



ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΑΚΟ ΠΡΟΓΡΑΜΜΑ

ΕΠΙΣΚΟΠΗΣΗ ΤΩΝ ΜΕΘΟΔΩΝ ΕΚΤΙΜΗΣΗΣ
ΤΟΥ ΚΕΝΤΡΙΚΟΥ ΥΠΟΧΩΡΟΥ ΜΕ ΣΚΟΠΟ
ΤΗ ΜΕΛΕΤΗ ΤΗΣ ΔΕΣΜΕΥΜΕΝΗΣ
ΚΑΤΑΝΟΜΗΣ y/x

Γεώργιος Δημητρίου-Δάμωνος Βλάσσης

ΕΡΓΑΣΙΑ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση

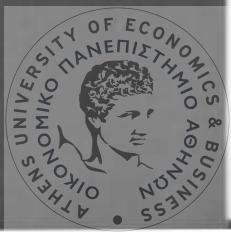
Μετακυριακού Διπλώματος
Συμπληρωματικής Εδίκευσης στη Στατιστική
Μερικής Παρακολούθησης (Part-time)

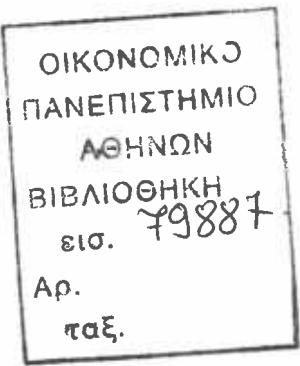
Αθήνα
Απρίλιος 2006

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΚΑΤΑΛΟΓΟΣ



0 000000 572132 0





ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΕΠΙΣΚΟΠΗΣΗ ΤΩΝ ΜΕΘΟΔΩΝ ΕΚΤΙΜΗΣΗΣ ΤΟΥ
ΚΕΝΤΡΙΚΟΥ ΥΠΟΧΩΡΟΥ ΜΕ ΣΚΟΠΟ ΤΗ ΜΕΛΕΤΗ
ΤΗΣ ΔΕΣΜΕΥΜΕΝΗΣ ΚΑΤΑΝΟΜΗΣ $y|x$.

Γεώργιος Δημητρίου-Δάμωνος Βλάσσης

ΕΡΓΑΣΙΑ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος
Συμπληρωματικής Ειδίκευσης στη Στατιστική
Μερικής Παρακολούθησης (Part-time).



Αθήνα

Φεβρουάριος 2006





ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Εργασία που υποβλήθηκε ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Συμπληρωματικής Ειδίκευσης στη Στατιστική
Μερικής Παρακολούθησης (Part-time)

ΕΠΙΣΚΟΠΗΣΗ ΤΩΝ ΜΕΘΟΔΩΝ ΕΚΤΙΜΗΣΗΣ ΤΟΥ ΚΕΝΤΡΙΚΟΥ ΥΠΟΧΩΡΟΥ ΜΕ ΣΚΟΠΟ ΤΗ ΜΕΛΕΤΗ ΤΗΣ ΔΕΣΜΕΥΜΕΝΗΣ ΚΑΤΑΝΟΜΗΣ y/x

Γεώργιος Δημητρίου-Δάμωνος Βλάσσης



Υπεύθυνο μέλος ΔΕΠ:
Παν. Τσιαμυρτζής
Λέκτορας

Ο Διευθυντής Μεταπτυχιακών Σπουδών

Μιχαήλ Ζαζάνης
Καθηγητής



ΕΥΧΑΡΙΣΤΙΕΣ

Ευχαριστώ τον καθηγητή μου και λέκτορα του Οικονομικού Πανεπιστημίου Αθηνών Παναγιώτη Τσιαμυρτζή για το ενδιαφέρον, την ενθάρυνση του και την εμπιστοσύνη με την οποία με περιέβαλε.





ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

Ο εκπονήσας την παρούσα διπλωματική εργασία είναι διπλωματούχος Μηχανολόγος Μηχανικός ΕΜΠ, και είναι δόκιμος τακτικός δημόσιος υπάλληλος σε κρατικό ασφαλιστικό οργανισμό (ΟΑΕΕ-ΤΣΑ). Έχει επίσης απασχοληθεί επί διετία στον τομέα Η/Μ μελετών και επί τριετία σαν Μηχανικός Παραγωγής στη βιομηχανία. Είναι κάτοχος Proficiency και φιλοδοξεί να συνεχίσει την ενασχόληση του με θέματα ακαδημαϊκού ενδιαφέροντος.





ABSTRACT

REVIEW OF THE CENTRAL SUBSPACE ESTIMATION METHODS IN PURSUE OF THE STUDY OF THE CONDITIONAL DISTRIBUTION $y|x$.

George Vlassis

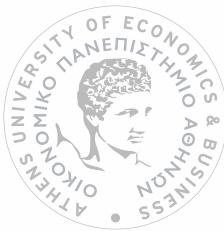
February 2006

The present thesis addresses the problem of regression, namely the study of the conditional distribution $y|x$, where $x = (x_1, \dots, x_p)^T$, via the use of methods seeking for a subspace of minimum dimension k upon which the projection of the data cloud (y, x) captures all the information that is available for the response y from the covariate x .

This subspace is called “central subspace” and its basis B , for which $y \perp\!\!\!\perp x | B^T x$, is to be estimated via the eigenvector decomposition of the matrix \hat{K} which is a consistent estimate of the matrix K for which $S(K) \subseteq S_{y|x}$, where $S_{y|x}$ is the central subspace.

The matrix K is different for each method and the linkage between K and the basis B of $S_{y|x}$ is established through theory.

It is in this fashion that the lack of sufficient data that is intrinsic to non-parametric regression methods, the dimension of the covariate x being high, is bypassed as well as the difficulty of choosing a suitable model as to the parametric regression methods. The only shortcoming of the central subspace estimation methods is related to the fulfillment of certain assumptions that must not be severely violated for the methods to be reliable. These assumptions concern the distribution of the covariate x and the features, in respect to the mean and covariance, of the projection of the data cloud (y, x) upon specific planes.





ΠΕΡΙΛΗΨΗ

**ΕΠΙΣΚΟΠΗΣΗ ΤΩΝ ΜΕΘΟΔΩΝ ΕΚΤΙΜΗΣΗΣ ΤΟΥ ΚΕΝΤΡΙΚΟΥ
ΥΠΟΧΩΡΟΥ ΜΕ ΣΚΟΠΟ ΤΗ ΜΕΛΕΤΗ ΤΗΣ ΔΕΣΜΕΥΜΕΝΗΣ
ΚΑΤΑΝΟΜΗΣ $y|x$.**

Γιώργος Βλάσσης

Φεβρουάριος 2006

Στην παρούσα εργασία αντιμετωπίζεται το πρόβλημα της παλινδρόμησης, δηλαδή της μελέτης της δεσμευμένης κατανομής $y|x$, όπου $x = (x_1, \dots, x_p)^T$, μέσω της χρήσης μεθόδων που σαν στόχο έχουν την εύρεση ενός υπόχωρου ελάχιστης διάστασης κ στον οποίο η προβολή των σημείων (y, x) να παρέχει όλη την πληροφορία που είναι διαθέσιμη για την απόκριση y μέσω του διανύσματος x των ανεξαρτήτων μεταβλητών.

Ο υπόχωρος αυτός καλείται «κεντρικός υπόχωρος» και η βάση του B εκτιμάται με βάση την ιδιοδιανυσματική ανάλυση της μήτρας \hat{K} η οποία αποτελεί συνεπή εκτιμήτρια της μήτρας K για την οποία ισχύει

$$S(K) \subseteq S_{y|x}, \text{όπου } S_{y|x} \text{ ο κεντρικός υπόχωρος, ισχύει δε } y \perp\!\!\!\perp x | B^T x.$$

Η μήτρα K είναι διαφορετική για κάθε μέθοδο και η διασύνδεση της με τη βάση του $S_{y|x}$ προκύπτει από τη θεωρία.

Παρακάμπτεται έτσι το πρόβλημα των μη παραμετρικών μεθόδων παλινδρόμησης που συνίσταται στην μη ύπαρξη επαρκούς ποσότητας δεδομένων η οποία οφείλει να αυξάνει εκθετικά αυξανομένης της διάστασης του x , αλλά και το πρόβλημα των παραμετρικών μεθόδων που συνίσταται στη δυσκολία επιλογής μοντέλου. Το μόνο μειονέκτημα των μεθόδων εκτίμησης του κεντρικού υπόχωρου είναι η ισχύς κάποιων συνθηκών οι οποίες δεν πρέπει να παραβιάζονται σοβαρά. Οι συνθήκες αυτές αφορούν την κατανομή του x καθώς και τα χαρακτηριστικά της προβολής των σημείων (y, x) σε συγκεκριμένα επίπεδα.





ΚΑΤΑΛΟΓΟΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

σελίδα

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1

ΚΕΦΑΛΑΙΟ 2

ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

2.1 ΚΕΝΤΡΙΚΟΣ ΥΠΟΧΩΡΟΣ (CENTRAL DIMENSION REDUCTION SUBSPACE)	7
2.2 ΔΙΑΣΥΝΔΕΣΗ ΤΗΣ ΜΕΘΟΔΟΥ OLS ΚΑΙ ΤΟΥ ΚΕΝΤΡΙΚΟΥ ΥΠΟΧΩΡΟΥ $Sy x$.	9
2.3 ΠΡΟΫΠΟΘΕΣΕΙΣ ΕΦΑΡΜΟΓΗΣ ΤΩΝ ΜΕΘΟΔΩΝ ΕΚΤΙΜΗΣΗ ΤΟΥ ΚΕΝΤΡΙΚΟΥ ΥΠΟΧΩΡΟΥ	11
2.3.1 ΥΠΟΘΕΣΗ ΓΡΑΜΜΙΚΟΤΗΤΑΣ	12
2.3.2 ΥΠΟΘΕΣΗ ΣΤΑΘΕΡΗΣ ΔΙΑΚΥΜΑΝΣΗΣ	13
2.4 ΣΥΝΘΗΚΕΣ ΕΛΕΓΧΟΥ ΤΟΥ ΜΕΣΟΥ ΚΑΙ ΤΗΣ ΔΙΑΚΥΜΑΝΣΗΣ (MEAN AND VARIANCE CHECKING CONDITIONS)	14

ΚΕΦΑΛΑΙΟ 3

ΜΕΘΟΔΟΙ ΡΟΠΗΣ ΠΡΩΤΗΣ ΤΑΞΗΣ ,*PIR*

3.1 ΜΕΘΟΔΟΣ <i>SIR</i> (<i>SLICED INVERSE REGRESSION</i>)	
3.1.1 ΘΕΩΡΗΤΙΚΗ ΘΕΜΕΛΙΩΣΗ	17
3.1.2 ΑΛΓΟΡΙΘΜΟΣ	18
3.1.3 ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΤΑΣΗΣ d	19
3.1.4 ΕΦΑΡΜΟΓΕΣ	20
3.1.5 ΣΧΟΛΙΑ – ΕΠΙΣΗΜΑΝΣΕΙΣ- ΠΡΟΕΚΤΑΣΕΙΣ	27
3.1.5 α ΠΑΡΑΒΙΑΣΗ ΤΗΣ ΥΠΟΘΕΣΗΣ ΓΡΑΜΜΙΚΟΤΗΤΑΣ ΤΟΥ ΘΕΩΡΗΜΑΤΟΣ 3.1	27
3.1.5 β ΣΤΑΘΜΙΣΜΕΝΟΣ ΕΛΕΓΧΟΣ χ^2	29
3.1.5 γ ΠΑΡΑΛΛΑΓΕΣ ΤΗΣ <i>SIR</i>	33
3.2 ΜΕΘΟΔΟΣ <i>PIR</i> (<i>PARAMETRIC INVERSE REGRESSION</i>)	
3.2.1 ΘΕΩΡΗΤΙΚΗ ΘΕΜΕΛΙΩΣΗ	

	σελίδα
3.2.1α ΠΕΡΙΠΤΩΣΗ ΜΗ ΣΤΑΘΕΡΗΣ ΔΙΑΚΥΜΑΝΣΗΣ	37
3.2.2 ΑΛΓΟΡΙΘΜΟΣ	39
3.2.3 ΕΦΑΡΜΟΓΕΣ	40
3.2.4 ΣΧΟΛΙΑ- ΕΠΙΣΗΜΑΝΣΕΙΣ	45
<u>ΚΕΦΑΛΑΙΟ 4</u>	
ΜΕΘΟΔΟΙ ΡΟΠΗΣ ΔΕΥΤΕΡΗΣ ΤΑΞΗΣ	
4.1 ΜΕΘΟΔΟΣ <i>SAVE</i> (<i>SLICED AVERAGE VARIANCE ESTIMATE</i>)	
4.1.1. ΘΕΩΡΗΤΙΚΗ ΘΕΜΕΛΙΩΣΗ	47
4.1.2 ΑΛΓΟΡΙΘΜΟΣ	48
4.1.3 ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΤΑΣΗΣ d	48
4.1.4 ΕΦΑΡΜΟΓΕΣ	49
4.1.5. ΣΧΟΛΙΑ – ΕΠΙΣΗΜΑΝΣΕΙΣ	50
4.2 ΜΕΘΟΔΟΣ <i>SIR II</i>	
4.2.1 ΘΕΩΡΗΤΙΚΗ ΘΕΜΕΛΙΩΣΗ	51
4.2.2 ΑΛΓΟΡΙΘΜΟΣ	52
4.2.3 ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΤΑΣΗΣ d	53
4.2.4 ΕΦΑΡΜΟΓΕΣ	54
4.2.5 ΣΧΟΛΙΑ – ΕΠΙΣΗΜΑΝΣΕΙΣ	55
4.3 ΜΕΘΟΔΟΣ <i>pHd</i> (<i>PRINCIPAL HESSIAN DIRECTIONS</i>)	
4.3.1 ΘΕΩΡΗΤΙΚΗ ΘΕΜΕΛΙΩΣΗ	56
4.3.2 ΑΛΓΟΡΙΘΜΟΣ	58
4.3.3 ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΤΑΣΗΣ d	59
4.3.3 α ΠΡΟΣΘΕΤΕΣ ΥΠΟΘΕΣΕΙΣ	61
4.3.4 ΕΦΑΡΜΟΓΕΣ	62
4.3.5 ΣΧΟΛΙΑ – ΕΠΙΣΗΜΑΝΕΙΣ	67

ΚΕΦΑΛΑΙΟ 5
ΑΛΛΕΣ ΜΕΘΟΔΟΙ

5.1 ΜΕΘΟΔΟΣ *SAT* (*SLICED AVERAGE THIRD – MOMENT ESTIMATION*)



σελίδα

5.1.1 ΘΕΩΡΗΤΙΚΗ ΘΕΜΕΛΙΩΣΗ	69
5.1.2 ΑΛΓΟΡΙΘΜΟΣ	71
5.1.3 ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΤΑΣΗΣ d	71
5.1.4 ΕΦΑΡΜΟΓΕΣ	72
5.1.5 ΣΧΟΛΙΑ - ΕΠΙΣΗΜΑΝΣΕΙΣ	75
5.2 ΜΕΘΟΔΟΣ COV_k	
5.2.1 ΘΕΩΡΗΤΙΚΗ ΘΕΜΕΛΙΩΣΗ	75
5.2.2 ΑΛΓΟΡΙΘΜΟΣ	78
5.2.3 ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΤΑΣΗΣ d	79
5.2.4 ΕΦΑΡΜΟΓΕΣ	80
5.2.5 ΣΧΟΛΙΑ – ΕΠΙΣΗΜΑΝΣΕΙΣ	84
5.2.5 α ΔΙΑΓΝΩΣΤΙΚΟΣ ΕΛΕΓΧΟΣ ΤΟΥ ΓΡΑΜΜΙΚΟΥ ΜΟΝΤΕΛΟΥ	84
5.2.5 β ΠΑΡΑΒΙΑΣΗ ΥΠΟΘΕΣΕΩΝ	85
5.2.5 γ ΣΥΝΔΕΣΗ ΜΕΘΟΔΩΝ SIR και COV_k	86
5.3 ΜΕΘΟΔΟΣ ΓΡΑΦΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ (GRAPHICAL REGRESSION)	
5.3.1 ΓΡΑΦΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΤΗΣ ΔΙΑΣΤΑΣΗΣ ΤΟΥ S_{yx}	87
5.3.2 ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΜΕΘΟΔΟΥ	88
5.3.3 ΕΦΑΡΜΟΓΕΣ	90
5.3.4 ΣΧΟΛΙΑ – ΕΠΙΣΗΜΑΝΣΕΙΣ	98
5.3.4 α ΧΡΗΣΗ ΚΑΤΑΛΟΙΠΩΝ ΤΕΤΡΑΓΩΝΙΚΗΣ ΠΡΟΣΑΡΜΟΓΗΣ	98
5.3.4 β ΧΡΗΣΗ ΤΗΣ ΜΕΘΟΔΟΥ ΓΙΑ ΤΟ ΔΙΑΓΝΩΣΤΙΚΟ ΕΛΕΓΧΟ ΜΟΝΤΕΛΟΥ	99
5.3.4 γ ΧΡΗΣΗ ΤΩΝ ΑΡΧΙΚΩΝ ΜΕΤΑΒΛΗΤΩΝ $x = (x_1, \dots, x_p)^T$	101
<u>REFERENCES</u>	103







ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 3.1 συντετογμένες(μέσος,τυπ.απόκλιση) του εκτιμώμενου

από τη SIR διανύσματος βάσης $\hat{\beta}_1$. 21

Πίνακας 3.2 συντελεστές προσδιορισμού των $\hat{\beta}_1^T x, \hat{\beta}_2^T x$,όπου $\hat{\beta}_1, \hat{\beta}_2$

οι εκτιμήσεις της SIR,στις $\beta_1^T x, \beta_2^T x$,όπου β_1, β_2 τα

πραγματικά διανύσματα βάσης(Μοντέλο 3.2). 22

Πίνακας 3.3 συντελεστές προσδιορισμού των $\hat{\beta}_1^T x, \hat{\beta}_2^T x$,όπου $\hat{\beta}_1, \hat{\beta}_2$

οι εκτιμήσεις της SIR,στις $\beta_1^T x, \beta_2^T x$,όπου β_1, β_2 τα

πραγματικά διανύσματα βάσης(Μοντέλο 3.3). 22

Πίνακας 3.4 τιμές της παρατηρούμενης ισχύος(L_0) και του παρατηρούμενου

επιπέδου σημαντικότητας(L_1) του κατά SIR ελέγχου για την

εκτίμηση της διάστασης του κεντρικού υπόχωρου για τα

δεδομένα του μοντέλου 3.11.(σε παρένθεση τιμές για

ονομαστικό επίπεδο σημαντικότητας 0.01). 41

Πίνακας 3.5 τιμές της παρατηρούμενης ισχύος(L_0) και του παρατηρούμενου

επιπέδου σημαντικότητας(L_1) του ελέγχου για την εκτίμηση της

rank (B) για τα δεδομένα του μοντέλου 3.11.(σε παρένθεση τιμές

για ονομαστικό επίπεδο σημαντικότητας 0.01). 42

Πίνακας 3.6 τιμές της παρατηρούμενης ισχύος(L_0, L_1) και του παρατηρούμενου

επιπέδου σημαντικότητας(L_2) του κατά SIR ελέγχου για την

εκτίμηση της διάστασης του κεντρικού υπόχωρου για τα

δεδομένα του μοντέλου 3.12.(σε παρένθεση τιμές για

ονομαστικό επίπεδο σημαντικότητας 0.01). 43

Πίνακας 3.7 τιμές της παρατηρούμενης ισχύος(L_0, L_1) και του παρατηρούμενου

επιπέδου σημαντικότητας(L_2) του ελέγχου για την εκτίμηση

της rank (B) για τα δεδομένα του μοντέλου 3.12.(σε παρένθεση

τιμές για ονομαστικό επίπεδο σημαντικότητας 0.01). 43



Πίνακας 3.8 τιμές της παρατηρούμενης ισχύος(L_0, L_1) και του παρατηρούμενου επιπέδου σημαντικότητας(L_2) του κατά SIR ελέγχου για την εκτίμηση της διάστασης του κεντρικού υπόχωρου για τα δεδομένα του μοντέλου 3.4-3.5.(σε παρένθεση τιμές για ονομαστικό επίπεδο σημαντικότητας 0.01).	44
Πίνακας 3.9 τιμές της παρατηρούμενης ισχύος(L_0, L_1) και του παρατηρούμενου επιπέδου σημαντικότητας(L_2) του ελέγχου για την εκτίμηση της rank (B) για τα δεδομένα του μοντέλου 3.4-3.5.(σε παρένθεση τιμές για ονομαστικό επίπεδο σημαντικότητας 0.01).	44
Πίνακας 4.1 τιμές της γωνίας(σε μοίρες) ανάμεσα στο ιδιοδιάνυσμα για τη μεγαλύτερη ιδιοτιμή με βάση τις μεθόδους SIR,SAVE,pHd, και την πραγματική βάση.	50
Πίνακας 5.1 εκτιμήσεις \hat{n}_1, \hat{n}_2 της βάσης του κεντρικού υπόχωρου (μέθοδος COV ₂).	81
Πίνακας 5.2 συντελεστές προσδιορισμού από την παλινδρόμηση των μεταβλητών που αντιστοιχούν στις δύο πρώτες κατευθύνσεις της SIR(για $h=3,4$),στις $\hat{n}_1^T x, \hat{n}_2^T x$.	82
Πίνακας 5.3 5,50,95 ποσοστιαία σημεία της κατανομής τιμών του συντελεστή προσδιορισμού από την παλινδρόμηση των μεταβλητών που αντιστοιχούν στις δύο πρώτες κατευθύνσεις της SIR και των μεταβλητών που αντιστοιχούν στις δύο πρώτες κατευθύνσεις της COV ₂ ,στις $\hat{n}_1^T x, \hat{n}_2^T x$,όπου (\hat{n}_1, \hat{n}_2) η βάση του κεντρικού υπόχωρου για τα δεδομένα του μοντέλου 5.1.	83
Πίνακας 5.4 συντελεστές των μεταβλητών Fit,gr1,gr2,gr3.	93
Πίνακας 5.5 συντελεστές των μεταβλητών Fit,gr1, gr2,gr3,gr4.	95

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ

Διάγραμμα 2.1 *scatterplot matrix* των περιθωρίων κατανομών $x_i | x_j$, 15

$$x_j|y \text{ και } y|x_j \text{ για } \mathbf{x} = (H, L, S, W)^T.$$

Διάγραμμα 2.2 *scatterplot matrix* των περιθωρίων κατανομών $x_i | x_j$, 16

$$x_j|y \text{ και } y|x_j \text{ για } \mathbf{x} = (\log H, \log L, \log S, \log W)^T.$$

Διάγραμμα 3.1 διάγραμμα $\{\mathbf{M}, \hat{\boldsymbol{\phi}}_1^T \mathbf{x}\}$ για το πρόβλημα με τα 23

$$\text{οστρακοειδή}(\hat{\boldsymbol{\phi}}_1 \text{ η πρώτη κατεύθυνση που εκτιμά } \eta \text{ SIR}).$$

Διάγραμμα 3.2 διάγραμμα $\{\mathbf{M}, \mathbf{b}_{ols}^T \mathbf{x}\}$ για το πρόβλημα με τα 25

$$\text{οστρακοειδή}(\mathbf{b}_{ols} \text{ η κατεύθυνση OLS}).$$

Διάγραμμα 3.3 διάγραμμα $\{\hat{\boldsymbol{\phi}}_1^T \mathbf{x}, \mathbf{b}_{ols}^T \mathbf{x}\}$ για το πρόβλημα με τα 25

$$\text{οστρακοειδή}.$$

Διάγραμμα 3.4 διάγραμμα $\{\hat{e}, \hat{\boldsymbol{\phi}}_1^T \mathbf{x}\}$ για το πρόβλημα με τα 26

$$\text{οστρακοειδή}(\hat{e} \text{ τα κατάλοιπα του μοντέλου}$$

$$\mathbf{M} | \mathbf{x} = \beta_0 + \boldsymbol{\beta}^T \mathbf{x} + \varepsilon).$$

Διάγραμμα 3.5 διάγραμμα $\{\hat{e}, \mathbf{b}_{ols}^T \mathbf{x}\}$ για το πρόβλημα με τα οστρακοειδή. 27

Διάγραμμα 4.1 διάγραμμα $\{y, \hat{\boldsymbol{\beta}}^T \mathbf{x}\}$ για το πρόβλημα της χημικής αντίδρασης ($\hat{\boldsymbol{\beta}}$ η κατεύθυνση OLS). 63

Διάγραμμα 4.2 διάγραμμα $\{y, \hat{\mathbf{u}}_1^T \mathbf{x}\}$ για το πρόβλημα της χημικής αντίδρασης ($\hat{\mathbf{u}}_1$ η πρώτη κατεύθυνση που εκτιμά η pHd). 65

Διάγραμμα 5.1 2D προβολή ελάχιστης διακύμανσης του 3D διαγράμματος για τις μεταβλητές H, Vol, D . 91

Διάγραμμα 5.2 2D προβολή (V, h_{unc}) του 3D διαγράμματος για τις μεταβλητές $Vol, -1.913 + 0.007294H + 0.1017D$, $, h_{unc}$ στη ζώνη 4 τιμών του y . 91



Διάγραμμα 5.3 2D προβολή (V, h_{unc}) του 3D διαγράμματος για τις μεταβλητές $Vol, -1.913 + 0.007294H + 0.1017D,$, h_{unc} στη ζώνη 10 τιμών του $y.$	92
Διάγραμμα 5.4 2D προβολή του 3D-AVP για τις μεταβλητές $M, gr2, gr3$ μετά την αφαίρεση της επίδρασης των $Fit, gr1.$	93
Διάγραμμα 5.5 2D προβολή ελάχιστης διακύμανσης του 3D διαγράμματος για τις μεταβλητές $M, Fit, gr1.$	94
Διάγραμμα 5.6 2D προβολή του 3D-AVP για τις μεταβλητές $y, gr3, gr4$ μετά την αφαίρεση της επίδρασης των $Fit, gr1, gr2.$	96
Διάγραμμα 5.7 2D προβολή ελάχιστης διακύμανσης του 3D-AVP για τις μεταβλητές $y, gr1, gr2$ μετά την αφαίρεση της επίδρασης της $Fit.$	96
Διάγραμμα 5.8 2D προβολή $\{y, Fit\}$ του 3D διαγράμματος για τις μεταβλητές $y, Fit, gr5.$	97
Διάγραμμα 5.9 2D προβολή $\{y, gr5\}$ του 3D διαγράμματος για τις μεταβλητές $y, Fit, gr5.$	98
Διάγραμμα 5.10 2D προβολή του 3D-AVP για τις μεταβλητές $y, gr3, gr4$ μετά την αφαίρεση της επίδρασης των $Fit, gr1, gr2$ για το μοντέλο 5.5.	99
Διάγραμμα 5.11 2D προβολή του 3D-AVP για τις μεταβλητές $y, gr3, gr4$ μετά την αφαίρεση της επίδρασης των $gr1, gr2$ για τα κατάλοιπα του μοντέλου $y x = \mathbf{n}_2^T \mathbf{u} + e.$	100



ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Η ανάλυση παλινδρόμησης είναι ως γνωστόν ένας δημοφιλής τρόπος για την μελέτη της δεσμευμένης κατανομής $y|x$ ο οποίος αποβλέπει στη σχέση ανάμεσα στην απόκριση y και το διάνυσμα των ανεξαρτήτων μεταβλητών $x = (x_1, \dots, x_p)^T$. Πρός το σκοπό αυτό χρησιμοποιούνται παραμετρικές και μη παραμετρικές τεχνικές. Οι παραμετρικές τεχνικές συνίστανται στην προσαρμογή κάποιου μοντέλου το οποίο υπαγορεύεται από τη φύση του προβλήματος η οποία συχνά άπτεται της σχέσης ανάμεσα στη συνάρτηση του μέσου $E(y|x)$ και στη συνάρτηση της διακύμανσης. Η προσαρμογή συνίσταται στην εκτίμηση των παραμέτρων του μοντέλου και στο σχετικό διαγνωστικό έλεγχο. Η πιο απλή και η πιο γνωστή περίπτωση μοντέλου είναι το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης (*multiple linear regression model*). Οι μη παραμετρικές τεχνικές χρησιμοποιούνται όταν η φύση του προβλήματος δεν είναι σε θέση να υπαγορεύει κάποιο μοντέλο, και αξιοποιούν καθεαυτά τα δεδομένα του προβλήματος. Πρόκειται για τεχνικές εξομάλυνσης (*smoothing*) που αναζητούν το μέσο $E(y|x)$ σε κάθε μία μεταξύ πολλών ζωνών (*slices*) για τις τιμές της απόκρισης y . Η επιτυχία των μη παραμετρικών αυτών μεθόδων εξομάλυνσης (*smoothing*) βασίζεται στην ύπαρξη επαρκούς ποσότητας δεδομένων η οποία αυξανομένης της διάστασης του x οφείλει να αυξάνει εκθετικά. Το φαινόμενο αυτό το οποίο είναι γνωστό και ως «κατάρα της διάστασης» (*curse of dimensionality*) είναι η αιτία της αποτυχίας των μη παραμετρικών μεθόδων όταν τα δεδομένα δεν είναι επαρκή, αποτέλεσε δε το έναυσμα για την ανάπτυξη των μεθόδων που αναπτύσσονται στα επόμενα κεφάλαια. Στόχος των μεθόδων αυτών είναι η εύρεση ενός υπόχωρου ελάχιστης διάστασης k στον οποίο η προβολή των σημείων (y, x) να παρέχει όλη την πληροφορία που είναι διαθέσιμη για την y μέσω του x κατά τρόπο δηλαδή ώστε η y να είναι της μορφής

$$y = f(\beta_1^T x, \beta_2^T x, \dots, \beta_k^T x)$$

όπου β_j διάνυσμα $px1, j= 1, 2, \dots, k$

χωρίς καμία προϋπόθεση ή περιορισμό για τη μορφή της f ή για τον τρόπο με τον οποίο η προβολή του x επιδρά στην y .

Ο υπόχωρος αυτός καλείται κεντρικός υπόχωρος (*central subspace*) συμβολίζεται δε $S_{y|x}$ και το αποτέλεσμα από την εφαρμογή των σχετικών



μεθόδων εκτίμησης των $\beta_1, \beta_2, \dots, \beta_k$ καλείται μείωση της διάστασης (*dimension reduction*), ενώ τα διανύσματα $\beta_1, \beta_2, \dots, \beta_k$ που αποτελούν τη βάση του κεντρικού υπόχωρου καλούνται αποτελεσματικές κατευθύνσεις μείωσης της διάστασης (*effective dimension reduction (e.d.r.) directions*). Πρόκειται για μεθόδους οι οποίες προετοιμάζουν τα δεδομένα ώστε αφενός να διαφανεί ευκρινέστερα η ενδεικνυόμενη συνέχεια και αφετέρου να είναι αποτελεσματική.

Η εν λόγω συνέχεια δύναται να είναι εφαρμογή παραμετρικής μοντελοποίησης, εκτίμηση της επιφάνειας απόκρισης, ανάλυση κατά συστάδες (*cluster analysis*), κ.λ.π.

Είναι σαφής η διαφοροποίηση των μεθόδων εκτίμησης του κεντρικού υπόχωρου από άλλες μεθόδους οι οποίες στοχεύουν στην προσέγγιση της συνάρτησης που συνδέει τα y και x . Τέτοιες μέθοδοι είναι οι *Projection pursuit regression* (Friedman και Stuetzle (1981)), *ACE* και *additive models* (Breiman & Friedman (1985), Stone (1986), Hastie και Tibshirani (1986)) και *partial splines* (Chen (1988), Cuzik (1987), Heckman (1986), Speckman (1987) κ.α.) οι οποίες στοχεύουν στην προσέγγιση του $E(y|x)$.

Στα επόμενα κεφάλαια παρουσιάζονται οι σημαντικότερες από τις μεθόδους εκτίμησης του κεντρικού υπόχωρου δηλαδή της βάσης και της διάστασής του, με σκοπό τη μείωση της διάστασης σε προβλήματα παλινδρόμησης της y στο διάνυσμα x . Οι μέθοδοι αυτές βασίζονται στην ιδέα ότι αν \hat{K} είναι μία συνεπής εκτίμηση της μήτρας K για την οποία ισχύει $S(K) \subseteq S_{y|x}$ τότε ο υπόχωρος με βάση τα αριστερά ιδιάζοντα διανύσματα (*left singular vectors*) που αντιστοιχούν στις μη μηδενικές ιδιάζουσες τιμές (*singular values*) της \hat{K} θα αποτελεί εκτίμηση του $S(K)$ (Cook και Yin (2001)). Ειδικότερα στο κεφάλαιο 2 παρουσιάζονται κάποιες βασικές έννοιες για την κατανόηση της σκοπιμότητας των εν λόγω μεθόδων, των προϋποθέσεων εφαρμογής τους και της πολύ σημαντικής διασύνδεσης του $S_{y|x}$ με τη μέθοδο *OLS*.

Στο Κεφάλαιο 3 παρουσιάζεται η μέθοδος *SIR* (*Sliced inverse regression*) (Li 1991)η σημαντικότερη μέθοδος η οποία βασίζεται στην ροπή πρώτης τάξης $E(x|y)$ της $x|y$. Η μέθοδος εκμεταλλεύεται τη σχέση ανάμεσα στις κατευθύνσεις στις οποίες μεγιστοποιείται η $Cov[E(z|y)]$ και τον κεντρικό υπόχωρο $S_{y|z}$, όπου z το τυποποιημένο διάνυσμα x . Οι κατευθύνσεις αυτές

προκύπτουν από την ιδιοδιανυσματική ανάλυση της $Cov[E(z|y)]$. Παρουσιάζεται επίσης η μέθοδος *PIR* (*Parametric inverse regression*) (Bura και Cook 2001b) η οποία βασίζεται στην προσαρμογή παραμετρικών καμπυλών στα δεδομένα των p αντίστροφων παλινδρομήσεων $x_j|y$, η οποία γίνεται μέσω του μοντέλου της πολλαπλής γραμμικής παλινδρόμησης. Η εκτίμηση της βάσης του κεντρικού υπόχωρου S_{yz} προκύπτει από το γινόμενο των ιδιοδιανυσμάτων της μήτρας των συντελεστών της παλινδρόμησης με γραμμικές συναρτήσεις του y .

Στο κεφάλαιο 4 παρουσιάζονται οι μέθοδοι *SAVE* (*Sliced average variance estimate*) (Cook και Weisberg (1991)) και *SIR II* (Li (1991b)) οι οποίες βασίζονται στη ροπή δεύτερης τάξης $Var(x|y)$ της $x|y$. Η *SIR II* εκμεταλλεύεται τη σχέση ανάμεσα στις κατευθύνσεις στις οποίες μεγιστοποιείται η $cov(z|y) - E[cov(z|y)]$ και τον κεντρικό υπόχωρο S_{yz} . Οι κατευθύνσεις αυτές προκύπτουν από την ιδιοδιανυσματική ανάλυση της $E(cov(z|y)-E[cov(z|y)])^2$. Η *SAVE* είναι μία ειδική περίπτωση της *SIR II* η οποία εκτιμά τη βάση του κεντρικού υπόχωρου S_{yz} με βάση την ιδιοδιανυσματική ανάλυση της $E[cov(z|y)-I]^2$. Οι δύο αυτές μέθοδοι χρησιμοποιούνται για την ανίχνευση κατευθύνσεων τις οποίες αδυνατεί να εκτιμήσει η *SIR*.

Παρουσιάζεται επίσης η μέθοδος *PHD* (*Principal Hessian Directions*) (Li (1992)) η οποία εκμεταλλεύεται τη σχέση ανάμεσα στις κατευθύνσεις στις οποίες μεγιστοποιείται η Εσιανή μήτρα της $E(y|x)$ και τον κεντρικό υπόχωρο S_{yx} . Χρησιμοποιώντας το λήμμα του Stein (1981, Λήμμα 4) η μέθοδος παρέχει τις κατευθύνσεις αυτές με βάση την ιδιοδιανυσματική ανάλυση της

$$\Sigma_{yxx} = E(y - \mu_y)(x - \mu_x)(x - \mu_x)^T.$$

Στο κεφάλαιο 5 παρουσιάζονται οι μέθοδοι *SAT* (*Sliced average third-moment estimation*) (Yin και Cook (2003)), *COV_k* (Yin και Cook (2002)) και η μέθοδος της γραφικής παλινδρόμησης (*Graphical Regression*) (Cook και Weisberg (1991)). Η μέθοδος *SAT* βασίζεται στη ροπή τρίτης τάξης της $z|y$ και στοχεύει στην ανίχνευση *edr* κατευθύνσεων τις οποίες αδυνατούν να ανιχνεύσουν οι αντίστροφες ροπές 1^{ης} και 2^{ης} τάξης. Έχει ευρεία εφαρμογή σε προβλήματα λογιστικής παλινδρόμησης και γενικότερα σε προβλήματα κατηγορικής απόκρισης y . Ειδικότερα αναζητά κατευθύνσεις στις οποίες

μεγιστοποιείται η διαφορά στην ασυμμετρία των $z|y$ για τις διάφορες τιμές του y . Οι κατευθύνσεις αυτές προκύπτουν από την ιδιοδιανυσματική ανάλυση της $E(\mathbf{M}_y \mathbf{M}_y^T)$ όπου \mathbf{M}_y η μήτρα των διαφορετικών στηλών της $M^{(3)}(z|y)^T$ όπου

$$M^{(3)}(z|y)^T = E_{z|y} [\{z-E(z|y)\} \otimes \{z-E(z|y)\} \{z-E(z|y)\}^T].$$

Η μέθοδος COV_k αποσκοπεί στην εκτίμηση ενός υπόχωρου ελάχιστης διάστασης στον οποίο η προβολή των σημείων (y, x) να περιέχει όλη την πληροφορία που είναι διαθέσιμη για την y μέσω των ροπών της $y|x$ από την 1^η έως και την k -οστή δηλαδή μέσω των $E(y|x), Var(y|x), \dots, M^{(k)}(y|x)$. Ο υπόχωρος αυτός καλείται κεντρικός υπόχωρος k -τάξεως (*Central k-th Moment Subspace*) και συνδέεται με τον γνωστό κεντρικό υπόχωρο $S_{y|x}$. Η εκτίμηση της βάσης του $CKMS$ προκύπτει από την ιδιοδιανυσματική ανάλυση της μήτρας $K = (E(yz), \dots, E(y^k z))$. Τέλος παρουσιάζεται η μέθοδος της γραφικής παλινδρόμησης η οποία αποτελεί το επόμενο βήμα στην εκτίμηση του κεντρικού υπόχωρου και η οποία χρησιμοποιείται σε συνδυασμό με κάποια από τις άλλες μεθόδους.

Να σημειωθεί ότι οι δύο μορφές του κεντρικού υπόχωρου $S_{y|x}$ και $S_{y|z}$ συνδέονται μέσω του μετασχηματισμού

$$S_{y|x} = \Sigma_x^{-1/2} S_{y|z}$$

όπου

$$\Sigma_x = Var(x).$$

Σε κάθε κεφάλαιο αναπτύσσεται η θεωρητική θεμελίωση των οικείων μεθόδων, καθώς και ο σχετικός αλγόριθμος για την εφαρμογή της μεθόδου. Ακολούθως παρουσιάζονται εφαρμογές με στόχο την παρουσίαση της λειτουργίας της μεθόδου και τη συγκριτική ανάλυση σε σχέση με άλλες μεθόδους, ενώ στο τέλος του κεφαλαίου παρατίθενται σχόλια, επισημάνσεις και συμπεράσματα.

Πρέπει να τονιστεί ότι η ανά χείρας επισκόπηση περιορίστηκε με εξαίρεση τη μέθοδο SAT σε περιπτώσεις συνεχών μεταβλητών y και x . Για την περίπτωση κατά την οποία η y είναι κατηγορική μεταβλητή ο αναγνώστης μπορεί να ανατρέξει στους Cook και Yin (2001), ενώ για την περίπτωση κατηγορικού x στους Chiaromonte, Cook και Li (2002). Ιδιαίτερη αναφορά πρέπει τέλος να γίνει στο πρόγραμμα Arc μέσω του οποίου μπορεί να γίνει εφαρμογή των

μεθόδων που παρουσιάζονται και το οποίο είναι διαθέσιμο στο site <http://www.stat.umn.edu/arc>. Η χρήση του προγράμματος υπό τη μορφή manual περιγράφεται από τους Cook και Weisberg (1999). Να σημειωθεί απλώς ότι το πρόγραμμα στη default μορφή του είναι σε θέση να εφαρμόσει τις μεθόδους *SIR*, *SAVE*, *PHD* και βέβαια τη γραφική την οποία εφαρμόζει σε συνδυασμό με κάποια από αυτές τις τρεις. Υπάρχει ωστόσο η δυνατότητα επέκτασης για τη χρήση και των υπόλοιπων μεθόδων.



ΚΕΦΑΛΑΙΟ 2

ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Στο κεφάλαιο αυτό αναπτύσσονται κάποιες έννοιες βασικές για την κατανόηση της σκοπιμότητας της εκτίμησης του κεντρικού υπόχωρου, των προϋποθέσεων εφαρμογής των σχετικών μεθόδων που αναλύονται στα επόμενα κεφάλαια και της χρησιμότητας της διασύνδεσης του κεντρικού υπόχωρου με τη γνωστή μέθοδο *OLS*. Παρουσιάζονται έτσι η έννοια της μείωσης της διάστασης (*dimension reduction*), η έννοια του κεντρικού υπόχωρου (*central subspace*), η σημασία της μοναδιαίας διάστασης του κεντρικού υπόχωρου μέσω του *1D estimation result* και η ισχύς των συνθηκών ελέγχου του μέσου και της διακύμανσης (*mean* και *variance checking conditions*) καθώς και η δυνατότητα εξασφάλισης της ισχύος.

2.1 ΚΕΝΤΡΙΚΟΣ ΥΠΟΧΩΡΟΣ (*CENTRAL DIMENSION REDUCTION SUBSPACE*)

Έστω \mathbf{B} μια $p \times q$ μήτρα βαθμού $q \leq p$ τέτοια ώστε

$$y \perp\!\!\!\perp x | \mathbf{B}^T x$$

όπου $x = (x_1, \dots, x_p)^T$ το διάνυσμα των ανεξάρτητων μεταβλητών οι οποίες εξηγούν τη μεταβλητότητα της απόκρισης y .

Στην περίπτωση αυτή οι δύο δεσμευμένες κατανομές $y|x$ και $y|\mathbf{B}^T x$ ταυτίζονται και η $q \times 1$ μήτρα $\mathbf{B}^T x$ θα περιέχει όλη την πληροφορία που είναι διαθέσιμη για το y μέσω του x .

Αυτός ο περιορισμός των μεταβλητών που χαρακτηρίζουν την $y|x$ ώστε η πληροφορία των p συνιστωσών x_1, \dots, x_p του x να παρέχεται επαρκώς από τους q γραμμικούς συνδυασμούς

$$\beta_1^T x, \beta_2^T x, \dots, \beta_q^T x$$

όπου $\beta_1, \beta_2, \dots, \beta_q$ τα διανύσματα που αποτελούν τις στήλες της μήτρας \mathbf{B} καλείται μείωση της διάστασης (*dimension reduction*). Ο δε υπόχωρος $S(\mathbf{B})$ που έχει σαν βάση τις στήλες αυτές καλείται υπόχωρος μείωσης της διάστασης (*dimension reduction subspace* ή *DRS*) για την $y|x$ ή ισοδύναμα για



την παλινδρόμηση του y στο x , συμβολίζεται δε ως S_{drs} και το διάγραμμα που προκύπτει από την προβολή των σημείων (y, x) στον υπόχωρο S_{drs} καλείται επαρκές διάγραμμα (*sufficient summary plot*) (Cook (1998)).

Βέβαια (Cook (1998)) ο υπόχωρος $S(B)$ δεν είναι ο μοναδικός *DRS*. Ωστόσο αν η τομή $\cap S_{drs}$ όλων των *DRS* είναι επίσης *DRS* τότε αυτός ο *DRS* θα έχει την ελάχιστη δυνατή διάσταση, θα είναι μοναδικός, καλείται δε κεντρικός υπόχωρος μείωσης της διάστασης (*Central dimension reduction subspace* ή *CDRS*) ή απλά κεντρικός υπόχωρος (*central subspace*) για την $y|x$ ή ισοδύναμα για την παλινδρόμηση του y στο x και συμβολίζεται $S_{y|x}$ (Cook (1999)). Αποδεικνύεται εξάλλου (Cook (1998)) ότι αν $S_{y|x}(n)$ είναι ο κεντρικός υπόχωρος για την $y|x$, τότε ο $S_{y|z}(A^{-1}n)$ θα είναι ο κεντρικός υπόχωρος για την $y|z$ όπου $z = A^T x$ και A $p \times p$ μήτρα πλήρους βαθμού (*full rank*).

Καθόλη την έκταση της παρούσας ανασκόπησης των μεθόδων εκτίμησης του εν λόγω κεντρικού υπόχωρου για την $y|x$ θεωρείται ότι ο κεντρικός υπόχωρος υπάρχει, θεωρείται δηλαδή ότι συντρέχουν οι προϋποθέσεις ισχύος των παρακάτω προτάσεων που παρουσιάζονται στον Cook (1998).

Πρόταση 2.1 Έστω $S(a)$ και $S(\phi)$ δύο *DRS* για την $y|x$. Αν το διάνυσμα x κατανέμεται με συνάρτηση πυκνότητας πιθανότητας $f(x) > 0$ $x \in \Omega_x \subset \mathbb{R}^p$ και $\mu(x) = 0$ για άλλες τιμές του x , όπου Ω_x ο δειγματικός χώρος των τιμών του x , και αν ο Ω_x είναι σύνολο κυρτό τότε η τομή $S(a) \cap S(\phi)$ θα είναι *DRS*.

Για την περίπτωση διακριτών ανεξαρτήτων μεταβλητών βλ. Cook (1998), Κ.6, πρόβλημα 6.5. Η πρόταση που ακολουθεί ισχύει μέσα στο πλαίσιο επαρκούς πληροφόρησης για την $y|x$ μέσω της $E(y|x)$. Ισχύει δηλαδή

$$y \perp\!\!\!\perp x | E(y|x).$$

Πρόταση 2.2 Έστω $S(a)$ και $S(\phi)$ δύο *DRS* για την $y|x$ η οποία είναι τέτοια ώστε $y \perp\!\!\!\perp x | E(y|x)$. Αν το διάνυσμα x κατανέμεται με συνάρτηση πυκνότητας πιθανότητας $f(x) > 0$ για $x \in \Omega_x \subset \mathbb{R}^p$ και η $E(y|x)$



μπορεί να εκφραστεί με τη μορφή συγκλίνουσας δυναμοσειράς γύρω από τις συντεταγμένες του $x = (x_1, \dots, x_p)^T$, δηλαδή

$$E(y|x) = \sum_{k_1, \dots, k_p}^{\infty} \alpha_{k_1, \dots, k_p} x_1^{k_1} \cdots x_p^{k_p}$$

τότε η τομή $S(\alpha) \cap S(\phi)$ θα είναι DRS.

Η παραπάνω πρόταση μπορεί πάντως να επεκταθεί και στην περίπτωση

$$y \perp\!\!\!\perp x | [E(y|x), Var(y|x)].$$

2.2 ΔΙΑΣΥΝΔΕΣΗ ΤΗΣ ΜΕΘΟΔΟΥ OLS ΚΑΙ ΤΟΥ ΚΕΝΤΡΙΚΟΥ ΥΠΟΧΩΡΟΥ $S_{y|x}$.

Αν δεχτούμε ότι η $y|x$ εκφράζεται από το μοντέλο

$$y = \alpha + b^T x + \varepsilon$$

όπου $\varepsilon \perp\!\!\!\perp x$, τότε οι τιμές των α, b για τις οποίες ελαχιστοποιείται μία συγκεκριμένη αντικειμενική συνάρτηση $L_n(\alpha, b)$ λαμβάνονται σαν εκτιμήσεις των α, b . Συνήθως η $L_n(\alpha, b)$ είναι της μορφής:

$$L_n(\alpha, b) = \frac{1}{n} \sum_{i=1}^n L(\alpha + b^T x_i, y_i)$$

όπου $L(u, v)$ κυρτή συνάρτηση του v .

Η πρόταση που ακολουθεί παρέχει διασύνδεση ανάμεσα στον OLS εκτιμητή \hat{b} του b και σε έναν οποιοδήποτε DRS με διάσταση 1 και είναι γνωστή ως πρόταση των Li και Duan (Cook (1998)).

Πρόταση 2.3 Έστω $S_{drs}(\gamma)$ ένας υπόχωρος μείωσης των διάστασης για την $y|x$ με $\dim S_{drs}(\gamma) = 1$, έστω $E(x|\gamma^T x)$ γραμμική συνάρτηση της $\gamma^T x$, και $\Sigma = Var(x)$ θετική ορισμένη. Εάν το διάνυσμα β για το οποίο ισχύει

$$(\tilde{\alpha}, \tilde{\beta}) = \arg \min_{a, b} E[L(a + b^T x, y)]$$

όπου $L(a + b^T x, y)$ κυρτή συνάρτηση του $a + b^T x$, είναι μοναδικό, τότε $\beta \in S_{drs}(\gamma)$.

Να σημειωθεί ότι $E[L(\alpha + \mathbf{b}^T \mathbf{x}, y)]$ είναι το όριο στο οποίο συγκλίνει σύμφωνα με το νόμο των μεγάλων αριθμών η ποσότητα

$$L_n(\alpha, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^n L(\alpha + \mathbf{b}^T \mathbf{x}_i, y_i)$$

Ο δειγματικός OLS εκτιμητής $\hat{\mathbf{b}}$ του \mathbf{b} για τον οποίο ισχύει

$$(\hat{\alpha}, \hat{\mathbf{b}}) = \arg \min_{\alpha, \mathbf{b}} L_n(\alpha, \mathbf{b})$$

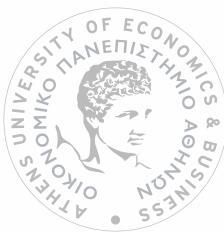
αποτελεί ταυτόχρονα και εκτιμητή του β .

Η παραπάνω πρόταση είναι γνωστή και ως *ID estimation result* (Cook και Weisberg (1999)) και αφορά τη γενική περίπτωση κατά την οποία η

$L(\alpha + \mathbf{b}^T \mathbf{x}, y)$ είναι κυρτή συνάρτηση του $\alpha + \mathbf{b}^T \mathbf{x}$. Ειδική περίπτωση αυτής είναι η $L(\alpha + \mathbf{b}^T \mathbf{x}, y) = (y - \alpha - \mathbf{b}^T \mathbf{x})^2$ βάσει της οποίας προκύπτουν οι OLS εκτιμητές των α, \mathbf{b} . Είναι αυτονόητο ότι η παραπάνω πρόταση είναι χρήσιμη μόνον εφόσον υπάρχει ο κεντρικός υπόχωρος. Στην περίπτωση αυτή και σύμφωνα με την παραπάνω πρόταση εάν $\dim S_{y|x}=1$ τότε η OLS εκτίμηση $\hat{\mathbf{b}}$ του διανύσματος \mathbf{b} θα είναι της μορφής $\hat{\mathbf{b}} = cy$. Να επισημανθεί ότι η πρόταση 2.3 αποτελεί εξειδίκευση της πρότασης 8.1 που παραθέτει ο Cook (1998) και η οποία γενικεύει την πρόταση των Li και Duan (1989, Θεώρημα 2.1) για την περίπτωση $\dim S_{drs}(y) \geq 1$. Επίσης να τονιστεί ότι η ισχύς του μοντέλου $y = \alpha + \mathbf{b}^T \mathbf{x} + \varepsilon$ δεν ελέγχεται διαγνωστικά ούτε προσαρμόζεται κατ' ανάγκη επαρκώς στα δεδομένα. Σ' αυτό το γεγονός άλλωστε συνίσταται το εύλογο ενδιαφέρον των παραπάνω προτάσεων των Li και Duan και του Cook. Ωστόσο εάν το παραπάνω μοντέλο δεν ισχύει ως προς τη γραμματικότητα του μέσου $E(y|x)$ τότε στην πράξη οι προτάσεις αυτές ενδεχομένως να μην παρουσιάζουν χρησιμότητα αφού είναι δυνατόν $\hat{\mathbf{b}} = \mathbf{0}$. Στις περιπτώσεις αυτές για τις οποίες υποτίθεται ότι

$$y = \alpha + \mathbf{b}^T z + z^T \mathbf{C}_k z + \varepsilon$$

όπου $\varepsilon \perp\!\!\!\perp x, z = \Sigma_x^{-1/2}(x - E(x)), \Sigma_x = Var(x)$ και \mathbf{C}_k πραγματική συμμετρική μήτρα βαθμού $k \leq p$, η διασύνδεση με τον κεντρικό υπόχωρο των εκτιμητών που παρέχει η μέθοδος OLS ή οποιαδήποτε άλλη μέθοδος η οποία χρησιμοποιεί κάποια αντικειμενική συνάρτηση η οποία είναι κυρτή



συνάρτηση του $(\alpha + \mathbf{b}^T z + z^T \mathbf{C}_k z)$, βασίζεται στην ακόλουθη πρόταση (Cook (1998)).

Πρόταση 2.4 Έστω $S_{y|z}(y)$ ο κεντρικός υπόχωρος για την $y|z$, και έστω ότι η z ακολουθεί πολυμεταβλητή κανονική κατανομή με $E(z)=\mathbf{0}$ και μήτρα συνδιακύμανσης \mathbf{I} . Εάν το διάνυσμα β και η μήτρα Γ_k για τα οποία ισχύει

$$(\tilde{\alpha}, \hat{\beta}, \hat{\Gamma}_k) = \arg \min_{a, b, C_k} E[L(\alpha + \mathbf{b}^T z + z^T \mathbf{C}_k z, y)]$$

όπου $L(\alpha + \mathbf{b}^T z + z^T \mathbf{C}_k z, y)$ κυρτή συνάρτηση του $(\alpha + \mathbf{b}^T z + z^T \mathbf{C}_k z)$, είναι μοναδικά τότε

$$\beta \in S_{y|z}(y)$$

και

$$S(\Gamma_k) \subset S_{y|z}(y)$$

Να σημειωθεί ότι $E[L(\alpha + \mathbf{b}^T z + z^T \mathbf{C}_k z, y)]$ είναι το όριο στο οποίο συγκλίνει, σύμφωνα με τον νόμο των μεγάλων αριθμών, η ποσότητα

$$L_n(\alpha, \mathbf{b}, \mathbf{C}_k) = \frac{1}{n} \sum_{i=1}^n L(\alpha + \mathbf{b}^T \hat{z}_i + \hat{z}_i^T \mathbf{C}_k \hat{z}_i, y_i)$$

όπου $\hat{z}_i = \hat{\Sigma}_x^{-1/2} (\mathbf{x}_i - \bar{\mathbf{x}})$ οι δειγματικές τιμές των z_i .

Οι δειγματικοί OLS εκτιμητές $\hat{\beta}$, $\hat{\mathbf{C}}_k$ των \mathbf{b} και \mathbf{C}_k για τους οποίους ισχύει

$$(\hat{\alpha}, \hat{\beta}, \hat{\mathbf{C}}_k) = \arg \min_{a, b, C_k} L_n(a, b, C_k)$$

αποτελούν ταυτόχρονα και εκτιμητές των β και Γ_k .

Η ισχύς του μοντέλου $y = \alpha + \mathbf{b}^T z + z^T \mathbf{C}_k z + \varepsilon$ έναντι του $y = \alpha + \mathbf{b}^T z + \varepsilon$ μπορεί να ελεγχθεί διαγνωστικά. Δύο διαφορετικοί τύποι ελέγχου για την περίπτωση $\kappa=1$ παρουσιάζονται από τους Cook και Weisberg (1999 K.14 2).

2.3 ΠΡΟΫΠΟΘΕΣΕΙΣ ΕΦΑΡΜΟΓΗΣ ΤΩΝ ΜΕΘΟΔΩΝ ΕΚΤΙΜΗΣΗΣ ΤΟΥ ΚΕΝΤΡΙΚΟΥ ΥΠΟΧΩΡΟΥ

Πριν ασχοληθούμε με τις συνθήκες ελέγχου του ενδεχομένου $\dim S_{y|x}=1$ το οποίο αν ισχύει επιτρέπει την αξιοποίηση όσων ειπώθηκαν στην προηγούμενη παράγραφο ας δούμε δύο πολύ βασικές προϋποθέσεις εφαρμογής των περισσοτέρων από τις μεθόδους εκτίμησης του κεντρικού υπόχωρου οι οποίες αναλύονται στα επόμενα κεφάλαια.



2.3.1 ΥΠΟΘΕΣΗ ΓΡΑΜΜΙΚΟΤΗΤΑΣ

Η υπόθεση γραμμικότητας ή γραμμικής συσχέτισης των ανεξάρτητων μεταβλητών εκφράζεται ως εξής

$$E(x_j | B^T x) = \alpha_j + b_j^T x (B^T x), \quad j=1,\dots,p$$

όπου $x=(x_1, \dots, x_p)^T$ και B η $p \times q$ μήτρα της οποίας οι q στήλες αποτελούν βάση του κεντρικού υπόχωρου $S_{y|x}$.

Ας δούμε τι σημαίνει η παραπάνω υπόθεση. Όταν η υπόθεση αυτή ισχύει τότε η γραμμή παλινδρόμησης που προσαρμόζεται στην προβολή των δεδομένων (y, x) σε κάθε επίπεδο το οποίο σχηματίζεται από κάθε ένα από τα διανύσματα βάσης $(1, 0, \dots, 0), (0, 1, \dots, 0), \dots, (0, 0, \dots, 1)$ του \mathbb{R}^p και κάθε ένα από τα διανύσματα βάσης του $S_{y|x}$, ή ισοδύναμα σε κάθε επίπεδο $B_j^T x$ όπου $j=1,\dots,q$ και B_j η στήλη j της μήτρας B , θα είναι ευθεία.

Να επισημανθεί ότι η υπόθεση αυτή θέτει έναν περιορισμό στην περιθώρια κατανομή του x και όχι στην δεσμευμένη κατανομή της $y|x$ όπως είναι η συνήθης πρακτική κατά την προσαρμογή ενός συγκεκριμένου μοντέλου στα δεδομένα.

Να επισημανθεί επίσης ότι αν και η υπόθεση γραμμικότητας πρέπει να ισχύει για τη βάση του κεντρικού υπόχωρου στην πράξη απαιτείται η ισχύς της για κάθε δυνατή μήτρα B , συνθήκη που ισοδυναμεί με ελλειπτική συμμετρία του διανύσματος x (Eaton(1986)). Για την επαγωγή ελλειπτικής συμμετρίας οι Cook και Nachtheim (1994) πρότειναν μία μέθοδο στάθμισης του διανύσματος x , ενώ στην πράξη πολύ καλά αποτελέσματα επιτυγχάνονται από τον ταυτόχρονο μετασχηματισμό των ανεξάρτητων μεταβλητών με σκοπό την επαγωγή πολυμεταβλητής κανονικότητας στο διάνυσμα x σύμφωνα με μία μέθοδο που πρότεινε ο Velilla (1993) και η οποία αποτελεί γενίκευση της μεθόδου Box-Cox για την περίπτωση μίας ανεξάρτητης μεταβλητής (Cook και Weisberg 1999 κ.5, 2). Σύμφωνα με τη μέθοδο αυτή του Velilla αναζητούνται μετασχηματισμοί δυνάμεων λ ώστε το διάνυσμα

$$T(x) = (x_j^{(\lambda_j)}) , \quad j=1,\dots,p$$

όπου

$$x_j^{(\lambda_j)} = (x_j^{(\lambda_j)} - 1)/\lambda, \text{ για } \lambda \neq 0$$



και

$$x_j^{(\lambda)} = \log x_j, \text{ για } \lambda = 0$$

να ακολουθεί κατά το δυνατόν κανονική κατανομή.

Στην πράξη η ισχύς της υπόθεσης γραμμικότητας ελέγχεται μέσω της γνωστής *scatterplot matrix* η οποία παρέχει όλα τα διδιάστατα διαγράμματα τα οποία απεικονίζουν τις περιθώριες σχέσεις κάθε πιθανού ζεύγους μεταβλητών. Εάν όλα τα διαγράμματα που αφορούν συνδυασμούς των μεταβλητών $x_j, j=1,\dots,p$ εμφανίζουν γραμμικό μέσο τότε μπορεί να θεωρηθεί ότι ισχύει η υπόθεση της γραμμικότητας. Εάν όχι τότε και με τη βοήθεια του προγράμματος Arc για το οποίο θα γίνει λόγος στη συνέχεια γίνεται χρήση της παραπάνω μεθόδου επαγωγής κανονικότητας. Παράδειγμα χρήσης της *scatterplot matrix* και μετασχηματισμού ώστε να ισχύει η υπόθεση της γραμμικότητας παρατίθεται στην επόμενη παράγραφο όπου φαίνεται ευκρινέστερα η σημασία της ισχύος της υπόθεσης αυτής.

Τέλος, να επισημανθεί ότι η παραπάνω υπόθεση δεν αποτελεί ιδιαίτερα αυστηρό περιορισμό δεδομένου ότι όπως απέδειξαν οι Diaconis και Freedman (1984), η προβολή ενός νέφους σημείων (y, x) , όπου $x=(x_1, \dots, x_p)^T$ τυχαίο διάνυσμα πολλών διαστάσεων, σε έναν υπόχωρο λίγων διαστάσεων κατανέμεται σχεδόν κανονικά οπότε η υπόθεση της γραμμικότητας ισχύει. Ούτε αποτελεί ιδιαιτέρως κρίσιμη υπόθεση εκτός αν παραβιάζεται σοβαρά γεγονός που καταδεικνύεται στα επόμενα επιμέρους κεφάλαια.

2.3.2 ΥΠΟΘΕΣΗ ΣΤΑΘΕΡΗΣ ΔΙΑΚΥΜΑΝΣΗΣ

Η υπόθεση αυτή εκφράζεται ως εξής

$$\text{Cov}(x_j x_k | \mathbf{B}^T x) = \sigma_{jk}, \quad j, k = 1, \dots, p$$

όπου $x=(x_1, \dots, x_p)^T$ και \mathbf{B} η $p \times q$ μήτρα της οποίας η q στήλης αποτελούν βάση του κεντρικού υπόχωρου $S_{y|x}$.

Η σημασία της υπόθεσης είναι ότι όταν αυτή ισχύει τότε η προβολή των δεδομένων (y, x) σε κάθε επίπεδο $\mathbf{B}_j^T x$, όπου $j=1, \dots, q$ και \mathbf{B}_j η στήλη j της μήτρας \mathbf{B} , θα παρουσιάζει σταθερή διακύμανση.

Η επαγωγή πολυμεταβλητής κανονικότητας με τη χρήση της μεθόδου που αναφέρθηκε στην προηγούμενη υποπαράγραφο με σκοπό την ενίσχυση της υπόθεσης γραμμικότητας εξασφαλίζει και την ισχύ της υπόθεσης σταθερής διακύμανσης.



Στην πράξη η ισχύς και αυτής της υπόθεσης ελέγχεται μέσω της *scatterplot matrix*. Εάν όλα τα διαγράμματα που αφορούν συνδυασμούς των μεταβλητών $x_j, j=1,...,p$ εμφανίζουν σταθερή διακύμανση τότε μπορεί να θεωρηθεί ότι η υπόθεση της σταθερής διακύμανσης ισχύει. Ούτε αυτή η υπόθεση είναι ιδιαιτέρως κρίσιμη εκτός αν παραβιάζεται σοβαρά.

2.4 ΣΥΝΘΗΚΕΣ ΕΛΕΓΧΟΥ ΤΟΥ ΜΕΣΟΥ ΚΑΙ ΤΗΣ ΔΙΑΚΥΜΑΝΣΗΣ (*MEAN AND VARIANCE CHECKING CONDITIONS*)

Οι συνθήκες που ακολουθούν αφορούν τις περιθώριες υπό συνθήκη κατανομές $x_j|y$ όπου x_j η j ανεξάρτητη μεταβλητή του τυχαίου διανύσματος $x=(x_1,...,x_p)^T$ των ανεξάρτητων μεταβλητών. Πρόκειται δηλαδή για την αντίστροφη παλινδρόμηση του x στο y δηλαδή την $x|y$. Οι ιδιότητες της $x|y$ και η σημασία της στη δυνατότητα εκτίμησης του κεντρικού υπόχωρου $S_{y|x}$ θα μας απασχολήσουν στο επόμενο κεφάλαιο. Προς το παρόν θα μας απασχολήσουν οι $E(x_j|y)$, $Var(x_j|y)$ όταν $dimS_{y|x}=1$. Συγκεκριμένα όταν ισχύει η υπόθεση της γραμμικότητας της παραγράφου 2.3.1 και επίσης $dimS_{y|x}=1$ τότε

$$E(x_j|y) = E(x_j) + a_j m(y)$$

$$Var(x_j|y) \approx Var(x_j) - a_j^2 s(y)$$

όπου a_j σταθερά η οποία έχει την ίδια τιμή και στις δύο παραπάνω εξισώσεις, και $m(y)$, $s(y)$ συναρτήσεις του y .

Οι παραπάνω εξισώσεις είναι γνωστές ως συνθήκες ελέγχου του μέσου και της διακύμανσης (*mean* και *variance checking conditions*) (Cook και Weisberg (1999), Cook (1998)).

Σύμφωνα με τη συνθήκη ελέγχου του μέσου $E(x_j|y)$ θα έχει την ίδια μορφή για όλα τα x_j . Επομένως ή θα είναι γραμμική σε όλα τα $2D$ διαγράμματα $x_j|y$ ή θα έχει την ίδια κοινή μορφή καμπυλότητας σε όλα τα διαγράμματα αυτά.

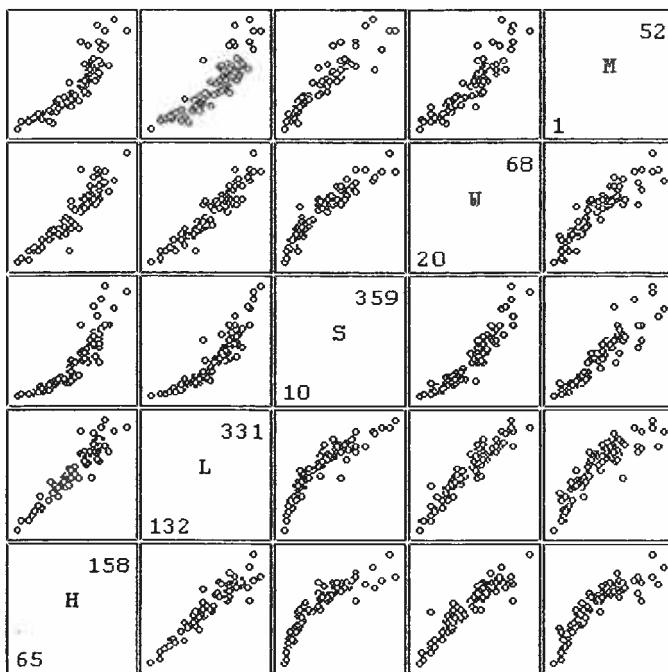
Σύμφωνα δε με τη συνθήκη ελέγχου της διακύμανσης και η $Var(x_j|y)$ θα έχει την ίδια μορφή σε όλα τα $2D$ διαγράμματα $x_j|y$.

Στο σημείο αυτό χρειάζεται προσοχή διότι οι παραπάνω συνθήκες είναι αναγκαίες δεν είναι όμως ικανές. Εάν δηλαδή δεν ισχύουν, εάν δηλαδή τα $2D$ διαγράμματα $x_j|y$ δεν παρουσιάζουν κοινή μορφή για την $E(x_j|y)$ ή την $Var(x_j|y)$, τότε δεν ισχύει $dimS_{y|x}=1$. Εάν όμως ισχύουν αυτό δεν σημαίνει απαραίτητα ότι $dimS_{y|x}=1$. Για την

απάντηση στο ερώτημα αν όντως $\dim S_{y|x} = 1$ σε μια τέτοια περίπτωση, μπορεί να γίνει χρήση του *1D estimation result*. Εάν δηλαδή η μεταβλητή $b_{ols}^T x$, όπου b_{ols} το διάνυσμα των συντελεστών της μεθόδου *OLS*, παρέχει επαρκώς την πληροφορία που είναι διαθέσιμη για την y μέσω του x τότε $\dim S_{y|x} = 1$. Εάν όχι τότε $\dim S_{y|x} \neq 1$. Η εν λόγω επάρκεια ελέγχεται με βάση όσα παρατίθενται στην παράγραφο 5.3.

Ας δούμε ένα παράδειγμα χρήσης της *scatterplot matrix*. Το παράδειγμα αφορά δεδομένα για οστρακοειδή τα οποία απετέλεσαν δείγμα το οποίο ελήφθη από την περιοχή Marlborough Sounds ανοιχτά των ακτών της Νέας Ζηλανδίας. Τα δεδομένα συνελέγησαν μέσα στα πλαίσια οικολογικής μελέτης για τα οστρακοειδή και με σκοπό τη μελέτη της εξάρτησης της μάζας M του πυρήνα από το μήκος L , το πλάτος W , το ύψος H και τη μάζα S του κέλυφουνς του οστρακοειδούς. Ακριβής περιγραφή των μεταβλητών υπάρχει στους (Cook και Weisberg (1999) πρόβλημα 14.3) ενώ τα δεδομένα διατίθενται στο αρχείο mussels.lsp το οποίο περιέχεται στο Arc.

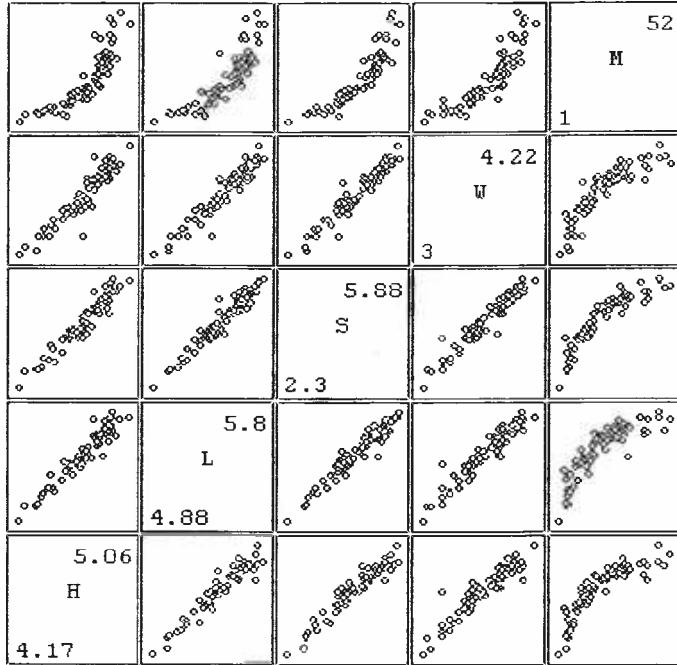
Στο διάγραμμα 2.1 που ακολουθεί παρατίθεται η *scatterplot matrix* όπου φαίνεται να υπάρχει απόκλιση τουλάχιστον από την υπόθεση της γραμμικότητας, δεδομένου ότι τα διαγράμματα των περιθωρίων σχέσεων των μεταβλητών x_i δεν εμφανίζουν γραμμικό μέσο.



Διάγραμμα 2.1 *scatterplot matrix* των περιθωρίων κατανομών $x_i | x_j$, $x_j | y$ και $y | x_j$

$$\text{για } x = (H, L, S, W)^T$$

Με χρήση του Arc για την επαγωγή πολυμεταβλητής κανονικότητας προκύπτει η *scatterplot matrix* όπως εμφανίζεται στο διάγραμμα που ακολουθεί όπου $x = (\log H, \log L, \log S, \log W)^T$



Διάγραμμα 2.2 *scatterplot matrix* των περιθωρίων κατανομών $x_i | x_j$, $x_j | y$ και $y | x_j$
για $x = (\log H, \log L, \log S, \log W)^T$

Είναι εμφανής η μεταβολή, ενώ είναι σαφές σύμφωνα και με όσα ειπώθηκαν ότι μπορεί να θεωρηθεί ότι ισχύει η υπόθεση της γραμμικότητας. Στη συνέχεια και παρατηρώντας τη δεξιά στήλη της *scatterplot matrix* βλέπουμε ότι η εικόνα των 2D διαγραμμάτων $x_j | y$ είναι συμβατή με το ενδεχόμενο $\dim S_{y|x} = 1$.

ΚΕΦΑΛΑΙΟ 3

ΜΕΘΟΔΟΙ ΡΟΠΗΣ ΠΡΩΤΗΣ ΤΑΞΗΣ, PIR (PARAMETRIC INVERSE REGRESSION)

Στο κεφαλαίο αυτό εξετάζονται δύο μέθοδοι. Η *SIR* (*Sliced inverse regression*) η οποία αποτελεί το σημαντικότερο εκπρόσωπο των μεθόδων που βασίζονται στη ροπή πρώτης τάξης της $x|y$ και η *PIR* (*Parametric inverse regression*) η οποία βασίζεται στην προσαρμογή παραμετρικών καμπυλών στα δεδομένα των p αντιστρόφων παλινδρομήσεων $x_j|y$.

3.1 ΜΕΘΟΔΟΣ SIR (SLICED INVERSE REGRESSION)

3.1.1 ΘΕΩΡΗΤΙΚΗ ΘΕΜΕΛΙΩΣΗ

Η μέθοδος αυτή εισήχθη από τον Li (1991a) ο οποίος θέλησε να εκτιμήσει τη βάση του $S_{y|x}$ μέσω της πρώτης ροπής $E(x|y)$ της αντίστροφης παλινδρόμησης, και βασίζεται σε ένα θεώρημα το οποίο απέδειξε. Το θεώρημα αυτό συνδέει τον κεντρικό υπόχωρο (*central subspace*) $S_{y|x}$ με τον $S_{E(x|y)} = \text{span}\{E(x|y) - E(x)|y \in \Omega_y\}$ και μπορεί να διατυπωθεί ως εξής (Cook (1998)).

Θεώρημα 3.1. Έστω η βάση του $S_{y|x}$ και $\Sigma = \text{Var}(x)$. Άν $E(x|\mathbf{n}^T x)$ είναι γραμμική συνάρτηση του $\mathbf{n}^T x$ τότε

$$S_{E(z|y)} \subset S(\Sigma \mathbf{n}) = \Sigma S_{y|x}$$

Άν τεθεί $z = \Sigma_x^{-1/2} (x - E(x))$, τότε προκύπτει (Cook (1998))

$$S_{E(z|y)} \subset S(\Sigma_x^{-1/2} \boldsymbol{\eta}) = S_{y|z}$$

Επομένως η εκτίμηση του υπόχωρου $S_{E(z|y)}$ της αντίστροφης παλινδρόμησης αποτελεί και εκτίμηση ενός τουλάχιστον μέρους του κεντρικού υπόχωρου (*central subspace*) $S_{y|z}$.

Η ακόλουθη πρόταση (Cook (1998)) αποτελεί τη βάση για την εκτίμηση του $S_{E(z|y)}$.

Πρόταση 3.1 $S\{Var[E(z|y)]\} = S_{E(z,y)}$

Σύμφωνα με την πρόταση αυτή τα ιδιοδιανύσματα της $Var[E(z|y)]$ που αντιστοιχούν σε μη μηδενικές ιδιοτιμές αποτελούν βάση του $S_{E(z,y)}$ (Cook (1998)).



Επομένως μία εκτίμηση του $S_{E(z,y)}$ μπορεί να προκύψει μέσω μίας εκτίμησης της $Var[E(z|y)]$.

Η χρήση των ιδιοδιανυσμάτων της $Var[E(z_i|y)]$ για την εκτίμηση της βάσης του $S_{y|z}$ δικαιολογείται και από το γεγονός ότι η $Var[E(z_i|y)]$ εκφυλίζεται σε οποιαδήποτε κατεύθυνση ορθογώνια στα διανύσματα βάσης του $S_{y|z}$. Επομένως οι κατευθύνσεις στις οποίες μεγιστοποιείται η $Var[E(z_i|y)]$ μπορούν να αποτελέσουν εκτίμηση των διανυσμάτων βάσης του $S_{y|z}$.

Για το σκοπό αυτό ο Li (1991a) προτείνει την αντικατάσταση της απόκριης y με μία διακριτοποιημένη μορφή \tilde{y} αυτής η οποία προκύπτει από τη δημιουργία h διαφορετικών ζωνών J_s , $s = 1, 2, \dots, h$ για τις τιμές της y . Ισχύει επομένως ότι $\tilde{y} = s$ όταν $y \in J_s$.

Αποδεικνύεται ότι $S_{\tilde{y}|x} \subset S_{y|x}$ (Cook (1998)). Όπως μάλιστα επισημαίνει ο Cook (1998) αν υποτεθεί ότι $S_{\tilde{y}|x} = S_{y|x}$, πράγμα που ισχύει όταν το h είναι μεγάλο, τότε σε συνδυασμό με την πρόταση 3.1 προκύπτει

$$S\{Var[E(z|\tilde{y})]\} = S_{E(z|\tilde{y})} \subset S_{\tilde{y}|z} \subset S_{y|z}$$

Πριν προχωρήσουμε στην παρουσίαση του αλγορίθμου για τη μέθοδο SIR πρέπει να γίνει αναφορά στις περιπτώσεις στις οποίες η SIR αποτυγχάνει να εκτιμήσει τον $S_{y|z}$. Αυτό συμβαίνει όταν για παράδειγμα $y = g(\beta_1 x) + \varepsilon$, όπου g συμμετρική συνάρτηση του $\beta_1 x$, για $\beta_1 x$ συμμετρικό περί το 0. Σε μια τέτοια περίπτωση $E(x|y)=0$ και η καμπύλη της τυποποιημένης αντίστροφης συνάρτησης παλινδρόμησης θα βρίσκεται μέσα σε ένα γνήσιο υποσύνολο του $S_{y|z}$ (Li (1991a)). Συνεπώς η SIR δεν θα μπορεί να εκτιμήσει τη βάση του $S_{y|z}$, να ανακαλύψει δηλαδή τις σημαντικές κατευθύνσεις του προβλήματος.

3.1.2 ΑΛΓΟΡΙΘΜΟΣ

Έστω $\hat{\Sigma}_x$ η εκτίμηση της Σ_x και $\hat{z}_i = \hat{\Sigma}_x^{-1/2}(x_i - \bar{x})$, όπου \bar{x} ο δειγματικός μέσος και $i = 1, 2, \dots, n$, όπου n το συνολικό πλήθος των παρατηρήσεων. Έστω επίσης n_s ο αριθμός των παρατηρήσεων στη ζώνη s . Ο αλγόριθμος της SIR προτάθηκε από τον Li (1991a) και έχει ως εξής.



Βήμα 1 Υπολογισμός σε κάθε ζώνη του δειγματικού μέσου των \hat{z}

$$\bar{z}_s = (\sum_{y_i \in J_s} \hat{z}_i) / n_s$$

Βήμα 2 Υπολογισμός της σταθμισμένης δειγματικής μήτρας συνδιακύμανσης

$$\hat{V} = \frac{I}{n} \sum_{s=1}^k n_s \bar{z}_s \bar{z}_s^T$$

Η μήτρα \hat{V} είναι συνεπής εκτίμηση της αντίστοιχης πληθυσμιακής μήτρας

$$Var [E(z|\tilde{y})] = \sum_{s=1}^h Pr(\tilde{y}=s) \mu_{z/s} \mu_{z/s}^T$$

όπου $\mu_{z/s} = E(z/\tilde{y}=s)$. Επομένως οι ιδιοτιμές και τα ιδιοδιανύσματα της \hat{V} είναι συνεπείς εκτιμήσεις των ιδιοτιμών λ_j και των ιδιοδιανυσμάτων I_j της $Var [E(z|\tilde{y})]$.

Βήμα 3 Εύρεση των ιδιοδιανυσμάτων $\hat{I}_1, \dots, \hat{I}_p$ που αντιστοιχούν στις ιδιοτιμές

$$\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p \text{ της μήτρας } \hat{V}.$$

Βήμα 4 Έστω $d = dim[S_{E(z/\tilde{y})}]$, τότε η SIR εκτίμηση $\hat{S}_{E(z/\tilde{y})}$ του $S_{E(z/\tilde{y})}$ θα είναι

$$\hat{S}_{E(z/\tilde{y})} = S(\hat{I}_1, \dots, \hat{I}_d)$$

Η SIR εκτίμηση του $S_{y/x}$ θα είναι τότε

$$\hat{S}_{y/x} = \hat{\Sigma}_x^{-1/2} \hat{S}_{E(z/\tilde{y})} = S(\hat{\Sigma}_x^{-1/2} \hat{I}_1, \dots, \hat{\Sigma}_x^{-1/2} \hat{I}_d)$$

Θέτοντας $\hat{\phi}_j = \hat{\Sigma}_x^{-1/2} \hat{I}_j$ η έκφραση της απόκρισης y συναρτήσει των d γραμμικών συνδυασμών $\hat{\phi}_1^T x, \dots, \hat{\phi}_d^T x$ παρέχει όλη την πληροφορία της παλινδρόμησης του y στο διάνυσμα x .

3.1.3 ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΤΑΣΗΣ d

Η εκτίμηση της διάστασης d του $S_{y/x}$ μπορεί να γίνει με δύο τρόπους. Ο πρώτος τρόπος είναι γραφικά και δεν θα μας απασχολήσει προς το παρόν. Ο δεύτερος τρόπος είναι μέσω ενός ελέγχου χ^2 ο οποίος χρησιμοποιεί το στατιστικό

$$\hat{\Lambda}_m = n \sum_{j=m+1}^p \hat{\lambda}_j$$

για τον έλεγχο της υπόθεσης $d = m$ έναντι της $d > m$.

Ξεκινώντας με την τιμή $m = 0$ συγκρίνεται η τιμή του παραπάνω στατιστικού με το επιθυμητό εκατοστιαίο σημείο της κατανομής του υπό την ισχύ της μηδενικής υπόθεσης. Εάν η τιμή του στατιστικού είναι μικρότερη τότε δεν μπορεί να απορριφθεί η υπόθεση $d = m$. Εάν είναι μεγαλύτερη τότε συνάγεται ότι $d > m$, αυξάνεται η τιμή του m κατά 1 και η διαδικασία επαναλαμβάνεται.

Για την εύρεση της ασυμπτωματικής κατανομής του $\hat{\Lambda}_d$ γίνονται κάποιες παραδοχές. Υπάρχουν δύο περιπτώσεις για τις παραδοχές αυτές. Η πρώτη προτάθηκε από τον Li (1991a) ο οποίος απέδειξε το παρακάτω θεώρημα.

Θεώρημα 3.2 Αν το διάνυσμα x κατανέμεται κανονικά τότε $n(p-d) \bar{\lambda}_{(p-d)}$ ακολουθεί

ασυμπτωματικά κατανομή χ^2 με $(p-d)(n-d-1)$ βαθμούς ελευθερίας.

Η δεύτερη απαντάται στον Cook(1998) καθώς και στους Bura-Cook(2001a) οι οποίοι πρότειναν τον λεγόμενο σταθμισμένο έλεγχο χ^2 (*weighted chi-squared test*) βασιζόμενοι σε πιο χαλαρές παραδοχές. Ο έλεγχος αυτός καθώς και τα πλεονεκτήματα του σε σχέση με τον έλεγχο του Li(1991a) θα μας απασχολήσουν αργότερα.

3.1.4 ΕΦΑΡΜΟΓΕΣ

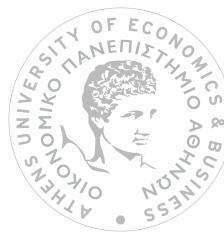
ΕΦΑΡΜΟΓΗ 1 Η εφαρμογή που ακολουθεί παρουσιάζεται από τον Li (1991a)
Έστω το μοντέλο

$$y = x_1 + x_2 + x_3 + x_4 + 0x_5 + \varepsilon \quad (3.1)$$

το οποίο χρησιμοποιήθηκε για την παραγωγή 100 iid ζευγών παρατηρήσεων (x_i, y_i) .

Ισχύει ότι x, ε ακολουθούν την τυποποιημένη κανονική κατανομή και $x \perp \!\! \perp \varepsilon$. Το μοντέλο δηλαδή πληροί τη συνθήκη γραμμικότητας, είναι ομοσκεδαστικό, μονοδιάστατο και η βάση του $S_{y|x}$ είναι $\beta = (1, 1, 1, 1, 0)$. Τυποποιώντας ώστε το μέτρο του διανύσματος β να είναι μοναδιαίο προκύπτει $\beta' = (0.5, 0.5, 0.5, 0.5, 0)$. Ο πίνακας 3.1 που ακολουθεί δείχνει την εκτίμηση $\hat{\beta}_1$, του διανύσματος β' όπως προκύπτει με χρήση της μεθόδου SIR για τρεις διαφορετικές τιμές του πλήθους ζωνών H. Η τιμή για κάθε μία από τις συντεταγμένες του $\hat{\beta}_1$ έχει προκύψει ως ο μέσος των τιμών για 100 επαναλήψεις, ενώ σε παρένθεση σημειώνεται η αντίστοιχη τυπική απόκλιση.

Όπως φαίνεται και για τις τρεις περιπτώσεις τιμών του h η κατεύθυνση που εκτιμά η SIR είναι πολύ κοντά στην πραγματική.



H	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	$\hat{\beta}_{14}$	$\hat{\beta}_{15}$
5	.505 (.052)	.498 (.049)	.494 (.056)	.488 (.056)	.002 (.066)
10	.502 (.046)	.500 (.045)	.492 (.055)	.491 (.049)	.001 (.060)
20	.500 (.048)	.502 (.046)	.497 (.053)	.487 (.054)	.003 (.060)

Πίνακας 3.1 συντετογμένες(μέσος,τυπ.απόκλιση) του εκτιμώμενου από τη SIR

διανύσματος βάσης $\hat{\beta}_1$

Ας δούμε τώρα δύο άλλες περιπτώσεις μοντέλων

$$y = x_1(x_1 + x_2 + 1) + \sigma \varepsilon \quad (3.2)$$

και

$$y = x_1 / (0.5 + (x_2 + 1.5)^2) + \sigma \varepsilon \quad (3.3)$$

Και για τα μοντέλα αυτά ισχύει η ανεξαρτησία των συνιστωσών του x , η ανεξαρτησία του x από το ε , η συνθήκη γραμμικότητας λόγω κανονικότητας και η ομοσκεδαστικότητα. Το μέγεθος του δείγματος είναι $n = 400$.

Αυτή τη φορά η διάσταση του S_{yx} είναι $d=2$.

Οι πίνακες 3.2 και 3.3 που ακολουθούν δίνουν το μέσο του $R^2(\hat{\beta}_1)$ και του $R^2(\hat{\beta}_2)$ για δύο διαφορετικές τιμές του σ και τρεις διαφορετικές τιμές του H για 100 επαναλήψεις και $p = 10$ συνιστώσεις του x . $\hat{\beta}_1$ και $\hat{\beta}_2$ είναι οι 2 πρώτες κατευθύνσεις που προκύπτουν από εφαρμογή της SIR και $R^2(\hat{\beta}_1), R^2(\hat{\beta}_2)$ είναι οι συντελεστές προσδιορισμού της παλινδρόμησης των $\hat{\beta}_1$ και $\hat{\beta}_2$ αντίστοιχα στα πραγματικά διανύσματα βάσης του S_{yx} . Τα πραγματικά διανύσματα βάσης και για τα δύο μοντέλα είναι $(1,0,\dots,0)$ και $(0,1,\dots,0)$.

Όπως φαίνεται από τους πίνακες αυτούς οι τιμές του $R^2(\hat{\beta}_1)$ είναι ψηλές. Η πρώτη κατεύθυνση δηλαδή που προκύπτει από τη SIR «εξηγεί» ικανοποιητικά την πραγματική βάση.

Για το $R^2(\hat{\beta}_2)$ όμως δεν ισχύει το ίδιο. Οι τιμές δεν είναι ικανοποιητικές πράγμα που σημαίνει ότι η δεύτερη κατεύθυνση που προκύπτει από τη SIR χάνει την πληροφορία της πραγματικής βάσης.

Τα αποτελέσματα δεν φαίνεται να επηρεάζονται από τις τιμές του H ενώ είναι κάπως καλύτερα για $\sigma = 0,5$ έναντι $\sigma = 1$

H	$\sigma = 0.5$		$\sigma = 1$	
	$R^2(\hat{\beta}_1) R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1) R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1) R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1) R^2(\hat{\beta}_2)$
5	.91 (.05)	.75 (.15)	.38 (.07)	.52 (.21)
10	.92 (.04)	.80 (.013)	.89 (.08)	.55 (.24)
20	.93 (.04)	.77 (.15)	.88 (.08)	.49 (.26)

Πίνακας 3.2 συντελεστές προσδιορισμού των $\hat{\beta}_1^T x, \hat{\beta}_2^T x$, όπου $\hat{\beta}_1, \hat{\beta}_2$

οι εκτιμήσεις της SIR, στις $\beta_1^T x, \beta_2^T x$, όπου β_1, β_2 τα πραγματικά διανύσματα βάσης (Μοντέλο 3.2)

H	$\sigma = 0.5$		$\sigma = 1$	
	$R^2(\hat{\beta}_1) R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1) R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1) R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1) R^2(\hat{\beta}_2)$
5	.96 (.02)	.83 (.08)	.89 (.06)	.51 (.23)
10	.96 (.02)	.88 (.06)	.90 (.06)	.56 (.23)
20	.9 (.02)	.89 (.06)	.90 (.06)	.53 (.24)

Πίνακας 3.3 συντελεστές προσδιορισμού των $\hat{\beta}_1^T x, \hat{\beta}_2^T x$, όπου $\hat{\beta}_1, \hat{\beta}_2$

οι εκτιμήσεις της SIR, στις $\beta_1^T x, \beta_2^T x$, όπου β_1, β_2 τα πραγματικά διανύσματα βάσης (Μοντέλο 3.3)

Επιβεβαιώνεται έτσι η αδυναμία της μεθόδου να «ανακαλύψει» την καμπυλότητα του μέσου.

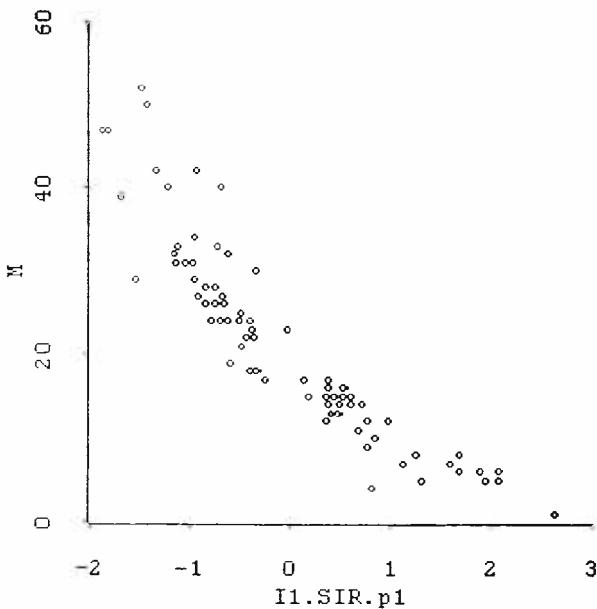


Στη συνέχεια παρουσιάζεται μία εφαρμογή (Cook 1998) η οποία δείχνει ότι αν και η SIR δεν είναι και πολύ κατάλληλη στο να αποκαλύπτει την καμπυλότητα του μέσου, είναι ευαίσθητη στην ετεροσκεδαστικότητα.

ΕΦΑΡΜΟΓΗ 2 Η εφαρμογή που ακολουθεί βασίζεται στα γνωστά δεδομένα για οστρακοειδή για τα οποία έγινε λόγος στην παράγραφο 2.4 και παρουσιάζεται από τον Cook (1998). Η απόκριση M είναι η μάζα του πυρήνα και οι ανεξάρτητες μεταβλητές που ενδιαφέρουν είναι το μήκος L του οστρακοειδούς, το πλάτος του W και η μάζα του S . Οι μεταβλητές W και S μετασχηματίστηκαν σε $W^{0.36}$ και $S^{0.11}$ με τη βοήθεια του προγράμματος Arc για την επαγωγή κανονικότητας ώστε να ικανοποιείται (κατά το δυνατόν) η συνθήκη γραμμικότητας.

Επομένως $x = (L, W^{0.36}, S^{0.11})^T$.

Η υπόθεση $d = 1$ δεν απορρίπτεται από τον έλεγχο χ^2 για τη διάσταση του S_{yx} για $h = 20$, ενώ το διάγραμμα που ακολουθεί δεν είναι άλλο από το διάγραμμα $\{M, \hat{\phi}_1^T x\}$ όπου $\hat{\phi}_1$ το διάνυσμα βάσης του κεντρικού υπόχωρου που εκτιμά η SIR .



Διάγραμμα 3.1 διάγραμμα $\{M, \hat{\phi}_1^T x\}$ για το πρόβλημα με τα οστρακοειδή ($\hat{\phi}_1$ η πρώτη κατεύθυνση που εκτιμά η SIR)

Παρατίθεται παράλληλα και το σχετικό ουτρυπ του προγράμματος Arc.

```

Inverse Regression SIR
Name of Dataset = Mussels
Name of Fit = I1.SIR
Response = M
Predictors = (L W^0.36 S^0.11)

Number of slices = 18
Slices sizes are: (5 5 4 5 6 4 5 6 4 5 5 4 4 6 4 4 4 2)
Std. coef. use predictors scaled to have SD equal to one.
          Lin Comb 1      Lin Comb 2      Lin Comb 3
Predictors    Raw     Std.     Raw     Std.     Raw     Std.
L            -0.000  -0.084   -0.007  -0.818   -0.001  -0.197
W^0.36       -0.108  -0.268    0.472   0.466   -0.287  -0.598
S^0.11       -0.994  -0.960    0.881   0.338    0.958   0.777

Eigenvalues           0.922           0.233           0.090
R^2(OLS| SIR)        0.999           1.000           1.000

```

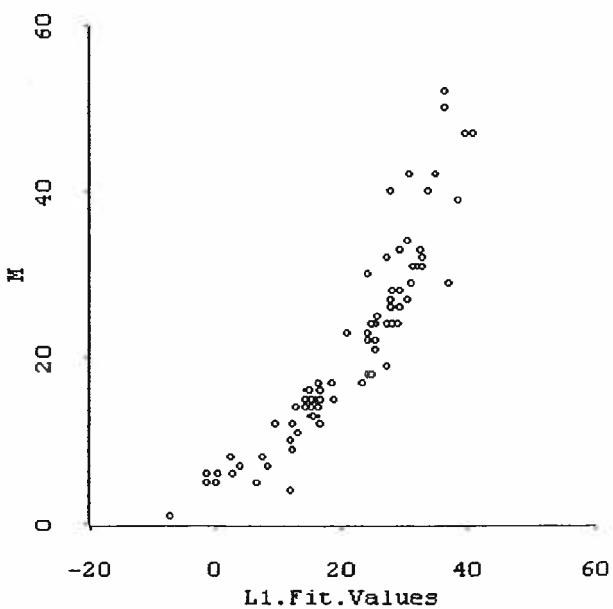
Approximate Chi-squared test statistics based on partial sums of eigenvalues times 82

Number of Components	Test Statistic	df	p-value
1	102.15	51	0.000
2	26.551	32	0.739
3	7.4195	15	0.945

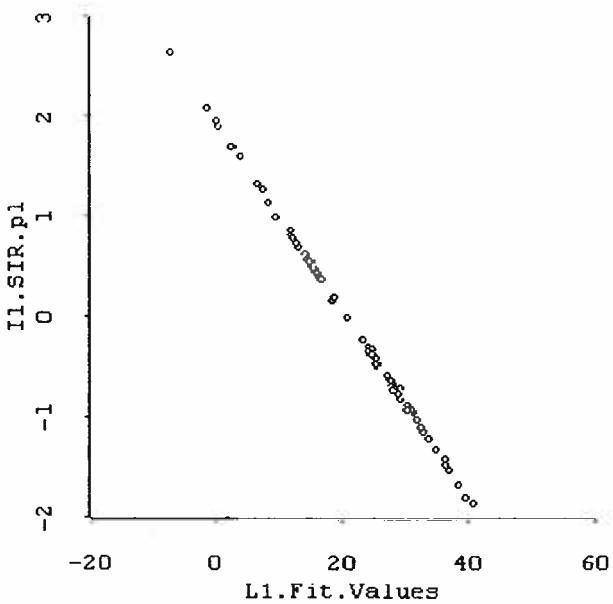
Plot: [Plot1]Mussels:(I1.SIR) SIR

Ο συντελεστής προσδιορισμού μεταξύ των μεταβλητών $b_{ols}^T x$ και $\hat{\phi}_1^T x$, όπου b_{ols} το διάνυσμα των συντελεστών της μεθόδου OLS, είναι 0.999 γεγονός που επιβεβαιώνει το γνωστό μας *1D estimation result*.

Χαρακτηριστική είναι και η εικόνα του διαγράμματος $\{\hat{\phi}_1^T x, b_{ols}^T x\}$ που ακολουθεί μαζί με το διάγραμμα $\{M, b_{ols}^T x\}$.



Διάγραμμα 3.2 διάγραμμα { M , $b_{ols}^T x$ } για το πρόβλημα με τα οστρακοειδή(b_{ols} η κατεύθυνση OLS



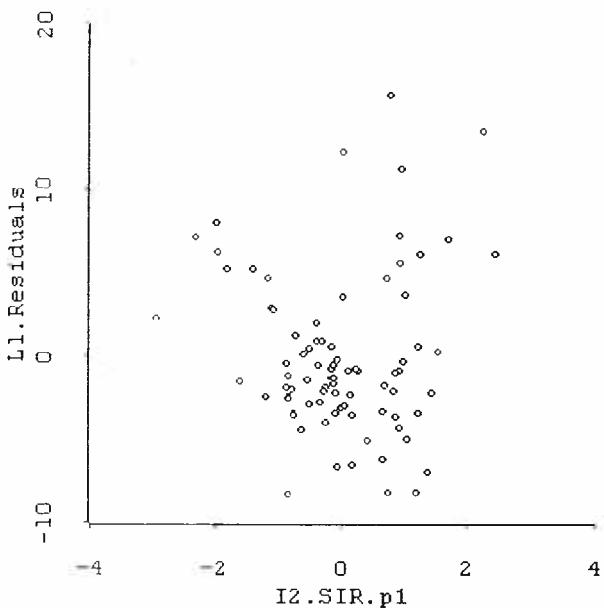
Διάγραμμα 3.3 διάγραμμα { $\hat{\phi}_1^T x$, $b_{ols}^T x$ } για το πρόβλημα με τα οστρακοειδή
Τέλος η εφαρμογή της μεθόδους SIR στα κατάλοιπα του μοντέλου

$$M | x = \beta_0 + \beta^T x + \varepsilon$$

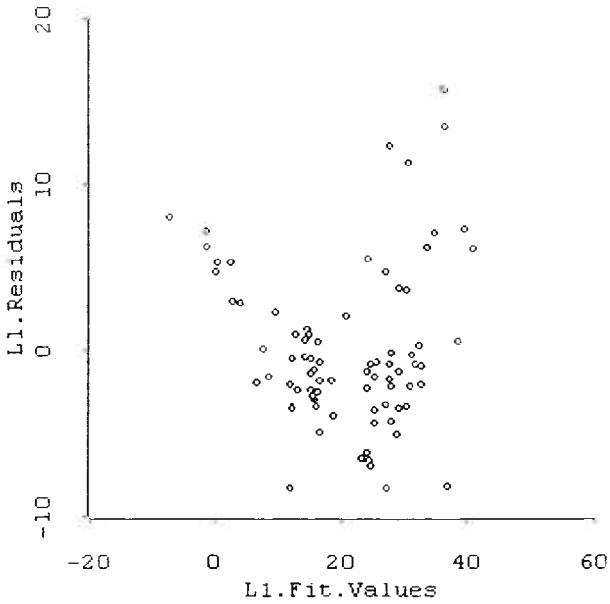


όπου $\mathbf{x} = (L, W^{0.36}, S^{0.11})^T$.

Έδειξε $\dim S_{y|x}=0$ ενώ σύγκριση του διαγράμματος $\{\hat{e}, \hat{\phi}_1^T \mathbf{x}\}$ για την πρώτη κατεύθυνση $\hat{\phi}_1$ που εκτιμά η SIR , και του διαγράμματος $\{\hat{e}, \mathbf{b}_{ols}^T \mathbf{x}\}$ δείχνει την αδυναμία της SIR στην ανίχνευση της καμπυλότητας του μέσου αλλά ταυτόχρονα και την ευαισθησία της στην ετεροσκεδαστικότητα. Τόσο η καμπυλότητα του μέσου όσο και η ύπαρξη ετεροσκεδαστικότητας ανιχνεύονται από το διάγραμμα $\{\hat{e}, \mathbf{b}_{ols}^T \mathbf{x}\}$. Τα δύο διαγράμματα εικονίζονται παρακάτω.



Διάγραμμα 3.4 διάγραμμα $\{\hat{e}, \hat{\phi}_1^T \mathbf{x}\}$ για το πρόβλημα με τα οστρακοειδή (\hat{e} τα κατάλοιπα του μοντέλου $M | \mathbf{x} = \beta_0 + \boldsymbol{\beta}^T \mathbf{x} + \varepsilon$)



Διάγραμμα 3.5 διάγραμμα { \hat{e} , $\beta_{ols}^T x$ } για το πρόβλημα με τα οστρακοειδή

3.1.5 ΣΧΟΛΙΑ – ΕΠΙΣΗΜΑΝΣΕΙΣ- ΠΡΟΕΚΤΑΣΕΙΣ

3.1.5α ΠΑΡΑΒΙΑΣΗ ΤΗΣ ΥΠΟΘΕΣΗΣ ΓΡΑΜΜΙΚΟΤΗΤΑΣ ΤΟΥ ΘΕΩΡΗΜΑΤΟΣ 3.1

Η μέθοδος *SIR* θεωρεί ότι

$$y = f(\beta_1^T x, \beta_2^T x, \dots, \beta_k^T x, \varepsilon)$$

ότι δηλαδή η απόκριση y εξαρτάται από το διάνυσμα x των p ανεξαρτήτων μεταβλητών μέσω k γραμμικών συνδυασμών των μεταβλητών αυτών χωρίς να κάνει καμία απολύτως υπόθεση για τη μορφή της εξάρτησης αυτής. Αυτή η απουσία υποθέσεων όπως επισημαίνουν οι Cook- Weisberg (1991) καθιστά τη *SIR* πολύ πρόσφορη για τη διαγνωστική αντιμετώπιση προβλημάτων τα οποία θα ήταν πολύ δύσκολο να προσεγγιστούν μέσω της ευθείας παλινδρόμησης.

Ωστόσο το θεώρημα 3.1 στο οποίο βασίζεται η *SIR* δεν είναι ελεύθερο υποθέσεων. Η υπόθεση της γραμμικότητας η οποία μπορεί να εκφραστεί ως

$$E(\mathbf{b}^T \mathbf{x} | \beta_1^T \mathbf{x}, \beta_2^T \mathbf{x}, \dots, \beta_k^T \mathbf{x}) = c_0 + c_1 \beta_1^T \mathbf{x} + c_2 \beta_2^T \mathbf{x} + \dots + c_k \beta_k^T \mathbf{x} \quad \forall \mathbf{b} \in \mathbb{R}^p$$

όπου c_0, c_1, \dots, c_k κάποιες σταθερές, δεν είναι βέβαιο ότι θα ισχύει πάντα.

Σίγουρα πάντως θα ισχύει όταν η κατανομή του x έχει ελλειπτική συμμετρία ακόμα δε περισσότερο όταν είναι κανονική.

Θέση στο θέμα πήρε και ο ίδιος ο Li (1991b) αντιμετωπίζοντας δύο περιπτώσεις. Για την περίπτωση της ήπιας παράβασης της υπόθεσης της γραμμικότητας χρησιμοποίησε το μοντέλο (3) - το οποίο παρουσιάστηκε στις εφαρμογές της παραγράφου 3.1.4 –θεωρώντας αυτή τη φορά ότι το διάνυσμα x κατανέμεται ομοιόμορφα στο διάστημα $[-\sqrt{3}, \sqrt{3}]^p$ και όχι κανονικά και επανέλαβε το πείραμα αντικαθιστώντας τα πραγματικά διανύσματα βάσης με δύο ορθογώνια διανύσματα τυχαία επιλεγμένα με βάση την ομοιόμορφη κατανομή στη μοναδιαία σφαίρα του \mathbb{R}^p . Τα αποτελέσματα έδειξαν ψηλές τιμές για τα $R^2(\hat{\beta}_1)$ και $R^2(\hat{\beta}_2)$ γεγονός που αποδεικνύει ότι η αποτελεσματικότητα της SIR έμεινε ανεπηρέαστη και επαληθεύει τους Diaconis και Freedman (1984) για τους οποίους έγινε λόγος στην παράγραφο 2.3.1.

Ας δούμε τώρα την περίπτωση της ευθείας παραβίασης της υπόθεσης της γραμμικότητας.

Ο Li (1990a) ασχολούμενος με μια τέτοια περίπτωση χρησιμοποίησε το μονοδιάστατο μοντέλο

$$y = \beta_1^T x + \varepsilon$$

και μεταξύ της πραγματικής βάσης $\beta_1 = (1, 0, 0, 0, 0)$ και της κατεύθυνσης $b = (0, 1, 0, 0, 0)$ επέβαλε τον περιορισμό

$$(\beta_1^T x)^2 - 0.5 \leq bx \leq (\beta_1^T x)^2 + 0.5$$

Το αποτέλεσμα είναι τα σημεία του επιπέδου που σχηματίζεται από τα διανύσματα β_1 , b δηλαδή του επιπέδου x_1, x_2 , να κατανέμονται κατά μήκος μιας καμπύλης αντί της επιθυμητής ευθείας.

Η μέθοδος SIR ωστόσο απέδωσε καλύτερα από ότι θα μπορούσε να περιμένει κανείς. Απεκάλυψε την πραγματική κατεύθυνση β_1 και επιπλέον αυτής την κατεύθυνση b . (Το μοντέλο ωστόσο είναι εκ κατασκευής μονοδιάστατο πράγμα που μας κάνει να υποπτευόμαστε ότι υπάρχει μία και μόνη κατεύθυνση στο επίπεδο που σχηματίζεται από τα b , β_1 η οποία να παρέχει όλη την πληροφορία που είναι διαθέσιμη για το y μέσω του x . Το θέμα μπορεί να προσεγγιστεί γραφικά μέσω του 3D Διαγράμματος $\{y, (SIR1, SIR2)\}$ σύμφωνα με όσα αναφέρονται στην παράγραφο 5.3.1.)



Ας δούμε τώρα τι προτείνει ο Li (1991b) για την αντιμετώπιση του προβλήματος της ευθείας παραβίασης της υπόθεσης της γραμμικότητας.

Η πρώτη πρόταση είναι η ιδέα του “*double-slicing*”. Σύμφωνα με αυτήν (για $\kappa=1$) η βάση του $S_{y|x}$ είναι το διάνυσμα b για το οποίο ελαχιστοποιείται η μέγιστη ιδιοτιμή της μήτρας

$$\text{Cov} [E(x|bx,y)] - \text{Cov} [E(x|bx)]$$

όπου x το τυποποιημένο διάνυσμα των ανεξαρτήτων μεταβλητών.

Μία άλλη πρόταση είναι αυτή σύμφωνα με την οποία (και πάλι για $\kappa=1$) η βάση του $S_{y|x}$ είναι το διάνυσμα b για το οποίο ελαχιστοποιείται η μέγιστη ιδιοτιμή της μήτρας

$$\text{Cov} \{E[x-E(x|bx)|y]\}.$$

3.1.5 β ΣΤΑΘΜΙΣΜΕΝΟΣ ΕΛΕΓΧΟΣ χ^2

Όπως είδαμε ο Li (1991a) απέδειξε ότι η ασυμπτωτική κατανομή του $\hat{\Lambda}_d$ είναι η χ^2 , υπό την προϋπόθεση ότι x κατανέμεται κανονικά.

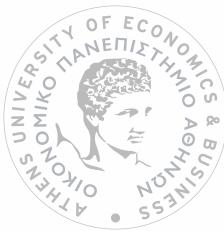
Ο Cook (1998) και οι Bura και Cook (2001a) ωστόσο απέδειξαν ότι δεν χρειάζεται το x να κατανέμεται κανονικά. Συγκεκριμένα απέδειξαν το παρακάτω θεώρημα (Cook (1998)).

Θεώρημα 3.3 Έστω y βάση του $S_{\bar{y}|z}$, $d = \dim[S_{E(z)\bar{y}}]$, h ο αριθμός των ζωνών, p ο αριθμός των ανεξάρτητων μεταβλητών και P_γ ο ορθογώνιος *projection operator* στον $S_{\bar{y}|z}$ (γ). Αν ισχύει ότι

1. $h > d + 1$ και $p > d$
2. $S_{E(z)\bar{y}} = S_{\bar{y}|z}$ (γ)
3. $E(z|\gamma^T z) = P_\gamma z$
4. $Var(z|\gamma^T z) = I_p - P_\gamma$

όπου z το τυποποιημένο διάνυσμα x , τότε η ασυμπτωτική κατανομή του $\hat{\Lambda}_d$ θα είναι χ^2 με $(p-d)(h-d-1)$ βαθμούς ελευθερίας.

Αν το διάνυσμα x κατανέμεται όπως υπέθεσε ο Li (1991a) τότε οι συνθήκες 3 και 4 ισχύουν.



Οι Bura και Cook (2001a) ισχυρίζονται ότι αποδεικνύεται ότι αν $\Sigma_{z|y}$ σταθερή μεταξύ των ζωνών και ισχύουν οι συνθήκες 1,2,3 τότε θα ισχύει και η 4. Δεδομένου ότι η 3 προϋποθέτει την ισχύ της υπόθεσης γραμμικότητας του θεωρήματος 3.1, η οποία είναι απαραίτητη για την εφαρμογή της *SIR*, καταλήγουμε στο συμπέρασμα ότι η καθοριστική συνθήκη για την εφαρμογή του ελέγχου χ^2 είναι η σταθερότητα της διακύμανσης της $z|y$ ή των $z|y^T z$. Η σταθερότητα αυτή των διακυμάνσεων μπορεί εύκολα να ελεγχθεί μέσω της *scatterplot matrix* των ανεξάρτητων μεταβλητών και της απόκρισης.

Ο Cook (1998) και οι Bura και Cook (2001a) εισήγαγαν επίσης έναν γενικότερο έλεγχο για την εκτίμηση της d , τον λεγόμενο σταθμισμένο έλεγχο χ^2 , ο οποίος κάνει ελάχιστες υποθέσεις. Συγκεκριμένα απέδειξαν το παρακάτω θεώρημα.

Θεώρημα 3.4 Έστω $d = \dim[S_{E(x,y)}]$. Αν $h > d + 1$ και $p > d$ τότε η ασυμπτωτική

κατανομή του $\hat{\Lambda}_d$ θα είναι ίδια με την κατανομή του

$$C = \sum_{k=1}^{(p-d)(h-d)} w_k C_k$$

όπου C_k είναι ανεξάρτητες χ^2 τυχαίες μεταβλητές με 1 βαθμό ελευθερίας και $w_1 \geq w_2 \geq \dots \geq w_{(p-d)(h-d)}$ οι ιδιοτιμές της μήτρας

$$\Delta_c = (\Gamma_{22}^T \otimes \Gamma_{12}^T) \Delta (\Gamma_{22} \otimes \Gamma_{12})$$

Οι $\Gamma_{22}^T, \Gamma_{12}^T, \Gamma_{22}, \Gamma_{12}$ προκύπτουν από τις Γ_1, Γ_2 για τις οποίες ισχύει

$$\mathbf{B} = \Gamma_1 \begin{bmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{bmatrix} \Gamma_2^T$$

ως εξής. Για την ορθοκανονική pxp μήτρα Γ_1 ισχύει

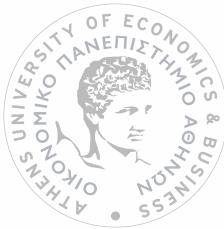
$$\Gamma_1 = (\Gamma_{11}, \Gamma_{12})$$

όπου

Γ_{11} μήτρα pxd και Γ_{12} μήτρα $px(p-d)$, για δε την ορθοκανονική hxh μήτρα Γ_2 ισχύει

$$\Gamma_2^T = \begin{bmatrix} \Gamma_{21}^T \\ \Gamma_{22}^T \end{bmatrix}$$

όπου Γ_{21}^T μήτρα dxh και Γ_{22}^T μήτρα $(h-d)xh$.



Επίσης

$$\mathbf{B} = (E(z | \tilde{y} = I) p_I^{1/2}, \dots, E(z | \tilde{y} = h) p_h^{1/2})$$

όπου $P_s = \Pr(\tilde{y} = s)$, $s=1,2,\dots,h$

και \mathbf{D} η $d \times d$ διαγώνιος μήτρα με διαγώνια στοιχεία τις θετικές ιδιάζουσες τιμές της \mathbf{B} .

Η μήτρα \mathbf{A} είναι $h \times h$ μπλόκ μήτρα με στοιχεία μήτρες A_{ts} διαστάσεων $p \times p$.

Για $t = s$ δηλαδή για τα διαγώνια στοιχεία της \mathbf{A} ισχύει

$$A_{ss} = \mathbf{I}_p p_s + (1 - 2p_s) \Sigma_{zs}, \text{ όπου } \Sigma_{zs} = \Sigma_x^{-1/2} \Sigma_{x|s} \Sigma_x^{-1/2}$$

για $t \neq s$

$$A_{ts} = (p_t p_s)^{1/2} (\mathbf{I}_p - \Sigma_{zt} - \Sigma_{zs})$$

Χρησιμοποιώντας τις τιμές που προκύπτουν από το εκάστοτε δείγμα για τις $\Sigma_{x|s}$, Σ_x και p_s προκύπτει η συνεπής εκτίμηση $\hat{\mathbf{A}}$ της \mathbf{A} .

Επίσης για την υποτιθέμενη τιμή της d και χρησιμοποιώντας την συνεπή εκτίμηση $\hat{\mathbf{Z}}_n$ της \mathbf{B} με βάση το δείγμα, προκύπτουν οι συνεπείς εκτιμήσεις $\hat{\Gamma}_{22}, \hat{\Gamma}_{12}$ των Γ_{22} , Γ_{12} και επομένως μπορεί να προκύψει και η σχετική εκτίμηση $\hat{\Lambda}_c$ της \mathbf{A}_c . Αν \hat{w}_k είναι οι ιδιοτιμές της $\hat{\Lambda}_c$ τότε η κατανομή του

$$\hat{C} = \sum_{k=1}^{(p-d)(h-d)} \hat{w}_k C_k$$

είναι συνεπής εκτίμηση της ασυμπτωτικής κατανομής του $\hat{\Lambda}_d$.

Ωστόσο, όπως αναφέρουν οι Bura και Cook (2001a), θα πρέπει να ληφθεί υπόψιν ότι μεγάλος αριθμός ζωνών h μπορεί να προκαλέσει αύξηση της διακύμανσης των βαρών \hat{w}_k και να θέσει υπό αμφισβήτηση τον χρησιμότητα της εκτίμησης της ασυμπτωτικής κατανομής του $\hat{\Lambda}_d$.

Ας δούμε κάποια παραδείγματα σύγκρισης του ελέγχου χ^2 και του σταθμισμένου ελέγχου χ^2 τα οποία παραθέτουν.

Το πρώτο από αυτά είναι το γνωστό παράδειγμα των οστρακοειδών της 3.1.4β όπου το αποτέλεσμα του σταθμισμένου ελέγχου χ^2 για την εύρεση της διάστασης d δίνει $d = 1$ αποτέλεσμα το οποίο συμφωνεί με την γραφική προσέγγιση του προβλήματος. Αντίθετα ο έλεγχος χ^2 δίνει $d = 2$ γεγονός που φαίνεται να είναι το «κόστος» από τη χρήση των σχετικών υποθέσεων που μάλλον δεν ισχύουν παρά το μετασχηματισμό



των ανεξάρτητων μεταβλητών. Στο δεύτερο παράδειγμα τους οι Bura και Cook (2001a) χρησιμοποιούν δύο διδιάστατα μοντέλα.

Το πρώτο έχει ως εξής

$$y = (4 + x_1)(2 + x_2 + x_3) + 0.5 \varepsilon \quad (3.4)$$

$$\begin{aligned} x_1 &= w_1 \\ x_2 &= v_1 + w_2 / 2 \\ x_3 &= -v_1 + w_2 / 2 \\ x_4 &= v_2 + v_3 \\ x_5 &= v_2 - v_3 \end{aligned} \quad (3.5)$$

Ισχύει ότι $v_1, v_2, v_4 \sim iid t_{(4)}$, $v_3 \sim t_{(3)}$, $v_5 \sim t_{(5)}$, w_1 και $w_2 \sim iid$ μεταβλητές με συνάρτηση πυκνότητας τη συνάρτηση γάμμα (0.25) και $\varepsilon \sim N(0,1)$ ανεξάρτητο από τις μεταβλητές v, w .

Ενώ το δεύτερο είναι

$$y = x_1 / (0.5 + (x_2 + 1.5)^2) + 0.5 \varepsilon \quad (3.6)$$

$$\begin{aligned} x_1 &= v_3 + v_4 + w / 6 \\ x_2 &= -v_3 + v_4 + w / 6 \\ x_3 &= -v_4 + w / 3 \\ x_4 &= v_1 + v_2 \\ x_5 &= -v_1 + v_2 \end{aligned} \quad (3.7)$$

ισχύει ότι v_1, v_2, v_3 και $v_4 \sim iid U(-4,4)$, $w \sim N(0,1)$ και $\varepsilon \sim N(0,1)$ ανεξάρτητο από τις μεταβλητές v, w .

Από τις *scatterplot matrices* για τα δύο μοντέλα προκύπτει ότι ισχύει η υπόθεση της γραμμικότητας δεν συμβαίνει όμως το ίδιο για την υπόθεση της σταθερής διακύμανσης των δεσμευμένων κατανομών των ανεξάρτητων μεταβλητών μεταξύ τους καθώς και της δεσμευμένης κατανομής $x|y$. Επομένως η συμπτωτική κατανομή του $\hat{\Lambda}_d$ δεν αναμένεται να είναι χ^2 αλλά σταθμισμένη χ^2 .

Κάθε ένα από τα δύο μοντέλα χρησιμοποιήθηκε για την παραγωγή δεδομένων με σκοπό την εφαρμογή του ελέγχου χ^2 και του σταθμισμένου ελέγχου χ^2 και τη σύγκρισή τους. Χρησιμοποιήθηκαν οι τιμές $n = 100,200,300$ για το μέγεθος του



δείγματος, και οι τιμές $h = 5, 10, 15$ για το πλήθος των ζωνών. Για κάθε συνδυασμό τιμών n, h υπολογίστηκε το ποσοστό των περιπτώσεων απόρριψης της υπόθεσης $d = 0,1$ ύστερα από 1000 επαναλήψεις, δηλαδή η ισχύς του αντίστοιχου ελέγχου, τόσο για τον έλεγχο χ^2 όσο και για τον σταθμισμένο έλεγχο χ^2 . Υπολογίστηκε επίσης το αυτό ποσοστό για τον έλεγχο της υπόθεσης $d = 2$ που δεν είναι άλλο από το παρατηρούμενο επίπεδο σημαντικότητας του ελέγχου. Τρία ήταν τα βασικά ευρήματα. Πρώτον, η ισχύς του σταθμισμένου ελέγχου χ^2 ήταν πολύ μεγαλύτερη για όλε τις περιπτώσεις τιμών n, h . Δεύτερον, το παρατηρούμενο επίπεδο σημαντικότητας του ελέγχου χ^2 ήταν πολύ μικρότερο από αυτό του σταθμισμένου ελέγχου χ^2 γεγονός που θέτει θέμα αξιοπιστίας του ελέγχου χ^2 όταν δεν ισχύουν οι προϋποθέσεις εφαρμογής του. Τρίτον, το παρατηρούμενο επίπεδο σημαντικότητας του σταθμισμένου ελέγχου χ^2 αυξάνει με την αύξηση του h γεγονός που επαληθεύει την επισήμανση των Bura και Cook (2001a) περί μείωσης του αξιοπιστίας του ελέγχου λόγω της συνεπαγόμενης αύξησης της διακύμανσης των βαρών \hat{w}_k .

Σύγκριση ανάμεσα στους δύο ελέγχους έγινε για το μοντέλο (3.6) υποθέτοντας ότι $x \sim N_5(\mathbf{0}, I_5)$ για $n = 100$ και $h = 5, 10, 15$. Το αποτέλεσμα ήταν εντυπωσιακό. Ο σταθμισμένος έλεγχος χ^2 ήταν εξίσου ισχυρός με τον έλεγχο χ^2 παρά το γεγονός ότι η κανονικότητα του x καλύπτει πλήρως τις προϋποθέσεις εφαρμογής του ελέγχου χ^2 . Οι Bura και Cook (2001a) αποφαίνονται ότι το πλήθος h των ζωνών δεν πρέπει να υπερβαίνει το 5-7% του μεγέθους του δείγματος ώστε το παρατηρούμενο επίπεδο σημαντικότητας να μην υπερβαίνει το ονομαστικό.

3.1.5 γ ΠΑΡΑΛΛΑΓΕΣ ΤΗΣ SIR

Οι Hardle και Tsybakov (1991) πρότειναν μία παραλλαγή της SIR. Συγκεκριμένα πρότειναν τη χρήση της μήτρας

$$\mathbf{B} = E_y[E(z|y) E(z^T|y)]$$

αντί της $Var [E(z|\tilde{y})]$

Τα στοιχεία της \mathbf{B} είναι

$$b_{jk} = \int m_j(y)m_k(y)Fdy$$

όπου $m_j(y)$ είναι η καμπύλη παλινδρόμησης του y στην j συνιστώσα του z και F είναι η αθροιστική συνάρτηση κατανομής του y . Για την εκτίμηση του b_{jk} μπορούν



να χρησιμοποιηθούν η εμπειρική αθροιστική συνάρτηση κατανομής F_n καθώς και οι μη παραμετρικές εκτιμήσεις \hat{m}_j, \hat{m}_k των m_j, m_k οπότε

$$\hat{b}_{jk} = \int \hat{m}_j(y) \hat{m}_k(y) F_n dy = \frac{1}{n} \sum_{i=1}^n \hat{m}_j(y_i) \hat{m}_k(y_i)$$

Σύμφωνα με την προσέγγιση αυτή ο μέσος της δεσμευμένης κατανομής $x|y$ θα είναι

$$E(x|y) = c_1(y) \beta_1 + \dots + c_k(y) \beta_k$$

όπου $c_j(y)$ κάποιες συναρτήσεις, και $\beta_1, \beta_2, \dots, \beta_k$ τα διανύσματα βάσης του $S_{y|x}$ για τα οποία ισχύει

$$B = \sum_{j,m=1}^k \tilde{c}_{jm} \beta_j \beta_m^T$$

όπου $\tilde{c}_{jm} = E [c_j(y) c_m(y)]$

Άλλες παραλλαγές της SIR μπορεί να αναζητήσει ο αναγνώστης στους Schott (1994) και Velilla (1998).

Πριν την παρουσίαση των μεθόδων $SIR II$ και $SAVE$ οι οποίες είναι μέθοδοι ροπής δεύτερης τάξης και οι οποίες χρησιμοποιούνται για να αντιμετωπίσουν προβλήματα στα οποία η SIR αποτυγχάνει, παρουσιάζεται στην 3.2 η μέθοδος PIR απαλλαγμένη από κάποιες αδυναμίες της SIR .

3.2 ΜΕΘΟΔΟΣ PIR (PARAMETRIC INVERSE REGRESSION)

3.2.1 ΘΕΩΡΗΤΙΚΗ ΘΕΜΕΛΙΩΣΗ

Η μέθοδος PIR η οποία εισήχθη από τους Bura και Cook (2001b) χρησιμοποιεί όπως και η SIR την αντίστροφη παλινδρόμηση με σκοπό την εκτίμηση του κεντρικού υπόχωρου $S_{y|x}$ εκμεταλλευόμενη την ισχύ του θεωρήματος 3.1. Η διαφορά τους συνίσταται στο γεγονός ότι η PIR βασίζεται στην προσαρμογή παραμετρικών καμπυλών στα δεδομένα των p αντίστροφων παλινδρομήσεων σε αντίθεση με την SIR που είναι μία μη παραμετρική μέθοδος, καθώς και στο γεγονός ότι η PIR δεν κάνει καμία υπόθεση για την κατανομή του διανύσματος x των ανεξάρτητων μεταβλητών. Η προσαρμογή των καμπυλών αυτών γίνεται με την βοήθεια του μοντέλου της πολλαπλής γραμμικής παλινδρόμησης το οποίο εξασφαλίζει την ισχύ της υπόθεσης γραμμικότητα του θεωρήματος 3.1. Το μοντέλο στη γενική του μορφή εξηγεί την απόκριση $y = (y_1, \dots, y_m)^T$ μέσω του τυποποιημένου διανύσματος

$z = (z_1, \dots, z_p)^T$ των ανεξάρτητων μεταβλητών. Έστω



$$E \begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix} | y = (f_1(y), \dots, f_q(y)) \begin{bmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1p} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{q1} & \beta_{q2} & \dots & \beta_{qp} \end{bmatrix} \quad (3.8)$$

όπου f_i οποιεσδήποτε πραγματικές και γραμμικά ανεξάρτητες γνωστές συναρτήσεις του y . Έστω επίσης τυχαίο δείγμα μεγέθους n παρατηρήσεων (y, x) από το οποίο προκύπτει η $n \times m$ μήτρα Y_n των παρατηρήσεων για τις αποκρίσεις, και η $n \times p$ μήτρα X_n των παρατηρήσεων για τις ανεξάρτητες μεταβλητές.

Το μοντέλο που προσαρμόζεται έχει ως εξής

$$Z_n | Y_n = F_n B + E_n$$

όπου

$$Z_n = (z_{ij}) = \{X_n - E(X_n)\} \Sigma_x^{-1/2}$$

η $n \times p$ μήτρα των τυποποιημένων παρατηρήσεων για τις ανεξάρτητες μεταβλητές,

$$F_n = (\tilde{f}_{il}) \text{ } n \times q \text{ μήτρα με στοιχεία } \tilde{f}_{il} = \tilde{f}_l(y_i) = f_{il} - \sum_{l=1}^n f_{il}/n,$$

$B = (\beta_{ij})$ $q \times p$ μήτρα των συντελεστών της παλινδρόμησης του z στο y , και

E_n η $n \times p$ μήτρα των σφαλμάτων για την οποία υποτίθεται ότι οι γραμμές της είναι ανεξάρτητες με μέσο $\mathbf{0}$ και σταθερή μήτρα διακύμανσης Σ_{zy} .

Οι συνθήκες αυτές για την E_n εκφράζονται ως εξής:

$$E(E_n | Y_n) = \mathbf{0}$$

$$\text{Cov}\{\text{vec}(E_n) | Y_n\} = \Sigma_{zy} \otimes I_n$$

όπου Σ_{zy} η $p \times p$ θετικά ορισμένη και ανεξάρτητη του y μήτρα διακύμανσης και $\text{vec}(E_n)$ το διάνυσμα που σχηματίζεται από την διάταξη των στηλών της E_n ώστε η i στήλη να είναι κάτω από την $i-1$ στήλη.

Θεωρείται επίσης ότι $n \geq \max(p, q)$

Ισχύει ότι

$$S_{E(z,y)} = S(B^T F_n^T)$$

και δεδομένου επίσης ότι

$$\text{rank}(B^T F_n^T) = \text{rank}(F_n B)$$

και ότι

$$\text{rank}(F_n B) = \text{rank}(B^T F_n^T F_n B) = \text{rank}(B)$$



δοθέντος ότι $\mathbf{F}_n^T \mathbf{F}_n$ θετικά ορισμένη μήτρα (Seber (1977)), η εκτίμηση της διάστασης του $S_{E(z,y)}$ δεν είναι άλλη από την εκτίμηση της $rank(\mathbf{B})$.

Σύμφωνα με τα όσα έχουν ήδη εκτεθεί η εκτιμώμενη διάσταση d του $S_{E(z,y)}$ θα αποτελεί εκτίμηση της διάστασης του S_{zy} , ενώ η εκτίμηση της βάσης του S_{zy} θα προκύψει από το γινόμενο της \mathbf{F}_n και της μήτρας των d ιδιοδιανυσμάτων που αντιστοιχούν στις d μεγαλύτερες ιδιοτιμές της εκτίμησης του \mathbf{B} .

Η εκτίμηση της \mathbf{B} που χρησιμοποιείται για το σκοπό αυτό είναι η

$$\hat{\mathbf{B}}_{std} = \mathbf{G}_n^{-1/2} \tilde{\mathbf{B}}_n \hat{\Sigma}_{zy}^{-1/2}$$

όπου

$$\tilde{\mathbf{B}}_n = (\mathbf{F}_n^T \mathbf{F}_n)^{-1} \mathbf{F}_n^T \hat{\mathbf{Z}}_n$$

η δειγματική OLS εκτίμηση της \mathbf{B} ,

$$\hat{\mathbf{Z}}_n = (\mathbf{X}_n - \bar{\mathbf{X}}_n) \hat{\Sigma}_x^{-1/2}$$

η δειγματική εκτίμηση της \mathbf{Z}_n ,

$$\mathbf{G}_n = (\mathbf{F}_n^T \mathbf{F}_n / n)^{-1}$$

και

$$\hat{\Sigma}_{zy} = (n - q)^{-1} (\mathbf{Z}_n - \mathbf{F}_n \hat{\mathbf{B}}_n)^T (\mathbf{Z}_n - \mathbf{F}_n \hat{\mathbf{B}}_n)$$

συνεπής και αμερόληπτη εκτίμηση της Σ_{zy} .

Επίσης

$$\hat{\mathbf{B}}_n = (\mathbf{F}_n^T \mathbf{F}_n)^{-1} \mathbf{F}_n^T \mathbf{Z}_n$$

είναι η OLS εκτίμηση της \mathbf{B} .

Χρησιμοποιώντας την $\hat{\mathbf{B}}_{std}$ οι Bura και Cook (2001b)

εκμεταλλεύτηκαν το γεγονός ότι

$$n^{1/2} \cdot \hat{\mathbf{H}}_n^{-1/2} vec(\hat{\mathbf{B}}_n - \mathbf{B}) \xrightarrow{D} N_{pq}(\mathbf{0}, \mathbf{I}_p \otimes \mathbf{I}_q) \quad (3.9)$$

όπου

$$\hat{\mathbf{H}}_n = \hat{\Sigma}_{zy} \otimes (\mathbf{F}_n^T \mathbf{F}_n / n)^{-1}$$

υπό την προϋπόθεση ότι η

$$\mathbf{H}_n = \Sigma_{zy} \otimes (\mathbf{F}_n^T \mathbf{F}_n / n)^{-1}$$

τείνει ασυμπτωτικά σε κάποια θετικά ορισμένη μήτρα \mathbf{H} .



Η ισχύς της (3.9) που προκύπτει από εφαρμογή του θεωρήματος του Slutsky (Bunke και Bunke(1986)) τους επέτρεψε να διατυπώσουν το παρακάτω θεώρημα.

Θεώρημα 3.5 Έστω ότι ισχύει το μοντέλο (3.8), ότι η \mathbf{G}_n συγκλίνει σημειακά σε μία

θετικά ορισμένη μήτρα και ότι $\hat{\Sigma}_{z|y}$ συνεπής εκτίμηση της $\Sigma_{z|y}$. Έστω επίσης $\hat{\phi}_j$, $j = 1, \dots, \min(q, p)$ οι ιδιάζουσες τιμές (singular values) της $\hat{\mathbf{B}}_{std}$. Τότε η ασυμπτωτική κατανομή του

$$\Lambda_d = n \sum_{j=d+1}^{\min(q,p)} \hat{\phi}_j^2$$

είναι η χ^2 με $(q-d)$ ($p-d$) βαθμούς ελευθερίας.

Το στατιστικό Λ_d χρησιμοποιείται για τον έλεγχο υποθέσεων για το βαθμό της μήτρας \mathbf{B} με τον ίδιο ακριβώς τρόπο που παρουσιάζεται στην 3.1.3. Για παράδειγμα για τον έλεγχο της υπόθεσης $d=1$ συγκρίνεται το Λ_1 με τα εκατοστιαία σημεία της χ^2 κατανομής με $(q-1)$ ($p-1$) βαθμούς ελευθερίας.

3.2.1α ΠΕΡΙΠΤΩΣΗ ΜΗ ΣΤΑΘΕΡΗΣ ΔΙΑΚΥΜΑΝΣΗΣ

Στην περίπτωση αυτή η $\Sigma_{z|y}$ είναι συνάρτηση του y , ισχύει δηλαδή

$$Cov(z|y) = \Sigma_{z|y}(y) = [\sigma_{ij}(y)]_{i,j=1}^p$$

με

$$\hat{\sigma}_{ij}(y_k) = Cov(z_{ki}, z_{kj} | y = y_k), \quad k = 1, \dots, n \text{ και } i, j = 1, \dots, p.$$

Επομένως η $Cov\{vec(Z_n) | Y_n\}$, και κατά συνέπεια και η $Cov\{vec(E_n) | Y_n\}$, θα είναι μία *pxp* συμμετρική μπλόκ μήτρα αποτελούμενη από p^2 μπλόκ τάξεως *nxn*, της οποίας το ij -οστό μπλόκ θα είναι μία *nxn* διαγώνιος μήτρα με στοιχεία $\sigma_{ij}(y_1), \dots, \sigma_{ij}(y_n)$ κατά μήκος της διαγωνίου, για $i, j = 1, \dots, p$.

Η εκτίμηση $\hat{\mathbf{B}}_{std}$ της \mathbf{B} που χρησιμοποιείται στην περίπτωση αυτή για την εκτίμηση της διάστασης d του $S_{E(z|y)}$ είναι τέτοια ώστε

$$vec(\hat{\mathbf{B}}_{std}) = \hat{\mathbf{H}}_n^{-1/2} vec(W_n \hat{\mathbf{Z}}_n)$$

όπου

$$\mathbf{W}_n = (\mathbf{F}_n^T \mathbf{F}_n)^{-1} \mathbf{F}_n^T, \text{και}$$

$\hat{\mathbf{H}}_n$ $qpxqp$ μπλόκ μήτρα της οποίας το ij -οστό μπλοκ θα είναι

$$n\mathbf{W}_n = diag\{\hat{\sigma}_{ij}(y_1), \dots, \hat{\sigma}_{ij}(y_n)\}\mathbf{W}_n^T$$

με

$\hat{\sigma}_{ij}(y_n)$ συνεπή εκτίμηση του $\sigma_{ij}(y_n)$.

Χρησιμοποιώντας την $\hat{\mathbf{B}}_{std}$ οι Bura και Cook (2001b)

εκμεταλλεύτηκαν το γεγονός ότι

$$n^{1/2} \cdot \hat{\mathbf{H}}_n^{-1/2} vec(\mathbf{W}_n \mathbf{Z}_n - \mathbf{B}) \xrightarrow{D} N_{pq}(\mathbf{0}, \mathbf{I}_p \otimes \mathbf{I}_q) \quad (3.10)$$

κάτι που ισχύει υπό την προϋπόθεση ότι η \mathbf{H}_n που προκύπτει από την $\hat{\mathbf{H}}_n$ με $\sigma_{ij}(y_n)$ αντί του $\hat{\sigma}_{ij}(y_n)$, και η $\hat{\mathbf{H}}_n$ τείνουν ασυμπτωτικά στην ίδια θετικά ορισμένη μήτρα \mathbf{H} .

Η ισχύς της (3.10) που προκύπτει από εφαρμογή του θεωρήματος του Slutsky (Bunke και Bunke(1986)) επέτρεψε στους Bura και Cook (2001b) να διατυπώσουν το παρακάτω θεώρημα που είναι ανάλογο του θεωρήματος 3.5.

Θεώρημα 3.6 Έστω ότι $\hat{\mathbf{H}}_n$ είναι συνεπής εκτίμηση της \mathbf{H} και $d = rank(\mathbf{B}) = dim S_{E(z|y)}$. Έστω επίσης $\hat{\phi}_j, j=1, \dots, min(q,p)$ οι ιδιάζουσες τιμές της $\hat{\mathbf{B}}_{std}$ και ότι ισχύουν οι συνθήκες (a) –(c) του λήματος 1 (βλ. Bura και Cook (2001b) Appendix A)

Τότε η ασυμπτωτική κατανομή του

$$\Lambda_d = n \sum_{j=d+1}^{min(q,p)} \hat{\phi}_j^2$$

είναι η χ^2 με $(q-d)(p-d)$ βαθμούς ελευθερίας.

Όπως φαίνεται το κρίσιμο θέμα που αφορά την αντιμετώπιση της περίπτωσης μη σταθερής διακύμανσης είναι η δυνατότητα συνεπούς εκτίμησης της $\Sigma_{z|y}(y)$.

Η $\hat{\sigma}_{ij}(y)$ μπορεί να προκύψει ως εξής

$$\hat{\sigma}_{ij}(y) = cov_n(z_i, z_j | y) = \hat{E}_n(z_i z_j | y) - \hat{E}_n(z_i | y) \hat{E}_n(z_j | y) \text{ για } i,j=1, \dots, p \quad (3.11)$$

όπου





$\hat{E}_n(\cdot|y)$ η OLS εκτίμηση του μέσου της παλινδρόμησης του (\cdot) στο y . Η επιλογή του προσαρμοζόμενου μοντέλου γίνεται με βάση τα δεδομένα και δεδομένου ότι οι OLS εκτιμητές είναι συνεπείς το ίδιο θα ισχύει και για το $\hat{\sigma}_{ij}(y)$.

Η $p \times p$ μήτρα $\hat{\Sigma}_{z,y}(y)$ με στοιχεία $\hat{\sigma}_{ij}(y)$ υπολογιζόμενα από την (3.11) θα αποτελεί συνεπή εκτίμηση της $\Sigma_{z,y}(y)$.

Τέλος όπως επισημαίνουν οι Bura και Cook (2001b) η προϋπόθεση ότι η \hat{H}_n τείνει ασυμπτωτικά σε μία θετικά ορισμένη μήτρα H , ώστε να ισχύει η (3.10), θα ισχύει εφόσον

$$\text{Cov}\{\hat{\sigma}_j(y_k), \hat{\sigma}_l(y_l)\} \xrightarrow{n \rightarrow \infty} 0$$

εφόσον δηλαδή η $\Sigma_{z,y}$ είναι ασυμπτωτικά ανεξάρτητη του y .

3.2.2 ΑΛΓΟΡΙΘΜΟΣ

Για τον έλεγχο της υπόθεσης $d = j$ ακολουθούνται τα παρακάτω βήματα.

Βήμα 1: Επιλογή των συναρτήσεων $f_1(y), \dots, f_q(y)$ για τις p αντίστροφες παλινδρομήσεις με βάση τη *scatterplot matrix* των y, x . Η διαδικασία μπορεί να διευκολυνθεί με την προσαρμογή πολυωνυμικών συναρτήσεων μέσω software.

Υπολογισμός της μήτρας F_n .

Βήμα 2: Υπολογισμός της δειγματικής εκτίμησης $\hat{Z}_n = (X_n - \bar{X}_n)\hat{\Sigma}_x^{-1/2}$

Βήμα 3: Υπολογισμός της μήτρας \tilde{B}_n των δειγματικών OLS εκτιμήσεων των συντελεστών των αντίστροφων παλινδρομήσεων των \hat{z}_i , $i=1,\dots,p$ στο y .

Βήμα 4: Διερεύνηση για την ισχύ σταθερής διακύμανσης μέσω της *scatterplot matrix*.

Αν ισχύει τότε

(α) Υπολογισμός της $\hat{\Sigma}_{z,y}$ σαν μήτρα των καταλούπων της παλινδρόμησης του z στο y διαιρεμένων με $n-q$. Εν συνεχείᾳ υπολογισμός της $G_n = (F_n^T F_n / n)^{-1}$.

(β) Υπολογισμός της τυποποιημένης μήτρας $\hat{B}_{sd} = G_n^{-1/2} \tilde{B}_n \hat{\Sigma}_{z,y}^{-1/2}$

Αν δεν ισχύει τότε:

(α) Υπολογισμός των $\hat{\sigma}_j(y_k)$ μέσω της (11) για $k=1,\dots,n$



(β) Υπολογισμός της $W_n = (F_n^T F_n)^{-1} F_n^{T^T}$ και της $pqxqr$ μπλόκ μήτρας \hat{H}_n με ij -οστό μπλοκ $nW_n \text{diag}\{\hat{\sigma}_{ij}(y_1), \dots, \hat{\sigma}_{ij}(y_n)\}W_n^T$

(γ) Υπολογισμός της τυποποιημένης μήτρας \hat{B}_{std} για την οποία

$$\text{vec}(\hat{B}_{std}) = \hat{H}_n^{-1/2} \text{vec}(W_n \hat{Z}_n),$$

Βήμα 5: Υπολογισμός των ιδιαζουσών τιμών $\hat{\phi}_j$, της \hat{B}_{std} και του στατιστικού Λ_d για $d=j$.

Βήμα 6: Σύγκριση της τιμής του Λ_j με τα εκατοστιαία σημεία της $\chi_{(p-j)(q-j)}^2$. Εάν το Λ_j είναι μικρότερο τότε $d=j$, ενώ εάν είναι μεγαλύτερο τότε τίθεται $j=j+1$ και επαναλαμβάνεται η διαδικασία.

Τέλος όπως ήδη αναφέρθηκε το γινόμενο των d διανυσμάτων της \hat{B}_{std} , τα οποία αντιστοιχούν στις d μεγαλύτερες ιδιοτιμές της \hat{B}_{std} (όπου $d = \dim S_{E(z,y)}$), με την F_n δίνει την εκτίμηση της βάσης του $S_{y|x}$. Πολλαπλασιασμός της βάσης του $S_{y|x}$ με $\hat{\Sigma}_x^{-1/2}$ από αριστερά δίνει την εκτίμηση της βάσης του $S_{y|x}$.

3.2.3 ΕΦΑΡΜΟΓΕΣ

Στη συνέχεια παρατίθεται η σύγκριση της ισχύος των δύο στατιστικών ελέγχων που χρησιμοποιούνται για την εκτίμηση της διάστασης d του κεντρικού υπόχωρου $S_{y|x}$, οι οποίοι βασίζονται ο μεν ένας εξ' αυτών στη *SIR* ο δε άλλος στην *PIR*. Η σύγκριση γίνεται μέσω τριών μοντέλων από τα οποία προέκυψαν δεδομένα μέσω προσομοίωσης. Χρησιμοποιούνται τρία μεγέθη δείγματος $n=50, 100, 250$. Για κάθε μέγεθος δείγματος και κατανομή του x οι παρατηρούμενες τιμές για την ισχύ και το επίπεδο σημαντικότητας των αντιστοίχων ελέγχων προέκυψαν ύστερα από 1000 επαναλήψεις. Ο βαθμός του πολυωνύμου που προσαρμόζεται προέκυψε με βάση την εικόνα της *scatterplot matrix* και με βάση την προσαρμογή των σχετικών καμπυλών μέσω του Arc (Cook and Weisberg 1999).

Το πρώτο μοντέλο είναι μονοδιάστατο και έχει ως εξής

$$y = x_1 + x_2 + x_4 + 0.5 \varepsilon, \quad \varepsilon \sim N(0, 1) \quad (3.12)$$

Για την κατανομή του x εξετάζονται δύο περιπτώσεις

(α) $x \sim N(0, I_4)$ και

(β) $x \sim \text{Pearson II}$ με παραμέτρους $m = -0.5$ και $\Sigma = I_4$ (Johnson 1987)

Η κατανομή Pearson II ανήκει στην οικογένεια των κατανομών με ελλειπτική συμμετρία των καμπυλών ίσης πιθανότητας και επομένως ικανοποιεί τη γνωστή συνθήκη γραμμικότητας που αποτελεί προϋπόθεση για την εφαρμογή τόσο της *SIR* όσο και της *PIR*.

Οι τιμές των πινάκων (3.4), (3.5) για L_o , L_i , αντιστοιχούν στο ποσοστό των περιπτώσεων που η υπόθεση $d=0$ απορρίφθηκε (χωρίς βέβαια να ισχύει). Αποτελούν δηλαδή τιμές της παρατηρούμενης ισχύος του αντίστοιχου ελέγχου για ονομαστικό επίπεδο σημαντικότητας 0.05. Οι τιμές σε παρένθεση αντιστοιχούν σε ονομαστικό επίπεδο σημαντικότητας 0.01.

Οι τιμές των πινάκων για L_l , A_l αντιστοιχούν στο ποσοστό των περιπτώσεων που η υπόθεση $d = 1$ απορρίφθηκε με τη διαφορά ότι η υπόθεση γνωρίζουμε ότι ισχύει. Αποτελούν δηλαδή τιμές του παρατηρούμενου επιπέδου σημαντικότητας με βάση το ονομαστικό επίπεδο σημαντικότητας 0.05. Οι τιμές σε παρένθεση αντιστοιχούν σε ονομαστικό επίπεδο σημαντικότητας 0.01.

Τα κενά του πίνακα (3.5) για πολυώνυμο 1^{ου} βαθμού οφείλονται στο γεγονός ότι στην περύπτωση πολυωνύμου 1^{ου} βαθμού η διάσταση του προβλήματος δεν μπορεί να είναι μεγαλύτερη από $d=1$. Αντίθετα για περιπτώσεις πολυωνύμων 2^{ου}, 3^{ου} και 4^{ου} βαθμού η διάσταση του προβλήματος θα μπορούσε (εσφαλμένα βέβαια) να προκύψει μεγαλύτερη από $d=1$.

Το σύμβολο H αφορά το πλήθος των ζωνών για την *SIR*.

Το σημαντικό εύρημα που προκύπτει από τους πίνακες είναι ότι για αυξημένες τιμές του H και ιδιαίτερα όταν $x \sim \text{Pearson II}$ και το μέγεθος του δείγματος είναι μικρό η ισχύς του ελέγχου της *SIR* φθίνει και είναι μικρότερη από την ισχύ του ελέγχου της *PIR*.

Επίσης σημαντικό είναι και το γεγονός ότι όταν $x \sim \text{Pearson II}$ και $n = 100, 250$ τα παρατηρούμενα επίπεδα σημαντικότητας του ελέγχου της *PIR* για πολυώνυμο 2^{ου} βαθμού είναι σαφώς μικρότερα των αντιστοίχων της *SIR*.

Results for normal x			Results for Pearson II x				
	$H = 5$	$H = 10$	$H = 15$		$H = 5$	$H = 10$	$H = 15$
$n=50$							
L_o	1.0(1.0)	1.0 (0.975)	0.964(0.640)	0.993(0.968)	0.941(0.687)	0.756(0.350)	
L_l	0.053 (0.009)	0.036(0.008)	0.025(0.003)	0.063(0.010)	0.041(0.006)	0.031(0.004)	
$n=100$							
L_o	1.0(1.0)	1.0(1.0)	1.0(1.0)	0.999(0.999)	0.997(0.996)	1.0(0.993)	
L_l	0.053(0.009)	0.053(0.008)	0.036(0.007)	0.068(0.013)	0.052(0.005)	0.045(0.009)	
$n=250$							
L_o	1.0(1.0)	1.0(1.0)	1.0(1.0)	1.0(1.0)	1.0(0.999)	0.999(0.999)	
L_l	0.053(0.011)	0.042(0.005)	0.047(0.008)	0.064(0.015)	0.053(0.008)	0.055(0.014)	

Πίνακας 3.4 τιμές της παρατηρούμενης ισχύος(L_0) και του παρατηρούμενου επιπέδου σημαντικότητας(L_1) του κατά SIR ελέγχου για την εκτίμηση της διάστασης του κεντρικού υπόχωρου για τα δεδομένα του μοντέλου 3.12.(σε παρένθεση τιμές για ονομαστικό επίπεδο σημαντικότητας 0.01)

Results for normal x					Results for Pearson II x	
	Degree1	Degree2	Degree3	Degree4	Degree1	Degree2
n=50						
Λ_0	1.0(1.0)	1.0(1.0)	1.0(1.0)	1.0(1.0)	1.0(0.999)	0.999(0.999)
Λ_1		0.087(0.022)	0.025(0.007)	0.061(0.022)		0.037(0.013)
n=100						
Λ_0	1.0(1.0)	1.0(1.0)	1.0(1.0)	1.0(1.0)	1.0(1.0)	1.0(1.0)
Λ_1		0.068(0.017)	0.025(0.005)	0.024(0.003)		0.033(0.003)
n=250						
Λ_0	1.0(1.0)	1.0(1.0)	1.0(1.0)	1.0(1.0)	1.0(1.0)	1.0(1.0)
Λ_1		0.052(0.005)	0.059(0.014)	0.061(0.016)		0.02(0)

Πίνακας 3.5 τιμές της παρατηρούμενης ισχύος(L_0) και του παρατηρούμενου επιπέδου σημαντικότητας(L_1) του ελέγχου για την εκτίμηση της $rank(B)$ για τα δεδομένα του μοντέλου 3.12.(σε παρένθεση τιμές για ονομαστικό επίπεδο σημαντικότητας 0.01).

Το δεύτερο μοντέλο είναι διδιάστατο και έχει ως εξής:

$$y = x_1 (x_2 + x_4 + 1) + 0.5 \varepsilon, \quad \varepsilon \sim N(0,1) \quad (3.13)$$

Για την κατανομή του x θεωρείται ότι $x \sim N(\mathbf{0}, I_4)$.

Ο πίνακας (3.6) που ακολουθεί παρέχει ένα σημαντικό εύρημα. Η μέθοδος SIR αποτυγχάνει να διαγνώσει την πραγματική διάσταση του προβλήματος που είναι $d=2$ με πιθανότητα 0.94 (όταν $n=50$ και $H=15$). Η μέθοδος παρουσιάζεται ισχυρή για $n=250$ και μικρό αριθμό ζωνών H , ενώ για την ακρίβεια της μεθόδου θα απαιτείτο σαφώς μεγαλύτερο μέγεθος δείγματος. Αντίθετα η PIR όπως φαίνεται από τον πίνακα



(3.7) είναι σαφώς πιο ισχυρή ακόμα και για μικρά δείγματα και μεγάλο βαθμό πολυωνύμου.

Results for the following values of H :				
	5	10	15	20
n=50				
L_0	0.677(0.486)	0.589 (0.319)	0.403(0.161)	
L_1	0.163 (0.049)	0.123(0.028)	0.062(0.009)	
L_2	0.005 (0.001)	0.010 (0)	0.005 (0)	
n=100				
L_0	0.951(0.885)	0.919(0.819)	0.879 (0.743)	
L_1	0.444(0.244)	0.357 (0.164)	0.305 (0.121)	
L_2	0.018 (0.004)	0.023(0.004)	0.013 (0.003)	
n=250				
L_0	1.0(0.999)	1.0(1.0)	1.0(0.998)	0.998(0.99)
L_1	0.934(0.832)	0.922(0.821)	0.844 (0.697)	0.705 (0.477)
L_2	0.038(0.004)	0.036 (0.001)	0.038 (0.005)	0.026 (0.002)

Πίνακας 3.6 τιμές της παρατηρούμενης ισχύος(L_0, L_1) και του παρατηρούμενου επιπέδου σημαντικότητας(L_2) του κατά SIR ελέγχου για την εκτίμηση της διάστασης του κεντρικού υπόχωρου για τα δεδομένα του μοντέλου 3.13.(σε παρένθεση τιμές για ονομαστικό επίπεδο σημαντικότητας 0.01).

Results for the following values of n and degrees :						
	$n = 50$		$n = 100$		$n = 250$	
	Degree 3	Degree 4	Degree 3	Degree 4	Degree 3	Degree 4
Λ_0	0.972 (0.935)	0.974 (0.938)	0.998 (0.994)	0.998 (0.993)	1.0 (1.0)	1.0 (1.0)
Λ_1	0.575 (0.388)	0.577 (0.392)	0.810 (0.668)	0.797 (0.643)	0.988 (0.969)	0.985 (0.966)
Λ_2	0.032 (0.006)	0.024 (0.007)	0.047 (0.008)	0.046 (0.009)	0.057 (0.008)	0.049 (0.007)

Πίνακας 3.7 τιμές της παρατηρούμενης ισχύος(Λ_0, Λ_1) και του παρατηρούμενου επιπέδου σημαντικότητας(Λ_2) του ελέγχου για την εκτίμηση της rank (B) για τα δεδομένα του μοντέλου 3.13.(σε παρένθεση τιμές για ονομαστικό επίπεδο σημαντικότητας 0.01).

Το τρίτο μοντέλο είναι το μοντέλο (3.4)-(3.5) της 3.1.5β, το οποίο είναι διδιάστατο. Η κατανομή του διανύσματος x ικανοποιεί την συνθήκη γραμμικότητας όπως αποδεικνύει ο Velilla (1998) (p.1092-1093) παρά το γεγονός ότι δεν παρουσιάζει καμπύλες ίσης πιθανότητας με ελλειπτική συμμετρία.

Ο πίνακας (3.8) που ακολουθεί δείχνει ότι η SIR εκτιμά τη διάσταση d του κεντρικού υπόχωρου ίση με 1 για όλες τις περιπτώσεις μεγεθών δείγματος και πλήθους ζωνών.

Αντίθετα, όπως φαίνεται από τον πίνακα (3.9), η *PIR* έχει παρατηρούμενο επίπεδο σημαντικότητας που κυμαίνεται από 56.2% για $n = 50$, ονομαστικό επίπεδο σημαντικότητας 0.01 και πολυώνυμο 3^ο βαθμού, έως 90.4% για $n = 250$, ονομαστικό επίπεδο σημαντικότητας 0.05 και πολυώνυμο 4^ο βαθμού.

Results for the following values of H :				
	5	10	15	20
$n=50$				
L_0	0.996(0.996)	0.951 (0.718)	0.831(0.5)	
L_1	0.039 (0.009)	0.075(0.014)	0.085 (0.019)	
L_2	0.001 (0)	0.002 (0)	0.007 (0.001)	
$n=100$				
L_0	1.0(1.0)	1.0(1.0)	1.0(1.0)	
L_1	0.032(0.008)	0.079 (0.026)	0.119 (0.035)	
L_2	0.001 (0)	0 (0)	0.007 (0)	
$n=250$				
L_0	1.0(1.0)	1.0(1.0)	1.0(1.0)	1.0(1.0)
L_1	0.049(0.014)	0.13(0.064)	0.164 (0.071)	0.183 (0.9)
L_2	0.002 (0)	0.004 (0)	0.007 (0)	0.011 (0.003)

Πίνακας 3.8 τιμές της παρατηρούμενης ισχύος (L_0, L_1) και του παρατηρούμενου επιπέδου σημαντικότητας (L_2) του κατά *SIR* ελέγχου για την εκτίμηση της διάστασης του κεντρικού υπόχωρου για τα δεδομένα του μοντέλου 3.4-3.5. (σε παρένθεση τιμές για ονομαστικό επίπεδο σημαντικότητας 0.01)

Results for the following values of n and degrees :						
$n = 50$		$n = 100$		$n = 250$		
Degree 3	Degree 4	Degree 3	Degree 4	Degree 3	Degree 4	Degree 4
Λ_0	1.0(1.0)	1.0(1.0)	1.0(1.0)	1.0(1.0)	1.0 (1.0)	1.0 (1.0)
Λ_1	0.726 (0.562)	0.808 (0.668)	0.786 (0.664)	0.84 (0.732)	0.871 (0.78)	0.904 (0.836)
Λ_2	0.113 (0.021)	0.182 (0.053)	0.087 (0.02)	0.149 (0.051)	0.092 (0.027)	0.137 (0.048)

Πίνακας 3.9 τιμές της παρατηρούμενης ισχύος (Λ_0, Λ_1) και του παρατηρούμενου επιπέδου σημαντικότητας (Λ_2) του ελέγχου για την εκτίμηση της *rank (B)* για τα δεδομένα του μοντέλου 3.4-3.5. (σε παρένθεση τιμές για ονομαστικό επίπεδο σημαντικότητας 0.01).

Συμπερασματικά όπως αναφέρουν οι Bura και Cook (2001b) επικαλούμενοι και άλλες δοκιμές τις οποίες δεν παραθέτουν, η *PIR* εμφανίζεται τουλάχιστον το ίδιο ισχυρή με τη *SIR* για μοντέλα με ανεξάρτητες μεταβλητές κανονικά ή περίπου κανονικά κατανεμημένες. Επί πλέον η *PIR* (και για την περίπτωση της μη σταθερής διακύμανσης) εμφανίζεται πολύ πιο ισχυρή από τη *SIR* για σύνθετα μοντέλα ή μοντέλα με ανεξάρτητες μεταβλητές των οποίων η κατανομή απέχει πολύ από την

κανονική. Ωστόσο η μέθοδος *PIR* για μη σταθερή διακύμανση θα πρέπει να ακολουθείται μόνο όταν η παραβίαση της υπόθεσης σταθερής διακύμανσης είναι σοβαρή, δεδομένου ότι απαιτεί ένα επιπλέον στάδιο εκτίμησης που αφορά την εκτίμηση των $\sigma_{ij}(y_n)$.

3.2.4 ΣΧΟΛΙΑ- ΕΠΙΣΗΜΑΝΣΕΙΣ

Όπως είδαμε η μέθοδος *PIR* κατατείνει στην εκτίμηση του κεντρικού υπόχωρου μέσω της προσαρμογής γραμμικών μοντέλων στις αντίστροφες παλινδρομήσεις του x στο y . Η μέθοδος περιλαμβάνει έναν έλεγχο χ^2 για την εκτίμηση της διάστασης d του $S_{E(x,y)}$. Η δομή του ελέγχου απορρέει από το γεγονός ότι η *OLS* εκτίμηση της μήτρας B των συντελεστών των αντίστροφων παλινδρομήσεων κατανέμεται ασυμπωτικά κανονικά. Υπό προϋποθέσεις η μέθοδος επεκτείνεται και στην περίπτωση της μη σταθερής διακύμανσης, ενώ δεν υπάρχει κανένας περιορισμός για την κατανομή του x . Παρά το γεγονός ότι η *PIR* βασίζεται σε προσαρμογή καμπυλών η οποία μπορεί να μην είναι τέλεια, η προσαρμογή αυτή προκύπτει από το σύνολο των δεδομένων. Αντίθετα η *SIR* και οι παραλλαγές της (Schott (1994), Velilla (1998)) βασίζονται στην προεπιλογή της παραμέτρου H η οποία γίνεται μάλλον αυθαίρετα χωρίς να λαμβάνονται υπόψιν τα δεδομένα.

Επομένως δεν προκαλεί έκπληξη το γεγονός ότι η ισχύς του ελέγχου χ^2 της *PIR* είναι μεγαλύτερη αυτού της *SIR*. Αυτό συμβαίνει ακριβώς λόγω απουσίας αυθαίρετων επιλογών και λόγω χρήσης της μεθόδου *OLS* για την εκτίμηση της μήτρας B των συντελεστών των αντίστροφων παλινδρομήσεων η οποία μέθοδος αποφέρει εκτιμήσεις με βέλτιστες ιδιότητες.

Ωστόσο δεν πρέπει να μας διαφεύγει το γεγονός ότι η *PIR* είναι προτιμητέα όταν τα διαγράμματα των αντίστροφων παλινδρομήσεων είναι ικανά να κατευθύνουν την επιλογή κατάλληλων καμπυλών προσαρμογής στα δεδομένα. Αντίθετα όταν αυτό δεν συμβαίνει και η επιλογή των καμπυλών είναι αμφίβολη, η *SIR* αποτελεί ενδεχομένως καλύτερη επιλογή.

Ο αλγόριθμος της μεθόδου *PIR* μπορεί να χρησιμοποιηθεί στο περιβάλλον του προγράμματος Arc (Cook και Weisberg 1999) μέσω κώδικα ο οποίος είναι διαθέσιμος στη διεύθυνση <http://gwis2/corc/gwn.edu/~ebura/publications.html>. Το Arc είναι διαθέσιμο στην διεύθυνση www.stat.nmn.edu/arc.



ΚΕΦΑΛΑΙΟ 4

ΜΕΘΟΔΟΙ ΡΟΠΗΣ ΔΕΥΤΕΡΗΣ ΤΑΞΗΣ

Στο κεφάλαιο αυτό εξετάζονται τρεις μέθοδοι ροπής δεύτερης τάξης. Η *SAVE* (*Sliced Average Variance Estimation*) και η *SIR II* οι οποίες βασίζονται στην ροπή δεύτερης τάξης της $x|y$ και η *pHd* (*Principal Hessian Directions*) η οποία βασίζεται στη συνδιακύμανση μεταξύ της απόκρισης y (ή των κατάλοιπων) και του γινομένου zz^T των τυποποιημένων ανεξάρτητων μεταβλητών x .

4.1 ΜΕΘΟΔΟΣ *SAVE* (*SLICED AVERAGE VARIANCE ESTIMATION*)

4.1.1. ΘΕΩΡΗΤΙΚΗ ΘΕΜΕΛΙΩΣΗ

Η μέθοδος αυτή εισήχθη από τους Cook και Weisberg (1991) και κάνει χρήση της αντίστροφης ροπής δεύτερης τάξης για την αντιμετώπιση προβλημάτων στα οποία η αντίστροφη ροπή πρώτης τάξης είναι σε θέση να ανασύρει μέρος μόνο της πληροφορίας που ενσωματώνει ο κεντρικός υπόχωρος. Για την εφαρμογή της μεθόδου απαιτείται η ισχύς των προϋποθέσεων 1.2 του παρακάτω θεωρήματος (Cook (1998)).

Θεώρημα 4.1 Έστω ότι οι στήλες της μήτρας γ αποτελούν βάση του S_{yz} . Εάν

$$1. E(z|\gamma^T z) = P_\gamma z$$

$$2. Var(z|\gamma^T z) = I_p - P_\gamma$$

όπου P_γ είναι ο *projection operator* για τον S_{yz} , και I_p η μοναδιαία μήτρα τάξεως p τότε

$$S(I_p - \Sigma_{zy}) \subset S_{yz}$$

όπου $\Sigma_{zy} = Var(z|y)$.

Όπως παρατηρεί ο Cook (1998) η ισχύς της γνωστής συνθήκης γραμμικότητας του θεωρήματος 3.1 συνεπάγεται την ισχύ της 1, ενώ όπως ήδη ειπώθηκε αν ισχύει η 1 και η Σ_{zy} είναι σταθερή τότε θα ισχύει και η 2. Επίσης η 2 θα ισχύει με ικανοποιητική προσέγγιση για πολλές περιπτώσεις κατανομής ανεξάρτητων μεταβλητών με ελλειπτική συμμετρία. Οι 1 και 2 μπορούν να ελεγχθούν μέσω της *scatterplot matrix*.



Οι Cook και Weisberg (1991) προτείνουν την εκτίμηση της βάσης του κεντρικού υπόχωρου μέσω των ιδιοδιανυσμάτων που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές της μήτρας.

$$SAVE = \sum_h (I - var(z | y \in I_h))^2$$

4.1.2 ΑΛΓΟΡΙΘΜΟΣ

Η διαδικασία είναι ανάλογη με την αντίστοιχη για τη *SIR* και εν μέρει περιγράφεται από τον Cook (2000).

Εστω n_s ο αριθμός των παρατηρήσεων για την απόκριση y στην ζώνη s , και έστω $\hat{z}_i = \hat{\Sigma}_x^{-1/2} (x_i - \bar{x})$, $i=1,2,\dots,n$ οι τυποποιημένες δειγματικές τιμές των παρατηρήσεων για τις ανεξάρτητες μεταβλητές x όπου $\hat{\Sigma}_x$ η δειγματική εκτίμηση της $\Sigma_x = Var(x)$ και \bar{x} ο δειγματικός μέσος των x .

Βήμα 1: Υπολογισμός σε κάθε ζώνη του δειγματικού μέσου των \hat{z}_i

$$\bar{z}_s = \frac{1}{n_s} \sum_{y_i \in J_s} \hat{z}_i$$

Βήμα 2: Υπολογισμός σε κάθε ζώνη της δειγματικής μήτρας συνδιακύμανσης

$$\hat{\Sigma}_{zs} = Cov(z | \tilde{y} = s) = \frac{1}{n_s - 1} \sum_{y_i \in J_s} z_i z_i^T - n_s \bar{z}_s \bar{z}_s^T$$

Βήμα 3: Υπολογισμός των ιδιοδιανυσμάτων $\hat{u}_1, \dots, \hat{u}_p$ που αντιστοιχούν στις ιδιοτιμές

$$\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p$$

$$\hat{M}_{save} = \sum_{s=1}^h n_s (I - \hat{\Sigma}_{zs})^2$$

Βήμα 4: Εστω $d = \dim S_{I-var(z|y)}$, τότε η *SAVE* εκτίμηση του $S_{I-var(z|y)}$ θα είναι

$$\hat{S}_{I-var(z|y)} = S(\hat{u}_1, \dots, \hat{u}_d)$$

και η *SAVE* εκτίμηση του $S_{y|x}$ θα είναι

$$\hat{S}_{y|x} = \hat{\Sigma}_x^{-1/2} \hat{S}_{I-var(z|y)} = S(\hat{\Sigma}_x^{-1/2} \hat{u}_1, \dots, \hat{\Sigma}_x^{-1/2} \hat{u}_d)$$

4.1.3 ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΤΑΣΗΣ d

Για την εκτίμηση της διάστασης του κεντρικού υπόχωρου οι Cook και Weisberg (1991) προτείνουν τη διαδικασία ελέγχου διατάξεων (*permutation test*). Σύμφωνα με



τη διαδικασία αυτή η οποία περιγράφεται από τον Cook (2000), για $d = 0$ συγκρίνεται

η τιμή του στατιστικού $\hat{L}_o = n \cdot \sum_{i=1}^P \hat{\lambda}_j$, η οποία προκύπτει με βάση τα δεδομένα, με τις

τιμές του στατιστικού αυτού οι οποίες προκύπτουν με βάση c τυχαίες διατάξεις των n τιμών της απόκρισης y . Η τιμή του p -value δεν είναι άλλη από το ποσοστό των τιμών του \hat{L}_o με βάση τις διατάξεις των τιμών του y που ξεπερνούν την τιμή του στατιστικού με βάση τα δεδομένα.

Για τον έλεγχο της υπόθεσης $d = 1$ ακολουθείται ανάλογη διαδικασία (Cook (2000)) με τη διαφορά ότι χρησιμοποιούνται δεδομένα που προκύπτουν από διατάξεις των δεικτών του ζεύγους $(y_i, \hat{\beta}_1^T x_i)$, όπου $\hat{\beta}_1 = \hat{\Sigma}^{-1/2} \hat{\lambda}_1$ όπου $\hat{\lambda}_1$ η μεγαλύτερη ιδιοτιμή της μήτρας \hat{M}_{save} . Για τον έλεγχο της υπόθεσης $d = d_o$ χρησιμοποιούνται δεδομένα που προκύπτουν από διατάξεις των δεικτών του διανύσματος $(y_i, \hat{\beta}_1^T x_i, \dots, \hat{\beta}_{d_o}^T x_i)$, $i = 1, \dots, n$ όπου $\hat{\beta}_{d_o} = \hat{\Sigma}^{-1/2} \hat{\lambda}_{d_o}$ και $\hat{\lambda}_{d_o}$ μεγαλύτερη ιδιοτιμή της μήτρας \hat{M}_{save} .

Αξιοσημείωτο πάντως είναι το γεγονός ότι αν το x κατανέμεται κανονικά η εκτίμηση της d με βάση τη διαδικασία που μόλις περιγράφηκε συμπίπτει πολύ συχνά με την εκτίμηση μέσω του ασυμπτωτικού ελέγχου χ^2 που βασίζεται στο θεώρημα 3.3.

4.1.4 ΕΦΑΡΜΟΓΕΣ

Στην εφαρμογή που ακολουθεί έγινε χρήση του μοντέλου

$$y = (\mu + 0.7071 z_1 + 0.7071 z_2)^2 \quad (4.1)$$

για διάφορες τιμές του μ , όπου $z_1, z_2 \text{ iid } N(0,1)$ μεταβλητές για τις οποίες ελήφθησαν 120 παρατηρήσεις. Για τα δεδομένα που παρήχθησαν έγινε χρήση αφενός της *SIR* και αφετέρου της *SAVE* με σκοπό την εκτίμηση του κεντρικού υπόχωρου. Ο πίνακας (4.1) που ακολουθεί δίνει τη γωνία σε μοίρες που σχηματίζεται ανάμεσα στο ιδιοδιάνυσμα που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή για κάθε μία από τις δύο μεθόδους και στην κατεύθυνση $n^T = (1, 1)$ που αποτελεί την πραγματική βάση του κεντρικού υπόχωρου.



μ	<i>SIR</i>	<i>SAVE</i>	<i>pHd</i>
0	87.82	0.74	8.90
0.25	13.04	1.79	12.92
0.50	7.15	1.97	6.93
1	4.20	1.32	18.19
2	2.00	1.60	13.84
4	0.19	0.71	21.31
8	0.56	0.81	0.93
100	0.03	0.27	33.46

Πίνακας 4.1 τιμές της γωνίας(σε μοίρες) ανάμεσα στο ιδιοδιάνυσμα για τη μεγαλύτερη ιδιοτιμή με βάση τις μεθόδους *SIR*,*SAVE* ,*pHd* ,και την πραγματική βάση.

Όπως φαίνεται η *SIR* αποτυγχάνει –όπως άλλωστε αναμενόταν – για μικρές τιμές του μ για τις οποίες η y είναι τετραγωνική συνάρτηση του $n^T z$, βελτιώνονται όμως αισθητά για μεγάλες τιμές του μ για τις οποίες η y τείνει να είναι γραμμική συνάρτηση του $n^T z$.

Αντίθετα η *SAVE* έχει πολύ καλή επίδοση ανεξαρτήτως τιμής του μ .

4.1.5. ΣΧΟΛΙΑ – ΕΠΙΣΗΜΑΝΣΕΙΣ

Οι μέθοδοι *SIR* και *PIR* προϋποθέτουν την ισχύ της υπόθεσης γραμμικότητας $E(z|y^T z) = P_y z$. Η μέθοδος *SAVE* προϋποθέτει την ισχύ μίας επιπλέον υπόθεσης που είναι η υπόθεση της σταθερής διακύμανσης $Var(z|y^T z) = I_p - P_y$. Επομένως η προσέγγιση της πολυδιάστατης κανονικότητας φαίνεται να είναι πιο επιτακτική. Ας επαναλάβουμε λοιπόν τις δύο επικρατέστερες μεθόδους που υπάρχουν. Η πρώτη προτάθηκε από τους Cook και Nachtheim (1994) και είναι μία μέθοδος στάθμισης του διανύσματος x ώστε η εμπειρική συνάρτηση κατανομής του να προσεγγίζει μία πολυμεταβλητή κανονική κατανομή στόχο με μήτρα διακύμανσης $\sigma^2 I$ όπου σ^2 κάποια προκαθορισμένη τιμή μεταξύ 0.5 και 1. Τα εν λόγω βάρη καθορίζονται μέσω ενός αλγορίθμου Monte Carlo ο οποίος διατίθεται για χρήση μέσω του Arc και ειδικότερα μέσω του αρχείου Reweight.lsp που πρέπει να φορτώνεται πριν από το φόρτωμα του αρχείου που περιέχει τα προς επεξεργασία δεδομένα. Η μέθοδος περιγράφεται συνοπτικά στον Cook (1998, p.185-190).

Η δεύτερη μέθοδος είναι η μέθοδος Box - Cox για την οποία έγινε λόγος στο δεύτερο κεφάλαιο καθώς και στην 3.1.5b. Η χρήση της συνιστάται σε περίπτωση εμφανούς μέσω της *scatterplot matrix* καμπυλότητας ή/και ετεροσκεδαστικότητας , ενώ η αποτελεσματικότητά της ελέγχεται και πάλι μέσο της *scatterplot matrix*. Ωστόσο, πρέπει να επισημανθεί ότι ο έλεγχος αυτός δεν είναι επαρκής για λόγους που ήδη έχουν εκτεθεί, φαίνεται ωστόσο να έχει ιδιαίτερη πρακτική αξία όπως επισημαίνει ο Cook (2000). Οι Cook και Weisberg (1999, section 19.4) παραθέτουν ένα παράδειγμα εφαρμογής της μεθόδου μεταξύ και άλλων τέτοιων παραδειγμάτων.

Τέλος να σημειωθεί, όπως αναφέρει ο Cook (2000), ότι κατευθύνσεις στις οποίες η $E(y|x)$ είναι γραμμική ή η $Var(x|y)$ δεν είναι σταθερή ανιχνεύονται ευκολότερα μέσω της *SIR*, κάτι άλλωστε που επαληθεύεται και μέσω του πίνακα 4.1.

Αυτό σημαίνει, όπως επισημαίνει ο Cook (2000), ότι οι *SIR* και *SAVE* πρέπει να εφαρμόζονται από κοινού και να αξιοποιούνται συμπληρωματικά, ώστε δηλαδή η μία από αυτές να ανιχνεύει κάποιες από τις σημαντικές κατευθύνσεις του κεντρικού υπόχωρου, και η άλλη τις υπόλοιπες. Επίσης η *SAVE* (όπως και η *SIR*) φαίνεται να εκτιμούν με ιδιαίτερη αξιοπιστία κατευθύνσεις του κεντρικού υπόχωρου σε προβλήματα με κατηγορική απόκριση. Πριν την παρουσίαση της μεθόδου *rHd* η οποία διαφοροποιείται ως προς τα δεδομένα τα οποία χρησιμοποιεί ακολουθεί στην 4.2 η παρουσίαση της μεθόδου *SIR II* που αποτελεί γενίκευση της *SAVE*.

4.2 ΜΕΘΟΔΟΣ *SIR II*

4.2.1 ΘΕΩΡΗΤΙΚΗ ΘΕΜΕΛΙΩΣΗ

Η μέθοδος αυτή προτάθηκε από τον Li(1991b) και βασίζεται σε ένα συμπέρασμα του ίδιου για τη μέθοδο *SIR* (Li(1991a)). Συγκεκριμένα ο Li(1991a) διαπίστωσε ότι αν το διάνυσμα x κατανέμεται κανονικά τότε σε οποιαδήποτε επίπεδο $b^T x$ κάθετο στα επίπεδα $\beta_i^T x$, $i = 1, \dots, k$, όπου β_i τα διανύσματα βάσης του κεντρικού υπόχωρου, θα είναι $Var(b^T x | y)$ σταθερή και επομένως

$$Var(b^T x | y) - E[Var(b^T x | y)] = 0.$$

Αναζητώντας έτσι τις διευθύνσεις στις οποίες μεγιστοποιείται το συναρτησιακό $Var(b^T x | y) - E[Var(b^T x | y)]$ πρότεινε σαν εκτιμήσεις των διανυσμάτων βάσης β_i , $i = 1, \dots, k$ του κεντρικού υπόχωρου τα διανύσματα b για τα οποία μεγιστοποιείται το μέσο τετραγωνικό μέτρο των διανύσματος $b\{Cov(z | y) - E[Cov(z | y)]\}$.

Δοθέντος ότι

$$E\|\mathbf{b}\{\text{Cov}(z|y) - E[\text{Cov}(z|y)]\}\|^2 = \mathbf{b} E\|\{\text{Cov}(z|y) - E[\text{Cov}(z|y)]\}\|^2 \mathbf{b}^T$$

τα εν λόγω διανύσματα \mathbf{b} δεν είναι άλλα από τα ιδιοδιανύσματα που αντιστοιχούν στις d μεγαλύτερες ιδιοτιμές της μήτρας

$$\text{SIR } II_s = E(\text{Cov}(z|y) - \text{airII})^2$$

όπου d η διάσταση του κεντρικού υπόχωρου και

$$\text{airII} = E[\text{Cov}(z|y)].$$

Εναλλακτικά ο Li (1990a) πρότεινε την εύρεση των διανυσμάτων \mathbf{b} για τα οποία μεγιστοποιείται ο μέσος της ποσότητας $\{\mathbf{b} (\text{Cov}(z|y) - E[\text{Cov}(z|y)]) \mathbf{b}^T\}^2$ και επομένως το συναρτησιακό $\text{Var}[\text{Var}(\mathbf{b}^T x|y)]$. Όπως αποδεικνύεται το ιδιοδιάνυσμα που αντιστοιχεί στη μέγιστη ιδιοτιμή της $\text{SIR } II_s$ αποτελεί μία καλή αρχική εκτίμηση της λύσης του προβλήματος. Επίσης επισήμανε τη δυνατότητα χρήσης της

$$\text{SIR } II_r = E [\text{airII}^{-1/2} \text{Cov}(z|y) \text{airII}^{-1/2} - \mathbf{I}]$$

αντί της $\text{SIR } II_s$, καθώς και τη δυνατότητα συμπληρωματικής χρήσης των μητρών $\text{SIR } II_s$ και $\text{SIR } I = \text{Cov}[E(z|y)]$ μέσω της $\text{SIR } II_a$, όπου

$$\text{SIR } II_a = (1-\alpha) \text{SIRI}^2 + \alpha \text{SIR } II_s$$

Η ιδέα αυτής της συμπληρωματικής χρήσης βασίζεται στην ταυτότητα

$$\Sigma_x = \text{SIR } II_s + \text{airII}$$

όπου $\Sigma_x = \text{Cov}(\mathbf{x})$, ενώ εύκολα αποδεικνύεται ότι για $\alpha = 0.5$

$$2 \text{SIR } II_a = E[\text{Cov}(z|y) - \mathbf{I}]^2 = \text{SAVE}$$

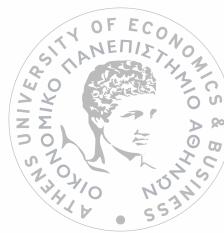
(Ο Li (1991b) μάλλον εκ παραδρομής αναφέρει ότι

$$2 \text{SIR } II_a = E[E(z|y) - \mathbf{I}]^2 = \text{SAVE)$$

4.2.2 ΑΛΓΟΡΙΘΜΟΣ

Η διαδικασία είναι ανάλογη με την αντίστοιχη για τη SIR .

Έστω n_s ο αριθμός των παρατηρήσεων για την απόκριση y στην ζώνη s , και έστω $\hat{z}_i = \hat{\Sigma}_x^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})$, $i=1,2,\dots,n$ οι τυποποιημένες δειγματικές τιμές των παρατηρήσεων για τις ανεξάρτητες x όπου $\hat{\Sigma}_x$ η δειγματική εκτίμηση της $\Sigma_x = \text{Var}(\mathbf{x})$ και $\bar{\mathbf{x}}$ ο δειγματικός μέσος του \mathbf{x} .



Βήμα 1: Υπολογισμός σε κάθε ζώνη του δειγματικού μέσου των \hat{z}_i

$$\bar{z}_s = \frac{1}{n_s} \sum_{y_i \in J_s} \hat{z}_i$$

Βήμα 2: Υπολογισμός σε κάθε ζώνη της δειγματικής μήτρας συνδιακύμανσης

$$\hat{\Sigma}_{zs} = \text{Cov}(z | \tilde{y} = s) = \frac{1}{n_s - 1} \sum_{y_i \in J_s} z_i z_i^T - n_s \bar{z}_s \bar{z}_s^T$$

Βήμα 3: Υπολογισμός του σταθμισμένου μέσου των $\text{Cov}(z | \tilde{y} = s)$

$$\bar{V} = \frac{1}{n} \sum_{s=1}^k n_s \text{Cov}(z | \tilde{y} = s)$$

Βήμα 4: Υπολογισμός της δειγματικής εκτίμησης

$$\hat{sirII}_s = \frac{1}{n} \sum_{s=1}^k n_s (\text{cov}(z | \tilde{y} = s) - \bar{V})^2$$

Βήμα 5: Εύρεση των ιδιοδιανυσμάτων $\hat{u}_1, \dots, \hat{u}_p$ που αντιστοιχούν στις ιδιοτιμές

$$\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p \quad \text{της μήτρας } \hat{sirII}_s$$

Βήμα 6: Εστω $d = \dim S_{y,x}$. Η *SIR II* εκτίμηση του $S_{y,x}$ θα είναι

$$\hat{S}_{y,x} = S(\hat{\Sigma}_x^{-1/2} \hat{u}_1, \dots, \hat{\Sigma}_x^{-1/2} \hat{u}_d)$$

4.2.3 ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΤΑΣΗΣ d

Ο Li (1991b) συμφωνεί με τη χρήση του ελέγχου διατάξεων (*permutation test*) για τον έλεγχο της υπόθεσης $d = 0$ έναντι της $d \geq 0$. Ο έλεγχος αυτός είναι ελεύθερος προϋποθέσεων και μπορεί να χρησιμοποιηθεί και κατά την εφαρμογή της *SIR* και της *SIR II*. Ωστόσο για την περίπτωση ελέγχου της $d = 1$ έναντι της $d > 1$ ο Li (1991b) πρότεινε τη χρήση ενός αλγορίθμου bootstrap που έχει ως εξής.

Βήμα 1: Υπολογισμός των εκτιμήσεων $\hat{\beta}_1, \dots, \hat{\beta}_p$ της βάσης του $S_{y,x}$ μέσω της

SIR II, και έστω

$$u_i = \hat{\beta}_1^T x_i$$

και

$$v_i = (\hat{\beta}_2^T x_i, \dots, \hat{\beta}_p^T x_i), i = 1, \dots, n$$

Βήμα 2: Διάταξη των u_i κατά αύξουσα σειρά ώστε

$$u_{(1)} \leq \dots \leq u_{(n)}$$

Βήμα 3: Δημιουργία πληθυσμού για τη λήψη δείγματος bootstrap: για $j=1, \dots, \kappa$

$$(y_{(i)}, u_{(i)}, v_{(i+j)}), \quad i=1, \dots, n-\kappa$$

$$(y_{(i)}, u_{(i)}, v_{(i-j)}), \quad i=\kappa+1, \dots, n$$

όπου κ γνωστός αριθμός

Βήμα 4: Λήψη iid δείγματος μεγέθους η από τον πληθυσμό του βήματος 3.

Βήμα 5: Υπολογισμός του μέσου των $p-1$ ιδιοτιμών της μήτρας $SIR II_s$ που προκύπτει με βάση τα δεδομένα του βήματος 4.

Βήμα 6: Πολλαπλή επανάληψη των βημάτων 4,5 για τη λήψη της κατανομής αναφοράς για τον έλεγχο της υπόθεσης $d=1$.

Ο Li (1991b) υποστηρίζει ότι ο παραπάνω αλγόριθμος μπορεί εύκολα να επεκταθεί και στην περίπτωση ελέγχου της $d = d_o > 1$ έναντι της $d \geq d_o$.

4.2.4 ΕΦΑΡΜΟΓΕΣ

Η πρώτη αναφέρεται από τον Li (1991b) και αφορά τα εξής δύο μοντέλα

$$y = sign(\varepsilon)[\log|\beta_1^T x| - 0.75] \quad (4.2)$$

και

$$y = sign(\beta_2^T x)[\Phi(\beta_2^T x) - 0.5] \quad (4.3)$$

Τα δύο αυτά μοντέλα χρησιμοποιήθηκαν για την παραγωγή δεδομένων μέσω προσομοίωσης. Για τα (4.2)-(4.3) αναφέρεται για το μέγεθος του δείγματος $n=300$ και για τον αριθμό των μεταβλητών $p=10$. Υπονοείται ότι τα δεδομένα είναι iid και ότι $\varepsilon \sim N(0,1)$. Επίσης είναι Φ η κανονική αθροιστική συνάρτηση κατανομής (για $N(0,1)$).

Η κατεύθυνση β_1 εκτιμάται πολύ ικανοποιητικά από την πρώτη κατεύθυνση της SIRII και για τα δύο μοντέλα. Ωστόσο δεν φαίνεται να ισχύει κάτι ανάλογο για την β_2 του μοντέλου (4.3) η οποία δεν εκτιμάται το ίδιο ικανοποιητικά από τη δεύτερη κατεύθυνση της SIR II. Η β_2 εκτιμάται ικανοποιητικά από τη μέθοδο “double – slicing” για την οποία έγινε λόγος στην 3.1.5b.

Η δεύτερη εφαρμογή αναφέρεται στη βιβλιγραφία και αφορά το εξής μοντέλο:

$$y = \beta_1^T x + (\beta_1^T x)^3 + 4(\beta_2^T x)^2 + \varepsilon \quad (4.4)$$

όπου $x \sim N_3(\mathbf{0}, I)$, $\varepsilon \sim N(0,1)$, $\beta_1 = (1, 1, 1)^T$ και $\beta_2 = (1, -1, -1)^T$.

Το μοντέλο αυτό χρησιμοποιήθηκε για την παραγωγή $n=300$ iid ζευγών (y_i, x_i) .



Από την εφαρμογή της *SIR* προέκυψε ότι η πρώτη κατεύθυνση $\hat{\beta}_1$ που εκτιμά η *SIR* προσεγγίζει ικανοποιητικά την πραγματική κατεύθυνση β_1 δοθέντος ότι το κανονικοποιημένο εσωτερικό γινόμενο των β_1 και $\hat{\beta}_1$ είναι 0.9894 και επομένως β_1 και $\hat{\beta}_1$ είναι σχεδόν παράλληλα διανύσματα. Η επιτυχία της *SIR* στην εκτίμηση της β_1 ήταν αναμενόμενη δοθέντος ότι ο όρος $\{\beta_1^T x + (\beta_1^T x)^3\}$ αυξάνει μονοτονικά. Ωστόσο η μέθοδος *SIR* δεν αποδίδει το ίδιο καλά στην εκτίμηση της β_2 . Η μέθοδος δεν αναγνωρίζει κατεύθυνση άλλη από τη β_1 κι αυτό ήταν αναμενόμενο δοθέντος ότι ο όρος $(\beta_2^T x)^2$ εμφανίζει συμμετρική κυρτότητα γύρω από το 0. Επομένως η κατεύθυνση β_2 αναμένεται να εκτιμηθεί από την κατεύθυνση b για την οποία μεγιστοποιείται η $Var(b^T x | y) - E[Var(b^T x | y)]$. Πράγματι το ιδιοδιάνυσμα $\hat{\beta}_1$ που αντιστοιχεί στη μέγιστη ιδιοτιμή της *SIR II_s* και η πραγματική κατεύθυνση β_2 έχουν κανονικοποιημένο εσωτερικό γινόμενο 0.9992.

Με δύο λόγια η *SIR* εκτιμά την κατεύθυνση β_1 για την οποία μεγιστοποιείται η αντίστροφη ροπή πρώτης τάξης, και η *SIR II* εκτιμά την κατεύθυνση β_2 για την οποία μεγιστοποιείται η αντίστροφη ροπή δεύτερης τάξης.

4.2.5 ΣΧΟΛΙΑ – ΕΠΙΣΗΜΑΝΣΕΙΣ

Η μέθοδος *SIR II* δεν αρκείται στην ισχύ της υπόθεσης γραμμικότητας και της υπόθεσης σταθερής διακύμανσης. Προϋποθέτει και την κανονικότητα του x . Για την περίπτωση της ελλειπτικής συμμετρίας του x που είναι μία προϋπόθεση λιγότερο αυστηρή, ο Li (1991b) βασιζόμενος στο θεώρημα 6.2 Li (1991b) ισχυρίζεται ότι αν το πλήθος p των ανεξάρτητων μεταβλητών είναι μεγάλο και η διάσταση k του κεντρικού υπόχωρου είναι μικρή τότε το ορθογώνιο συμπλήρωμα του κεντρικού υπόχωρου θα περιέχεται στον υπόχωρο με βάση ιδιοδιάνυσμα που αντιστοιχεί σε ιδιοτιμή της $Cov(z | y) - airII$ που δεν θα είναι μεν μηδενική θα είναι όμως μικρή. Με άλλα λόγια δηλαδή οι κατευθύνσεις $\hat{\beta}_i$, $i = 1, \dots, k$ που εκτιμά η *SIR II* θα προσεγγίζουν και σ' αυτήν την περίπτωση την πραγματική βάση του κεντρικού υπόχωρου με μεγάλη πιθανότητα. Τέλος για την περίπτωση ισχύος της υπόθεσης γραμμικότητας και μόνο, ο Li (1991b) παραπέμπει στο θεώρημα 6.1 (Li (1991b)) το οποίο όπως επισημαίνει μπορεί να αποτελέσει βάση συζήτησης.

4.3 ΜΕΘΟΔΟΣ *pHd* (*PRINCIPAL HESSIAN DIRECTIONS*)

4.3.1 ΘΕΩΡΗΤΙΚΗ ΘΕΜΕΛΙΩΣΗ

Η μέθοδος αυτή εισήχθη από τον Li (1992) και βασίζεται στην παρατήρηση ότι η r_{xp} χεισιανή μήτρα $H(x)$ της $E(y|x)$ εκφυλίζεται σε οποιοδήποτε επίπεδο $b^T x$ κάθετο στα $\beta_i^T x$, $i = 1, \dots, k$ όπου β_i τα διανύσματα βάσης του κεντρικού υπόχωρου.

Υπενθυμίζεται ότι

$$H(x) = \frac{\partial^2 E(y|x)}{\partial x \partial x^T}$$

Επομένως σύμφωνα και με το λήμμα 2.2 του Li (1992), οι *principal Hessian directions* οι οποίες ορίζονται ως τα ιδιοδιανύσματα της μήτρας

$$\bar{H}_x \Sigma_x$$

όπου

$$\bar{H}_x = EH(x)$$

και $\Sigma_x = Var(x)$, και αντιστοιχούν ταυτόχρονα σε ιδιοτιμές σημαντικά διάφορες του μηδενός μπορούν ν' αποτελέσουν εκτιμήσεις της βάσης του κεντρικού υπόχωρου $S_{y|x}$. Αυτές οι *principal Hessian directions* αποτελούν ένα σύστημα αξόνων κατά μήκος των οποίων η μέση καμπυλότητα ως προς τις δεύτερες παραγώγους της $E(y|x)$ μεγιστοποιείται.

Συνεπώς το κρίσιμο ζήτημα είναι η δυνατότητα εκτίμησης της \bar{H}_x η οποία επιτυγχάνεται μέσω του ακόλουθου πορίσματος που απορρέει από το λήμμα του Stein(1981, Λήμμα 4).

Πόρισμα 4.1 Έστω ότι x κατανέμεται κανονικά με μέσο μ_x και μήτρα διακύμανσης Σ_x , και έστω μ_y ο μέσος του y . Τότε θα ισχύει

$$\bar{H}_x = \Sigma_x^{-1} \Sigma_{yx} \Sigma_x^{-1}$$

όπου

$$\Sigma_{yx} = E(y - \mu_y)(x - \mu_x)(x - \mu_x)^T$$

Επομένως οι *pHd* (*Principal Hessian directions*) μπορούν να προκύψουν μέσω της Σ_{yx} σύμφωνα με το θεώρημα που ακολουθεί.



Θεώρημα 4.2 Αν x κατανέμεται κανονικά τότε οι pHd θα είναι τα ιδιοδιανύσματα

b_j της μήτρας Σ_{yxx} σύμφωνα με την εξίσωση

$$\Sigma_{yxx} b_j = \lambda_j \Sigma_x b_j, \quad j = 1, \dots, p$$

O Cook (1998) θεωρώντας $z = \Sigma_x^{-1/2}(x - E(x))$, καθώς και τα OLS κατάλοιπα

$$e = y - E(y) - \beta^T z$$

όπου

$$\beta = Cov(z, y),$$

καταλήγει στη σχέση

$$\Sigma_{ezz} = E\left(\frac{\partial^2 E(e|z)}{\partial z \partial z^T}\right)$$

όπου

$$\Sigma_{ezz} = E(ezz^T)$$

εκφράζοντας μία παραλλαγή του πορίσματος 4.1, οπότε δεδομένου ότι

$$\Sigma_{ezz} = \Sigma_x^{-1/2} \Sigma_{exx} \Sigma_x^{-1/2}$$

προκύπτει ότι οι pHd της $E(e|z)$ οι εκτιμήσεις δηλαδή της βάσης του S_{ez} θα είναι τα ιδιοδιανύσματα b_j της μήτρας Σ_{ez} (Cook (1998)) σύμφωνα με την εξίσωση

$$\Sigma_{ezz} b_j = \lambda_j b_j, \quad j = 1, \dots, p$$

Αυτή η παραλλαγή του πορίσματος 4.1 για τον Cook δεν είναι παρά μία ειδική περίπτωση της πρότασης 4.3 που ακολουθεί και σύμφωνα με την οποία ο S_{ezz} μπορεί να χρησιμοποιηθεί για την εκτίμηση του S_{ez} ακόμα και όταν το x δεν κατανέμεται κανονικά.

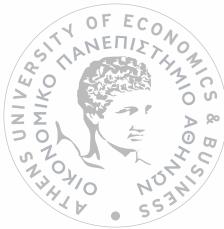
Πρόταση 4.3 Αν ρ αποτελεί βάση του S_{ez} και

$$1. E(z|\rho^T z) = P_\rho z$$

$$2. Var(z|\rho^T z) = I_p - P_\rho$$

τότε

$$S_{ezz} \subset S_{ez} \text{ όπου } S_{ezz} = S(\Sigma_{ez})$$



Ο υπόχωρος $S_{e|z}$ συνδέεται με τον $S_{y|z}$ σύμφωνα με την παρακάτω πρόταση.

Πρόταση 4.4 Αν οι στήλες της μήτρας γ αποτελούν βάση του υπόχωρου $S_{y|z}$,

$\beta = \text{Cov}(z, y)$ και $E(z|\gamma^T z) = P_z z$ όπου P_z η *projection matrix* στον $S(\gamma)$ τότε

$$S_{y|z} = S_{e|z} + S(\beta)$$

Η συνθήκη $E(z|\gamma^T z) = P_z z$ εξασφαλίζει ότι $\beta \in S_{y|z}$

Ο Cook χρησιμοποιεί την παραπάνω πρόταση με σκοπό την εφαρμογή της μεθόδου *rHd* στα *OLS* κατάλοιπα και όχι στην απόκριση y , για λόγους που θα εξηγηθούν στη συνέχεια.

4.3.2 ΑΛΓΟΡΙΘΜΟΣ

Έστω $\hat{z}_i = \hat{\Sigma}_x^{-1/2} (x_i - \bar{x})$, $i=1,2,\dots,n$ όπου \bar{x} , ο δειγματικός μέσος και

$\hat{\Sigma}_x = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$ η δειγματική εκτίμηση της Σ_x .

Βήμα 1: Υπολογισμός των δειγματικών *OLS* κατάλοιπων

$$\hat{e}_i = y_i - \bar{y} - \hat{\beta}^T \hat{z}_i$$

όπου

$$\hat{\beta} = (Z^T Z)^{-1} Z^T Y$$

και

$$Z = \begin{bmatrix} 1 & z_{11} & z_{12} & \cdots & z_{1,p-1} \\ 1 & z_{21} & z_{22} & \cdots & z_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & z_{n1} & z_{n2} & \cdots & z_{n,p-1} \end{bmatrix}$$

και

$$Y^T = (y_1, \dots, y_n)$$

Βήμα 2: Υπολογισμός της δειγματικής εκτίμησης $\hat{\Sigma}_{ezz}$

$$\hat{\Sigma}_{ezz} = \hat{\Sigma}_x^{-1/2} \hat{\Sigma}_{exx} \hat{\Sigma}_x^{-1/2}$$

όπου

$$\hat{\Sigma}_{\epsilon\epsilon} = \frac{1}{n} \sum_{i=1}^n \hat{e}_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

Βήμα 3: Εύρεση των διανυσμάτων $\hat{\mathbf{l}}_1, \dots, \hat{\mathbf{l}}_p$ που αντιστοιχούν στις ιδιοτιμές

$$|\hat{\lambda}_1| \geq \dots \geq |\hat{\lambda}_p| \text{ της μήτρας } \hat{\Sigma}_{\epsilon\epsilon}.$$

Βήμα 4: Έστω $d = \dim S_{\epsilon\epsilon}$, τότε για την εκτίμηση του $S_{\epsilon\epsilon}$ θα ισχύει

$$\hat{S}_{\epsilon\epsilon} = S(\hat{\mathbf{l}}_1, \dots, \hat{\mathbf{l}}_d)$$

υπό την προϋπόθεση ισχύος του πορίσματος 4.1 με την μορφή

$$\Sigma_{\epsilon\epsilon} = E\left(\frac{\partial^2 E(e | z)}{\partial z \partial z^T}\right)$$

4.3.3 ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΤΑΣΗΣ d

Για τον έλεγχο της υπόθεσης $d = m$ έναντι της $d > m$ για τη διάσταση d του $S_{\epsilon\epsilon}$ ο Li (1992) πρότεινε τη χρήση του παρακάτω στατιστικού

$$\hat{A}_m = \frac{n \sum_{j=m+1}^p \hat{\lambda}_j^2}{2V\hat{a}r(e)}$$

όπου $V\hat{a}r(e)$ συνεπής εκτιμήτρια της περιθώριας διακύμανσης του e . Έστω Θ_0 η $p \times (p-d)$ μήτρα με στήλες τα ιδιοδιανύσματα που αντιστοιχούν στις μηδενικές ιδιοτιμές της $\Sigma_{\epsilon\epsilon}$, και το $(p-d) \times 1$ διάνυσμα $v = \Theta_0^T z$ με j -οστό στοιχείο v_j . Έστω επίσης το $((p-d)(p-d+1)/2) \times 1$ διάνυσμα

$$w = \begin{pmatrix} \begin{pmatrix} v_1^2 - 1 \\ \sqrt{2}v_1v_2 \\ \sqrt{2}v_1v_3 \\ \vdots \\ \sqrt{2}v_1v_{p-d} \\ \vdots \\ \begin{pmatrix} v_j^2 - 1 \\ \sqrt{2}v_jv_{j+1} \\ \vdots \\ \sqrt{2}v_jv_{p-d} \\ \vdots \\ \begin{pmatrix} v_{p-d-1}^2 - 1 \\ \sqrt{2}v_{p-d-1}v_{p-d} \\ \begin{pmatrix} v_{p-d-1}^2 \end{pmatrix} \end{pmatrix} \end{pmatrix} \end{pmatrix}$$

Για την ασυμπτωτική κατανομή του $\hat{\Lambda}_d$ ισχύει η παρακάτω πρόταση (Cook (1998)).

Πρόταση 4.5 Η ασυμπτωτική κατανομή του $\hat{\Lambda}_d$ είναι ίδια με την κατανομή του

$$C = \frac{1}{2\text{var}(e)} \cdot \sum_{j=1}^{(p-d)(p-d+1)/2} \omega_j c_j$$

όπου c_j είναι ανεξάρτητες χ^2 τυχαίες μεταβλητές με 1 βαθμό ελευθερίας και $\omega_1 \geq \omega_2 \geq \dots \geq \omega_{(p-d)(p-d+1)/2}$

οι ιδιοτιμές της $\text{Var}(ew)$.

Στο σημείο αυτό πρέπει να τονιστεί όπως επισημαίνει ο Cook (1998) ότι η ασυμπτωτική κατανομή του $\hat{\Lambda}_d$ δεν θα ήταν η παραπάνω εάν δεν ισχυει $\text{Cov}(e, z) = 0$. Ας δούμε τώρα έναν αλγόριθμο τον οποίο παραθέτει ο Cook (1998) για την εύρεση της διάστασης d μέσω στατιστικού ελέγχου.

Βήμα 1: Υπολογισμός

(α) της $p \times (p-d)$ μήτρας $\hat{\Theta}_0 = (\hat{l}_{d+1}, \dots, \hat{l}_p)$

(β) των $(p-d) \times 1$ διανυσμάτων $\hat{v}_i = \hat{\Theta}_0^T \hat{z}_i$, $i = 1, \dots, n$

(γ) των $(p-d)(p-d+1)/2 \times 1$ διανυσμάτων \hat{w}_i , $i = 1, \dots, n$ με χρήση των στοιχείων \hat{v}_i .



Βήμα 2: Υπολογισμός της δετγματικής εκτίμησης $\hat{\Sigma}_{ew}$ της $Var(\mathbf{ew})$ μέσω των $\hat{\omega}_i$ και των \hat{w}_i .

Βήμα 3: Υπολογισμός των ιδιοτιμών $\hat{\omega}_1 \geq \hat{\omega}_2 \geq \dots \geq \hat{\omega}_{(p-d)(p-d+1)/2}$ της $\hat{\Sigma}_{ew}$.

Βήμα 4: Εκτίμηση της κατανομής του \hat{A}_d μέσω της κατανομής του

$$\hat{C} = \frac{1}{2\hat{var}(e)} \cdot \sum_{j=1}^{(p-d)(p-d+1)/2} \hat{\omega}_j c_j$$

όπου c_j είναι ανεξάρτητες χ^2 τυχαίες μεταβλητές με 1 βαθμό ελευθερίας.

Για τον υπολογισμό ποσοστιαίων σημείων κατανομών που ακολουθούν γραμμικοί συνδυασμοί χ^2 τυχαίων μεταβλητών υπάρχει εκτενής βιβλιογραφία (Farebrother (1990), Field (1993) και Wood (1989)).

Ο Cook (1998) με σκοπό τον έλεγχο της ανεξαρτησίας των e , z χρησιμοποιεί την κατανομή του \hat{A}_0 με βάση το παρακάτω πόρισμα.

Πόρισμα 4.5 Έστω $e \perp\!\!\!\perp z$, τότε η ασυμπτωτική κατανομή του \hat{A}_0 θα είναι ίδια

με την κατανομή του

$$D = \frac{1}{2} \sum_{j=1}^{p(p+1)/2} \omega_j c_j$$

όπου c_j είναι ανεξάρτητες χ^2 τυχαίες μεταβλητές με 1 βαθμό ελευθερίας και $\omega_1 \geq \omega_2 \geq \dots \geq \omega_{p(p+1)/2}$ οι ιδιοτιμές της $Var(\mathbf{w})$,

όπου \mathbf{w} προκύπτει για $\Theta_0 = I$.

4.3.3α ΠΡΟΣΘΕΤΕΣ ΥΠΟΘΕΣΕΙΣ

Έστω ότι η pxd μήτρα Θ αποτελεί ορθογώνια βάση του S_{ezz} και ότι η μήτρα (Θ, Θ_0) αποτελεί ορθογώνια βάση του \mathbb{R}^p όπου Θ_0 η μήτρα που ορίστηκε στην προηγούμενη παράγραφο.

Η πρόταση που ακολουθεί παρατίθεται από τον Cook (1998) και δείχνει ότι η ασυμπτωτική κατανομή του \hat{A}_d απλοποιείται όταν το διάνυσμα x των ανεξάρτητων μεταβλητών κατανέμεται κανονικά και όταν οι κατευθύνσεις των διανυσμάτων που αποτελούν τις στήλες της Θ επαρκούν για την εύρεση της δομής των $E(e|z)$ και



$Var(e|z)$.

Πρόταση 4.6 Έστω ότι

1. x κατανέμεται κανονικά
2. $E(e|z) = E(e|\Theta^T z)$
3. $Var(e|z) = Var(e|\Theta^T z)$

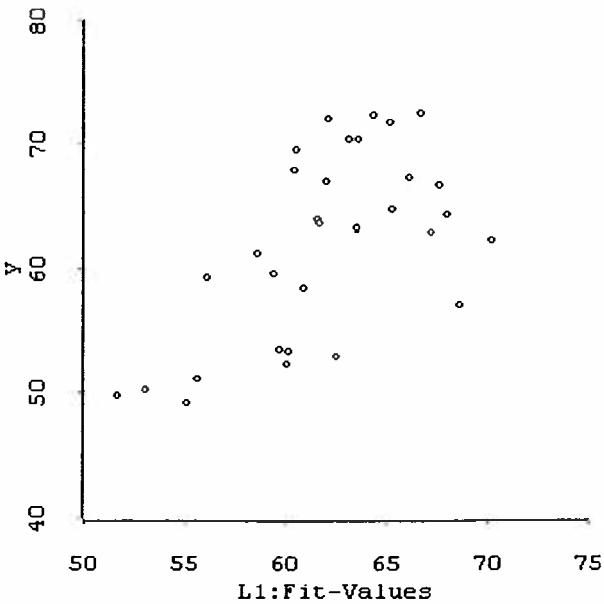
τότε η ασυμπτωτική κατανομή του \hat{A}_d θα είναι χ^2 με $(p-d)(p-d+1)/2$ βαθμούς ελευθερίας.

Στο σημείο αυτό θα πρέπει να επισημανθεί ότι όταν ο έλεγχος της υπόθεσης $d = 0$ οδηγεί σε απόρριψη με βάση την κατανομή της πρότασης 4.5 τότε αυτό θέτει υπό αμφισβήτηση την ισχύ των 1,2 και 3.

4.3.4 ΕΦΑΡΜΟΓΕΣ

ΕΦΑΡΜΟΓΗ 1 Η εφαρμογή που ακολουθεί αφορά δεδομένα τα οποία υπάρχουν στο αρχείο ryield.lsp που διατίθεται μέσω του Arc. Τα δεδομένα αφορούν παρατηρήσεις για την απόκριση για την σχετίζεται με το αποτέλεσμα μιας χημικής αντίδρασης δύο σταδίων, καθώς και για 5 ανεξάρτητες μεταβλητές που σχετίζονται με τη χρονική διάρκεια και τη θερμοκρασία των δύο σταδίων του πειράματος. Η επεξεργασία των δεδομένων και ο σχολιασμός των αποτελεσμάτων παρατίθεται από τον Cook (1998).

Το διάγραμμα 4.1 που ακολουθεί απεικονίζει τη σχέση των $y, \hat{\beta}^T x$, όπου $\hat{\beta}$ η OLS εκτίμηση των συντελεστών παλινδρόμησης του y στο x .



Διάγραμμα 4.1 διάγραμμα { $y, \hat{\beta}^T x$ } για το πρόβλημα της χημικής αντίδρασης ($\hat{\beta}$ η κατεύθυνση OLS)

Η εικόνα του διαγράμματος προδίδει την ακαταλληλότητα του OLS γραμμικού μοντέλου γεγονός που αποδεικνύεται και από την απόρριψη της υπόθεσης $d = 0$ για τη διάσταση του S_{ez} . Η απόρριψη αυτή ισοδυναμεί με απόρριψη της υπόθεσης για έλλειψη συσχέτισης μεταξύ των OLS κατάλοιπων e και του γινομένου zz^T . Ωστόσο δεν μπορεί να εξαχθεί συμπέρασμα για το αν αυτή η ύπαρξη συσχέτισης οφείλεται στην αστοχία της OLS μοντελοποίησης της $E(y|z)$ ή της $Var(y|z)$. Για την εύρεση της απάντησης η χρήση της SIR φαίνεται να βοηθάει. Συγκεκριμένα για

$h = 7$ ζώνες το $p\text{-value}$ του $\hat{\Lambda}_0$ είναι 0.216 και συνεπώς η υπόθεση $d = 0$ για τη διάσταση του S_{yz} δεν μπορεί να απορριφθεί. Δοθέντος όμως ότι η SIR είναι ευαίσθητη στην ετεροσκεδαστικότητα δεν θα προέκυπτε $d=0$ σε περίπτωση αστοχίας της OLS μοντελοποίησης της $Var(y|z)$. Επομένως το OLS μοντέλο αστοχεί στη μοντελοποίηση της $E(y|z)$ η οποία στην πραγματικότητα δεν αυξάνει μονότονα κάτι που αν συνέβαινε δεν θα προέκυπτε $d = 0$ δοθέντος ότι η SIR είναι αποτελεσματική στην ανίχνευση μη γραμμικής τάσης, εκτός αν δεν είναι μονότονη.

Τα δύο ως τώρα κρίσιμα συμπεράσματα είναι η έλλειψη ετεροσκεδαστικότητας της $y|z$ και η έλλειψη γραμμικότητας της $y|z$. Η έλλειψη ετεροσκεδαστικότητας της $y|z$ μας επιτρέπει να εκτιμήσουμε τον S_{ez} μέσω του S_{ez} εφόσον ισχύουν οι

προϋποθέσεις της πρότασης 4.3 κάτι που σύμφωνα με την εικόνα της *scatterplot matrix* φαίνεται να συμβαίνει. Πρέπει να τονιστεί ότι αν αντίθετα υπήρχε ετεροσκεδαστικότητα τότε ο S_{ez} δεν θα εκτιμούσε επαρκώς τον S_{ez} του οποίου θα ήταν γνήσιο υποσύνολο, θα υπολείπετο δηλαδή σε πληροφορία κάτι που φαίνεται χαρακτηριστικά στην περίπτωση κανονικής κατανομής του x οπότε

$$\Sigma_{ez} = E\left(\frac{\partial^2 E(e|z)}{\partial z \partial z^T}\right).$$

Η έλλειψη γραμμικότητας της $y|z$ μας απαλλάσσει από την ανάγκη προσθήκης του $S(\beta)$ στον S_{ez} . Πρέπει δε να σημειωθεί ότι η ισχύς των προϋποθέσεων της πρότασης 4.3 εξασφαλίζει ότι $\beta \in S_{yz}$ σύμφωνα με την πρόταση 4.4. Πράγματι η τιμή του συντελεστή προσδιορισμού R^2 για την παλινδρόμηση της $\hat{\beta}^T x$ στις $\hat{u}_1^T x, \hat{u}_2^T x$ και $\hat{u}_3^T x$ όπου $\hat{u}_1, \hat{u}_2, \hat{u}_3$ τα μετασχηματισμένα ιδιοδιανύσματα που αντιστοιχούν στις τρεις μεγαλύτερες ιδιοτιμές της Σ_{ez} είναι 0.947 πράγμα που σημαίνει ότι $S(\beta) \in S_{ez}$ και ότι η μέθοδος *pHd* πέτυχε να ανιχνεύσει την ύπαρξη καμπυλότητας στην $E(y|z)$. Να σημειωθεί ότι έγινε χρήση των τριών μεγαλυτέρων ιδιοτιμών δοθέντος ότι $\hat{d} = 3$ με βάση τις *p-value* του \hat{A}_d που προτείνει η μέθοδος *pHd*. Πάντως ο δειγματικός συντελεστής συσχέτισης μεταξύ $\hat{\beta}^T x$ και $\hat{u}_1^T x$ είναι μόλις 0.117, όπως φαίνεται και από το *Output* του *Arc* που ακολουθεί, πράγμα που σημαίνει ότι $\dim S_{yz} \geq 2$, το δε διάγραμμα 4.2 απεικονίζει τη σχέση των $y, \hat{u}_1^T x$.

Inverse Regression *pHd* (*OLS residuals*)

Name of Dataset = ReacYield

Name of Fit = I1.pHd

Response = *OLS residuals based on y*

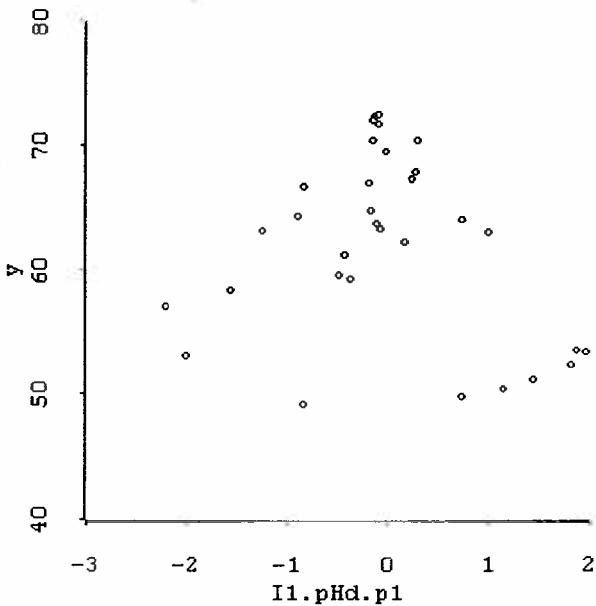
Predictors = (T1 T2 Lt1 Lt2 C)

Std. coef. use predictors scaled to have SD equal to one.

Predictors	Lin Comb 1		Lin Comb 2		Lin Comb 3	
	Raw	Std.	Raw	Std.	Raw	Std.
T1	0.336	0.356	-0.580	-0.605	0.657	0.681
T2	-0.531	-0.528	0.516	0.505	0.512	0.498
Lt1	0.192	0.190	0.386	0.375	0.505	0.488
Lt2	-0.021	-0.022	0.108	0.109	0.212	0.211
C	-0.753	-0.747	-0.487	-0.475	0.080	0.078



Eigenvalues	-6.110	3.376	-2.948
R^2(OLS pHd)	0.117	0.139	0.947
Tests from sums of squared eigenvalues times 0.468264			
Number of Components	Eigenvalue Partial Sum	DF	ChiSq p-value Adj p-value
1	28.35	15	0.019 0.003
2	10.87	10	0.368 0.047
3	5.534	6	0.477 0.051
4	1.465	3	0.690 0.241



Διάγραμμα 4.2 διάγραμμα $\{y, \hat{u}_1^T x\}$ για το πρόβλημα της χημικής αντίδρασης (\hat{u}_1 η πρώτη κατεύθυνση που εκτιμά η pHd)

Ενδιαφέρον, στο σημείο αυτό, παρουσιάζει η εφαρμογή της μεθόδου pHd στα OLS κατάλοιπα r από την πολινδρόμηση του $\hat{\beta}^T x$ στο x . Στην περίπτωση αυτή είναι σαφές ότι $S(\beta) \notin S_{rz}$ ενώ προκύπτει ότι $\dim S_{rz} = 1$. Όσο για τον δειγματικό συντελεστή συσχέτισης μεταξύ της αντίστοιχης pHd κατεύθυνσης και της $\hat{u}_1^T x$ είναι 0.99 πράγμα που σημαίνει ότι $\dim \hat{S}_{y|x} = 2$ και ότι τα διανύσματα βάσης του $\hat{S}_{y|x}$ θα είναι $\hat{\Sigma}_x^{-1/2} \hat{\beta}$ και \hat{u}_1 .

ΕΦΑΡΜΟΓΗ 2 Η εφαρμογή αυτή παρατίθεται από τον Li(1992) και βασίζεται σε ένα σετ δεδομένων από τους Breiman και Friedman (1985). Σκοπός της εφαρμογής



είναι να δείξει κάτι που φάνηκε ήδη από την προηγούμενη εφαρμογή. Την αποτελεσματικότητας της συμπληρωματικής εφαρμογής των μεθόδων *SIR* και *rHd*.

Τα δεδομένα είναι παρατηρήσεις για την συγκέντρωση όζοντος της ατμόσφαιρας καθώς και για οκτώ ακόμη μεταβλητές με στόχο τη διερεύνηση της επίδρασης των οκτώ αυτών μεταβλητών στη συγκέντρωση όζοντος. Το μέγεθος του δείγματος των παρατηρήσεων είναι $n = 330$.

Αρχικά εφαρμόστηκε η *SIR* η οποία αποκάλυψε μία σημαντική κατεύθυνση πολύ κοντά στην *OLS* κατεύθυνση \hat{b}_{ols} . Η χρήση αλγορίθμου επιλογής των πιο σημαντικών ανεξάρτητων μεταβλητών και η εν συνεχείᾳ επανάληψη της *SIR* απέφερε τη μεταβλητή $\hat{b}_s^T x$ που εξηγεί πλήρως την $\hat{b}_{ols}^T x$ διθέντος ότι ο δειγματικός συντελεστής συσχέτισης μεταξύ τους είναι 0.99. Το διάγραμμα $\{y, \hat{b}_s^T x\}$ παρουσιάζει σαφή τετραγωνική μορφή κάτι που υπαγορεύει την προσαρμογή του πολυωνύμου

$$y = c_0 + c_1 u_1 + c_2 u_1^2 + \varepsilon$$

όπου $u_1 = \hat{b}_s^T x$

Στη συνέχεια εφαρμόζεται η *rHd* στα κατάλοιπα της προσαρμογής του πολυωνύμου και μετά την επιλογή των πιο σημαντικών μεταβλητών και την επανάληψη της μεθόδου *rHd* στα αντίστοιχα κατάλοιπα προκύπτει η κατεύθυνση \hat{b}_{phd} σαν η μοναδική σημαντική. Το διάγραμμα $\{e, \hat{b}_{phd}^T x\}$ παρουσιάζει σαφή τετραγωνική μορφή όχι όμως και ετεροσκεδαστικότητα.

Αντίθετα το διάγραμμα $\{e, \hat{b}_{sir}^T x\}$ για τη μεταβλητή $\hat{b}_{sir}^T x$ που αντιστοιχεί στη μοναδική κατεύθυνση \hat{b}_{sir} που προκύπτει από αντίστοιχη εφαρμογή της *SIR* στα κατάλοιπα της προσαρμογής του πολυωνύμου παρουσιάζει ετεροσκεδαστικότητα.

Ο δε δειγματικός συντελεστής συσχέτισης μεταξύ των $\hat{b}_{phd}^T x$ και $\hat{b}_{sir}^T x$ είναι μόλις 0.2 πράγμα που σημαίνει ότι οι κατευθύνσεις $\hat{b}_{phd}, \hat{b}_{sir}$ είναι διαφορετικές. Η $\hat{b}_{phd}^T x$ επιτυγχάνει να ανιχνεύσει την καμπυλότητα της $E(y|u_1)$ αποτυγχάνει όμως στην ανίχνευση της ετεροσκεδαστικότητας για λόγους που έχουν ήδη συζητηθεί. Αντίθετα η $\hat{b}_{sir}^T x$ αποτυγχάνει να ανιχνεύσει την καμπυλότητα της $E(y|u_1)$, επιτυγχάνει όμως στην ανίχνευση της ετεροσκεδαστικότητας. Επομένως η συνδυαστική χρήση των *rHd* και *SIR* μας επιτρέπει να αντλήσουμε την πληροφορία της $y|x$ όπως αυτή η πληροφορία αποτυπώνεται στην πρώτη $E(y|x)$ και τη δεύτερη $Var(y|x)$ ροπή της $y|x$.

4.3.5 ΣΧΟΛΙΑ – ΕΠΙΣΗΜΑΝΕΙΣ

Όπως αναφέρει ο Li(1992) η pHd αδυνατεί να ανιχνεύσει οποιαδήποτε edr κατεύθυνση b για την οποία

$$Cov(y, (b^T(x - \mu_x))^2) = 0$$

Το ενδεχόμενο να πρόκειται για edr προκύπτει είτε με τη βοήθεια κάποιας άλλης μεθόδου είτε γραφικά. Για την αντιμετώπιση του προβλήματος ώστε να μπορεί να γίνει χρήση της pHd για τη διερεύνηση του ενδεχομένου να πρόκειται για edr ο Li (1992) προτείνει τη χρήση μετασχηματισμού που υποδεικνύεται από την αντίστροφη παλινδρόμηση. Συγκεκριμένα ο Li (1992) προτείνει την εφαρμογή της μεθόδου στα δεδομένα $\{T(y), (b^T(x - \mu_x))^2\}$

όπου

$$T(y) = c_b^{-1} E((b^T(x - \mu_x))^2 | y)$$

$$c_b^2 = Var(E((b^T(x - \mu_x))^2 | y))$$

και

$$Var(b^T x) = 1$$

Η διασύνδεση ανάμεσα στην pHd και τη SIR ή τη $SIR II$ που υπονοείται από τον παραπάνω μετασχηματισμό επισημαίνεται και από τον Li (1992) ο οποίος παρατήρησε ότι

$$\Sigma_{yxx} = p_h(E((x - \mu_x)(x - \mu_x)^T | y \in I_h) - \Sigma_x)$$

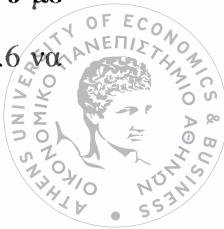
όπου $p_h = P(y \in J_s)$, J_s η ζώνη τιμών του y για $s = 1, 2, \dots, h$.

Η βάση του κεντρικού υπόχωρου θα μπορεί σ' αυτήν την περίπτωση να εκτιμηθεί από τα ιδιοδιανύσματα της $E((x - \mu_x)(x - \mu_x)^T | y \in I_h)$ που αντιστοιχούν στις σημαντικά διάφορες της μονάδας ιδιοτιμές της ως προς τη Σ_x .

Παρόμοιο πρόβλημα αδυναμίας της pHd στην ανίχνευση της β επισημαίνει ο Cook (1998) για την εφαρμογή της pHd στην απόκριση y υπό την ισχύ του γραμμικού μοντέλου

$$y | z = \beta_0 + \beta^T z + \varepsilon$$

Σε μια τέτοια περίπτωση η συνθήκη $Cov(e, z) = \mathbf{0}$ η οποία αποτελεί βασική προϋπόθεση για την ισχύ της πρότασης 4.6 δεν θα ισχύει αφού $\beta = Cov(y, z) \neq \mathbf{0}$ με αποτέλεσμα η κατανομή του $\hat{\Delta}_o$ που προκύπτει από εφαρμογή της πρότασης 4.6 να διανέπισθη.



είναι μετατοπισμένη προς τα αριστερά σε σχέση με την πραγματική. Κατά συνέπεια η ισχύς του ελέγχου $\kappa = 0$ περιορίζεται αφού σε κάποιες περιπτώσεις γίνεται δεκτή η υπόθεση $\kappa = 1$ παρ' ότι $\Sigma_{yz} = \mathbf{0}$ ώστε $\kappa = 0$. Με άλλα λόγια η εφαρμογή της μεθόδου pHd θα ανιχνεύει κάποια κατεύθυνση η οποία στην πραγματικότητα δεν θα είναι edr κατεύθυνση. Ο Cook (1998) καταλήγει στο συμπέρασμα ότι καλό είναι να αποφεύγεται η εφαρμογή της μεθόδου στην απόκριση y . Περισσότερα επί του θέματος παρατίθενται στον Cook (1998b).

ΚΕΦΑΛΑΙΟ 5

ΑΛΛΕΣ ΜΕΘΟΔΟΙ

Στο κεφάλαιο αυτό εξετάζονται τρεις ακόμη μέθοδοι μέσα στα πλαίσια της dimension reduction μέσω εκτίμησης του κεντρικού υπόχωρου. Η *SAT* (Sliced Average Third-Moment Estimation) η οποία βασίζεται στη ροπή τρίτης τάξης της $x|y$, η *Cov_k* οποία στοχεύει στην εκτίμηση του *CKMS* (*Central K-th Moment Subspace*) του υπόχωρου δηλαδή που περιέχει όλη την πληροφορία η οποία διατίθεται για την γ μέσω των ροπών της $y|x$ από την 1^η έως και την κ -οστή, και η γραφική μέθοδος η οποία χρησιμοποιείται σε συνδυασμό με κάποια άλλη μέθοδο.

5.1 ΜΕΘΟΔΟΣ *SAT* (*SLICED AVERAGE THIRD – MOMENT ESTIMATION*)

5.1.1 ΘΕΩΡΗΤΙΚΗ ΘΕΜΕΛΙΩΣΗ

Η μέθοδος αυτή εισήχθη από τους Yin και Cook (2003) και στοχεύει στην ανίχνευση *edr* κατευθύνσεων οι οποίες δεν ανιχνεύονται μέσω των αντίστροφων ροπών 1^{ης} τάξης και 2^{ης} τάξης. Για το σκοπό αυτό η μέθοδος χρησιμοποιεί την αντίστροφη ροπή 3^{ης} τάξης και βασίζεται στο παρακάτω θεώρημα (Yin και Cook (2003)).

Θεώρημα 5.1 Εάν οι στήλες της μήτρας γ αποτελούν βάση του $S_{y|x}$ και

$$3. \quad E(z|\gamma^T z) = P_\gamma z$$

$$4. \quad Var(z|\gamma^T z) = I_p - P_\gamma$$

$$5. \quad M^{(3)}(z|\gamma^T z) = 0$$

όπου P_γ ο *projection operator* για τον $S_{y|x}$, και I_p η μοναδιαία μήτρα τάξεως p τότε

$$M^{(3)}(z|y) = (P_\gamma \otimes P_\gamma) M^{(3)}(z|y) P_\gamma$$

Υπενθυμίζεται ότι για $x = (x_1, x_2, \dots, x_p)^T$

$$M^{(3)}(x) = E[\{x - E(x)\} \otimes \{x - E(x)\} \{x - E(x)\}^T]$$

ενώ υπάρχουν τρεις εναλλακτικοί τρόποι έκφρασης της $M^{(3)}(x)$.

Σύμφωνα με τον πρώτο η $M^{(3)}(x)$ είναι μία τρισδιάστατη *pxpxp* μήτρα της οποίας το n -οστό πρόσωπο δηλαδή η *pxp* μήτρα που είναι κάθετη στο επίπεδο των γραμμών και των στηλών είναι

$$M_k^{(3)}(\mathbf{x}) = E[\{x_k - E(x_k)\} \{\mathbf{x} - E(\mathbf{x})\} \{\mathbf{x} - E(\mathbf{x})\}^T], \quad k = 1, 2, \dots, p.$$

Στην περίπτωση αυτή

$$M^{(3)}(\mathbf{x}) = \begin{pmatrix} M_1^{(3)}(\mathbf{x}) \\ \vdots \\ M_p^{(3)}(\mathbf{x}) \end{pmatrix}$$

κάτι που σημαίνει ότι η $M^{(3)}(\mathbf{x})$ μπορεί εναλλακτικά να εκφραστεί και σαν $p^2 \times p$ block μήτρα με στοιχεία - μήτρες τα $p \times p$ πρόσωπα $M_k^{(3)}(\mathbf{x}) \quad k = 1, 2, \dots, p$.

Σύμφωνα με τον τρίτο τρόπο έκφρασης η $M^{(3)}(\mathbf{x})$ είναι μία $p \times p$ μήτρα με στοιχεία στήλες. Η ij -οστή στήλη είναι

$$M_{ij}^{(3)}(\mathbf{x}) = E[\{\mathbf{x} - E(\mathbf{x})\} \{x_i - E(x_i)\} \{x_j - E(x_j)\}], \quad i, j = 1, 2, \dots, p$$

Ισχύει στην περίπτωση αυτή ότι

$$M_{ij}^{(3)}(\mathbf{x}) = (\mathbf{I} \otimes \mathbf{e}_i^T) M^{(3)}(\mathbf{x}) \mathbf{e}_j$$

όπου \mathbf{e}_j το μονοδιαίο $p \times 1$ διάνυσμα με 1 στην i -οστή θέση και 0 στις υπόλοιπες.

Για τη δεσμευμένη κατανομή $z|y$ ισχύει

$$M^{(3)}(z|y) = E_{z|y}[\{z - E(z|y)\} \otimes \{z - E(z|y)\} \{z - E(z|y)\}^T]$$

Σύμφωνα με το παραπάνω θεώρημα 5.1 και υπό την ισχύ των προϋποθέσεων γραμμικότητας 1, σταθερής διακύμανσης 2 και συμμετρίας 3 οι οποίες ισχύουν όταν z ή $z|y$ κατανέμεται κανονικά, οι διανυσματικοί χώροι με βάση αντίστοιχα τις γραμμές, τις στήλες και τα πρόσωπα της $M^{(3)}(z|y)$ θα ανήκουν στον κεντρικό υπόχωρο S_{yz} . Αυτό σημαίνει ότι η $M^{(3)}(z|y)$ μπορεί να χρησιμεύσει για την εκτίμηση του κεντρικού υπόχωρου S_{yz} . Ειδικότερα για τη μήτρα

$$\mathbf{M}_{SAT} = E(\mathbf{M}_y \mathbf{M}_y^T)$$

όπου \mathbf{M}_y η $p \times p(p+1)/2$ μήτρα των διαφορετικών στηλών της $p \times p^2$ μήτρας

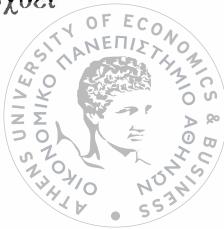
$M^{(3)}(z|y)^T$, ισχύει το παρακάτω θεώρημα που διατυπώθηκε από τους

Yin και Cook (2003).

Θεώρημα 5.2 Ισχύει

1. $S(\mathbf{M}_{SAT}) = S \{ \mathbf{M}_y, y \in R(y) \}$ και

υπό την ισχύ των προϋποθέσεων γραμμικότητας (1), σταθερής διακύμανσης (2), και συμμετρίας (3) του θεωρήματος 5.1 ισχύει



$$2. S(\mathbf{M}_{SAT}) \subseteq S_{yk}$$

5.1.2 ΑΛΓΟΡΙΘΜΟΣ

Ο αλγόριθμος που ακολουθεί παρατίθεται από τους Yin και Cook (2003)

Βήμα 1: Διακριτοποίηση της απόκρισης y με βάση τη δημιουργία h διαφορετικών

ζωνών J_s , $s=1,2,\dots,h$ για τις τιμές της. Ισχύει ότι $\tilde{y}=s$ όταν $y \in J_s$.

Βήμα 2: Υπολογισμός των τυποποιημένων δειγματικών τιμών

$$\hat{z}_i = \hat{\Sigma}_x^{-1/2} (\mathbf{x}_i - \bar{\mathbf{x}}), i=1,2,\dots,n.$$

των παρατηρήσεων για τις ανεξάρτητες μεταβλητές x , όπου $\hat{\Sigma}_x$ η δειγματική εκτίμηση της $\Sigma_x = Var(x)$ και \bar{x} ο δειγματικός μέσος του x .

Βήμα 3: Υπολογισμός σε κάθε ζώνη s των $p(p+1)/2$ διανυσμάτων της δειγματικής εκτίμησης $\hat{M}^{(3)}(\hat{z} | \tilde{y} = s)$ ως εξής

$$\hat{M}_{jk}^{(3)}(\hat{z} | \tilde{y} = s) = \frac{1}{n_s} \sum_{i \in J_s}^n \hat{z}_i \hat{z}_{ij} \hat{z}_{ik}, j \geq k = 1, \dots, p$$

όπου \hat{z}_{ij} το j -οστό στοιχείο του διανύσματος \hat{z}_i , και n_s ο αριθμός των παρατηρήσεων που εμπίπτουν στη ζώνη s . Τα διανύσματα αυτά δεν είναι άλλα από τις στήλες της $p \times p(p+1)/2$ μήτρας \hat{M}_y , που αποτελεί δειγματική εκτίμηση της M_y .

Βήμα 4: Υπολογισμός των ιδιοδιανυσμάτων \hat{l}_j που αντιστοιχούν στις ιδιοτιμές $\hat{\lambda}_j$ της δειγματικής εκτίμησης της μήτρας \mathbf{M}_{SAT} .

$$\hat{M}_{SAT} = \sum_{s=1}^h \frac{n_s}{n} \hat{M}_s \hat{M}_s^T$$

όπου \hat{M}_s είναι η \hat{M}_y για τη ζώνη s που υπολογίστηκε στο βήμα 3.

Βήμα 5: Έστω $d = \dim S(\mathbf{M}_{SAT})$, τότε θα είναι

$$\hat{S}(\mathbf{M}_{SAT}) = S(\hat{l}_1, \dots, \hat{l}_d)$$

και η SAT εκτίμηση του S_{yk} θα είναι

$$\hat{S}_{y|x} = S(\hat{\Sigma}_x^{-1/2} \hat{l}_1, \dots, \hat{\Sigma}_x^{-1/2} \hat{l}_d)$$

5.1.3 ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΤΑΣΗΣ d



Για την εκτίμηση της διάστασης d του $S(M_{SAT})$ οι Yin και Cook (2003) προτείνουν τη χρήση του στατιστικού

$$\hat{\Lambda}_m = n \sum_{j=m+1}^p \hat{\lambda}_j$$

και τη διεξαγωγή ελέγχου διατάξεων (*permutation test*) σύμφωνα με τη διαδικασία που παρουσιάστηκε μέσα στα πλαίσια ανάπτυξης της μεθόδου *SAVE* και συζητήθηκε ήδη στην παράγραφο 4.1.3. Η ιδέα της διεξαγωγής τέτοιων ελέγχων μελετήθηκε περαιτέρω από τους Cook και Yin (2001), ενώ οι Yin και Cook (2002) μέσα στα πλαίσια ανάπτυξης της μεθόδου *COV_k* η οποία θα συζητηθεί στην επόμενη παράγραφο παραθέτουν τη θεωρητική βάση του ελέγχου αυτού.

5.1.4 ΕΦΑΡΜΟΓΕΣ

Η μέθοδος *SAT* φαίνεται να βρίσκει ευρεία εφαρμογή σε προβλήματα λογιστικής πολινδρόμησης (*logistic regression*) η οποία αποτελεί χαρακτηριστική περίπτωση πολινδρόμησης για την οποία η $y|x$ μπορεί να προσεγγιστεί μέσω της αντίστροφης $x|y$. Υπενθυμίζεται ότι σε τέτοια προβλήματα

$$y_i|x_i \sim Bin(m_i, \theta(x_i))$$

Όπου y_i ο αριθμός των επιτυχιών σε m_i δοκιμές και $\theta(x_i)$ η πιθανότητα επιτυχίας για την οποία θεωρείται ότι ακολουθεί τη λογιστική συνάρτηση

$$\theta(x_i) = \frac{\exp(\mathbf{n}^T \mathbf{u}_i)}{1 + \exp(\mathbf{n}^T \mathbf{u}_i)}$$

όπου \mathbf{u}_i συναρτήσεις των \mathbf{x}_i .

Σε μία τέτοια περίπτωση θα είναι

$$E(y_i|m_i|x_i) = \theta(x_i)$$

$$Var(y_i|m_i|x_i) = \theta(x_i)(1 - \theta(x_i))| m_i$$

οπότε η τιμής της πιθανότητας θ θα καθορίζει αμφότερες τις συναρτήσεις του μέσου και της διακύμανσης. Συνήθως $m_i = 1$, οπότε $y_i = 0$ ή 1.

Σύμφωνα με το θεώρημα του Bayes προκύπτει ότι

$$\log\left(\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})}\right) = logc + \log\left(\frac{f(\mathbf{x} | y=1)}{f(\mathbf{x} | y=0)}\right)$$

κάτι που σημαίνει ότι η σχέση ανάμεσα στις δύο δεσμευμένες κατανομές του \mathbf{x} για $y = 0$ και 1 δηλαδή η διαφορά τους ως προς το μέσο, τη διακύμανση και τη

ασυμμετρία μπορεί να αποτελέσει βάση για την εκτίμηση της διάστασης και της βάσης του κεντρικού υπόχωρου και να χρησιμεύσει για την κατάλληλη μοντελοποίηση της πιθανότητας θ .

Συνιστάται στον αναγνώστη παράλληλα με την μελέτη των εφαρμογών που ακολουθούν και παρατίθενται από τους Yin και Cook (2003), η μελέτη των Κεφαλαίων 21,22 από τους Cook και Weisberg (1999) καθώς και η μελέτη της μεθοδολογίας των Cook και Lee (1999).

ΕΦΑΡΜΟΓΗ 1

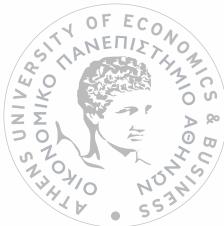
Τα δεδομένα της πρώτης εφαρμογής έχουν προκύψει από προσομοίωση ενός μοντέλου για το οποίο $y = 0,1$ με $Pr(y=0) = Pr(y=1) = 0.5$ ενώ οι αντίστροφες ροπές πρώτης και δεύτερης τάξης είναι κοινές μεταξύ των $x|(y=0)$ και $x|(y=1)$ η δε αντίστροφη ροπή τρίτης τάξης διαφέρει. Συγκεκριμένα:

$$x|(y=0) \sim 0.25 N(6J_1, I) + 0.75N(2J_1, I)$$

$$x|(y=1) \sim 0.5 N(3J_1, J_2) + 0.5N(3J_1, J_3)$$

όπου $J_1 = (1,1,0)^T$, $J_2 = 2I - e_3e_3^T$, $J_3 = 6J_1J_1^T + e_3e_3^T$ με I την 3×3 μοναδιαία μήτρα και $e_3 = (0,0,1)^T$. Ισχύουν δηλαδή οι προϋποθέσεις εφαρμογής του θεωρήματος 5.1.

Η πραγματική βάση του κεντρικού υπόχωρου είναι $(1,0,0)^T, (0,1,0)^T$. Για την διεξαγωγή των ελέγχων που προβλέπουν οι μέθοδοι *SIR*, *SAVE* και *SAT* για την εκτίμηση της διάστασης του κεντρικού υπόχωρου, χρησιμοποιήθηκαν $n = 600$ iid παρατηρήσεις. Διεξήχθη ο ασυμπτωτικός έλεγχος *SIR* ενώ για τις άλλες δύο μεθόδους χρησιμοποιήθηκε έλεγχος διατάξεων (*permutation test*) ο οποίος βασίστηκε σε 500 αναδιατάξεις των τιμών της απόκρισης y . Σύμφωνα με τα αποτελέσματα η *SIR* και η *SAVE* δεν ανίχνευσαν καμία *edr* κατεύθυνση κάτι που ήταν αναμενόμενο δοθέντος ότι ο μέσος και η διακύμανση της $x|y$ δεν διαφοροποιούνται για $y=0$ και 1. Αντίθετα η αναμενόμενη ενασθησία της *SAT* στη διαφοροποίηση της τρίτης ροπής της $x|y$ επαληθεύεται δεδομένου ότι η *SAT* ανίχνευσε δύο *edr* κατευθύνσεις. Ο συντελεστής προσδιορισμού R^2 από την γραμμική παλινδρόμηση μεταξύ καθεμιάς από τις δύο αυτές κατευθύνσεις και των $(1,0,0)^T x, (0,1,0)^T x$ δηλαδή των x_1, x_2 είναι αντίστοιχα 0.9991 και 0.993 κάτι που σημαίνει ότι οι δύο κατευθύνσεις που ανίχνευσε η *SAT* προσεγγίζουν πολύ ικανοποιητικά τη βάση του κεντρικού υπόχωρου.



ΕΦΑΡΜΟΓΗ 2

Η δεύτερη εφαρμογή είναι μία παραλλαγή της πρώτης ως προς τη μορφή της $x|y$ την οποία επιχειρούμε να ανιχνεύσουμε μέσω εφαρμογής των μεθόδων SIR , $SAVE$ και SAT σε $400 iid$ παρατηρήσεις που έχουν προέλθει από προσομοίωση. Στόχος βέβαια και πάλι είναι να ελεγχθεί η ικανότητα των παραπάνω μεθόδων να εκτιμούν τη βάση του κεντρικού υπόχωρου. Στη δεύτερη αυτή εφαρμογή ισχύει $x|(U=j) \sim N(\mu_j, I)$ όπου U τυχαία μεταβλητή που παίρνει τις τιμές $\{1,2,3,4\}$ με πιθανότητα αντίστοιχα $\{0.1,0.4,0.1,0.4\}$. Ισχύει επίσης

$$\mu_1 = (1,1,1,1,1)^T$$

$$\mu_2 = (9,8,7,8,6,2)^T$$

$$\mu_3 = (1,8,1,6,1,7)^T$$

$$\mu_4 = (9,15,7,13,6,8)^T$$

ενώ $y=0$ όταν $U=1,2$ και $y=1$ όταν $U=3,4$. Ισχύουν δηλαδή και σ' αυτήν την εφαρμογή οι προϋποθέσεις εφαρμογής του θεωρήματος 5.1.

Η μέθοδος $SAVE$ δεν ανίχνευσε καμία edr κατεύθυνση όπως ήταν άλλωστε αναμενόμενο αφού η διακύμανση των $x|U$ είναι σταθερή. Αντίθετα η SIR ανιχνεύει μία edr κατεύθυνση στην οποία μεγιστοποιείται η διαφορά των $E(x|y)$ για $y=0,1$ ενώ η SAT ανιχνεύει και αυτή μία edr κατεύθυνση στην οποία μεγιστοποιείται η διαφορά των τρίτων ροπών της $x|y$ για $y=0,1$ κάτι που ήταν αναμενόμενο δοθέντος ότι

$$M^{(3)}(z|y=0) = \frac{f_1 f_2}{(f_1 + f_2)^3} (f_2 - f_1) \mathbf{b} \otimes \mathbf{b} \mathbf{b}^T$$

και

$$M^{(3)}(z|y=1) = \frac{f_3 f_4}{(f_3 + f_4)^3} (f_4 - f_3) \mathbf{b} \otimes \mathbf{b} \mathbf{b}^T$$

όπου $\mathbf{b} = \mu_1 - \mu_2$ και $f_j = Pr(U=j)$, $j=1,\dots,4$ με $f_1 \neq f_2$ και $f_4 \neq f_3$.

Το διάγραμμα $\{\hat{b}_{SAT1}^T x, \hat{b}_{SIR1}^T x\}|y$ όπου $\hat{b}_{SAT1}, \hat{b}_{SIR1}$ οι παραπάνω πρώτες κατευθύνσεις που εκτιμούν οι μέθοδοι SAT και SIR αναπαριστά τη δομή του προβλήματος. Ειδικότερα αναπαριστά τέσσερις κανονικούς υποπληθυσμούς με σταθερή περίπου διακύμανση, δύο διαφορετικές τιμές για το μέσο, μία για τους υποπληθυσμούς που αντιστοιχούν σε $U=1,2$ και μία για αυτούς που αντιστοιχούν σε $U=3,4$ και επίσης



δύο διαφορετικές ασυμμετρίες μία για τους υποπληθυσμούς που αντιστοιχούν σε $U=1,2$ και μία για αυτούς που αντιστοιχούν σε $U=3,4$.

Η εικόνα αυτή είναι συμβατή με την ύπαρξη κεντρικού υπόχωρου δύο διαστάσεων που είναι και η πραγματική διάσταση του προβλήματος.



5.1.5 ΣΧΟΛΙΑ - ΕΠΙΣΗΜΑΝΣΕΙΣ

Δύο ήταν τα σύνολα των υποθέσεων που έγιναν δεκτά κατά την ανάπτυξη της μεθόδου *SAT* (Yin και Cook (2003)). Το πρώτο σύνολο αφορά την υπόθεση ότι η απόκριση y και το διάνυσμα x ακολουθούν από κοινού κατανομή, ενώ το δεύτερο αφορά την υπόθεση της γραμμικότητας της σταθερής διακύμανσης και της συμμετρίας. Η παραβίαση της υπόθεσης της από κοινού κατανομής των y και x η οποία αφορά τη σχεδίαση του πειράματος δεν φαίνεται να επιδρά σημαντικά στα αποτελέσματα της μεθόδου, ωστόσο η σοβαρή παραβίαση οποιασδήποτε από τις υπόλοιπες υποθέσεις μπορεί να οδηγήσει σε παραπλανητικά αποτελέσματα. Δεδομένου, όπως ειπώθηκε, ότι το δεύτερο σετ υποθέσεων ισχύει όταν η περιθώρια ή η δεσμευμένη κατανομή του x είναι κανονική, συνιστάται η χρήση μεθόδων για την επαγωγή πολυμεταβλητής κανονικότητας για τις οποίες έγινε ήδη λόγος. Να σημειωθεί επίσης, όπως επισημαίνουν οι Yin και Cook (2003), ότι η χρήση της μεθόδου *SAT* στην εφαρμογή 2 ισοδυναμεί με τη χρήση της μεθόδου της διακριτής ανάλυσης(discriminant analysis) (Cook και Yin (2001)).

5.2 ΜΕΘΟΔΟΣ COV_k

5.2.1 ΘΕΩΡΗΤΙΚΗ ΘΕΜΕΛΙΩΣΗ

Η μέθοδος αυτή εισήχθη από τους Yin και Cook (2002) και βασίζεται στην έννοια του κεντρικού υπόχωρου ροπής k - τάξεως (*Central k-th Moment Subspace*) στις ιδιότητες του και στη σχέση του με τον γνωστό κεντρικό υπόχωρο.

Η προσπάθεια αναζήτησης ενός *DRS* ροπής k - τάξεως συνίσταται στην προσπάθεια αναζήτησης μιας $p \times q$ μήτρας n με $q \leq p$ τέτοιας ώστε το τυχαίο διάνυσμα $n^T x$ να περιέχει όλη την πληροφορία που διατίθεται για την απόκριση y μέσω των ροπών της $y|x$ από την 1^η έως και την k -οστή δηλαδή μέσω των

$E(y|x)$, $Var(y|x)$, ..., $M^{(k)}(y|x)$ όπου

$$M^{(k)}(y|x) = E[\{y - E(y|x)\}^k | x]$$



Προκύπτει έτσι ο ακόλουθος ορισμός (Yin και Cook (2002)).

Ορισμός 5.1 Εάν

$$y \perp\!\!\!\perp \{M^{(1)}(y|x), \dots, M^{(k)}(y|x)\} \mid \mathbf{n}^T \mathbf{x}$$

Τότε ο $S(n)$ ορίζεται ως DRS ροπής k -τάξεως για την $y|x$.

Σύμφωνα με τον παραπάνω ορισμό ένας DRS θα είναι αναγκαία και DRS ροπής k -τάξεως ο οποίος με τη σειρά του θα είναι και DRS i -τάξεως για οποιοδήποτε $i \leq k$. Επίσης όταν $k \rightarrow \infty$ τότε ο $S(n)$ θα είναι DRS εφόσον βέβαια υπάρχει η ροπογεννήτρια συνάρτηση της $y|x$. Επομένως η αναζήτηση του $S(n)$ όταν $k \rightarrow \infty$ ισοδυναμεί με την αναζήτηση του $S(n)$ ώστε $y \perp\!\!\!\perp \mathbf{x} \mid \mathbf{n}^T \mathbf{x}$. Η πρόταση που ακολουθεί (Yin και Cook (2002)), παρέχει συνθήκες ισοδύναμες με τη συνθήκη ανεξαρτησίας του παραπάνω ορισμού.

Πρόταση 5.1 Οι παρακάτω συνθήκες είναι ισοδύναμες

1. $y \perp\!\!\!\perp \{M^{(1)}(y|x), \dots, M^{(k)}(y|x)\} \mid \mathbf{n}^T \mathbf{x}$
2. $Cov\{y^i, M^{(j)}(y|x) \mid \mathbf{n}^T \mathbf{x}\} = \mathbf{0}$ για $j=1, \dots, k$.
3. $M^{(j)}(y|x)$ είναι συνάρτηση του $\mathbf{n}^T \mathbf{x}$. Ισοδύναμα $E(y^i | \mathbf{x})$ είναι συνάρτηση του $\mathbf{n}^T \mathbf{x}$ για $j=1, \dots, k$.
4. $Cov\{y^i, f(\mathbf{x}) \mid \mathbf{n}^T \mathbf{x}\} = \mathbf{0}$ για $j=1, \dots, k$ και οποιαδήποτε συνάρτηση $f(\mathbf{x})$.

Ας δούμε τώρα την έννοια του κεντρικού DRS ροπής k -τάξεως την οποία παραθέτουν οι Yin και Cook (2002), μία έννοια ανάλογη του κεντρικού DRS γνωστού και ως κεντρικού υπόχωρου.

Ορισμός 5.2 Έστω $S_{y|x}^{(k)} \cap S^{(k)}$ η τομή όλων των DRS ροπής k -τάξεως $S^{(k)}$. Εάν

$S_{y|x}^{(k)}$ είναι επίσης DRS ροπής k -τάξεως τότε ο $S_{y|x}^{(k)}$ ορίζεται ως ο κεντρικός DRS ροπής k -τάξεως ή κεντρικός υπόχωρος ροπής k -τάξεως (*Central k-th Moment Subspace*) ή *CKMS*.



Ο *CKMS* δεν υπάρχει πάντα δεδομένου ότι η τομή δύο *DRS* ροπής κ -τάξεως δεν είναι πάντα *DRS* ροπής κ -τάξεως. Ωστόσο η ύπαρξη του *CKMS* μπορεί να εξασφαλιστεί υπό την ισχύ των προϋποθέσεων για τις οποίες υπάρχει ο κεντρικός *DRS* ή κεντρικός υπόχωρος και για τις οποίες έγινε λόγος στο Κεφάλαιο 1. Στα πλαίσια παρουσίασης της παρούσας μεθόδου δεν τίθεται θέμα ύπαρξης του *CKMS*, θεωρείται δηλαδή ότι υπάρχει.

Επακόλουθο όσων ειπώθηκαν είναι η παρακάτω συνθήκη (Yin και Cook (2002)), που χαρακτηρίζει τη σχέση ανάμεσα στους κεντρικούς υπόχωρους ροπής 1- έως κ - τάξεως και τον κεντρικό υπόχωρο. Ισχύει λοιπόν ότι

$$S_{y|x}^{(1)} \subseteq \dots \subseteq S_{y|x}^{(\kappa)} \subseteq \dots \subseteq S_{y|x}$$

και

$$S_{y|x} = \lim_{k \rightarrow \infty} (S_{y|x}^{(k)})$$

Εάν η $y|x$ χαρακτηρίζεται πλήρως από τις ροπές της έως και την κ - αστή τότε $S_{y|x}^{(\kappa)} = S_{y|x}$ ενώ αν το διάνυσμα x τυποποιηθεί ώστε να έχει μέσο το 0 και διακύμανση I τότε θα ισχύει $S_{y|x}^{(\kappa)} = \Sigma_x^{-1/2} S_{y|z}^{(\kappa)}$ ανάλογα με τη σχέση $S_{y|x} = \Sigma_x^{-1/2} S_{y|z}$ που χαρακτηρίζει τους κεντρικούς υπόχωρους.

Τώρα που είναι σαφής η έννοια του *CKMS* και η σχέση του με τον κεντρικό υπόχωρο $S_{y|z}$ (ή $S_{y|x}$) ας δούμε μία σημαντική ιδιότητα του *CKMS* που μας επιτρέπει την εκτίμηση του *CKMS* και κατά συνέπεια την εκτίμηση του $S_{y|z}$. Η ιδιότητα αυτή περιγράφεται από την ακόλουθη πρόταση (Yin και Cook (2002)).

Πρόταση 5.2 Έστω g βάση του *CKMS* $S_{y|z}^{(\kappa)}$ και έστω $E(z | g^T z)$ γραμμική. Εάν

$f^{(k)}(y)$ είναι οποιοδήποτε κ το πολύ βαθμού πολυώνυμο του y τότε

$$E\{f^{(k)}(y)z\} \in S_{y|z}^{(\kappa)} \subseteq S_{y|z}.$$

Υπάρχουν πολλοί τρόποι επιλογής του $f^{(k)}(y)$ αλλά η επιλογή $f^{(k)}(y) = y^k$ φαίνεται κατάλληλη. Στη συνέχεια ορίζουμε τη μήτρα K

$$K = (E(yz), \dots, E(y^k z))$$

και ως $S_{cov}^{(k)}$ τον υποχώρο $S(K)$.

Με βάση την παραπάνω πρόταση και το γεγονός ότι



$$S_{y|z}^{(1)} \subseteq \dots \subseteq S_{y|z}^{(k)} \subseteq \dots \subseteq S_{y|z}$$

προκύπτει ότι

$$S_{cov}^{(k)} \subseteq S_{y|z}^{(k)}$$

Επομένως αν \hat{K} αποτελεί συνεπή εκτίμηση της K και $d = \dim S_{cov}^{(k)}$ τότε ο υπόχωρος με βάση τα αριστερά ιδιάζοντα διανύσματα (*left singular vectors*) της \hat{K} που αντιστοιχούν στις d μεγαλύτερες ιδιάζουσες τιμές (*singular values*) της \hat{K} θα αποτελεί συνεπή εκτίμηση του $S_{cov}^{(k)}$.

Συνήθως η διάσταση d δεν είναι γνωστή κάτι που οδηγεί στην εκτίμηση της μέσω μεθόδων που έχουν ήδη συζητηθεί. Επίσης αντί της K χρησιμοποιείται εναλλακτικά η

$$K_c = \{E(wz), \dots, E(w^k z)\}, \text{ όπου } w = \{y - E(y)\} / \sqrt{Var(y)}.$$

5.2.2 ΑΛΓΟΡΙΘΜΟΣ

Ο αλγόριθμος που ακολουθεί (Yin και Cook (2002)), αποσκοπεί στην εκτίμηση του

$$S[E\{f_j^{(k)}(y)z\}], \quad j=1,\dots,h$$

Όπου $f_1^{(k)}(y), \dots, f_h^{(k)}(y)$ είναι γραμμικώς ανεξάρτητα και γνωστά πολυώνυμα κ το πολύ βαθμού, με $h \leq \min(p, k)$. Στην πράξη επιλέγεται $h = \min(p, k)$. Ο παραπάνω υπόχωρος αποτελεί γενίκευση του $S(K_c) = S_{cov}^{(k)}$.

Βήμα 1: Υπολογισμός των τυποποιημένων δειγματικών τιμών

$$\hat{z}_i = \hat{\Sigma}_x^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}), \quad i=1,2,\dots,n.$$

των παρατηρήσεων για τις ανεξάρτητες μεταβλητές x , όπου $\hat{\Sigma}_x$ η δειγματική εκτίμηση της $\Sigma_x = Var(x)$ και $\bar{\mathbf{x}}$ ο δειγματικός μέσος του x .

Βήμα 2: Υπολογισμός της δειγματικής εκτίμησης \hat{K}_h της μήτρας

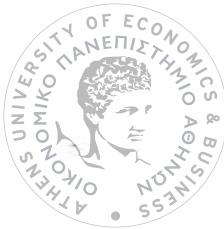
$$\begin{aligned} \hat{K}_h &= (E[\{f_1^{(k)}(y) - M_1\}z], \dots, E[\{f_h^{(k)}(y) - M_h\}z]) \\ &= (E\{f_1^{(k)}(y)z\}, \dots, E\{f_h^{(k)}(y)z\}) \end{aligned}$$

όπου

$$M_j = E\{f_j^{(k)}(y)\}, \quad j=1,\dots,h$$

ως εξής

$$\hat{K}_h = \left(\frac{1}{n} \sum_{i=1}^n f_1^{(k)}(y_i) \hat{z}_i, \dots, \frac{1}{n} \sum_{i=1}^n f_h^{(k)}(y_i) \hat{z}_i \right)$$



Βήμα 3: Υπολογισμός των αριστερών ιδιαζόντων διανυσμάτων (*left singular vectors*) $\hat{\mathbf{l}}_j$, $j = 1, \dots, h$ που αντιστοιχούν στις ιδιάζουσες τιμές (*singular values*) της $\hat{\mathbf{K}}_h$.

Βήμα 4: Έστω $d = \dim S(\mathbf{K}_h)$, τότε θα είναι

$$\hat{S}(\mathbf{K}_h) = S(\hat{\mathbf{l}}_1, \dots, \hat{\mathbf{l}}_d)$$

και η COV_k εκτίμηση του $S_{y|x}^{(k)}$ θα είναι

$$\hat{S}_{y|x}^{(k)} = S(\hat{\Sigma}_x^{-1/2} \hat{\mathbf{l}}_1, \dots, \hat{\Sigma}_x^{-1/2} \hat{\mathbf{l}}_d)$$

5.2.3 ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΤΑΣΗΣ d

Για την εκτίμηση της διάστασης d του $S(\mathbf{K}_h)$ ακολουθείται η διαδικασία του ελέγχου διατάξεων (*permutation test*) όπως παρουσιάστηκε στην 4.1.3. Οι Yin και Cook (2002) παραθέτουν τη θεωρητική βάση αυτής της διαδικασίας η οποία άπτεται της ακόλουθης πρότασης.

Πρόταση 5.3 Έστω \mathbf{U} η $p \times p$ μήτρα των ιδιοδιανυσμάτων \mathbf{u}_j της $p \times p$ θετικά ημιορισμένης και συμμετρικής μήτρας \mathbf{K} για την οποία $S(\mathbf{K}) = S_{y|x}$ και έστω $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)$ όπου \mathbf{U}_1 μήτρα διαστάσεων $p \times m$ και m η διάσταση του $S(\mathbf{K})$. Έστω επίσης $\mathbf{U}_1^T \mathbf{z} \perp\!\!\!\perp \mathbf{U}_2^T \mathbf{z}$.

Τότε ο $S(\mathbf{U}_1)$ θα αποτελεί DRS για την $y|z$ αν και μόνο αν

$$(y, \mathbf{U}_1^T \mathbf{z}) \perp\!\!\!\perp \mathbf{U}_2^T \mathbf{z}$$

Εάν δηλαδή η τιμή του $\hat{\Lambda}_m$ με βάση τα δεδομένα, προέρχεται από την κατανομή των τιμών του $\hat{\Lambda}_m$ που προκύπτουν από την αναδιάταξη των δειγματικών τιμών $(y, \mathbf{U}_1^T \mathbf{z})$ όπου \mathbf{U}_1 η μήτρα των m πρώτων ιδιοδιανυσμάτων της \mathbf{K} , τότε δεν μπορούμε να αντικρούσουμε την υπόθεση $(y, \mathbf{U}_1^T \mathbf{z}) \perp\!\!\!\perp \mathbf{U}_2^T \mathbf{z}$ και συνεπώς σύμφωνα με την παραπάνω πρόταση ο $S(\mathbf{U}_1)$ θα αποτελεί DRS για την $y|z$ και συνακόλουθα εκτίμηση του $S_{y|x}$. Επομένως για τον έλεγχο της υπόθεσης $d = m$ μπορεί να γίνει σύγκριση της τιμής του στατιστικού $\hat{\Lambda}_m = n \cdot \sum_{j=m+1}^p \hat{\lambda}_j$ όπου $\hat{\lambda}_j$ οι ιδιοτιμές της $\hat{\mathbf{K}}$ με το αντίστοιχο εκατοστιαίο σημείο της κατανομής του $\hat{\Lambda}_m$ η οποία προκύπτει με βάση τιμές του

στατιστικού για συγκεκριμένο αριθμό διατάξεων των δειγματικών τιμών $(y, U_1^T z)$. Εάν η τιμή του στατιστικού με βάση τα αρχικά δεδομένα υπερβαίνει το εκατοστιαίο σημείο της κατανομής τότε η υπόθεση $d=m$ απορρίπτεται, τίθεται $d = m+1$ και επαναλαμβάνεται η διαδικασία.

Το πλεονέκτημα του ελέγχου αυτού είναι ότι η υπόθεση $U_1^T z \neq U_2^T z$ στην οποία βασίζεται είναι ασθενέστερη της υπόθεσης κανονικότητας του z στην οποία βασίζεται το θεώρημα 3.2 για τον ασυμπτωτικό έλεγχο. Να προστεθεί τέλος ότι η πραγματοποίηση του ελέγχου διατάξεων μέσω του Arc είναι δυνατή φορτώνοντας το αρχείο PermTest.lsp που διατίθεται μέσω του προγράμματος. Το φόρτωμα του αρχείου πρέπει να γίνεται πρίν από το φόρτωμα του αρχείου που περιέχει τα προς επεξεργασία δεδομένα.

5.2.4 ΕΦΑΡΜΟΓΕΣ

ΕΦΑΡΜΟΓΗ 1 Η εφαρμογή αυτή παρατίθεται από τους Yin και Cook (2002) και αφορά δεδομένα που παρουσιάζονται από τον Freund (1979). Τα δεδομένα αυτά είναι παρατηρήσεις για την ημερήσια ποσότητα εξάτμισης από το έδαφος (απόκριση y) καθώς και για 4 ανεξάρτητες μεταβλητές που χαρακτηρίζουν τις καμπύλες της ημερήσιας μεταβολής της θερμοκρασίας και της υγρασίας του αέρα. Στόχος είναι η μελέτη της $y|z$ μέσω του $S_{cov}^{(2)}$. Η οπτική εξέταση της scatterplot matrix για τις ανεξάρτητες μεταβλητές συνηγορεί υπέρ της ισχύος της υπόθεσης γραμμικότητας.

Γίνεται χρήση της μήτρας

$$\hat{K}_c = \left(\frac{1}{n} \sum_{i=1}^n \hat{w}_i \hat{z}_i, \frac{1}{n} \sum_{i=1}^n \hat{w}_i^2 \hat{z}_i \right)$$

όπου $\hat{w}_i = (y_i - \bar{y}) / \hat{\sigma}(y)$. Τα αποτελέσματα του ελέγχου διατάξεων (*permutation test*) βασίστηκαν σε 1000 αναδιατάξεις από όπου προέκυψε ότι $\dim S_{cov}^{(2)} = 2$. Το ίδιο προέκυψε και για τη διάσταση του $S_{cov}^{(3)}$ γεγονός που οδηγεί στο συμπέρασμα ότι $S_{cov}^{(2)} = S_{y|z}^{(2)} = S_{y_1 z}$ ότι δηλαδή η πληροφορία για την y καλύπτεται από τις δύο πρώτες ροπές της $y|z$.

Στον πίνακα 5.1 που ακολουθεί φαίνονται οι συντεταγμένες των δύο κατευθύνσεων \hat{n}_1 και \hat{n}_2 , που ανιχνεύει η COV_2 καθώς και οι τυποποιημένες τιμές τους. Σύμφωνα με



αυτές η υγρασία του αέρα είναι η κυρίαρχη συνιστώσα στην πρώτη κατεύθυνση, ενώ στην δεύτερη κατεύθυνση όλες οι συνιστώσες συμμετέχουν εξίσου.

Predictor x	$\hat{\eta}_1$	Standardized $\hat{\eta}_1$	$\hat{\eta}_2$	Standardized $\hat{\eta}_2$
Air temperature area	0.096	0.114	0.242	0.531
Air temperature range	0.621	0.148	-0.797	-0.353
Humidity area	-0.562	-0.935	0.177	0.547
Humidity range	-0.538	-0.301	0.524	0.542

Πίνακας 5.1 εκτιμήσεις $\hat{\eta}_1, \hat{\eta}_2$ της βάσης του κεντρικού υπόχωρου(μέθοδος COV_2)

Με σκοπό τη σχετική σύγκριση εφαρμόστηκε στα δεδομένα και η μέθοδος SIR η οποία ωστόσο παρουσίασε ευαισθησία στον αριθμό των ζωνών h . Ο πίνακας 5.2 που ακολουθεί είναι χαρακτηριστικός της σχέσης ανάμεσα στις δύο μεθόδους και δείχνει τους συντελεστές προσδιορισμού R^2 από την παλινδρόμηση καθεμιάς από τις δύο κατευθύνσεις που ανιχνεύουν οι $SIR_3(h=3)$ και $SIR_4(h=4)$ στις μεταβλητές $\hat{\eta}_1^T x$ και $\hat{\eta}_2^T x$ που αντιστοιχούν στις κατευθύνσεις $\hat{\eta}_1, \hat{\eta}_2$ που ανιχνεύει η COV_2 . Όπως φαίνεται οι μεταβλητές που αντιστοιχούν στις πρώτες κατευθύνσεις των SIR_3 και SIR_4 δεν είναι παρά γραμμικοί συνδυασμοί των $\hat{\eta}_1^T x$ και $\hat{\eta}_2^T x$ κάτι όμως που δεν ισχύει για τις δεύτερες κατευθύνσεις. Επιπλέον έρευνα έδειξε ότι οι μεταβλητές που αντιστοιχούν στις πρώτες κατευθύνσεις των SIR_3 και SIR_4 είναι παραπλήσιες με την $\hat{\eta}_1^T x$, δεν ισχύει όμως το ίδιο για τις μεταβλητές που αντιστοιχούν στις δεύτερες κατευθύνσεις των SIR_3 και SIR_4 σε σχέση με την $\hat{\eta}_2^T x$.

Method	R^2 – values for the following SIR predictors	
	1st	2nd
SIR_3	0.982	0.018
SIR_4	0.999	0.510

Πίνακας 5.2 συντελεστές προσδιορισμού από την παλινδρόμηση των μεταβλητών που αντιστοιχούν στις δύο πρώτες κατευθύνσεις της SIR (για $h=3,4$),

στις $\hat{n}_1^T x, \hat{n}_2^T x$

Στη συνέχεια με σκοπό τον έλεγχο της αξιοπιστίας των μεθόδων έγινε προσομοίωση με βάση το μοντέλο:

$$y_{sim,i} = \hat{y}_i + 4\epsilon_i, i = 1, \dots, 46 \quad (5.1)$$

όπου \hat{y}_i οι τιμές (*fitted values*) από την προσαρμογή ενός πλήρως τετραγωνικού (*full quadratic*) μοντέλου στις μεταβλητές $\hat{n}_1^T x, \hat{n}_2^T x$ και $\epsilon_i \sim N(0,1)$. Είναι γνωστό ότι $S_{y_{sim},x} = S_{cov}^{(2)}$ και ότι η βάση του κεντρικού υπόχωρου είναι (\hat{n}_1, \hat{n}_2) . Για την αξιολόγηση της ακρίβειας των μεθόδων SIR και COV , εφαρμόστηκαν οι μέθοδοι COV_2 και SIR_3 στα δεδομένα που προέκυψαν από το παραπάνω μοντέλο προσομοίωσης και υπολογίστηκαν τα 5,50 και 95 ποσοστιαία σημεία της κατανομής του συντελεστή προσδιορισμού R^2 από την παλινδρόμηση καθεμίας από τις μεταβλητές που αντιστοιχούν στις δύο κατευθύνσεις που ανιχνεύουν οι μέθοδοι, στις δύο μεταβλητές $\hat{n}_1^T x$ και $\hat{n}_2^T x$. Όπως φαίνεται από τον πίνακα 5.3 που ακολουθεί οι μεταβλητές που αντιστοιχούν στις πρώτες κατευθύνσεις των COV_2 και SIR_3 αποτελούν γραμμικούς συνδυασμούς των $\hat{n}_1^T x, \hat{n}_2^T x$ όπως επίσης και η μεταβλητή που αντιστοιχεί στη δεύτερη κατεύθυνση της COV_2 , κάτι όμως που δεν ισχύει για τη δεύτερη κατεύθυνση της SIR_3 .

Method	Predictor	5% R ²	Median R ²	95% R ²
COV_2	1 st	0.989	0.995	0.999
COV_2	2 nd	0.890	0.979	0.998
SIR_3	1 st	0.973	0.984	0.995
SIR_3	2 nd	0.016	0.308	0.936
COV_2	1 st	0.987	0.996	0.999
COV_2	2 nd	0.878	0.971	0.998
SIR_4	1 st	0.972	0.987	0.998
$SIR_3, y_{sim} \perp\!\!\!\perp x$	2 nd	0.102	0.213	0.741
	2 nd	0.085	0.516	0.931

Πίνακας 5.3 5,50,95 ποσοστιαία σημεία της κατανομής τιμών του συντελεστή προσδιορισμού από την παλινδρόμηση των μεταβλητών που αντιστοιχούν στις δύο πρώτες κατευθύνσεις της SIR και των μεταβλητών που αντιστοιχούν στις δύο πρώτες κατευθύνσεις της COV_2 , στις $\hat{n}_1^T x, \hat{n}_2^T x$, όπου (\hat{n}_1, \hat{n}_2) η βάση του κεντρικού υπόχωρου για τα δεδομένα του μοντέλου 5.1

Οι τέσσερις επόμενες γραμμές του πίνακα αφορούν δεδομένα που προέκυψα από ανεξάρτητη προσομοίωση από όπου φαίνεται ότι η συμπεριφορά της δεύτερης κατεύθυνσης που ανιχνεύει η SIR_4 δεν είναι ικανοποιητική.

Τέλος η τελευταία γραμμή του πίνακα αφορά δεδομένα που προέκυψαν ως εξής

$$y_{sim,i} = 4\varepsilon_i$$

ώστε $y_{sim} \perp\!\!\!\perp x$.

Ωστόσο η δεύτερη κατεύθυνση που ανιχνεύει η SIR_3 κάθε άλλο παρά τυχαία φαίνεται να είναι.

5.2.5 ΣΧΟΛΙΑ – ΕΠΙΣΗΜΑΝΣΕΙΣ

5.2.5α ΔΙΑΓΝΩΣΤΙΚΟΣ ΕΛΕΓΧΟΣ ΤΟΥ ΓΡΑΜΜΙΚΟΥ ΜΟΝΤΕΛΟΥ

Στην προηγούμενη εφαρμογή αλλά και σε αρκετές άλλες περιπτώσεις προβλημάτων επιχειρείται η μελέτη της $y|z$ μέσω του $S_{cov}^{(2)}$. Όταν η πληροφορία για την $y|z$ πράγματι εξαντλείται από τις $E(y|z)$ και $Var(y|z)$ τότε τα ιδιάζοντα διανύσματα (*singular vectors*) που αντιστοιχούν στις δύο μεγαλύτερες ιδιάζουσες τιμές της μήτρας

$$\hat{\mathbf{K}}_c = \left(\frac{1}{n} \sum_{i=1}^n \hat{w}_i \hat{z}_i, \frac{1}{n} \sum_{i=1}^n \hat{w}_i^2 \hat{z}_i \right)$$

αποτελούν εκτίμηση της βάσης του $S_{y|z}$. Αν παρατηρήσει κανείς τις στήλες της $\hat{\mathbf{K}}_c$ θα δει κάτι πολύ ενδιαφέρον (Yin και Cook (2002)). Ότι το τετραγωνικό μήκος της πρώτης στήλης είναι n^{-1} φορές το στατιστικό του *score test* για την υπόθεση $\beta = \mathbf{0}$ για το ομοσκεδαστικό μονοδιάστατο γραμμικό μοντέλο

$$y = \beta_0 + g(\boldsymbol{\beta}^T z) + \varepsilon \quad (5.2)$$

και ότι το τετραγωνικό μήκος της δεύτερης στήλης είναι $2n^{-1}$ φορές το στατιστικό του *score test* για την υπόθεση $\alpha = \mathbf{0}$ για το ετεροσκεδαστικό γραμμικό μοντέλο

$$y = \beta_0 + \exp(\boldsymbol{a}^T z)\varepsilon \quad (5.3)$$

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

Επομένως η πρώτη στήλη της μήτρας περιέχει διαγνωστική πληροφορία για τη γραμμικότητα του μέσου υπό την προϋπόθεση ύπαρξης ομοσκεδαστικότητας, ενώ η δεύτερη στήλη περιέχει διαγνωστική πληροφορία για τη διακύμανση δηλαδή για την ύπαρξη ετεροσκεδαστικότητας υπό την προϋπόθεση ότι ο μέσος μοντελοποιείται σωστά από το γραμμικό μοντέλο ή για την γραμμικότητα του μέσου υπό την προϋπόθεση ότι το μοντέλο είναι ομοσκεδαστικό. Σε περίπτωση που το γραμμικό μοντέλο δεν επαρκεί για τη μοντελοποίηση ούτε του μέσου ούτε της διακύμανσης, τότε η δεύτερη στήλη περιέχει πληροφορία τόσο για την ύπαρξη ετεροσκεδαστικότητας όσο και για τη γραμμικότητα του μέσου. Όσα ειπώθηκαν μέχρι τώρα για την σημασία της δεύτερης στήλης της $\hat{\mathbf{K}}_c$ βασίζονται στην ιδέα του διαγνωστικού ελέγχου *score test* (Cook και Weisberg (1983)) για την υπόθεση $\alpha = \mathbf{0}$ στο γραμμικό μοντέλο

$$y = \beta_0 + \boldsymbol{\beta}^T v + \exp(\boldsymbol{a}^T z)\varepsilon$$



όπου $\varepsilon \sim N(0, \sigma^2)$ και \mathbf{v} διάνυσμα συναρτήσεων των στοιχείων του \mathbf{z} , ενώ για τη χρήση του ελέγχου αυτού παράλληλα με τη χρήση των OLS κατάλοιπων υπάρχει ανάλυση από τον Cook (1998) στο κεφάλαιο 14. Ειδικότερα προτείνεται σαν εκτίμηση του α το διάνυσμα β_{ols} το οποίο προκύπτει από την OLS προσαρμογή του γραμμικού μοντέλου

$$\mathbf{r}^2 = \beta_0 + \boldsymbol{\beta}^T \mathbf{z} + \varepsilon$$

στις τιμές (r_i^2, \hat{z}_i) , όπου r_i το OLS κατάλοιπο από την προσαρμογή του γραμμικού μοντέλου

$$y = \beta_0 + \boldsymbol{\beta}^T \mathbf{z} + \varepsilon$$

Αποδεικνύεται δε (Cook και Weisberg (1983)) ότι η $\hat{\mathbf{K}}_2$ η οποία είναι τέτοια ώστε

$$\hat{\mathbf{K}}_c = (\hat{\mathbf{K}}_1, \hat{\mathbf{K}}_2) = \left(\frac{1}{n} \sum_{i=1}^n \hat{r}_i \hat{z}_i, \frac{1}{n} \sum_{i=1}^n \hat{r}_i^2 \hat{z}_i \right)$$

αποτελεί συνεπή εκτίμηση του α υπό την προϋπόθεση ότι $E(\mathbf{z}|\boldsymbol{\alpha}^T \mathbf{z})$ γραμμική. Καταδεικνύεται έτσι η δυνατότητα διασύνδεσης του διαγνωστικού ελέγχου για την ύπαρξη ετεροσκεδαστικότητας μέσω των OLS κατάλοιπων, με τις στήλες της μήτρας $\hat{\mathbf{K}}_c = (\hat{\mathbf{K}}_1, \hat{\mathbf{K}}_2)$

5.2.5 β ΠΑΡΑΒΙΑΣΗ ΥΠΟΘΕΣΕΩΝ

Τρεις ήταν οι υποθέσεις που έγιναν κατά την ανάπτυξη της μεθόδου COV_k (Yin και Cook (2002)). Σύμφωνα με την πρώτη η απόκριση y και το διάνυσμα \mathbf{x} ακολουθούν από κοινού κατανομή. Η παραβίαση της υπόθεσης αυτής δεν είναι ωστόσο κρίσιμη καθότι τα αποτελέσματα που προέκυψαν ισχύουν και στην περίπτωση σχεδίασης πειράματος για το οποίο το \mathbf{x} παίρνει προκαθορισμένες τιμές (Yin και Cook (2002)). Σύμφωνα με τη δεύτερη υπόθεση ισχύει η γραμμικότητα της $E(\mathbf{z}|\boldsymbol{\gamma}^T \mathbf{z})$. Η παραβίαση της υπόθεσης αυτής έχει σαν αποτέλεσμα ότι $E(y^k z) \in L$ όπου $L = S\{E(z|\boldsymbol{\gamma}^T \mathbf{z})\}$.

Με άλλα λόγια ο $S_{cov}^{(k)}$ θα αποτελεί εκτίμηση ενός υπόχωρου που περιλαμβάνει κατευθύνσεις πλέον των όσων περιλαμβάνει ο $S_{yz}^{(k)}$, η COV_k δηλαδή θα υπερεκτιμά το στόχο της (Yin και Cook (2002)).

Σύμφωνα τέλος με την τρίτη υπόθεση που απαιτείται για την εφαρμογή του ελέγχου διατάξεων (*permutation test*) θα πρέπει $\mathbf{U}_1^T \mathbf{z} \perp\!\!\!\perp \mathbf{U}_2^T \mathbf{z}$. Εάν η υπόθεση αυτή δεν ισχύει,

τότε δεδομένου ότι $Cov(\mathbf{U}_1^T \mathbf{z}, \mathbf{U}_2^T \mathbf{z}) = 0$ υπάρχουν δύο ενδεχόμενα. Το πρώτο ενδεχόμενο είναι να ισχύει $E(\mathbf{U}_2^T \mathbf{z} | \mathbf{U}_1^T \mathbf{z}) = 0$ και η $\mathbf{U}_2^T \mathbf{z} | \mathbf{U}_1^T \mathbf{z}$ να εξαρτάται από την $\mathbf{U}_1^T \mathbf{z}$ μέσω κάποιας ανώτερης ροπής, και το δεύτερο είναι η $E(\mathbf{U}_2^T \mathbf{z} | \mathbf{U}_1^T \mathbf{z})$ να αποτελεί μη γραμμική συνάρτηση της $\mathbf{U}_1^T \mathbf{z}$. Εάν ισχύει η γραμμικότητα της $E(z | \gamma^T z)$ και επιπλέον $y \perp\!\!\!\perp \mathbf{U}_2^T \mathbf{z} | \mathbf{U}_1^T \mathbf{z}$ τότε συμβαίνει το πρώτο ενδεχόμενο γεγονός που δεν επηρεάζει την ισχύ των αποτελεσμάτων του ελέγχου διατάξεων. Εάν όμως δεν ισχύει η γραμμικότητα της $E(z | \gamma^T z)$ τότε θα ισχύει το δεύτερο ενδεχόμενο γεγονός που επηρεάζει δυσμενώς την ισχύ των αποτελεσμάτων του ελέγχου διατάξεων (Yin και Cook (2002)).

5.2.5 γ ΣΥΝΔΕΣΗ ΜΕΘΟΔΩΝ SIR και COV_k

Η σύνδεση των δύο μεθόδων επιτυγχάνεται μέσω της ακόλουθης πρότασης (Yin και Cook (2002)) η οποία συνδέει τον $S_{cov}^{(k)}$ και τον $S_{E(z|y)}$.

Πρόταση 5.4 (a) Αν η απόκριση y ορίζεται επί του πεπερασμένου συνόλου

$$R(y) = \{a_0, a_1, \dots, a_n\}$$

τότε

$$S_{E(z|y)} = \text{span}\{E(y^i z), i=1, \dots, n\} = S_{cov}^{(k)}$$

(b) Αν η απόκριση y και η $\mu_y = E(z|y)$ είναι συνεχείς επί του πεδίου ορισμού $R(y)$ τότε

$$S_{E(z|y)} = \text{span}\{E(y^i z), i=1, \dots, n\} = \lim_{k \rightarrow \infty} S_{cov}^{(k)}$$

Επομένως (Yin και Cook (2002)) για κάποια κατάλληλη τιμή του k οι δύο υπόχωροι $S_{E(z|y)}$ και $S_{cov}^{(k)}$ ταυτίζονται. Η τιμή αυτή που είναι η πλέον κατάλληλη σαν τιμή του αριθμού ζωνών h , δεν μπορεί να είναι άλλη από τη μέγιστη τάξη k των ροπών μέσω των οποίων η $y|z$ εξαρτάται από τη z . Επομένως όταν ενδιαφέρει η μελέτη της $y|z$ μέσω των χαμηλής τάξης ροπών αυτής επιλέγεται έξισου χαμηλός αριθμός ζωνών h , ενώ αντίθετα όταν ανιχνεύεται η πληροφορία για την $y|z$ που παρέχεται από υψηλής τάξης ροπές επιλέγεται εξισου υψηλός αριθμός ζωνών h .



Επίσης η παραπάνω πρόταση μπορεί να χρησιμεύσει για την κατανόηση της αυτίας για την οποία η *SIR* αποτυγχάνει στην ανίχνευση μέρους του κεντρικού υπόχωρου S_{yx} όταν η y είναι συμμετρική συνάρτηση του z (Yin και Cook (2002)). Σε μια τέτοια περίπτωση όπως για παράδειγμα όταν $y = z^2 + \varepsilon$, όπου $z \sim N(0,1)$ και

$z \perp\!\!\!\perp \varepsilon$, θα ισχύει $E(y|z) = 0$ πράγμα που σημαίνει ότι $S_{E(z,y)} = S_{cov}^{(k)} = S(0)$.

5.3 ΜΕΘΟΔΟΣ ΓΡΑΦΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ (*GRAPHICAL REGRESSION*)

Η μέθοδος που ακολουθεί αναλύεται από τους Cook και Weisberg (1999, K20) και δεδομένου ότι χρησιμοποιείται σε συνδυασμό με κάποια από τις άλλες μεθόδους μπορεί να θεωρηθεί ότι αποτελεί το επόμενο βήμα στην εκτίμηση της βάσης του κεντρικού υπόχωρου. Πρακτικά οι μέθοδοι των οποίων τα αποτελέσματα χρησιμοποιεί, μπορεί να είναι η *SAVE* και η *pHd* οι οποίες ανιχνεύουν την *OLS* κατεύθυνση ακόμα και όταν η απόκριση y παρουσιάζει συμμετρία στο *OLS* επίπεδο. Η μέθοδος της γραφικής παλινδρόμησης βασίζεται στη δυνατότητα γραφικής προσέγγισης της διάστασης του κεντρικού υπόχωρου για την οποία θα γίνει λόγος στην επόμενη παράγραφο, ενώ εφαρμόζεται με τη βοήθεια του προγράμματος Arc για το οποίο έγινε λόγος στην εισαγωγή. Τέλος να τονιστεί ότι για όσα ακολουθούν θεωρείται ότι ισχύει η υπόθεση της γραμμικότητας.

5.3.1 ΓΡΑΦΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΤΗΣ ΔΙΑΣΤΑΣΗΣ ΤΟΥ S_{yx}

Η ιδέα βασίζεται στο γεγονός ότι αν η μεταβλητή y εξαρτάται από τις μεταβλητές x_1, x_2 μόνο μέσω του γραμμικού συνδυασμού

$$h(\theta) = b(\cos\theta)x_1 + c(\sin\theta)x_2, \text{ όπου } b, c \text{ σταθερές}$$

ο οποίος απεικονίζεται από τον οριζόντιο άξονα του *2D* διαγράμματος (y, h) που προκύπτει από περιστροφή του *3D* διαγράμματος x_1, x_2 γύρω από τον άξονα *V* κατά γωνία θ , τότε το εν λόγω *2D* διάγραμμα θα παρουσιάζει τη μικρότερη διακύμανση γύρω από το μέσο. Επιπλέον δοθέντος ότι σε μία τέτοια περίπτωση $y \perp\!\!\!\perp h$ τότε το y θα είναι ανεξάρτητο οποιουδήποτε γραμμικού συνδυασμού των x_1, x_2 ο οποίος είναι ασυσχέτιστος με τον h . Με άλλα λόγια η προβολή των σημείων που αντιστοιχούν σε τιμές της y σε μία κάθετη ζώνη του *2D* διαγράμματος στο επίπεδο που σχηματίζεται από την y και το γραμμικό συνδυασμό h_{unc} ο οποίος είναι ασυσχέτιστος δηλαδή



ορθογώνιος με τον h , θα κατανέμονται με σταθερή διακύμανση γύρω από μία οριζόντια γραμμή, και βέβαια αυτό θα ισχύει για κάθε τέτοια ζώνη.

Επομένως η τεχνική που ακολουθείται για την εκτίμηση της διάστασης του S_{yx} έχει ως εξής.

Περιστρέφεται το $3D$ διάγραμμα γύρω από τον άξονα y και επιλέγεται το αντίστοιχο $2D$ διάγραμμα με την ελάχιστη δυνατή διακύμανση γύρω από το μέσο. Εάν στο σχετικό $2D$ διάγραμμα (y, h_{unc}) για τιμές της y σε μία κάθετη ζώνη, η y εμφανίζεται ανεξάρτητη του h_{unc} και αυτό ισχύει για κάθε ζώνη, τότε η διάσταση του S_{yx} είναι μονάδα και το διάγραμμα $(y, h(\theta))$ θα είναι ένα $2D$ διάγραμμα επαρκούς πληροφόρησης (*sufficient summary plot*).

Εάν όμως υπάρχει κάποια ζώνη για την οποία το σχετικό διάγραμμα δείχνει την ύπαρξη εξάρτησης τότε ο S_{yx} είναι διδιάστατος, ενώ εάν για κανένα $2D$ περιστροφής διάγραμμα δεν εμφανίζεται να υπάρχει εξάρτηση ανάμεσα στην y και στο h , τότε ο S_{yx} έχει μηδενική διάσταση.

Η παραπάνω διαδικασία υλοποιείται μέσω του προγράμματος Arc (Cook (1999), K 184).

5.3.2 ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΜΕΘΟΔΟΥ

Έστω $x = (x_1, \dots, x_p)^T$ και έστω οι τρεις μεταβλητές x_1, x_2 και x_3 όπου x_3 το σύνολο των υπολοίπων ($p-2$) μεταβλητών. Το ερώτημα στο οποίο καλείται να απαντήσει η μέθοδος είναι κατά πόσον οι x_1, x_2 μπορούν να αντικατασταθούν από μία άλλη μεταβλητή x_{12} που αποτελεί γραμμικό συνδυασμό των x_1, x_2 , δηλαδή

$$x_{12} = b_1 x_1 + b_2 x_2$$

ώστε

$$y \perp\!\!\! \perp x | (x_{12}, x_3)$$

Εάν αυτό μπορεί να γίνει και $b_1 = b_2 = 0$ τότε οι μεταβλητές x_1, x_2 δεν συνεισφέρουν στην εξήγηση της μεταβλητότητας της απόκρισης y και επομένως μπορούν να απαλειφθούν. Εάν αυτό μπορεί να γίνει και $b_1 \neq b_2 \neq 0$ τότε επαναλαμβάνεται η διαδικασία για τις μεταβλητές x_{12}, x_3 με σκοπό την περαιτέρω μείωση τους.



Εάν δεν μπορεί να γίνει αυτό τότε η διάσταση της $y|x$ θα είναι τουλάχιστον 2 και το αν θα είναι μεγαλύτερη από 2 ή όχι θα εξαρτηθεί από τη δυνατότητα μείωσης των μεταβλητών της x_3 .

Η διαδικασία σταματάει όταν δεν υπάρχει η δυνατότητα περαιτέρω μείωσης όταν δηλαδή εξαντληθούν τα ζεύγη των προς συνδυασμό μεταβλητών.

Ο έλεγχος της δυνατότητας συνδυασμού των μεταβλητών x_1, x_2 γίνεται μέσω διαγράμματος $3D$ *added variable plot* δηλαδή μέσω του $3D$ διαγράμματος των κατάλοιπων $\hat{e}(y | x_3)$ του μοντέλου $y | x_3 = \mathbf{n}^T \mathbf{u} + e$ σε σχέση με τα $(\hat{e}(x_1 | x_3), \hat{e}(x_2 | x_3))$, και με βάση την τεχνική που περιγράφηκε στην προηγούμενη παράγραφο της οποίας η παρούσα μέθοδος αποτελεί γενίκευση για περισσότερες των δύο ανεξάρτητες μεταβλητών. Αντί των αρχικών μεταβλητών x_1, \dots, x_p συνιστάται η χρήση των μεταβλητών $\hat{\beta}_1^T x, \dots, \hat{\beta}_p^T x$ όπου $\hat{\beta}_1, \dots, \hat{\beta}_p$ τα μετασχηματισμένα ιδιοδιανύσματα που αντιστοιχούν στις ιδιοτιμές της μήτρας που χρησιμοποιεί κάποια συγκεκριμένη μέθοδος για την εκτίμηση του κεντρικού υπόχωρου. Τα ιδιοδιανύσματα μετασχηματίζονται έτσι ώστε οι μεταβλητές $\hat{\beta}_1^T x, \dots, \hat{\beta}_p^T x$ να είναι ορθογώνιες.

Γενικά (Cook (1998)) η μήτρα

$$\mathbf{A} = \begin{bmatrix} a_1 & a_2 \\ c_1 & c_2 \end{bmatrix}$$

για την οποία οι συντεταγμένες του διανύσματος $z = Ax$ είναι ορθογώνιες είναι η μήτρα A για την οποία μεγιστοποιείται η ορίζουσα $|M^{-1}AA^TM^{-1}|$ όπου

$$M = diag(m_1, m_2)$$

$$m_1 = \max_i |a_1 x_{i1} + a_2 x_{i2}|$$

$$m_2 = \max_i |c_1 x_{i1} + c_2 x_{i2}|$$

υπό την προϋπόθεση ότι $xx^T = I$.

Εναλλακτικές προσεγγίσεις για την μήτρα A και περισσότερα επί του θέματος της ορθογωνιοποίησης μπορεί να αναζητήσεις ο αναγνώστη στον Cook (1998) και στους Cook και Weisberg (1990).

Η προαναφερθείσα συγκεκριμένη μέθοδος μπορεί να είναι γενικά οποιαδήποτε αν και οι Cook και Weisberg (1999, Κ 20) προτείνουν τη χρήση της *SAVE* ή της *pHd*.



σημαντικό πλεονέκτημα της χρήσης των $\hat{\beta}_1^T x, \dots, \hat{\beta}_p^T x$ σε σχέση με τη χρήση των x_1, \dots, x_p είναι ότι οι $\hat{\beta}_1^T x, \dots, \hat{\beta}_p^T x$ είναι διατεταγμένες κατά φθίνουσα σειρά σημαντικότητας δεδομένου ότι στα ιδιοδιανύσματα $\hat{\beta}_1, \dots, \hat{\beta}_p$ αντιστοιχούν ιδιοτιμές κατά φθίνουσα σειρά μεγέθους. Οι Cook και Weisberg (1990, K20) καλούν τις μεταβλητές αυτές *Fit*, *gr1, gr2, ..., grp*. Η *Fit* είναι η $b_{ols}^T x$ και το b_{ols} θα είναι το μοναδικό σημαντικό διάνυσμα βάσης του κεντρικού υπόχωρου σε περίπτωση που ο υπόχωρος αυτός είναι μονοδιάστατος σύμφωνα και με το *1D-estimation result*, ενώ οι γραμμικοί συνδυασμοί που προκύπτουν σαν αποτέλεσμα εφαρμογής της μεθόδου γραφικής παλινδρόμησης στις μεταβλητές *Fit*, *gr1, ..., grp* και της κατά το δυνατόν μείωσής τους αποτελούν μία πληρέστερη εκτίμηση της βάσης του κεντρικού υπόχωρου S_{yk} .

5.3.3 ΕΦΑΡΜΟΓΕΣ

ΕΦΑΡΜΟΓΗ 1 Γίνεται εφαρμογή της τεχνικής που περιγράφεται στην 5.3.1 στα δεδομένα που αφορούν τον όγκο V (απόκριση) 70 κωνοφόρων δένδρων, το ύψος τους H και τη διάμετρο D του κορμού στο ύψος του στήθους. Η εφαρμογή παρατίθεται από τους Cook και Weisberg (1999,K.18 1) ενώ τα δεδομένα υπάρχουν στο αρχείο pines. Lsp που διατίθεται με το Arc .

Το διάγραμμα 5.1 που ακολουθεί αποτελεί το $2D$ διάγραμμα κατ' εκτίμησην ελάχιστης διακύμανσης που προκύπτει από περιστροφή του $3D$ διαγράμματος $(V, (H, D))$.

Ο γραμμικός συνδυασμός που αντιστοιχεί στον οριζόντιο άξονα είναι

$$-1.913 + 0.007294H + 0.1017D$$



Rem lin trend

O to e(O|H)

aaa Scaling

OLS NIL
▼ [progress bar]

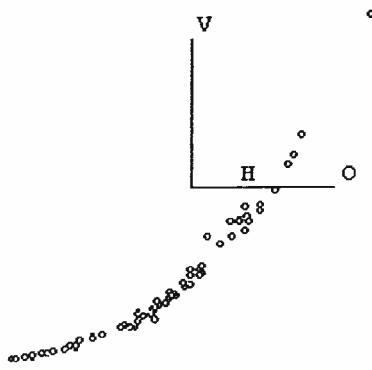
▼ Recall/Extract

▼ Case deletions

H: H

V: Vol

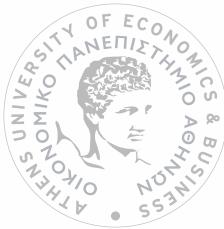
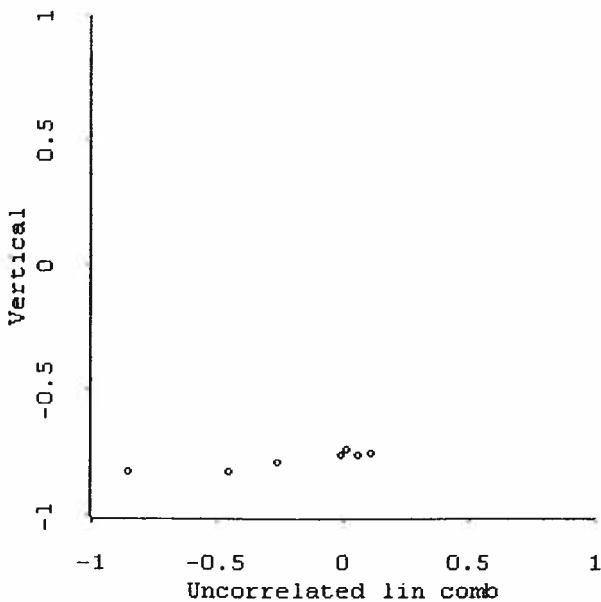
O: D



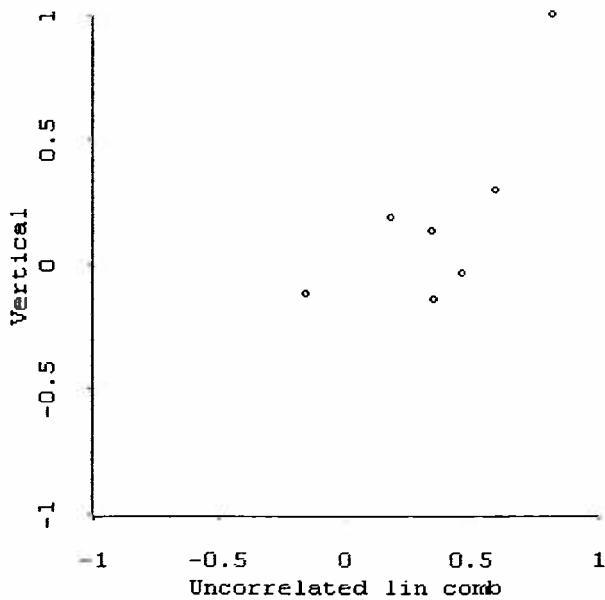
Rock Pitch Roll Yaw

Διάγραμμα 5.1 2D προβολή ελάχιστης διακύμανσης του 3D διαγράμματος για τις μεταβλητές H, Vol, D

Στη συνέχεια παρατίθενται κάποια από τα διαγράμματα (V, h_{unc}) η εικόνα των οποίων συνηγορεί υπέρ της απόρριψης της 1D για τον S_{jk} δεδομένου ότι στο διάγραμμα (V, h_{unc}) για τη ζώνη 10 φαίνεται να υπάρχει συσχέτιση μεταξύ V και h_{unc} .



Διάγραμμα 5.2 2D προβολή (V, h_{unc}) του 3D διαγράμματος για τις μεταβλητές $Vol, -1.913 + 0.007294H + 0.1017D, h_{unc}$ στη ζώνη 4 τιμών του y



Διάγραμμα 5.3 2D προβολή (V, h_{unc}) του 3D διαγράμματος για τις μεταβλητές Vol, h_{unc} στη ζώνη 10 τιμών του y

Τα παραπάνω διαγράμματα προέκυψαν με χρήση του Arc (Cook (1999), K 181).

ΕΦΑΡΜΟΓΗ 2 Η εφαρμογή αυτή Cook και Weisberg (1999) αφορά τα γνωστά δεδομένα για τα οστρακοειδή που χρησιμοποιήθηκαν στα πλαίσια της εφαρμογής 2 για τη μέθοδο SIR.

Το πρώτο βήμα αφορά τον έλεγχο ισχύος της υπόθεσης γραμμικότητας ο οποίος γίνεται μέσων της *scatterplot matrix*. Με χρήση της γνωστής μεθόδου Box-Cox για την επαγωγή πολυμεταβλητής κανονικότητας προκύπτει το διάνυσμα

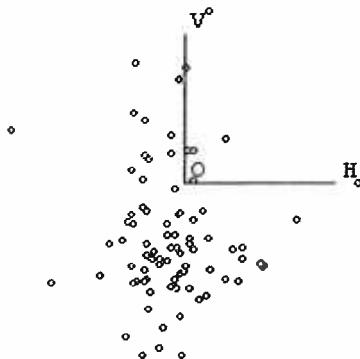
$x = (logH, logL, logW, logS)^T$ για το οποίο ισχύει η υπόθεση της γραμμικότητας. Οι συντελεστές των μεταβλητών $Fit, gr1, gr2, gr3$ παρατίθενται στον πίνακα 5.4 που ακολουθεί.

	<i>LogH</i>	<i>logL</i>	<i>logS</i>	<i>LogW</i>
<i>G1.Fit</i>	-0.252	-0.595	0.472	0.599
<i>G1.gr1</i>	0.906	0.322	-0.275	-0.017
<i>G1.gr2</i>	0.625	-0.743	0.111	-0.212
<i>G1.gr3</i>	0.092	-0.592	-0.139	0.789

Πίνακας 5.4 συντελεστές των μεταβλητών *Fit,gr1,gr2,gr3*

Στη συνέχεια παρατίθεται το διάγραμμα 5.4 το οποίο αποτελεί μία 2D προβολή του 3D - *AVP* για τις μεταβλητές *gr2,gr3* μετά την αφαίρεση της επίδρασης των *Fit* και *gr1*, από όπου φαίνεται ότι $M\ll(gr2, gr3)|(Fit, gr1)$ δεδομένου ότι το 3D - *AVP* παρουσιάζει την εικόνα ενός σφαιρικού νέφους σημείων, και επομένως οι μεταβλητές *gr2,gr3* μπορούν να απαλειφθούν.

- Rem lin trend
 O to e(O|H)
 aaa Scaling
 OLS NIL
 V
 Recall/Extract
 Case deletions
 H: e(G1.gr2|rest)
 V: e(M|rest)
 O: e(G1.gr3|rest)
 Greg methods



Rock Pitch Roll Yaw

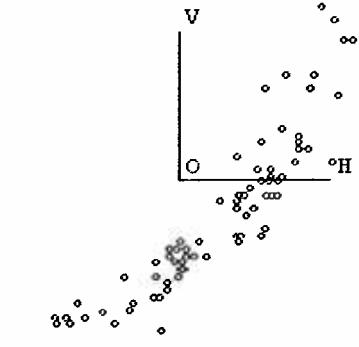
Διάγραμμα 5.4 2D προβολή του 3D - *AVP* για τις μεταβλητές *M,gr2,gr3* μετά την αφαίρεση της επίδρασης των *Fit,gr1*.

Τέλος ακολουθεί το διάγραμμα 5.5 το οποίο αποτελεί των 2D προβολή του 3D διαγράμματος για τις εναπομένουσες μεταβλητές *Fit, gr1* από όπου φαίνεται ότι $\dim S_{M,gr1,Fit} = 1$



Rem lin trend
 O to e(O|H)
 aaa Scaling
 OLS NIL

 Recall/Extract
 Case deletions
 H: G1.Fit
 V: M
 O: G1.gr1
 Greg methods



Rock Pitch Roll yaw

Διάγραμμα 5.5 2D προβολή ελάχιστης διακύμανσης του 3D διαγράμματος για τις μεταβλητές M , Fit , $gr1$

Επομένως $M \perp\!\!\!\perp x|gr4$ όπου

$$gr4 = b_0 Fit + b_1 gr1 = -0.167 logH - 0.624 logL + 0.563 logS + 0.516 logW$$

με

$$\hat{\beta} = (-0.167, -0.624, 0.563, 0.516)$$

το διάνυσμα βάσης του κεντρικού υπόχωρου $S_{M|x}$. Οι μεταβλητές $Fit, gr1, gr2, gr3$ προέκυψαν με χρήση της μεθόδου rHd και ο συντελεστής συσχέτισης μεταξύ της μεταβλητής $b_{ols}^T x$ που εκτιμά η μέθοδος OLS και της $gr4$ δείχνει ότι η κατεύθυνση που ανιχνεύει η γραφική μέθοδος επί των μεταβλητών rHd συμπίπτει σχεδόν με την κατεύθυνση OLS .

Όπως ήδη ειπώθηκε τα δεδομένα για την παρούσα εφαρμογή χρησιμοποιήθηκαν στα πλαίσια της εφαρμογής 2 για τη μέθοδο SIR , με τη διαφορά όμως ότι για το διάνυσμα x είχε θεωρηθεί $x = (L, W^{0.36}, S^{0.11})^T$. Θεωρώντας

$x = (logH, logL, logW, logS)^T$ και εφαρμόζοντας τη μέθοδο SIR προκύπτει η μοναδική κατεύθυνση

$$\hat{\beta} = (-0.733, 0.34, -0.406, -0.427).$$



Και στην περίπτωση αυτή για το x , προκύπτει ότι η κατεύθυνση SIR είναι πολύ κοντά στην OLS δοθέντος ότι ο συντελεστής συσχέτισης μεταξύ των αντιστοίχων μεταβλητών $\hat{\beta}^T x$ και $b_{ols}^T x$ είναι 0.987 και επομένως οι τρεις μέθοδοι ανιχνεύουν την ίδια κατεύθυνση.

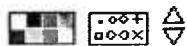
ΕΦΑΡΜΟΓΗ 3 Η εφαρμογή αυτή (Cook και Weisberg (1999)) αφορά δεδομένα τα οποία χρησιμοποιήθηκαν στα πλαίσια της εφαρμογής 1 για τη μέθοδο pHd .

Η εικόνα της *scatterplot matrix* δεν δείχνει να υπάρχει παραβίαση της υπόθεσης γραμμικότητας και έτσι δεν υπάρχει ανάγκη μετασχηματισμού των ανεξάρτητων μεταβλητών. Συνεπώς λαμβάνεται $x = (T_1, T_2, Lt_1, Lt_2, C)^T$ ενώ στον πίνακα 5.5 που ακολουθεί παρατίθενται οι συντελεστές των μεταβλητών $Fit, gr1, gr2, gr3, gr4$.

	T_1	T_2	Lt_1	Lt_2	C
$G1.Fit$	0.526	0.585	0.186	0.388	0.443
$G1.gr1$	-0.531	0.401	-0.269	-0.097	0.689
$G1.gr2$	-0.490	0.572	0.427	0.163	-0.473
$G1.gr3$	-0.355	-0.219	-0.622	0.650	-0.129
$G1.gr4$	0.177	0.460	-0.539	-0.637	-0.246

Πίνακας 5.5 συντελεστές των μεταβλητών $Fit, gr1, gr2, gr3, gr4$

Το διάγραμμα 5.6 που ακολουθεί είναι μία $2D$ προβολή του $3D - AVP$ για τις $gr3, gr4$ μετά την αφαίρεση της επίδρασης των υπολοίπων μεταβλητών $Fit, gr1, gr2$ από όπου φαίνεται ότι $y \perp\!\!\!\perp (gr3, gr4) | (Fit, gr1, gr2)$ δεδομένου ότι το $3D - AVP$ παρουσιάζει την εικόνα ασθενούς εξάρτησης μεταξύ των μεταβλητών και επομένως οι μεταβλητές $gr3, gr4$ μπορούν να απαλειφθούν. Ισχύει δηλαδή $dimS=0$ όπου S ο υπόχωρος της $(y | (Fit, gr1, gr2)) | ((gr3, gr4) | (Fit, gr1, gr2))$



Rem lin trend

O to e(O|H)

aaa Scaling

OLS

NIL



Recall/Extract

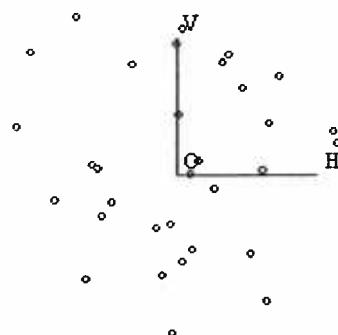
Case deletions

H: e(G1.gr3|rest)

V: e(y|rest)

O: e(G1.gr4|rest)

Greg methods



Rock Pitch Roll Yaw

Διάγραμμα 5.6 2D προβολή του 3D -AVP για τις μεταβλητές $y, gr3, gr4$ μετά την αφαίρεση της επίδρασης των $Fit, gr1, gr2$

Στη συνέχεια το διάγραμμα 5.7 αποτελεί την 2D προβολή του 3D - AVP για τις μεταβλητές $gr1, gr2$ μετά την αφαίρεση της επίδρασης της Fit από όπου φαίνεται ότι $\dim S=1$ όπου S ο υπόχωρος της $(y|Fit)|(gr1, gr2)|Fit$. Με άλλα λόγια

$$y \perp\!\!\!\perp (gr1, gr2) | Fit, gr5$$

όπου

$$gr5 = b_0 gr1 + b_1 gr2 = -0.531 T_1 + 0.401 T_2 - 0.269 Lt_1 - 0.097 Lt_2 + 0.689 C$$



Rem lin trend

O to e(O|H)

aaa Scaling

OLS

NIL



Recall/Extract

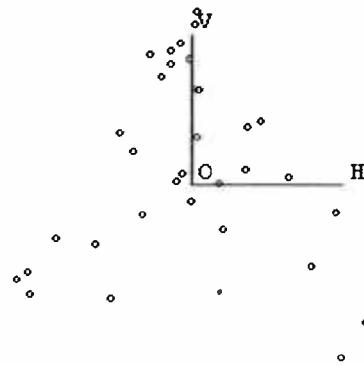
Case deletions

H: e(G1.gr1|rest)

V: e(y|rest)

O: e(G1.gr2|rest)

Greg methods



Rock Pitch Roll Yaw

Διάγραμμα 5.7 2D προβολή ελάχιστης διακύμανσης του 3D -AVP για τις μεταβλητές $y, gr1, gr2$ μετά την αφαίρεση της επίδρασης της Fit .

Τέλος από την εικόνα του 3D διαγράμματος της y ως προς τις εναπομένουσες μεταβλητές ($Fit, gr5$) φαίνεται ότι δεν μπορεί να υπάρξει περαιτέρω μείωση των μεταβλητών πράγμα που σημαίνει ότι $\dim S_{y(Fit, gr5)} = 2$. Επομένως

$$y \perp\!\!\!\perp x | (Fit, gr5)$$

και δοθέντος ότι

$$Fit = 0.526 T_1 + 0.585 T_2 + 0.186 Lt_1 + 0.388 Lt_2 + 0.443 C$$

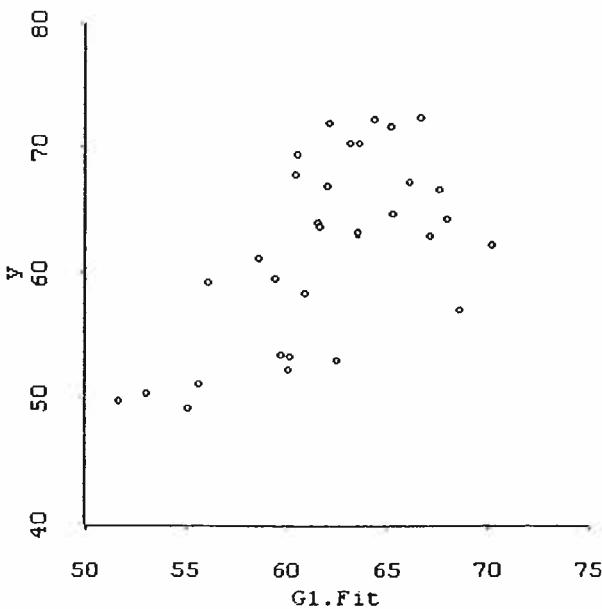
$$gr5 = -0.531 T_1 + 0.401 T_2 - 0.269 Lt_1 - 0.097 Lt_2 + 0.689 C$$

τα δύο διανύσματα βάσης του κεντρικού υπόχωρου S_{yk} θα είναι

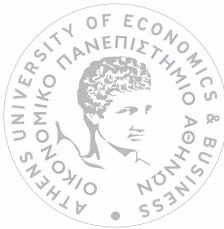
$$\hat{\beta}_1 = (0.526, 0.585, 0.186, 0.388, 0.443)^T$$

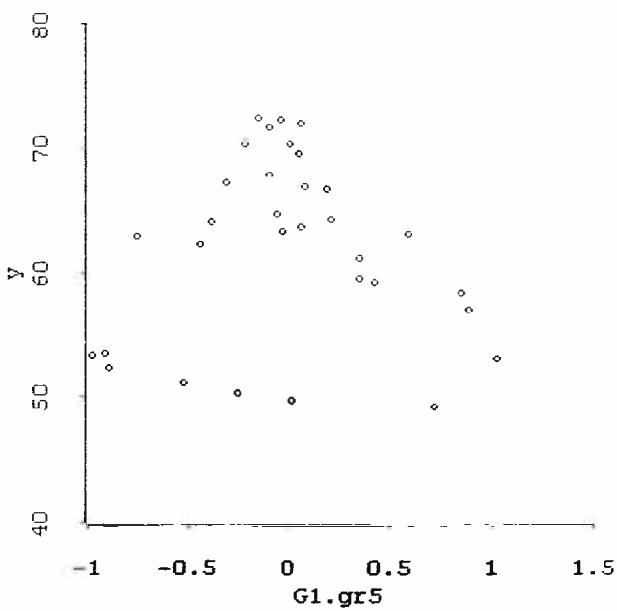
$$\hat{\beta}_2 = (-0.531, 0.401, -0.269, -0.097, 0.689)^T.$$

Τα δύο διαγράμματα 5.8 και 5.9 που ακολουθούν αποτελούν δύο 2D προβολές του παραπάνω 3D διαγράμματος. Το πρώτο είναι της y ως προς την Fit και το δεύτερο της y ως προς την $gr5$.



Διάγραμμα 5.8 2D προβολή $\{y, Fit\}$ του 3D διαγράμματος για τις μεταβλητές $y, Fit, gr5$.





Διάγραμμα 5.9 2D προβολή $\{y, gr5\}$ του 3D διαγράμματος για τις μεταβλητές $y, Fit, gr5$.

5.3.4 ΣΧΟΛΙΑ – ΕΠΙΣΗΜΑΝΣΕΙΣ

5.3.4 α ΧΡΗΣΗ ΚΑΤΑΛΟΙΠΩΝ ΤΕΤΡΑΓΩΝΙΚΗΣ ΠΡΟΣΑΡΜΟΓΗΣ (Cook και Weisberg (1999)).

Στα 3D -AVP τα οποία χρησιμοποιήθηκαν ως τώρα ο κάθετος άξονας έδινε τιμές των κατάλοιπων από τη γραμμική παλινδρόμηση της απόκρισης y στις μεταβλητές εκτός αυτών των οποίων εξετάζεται το ενδεχόμενο συνδυασμού τους. Έτσι αν οι μεταβλητές του προβλήματος είναι οι x_1, x_2, x_3 και εξετάζεται η δυνατότητα συνδυασμού των x_1, x_2 ο κάθετος άξονας παρέχει τιμές των $\hat{e}(y | x_3)$ από την προσαρμογή του μοντέλου

$$y | x_3 = \mathbf{n}^T \mathbf{u} + e \quad (5.4)$$

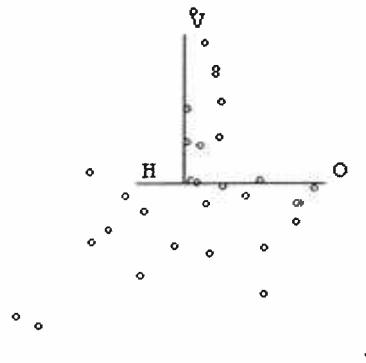
Εάν στις περιπτώσεις στις οποίες είναι εμφανής η μη γραμμικότητα της $E(y | x_3)$ χρησιμοποιηθούν αντί των κατάλοιπων από την προσαρμογή του παραπάνω μοντέλου τα κατάλοιπα από την προσαρμογή του πλήρως τετραγωνικού μοντέλου (full quadratic) που στην απλούστερη 2D μορφή του είναι

$$y | x_3 = a_0 + a_1(\mathbf{n}_1^T \mathbf{u}) + a_2(\mathbf{n}_2^T \mathbf{u}) + a_3(\mathbf{n}_1^T \mathbf{u})^2 + a_4(\mathbf{n}_2^T \mathbf{u})^2 + a_5(\mathbf{n}_1^T \mathbf{u})(\mathbf{n}_2^T \mathbf{u}) + e \quad (5.5)$$

τότε η εικόνα του διαγράμματος $3D -AVP$ μπορεί να γίνει σαφέστερη και τα συμπεράσματα είναι δυνατόν να διαφοροποιηθούν.

Η εφαρμογή 3 αποτελεί χαρακτηριστικό παράδειγμα. Έτσι αν ο κάθετος άξονας του $3D -AVP$ για τις $gr3, gr4$ δίνει τιμές των κατάλοιπων από την προσαρμογή του μοντέλου (5.5) αντί των κατάλοιπων από την προσαρμογή του μοντέλου (5.4) τότε όπως φαίνεται από το $2D$ διάγραμμα που ακολουθεί $dimS = 1$ αντί $dimS = 0$. Υπενθυμίζεται ότι S είναι ο υπόχωρος της $(y|(gr3, gr4))|(Fit, gr1, gr2)$.

- Rem lin trend
 O to e(O|H)
 aaa Scaling
 OLS NIL
 Recall/Extract
 Case deletions
 H: e(G1.gr3|rest)
 V: e(y|Quad rest)
 O: e(G1.gr4|rest)
 Greg methods



Rock Pitch Roll Yaw

Διάγραμμα 5.10 $2D$ προβολή του $3D -AVP$ για τις μεταβλητές $y, gr3, gr4$

μετά την αφαίρεση της επίδρασης των $Fit, gr1, gr2$ για το μοντέλο 5.5

5.3.4β ΧΡΗΣΗ ΤΗΣ ΜΕΘΟΔΟΥ ΓΙΑ ΤΟ ΔΙΑΓΝΩΣΤΙΚΟ ΕΛΕΓΧΟ ΜΟΝΤΕΛΟΥ (Cook και Weisberg (1999))

Η μέθοδος της γραφικής παλινδρόμησης μπορεί να χρησιμοποιηθεί και για το διαγνωστικό έλεγχο ενός υπό εξέταση μοντέλου για την περιγραφή της $y|x$. Έτσι αν οι μεταβλητές του προβλήματος είναι οι x_1, x_2, x_3 το $3D -AVP$ για τις μεταβλητές x_1, x_2 με κάθετο άξονα τα κατάλοιπα από την προσαρμογή του γραμμικού μοντέλου

$$L_1 | x_3 = \mathbf{n}_1^T \mathbf{u} + e$$

όπου L_1 τα κατάλοιπα από την προσαρμογή του υπό εξέταση μοντέλου, έστω



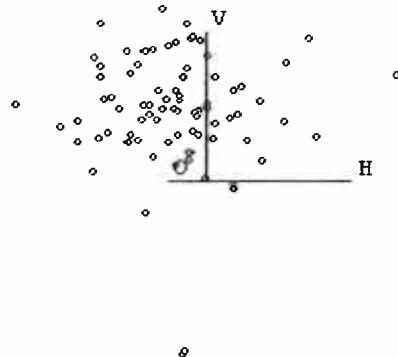
$$y|x = \mathbf{n}_2^T \mathbf{u} + e$$

και εφόσον βέβαια το υπό εξέταση μοντέλο είναι επαρκές, θα παρουσιάζει την εικόνα ενός περίπου σφαιρικού νέφους σημείων, πράγμα που θα ισχύει για όλα τα $3D$ -AVP και το οποίο σημαίνει $\dim S_{L_1x} = 0$. Αν κάποιο από τα $3D$ -AVP δεν παρουσιάζει αυτήν την εικόνα τότε $\dim S_{L_1x} \neq 0$ και άρα το υπό εξέταση μοντέλο είναι ανεπαρκές.

Πρέπει πάντως να επισημανθεί ότι κατά την χρήση των κατάλοιπων αντί της απόκρισης y με σκοπό το διαγνωστικό έλεγχο του υπό εξέταση μοντέλου, η μεταβλητή Fit δεν θα ανιχνεύεται από τη μέθοδο pHd εφόσον το μοντέλο είναι επαρκές όπως ήδη επώθηκε κατά την ανάλυση της μεθόδου. Για το λόγο αυτό το Arc δεν υπολογίζει τη μεταβλητή αυτή.

Εφαρμογή της μεθόδου στα δεδομένα της εφαρμογής 2 με σκοπό το διαγνωστικό έλεγχο του γραμμικού μοντέλου δείχνει ότι το εν λόγω μοντέλο είναι επαρκές. Ενδεικτικά παρατίθεται το ακόλουθο διάγραμμα 5.11 που είναι μία $2D$ προβολή του $3D$ -AVP για τις μεταβλητές $y, gr3, gr4$ μετά την αφαίρεση της επίδρασης των $gr1, gr2$ όπου $gr1, gr2, gr3, gr4$ οι ορθογωνιοποιημένες μεταβλητές που ανιχνεύει η pHd εφαρμοζόμενη επί των κατάλοιπων του μοντέλου $y|x = \mathbf{n}_2^T \mathbf{u} + e$.

-
- Rem lin trend
- O to e(O|H)
- aaa Scaling
- OLS NIL
▼
- Recall/Extract
- Case deletions
- H: e(G18.gr3|rest)
- V: e(L3.Residuals|rest)
- O: e(G18.gr4|rest)
- Greg methods

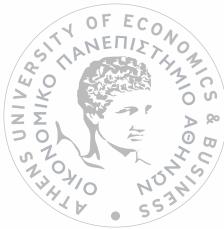


Rock Pitch Roll Yaw

Διάγραμμα 5.11 $2D$ προβολή του $3D$ -AVP για τις μεταβλητές $y, gr3, gr4$

μετά την αφαίρεση της επίδρασης των $gr1, gr2$ για τα κατάλουπα

του μοντέλου $y|x = \mathbf{n}_2^T \mathbf{u} + e$



5.3.4 γ ΧΡΗΣΗ ΤΩΝ ΑΡΧΙΚΩΝ ΜΕΤΑΒΛΗΤΩΝ $x = (x_1, \dots, x_p)^T$ (Cook και Weisberg (1999)).

Όπως ήδη ειπώθηκε είναι προτιμότερη η χρήση των ορθογωνιοποιημένων μεταβλητών που ανιχνεύει κάποια από τις μεθόδους εκτίμησης του κεντρικού υπόχωρου S_{yx} .

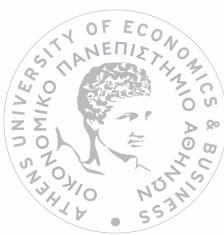
Ωστόσο δεν αποκλείεται η χρήση των αρχικών μεταβλητών $x = (x_1, \dots, x_p)^T$. Στην περίπτωση αυτή συνιστάται η εξέταση όλων των 2D -AVP και η διερεύνηση αρχικά της δυνατότητας συνδυασμού των μεταβλητών που εμφανίζουν την ισχυρότερη εξάρτηση στο αντίστοιχο 2D - AVP. Η διαδικασία υλοποιείται μέσω του Arc.





REFERENCES

- Breiman, L., and Friedman, J. (1985).** Estimating Optimal transformations for Multiple Regression and Correlation,
Journal of the American Statistical Association,80, 580-597.
- Bunke, H. and Bunke, O. (eds) (1986).** Statistical Inference in Linear Models, Vol. I, Statistical Methods of Model building.Chichester: Wiley.
- Bura, E. & Cook, R. D. (2001a).** Extending SIR:The Weighted Chi-Square Test,
Journal of the American Statistical Association, 96, 996-1003.
- Bura, E. and Cook, R.D (2001b).** Estimating the Structural Dimension of Regressions via Parametric inverse regression,
Journal of the Royal Statistical Society , B 63, Part 2, 393-410.
- Chen, H. (1988).** Convergence Rates for Parametric Components in a Partly Linear Model , The Annals of Statistics, 16, 136-146.
- Chiaromonte F., Cook, R.D., and Li, B (2002).** Sufficient Dimension Reduction in Regressions With Categorical Predictors,Annals of Statistics, 30, 475-97.
- Cook, R.D. (1998).** Regression Graphics: Ideas for studying Regression Through Graphics. New York:Wiley.
- Cook, R.D. (1998b).** Principal Hessian Directions Revisited (with discussion),
Journal of the American Statistical Association,93, 84-100.
- Cook, R. D (2000).** Using Arc for Dimension Reduction and Graphical Exploration in Regression, School of Statistics, 1994Buford Ave., University of Minnesota, St. Paul, MN 55108,USA.
- Cook, R.D., and Lee, H. (1999).** Dimension Reduction in Regressions with a Binary Response,
Journal of the American Statistical Association,94, 1187-1200.
- Cook, R. D. and Nachtsheim, C.J. (1994).** Reweighting to Achieve Elliptically Contoured Covariates in Regression,
Journal of the American Statistical Association,89, 592-599.
- Cook, R.D. and Weisberg, S (1983).** Diagnostics for Heteroscedasticity in Regression, Biometrika, 70, 1-10.



- Cook, R.D., and Weisberg, S. (1990).** Three Dimensional Residual Plots,
 In K. Berk and L. Malone (Eds)
 Proceedings of the 21st Symposium On the Interface:
 Computing Science and Statistics, pp.162-166.
 Washington: American Statistical Association
- Cook, R.D. and Weisberg, S.(1991).** Comment on “Sliced inverse regression”
 (by K.C. Li),Journal of the American Statistical Association, 86, 328-332.
- Cook, R.D. and Weisberg, S. (1999).** Applied Regression including Computing
 and Graphics, New York: Wiley.
- Cook, R.D., and Yin, X. (2001).** Dimension Reduction and Visualization in
 Discriminant Analysis (with discussion),
 Anst. New Zeal. S. Statist. , 43, 147-199.
- Cuzik, J. (1987).** “Semiparametric Additive Regression”, αδημοσίευτο χειρόγραφο.
- Diaconis, P. and Freedman, D. (1984).** Asymptotics of Graphical Projection Pursuit,
 The Annals of Statistics, 12. 793-815.
- Eaton, M.L. (1986).** A Characterization of Spherical Distributions,
 Journal of Multivariate Analysis, 20, 272-276.
- Farebrother, R. (1990).** The distributions of a Quadratic Form in Normal Variables,
 Applied Statistics, 39, 294-309.
- Field, C. (1993).** Tail Areas of Linear Combinations of Chi-Squared and Non Central
 Chi-Squared, Journal of Statistical Computation and Simulation, 45, 243-248.
- Freud, R.J. (1979).** Proc. Statist. Comput. Sect. Am. Statist. Ass., 111-112.
- Friedman, J., and Stuetzle, W. (1981).** Projection pursuit regression,
 Journal of the American Statistical Association, 76, 817-823.
- Härdle, W. and Tsybakov, A.B. (1991).** Comment on “Sliced inverse regression”
 (by K.C. Li),Journal of the American Statistical Association, 86, 333-335.
- Hastie, T. and Tibshirani, R. (1986).** Generalized Additive Models,
 Statistical Science, 1, 1 297-318.
- Heckman, N. (1986).** Spline Smoothing in Party Linear Models,
 Journal of the Royal Statistical Society Ser. B, 48, 244-248.
- Johnson, M.E. (1987).** Multivariate Statistical Simulation. New York: Wiley.
- Li,K.C. (1990a).**On Principal Hessian Directions for Data Visualization and
 Dimension Reduction: another application of Stein’s lemma,
 UCLA technical report, Dept., of Mathematics.



- Li, K.C. (1990b).** Uncertainty Analysis for Mathematical Models with *SIR*,
UCLA technical report, Dept. of Mathematics.
- Li,K.C. (1991a).** Sliced Inverse Regression for Dimension Reduction,
Journal of the American Statistical Association ,86, 316-327.
- Li,K.C. (1991b).** Rejoinder on “Sliced inverse regression” (by K.C. Li)
Journal of the American Statistical Association , 86, 337-342.
- Li, K.C. (1992).** On Principal Hessian Directions for Data Visualization and
Dimension Reduction:Another application of Stein’s lemma,
Journal of the American Statistical Association,87, 1025-1039.
- Li, K.C. and Duan, N. (1989).** Regression Analysis under Link Violation,
The Annals of Statistics, 17, 1009-1052.
- Schott , J. (1994).** Determining the Dimensionality in Sliced inverse regression,
Journal of the American Statistical Association,89, 141-148.
- Seber, G.A.F. (1977).** Linear Regression Analysis. New York:Wiley.
- Speckman,P.(1987).** Kernel Smoothing in Partial Linear Models, αδημοσίευτο
χειρόγραφο.
- Stein, C. (1981).** Estimation of the Mean of a Multivariate Normal Distribution,
The Annals of Statistics, 9, 1135-1151.
- Stone, C. (1986).** The Dimensionality Reduction Principle for Generalized
Additive Models, The Annals of Statistics, 13, 689-705.
- Vellila, S. (1998).** Assesing the Number of Linear Components in a General
Regression Problem, Journal of the American Statistical Association,93, 1088-
1089.
- Wood, A. (1989).** An F-Approximation to the Distribution of a Linear
Combination of Chi-Squared Random Variables.
Communication in Statistics, Part B-Simulation and
Computation,18,1439-1456.
- Yin, X., and Cook, R.D. (2002).** Dimension Reduction for the Conditional k^{th}
Moment in Regression,
Journal of the Royal Statistical Society,B, 64. Part 2,159-175.
- Yin, X., and Cook, R.D. (2003).** Estimating Central Subspaces via Inverse
Third Moments, Biometrika, 90, 1 , pp 113-125.

