλυση βασίστηκε σε δεδομένα από 31 φερέγγυες εταιρίες για την επταετία από το 1996 μέχρι και το 2002. Είχαμε δηλαδή 217 παρατηρήσεις για 10 συνεχείς μεταβλητές. Οι μεταβλητές αυτές είναι οι εξής:

► *OF*: Ίδια Κεφάλαια (Own Funds)
► *TC*: Συνολικές Απαιτήσεις (Total Claims)
► *TA*: Συνολικά Περιουσιακά Στοιχεία (Total Assets)
► *CR*: Τρέχων κίνδυνος (Current Risk)
► *OC*: Εκκρεμείς Απαιτήσεις (Outstanding Claims)
► *TP*: Συνολικές Προβλέψεις (Total Provisions)
► *L*: Χρέη (Liabilities)
► *I*: Έσοδα (Incoming)
► *PC*: Πληρωμένες Απαιτήσεις (Paid Claims)
► *EX*: Έξοδα (Expenses)

Παρακάτω έχουμε τον πίνακα με κάποια περιγραφικά στατιστικά των 10 μεταβλητών:

Descriptive Statistics			
Var. Name	Mean	Median	St.Dev.
OF	13781.87	4344.95	24722.04
TC	23292.19	7461.29	42392.41
TA	83082.78	17992.65	199472.86
CR	9249.03	2454.51	18776.73
OC	24929.96	5885.68	75557.13
TP	57491.00	10027.31	159604.78
L	10760.99	4609.84	18291.26
I	29618.66	10338.76	54520.33
PC	16842.63	5956.74	31434.69
EX	26370.51	9055.46	47377.26

ΠΙΝΑΚΑΣ 2-7: Πίνακας περιγραφικών στατιστικών για τα δεδομένα των ασφαλιστικών.

Σκοπός μας είναι να μελετήσουμε τον τρόπο που επηρεάζουν συγκεκριμένα οικονομικά χαρακτηριστικά μίας εταιρίας τα Ίδια Κεφάλαιά της, όχι στη μέση τους τιμή αλλά σε όλο το εύρος της κατανομής τους, δηλαδή για εταιρίες με χαμηλά, μεσαία και υψηλά Ίδια Κεφάλαια. Επομένως θα εφαρμόσουμε την ΠΠΣ με ανεξάρτητες μεταβλητές τις TC, TA, CR, OC, TP, L, I, PC, EX και εξαρτημένη την OF. Κάνοντας τα διαγράμματα διασποράς της OF με κάθε μία από τις ανεξάρτητες μεταβλητές (βλ.

Παράρτημα) θα παρατηρήσουμε ότι τα δεδομένα μας παρουσιάζουν ετεροσχεδαστικότητα, οπότε η ΠΠΣ θα μας δώσει μια πιο πλήρη εικόνα από εκείνη που θα μας έδινε η εφαρμογή της ΚΓΠ. Το πολλαπλό μοντέλο ΠΠΣ που θα προσαρμόσουμε για το τ-ποσοστιαίο σημείο είναι το παρακάτω:

$$Q_t(OF_i \mid TC_i, TA_i, CR_i, OC_i, TP_i, L_i, I_i, PC_i, EX_i) = b_0(t) + b_{TC}(t)\cdot TC_i + b_{TA}(t)\cdot TA_i +$$
$$+ b_{CR}(t)\cdot CR_i + b_{OC}(t)\cdot OC_i + b_{TP}(t)\cdot TP_i + b_L(t)\cdot L_i + b_I(t)\cdot I_i + b_{PC}(t)\cdot PC_i +$$
$$+ b_{EX}(t)\cdot EX_i + u_i(t) \qquad (2.6.1)$$

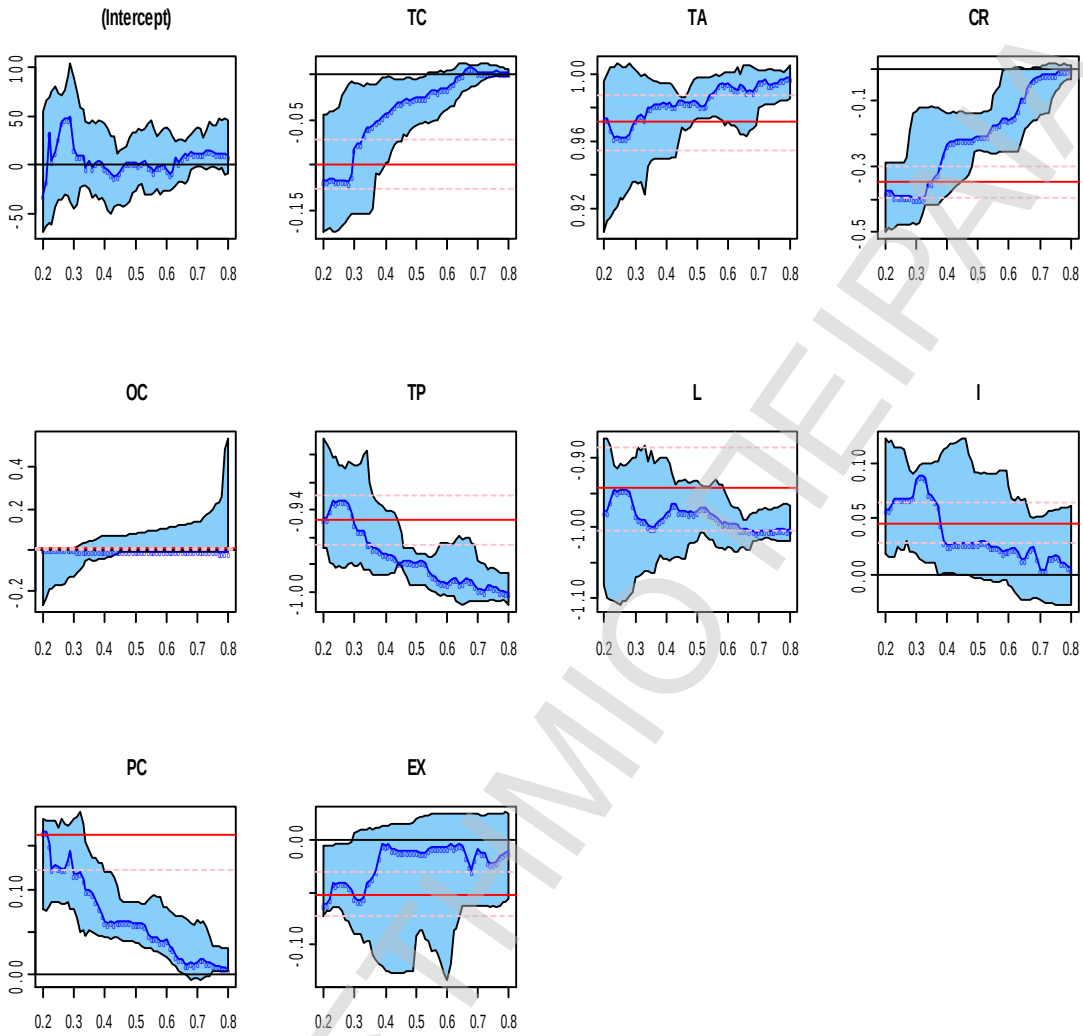
Με τη βοήθεια του στατιστικού πακέτου της **R**, εφαρμόζουμε την ΠΠΣ για τα ποσοστιαία σημεία με *τ = 0.05, 0.25, 0.5, 0.75, 0.95* και παίρνουμε τον παρακάτω πίνακα με αποτελέσματα των παλινδρομήσεων αυτών. Μέσα σε παρένθεση είναι οι τιμές των στατιστικών (t-values) βάσει των οποίων ελέγχουμε τη σημαντικότητα των αντίστοιχων εκτιμήσεων. Οι στατιστικά σημαντικές εκτιμήσεις συνοδεύονται με "*":

Παλινδρόμηση Ποσοστιαίων Σημείων (Μοντέλο 2.6.1)					
	τ = 0.05	*τ = 0.25*	*τ = 0.5*	*τ = 0.75*	*τ = 0.95*
$\hat{b}_0(t)$	-145.336 * (-1.53)	27.535 (0.63)	2.402 (0.27)	14.027 (1.09)	1.878 (0.09)
$\hat{b}_{TC}(t)$	-0.174 * (-4.81)	-0.117 * (-4.61)	-0.026 * (-1.29)	0.003 (0.19)	0.006 * (1.63)
$\hat{b}_{TA}(t)$	0.957 * (32.14)	0.962 * (116.57)	0.983 * (72.53)	0.995 * (98.86)	1.001 *(240.07)
$\hat{b}_{CR}(t)$	-0.366 * (-4.93)	-0.388 * (-6.49)	-0.209 * (-2.06)	-0.014 (-0.34)	0.005 (0.42)
$\hat{b}_{OC}(t)$	-0.014 (-0.36)	0.005 (0.52)	0.000 (0.11)	-0.004 (-0.28)	-0.005 (-30.96)
$\hat{b}_{TP}(t)$	-0.918 * (-24.78)	-0.933 * (-69.67)	-0.978 * (-67.08)	-0.996 * (-83.16)	-1.006 * (-222.93)
$\hat{b}_L(t)$	-0.955 * (-29.53)	-0.947 * (-33.44)	-0.976 * (-27.13)	-1.003 * (-65.41)	-0.976 * (-76.97)
$\hat{b}_I(t)$	0.039 * (0.92)	0.067 * (3.08)	0.029 (0.85)	0.017 (1.16)	0.008 (1.95)
$\hat{b}_{PC}(t)$	0.199 * (1.15)	0.126 * (4.07)	0.060 * (2.59)	0.013 * (0.82)	0.006 (-1.14)
$\hat{b}_{EX}(t)$	-0.031 (-2.12)	-0.041 * (-2.04)	-0.011 (-0.28)	-0.022 (-1.35)	0.003 (-1.59)

ΠΙΝΑΚΑΣ 2-8: Πίνακας Παλινδρόμησης Ποσοστιαίων Σημείων για τ= *0.05, 0.25, 0.5, 0.75, 0.95* για τα δεδομένα των ασφαλιστικών.

Στον παραπάνω πίνακα παρατηρούμε ότι οι περισσότερες εκτιμήσεις είναι στατιστικά σημαντικές, κάτι που συμβαίνει κυρίως για τα χαμηλά ποσοστιαία σημεία, δηλαδή οι εταιρίες με χαμηλά Ίδια Κεφάλαια δείχνουν γενικά να επηρεάζονται περισσότερο από την πλειοψηφία των οικονομικών χαρακτηριστικών του ισολογισμού τους. Επίσης βλέπουμε ότι οι εκτιμήσεις στα διάφορα ποσοστιαία σημεία για τη μεταβλητή OC είναι στατιστικά μη σημαντικές, επομένως τα Ίδια Κεφάλαια δεν επηρεάζονται από τις Εκκρεμείς Απαιτήσεις είτε πρόκειται για εταιρίες με χαμηλά, μεσαία ή υψηλά Ίδια Κεφάλαια.

Μια πιο πλήρη εικόνα μας δίνουν τα παρακάτω διαγράμματα, τα οποία μας δείχνουν πως συμπεριφέρεται όλο το φάσμα των ποσοστιαίων σημείων της κατανομής των Ιδίων Κεφαλαίων. Σε αυτά παριστάνονται οι εκτιμήσεις της ΠΠΣ για τους συντελεστές του μοντέλου (2.6.1) μαζί με τα 95% διαστήματα εμπιστοσύνης. Η οριζόντια γραμμή μας δείχνει που βρίσκεται η εκτίμηση της ΚΓΠ για τη μέση τιμή της OF:



ΣΧΗΜΑ 2-13: Παλινδρόμηση Ποσοστιαίων Σημείων για τα δεδομένα των ασφαλιστικών.

Το πρώτο διάγραμμα αναφέρεται στον σταθερό όρο του μοντέλου και η ερμηνεία του δεν παρουσιάζει κάποιο ενδιαφέρον.

Στο δεύτερο διάγραμμα παρατηρούμε αρχικά ότι όλες οι εκτιμήσεις είναι κάτω από το 0, πράγμα που σημαίνει ότι όσο υψηλότερες Συνολικές Απαιτήσεις έχει μία εταιρία τόσο χαμηλότερα θα είναι τα Ίδια Κεφάλαιά της. Όσον αφορά τη συμπεριφορά της OF βλέπουμε ότι στα χαμηλά ποσοστιαία της σημεία επηρεάζεται σαφώς περισσότερο από ότι στα υψηλά, καθώς οι εκτιμήσεις σε αυτά τα σημεία της κατανομής

είναι κατά απόλυτη τιμή μεγαλύτερες. Επομένως όσο χαμηλότερα Ίδια Κεφάλαια έχει μία εταιρία τόσο περισσότερο θα επηρεάζεται από τις Συνολικές Απαιτήσεις της.

Στο τρίτο διάγραμμα παρατηρούμε μία αρκετά μεγάλη επιρροή των Συνολικών Περιουσιακών Στοιχείων στα Ίδια Κεφάλαια σε όλο το εύρος της κατανομής των τελευταίων. Οι διαφορές για εταιρίες με χαμηλά και υψηλά Ίδια Κεφάλαια δεν είναι μεγάλες, αν και οι εταιρίες με υψηλά Ίδια Κεφάλαια δείχνουν να επηρεάζονται ελαφρώς περισσότερο από τα Συνολικά Περιουσιακά Στοιχεία τους. Το θετικό πρόσημο των εκτιμήσεων δηλώνει τη θετική συνάφεια των δύο μεταβλητών. Δηλαδή όσο υψηλότερα είναι τα Συνολικά Περιουσιακά Στοιχεία των εταιριών, τόσο υψηλότερα θα είναι και τα Ίδια Κεφάλαιά τους.

Στο τέταρτο διάγραμμα βλέπουμε μία κατάσταση παρόμοια με αυτή του δεύτερου. Φαίνεται ότι όσο υψηλότερος είναι ο Τρέχων κίνδυνος που έχει μία εταιρία τόσο χαμηλότερα θα είναι τα Ίδια Κεφάλαιά της. Αυτή η επιρροή είναι περισσότερο έντονη στις εταιρίες με χαμηλά Ίδια Κεφάλαια παρά σε αυτές με υψηλά, στις οποίες φαίνεται να μην επηρεάζονται τα Ίδια Κεφάλαιά τους από τον Τρέχοντα κίνδυνο.

Στο πέμπτο διάγραμμα φαίνεται αυτό που είχαμε διαπιστώσει και από τον πίνακα 2-8, δηλαδή ότι για όλες τις εταιρίες οι Εκκρεμείς Απαιτήσεις δεν επηρεάζουν τα Ίδια Κεφάλαια.

Στο έκτο διάγραμμα παρατηρούμε μία σημαντική αρνητική συνάφεια μεταξύ των μεταβλητών TP και OF, πράγμα που σημαίνει ότι όσο υψηλότερες Συνολικές Προβλέψεις έχει μία εταιρία τόσο χαμηλότερα θα είναι τα Ίδια Κεφάλαιά της. Εδώ οι εταιρίες με υψηλά Ίδια Κεφάλαια επηρεάζονται σαφώς περισσότερο από τις Συνολικές Προβλέψεις τους, αφού οι αντίστοιχες εκτιμήσεις είναι κατά απόλυτη τιμή μεγαλύτερες από τις αντίστοιχες των εταιριών με χαμηλά Ίδια Κεφάλαια.

Στο έβδομο διάγραμμα βλέπουμε ότι τα Χρέη επηρεάζουν αρνητικά τα Ίδια Κεφάλαια μίας εταιρίας, αλλά δε φαίνεται κάποια σημαντική διαφορά όσον αφορά το μέγεθος αυτής της επιρροής στις εταιρίες με χαμηλά κι υψηλά Ίδια Κεφάλαια. Ίσως οι εταιρίες με υψηλά Ίδια Κεφάλαια να επηρεάζονται λίγο περισσότερο από τα Χρέη τους σε σχέση με τις εταιρίες με χαμηλά Ίδια Κεφάλαια.

Στο όγδοο διάγραμμα παρατηρούμε ότι μόνο οι εταιρίες με χαμηλά Ίδια Κεφάλαια επηρεάζονται θετικά από τα Έσοδά τους καθώς στο συγκεκριμένο διάγραμμα από το *0.35*-ποσοστιαίο σημείο και μετά το 0 περιέχεται στο 95% δ.ε..

Στο ένατο διάγραμμα έχουμε θετικές εκτιμήσεις, επομένως όσο υψηλότερες Πληρωμένες Απαιτήσεις έχει μία εταιρία τόσο υψηλότερα θα είναι και τα Ίδια Κεφάλαιά της. Όσον αφορά τη συμπεριφορά της OF δοθείσης της PC, βλέπουμε ότι στα χαμηλότερα ποσοστιαία σημεία επηρεάζεται σαφώς περισσότερο από ότι στα υψηλά. Δηλαδή όσο χαμηλότερα Ίδια Κεφάλαια έχει μία εταιρία τόσο περισσότερο θα επηρεάζεται από τις Πληρωμένες Απαιτήσεις της.

Τέλος στο δέκατο διάγραμμα παρατηρούμε μία αρνητική επιρροή των Εξόδων στα Ίδια Κεφάλαια, που υπάρχει όμως μόνο στα πολύ χαμηλά ποσοστιαία σημεία της κατανομής της OF, καθώς από το *0.25*-ποσοστιαίο σημείο και μετά το 0 περιέχεται στο 95% δ.ε.. Άρα για τις εταιρίες με πολύ χαμηλά Ίδια Κεφάλαια όσο υψηλότερα είναι τα Έξοδα τόσο χαμηλότερα θα είναι τα Ίδια Κεφάλαια. Στις υπόλοιπες εταιρίες δεν επηρεάζονται τα Ίδια Κεφάλαια από τα Έξοδά τους.

Εκτός από τα παραπάνω συμπεράσματα, από τα διαγράμματα του σχήματος 2-13 διαπιστώνουμε και την ελλιπή εικόνα που μας δίνει η ΚΓΠ. Στην πλειοψηφία των διαγραμμάτων το διάστημα εμπιστοσύνης για τη μέση επίδραση δεν είναι αντιπροσωπευτικό.

ΚΕΦΑΛΑΙΟ 3

ΜΙΚΤΑ ΜΟΝΤΕΛΑ ΚΑΙ ΠΑΛΙΝΔΡΟΜΗΣΗ ΠΟΣΟΣΤΙΑΙΩΝ ΣΗΜΕΙΩΝ

3.1 Εισαγωγή

Πολλές φορές, όταν εξετάζουμε τον τρόπο που επηρεάζεται μία μεταβλητή που μας ενδιαφέρει, είναι τέτοιος ο σχεδιασμός του πειράματος, που εκτός από τις επεξηγηματικές μεταβλητές, πρέπει να λάβουμε υπόψη και τους τομείς στους οποίους χωρίζονται τα δεδομένα. Για παράδειγμα, σε ένα ιατρικό πείραμα με μία μεταβλητή που μελετάμε βάσει διαφόρων επεξηγηματικών μεταβλητών, τομείς μπορεί να είναι τα διάφορα ιατρικά κέντρα στα οποία έγιναν οι μετρήσεις. Ωστόσο, στις περισσότερες πρακτικές εφαρμογές τα μεγέθη δείγματος των τομέων δεν είναι αρκετά μεγάλα ώστε να επιτρέπουν εκτίμηση που θα βασίζεται μόνο σε δειγματικές μονάδες του τομέα. Τέτοιοι τομείς ονομάζονται «*μικρές περιοχές*».

Όταν έχουμε μικρές περιοχές, τότε πρέπει να βασιστούμε σε εναλλακτικές μεθόδους για την εκτίμηση των παραμέτρων που μας ενδιαφέρουν. Οι πιο συνήθεις από αυτές τις μεθόδους είναι οι μοντελοκεντρικές μέθοδοι, που ταξινομούνται σε δύο κατηγορίες: σε αυτές που βασίζονται σε Μοντέλα Σταθερών Επιδράσεων (Fixed Effects models), που εξηγούν τη μεταξύ των περιοχών διασπορά της εξαρτημένης μεταβλητής, χρησιμοποιώντας τις επεξηγηματικές και σε αυτές που βασίζονται σε Μικτά Μοντέλα (Random Effects models ή Mixed models), που περιλαμβάνουν και τις τυχαίες επιδράσεις της κάθε περιοχής. Τα μοντέλα που χρησιμοποιούνται ευρέως στην εκτίμηση σε μικρές περιοχές είναι τα Μικτά Μοντέλα. Ωστόσο, τέτοια μοντέλα απαιτούν

αποσαφήνιση του τυχαίου μέρους τους κάνοντας δυσνόητη την συμπερασματολογία, εξαρτώνται από αυστηρές παραμετρικές υποθέσεις και υποθέσεις γύρω από την κατανομή και δεν προσφέρουν ανθεκτικές εκτιμήσεις.

Μία νέα προσέγγιση στην εκτίμηση σε μικρές περιοχές αντιμετωπίζει αποτελεσματικά τα παραπάνω προβλήματα. Βασίζεται στην Παλινδρόμηση Ποσοστιαίων Σημείων και προτάθηκε το 2006 από τους Ray Chambers και Niko Tzavidi στην εργασία τους με τίτλο «M-Quantile models for small area estimation» που δημοσίευσε το επιστημονικό περιοδικό Biometrika. Στη μέθοδο αυτή αντί του Μικτού Μοντέλου, χρησιμοποιείται το μοντέλο της Μ-Παλινδρόμησης Ποσοστιαίων Σημείων για την εκτίμηση σε μικρές περιοχές, που αποφεύγει τις τυχαίες επιδράσεις. Τα νέα αυτά μοντέλα έχουν μία σειρά από πλεονεκτήματα σε σχέση με τα Μικτά Μοντέλα, όπως μη παραμετρική αποσαφήνιση, ανθεκτικές εκτιμήσεις και δυνατότητα εκτιμήσεων και άλλων χαρακτηριστικών των μικρών περιοχών όπως διάμεσοι και λοιπά ποσοστιαία σημεία.

Παρακάτω, αφού αναφερθούμε στα Μικτά Μοντέλα και στους ανθεκτικούς εκτιμητές, θα παρουσιάσουμε τη μέθοδο της Μ-Παλινδρόμησης Ποσοστιαίων Σημείων για την εκτίμηση σε μικρές περιοχές και θα τη συγκρίνουμε με την παραδοσιακή τεχνική των Μικτών Μοντέλων, που μέχρι σήμερα χρησιμοποιείται για την εκτίμηση σε τέτοιες περιοχές. Τέλος γίνεται και μία αναφορά σε μία προσέγγιση του Roger Koenker (2005) όσον αφορά το ζήτημα της σύνδεσης μεταξύ της Παλινδρόμησης Ποσοστιαίων Σημείων με τα Μικτά Μοντέλα με τη χρήση των L-στατιστικών.

3.2 Γραμμικά Μικτά Μοντέλα

Έστω ότι έχουμε μία μεταβλητή y που μας ενδιαφέρει και ένα διάνυσμα p επεξηγηματικών μεταβλητών x_{ij}, με γνωστές τιμές για κάθε πληθυσμιακή μονάδα i σε κάθε μικρή περιοχή j. Σκοπός μας είναι να χρησιμοποιήσουμε τα δεδομένα αυτά για να εκτιμήσουμε ποσότητες, όπως για παράδειγμα η μέση τιμή m_j της y, για τις διάφορες μικρές περιοχές. Καθεμία από τις περιοχές αυτές έχει μικρό μέγεθος δείγματος, οπότε

πρέπει να καταφύγουμε σε εναλλακτικές μεθόδους εκτίμησης που «δανείζονται πληροφορία» από όλες τις μικρές περιοχές, ώστε να παράγουν μία εκτίμηση για μία συγκεκριμένη μικρή περιοχή. Η πιο γνωστή από αυτές τις μεθόδους είναι τα Γραμμικά Μικτά Μοντέλα (ΓΜΜ). Στη γενική περίπτωση ένα ΓΜΜ έχει την εξής μορφή:

$$y_{ij} = x_{ij}^T b + z_{ij}^T g_j + u_{ij}, \qquad (i=1,\ldots,n\ , j=1,\ldots d) \qquad (3.2.1)$$

ή αλλιώς $$E(y_{ij} \mid g_j) = x_{ij}^T b + z_{ij}^T g_j, \qquad (3.2.2)$$

όπου ο όρος $x_{ij}^T b$ αντιπροσωπεύει τις σταθερές επιδράσεις και ο όρος $z_{ij}^T g_j$ αποτελεί το τυχαίο κομμάτι του μοντέλου. Το g_j δηλώνει ένα διάνυσμα τυχαίων επιδράσεων και το z_{ij} δηλώνει ένα διάνυσμα επεξηγηματικών μεταβλητών των οποίων οι τιμές είναι γνωστές για όλες τις μονάδες του πληθυσμού.

Όμως το μοντέλο (3.2.2) για να έχει ισχύ πρέπει να ικανοποιούνται και οι παρακάτω συνθήκες:

► Οι x_{ij} και z_{ij} είναι μη στοχαστικές μεταβλητές.

► Δοθέντων των g_j, οι y_{ij} είναι ανεξάρτητες τυχαίες μεταβλητές.

► Δοθέντων των g_j, οι y_{ij} ακολουθούν την κανονική κατανομή.

► $Eg_j = 0$ και $Var\left(g_j\right) = D$, όπου D ένας θετικά ορισμένος πίνακας.

► Τα g_j ακολουθούν την κανονική κατανομή.

Κάτω από τις παραπάνω συνθήκες, μπορούμε να προχωρήσουμε στην εκτίμηση των παραμέτρων του ΓΜΜ χρησιμοποιώντας τη μέθοδο τής μέγιστης πιθανοφάνειας. Αν θεωρήσουμε ότι το διάνυσμα των επεξηγηματικών μεταβλητών είναι γνωστό για κάθε μονάδα του πληθυσμού *i* στη μικρή περιοχή *j* και οι τιμές της *y* είναι διαθέσιμές για τις δειγματικές μονάδες, τότε σύμφωνα με την προσέγγιση του Henderson (1953, ***Method 3***) οι μέσοι των διαφόρων περιοχών εκτιμώνται ως εξής:

$$\hat{m}_j = N_j^{-1}\left\{\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j}\left(x_{ij}^T\hat{b} + z_{ij}^T\hat{g}_j\right)\right\}, \qquad (3.2.3)$$

όπου s_j είναι το σύνολο των n_j δειγματικών μονάδων της περιοχής και r_j το σύνολο των υπόλοιπων $N_j - n_j$ μονάδων.

Ο παραπάνω εκτιμητής μπορεί να χωριστεί σε δύο ειδικές περιπτώσεις. Στην πρώτη περίπτωση τα z_{ij} αποτελούνται από δείκτες για τις διάφορες μικρές περιοχές και ο εκτιμητής (3.2.3) ονομάζεται εκτιμητής «τυχαίων σταθερών όρων». Στη δεύτερη περίπτωση έχουμε το πιο σύνηθες φαινόμενο, όπου τα z_{ij} εκτός από δείκτες για τις διάφορες μικρές περιοχές αποτελούνται και από άλλες επεξηγηματικές μεταβλητές. Στην περίπτωση αυτή ο εκτιμητής (3.2.3) ονομάζεται εκτιμητής «τυχαίων κλίσεων».

Γενικά, ο ρόλος των τυχαίων επιδράσεων στα ΓΜΜ είναι να «εξηγούν» τις διαφορές μεταξύ των μικρών περιοχών στη δεσμευμένη κατανομή της *y* δοθέντων των *x*. Παρόλαυτα, όπως θα δούμε παρακάτω, δεν αποτελούν το μόνο τρόπο για κάτι τέτοιο.

3.3 Ανθεκτικοί εκτιμητές

Η θεωρία των ανθεκτικών εκτιμητών αναπτύχθηκε τη δεκαετία του 70 σε μία προσπάθεια μείωσης των επιδράσεων των έκτροπων παρατηρήσεων που συχνά υπάρχουν σε διάφορα σετ δεδομένων. Οι πιο γνωστοί από αυτούς τους εκτιμητές είναι οι L, R και M. Εκτενέστερα θα ασχοληθούμε με τους τελευταίους, καθώς χρησιμοποιούνται στην μέθοδο της M-Παλινδρόμησης Ποσοστιαίων Σημείων που θα δούμε παρακάτω, αλλά θα περιγράψουμε συνοπτικά και τους L-εκτιμητές.

3.3.1 Συνάρτηση επίδρασης

Πριν δώσουμε τον ορισμό των M-εκτιμητών, θα πρέπει να αναφερθούμε συνοπτικά στη συνάρτηση επίδρασης (influence function) ως μέτρο ανθεκτικότητας ενός

εκτιμητή. Η συνάρτηση επίδρασης πρωτοεισήχθη από τον Hampel το 1968 και περιγράφει την επίδραση που έχει στον εκτιμητή μία υποτυπώδης αλλοίωση Δx στο σημείο x, μετρώντας την ασυμπτωτική αμεροληψία που προκαλείται. Παρακάτω έχουμε τη μαθηματική έκφραση της συνάρτησης επίδρασης ενός εκτιμητή T της κατανομής F:

$$IF(x;T,F)=\lim_{t\to 0}\frac{T\left((1-t)F+t\Delta x\right)-T(F)}{t}, \qquad (3.3.1)$$

όπου *x* τα σημεία του δειγματικού χώρου, όπου το παραπάνω όριο υπάρχει.

Αποδεικνύεται ότι η ασυμπτωτική διασπορά της παραπάνω συνάρτησης επίδρασης είναι:

$$V(T,F)=\int IF^2(x;T,F)\,dF(x). \qquad (3.3.2)$$

3.3.2 Μ-εκτιμητές

Οι Μ-εκτιμητές θεμελιώθηκαν από τον Huber το 1973 και θεωρούνται από τις πιο απλές προσεγγίσεις τόσο θεωρητικά όσο και υπολογιστικά. Χρησιμοποιούνται ευρέως στην ανάλυση δεδομένων με έκτροπες παρατηρήσεις κυρίως της εξαρτημένης μεταβλητής *y*.

Όπως είναι γνωστό ο εκτιμητής μεγίστης πιθανοφάνειας (ΕΜΠ) ορίζεται ως η τιμή $T_n=T_n(X_1,...,X_n)$ που μεγιστοποιεί την ποσότητα $\prod_{i=1}^{n} f_{T_n}(X_i)$. Ισοδύναμα ο ΕΜΠ παράγεται από το παρακάτω πρόβλημα ελαχιστοποίησης:

$$\min_{T_n}\sum_{i=1}^{n}\left[-\log f_{T_n}(X_i)\right]. \qquad (3.3.3)$$

Ο Huber το 1964 πρότεινε τη γενίκευση του παραπάνω προβλήματος σε ένα άλλο πρόβλημα ελαχιστοποίησης, όπως παρακάτω:

$$\min_{T_n} \sum_{i=1}^{n} r\left(X_i, T_n\right). \qquad (3.3.4)$$

Η ρ είναι μία συνάρτηση του $\Omega \times \Theta$, όπου Ω είναι ο δειγματικός χώρος και Θ ο παραμετρικός χώρος. Έστω ότι η ρ έχει παράγωγο $y\left(x,q\right) = \frac{\partial}{\partial q} r\left(x,q\right)$. Τότε το πρόβλημα ελαχιστοποίησης (3.3.4) μπορεί να πάρει τη μορφή της παρακάτω εξίσωσης:

$$\sum_{i=1}^{n} y\left(X_i, T_n\right) = 0. \qquad (3.3.5)$$

Κάθε εκτιμητής που ορίζεται από το πρόβλημα ελαχιστοποίησης (3.3.4) ή από την εξίσωση (3.3.5) ονομάζεται *Μ-εκτιμητής*. Η συνάρτηση επίδρασης ενός Μ-εκτιμητή δίνεται από τη σχέση:

$$IF\left(x;T,F\right) = \frac{y\left[x, T\left(F\right)\right]}{-\int \frac{\partial}{\partial q} y\left(y,q\right)\big|_{q=T(F)}\, dF\left(y\right)} \qquad (3.3.6)$$

και έχει ασυμπτωτική διασπορά:

$$V\left(T,F\right) = \frac{\int y^{2}\left[x, T\left(F\right)\right] dF\left(x\right)}{\left[\int \frac{\partial}{\partial q} y\left(y,q\right)\big|_{q=T(F)}\, dF\left(y\right)\right]^{2}}. \qquad (3.3.7)$$

Έχουν κατά καιρούς προταθεί διάφοροι Μ-εκτιμητές, από τους οποίους βασικότεροι είναι οι παρακάτω:

► *Huber minimax*

$$y\left(t\right) = \begin{cases} t & , \alpha\nu\ |t| < b \\ b\,\mathrm{sgn}\left(t\right) & , \alpha\nu\ |t| \geq b \end{cases}, \qquad (3.3.8)$$

όπου b μία σταθερά.

► *Descending minimax*

$$y\ (t)=\begin{cases} t & ,\ \text{αν}\ |t|<a \\ b\,\mathrm{sgn}(t)\tanh\left[\dfrac{1}{2}b(c-|t|)\right] & ,\ \text{αν}\ a\le|t|<c \\ 0 & ,\ \text{αλλού} \end{cases}, \tag{3.3.9}$$

όπου α, b και c σταθερές.

► *Hampel*

$$y\ (t)=\begin{cases} t & ,\ \text{αν}\ |t|<a \\ a\,\mathrm{sgn}(t) & ,\ \text{αν}\ a\le|t|<b \\ \left\{(c-|t|)/(c-b)\right\}a\,\mathrm{sgn}(t) & ,\ \text{αν}\ b\le|t|<c \\ 0 & ,\ \text{αλλού} \end{cases}, \tag{3.3.10}$$

όπου α, b και c σταθερές.

► *Andrew*

$$y\ (t)=\begin{cases} \sin(t) & ,\ \text{αν}\ -p\le t<p \\ 0 & ,\ \text{αλλού} \end{cases}. \tag{3.3.11}$$

► *Tukey*

$$y\ (t)=\begin{cases} t\left(1-(t/c)^2\right)^2 & ,\ \text{αν}\ t\le c \\ 0 & ,\ \text{αλλού} \end{cases}, \tag{3.3.12}$$

όπου c μία σταθερά.

3.3.3 L-εκτιμητές

Οι L-εκτιμητές είναι γραμμικοί συνδυασμοί του διατεταγμένου δείγματος και εφαρμόζονται ευρέως στην εκτίμηση παραμέτρων. Οι πιο βασικοί L-εκτιμητές που έχουν προταθεί είναι οι παρακάτω:

► *Διάμεσος δείγματος*

$$Q(1/2) = \inf\{x : F(X) \geq 1/2\}. \qquad (3.3.13)$$

► *Μέση τιμή μετά από ποσοστιαία αποκοπή του Tukey*

$$\hat{a} = 0.25 \cdot Q(1/4) + 0.5 \cdot Q(1/2) + 0.25 \cdot Q(1/4), \qquad (3.3.14)$$

όπου $Q(t) = \inf\{x : F(X) \geq t\}$.

3.4 Μ-Παλινδρόμηση Ποσοστιαίων Σημείων

Στο πρώτο κεφάλαιο αναλύσαμε όλα τα βασικά στοιχεία της μεθόδου της ΠΠΣ, όπως θεμελιώθηκε από τους Koenker και Basset και συμπληρώθηκε από άλλους σπουδαίους στατιστικούς. Το 1988 οι Breckling και Chambers στην εργασία τους «M-Quantiles» που δημοσιεύτηκε στο περιοδικό Biometrika παρουσίασαν τη μέθοδο της Μ-Παλινδρόμησης Ποσοστιαίων Σημείων (Μ-ΠΠΣ). Στη μέθοδο αυτή η ΠΠΣ συνδέεται με την ανθεκτική παλινδρόμηση που βασίζεται στους Μ-εκτιμητές και στις συναρτήσεις επίδρασης.

Το Μ-ποσοστιαίο σημείο τάξης τ της δεσμευμένης πυκνότητας της y δοθέντων των x, ορίζεται ως η λύση $Q_t\left(y \mid x; y\right)$ της παρακάτω εξίσωσης:

$$\int y_{\ t}\left(y-Q\right) f\left(y \mid x\right) dy = 0, \qquad (3.4.1)$$

όπου ψ η συνάρτηση επίδρασης του αντίστοιχου Μ-ποσοστιαίου σημείου.

Ένα γραμμικό μοντέλο Μ-ΠΠΣ είναι αυτό για το οποίο το δεσμευμένο Μ-ποσοστιαίο σημείο τάξης τ είναι:

$$Q_t\left(y \mid x; y\right) = x^T b_y\left(t\right). \qquad (3.4.2)$$

Για καθορισμένο ποσοστιαίο σημείο τ και συγκεκριμένη συνάρτηση επίδρασης ψ, οι εκτιμητές των παραμέτρων της Μ-ΠΠΣ μπορούν να βρεθούν αν λύσουμε τις παρακάτω εξισώσεις:

$$\sum_{i=1}^{n} y_{\ t}\left(r_{ity}\right) x_i = 0, \qquad (3.4.3)$$

όπου $r_{ity} = y_i - x_i^T b_y\left(t\right)$,

$$y_{\ t}\left(r_{ity}\right) = 2y\left(s^{-1} r_{ity}\right) \cdot \left\{ qI\left(r_{ity} > 0\right) + \left(1-q\right) I\left(r_{ity} \le 0\right)\right\}$$

και s ένας κατάλληλος ανθεκτικός εκτιμητής κλίσης, όπως για παράδειγμα η μέση απόλυτη απόκλιση.

Εύκολα διαπιστώνει κανείς ότι κάθε μία από τις συναρτήσεις επίδρασης που αναφέραμε στην παράγραφο 3.3 μπορούν να χρησιμοποιηθούν στην Μ-ΠΠΣ. Εδώ χρησιμοποιούμε τη συνάρτηση επίδρασης (3.3.8) του Huber. Η λύση των (3.4.3) γίνεται μέσω ενός κατάλληλου αλγορίθμου.

Συνήθως χρησιμοποιείται η Μ-ΠΠΣ αντί της ΠΠΣ κυρίως για πρακτικούς λόγους. Οι αλγόριθμοι που χρησιμοποιούνται στην ΠΠΣ, τους οποίους αναφέραμε στην παράγραφο 1.4, δε συγκλίνουν απαραίτητα σε μία μοναδική λύση. Αντίθετα, ο Kokic το 1997 απέδειξε ότι οι αλγόριθμοι της Μ-ΠΠΣ, που βασίζονται σε σταθμισμένα ελάχιστα

τετράγωνα, συγκλίνουν σε μία μοναδική λύση, αρκεί να χρησιμοποιούμε μία συνεχή, μονότονη συνάρτηση επίδρασης.

3.5 Εφαρμογή της Μ-ΠΠΣ σε μικρές περιοχές

Τα ΓΜΜ θεωρούν ότι η μεταβλητότητα της δεσμευμένης κατανομής της *y* δοθέντων των *x* εξηγείται ως ένα βαθμό από μία ιεραρχική δομή, που ορίζεται από τις μικρές περιοχές. Παρόλαυτα είδαμε στην προηγούμενη παράγραφο ότι η μεταβλητότητα αυτή μπορεί να μοντελοποιηθεί και μέσω της Μ-ΠΠΣ, η οποία δεν εξαρτάται από κάποιας μορφής ιεραρχική δομή.

Και εδώ θα θεωρήσουμε ότι το διάνυσμα των επεξηγηματικών μεταβλητών είναι γνωστό για κάθε μονάδα του πληθυσμού *i* στη μικρή περιοχή *j* και οι τιμές της *y* είναι διαθέσιμές για τις δειγματικές μονάδες. Κατατάσσουμε τις πληθυσμιακές μονάδες μόνο με βάση το *i=1,…,n* και θεωρούμε τους συντελεστές των Μ-ποσοστιαίων σημείων, που είναι οι τιμές t_i, για τις οποίες $Q_{t_i}(x_i, y) = y_i$. Έστω q_j ο μέσος όρος των t_i της περιοχής j. Θεωρούμε το γραμμικό μοντέλο (3.4.2) της Μ-ΠΠΣ, για το οποίο υπάρχει η παρακάτω προσέγγιση πρώτης τάξης μέσω του αναπτύγματος Taylor:

$$\hat{m}_j = N_j^{-1}\left\{\sum_{i\in s_j} y_i + \sum_{i\in r_j} x_i^T b_y\left(t_i\right)\right\}$$

$$\approx N_j^{-1}\left\{\sum_{i\in s_j} y_i + \sum_{i\in r_j} x_i^T b_y\left(q_j\right)\right\} + N_j^{-1}\sum_{i\in r_j} x_i^T\left\{\frac{\partial b_y\left(q_j\right)}{\partial q_j}\right\}\left(t_i - q_j\right),$$

όπου s_j είναι το σύνολο των n_j δειγματικών μονάδων της περιοχής, r_j το σύνολο των υπόλοιπων $N_j - n_j$ μονάδων και b_y ο συντελεστής της Μ-ΠΠΣ που βασίζεται στη συνάρτηση επίδρασης του Huber. Στην παραπάνω προσέγγιση επικρατεί ο πρώτος όρος,

και με αυτόν τον τρόπο προκύπτει ο παρακάτω εκτιμητής $\hat{m}_j$ του μέσου m_j της περιοχής j:

$$\hat{m}_j = N_j^{-1}\left\{\sum_{i\in s_j} y_i + \sum_{i\in r_j} x_i^T \hat{b}_y\left(\hat{q}_j\right)\right\}. \qquad (3.5.1)$$

Η παραπάνω έκφραση (3.5.1) παριστάνει τις εκτιμήσεις για τις μικρές περιοχές που βασίζονται στην Μ-ΠΠΣ.

Μπορούν να δοθούν διάφορες εκδοχές στο q_j και κατά συνέπεια να προκύψουν διάφοροι εκτιμητές $\hat{q}_j$, όπως η μέση τιμή των t_i της περιοχής j ή η διάμεσός τους, αν θέλουμε έναν πιο ανθεκτικό ορισμό. Εκτός όμως από την προσέγγιση αυτή, για τον εκτιμητή $\hat{m}_j$ υπάρχει και μία άλλη προσέγγιση. Θα μπορούσαμε αντί να πάρουμε το μέσο όρο των t_i και στη συνέχεια να προσαρμόσουμε την Μ-ΠΠΣ, να προσαρμόσουμε πρώτα για όλα τα i τους συντελεστές $b_y(t_i)$ της Μ-ΠΠΣ που αντιστοιχούν στα t_i, στη συνέχεια να πάρουμε τον μέσο όρο τους και να χρησιμοποιήσουμε αυτή την ποσότητα στην έκφραση (3.5.1). Έχει παρατηρηθεί ότι οι δύο προσεγγίσεις δίνουν παρόμοια αποτελέσματα όταν οι συντελεστές των Μ-ποσοστιαίων σημείων, t_i μοιάζουν πιο πολύ μέσα στις περιοχές απ'ότι μεταξύ των περιοχών.

Ας θεωρήσουμε και πάλι την αρχική μας προσέγγιση που καταλήγει στον εκτιμητή (3.5.1). Η προσέγγιση αυτή μας οδηγεί σε μια σύνδεση με το ΓΜΜ. Η τιμή q_j για κάθε περιοχή μπορεί να ερμηνευθεί σαν μια ψευδοτυχαία επίδραση για την περιοχή j. Για παράδειγμα, αν το q_j είναι ίσο με 0.5 για όλα τα j, συμπεραίνουμε ότι δεν υπάρχει μεταβλητότητα μεταξύ των περιοχών, πέρα από αυτή που εξηγείται από τις επεξηγηματικές μεταβλητές του μοντέλου. Δηλαδή ο παράγοντας «περιοχή» δεν επιδρά στη μεταβλητή y που μας ενδιαφέρει.

Με όποιο τρόπο και να ορίσουμε το q_j για την περιοχή j, η έκφραση (3.5.1) είναι ισοδύναμη με την πρόβλεψη μιας μη παρατηρηθείσας τιμής y_i για μια εκτός

δείγματος μονάδα i της περιοχής j, χρησιμοποιώντας το $x_i^T b_y\left(q_j\right)$. Επομένως μπορούμε να υπολογίσουμε τις προβλεπόμενες τιμές και για άλλα χαρακτηριστικά των μικρών περιοχών χρησιμοποιώντας τις παραπάνω προβλέψεις. Για παράδειγμα ένας εκτιμητής για τη συνάρτηση κατανομής του πληθυσμού που ορίζεται από τις τιμές του y στην περιοχή j, σύμφωνα με την παραπάνω προσέγγιση είναι ο παρακάτω:

$$\hat{F}_j(t) = N_j^{-1}\left[\sum_{i\in s_j} I(y_i \le t) + \sum_{i\in r_j} I\left\{x_i^T \hat{b}_y\left(\hat{q}_j\right) \le t\right\}\right]. \qquad (3.5.2)$$

Από την παραπάνω έκφραση εύκολα προκύπτουν και οι εκτιμητές για τα διάφορα ποσοστιαία σημεία της y στην περιοχή j. Συγκεκριμένα,. η εκτιμώμενη διάμεσος των τιμών της y στην περιοχή j είναι η διάμεσος του συνόλου $\left\{y_i; i\in s_j\right\}\cup\left\{x_i^T \hat{b}_y\left(q_j\right); i\in r_j\right\}$ και γενικότερα το τ-ποσοστιαίο σημείο των τιμών της y στην περιοχή j είναι το τ-ποσοστιαίο σημείο του ίδιου συνόλου.

3.6 Σύγκριση της Μ-ΠΠΣ με τα ΓΜΜ για εκτίμηση σε μικρές περιοχές

Όπως αναφέραμε στην παράγραφο 3.5 υπάρχει μία σύνδεση μεταξύ του μοντέλου της Μ-ΠΠΣ και του ΓΜΜ. Αυτή η σύνδεση γίνεται πιο ξεκάθαρη αν γράψουμε το μοντέλο (3.4.2) της Μ-ΠΠΣ με την παρακάτω μορφή:

$$y_{ij} = x_i^T b_y\left(0,5\right) + x_i^T\left\{b_y\left(q_j\right) - b_y\left(0,5\right)\right\}. \qquad (3.6.1)$$

Το ΓΜΜ αποτελείται από έναν όρο σταθερών επιδράσεων, που είναι ίδιο για κάθε μικρή περιοχή, και από έναν όρο τυχαίων επιδράσεων, που είναι διαφορετικό για κάθε μικρή περιοχή και τη χαρακτηρίζει. Στην παραπάνω μορφή (3.6.1) βλέπουμε ότι το μοντέλο της Μ-ΠΠΣ εκφράζει αυτή τη μεταβλητότητα μεταξύ των περιοχών μέσω της

απόκλισης του συντελεστή $\hat{b}_y\left(\hat{q}_j\right)$ από τη διάμεσο τιμή $\hat{b}_y(0,5)$. Ο όρος $\left\{b_y\left(q_j\right)-b_y(0,5)\right\}$ παίζει το ρόλο της ψευδοτυχαίας επίδρασης για την περιοχή j.

Η χρησιμοποίηση των μοντέλων της Μ-ΠΠΣ για την εκτίμηση σε μικρές περιοχές έχει μία σειρά από πλεονεκτήματα σε σχέση με τα ΓΜΜ. Παρακάτω απαριθμούμε τα πλεονεκτήματα αυτά:

► Τα μοντέλα Μ-ΠΠΣ δεν υπακούουν σε αυστηρές παραμετρικές και γύρω από την κατανομή υποθέσεις σε αντίθεση με τα ΓΜΜ που τις προϋποθέτουν λόγω της παρουσίας του τυχαίου μέρους σε αυτά.

► Ένα σημαντικό σημείο στα ΓΜΜ είναι η αποσαφήνιση του τυχαίου μέρους του μοντέλου, όσον αφορά τη χρήση των διαγνωστικών τεστ όπως το τεστ λόγου πιθανοφανειών. Επίσης η χρήση των μοντέλων τυχαίων κλίσεων είναι δύσκολη όταν έχουμε μεγάλο αριθμό επεξηγηματικών μεταβλητών, σε αντίθεση με τα μοντέλα Μ-ΠΠΣ που το μόνο που απαιτούν είναι εφαρμογή των μεθόδων μοντελοποίησης για να ενσωματώσουν τις x.

► Τα μοντέλα Μ-ΠΠΣ επιτρέπουν την ανθεκτική συμπερασματολογία γύρω από τις έκτροπες παρατηρήσεις που συχνά παρατηρούνται. Μάλιστα έχει αναπτυχθεί κατάλληλο λογισμικό γι' αυτό το σκοπό, σε αντίθεση με τα ΓΜΜ που στον τομέα αυτόν υστερούν.

► Τέλος στα μοντέλα Μ-ΠΠΣ έχει αποδειχτεί ότι μπορούμε να ενσωματώσουμε μια μη παραμετρική σχέση μεταξύ y και x. Αντίθετα δεν έχει βρεθεί μια μη παραμετρική προσέγγιση για τα ΓΜΜ.

Εκτός όμως από τα παραπάνω πλεονεκτήματα, η εκτίμηση σε μικρές περιοχές μέσω της Μ-ΠΠΣ έχει και κάποια μειονεκτήματα τα οποία είναι τα εξής.

► Τα μοντέλα Μ-ΠΠΣ απαιτούν δεδομένα, όπου τα Μ-ποσοστιαία σημεία για τη δεσμευμένη κατανομή της *y* δοθέντων των *x* είναι καλά ορισμένα, κάτι που δεν γίνεται όταν για παράδειγμα η *y* είναι ονομαστική μεταβλητή. Αντιθέτως μέσω της θεωρίας που έχει αναπτυχθεί γύρω από τα γενικευμένα ΓΜΜ δεν αντιμετωπίζουμε παρόμοιο πρόβλημα αν χρησιμοποιήσουμε τη συγκεκριμένη μέθοδο.

► Όταν η μεταβλητή απόκρισης *y* είναι πολυδιάστατη, δε μπορούμε να ορίσουμε τα πολυδιάστατα ποσοστιαία σημεία, αφού μέχρι σήμερα οι στατιστικοί δεν έχουν συμφωνήσει σε έναν κοινά αποδεκτό ορισμό. Ως εκ τούτου η εφαρμογή της Μ-ΠΠΣ στην περίπτωση αυτή είναι αδύνατη.

Γενικότερα δε μπορούμε να πούμε ότι η Μ-ΠΠΣ είναι πιο αξιόπιστη από τα ΓΜΜ, όταν οι υποθέσεις που τα τελευταία απαιτούν ικανοποιούνται. Η χρήση, λοιπόν, της Μ-ΠΠΣ ενδείκνυται κυρίως σε περιπτώσεις που επιθυμούμε πιο ανθεκτικές εκτιμήσεις και κατά προτίμηση σε περιπτώσεις που δεν ικανοποιούνται οι υποθέσεις για τη χρήση των ΓΜΜ. Στην συνέχεια θα δούμε μία εφαρμογή που θα μας βοηθήσει να βγάλουμε χρήσιμα συμπεράσματα στην προσπάθειά μας για σύγκριση των δύο μεθόδων.

3.7 Μελέτη προσομοίωσης για τη σύγκριση Μ-ΠΠΣ και ΓΜΜ

Στη μελέτη αυτή των Chambers και Tzavidis (2006) χρησιμοποιήθηκε ένα δείγμα από 1652 φάρμες που βρίσκονται σε 29 περιοχές της Αυστραλίας (μεταβλητή ***Region***). Η εξαρτημένη μεταβλητή ήταν η *TCC* που εκφράζει το συνολικό κόστος για κάθε φάρμα. Άλλα διαθέσιμα στοιχεία ήταν το μέγεθος δείγματος κάθε φάρμας, η συνολική της έκταση σε εκτάρια και η κλιματική ζώνη στην οποία βρίσκεται. Με βάση τα παραπάνω στοιχεία οι φάρμες κατατάχθηκαν σε 6 ομάδες που αποτελούν τις αντίστοιχες κατηγορίες της μεταβλητής ***SizeZone***:

►Κλιματική Ζώνη 1 και έκταση μικρότερη από 50000 εκτάρια.
►Κλιματική Ζώνη 1 και έκταση μεγαλύτερη από 50000 εκτάρια.
►Κλιματική Ζώνη 2 και έκταση μικρότερη από 1500 εκτάρια.
►Κλιματική Ζώνη 2 και έκταση μεγαλύτερη από 1500 εκτάρια.
►Κλιματική Ζώνη 3 και έκταση μικρότερη από 750 εκτάρια.
►Κλιματική Ζώνη 3 και έκταση μεγαλύτερη από 750 εκτάρια.

Επίσης ως επεξηγηματική μεταβλητή χρησιμοποιήθηκε και η *FarmArea* που εκφράζει τη συνολική έκταση σε εκτάρια, ενώ οι μικρές περιοχές καθορίστηκαν βάσει της μεταβλητής Region. Οι τιμές για τις παραπάνω μεταβλητές (FarmArea, SizeZone, Region) θεωρήθηκαν γνωστές για όλες τις μονάδες του πληθυσμού.

Σκοπός της μελέτης αυτής ήταν να αξιολογηθούν οι εκτιμήσεις των μέσων και των διαμέσων της TCC για κάθε περιοχή βάσει των ΓΜΜ και των μοντέλων Μ-ΠΠΣ και να συγκριθούν οι δύο μέθοδοι. Προσαρμόστηκαν τα παρακάτω 4 γραμμικά μοντέλα:

1. Μοντέλο τυχαίων σταθερών όρων (*random intercept*).
2. Μοντέλο τυχαίων κλίσεων (*random slope*).
3. Μοντέλο Μ-ΠΠΣ με b=1.345 στη συνάρτηση (3.3.8) του Huber (*M-quantile*).
4. Μοντέλο Μ-ΠΠΣ με b= 2×10^5 στη συνάρτηση (3.3.8) του Huber (*Expectile*).

Και τα 4 μοντέλα περιέχουν τις κύριες επιδράσεις των μεταβλητών FarmArea και SizeZone, καθώς και την αλληλεπίδρασή τους, ενώ για το μοντέλο τυχαίων κλίσεων έχουμε τυχαία κλίση για τη μεταβλητή FarmArea.

Οι εκτιμήσεις των μέσων υπολογίστηκαν μέσω των σχέσεων (3.2.3) για τα ΓΜΜ και (3.5.1) για τα Μ-ΠΠΣ. Όσον αφορά τις εκτιμήσεις των διαμέσων, αυτές υπολογίστηκαν για τα ΓΜΜ μέσω της σχέσης $med\left[\{y_i, i\in s_j\}\cup\{x_i^T\hat{b}+z_i^T\hat{g}_j, i\in r_j\}\right]$ και για τα μοντέλα Μ-ΠΠΣ μέσω της σχέσης $med\left[\{y_i, i\in s_j\}\cup\{x_i^T\hat{b}_y(\hat{q}_j), i\in r_j\}\right]$.

Τα μέτρα σύγκρισης της αποτελεσματικότητας των παραπάνω μεθόδων ήταν οι σχετικές μεροληψίες (relative bias) και οι σχετικές ρίζες των μέσων τετραγωνικών σφαλμάτων (relative RMSE)

Η μελέτη ακολούθησε δύο βήματα. Αρχικά δημιουργήθηκε ένας πληθυσμός N=81982 φαρμών μέσω δειγματοληψίας με αντικατάσταση N φορές από το αρχικό δείγμα των 1652 φαρμών και με πιθανότητα ανάλογη με το μέγεθος δείγματος της κάθε φάρμας. Στη συνέχεια από τον προσομοιωμένο αυτό πληθυσμό επιλέχθηκαν 500 ανεξάρτητα τυχαία δείγματα ίδιου μεγέθους με το αρχικό (1652). Τα μεγέθη δείγματος των 29 περιοχών και αυτά ρυθμίστηκαν να είναι ίδια με το αρχικό δείγμα (από 6 ως 117).

Στον παρακάτω πίνακα βλέπουμε τις σχετικές μεροληψίες και τις σχετικές ρίζες των μέσων τετραγωνικών σφαλμάτων για τις 4 μεθόδους όσον αφορά τις εκτιμήσεις των μέσων για κάθε περιοχή:

Region	Random Intercept	Random Slope	*M*-quantile	Expectile	Random Intercept	Random Slope	*M*-quantile	Expectile
	Relative Bias				Relative Root Mean Squared Error			
1	3·57	-6·99	-14·12	-0·14	19·63	31·41	24·32	24·36
2	12·12	15·50	-22·20	-12·16	22·13	41·71	31·89	25·64
3	-4·77	0·64	-29·13	-14·65	21·52	28·72	30·88	23·48
4	30·66	10·21	6·81	15·91	36·66	20·71	15·51	20·98
5	9·15	5·70	-12·90	-0·20	17·43	15·03	14·81	10·04
6	50·54	22·81	-21·31	0·73	98·51	57·86	37·38	40·10
7	1·94	-2·69	-10·71	2·95	17·84	26·41	13·05	16·10
8	-10·51	-7·72	-16·61	-15·85	14·21	12·41	17·67	16·86
9	-7·10	4·93	-22·90	-21·38	14·43	16·43	23·92	24·09
10	-4·59	-2·45	-14·20	-10·95	10·37	10·04	15·88	13·72
11	15·36	14·36	2·84	12·30	21·21	21·78	7·02	13·74
12	-7·87	-8·72	-31·51	-24·87	15·24	14·40	31·90	25·39
13	7·08	8·19	-12·27	-5·31	15·51	16·16	13·19	7·60
14	1·08	2·45	-10·06	-3·77	11·45	19·56	13·74	11·43
15	2·10	0·09	-14·56	-7·18	18·07	17·18	15·44	9·33
16	-6·21	-7·02	-30·97	-25·09	11·27	11·93	31·23	25·43
17	4·83	5·40	-14·89	-8·70	16·24	16·15	15·56	10·51
18	25·62	32·44	10·87	18·93	34·91	43·88	14·21	21·58
19	-2·55	-2·54	-21·29	-14·96	7·72	8·27	21·78	15·74
20	12·35	12·77	-4·43	12·35	20·55	21·99	12·65	18·27
21	-10·49	-8·67	-26·39	-20·14	29·85	28·47	26·74	21·35
22	4·55	1·00	-15·87	-3·06	12·20	9·59	17·18	12·12
23	6·26	6·34	-8·20	-1·59	17·54	17·68	9·11	4·75
24	-7·78	-5·44	-31·29	-25·13	17·05	18·49	31·50	25·43
25	-6·33	-3·01	-29·46	-24·31	11·13	9·81	29·62	24·54
26	2·44	3·53	-15·17	-9·10	11·65	11·79	15·58	9·83
27	-2·72	-4·25	-23·76	-18·50	8·98	11·28	24·07	19·05
28	-0·07	0·65	-16·18	-9·12	7·68	7·42	16·81	10·32
29	-1·60	-2·20	-19·05	-13·48	7·46	7·02	19·41	14·04
Mean	4·04	2·94	-16·17	-7·81	19·60	19·78	20·41	17·79
Median	1·94	0·65	-15·87	-9·10	16·24	16·43	17·18	16·86

ΠΙΝΑΚΑΣ 2-9: Πίνακας με τα μέτρα σύγκρισης των εκτιμήσεων των μέσων μέσω ΓΜΜ και Μ-ΠΠΣ. (Πηγή: Chambers, Tzavidis,2006)

Στον παραπάνω πίνακα παρατηρούμε ότι οι εκτιμήσεις μέσω των ΓΜΜ έχουν μικρότερες σχετικές μεροληψίες. Ωστόσο αυτές οι μικρότερες μεροληψίες συνοδεύονται και από μεγαλύτερη μεταβλητότητα των εκτιμητών. Συνεπώς η εικόνα θα γίνει καθαρότερη αν συγκρίνουμε τις σχετικές ρίζες των μέσων τετραγωνικών σφαλμάτων που λαμβάνουν υπόψη και τη μεταβλητότητα των εκτιμητών. Αυτή η σύγκριση δείχνει ότι οι 4 μέθοδοι είναι σχεδόν το ίδιο αποτελεσματικές.

Εντελώς διαφορετική εικόνα έχουμε όσον αφορά την εκτίμηση των διαμέσων των περιοχών. Στον παρακάτω πίνακα βλέπουμε τις τιμές για τα ίδια μέτρα για τις 4 μεθόδους όσον αφορά τις εκτιμήσεις των διαμέσων για κάθε περιοχή:

Region	Random Intercept	Random Slope	*M*-quantile	Expectile	Random Intercept	Random Slope	*M*-quantile	Expectile
	Relative Bias				Relative Root Mean Squared Error			
1	44·54	33·01	35·17	37·87	52·16	54·11	75·28	75·19
2	168·74	180·47	10·67	-0·81	182·04	209·03	68·38	76·09
3	51·74	54·59	7·91	12·00	56·07	62·07	19·53	21·49
4	35·02	18·64	-10·48	-11·78	43·16	28·77	24·66	27·89
5	-9·05	9·27	-11·01	-20·95	27·30	19·25	19·26	27·05
6	-4·16	36·28	-0·27	-0·67	58·83	59·45	5·24	7·10
7	5·06	12·61	8·38	5·83	24·30	24·03	11·28	12·78
8	-3·73	-4·30	-5·88	-3·96	10·88	10·93	12·65	12·27
9	10·89	11·15	-12·21	-14·21	17·87	17·10	20·87	22·86
10	-14·56	-14·19	-22·31	-17·61	17·68	17·70	25·97	21·42
11	71·73	62·21	39·27	40·21	76·12	67·71	44·13	46·27
12	35·05	22·09	6·48	6·42	40·16	27·27	17·48	16·54
13	15·18	18·79	-12·48	-13·63	25·70	27·09	15·17	16·25
14	7·99	7·23	1·11	1·59	13·90	13·67	15·20	14·60
15	4·32	4·51	-3·92	-4·77	34·73	32·65	11·57	12·00
16	58·15	50·72	6·36	5·01	61·45	53·76	15·61	13·97
17	28·28	29·52	-0·51	-0·62	37·90	38·22	6·70	7·85
18	17·82	34·61	-15·52	-16·91	37·97	53·96	23·36	24·26
19	30·91	29·72	6·69	8·10	33·09	31·76	11·42	13·16
20	68·38	77·64	59·60	60·68	80·66	85·74	67·38	68·66
21	55·26	67·53	10·40	11·59	92·25	96·63	17·12	18·34
22	94·74	86·35	53·58	52·79	104·68	90·06	56·77	59·34
23	7·06	15·12	-9·14	-10·16	31·01	31·70	12·70	14·03
24	28·40	23·99	-3·50	-6·83	36·82	35·90	8·46	11·56
25	29·56	25·51	7·59	2·51	33·30	28·18	11·60	10·59
26	30·50	31·67	13·29	10·68	35·50	36·00	14·95	13·78
27	15·47	9·65	-11·11	-11·77	21·48	19·42	12·46	13·44
28	8·01	7·08	-1·64	-2·74	13·78	12·24	5·60	6·09
29	6·56	6·37	-10·53	-11·15	11·96	10·29	11·72	12·18
Mean	30·96	32·68	4·69	3·68	45·27	44·65	22·85	24·04
Median	28·30	24·00	-0·27	-0·67	35·50	31·76	15·20	14·60

ΠΙΝΑΚΑΣ 2-10: Πίνακας με τα μέτρα σύγκρισης των εκτιμήσεων των διαμέσων μέσω ΓΜΜ και Μ-ΠΠΣ. (Πηγή: Chambers, Tzavidis,2006)

Εδώ βλέπουμε ότι οι εκτιμήσεις των διαμέσων βάσει των μεθόδων της Μ-ΠΠΣ έχουν πολύ μικρότερη σχετική μεροληψία σε σχέση με τις αντίστοιχες εκτιμήσεις των ΓΜΜ. Επίσης οι εκτιμήσεις της Μ-ΠΠΣ παρουσιάζουν και μικρότερη διακύμανση. Γι' αυτό και οι σχετικές ρίζες των μέσων τετραγωνικών σφαλμάτων των ΓΜΜ είναι σχεδόν

διπλάσιες από αυτές των μεθόδων της Μ-ΠΠΣ. Επομένως στην εκτίμηση των διαμέσων των περιοχών η Μ-ΠΠΣ είναι πιο αποτελεσματική από τα ΓΜΜ.

3.8 Μία εναλλακτική προσέγγιση στη σύνδεση μεταξύ ΠΠΣ και ΓΜΜ

Όπως είδαμε στις προηγούμενες παραγράφους, τα μικτά μοντέλα αποτελούν τον πλέον διαδεδομένο τρόπο για τη μοντελοποίηση δεδομένων σαν αυτά που περιγράψαμε στην παράγραφο 3.1. Ο Roger Koenker σε περσινή εργασία του με τίτλο «Quantile Regression for longitudinal data», που δεν έχει δημοσιευτεί προς το παρόν σε κάποιο περιοδικό αλλά είναι διαθέσιμη στο διαδύκτιο, επιχειρεί να κάνει μία σύνδεση του κλασικού ΓΜΜ με το μοντέλο της ΠΠΣ, με σκοπό να πετύχει μία πιο ανθεκτική προσέγγιση στην ανάλυση τέτοιων δεδομένων.

Θεωρούμε και πάλι το ΓΜΜ (3.2.1) όπως το ορίσαμε στην παράγραφο 3.2 κάτω από τους γνωστούς περιορισμούς:

$$y_{ij} = x_{ij}^T b + z_{ij}^T g_j + u_{ij}, \qquad (i=1,\ldots,n\ , j=1,\ldots d)$$

ή αλλιώς:

$$E\left(y_{ij} \mid x_{ij}\right) = a_i + x_{ij}^T b\ , \qquad (3.7.1)$$

όπου τα a_i εκφράζουν την ειδική μεταβλητότητα κάθε υποκειμένου και αυτή που δεν ενσωματώθηκε στις επεξηγηματικές μεταβλητές.

Το παραπάνω μοντέλο μπορούμε, για παράδειγμα, να το χρησιμοποιήσουμε σε μία ιατρική έρευνα, όπου ο δείκτης i κατατάσσει τους ασθενείς και ο δείκτης j κατατάσσει σε περιοχές τις d ξεχωριστές μετρήσεις που έκανε κάθε ασθενής.

Για να επεκτείνουμε το μοντέλο (3.7.1) σε ένα μοντέλο ΠΠΣ θα πρέπει να σκεφτούμε τι ρόλο θα παίζουν τα a_i. Όπως τονίσαμε και σε προηγούμενη παράγραφο στις περισσότερες εφαρμογές ο αριθμός d των παρατηρήσεων i είναι σχετικά μικρός στις διάφορες περιοχές j, άρα στην ΠΠΣ δε θα ήταν ρεαλιστικό να επιχειρήσουμε να

εκτιμήσουμε τις ειδικές επιδράσεις κάθε υποκειμένου που εξαρτώνται από το εκάστοτε *τ*-ποσοστιαίο σημείο που μας ενδιαφέρει. Έτσι θεωρούμε το μοντέλο της ΠΠΣ όπως παρακάτω:

$$Q_t\left(y_{ij} \mid x_{ij}\right) = a_i + x_{ij}^T b\left(t\right), \qquad (3.7.2)$$

όπου τα a_i ενσωματώνουν μόνο την επίδραση θέσης κάθε υποκειμένου στα δεσμευμένα ποσοστιαία σημεία της εξαρτημένης μεταβλητής *y*. Αντίθετα με τα a_i, οι επιδράσεις των επεξηγηματικών x_{ij} επιτρέπεται να εξαρτώνται από το εκάστοτε ποσοστιαίο σημείο ενδιαφέροντος, *τ*.

Για να εκτιμήσουμε το μοντέλο (3.7.2) για διαφορετικά ποσοστιαία σημεία ταυτόχρονα, προτείνεται ο παρακάτω εκτιμητής, που είναι παρόμοιας λογικής με τον εκτιμητή (1.3.5):

$$\min_{(a,b)} \sum_{k=1}^{q} \sum_{j=1}^{n} \sum_{i=1}^{d} w_k \cdot r_{t_k}\left(y_{ij} - a_i - x_{ij}^T b\left(t_k\right)\right), \qquad (3.7.3)$$

όπου r_{t_k} η «συνάρτηση ελέγχου» (1.3.2) για το ποσοστιαίο σημείο t_k. Τα βάρη w_k σταθμίζουν τις σχετικές επιρροές καθενός από τα q ποσοστιαία σημεία $\{t_1,...t_q\}$ με βάση την εκτίμηση των παραμέτρων a_i. Η επιλογή των βαρών w_k και των αντίστοιχων ποσοστιαίων σημείων είναι παρόμοια με την επιλογή του σταθμισμένου L-εκτιμητή (3.3.14).

Η λύση του προβλήματος ελαχιστοποίησης (3.7.3) προκύπτει με τη χρήση κατάλληλου αλγορίθμου σαν αυτούς που περιγράψαμε στην παράγραφο 1.4. Συνήθως χρησιμοποιείται η Μέθοδος Εσωτερικού Σημείου (Portnoy και Koenker, 1997).

ΠΑΡΑΡΤΗΜΑ

Παράρτημα Α: Στατιστικό πακέτο

Η στατιστική ανάλυση μέσω της ΠΠΣ απαιτεί ένα κατάλληλο στατιστικό πακέτο. Η στατιστική γλώσσα που έχει αναπτυχθεί περισσότερο και είναι πιο εύχρηστη στον τομέα αυτόν είναι η **R**.

Η **R** είναι ένα στατιστικό πακέτο που μπορεί να κατεβάσει κανείς ελεύθερα αν επισκεφτεί τη διεύθυνση ***http://lib.stat.cmu.edu/R/CRAN/***. Έχοντας μπει στην ιστοσελίδα αυτή, ακολουθούμε τα παρακάτω βήματα:

1. Κάνουμε κλικ στο ***Windows (95 and later)*** που βρίσκεται στα δεξιά της οθόνης μας
2. Κάνουμε κλικ στο ***base***.
3. Κάνουμε κλικ στο ***R-2.4.1-win32.exe*** (τελευταία έκδοση που έχει κυκλοφορήσει) και σώζουμε το αρχείο στον υπολογιστή μας.
4. Κάνουμε διπλό κλικ στο exe αρχείο που έχουμε σώσει και αρχίζει αυτόματα η εγκατάσταση του προγράμματος.

Με την ολοκλήρωση των παραπάνω βημάτων θα πρέπει να εγκαταστήσουμε ένα ειδικό πακέτο για την ΠΠΣ. Αυτό γίνεται αν στην ίδια ιστοσελίδα κάνουμε κλικ στο “R packages” και από τα διαθέσιμα πακέτα που εμφανίζονται κάνουμε κλικ στο “quantreg” για να το εγκαταστήσουμε. Εναλλακτικά αυτό μπορεί να γίνει και από το περιβάλλον της ίδιας της **R**, εφόσον είμαστε συνδεδεμένοι στο ίντερνετ, αν πληκτρολογήσουμε την εντολή:

```
> install.packages(“quantreg”)
```

Εφόσον το συγκεκριμένο πακέτο έχει εγκατασταθεί, μπορούμε να το ενεργοποιήσουμε παραθυρικά κάνοντας κλικ στο “Packages” → Load package → “quantreg” ή μέσω της εντολής:

> library(quantreg)

Σε περίπτωση που ο χρήστης θέλει να μάθει λεπτομέρειες για τις διάφορες εντολές που χρησιμοποιούνται για την ΠΠΣ, τότε μπορεί να πληκτρολογήσει:

> help(package = “quantreg”) ή
> help(rq)

Επίσης η **R** προσφέρει και κάποια παραδείγματα εφαρμογών της ΠΠΣ μέσω των εντολών της. Για να τα δει κανείς αυτά, αρκεί να δώσει την εντολή:

> example(rq)

Ο αλγόριθμος που χρησιμοποιεί η R για την εκτίμηση των παραμέτρων είναι η Μέθοδος Εσωτερικού Σημείου (Portnoy & Koenker, 1997).

Παράρτημα Β: Δεδομένα

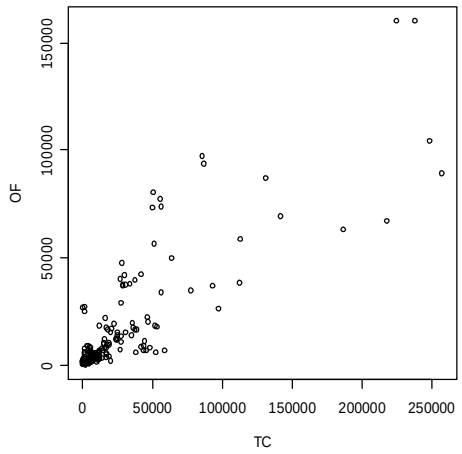
2.5

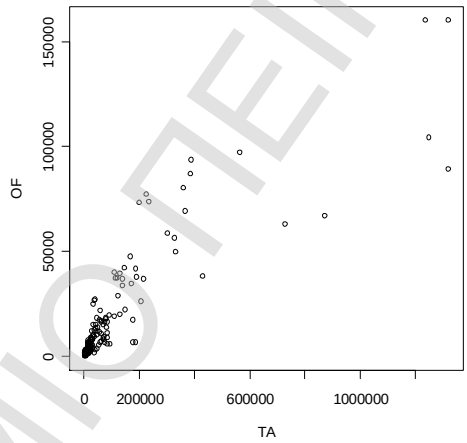
ΒΟΥΛΕΥΤΙΚΕΣ ΕΚΛΟΓΕΣ 7ης ΜΑΡΤΙΟΥ 2004

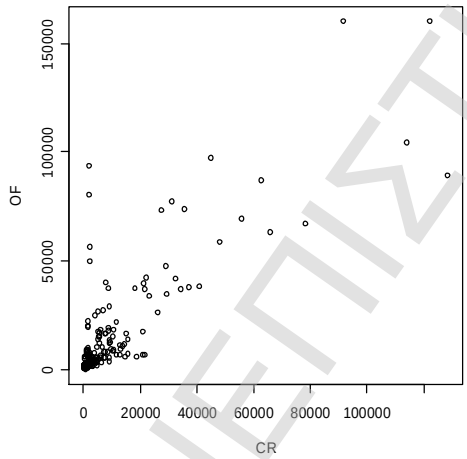
ΠΑΡΑΡΤΗΜΑ Γ: Διαγράμματα

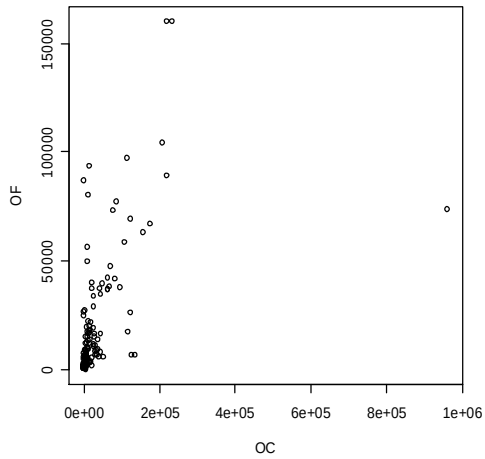
2.6

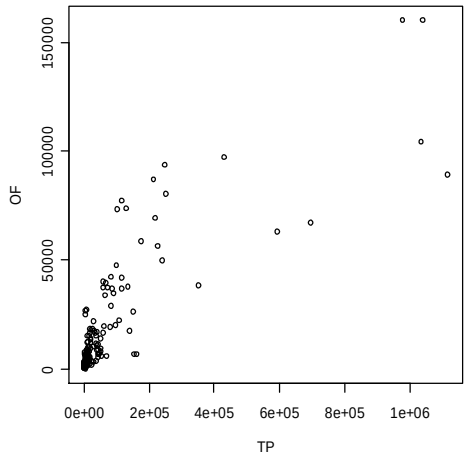
Διαγράμματα διασποράς

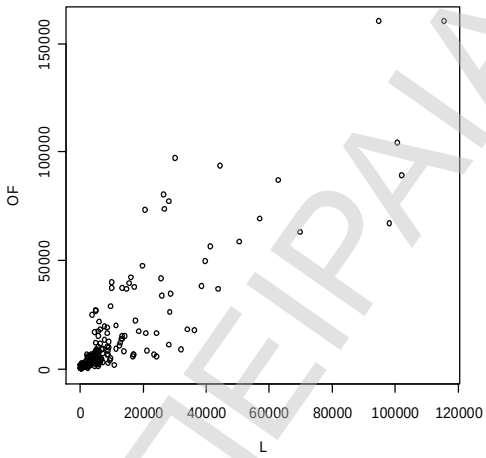


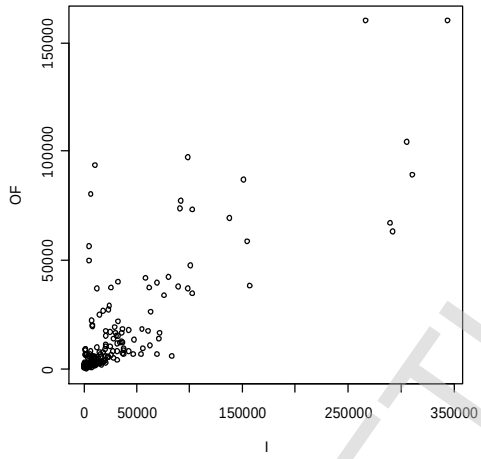


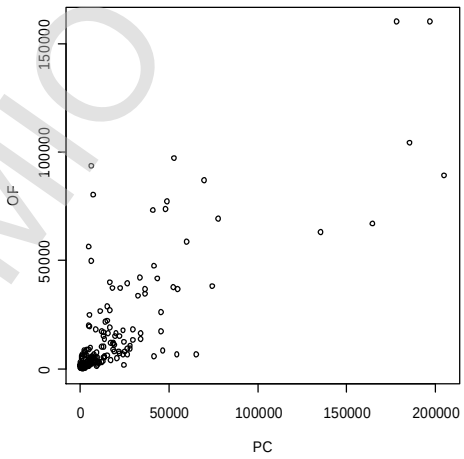


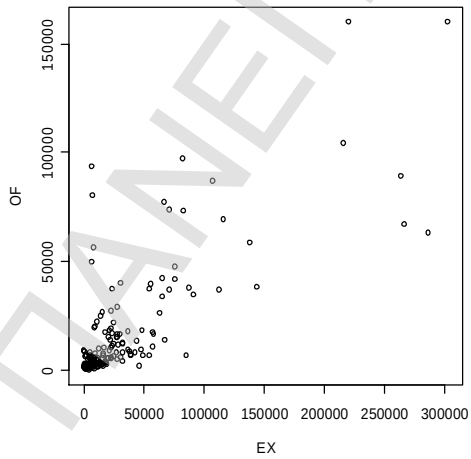




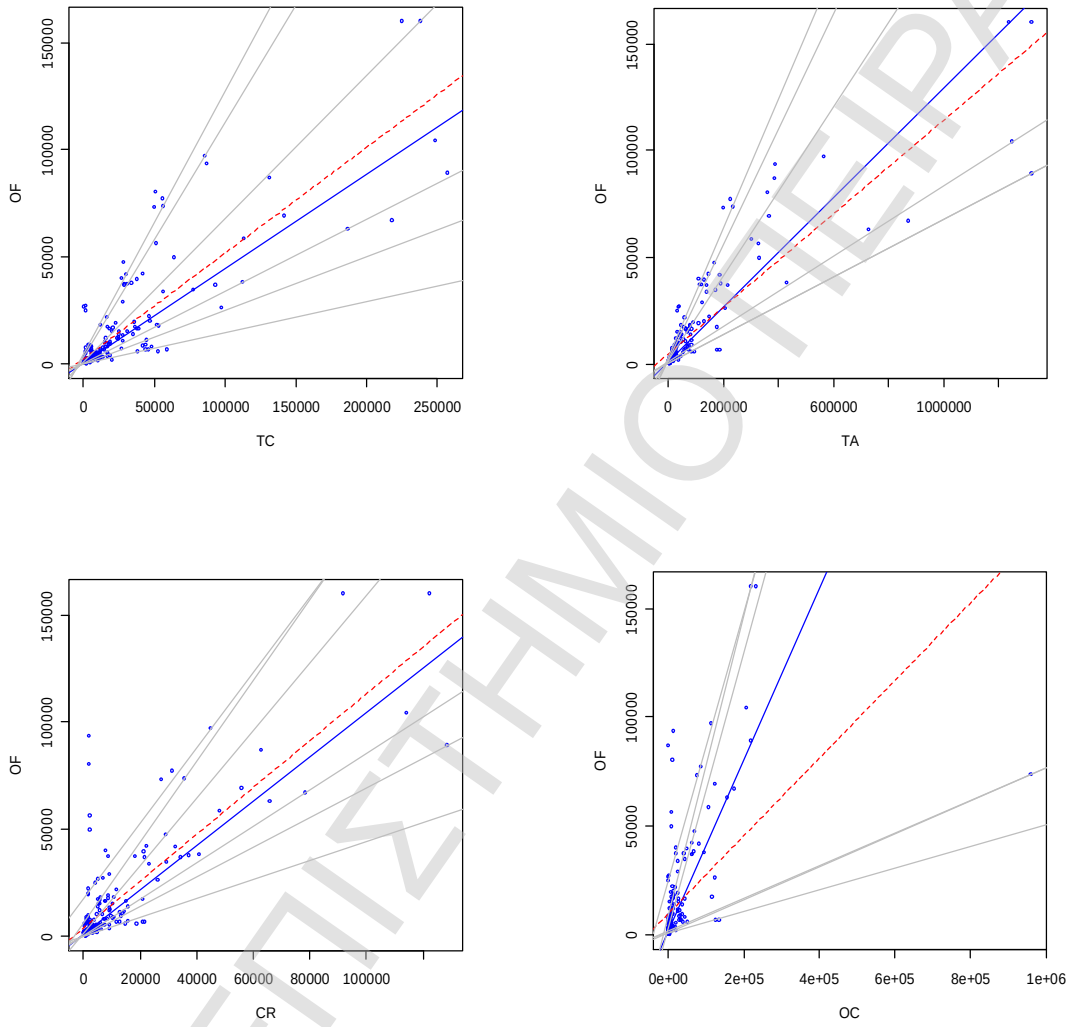


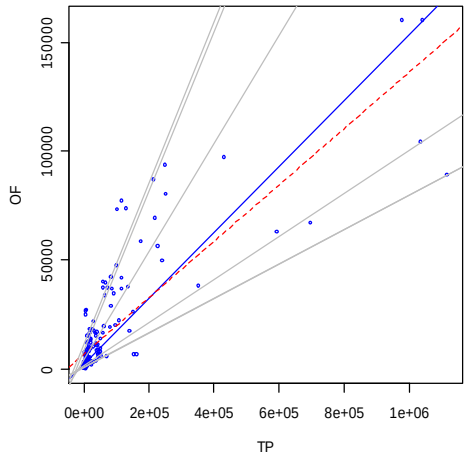


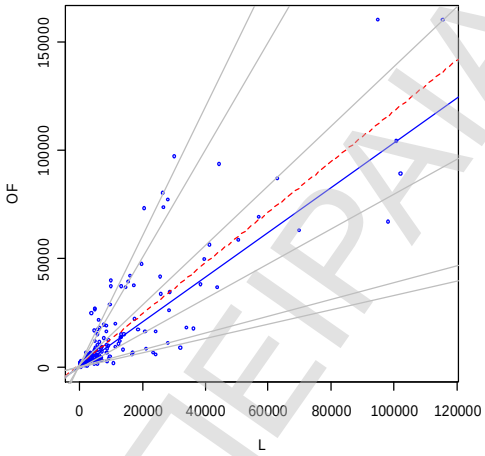


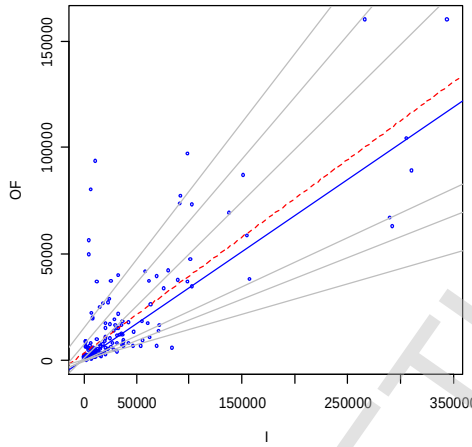


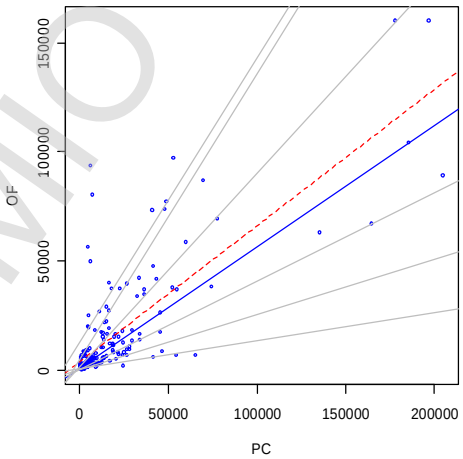
Διαγράμματα διασποράς με τις προσαρμοσμένες ευθείες ΠΠΣ ($\tau = 0.05, 0.25, 0.5, 0.75, 0.95$) και ΚΓΠ

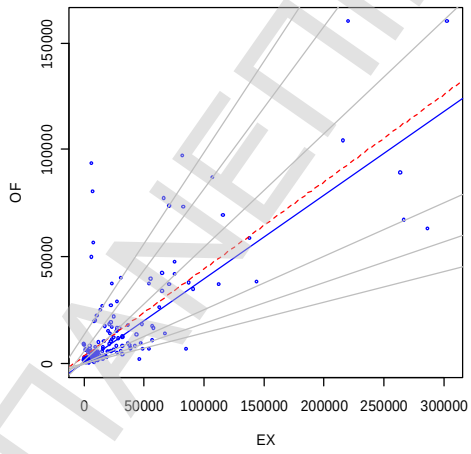












ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Abreveya, J. (2001). The Effects of Demographics and Maternal Behavior on the Distribution of Birth Outcomes. *Empirical Economics*, **26**, 246-257.

[2] Barrodale, I. and Roberts, F. (1973). An Improved Algorithm for Discrete L_1 Linear Approximation. *SIAM Journal of Numerical Analysis*, **10**, 839-848.

[3] Basset, G., Tam, M. and Knight, K. (2002). Quantile Models and Estimators for Data Analysis. *Metrika*, **55**, 17-26.

[4] Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, **75**, 761-771.

[5] Buchinsky, M. (1994). Changes in U.S. Wage Structure 1963-1987. An Application of Quantile Regression. *Econometrika*, **62**, 405-458.

[6] Buchinsky, M. (1997). The Dynamics of Changes in the Female Wage Distribution in the U.S.A.: A Quantile Regression Approach. *Journal of Applied Econometrics*, **13**, 1-30.

[7] Buchinsky, M. (1998). Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research. *The Journal of Human Resources*, **33**, 88-126.

[8] Cade, B. (2003). Quantile Regression Models of Animal Habitat Relationships. *Fort Colins Center Science Center Publication (http://www.fort.usgs.gov/).*

[9] Chamberlain, G. (1994). Quantile Regression, Censoring and the Structure of Wage. *In Proceeding of the Sixth World Congress of the Econometric Society, ed. Sims, C., New York: Cambridge University Press.*

[10] Chambers, R. and Tzavidis, N. (2006). M-quantile Models for Small Area Estimation. *Biometrika Trust*, **93**, 255-268.

[11] Chaudhuri, P., Doksum, K. and Samarov A. (1997). On Average Derivative Quantile Regression, *Annals of Statistics*, **25**, 715-744.

[12] Chen, C. (2004). An Adaptive Algorithm for Quantile Regression. Theory and Applications of Recent Robust Methods, *ed. Hubert, M., Pison, G., Struyf, A. and Van Aeist, S., Series: Statistics for Industry and Technology, Birkhauser, Basel*, 39-48.

[13] Donald, S.G. and Paarch, H.J. (1993). Piecewise Pseudo-Maximum Likelihood Estimation in Empirical Models of Auctions. *International Economic Review*, **34**, 121-148.

[14] Eide, E. and Showalter, M. (1998). The Effect of School Quality on Student Performance: A Quantile Regression Approach. *Economics Letters*, **58**, 345-350.

[15] Engel, E. (1857). Die Produktions und Konsuptionverhaltnisse des Konigreichs Sachsen. Reprinted in Die Lebencosten Blgischer Arbeiter-Familien Fruher und Jetzt. *International Statistical Bulletin*, **9**, 1-125.

[16] Fitzenberger, B. (1999). Wages and Employment Across Skill Groups. Heidelberg: Phisica-Verlag.

[17] Garcia, J., Hernadez, P. and Lopez-Nicolas, A. (2001). How wide is the gap? An Investigation of Gender Wage Differences Using Quantile Regression. *Empirical Economics*, **26**, 149-167.

[18] Gutembrunner, C. and Jureckova, J. (1992). Regression Quantile and Regression Rank Score Process in the Linear Model and Derived Statistics. *Annals of Statistics*, **20**, 305-330.

[19] Hampel, F.R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, **42**, 1887-1896.

[20] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). Robust Statistics: The Approach based on Influence functions, *Wiley, New York*.

[21] Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics* **9**, 226-52.

[22] Hendricks, N. and Koenker, R. (1991). Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity. *Journal of the American Statistical Association*, **87**, 58-68.

[23] Huber , P.J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, **35**, 73-101.

[24] Knight, K. (1998). Limiting Distributions of L_1 Regression Estimators under General Conditions. *Annals of Statistics*, **26**, 755-770.

[25] Koenker, R. (2005). Quantile Regression, *Economic society monographs, Cambridge*.

[26] Koenker, R. (2005). Quantile Regression for Longitudinal Data. *(http://www..econ.uiuc.edu/~roger/research/panel/long.pdf)*

[27] Koenker, R. and Basset, G. (1978). Regression Quantiles. *Econometrica*, **46**, 33-50.

[28] Koenker, R. and D'Orey, V. (1987). Computing Regression Quantiles. *Journal of the Royal Statistical Society, Applied Statistics*, **36**, 383-393.

[29] Koenker, R. and Gerling, O. (2001). Reappraising Medfly Longevity: A Quantile Regression Survival Analysis. *Journal of the American Statistical Association*, **96**, 458-468.

[30] Koenker, R. and Hallock, K.F. (2001). *Journal of Economic Perspectives*, **15**, 143-156.

[31] Koenker, R. and Portnoy, S. (1987). L-Estimation for Linear Models. *Journal of the American Statistical Association*, **82**, 851-857.

[32] Koenker, R. and Portnoy, S. (1990). M-Estimation of Multivariate Regressions. *Journal of the American Statistical Association*, **85**, 1060-1068.

[33] Kokic, P., Chambers, R., Breckling, J. and Beare, S. (1997). A Measure of Production Performance. *Journal of Business and Economic Statistics*, **15**, 445-451.

[34] Machado, J. (1993). Robust Model Selection and *M*-estimation. *Economic Theory*, **9**, 478-493.

[35] Machado, J. and Mata, J. (2001). Counter-factual Decomposition of Changes in Wage Distributions Using Quantile Regression. *Empirical Economics*, **26**, 115-134.

[36] Pitselis, G. (2006). Solvency Supervision, Regulations and Insolvency Prediction: The Case of Greece. *Presented in International Congress Euro2006 on Operation Research, Reykjavik, Iceland*.

[37] Pitselis, G. (2007). Risk Based Capital, Supervision of Solvency and Cross-Section Effect models. *Presented in International Congress IME 2006, Leuven, Belgium.*

[38] Portnoy, S. (1991). Asymptotic Behavior of Regression Quantiles in Non-Stationary, Depended Cases. *Journal of Multivariate Analysis,* **38**, 100-113.

[39] Portnoy, S. and Koenker, R. (1997). The Gaussian Hare and the Laplacian Tortoise: Computability of Squared-Error Versus Absolute-Error Estimators, with Discussion. *Statistical Science*, **12**, 279-300.

[40] Powell, J. (1986). Cencored Regression Quantiles. *Journal of Econometrics,* **25**, 303-325.

[41] Schultz, T.P. and Mwabu, G. (1998). Labor Unions and the Distribution of Wages and Employment in South Africa. *Industrial and Labor Relations Review*, **51**, 680-703.

[42] Tsay, R.S. (2002). Analysis of Financial Time Series, *John Wiley and Sons, New York.*