ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ PLS

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΑΥΡΟΥΤΣΟΣ ΝΙΚΟΛΑΟΣ

Επιβλέπουσα: Καρώνη Χρυσηίς, Καθηγήτρια Ε.Μ.Π

Αθήνα, Οκτώβριος 2014

ABSTRACT

Partial Least Squares (PLS) is a commonly used method for predictive modelling, mostly applicable in the manufacturing industry. In this industry, the number of factors is often large and correlations between factors occur. In such cases, when the main goal of the analyst is prediction rather than the analysis of the relationship between variables, PLS is recommended as a very useful tool.

As far as the present work is concerned, the focus falls particularly on the statistical study of multicollinearity problems in regression analysis and on data analysis with PLSR models. In the first chapter, reference is made to the causes and effects of multicollinearity and to methods of detecting the presence of multicollinearity. Principal Component Regression (PCR) and Ridge Regression (RR) are also proposed as methods for dealing with multicollinearity. In the second chapter, the PLSR model is described as well as methods for selecting models with the highest predictive value. The second chapter also gives a brief overview of how PLSR works, relating it to the preceding multivariate techniques (PCR and RR). Finally, in the third chapter, three examples are presented that demonstrate how PLSR models are evaluated and how their components are interpreted.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά την Καθηγήτρια
του Εθνικού Μετσόβιου Πολυτεχνείου κ. Καρώνη Χρυσηίς
για την συνεχή ενθάρρυνση, συμπαράσταση και εμπιστοσύνη
που μου έδειξε σε όλη τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας.
Επίσης, θα ήθελα να ευχαριστήσω τους γονείς μου
για την πολύτιμη βοήθεια τους καθ' όλη την διάρκεια των σπουδών μου.
Τέλος θα ήθελα να αφιερώσω όλη αυτή την προσπάθεια
στην σύζυγο μου Ζωή και τον γιό μου Παναγιώτη
και να τους ευχαριστήσω
για την υπομονή που έδειξαν αυτά τα δύο χρόνια.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1
Πολυσυγγραμμικότητα

1.1 Παλινδρόμηση

Σε διάφορα Στατιστικά προβλήματα, συνήθως δεν είναι αρκετό να ασχολούμαστε με την μελέτη των χαρακτηριστικών κάθε μεταβλητής ξεχωριστά (μέση τιμή, διασπορά κ.λ.π.). Σε αρκετές περιπτώσεις, εξίσου ενδιαφέρουσα είναι και η ταυτόχρονη μελέτη δύο ή περισσοτέρων μεταβλητών για να προσδιορίσουμε με ποιο τρόπο οι μεταβλητές αυτές σχετίζονται μεταξύ τους. Για παράδειγμα:

(α) Η ηλικία και το βάρος ενός παιδιού είναι προφανές ότι έχουν κάποια θετική εξάρτηση μεταξύ τους. Δηλαδή όσο πιο μεγάλο είναι το παιδί τόσο πιο μεγάλο βάρος θα έχει.

(β) Η συνολική παραγωγή σιταριού εξαρτάται από τη θέση του χωραφιού, την ποσότητα λιπάσματος, την επίδραση της θερμοκρασίας κ.α.

(γ) το ύψος των μηνιαίων αποδοχών των υπαλλήλων μιας εταιρείας δεν εξαρτάται από το βάρος τους.

Στη στατιστική μια από τις πιο γνωστές τεχνικές που χρησιμοποιείται για την ανάλυση δεδομένων με πολλές μεταβλητές είναι η Ανάλυση Παλινδρόμησης *(Regression Analysis)*. Ειδικότερα μπορούμε να πούμε ότι η Ανάλυση Παλινδρόμησης είναι εκείνη η τεχνική που ερευνά και μοντελοποιεί τη σχέση μεταξύ μεταβλητών. Για τον σκοπό αυτό χρησιμοποιούμε μια ισότητα η οποία προκύπτει από μια λογική διαδικασία έτσι ώστε να εκφράσουμε τη σχέση μεταξύ μιας μεταβλητής που μας ενδιαφέρει και ενός συνόλου άλλων μεταβλητών.

Σε κάθε πρόβλημα παλινδρόμησης διακρίνουμε συνήθως δύο είδη μεταβλητών: τις ανεξάρτητες *(independent, predictor, regressor variable)* που τις συμβολίζουμε με x και τις εξαρτημένες *(response variables)* που τις συμβολίζουμε με y. Όπως θα διαπιστώσουμε και παρακάτω η προσαρμογή ενός μοντέλου παλινδρόμησης απαιτεί πολλές και χρονοβόρες πράξεις. Για το λόγο αυτό οι υπολογιστές παίζουν σημαντικό ρόλο στην εφαρμογή της παλινδρόμησης με τη χρήση στατιστικών πακέτων *(MINITAB, R, Statgraphics)*.

Η πιο απλή περίπτωση παλινδρόμησης είναι η απλή γραμμική παλινδρόμηση όπου υπάρχει μια μόνο ανεξάρτητη μεταβλητή x η οποία προσεγγίζει γραμμικά την εξαρτημένη μεταβλητή y. Όταν όμως χρησιμοποιούμε περισσότερες από μια ανεξάρτητες μεταβλητές έτσι ώστε με κάποιο γραμμικό συνδυασμό τους να εκφράσουμε την εξαρτημένη μεταβλητή y τότε έχουμε την πολλαπλή γραμμική παλινδρόμηση.

1.2 Πολλαπλή γραμμική παλινδρόμηση

Στόχος της μελέτης ενός γραμμικού μοντέλου είναι η πρόβλεψη μιας μεταβλητής Y με βάση τα στοιχεία που διαθέτουμε για ένα σύνολο άλλων μεταβλητών $X_1, X_2, X_3, \dots, X_k$ καθώς και η διαπίστωση κατά πόσο οι μεταβλητές αυτές συμμετέχουν στην διαμόρφωση των τιμών της Y. Στο κλασικό γραμμικό μοντέλο η Y θεωρείται τυχαία μεταβλητή, η οποία αποτελείται από ένα γραμμικό κομμάτι που περιέχει τα $X_1, X_2, X_3, \dots, X_k$ και ένα τυχαίο σφάλμα $\boldsymbol{\varepsilon}$. Οι τιμές που παίρνουν τα X_j θεωρούνται μη στοχαστικές ενώ τα σφάλματα $\boldsymbol{\varepsilon}$ τυχαίες μεταβλητές που ικανοποιούν κάποιες στατιστικές υποθέσεις.

Η γενική μορφή του γραμμικού μοντέλου είναι:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + \boldsymbol{\varepsilon}$$

Ειδικότερα αν έχουμε ένα δείγμα n ανεξάρτητων παρατηρήσεων με k ανεξάρτητες μεταβλητές η προηγούμενη εξίσωση για την i παρατήρηση γίνεται:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_k x_{ik} + \varepsilon_i \text{ για } \quad 1 \le i \le n \qquad (1.2.1)$$

όπου:

y_i: η i τιμή της εξαρτημένης μεταβλητής *(response)* Y για $1 \le i \le n$

x_{ij}: η τιμή της ανεξάρτητης μεταβλητής X_j για την i παρατήρηση με $1 \le i \le n$ και $1 \le j \le k$

ε_i: το τυχαίο σφάλμα της i παρατήρησης από την αναμενόμενη τιμή της $E(y_i)$

Επίσης η σχέση $(1.2.1)$ μπορεί να γραφεί και σε μορφή πίνακα ως εξής:

$$\boldsymbol{y} = X \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

με $\boldsymbol{y} = (y_1, y_2, \dots\dots, y_n)', \boldsymbol{\beta} = (\beta_0, \beta_1, \dots\dots, \beta_k)', \boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_1, \dots\dots., \varepsilon_n)'$ και X ο πίνακας $n \times (k+1)$ που είναι της μορφής:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

και ονομάζεται πίνακας σχεδιασμού *(Design Matrix)*

Τα τυχαία σφάλματα ε_i επίσης, ως τυχαίες μεταβλητές θα πρέπει να ακολουθούν την Κανονική κατανομή με:

1. $E(\varepsilon_i) = 0$
2. $V(\varepsilon_i) = \sigma^2$ (δηλαδή να έχουν σταθερή διασπορά ή όπως λέμε να ισχύει η υπόθεση της ομοσκεδαστηκότητας)
3. $Cov(\varepsilon_i, \varepsilon_j) = 0$ όπου $1 \leq i, j \leq n$

Εδώ θα πρέπει να πούμε ότι ο όρος «γραμμικός» αναφέρεται στους άγνωστους συντελεστές $\beta_0, \beta_1, \dots \dots, \beta_k$ και όχι στις ανεξάρτητες μεταβλητές $X_1, X_2, X_3, \dots ., X_k$.

Γίνεται κατανοητό ότι ο υπολογισμός ενός γραμμικού μοντέλου αρχικά πραγματοποιείται με την εκτίμηση των παραμέτρων $\beta_0, \beta_1, \dots \dots, \beta_k$ των ανεξάρτητων μεταβλητών. Έπειτα με την βοήθεια συγκεκριμένων στατιστικών ελέγχων αλλά και άλλων τεχνικών που θα δούμε στην συνέχεια, μπορούμε να το εξετάσουμε ως προς την καταλληλότητα του και να το βελτιώσουμε.

1.2.1 Εκτίμηση των παραμέτρων $\boldsymbol{\beta_0, \beta_1, \dots \dots, \beta_k}$

Για την εκτίμηση των $\beta_0, \beta_1, \dots \dots, \beta_k$ εφαρμόζουμε την μέθοδο των ελαχίστων τετραγώνων *(OLS - Ordinary Least-Squares)* έτσι ώστε το άθροισμα τετραγώνων

$$\sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}' \cdot \boldsymbol{\varepsilon}$$

των τυχαίων σφαλμάτων να ελαχιστοποιείται. Και τελικά καταλήγουμε στο συμπέρασμα ότι οι εκτιμήτριες ελαχίστων τετραγώνων

$\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots \dots, \hat{\beta}_k)$ υπολογίζονται από την λύση της εξίσωσης:

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1} \cdot X' \cdot \boldsymbol{y} \qquad (1.2.2)$$

Αρκεί βέβαια ο πίνακας $X'X$ να είναι αντιστρέψιμος.

Έτσι το γραμμικό μοντέλο δίνεται από την σχέση $\widehat{\boldsymbol{y}} = X \cdot \widehat{\boldsymbol{\beta}}$

και τα υπόλοιπα *(residuals)* που προκύπτουν από τις n παρατηρήσεις από την σχέση

$$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$$

Εύκολα αποδεικνύεται ότι το $\widehat{\boldsymbol{\beta}}$ είναι αμερόληπτη εκτιμήτρια του $\boldsymbol{\beta}$ δηλαδή

$$E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

Αν εκτιμήσουμε το $\boldsymbol{\beta}$ με τη μέθοδο μέγιστης πιθανοφάνειας θα καταλήξουμε πάλι στην εξίσωση (1.2.2). Το ίδιο θα συμβεί και για την εκτίμηση της διασποράς των σφαλμάτων σ^2 όπου η εκτιμήτρια $\hat{\sigma}^2 = \frac{\boldsymbol{e}'\cdot\boldsymbol{e}}{n}$ είναι μια μεροληπτική εκτιμήτρια του σ^2 ενώ η $S^2 = \frac{SSE}{n-k-1}$ είναι αμερόληπτη εκτιμήτρια.

Αξίζει να αναφέρουμε ότι για την εκτιμήτρια ελαχίστων τετραγώνων του $\boldsymbol{\beta}$ ισχύει το θεώρημα των *Gauss – Markov* σύμφωνα με το οποίο, η παραπάνω εκτιμήτρια είναι η αμερόληπτη γραμμική εκτιμήτρια του $\boldsymbol{\beta}$ με την ελάχιστη δυνατή διασπορά. Γι' αυτό το λόγο οι εκτιμήτριες ελαχίστων τετραγώνων χαρακτηρίζονται ως *BLUE (Best Linear Unbiased Estimators).*

Βρίσκοντας την εκτιμήτρια του διανύσματος $\boldsymbol{\beta}$ δηλαδή τους συντελεστές των ανεξάρτητων μεταβλητών θα πρέπει να εξεταστεί αν το μοντέλο στο οποίο καταλήξαμε είναι κατάλληλο. Οι τρόποι με τους οποίους γίνεται κάτι τέτοιο περιγράφονται στις επόμενες παραγράφους.

1.2.2 Συντελεστής προσδιορισμού (R^2) και έλεγχοι t και F

Για τον έλεγχο της καλής προσαρμογής του μοντέλου μπορεί να χρησιμοποιηθεί το πηλίκο

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

το οποίο εκφράζει το ποσοστό της ολικής μεταβολής του Y που εξηγείται από την παλινδρόμηση. Το R^2 λέγεται συντελεστής προσδιορισμού και εκφράζεται σε ποσοστό επί τις εκατό. Οπότε:

$$0 \leq R^2 \leq 1$$

Αν $R^2 = 0$ σημαίνει ότι η εξαρτημένη μεταβλητή δεν μπορεί να προβλεφθεί από τις ανεξάρτητες μεταβλητές δηλαδή $\beta_1 = \beta_2 = \cdots = \beta_k = 0$. Ενώ αν $R^2 = 1$ σημαίνει ότι όλες οι εκτιμημένες τιμές συμπίπτουν με τις πραγματικές τιμές (τέλεια προσαρμογή). Επομένως γίνεται φανερό ότι ο συντελεστής προσδιορισμού θέλουμε να πλησιάζει το 1. Το R^2 όπως θα δούμε παρακάτω χρησιμοποιείται και για την επιλογή των κατάλληλων επεξηγηματικών μεταβλητών.

Για τους συντελεστές $\boldsymbol{\beta}$ πραγματοποιούνται οι έλεγχοι t και F όπου προσπαθούμε να διαπιστώσουμε αν κάποιοι από αυτούς τους συντελεστές είναι μηδέν. Αυτό θα σημαίνει ότι οι τιμές της εξαρτημένης μεταβλητής *(response)* δεν επηρεάζονται από τις μεταβλητές *(regressors)* εκείνες, των οποίων οι συντελεστές είναι μηδέν.

Με δεδομένο πάντα ότι τα τυχαία σφάλματα $\boldsymbol{\varepsilon}$ ακολουθούν την πολυδιάστατη Κανονική κατανομή, $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I)$ τότε οι τιμές $\boldsymbol{y} = X \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$ ακολουθούν και αυτές την πολυδιάστατη Κανονική κατανομή, $\boldsymbol{y} \sim N_n(X \cdot \boldsymbol{\beta}, \sigma^2 I)$ όπως και οι εκτιμήσεις των συντελεστών $\boldsymbol{\beta}$ δηλαδή $\widehat{\boldsymbol{\beta}} \sim N_{k+1}(\boldsymbol{\beta}, \sigma^2 (X'X)^{-1})$. Αυτό σημαίνει ότι και η κατανομή των στοιχείων του διανύσματος $\boldsymbol{\beta}$ ακολουθούν τη μονοδιάστατη Κανονική κατανομή δηλαδή $\hat{\beta}_j \sim N(\beta_j, \sigma^2 d_{jj})$ όπου d_{jj}το j-οστό διαγώνιο στοιχείο του πίνακα $(X'X)^{-1}$.

Για τον έλεγχο t εξετάζουμε τις υποθέσεις:

H_0: $\beta_j = 0$ έναντι της H_1: $\beta_j \neq 0$

με ελεγχοσυνάρτηση την

$$t = \frac{\beta_j}{se(\beta_j)} \sim t_{n-k-1}(Wald) \qquad \text{όπου} \quad se(\beta_j) = S\sqrt{d_{jj}}$$

Για τον έλεγχο F εξετάζουμε τις υποθέσεις:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ έναντι της

H_1: $τουλάχιστον\ ένα\ από\ τα\ \beta_j \neq 0$

με ελεγχοσυνάρτηση την

$$F = \frac{SSR/_k}{S^2} \sim F_{k,n-k-1}$$

Με αυτό τον τρόπο ελέγχουμε αν κάποιος από τους συντελεστές είναι μηδέν (έλεγχος t) ή αν υπάρχει τουλάχιστον ένας συντελεστής διάφορος του μηδενός (έλεγχος F). Οι έλεγχοι t δεν θεωρούνται αξιόπιστοι διότι ο έλεγχος για το αν ένας συντελεστής είναι μηδέν, γίνεται με δεδομένο ότι οι υπόλοιποι είναι διαφορετικοί του μηδενός. Αυτό συμβαίνει όταν δύο συντελεστές έχουν σημαντική συσχέτιση με την εξαρτημένη μεταβλητή αλλά και μεταξύ τους. Λαμβάνοντας λοιπόν υπόψη, ότι οι έλεγχοι t μπορεί να μην δώσουν αξιόπιστα αποτελέσματα και αν απορριφθεί η μηδενική υπόθεση στον έλεγχο F, θα θέλαμε να γνωρίζουμε ποιοι από τους συντελεστές είναι διάφοροι του μηδενός. Κάτι τέτοιο θα μας βοηθούσε στην επιλογή των καταλληλότερων επεξηγηματικών μεταβλητών. Αναφορικά, μερικές μέθοδοι με τις οποίες μπορούμε να ελέγξουμε αν κάποιος από τους συντελεστές β_j είναι μηδέν είναι:

α. Η διαδικασία της διαδοχικής αφαίρεσης *(Backward Elimination)*

β. Η διαδικασία διαδοχικής πρόσθεσης ή της προς τα εμπρός επιλογής *(Forward Selection)*

γ. Κατά βήματα εμπρός πίσω επιλογή *(Stepwise selection)*

1.2.3 Έλεγχος των υπολοίπων

Όπως έχει ήδη αναφερθεί απαραίτητη προϋπόθεση για τη δημιουργία ενός αξιόπιστου γραμμικού μοντέλου είναι να τηρούνται οι τρεις συνθήκες για τα σφάλματα που αναφέραμε παραπάνω. Δηλαδή:

1. $E(\varepsilon_i) = 0$
2. $V(\varepsilon_i) = \sigma^2$
3. $Cov(\varepsilon_i, \varepsilon_j) = 0$ όπου $1 \leq i, j \leq n$

Και γενικότερα θα πρέπει $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I)$

Επειδή, όπως είναι φυσικό, δεν μπορούμε να γνωρίζουμε το διάνυσμα των σφαλμάτων $\boldsymbol{\varepsilon}$ γι' αυτό καταφεύγουμε στον έλεγχο των υπολοίπων $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ τα οποία μπορούμε να υποθέσουμε ότι αποτελούν μια εκτίμηση των σφαλμάτων $\boldsymbol{\varepsilon}$. Διακρίνουμε τέσσερα είδη σφαλμάτων όπου εκτός από τα παραπάνω, που ονομάζονται συνήθη υπόλοιπα *(regular residuals)*, είναι τα τυποποιημένα υπόλοιπα *(standardized residuals)*, τα *"Deleted"* υπόλοιπα και τα υπόλοιπα *PRESS*.

Η ύπαρξη διαφόρων ειδών υπολοίπων οφείλεται στο γεγονός ότι τα κανονικά υπόλοιπα ενδέχεται να παρουσιάζουν ετεροσκεδαστικότητα και αυτή να μην οφείλεται σε κακή προσαρμογή του μοντέλου. Για να το ξεπεράσουμε λοιπόν αυτό, απαιτούμε τα υπόλοιπα να έχουν την ίδια κατανομή. Τα υπόλοιπα *"Deleted"* και *PRESS* είναι ειδικές περιπτώσεις των τυποποιημένων υπολοίπων και των συνήθων υπολοίπων αντίστοιχα και χρησιμοποιούνται για πιο ασφαλείς ελέγχους όταν κρίνεται αναγκαίο.

Υπολογίζοντας αυτά τα υπόλοιπα ο έλεγχος των αρχικών προϋποθέσεων γίνεται χρησιμοποιώντας τις κατάλληλες γραφικές παραστάσεις που είναι:

1. Γραφικός έλεγχος της κανονικότητας των υπολοίπων *(Normal Probability Plot)*.
2. Γραφική παράσταση των e_i και των $\hat{y}_i$.
3. Γραφική παράσταση για τον έλεγχο της συσχέτισης των υπολοίπων.
4. Γραφική παράσταση για τον έλεγχο ανεξαρτησίας των e_i από κάθε μία επεξηγηματική μεταβλητή x ξεχωριστά.

Γενικά αν κάποια από τις προϋποθέσεις, φαίνεται να μην ισχύει μπορούμε με κατάλληλο μετασχηματισμό του $\boldsymbol{y}$ ή του X να αντιστρέψουμε την εικόνα.

1.2.4 Τρόποι επιλογής των κατάλληλων επεξηγηματικών μεταβλητών

Στην αρχή της μελέτης ενός προβλήματος παλινδρόμησης ενδέχεται να μην γνωρίζουμε ποιες από τις επεξηγηματικές μεταβλητές συμμετέχουν πραγματικά στην πρόβλεψη της εξαρτημένης μεταβλητής. Αυτό έχει ως αποτέλεσμα να χρησιμοποιούμε παραπάνω μεταβλητές από αυτές που πραγματικά θα έπρεπε να έχουμε χρησιμοποιήσει. Είναι επίσης προφανές ότι όσο λιγότερες μεταβλητές χρησιμοποιούμε, κερδίζουμε σε χρόνο και σε χρήμα. Για αυτούς λοιπόν τους λόγους έχουν αναπτυχθεί έλεγχοι και κριτήρια που βοηθούν στη σωστή επιλογή των καταλληλότερων μεταβλητών, τα οποία κατατάσσονται στις ακόλουθες κατηγορίες:

1. Μέτρα καταλληλότητας

α. Συντελεστής προσδιορισμού R^2.

Όπως έχουμε ήδη πει ο συντελεστής προσδιορισμού φανερώνει το ποσοστό συμμετοχής των επεξηγηματικών μεταβλητών στην μεταβολή της εξαρτημένης μεταβλητής. Συνηθίζουμε, βέβαια, να χρησιμοποιούμε αντί του R^2 τον διορθωμένο συντελεστή προσδιορισμού ${R^2}_{adj}$ *(adjusted)*, ο οποίος ορίζεται από τον τύπο

$$ {R^2}_{adj} = 1 - \frac{MSE}{MST} $$

β. C_p – *Mallows*. . Είναι ένα μέτρο το οποίο βασίζεται στο μέσο τετραγωνικό σφάλμα και σκοπό έχει να βελτιώσει την προβλεπτική αξία του μοντέλου. Ορίζεται από τον τύπο:

$$ C_p = \frac{SSE(p)}{\hat{\sigma}^2} + 2p - n $$

όπου $p = k + 1$. Καλύτερο μοντέλο θεωρείται εκείνο για το οποίο ισχύει $C_p \cong p$, ενώ αν υπάρχουν περισσότερα μοντέλα για τα οποία ισχύει $C_p \cong p$ επιλέγουμε εκείνο με το μικρότερο p.

γ. *AIC (Akaike's information criterion).* Το μέτρο αυτό ορίζεται από τον τύπο:

$$ AIC = 2d - 2lnL $$

Όπου L η μεγιστοποιημένη τιμή της συνάρτησης πιθανοφάνειας για το εκτιμώμενο μοντέλο και d το πλήθος των παραμέτρων. Η πρόσθεση

επιπλέον μεταβλητών μειώνουν την τιμή του *AIC* μόνο αν πετυχαίνουμε καλή προσαρμογή του μοντέλου.

δ. *BIC (Bayesian information criterion)*. Το μέτρο αυτό έχει παρόμοια χρήση με το *AIC* και ο τύπος του είναι:

$$BIC = d \cdot lnn - 2lnL$$

2. Στατιστικοί έλεγχοι

Για τον προσδιορισμό καταλληλότερων μεταβλητών χρησιμοποιούνται, όπως αναλυτικά έχουμε πει στην υποπαράγραφο 1.2.2, οι έλεγχοι t και F και η παλινδρόμηση με βήματα.

3. Γραφικές παραστάσεις

α. Διάγραμμα πρόσθετων μεταβλητών. Ας υποθέσουμε ότι στο μοντέλο μας θέλουμε να εισάγουμε μια καινούρια μεταβλητή. Θεωρούμε την καινούρια μεταβλητή ως εξαρτημένη και κατασκευάζουμε ένα γραμμικό μοντέλο παλινδρόμησης, με ανεξάρτητες μεταβλητές αυτές που ήδη έχουμε. Υπολογίζουμε τα υπόλοιπα αυτού του μοντέλου αλλά και του αρχικού. Τέλος κατασκευάζουμε το γράφημα αυτών των δύο κατηγοριών υπολοίπων. Αν τα σημεία που προκύπτουν εμφανίζουν γραμμική σχέση τότε η νέα μεταβλητή πρέπει να χρησιμοποιηθεί στο μοντέλο.

β. Διάγραμμα μερικών υπολοίπων. Υποθέτουμε ξανά ότι θέλουμε να εισάγουμε στο μοντέλο μια καινούρια ανεξάρτητη μεταβλητή. Αρχικά κατασκευάζουμε το μοντέλο παλινδρόμησης που περιέχει και την καινούρια μεταβλητή και στην συνέχεια καταγράφουμε τα υπόλοιπα. Έπειτα υπολογίζουμε τα μερικά υπόλοιπα της καινούριας μεταβλητής ως εξής: στην στήλη των αρχικών υπολοίπων που έχουμε βρει προσθέτουμε την στήλη με τις τιμές της καινούριας μεταβλητής πολλαπλασιάζοντας την με τον συντελεστή παλινδρόμησης της στο αρχικό μοντέλο. Κατασκευάζουμε το γράφημα των μερικών υπολοίπων και των τιμών της καινούριας μεταβλητής. Αν και εδώ τα σημεία του γραφήματος εμφανίζουν γραμμική εξάρτηση τότε η νέα μεταβλητή πρέπει να χρησιμοποιηθεί στο μοντέλο.

1.2.5 Εφαρμογή

Όλα όσα αναφέρθηκαν παραπάνω μπορούν να γίνουν ευκολότερα κατανοητά με το ακόλουθο παράδειγμα:

Σε ένα σχολικό έτος 28 μαθητές της Γ Λυκείου εξετάστηκαν σε τρία διαγωνίσματα (D1, D2, D3) και τέσσερα τεστ (T1, T2, T3, T4) στο μάθημα των Μαθηματικών Κατεύθυνσης. Οι βαθμοί τους σε αυτές τις εξετάσεις αποτελούν τις ανεξάρτητες μεταβλητές ενώ ο τελικός τους βαθμός στις Πανελλήνιες εξετάσεις την εξαρτημένη μεταβλητή (F). Θέλουμε να προσαρμόσουμε ένα γραμμικό μοντέλο παλινδρόμησης ώστε να διαπιστώσουμε ποια από τα επτά διαγωνίσματα και τεστ επηρεάζουν την τελική βαθμολογία και να επιτύχουμε την καλύτερη δυνατή πρόβλεψη του βαθμού των Μαθηματικών Κατεύθυνσης στις Πανελλήνιες εξετάσεις. Προσαρμόζοντας ένα γραμμικό μοντέλο στα δεδομένα έχουμε τα παρακάτω αποτελέσματα:

Predictor	Coef	SE Coef	T	P
Constant	-3,584	1,135	-3,16	0,005
D1	0,0934	0,1432	0,65	0,522
D2	-0,0925	0,2496	-0,37	0,715
D3	0,6310	0,2949	2,14	0,045
T1	0,0103	0,1697	0,06	0,952
T2	0,4732	0,1719	2,75	0,012
T3	0,1045	0,1733	0,60	0,553
T4	0,0962	0,1511	0,64	0,532

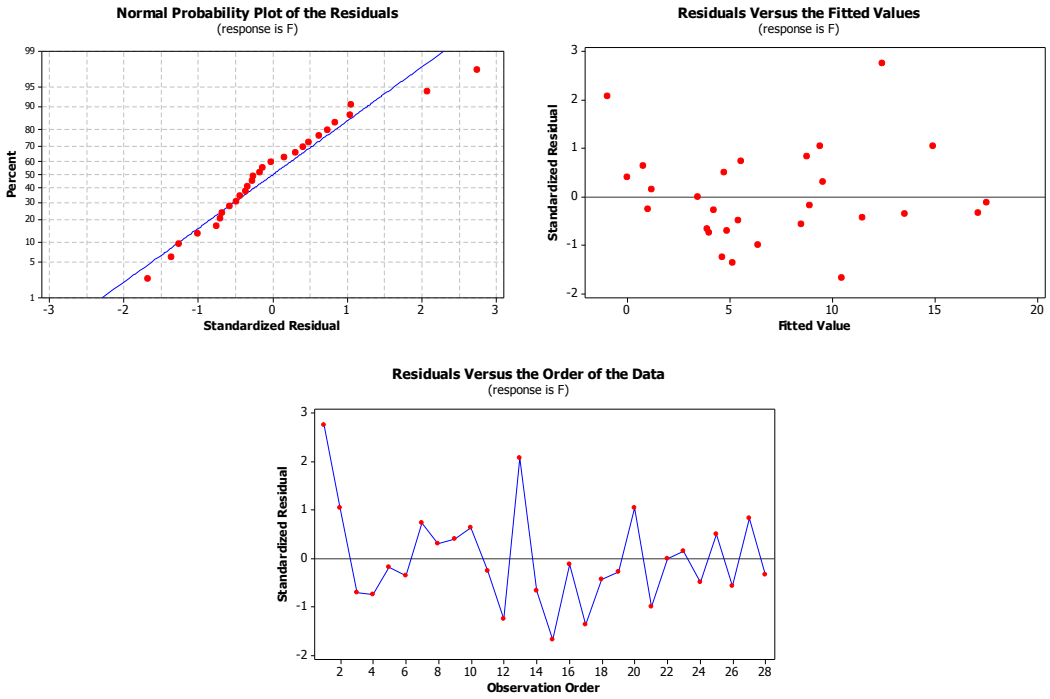
S = 2,17135 R-Sq = 87,9% R-Sq(adj) = 83,6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	7	683,698	97,671	20,72	0,000
Residual Error	20	94,295	4,715		
Total	27	777,993			

Παρατηρούμε από τον έλεγχο t ότι μόνο οι μεταβλητές D3 και T2 είναι στατιστικά σημαντικές. Παρόλα αυτά από την ανάλυση διασποράς ο έλεγχος F δείχνει ότι το μοντέλο μας προσαρμόζεται καλά. Οι συντελεστές προσδιορισμού R^2 και $R^2{}_{adj}$ είναι αρκετά υψηλοί γεγονός που φανερώνει ότι οι μεταβλητές συμμετέχουν σε μεγάλο

ποσοστό στη μεταβολή της εξαρτημένης μεταβλητής. Ακολουθούν οι γραφικοί έλεγχοι για τα υπόλοιπα (χρησιμοποιούμε τα τυποποιημένα υπόλοιπα):



Στο πρώτο γράφημα *(Normal Probability Plot)* βλέπουμε ότι εκτός από δύο τελευταία σημεία, τα υπόλοιπα παρουσιάζουν καλή γραμμικότητα πράγμα που σημαίνει ότι η υπόθεση της γραμμικότητας των υπολοίπων δεν παραβιάζεται σε μεγάλο βαθμό. Στο επόμενο γράφημα *(Residuals Versus the Fitted Values)* τα σημεία κατανέμονται τυχαία γύρω από το μηδέν, κάτι που μας εξασφαλίζει την ομοσκεδαστικότητα των υπολοίπων. Τέλος στο τέταρτο γράφημα *(Residuals Versus the Order of the Data)* παρουσιάζονται τα υπόλοιπα σε σχέση με τη σειρά των δεδομένων. Και εδώ τα σημεία κατανέμονται τυχαία γύρω από το μηδέν. Έτσι καταλήγουμε στο συμπέρασμα ότι δεν υπάρχει συσχέτιση μεταξύ των υπολοίπων.

Τέλος, θα προσπαθήσουμε να εντοπίσουμε τις κατάλληλες μεταβλητές στο μοντέλο. Αυτό θα γίνει χρησιμοποιώντας το συντελεστή προσδιορισμού και το C_p – *Mallows* όπως βλέπουμε στον ακόλουθο πίνακα.

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	D1	D2	D3	T1	T2	T3	T4
1	79,2	78,4	10,4	2,4972			X				
1	64,9	63,5	33,9	3,2413					X		
2	85,9	84,7	1,3	2,0967			X		X		
2	82,4	81,0	7,0	2,3399					X	X	
3	87,1	85,5	1,3	2,0450			X		X		X
3	87,0	85,4	1,4	2,0510			X		X	X	
4	87,6	85,4	2,5	2,0520	X		X		X	X	
4	87,5	85,4	2,6	2,0530	X		X		X		X
5	87,8	85,0	4,1	2,0774	X		X		X	X	X
5	87,6	84,8	4,4	2,0914	X	X	X		X		X
6	87,9	84,4	6,0	2,1192	X	X	X		X	X	X
6	87,8	84,3	6,1	2,1263	X		X	X	X	X	X
7	87,9	83,6	8,0	2,1714	X	X	X	X	X	X	X

Από αυτά τα αποτελέσματα παρατηρούμε ότι η τιμή που επηρεάζει περισσότερο το μοντέλο είναι η D3 αφού $R^2{}_{adj} = 78.4\%$ αλλά η τιμή $C_p = 10.4$ απέχει αρκετά από την επιθυμητή τιμή που είναι περίπου 2. Άρα η D3 από μόνη της μάλλον δεν δίνει το βέλτιστο μοντέλο. Παρατηρούμε επίσης ότι αν συμπεριλάβουμε επιπλέον και την T2 το $R^2{}_{adj} = 84.7\%$ μεταβάλλεται ικανοποιητικά. Αυτό σημαίνει ότι και η T2 θα πρέπει να προστεθεί στο μοντέλο. Εξάλλου και το C_p σε αυτή την περίπτωση βελτιώνεται και γίνεται 1.3 με αναμενόμενη τιμή το 3. Επιπλέον όπως βλέπουμε προσθέτοντας και άλλες μεταβλητές στο μοντέλο η μεταβολή του $R^2{}_{adj}$ είναι αμελητέα ενώ δεν μπορούμε να πούμε ότι το C_p παρουσιάζει αισθητή βελτίωση. Άρα καταλήγουμε στο συμπέρασμα ότι το βέλτιστο μοντέλο πρέπει να περιλαμβάνει μόνο τις μεταβλητές D3 και T2.

1.3 Πολυσυγγραμμικότητα

Στην γραμμική παλινδρόμηση είναι συχνό το φαινόμενο, οι επεξηγηματικές μεταβλητές να εμφανίζουν μέτριου ή υψηλού βαθμού συσχέτιση. Σε αυτές τις περιπτώσεις λέμε ότι στο μοντέλο υφίσταται το πρόβλημα της πολυσυγγραμμικότητας.

Αναλυτικότερα, γνωρίζουμε ότι αν μεταξύ των επεξηγηματικών μεταβλητών δεν υπάρχει συσχέτιση, οι μεταβλητές αυτές ονομάζονται ορθογώνιες. Τότε το μοντέλο κρίνεται κατάλληλο για προβλέψεις , για εκτιμήσεις, για την επιλογή των κατάλληλων μεταβλητών και γενικότερα για οτιδήποτε μπορεί να προσφέρει ένα μοντέλο παλινδρόμησης. Στις περισσότερες περιπτώσεις όμως, οι μεταβλητές δεν είναι ορθογώνιες. Αυτό σημαίνει ότι μεταξύ των μεταβλητών υπάρχει συσχέτιση. Αν η συσχέτιση αυτή είναι υψηλή τότε λέμε ότι έχει εμφανιστεί το φαινόμενο της πολυσυγγραμμικότητας.

Στις επόμενες παραγράφους θα αναφερθούμε λεπτομερώς στις αιτίες της πολυσυγγραμμικότητας και στις μεθόδους διάγνωσης και αντιμετώπισης της.

1.3.1 Μετασχηματισμός των μεταβλητών.

Για την αντιμετώπιση του φαινομένου της πολυσυγγραμμικότητας έχουν αναπτυχθεί μοντέλα παλινδρόμησης στα οποία κρίνεται (από τους περισσότερους) απαραίτητος ο μετασχηματισμός όλων των μεταβλητών. Ο λόγος που γίνονται οι μετασχηματισμοί αυτοί είναι να αποκτήσουν οι μεταβλητές (στήλες) την ίδια βαρύτητα. Υπάρχουν αρκετοί μετασχηματισμοί, αλλά αυτοί που χρησιμοποιούνται περισσότερο είναι οι δύο που περιγράφουμε αμέσως παρακάτω.

Α. *Centering and Scaling*. Υποθέτουμε ότι κάθε ανεξάρτητη μεταβλητή X_j και η εξαρτημένη μεταβλητή $\boldsymbol{y}$ μετασχηματίζονται σύμφωνα με τους τύπους:

$$x_{ij} \leftrightarrow \frac{x_{ij} - \overline{\boldsymbol{x}}_{\boldsymbol{j}}}{\sqrt{\sum_{i=1}^{n}\left(x_{ij} - \overline{\boldsymbol{x}}_{\boldsymbol{j}}\right)^2}} \qquad (1.3.1)$$

και

$$y_i \leftrightarrow \frac{y_i - \overline{\boldsymbol{y}}}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{\boldsymbol{y}})^2}} \qquad (1.3.2)$$

Έχοντας προβεί στους παραπάνω μετασχηματισμούς και αφού προσαρμόσουμε στα δεδομένα μας το κατάλληλο μοντέλο, μπορούμε να υπολογίσουμε το διάνυσμα $\boldsymbol{b}$ που αντιστοιχεί στα πραγματικά δεδομένα, δηλαδή στις μη μετασχηματισμένες μεταβλητές όπως και την αντίστοιχη σταθερά b_0. Αυτό γίνεται με την χρήση των ακόλουθων τύπων:

$$\hat{b}_j = \frac{\hat{\beta}_j}{\sqrt{\sum_{i=1}^{n}\left(x_{ij} - \bar{\boldsymbol{x}}_{\boldsymbol{j}}\right)^2}}, \qquad 1 \le j \le k \qquad (1.3.3)$$

και

$$\hat{b}_0 = \hat{\beta}_0 - \sum_{j=1}^{k} \frac{\hat{\beta}_j \cdot \bar{x}_j}{\sqrt{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}} = \hat{\beta}_0 - \sum_{j=1}^{k} \hat{b}_j \cdot \bar{x}_j$$

$$= -\sum_{j=1}^{k} \hat{b}_j \cdot \bar{x}_j \qquad , \left(\hat{\beta}_0 = \bar{\boldsymbol{y}} = 0\right) \qquad (1.3.4)$$

Β. <u>Τυποποίηση δεδομένων *(Standardization)*</u>. Οι τύποι μετασχηματισμού στην περίπτωση αυτή ακολουθούν αμέσως παρακάτω, ενώ ο υπολογισμός του διανύσματος των πραγματικών συντελεστών πραγματοποιείται με τύπους ανάλογους με τους (1.3.3) και (1.3.4).

$$x_{ij} \leftrightarrow \frac{x_{ij} - \bar{x}_j}{sd(\boldsymbol{x_j})} \qquad (1.3.5)$$

και

$$y_i \leftrightarrow \frac{y_i - \bar{\boldsymbol{y}}}{sd(\boldsymbol{y})} \qquad (1.3.6)$$

Γενικά ισχύει ότι $\hat{\beta}_0 = \bar{\boldsymbol{y}}$ και στους δύο μετασχηματισμούς. Και επειδή από τις τιμές της μεταβλητής $\boldsymbol{y}$ αφαιρείται η μέση τιμή τους $(y_i - \bar{\boldsymbol{y}})$, η νέα μέση τιμή μετά τους μετασχηματισμούς θα είναι μηδέν, οπότε $\hat{\beta}_0 = 0$. Επομένως δεν θα υπάρχει σταθερός όρος στο μοντέλο παλινδρόμησης. Για τον λόγο αυτό θεωρούμε ότι ο πίνακας X θα είναι ένας $n \times k$ πίνακας που θα περιέχει μόνο τις ανεξάρτητες μεταβλητές και όχι την πρώτη στήλη με τις μονάδες που αντιστοιχούν στην σταθερά β_0.

Επίσης, εάν χρησιμοποιήσουμε τον πρώτο μετασχηματισμό, ο πίνακας $X'X$ θα είναι ο πίνακας συνδιακύμανσης των ανεξάρτητων μεταβλητών με διαγώνια στοιχεία την μονάδα και ο πίνακας $(X'X)^{-1}$ θα έχει σαν διαγώνια στοιχεία του τις τιμές *VIF*, οι οποίες περιγράφονται αναλυτικά σε επόμενη παράγραφο. Στον δεύτερο

μετασχηματισμό ο πίνακας $X'X$ αντιστοιχεί στον πίνακα διασποράς συνδιασποράς των ανεξάρτητων μεταβλητών. Από εδώ και στο εξής θα θεωρείται δεδομένο ότι όλες οι μεταβλητές έχουν μετασχηματιστεί με τον ένα ή τον άλλο τρόπο. Συγκεκριμένα στο παράδειγμα με τους βαθμούς των μαθητών και στις εφαρμογές του τρίτου κεφαλαίου όλες οι μεταβλητές έχουν μετασχηματιστεί σύμφωνα με τον δεύτερο τρόπο (τύποι (1.3.5) και (1.3.6)).

1.3.2 Αιτίες που προκαλούν πολυσυγγραμμικότητα

Γνωρίζουμε ότι το γραμμικό μοντέλο παλινδρόμησης είναι:

$$\boldsymbol{y} = X \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

όπου με δεδομένο ότι έχουμε προβεί σε έναν από τους προηγούμενους μετασχηματισμούς, ο X είναι ο $n \times k$ πίνακας που έχει σαν στήλες τις επεξηγηματικές μεταβλητές, $\boldsymbol{y}$ το $n \times 1$ διάνυσμα των εξαρτημένων μεταβλητών και $\boldsymbol{\beta}$ το $k \times 1$ διάνυσμα των συντελεστών των μεταβλητών αυτών, που έχουν εκτιμηθεί με την μέθοδο των ελαχίστων τετραγώνων. Δηλαδή:

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1} \cdot X' \cdot \boldsymbol{y}$$

Οι στήλες $X_1, X_2, X_3, \ldots, X_k$ του πίνακα X θα είναι γραμμικώς εξαρτημένες αν υπάρχουν αριθμοί $t_1, t_2, t_3, \ldots, t_k$ όχι όλοι μηδέν ώστε:

$$\sum_{j=1}^{k} t_j \cdot X_j = \mathbf{0} \qquad (1.3.7)$$

Εάν ο τύπος (1.3.7) ισχύει για τουλάχιστον δύο από τις στήλες του πίνακα X τότε η τάξη *(rank)* του πίνακα $X'X$ θα είναι μικρότερη από την τιμή k και δεν θα ορίζεται ο $(X'X)^{-1}$, με αποτέλεσμα η εκτίμηση του διανύσματος $\boldsymbol{\beta}$ να είναι αδύνατη. Κάποιες φορές όμως ο τύπος (1.3.7), για κάποιες από τις στήλες του πίνακα X, δίνει τιμές πολύ κοντά στο μηδέν. Τότε λέμε ότι ο πίνακας $X'X$ εμφανίζει σχεδόν γραμμική εξάρτηση και ταυτόχρονα εμφανίζεται και το πρόβλημα της πολυσυγγραμμικότητας. Αξίζει να σημειωθεί εδώ ότι το φαινόμενο της πολυσυγγραμμικότητας εμφανίζεται πάντα σε μικρό ή μεγάλο βαθμό, εκτός εάν οι επεξηγηματικές μεταβλητές είναι ορθογώνιες. Τότε ο πίνακας $X'X$ θα είναι διαγώνιος και φυσικά αντιστρέψιμος.

Οι βασικότεροι λόγοι που το φαινόμενο της πολυσυγγραμμικότητας κάνει την εμφάνιση του είναι οι παρακάτω:

Α. **Η μέθοδος που θα επιλέξουμε να συλλέξουμε τα δεδομένα** μπορεί να αποτελέσει αιτία για να εμφανιστεί το φαινόμενο της πολυσυγγραμμικότητας. Αυτό συμβαίνει όταν κάποιες από τις επεξηγηματικές μεταβλητές συνδυάζονται με συγκεκριμένο τρόπο. Για παράδειγμα, μπορεί να έχουμε επιλέξει, όταν η μεταβλητή X_i έχει υψηλή τιμή η μεταβλητή X_j να έχει χαμηλή τιμή και αντίστροφα. Αυτό θα έχει σαν αποτέλεσμα η μια μεταβλητή να εξαρτάται από την άλλη και αν η εξάρτηση αυτή είναι ισχυρή τότε είναι πολύ πιθανό να εμφανιστεί το πρόβλημα της πολυσυγγραμμικότητας.

Β. **Οι περιορισμοί στο μοντέλο ή τον πληθυσμό.** Οι περιορισμοί μπορεί να τεθούν από τον μελετητή και να αφορούν στο μοντέλο ή στον πληθυσμό άλλα μπορεί και να υπάρχουν λόγω της φύσης του θέματος που μελετάμε. Ας υποθέσουμε, για παράδειγμα ότι η φυσική κατάσταση ενός πληθυσμού ενηλίκων εξαρτάται από τις ώρες που προπονούνται ανά εβδομάδα (X_1) και από τις ώρες που δουλεύουν ανά εβδομάδα (X_2). Εύκολα καταλαβαίνουμε ότι οι άνθρωποι που αθλούνται περισσότερες ώρες θα έχουν καλύτερη φυσική κατάσταση. Αυτοί όμως που αθλούνται περισσότερες ώρες θα δουλεύουν και λιγότερες ώρες. Αυτός ο φυσικός περιορισμός είναι πολύ πιθανό να προκαλέσει έντονη αρνητική συσχέτιση μεταξύ των μεταβλητών X_1 και X_2. Σε αυτή την περίπτωση εμφανίζεται το φαινόμενο της πολυσυγγραμμικότητας.

Γ. **Η επιλογή του μοντέλου.** Για παράδειγμα, η πρόσθεση όλο και περισσότερων όρων στο μοντέλο παλινδρόμησης μπορεί να δυσκολέψει σταδιακά την αντιστροφή του πίνακα $X'X$. Ή κάποιες φορές αν το εύρος των τιμών μιας επεξηγηματικής μεταβλητής X είναι μικρό και στο μοντέλο προσθέσουμε και την X^2 τότε αυτό μπορεί να προκαλέσει πολυσυγγραμμικότητα. Τέλος η γραμμική εξάρτηση δύο ή περισσοτέρων μεταβλητών ίσως δημιουργήσει πολυσυγγραμμικότητα. Σε τέτοιες περιπτώσεις όπως οι παραπάνω, επιλέγουμε ένα μοντέλο που περιέχει ένα μέρος των επεξηγηματικών μεταβλητών.

Δ **Η ύπαρξη πολλών επεξηγηματικών μεταβλητών σε μικρό πλήθος παρατηρήσεων.** Για την αντιμετώπιση του παραπάνω προβλήματος οι *Mason, Gunst,* και *Webster (1973)* πρότειναν τρεις συγκεκριμένες μεθόδους:

1. Αντικατάσταση του μοντέλου με ένα αντίστοιχο που θα περιέχει λιγότερες μεταβλητές
2. πραγματοποίηση προκαταρκτικών μελετών χρησιμοποιώντας ένα μέρος των επεξηγηματικών μεταβλητών
3. χρήση της Ανάλυσης Κύριων Συνιστωσών *(Principal Components Analysis)* στην παλινδρόμηση.

Οι δύο πρώτες μέθοδοι δεν λαμβάνουν υπόψη τους πιθανή αλληλεξάρτηση κάποιων από τις επεξηγηματικές μεταβλητές και έτσι οδηγούν σε μη ικανοποιητικά αποτελέσματα. Με την τρίτη μέθοδο θα ασχοληθούμε αναλυτικότερα παρακάτω.

1.3.3 Επιδράσεις της πολυσυγγραμμικότητας

Έστω $C = (X'X)^{-1}$. Τότε μπορεί να αποδειχθεί ότι τα διαγώνια στοιχεία του πίνακα C δίνονται από τον τύπο

$$c_{jj} = \frac{1}{1 - R_j^2}$$

όπου R_j^2 είναι ο συντελεστής προσδιορισμού που προκύπτει αν εφαρμόσουμε την παλινδρόμηση στο μοντέλο με εξαρτημένη μεταβλητή την X_j και επεξηγηματικές όλες τις υπόλοιπες. Θα πρέπει να σημειωθεί ότι ο παραπάνω τύπος προκύπτει αφού έχουμε μετασχηματίσει όλες τις ανεξάρτητες μεταβλητές σύμφωνα με τους τύπους (1.3.1) και (1.3.2).

Γνωρίζουμε επίσης ότι η διασπορά του εκτιμημένου συντελεστή $\hat{\beta}_j$ δίνεται από τον τύπο:

$$V(\hat{\beta}_j) = c_{jj} \cdot \sigma^2$$

Και επομένως:

$$V(\hat{\beta}_j) = \frac{\sigma^2}{1 - R_j^2}$$

Έτσι λοιπόν καθώς το $R_j^2 \to 1$, δηλαδή η μεταβλητή X_j μπορεί να προβλεφθεί (εξαρτάται) σε πολύ μεγάλο βαθμό από τις υπόλοιπες, η διασπορά $V(\hat{\beta}_j) \to \infty$ (πολύ μεγάλη διασπορά).

Τότε όμως, $$se(\hat{\beta}_j) = (V(\hat{\beta}_j))^{1/2} \to \infty,$$

με συνέπεια

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \to 0$$

Κάτι τέτοιο λοιπόν, μπορεί να μας οδηγήσει σε λανθασμένα συμπεράσματα που έχουν να κάνουν με τον εντοπισμό των στατιστικά σημαντικών μεταβλητών αφού η τιμή της ελεγχοσυνάρτησης t θα εμφανίζεται πάντα μικρή. Επίσης αποδεικνύεται ότι η συνδιακύμανση των $\hat{\beta}_i$ και $\hat{\beta}_j$ θα εμφανίζεται μεγάλη αν οι αντίστοιχες μεταβλητές των παραπάνω συντελεστών συμβάλουν στην πολυσυγγραμμικότητα του μοντέλου.

Επίσης, εξαιτίας της πολυσυγγραμμικότητας οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_i$ εμφανίζουν υψηλές απόλυτες τιμές. Αυτός ο ισχυρισμός

αποδεικνύεται αν υπολογίσουμε το τετράγωνο της απόστασης d της εκτιμήτριας $\widehat{\boldsymbol{\beta}}$ από το διάνυσμα των πραγματικών συντελεστών $\boldsymbol{\beta}$. Θα έχουμε λοιπόν:

$$\boldsymbol{d} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$$

$$\boldsymbol{d}^2 = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \cdot \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$$

Αν υπολογίσουμε την αναμενόμενη τιμή του d^2 θα έχουμε:

$$E(\boldsymbol{d}^2) = E\left[\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \cdot \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right] = \sum_{j=1}^{k} E\left(\hat{\beta}_j - \beta_j\right)^2 = \sum_{j=1}^{k} V\left(\hat{\beta}_j\right) = \sum_{j=1}^{k} c_{jj} \cdot \sigma^2$$

$$= \sigma^2 \cdot \sum_{j=1}^{k} c_{jj} = \sigma^2 \cdot Tr(X'X)^{-1}$$

όπου,

$$Tr(X'X)^{-1} = \sum_{j=1}^{k} c_{jj}$$

Γνωρίζουμε, επίσης, ότι το άθροισμα των διαγώνιων στοιχείων ενός πίνακα $Tr(X'X)^{-1}$ είναι ίσο με το άθροισμα των ιδιοτιμών του και αν λ_i , $i = 1,2, \dots, k$ οι ιδιοτιμές του $X'X$ τότε οι ιδιοτιμές του αντιστρόφου του $(X'X)^{-1}$ θα είναι $\frac{1}{\lambda_i}$. Έτσι λοιπόν θα έχουμε:

$$E(\boldsymbol{d}^2) = \sigma^2 \cdot \sum_{j=1}^{k} \frac{1}{\lambda_j} \qquad (1.3.8)$$

Λόγω της πολυσυγγραμμικότητας τουλάχιστον μια από τις ιδιοτιμές του $X'X$ θα είναι πολύ μικρή. Άρα από τον τύπο (1.3.8) συνεπάγεται ότι η απόσταση $\boldsymbol{d}$ πιθανόν να είναι μεγάλη.

Αξίζει να σημειωθεί εδώ ότι, αν και οι εκτιμήτριες ελαχίστων τετραγώνων μπορεί να απέχουν πολύ από την πραγματικότητα εντούτοις η προβλεπτική ικανότητα του μοντέλου μπορεί να είναι αρκετά ικανοποιητική.

1.3.4 Τρόποι διάγνωσης της πολυσυγγραμμικότητας

Α. Εξέταση του πίνακα συσχέτισης *(Correlation Matrix)*. Σύμφωνα με αυτόν τον τρόπο ελέγχουμε τα μη διαγώνια στοιχεία r_{ij} του πίνακα $X'X$. Αν για τις μεταβλητές X_i , X_j υπάρχει γραμμική εξάρτηση τότε η τιμή του συγκεκριμένου στοιχείου θα είναι κοντά στην μονάδα. Ο τρόπος αυτός θεωρείται κατάλληλος για να υποδείξει ότι υπάρχει γραμμική σχέση μόνο μεταξύ δύο ανεξάρτητων μεταβλητών. Αυτό σημαίνει πως δεν είναι σίγουρο ότι θα εντοπίσει γραμμική σχέση μεταξύ τριών ή παραπάνω μεταβλητών.

Στο παράδειγμα με τους βαθμούς των μαθητών ο πίνακας $X'X$ είναι:

	D1	D2	D3	T1	T2	T3	T4
D1	1.0000003	0.6112304	0.7227777	0.7104033	0.5882982	0.7124129	0.6859952
D2	0.6112304	1.0000002	0.8950220	0.7494988	0.6870437	0.7454560	0.6469854
D3	0.7227777	0.8950220	1.0000000	0.7758852	0.6963788	0.8441568	0.6897531
T1	0.7104033	0.7494988	0.7758852	0.9999999	0.7049928	0.7792620	0.7597561
T2	0.5882982	0.6870437	0.6963788	0.7049928	1.0000001	0.5631916	0.6021371
T3	0.7124129	0.7454560	0.8441568	0.7792620	0.5631916	0.9999998	0.7655332
T4	0.6859952	0.6469854	0.6897531	0.7597561	0.6021371	0.7655332	0.9999989

Παρατηρούμε υψηλές τιμές συσχέτισης μεταξύ των μεταβλητών $D2, D3$ $(0{,}895022)$ και $T3, D3$ $(0{,}8441568)$. Μπορούμε να συμπεράνουμε, λοιπόν, την ύπαρξη του φαινομένου της πολυσυγγραμμικότητας στο μοντέλο.

Β. Παράγοντας Μεγέθυνσης Διασποράς *(Variance Inflation Factor, VIF).* Στην προηγούμενη παράγραφο είπαμε ότι κάθε διαγώνιο στοιχείο c_{jj} του πίνακα $C = (X'X)^{-1}$ δίνεται από τον τύπο

$$c_{jj} = \frac{1}{1 - R_j^2}$$

όπου R_j^2 είναι ο συντελεστής προσδιορισμού που προκύπτει αν εφαρμόσουμε την παλινδρόμηση στο μοντέλο με εξαρτημένη μεταβλητή την X_j και επεξηγηματικές όλες τις υπόλοιπες. Παρατηρούμε ότι, αν η μεταβλητή X_j είναι σχεδόν ορθογώνια σε σχέση με τις υπόλοιπες μεταβλητές τότε η τιμή του R_j^2, που δείχνει σε τι ποσοστό η X_j μπορεί να προβλεφθεί από τις υπόλοιπες μεταβλητές, θα πλησιάζει την τιμή μηδέν με συνέπεια η τιμή του c_{jj} να πλησιάζει την μονάδα. Αν όμως η X_j εξαρτάται

γραμμικά (είναι δυνατόν να προβλεφθεί) από τουλάχιστον μια από τις υπόλοιπες μεταβλητές τότε η τιμή του R_j^2 θα πλησιάζει την μονάδα και επομένως το στοιχείο c_{jj} θα έχει αρκετά μεγάλη τιμή. Άρα αν κατά την προσαρμογή ενός μοντέλου παλινδρόμησης υπολογίσουμε επιπλέον τις τιμές των διαγώνιων στοιχείων c_{jj} και διαπιστώσουμε υψηλές τιμές για τουλάχιστον μια μεταβλητή τότε μπορούμε να συμπεράνουμε ότι το φαινόμενο της πολυσυγγραμμικότητας υφίσταται για το συγκεκριμένο μοντέλο. Έτσι λοιπόν, τις τιμές των διαγώνιων στοιχείων c_{jj} τις συμβολίζουμε με VIF_j και θεωρούμε ως ένδειξη πολυσυγγραμμικότητας, τιμές των $VIF_j > 5$

Αν στο παράδειγμα με τους βαθμούς των μαθητών υπολογίσουμε την τιμή VIF κάθε μεταβλητής έχουμε τα ακόλουθα αποτελέσματα:

D1	D2	D3	T1	T2	T3	T4
2.675491	5.616847	8.980878	4.091119	2.476230	5.075767	3.101344

Τα αποτελέσματα αυτά μας οδηγούν σε ανάλογο συμπέρασμα με αυτό που προέκυψε από την εξέταση του πίνακα συσχέτισης. Δηλαδή, ότι οι μεταβλητές $D2, D3, T3$ εμφανίζουν σχεδόν γραμμική εξάρτηση.

Γ. <u>Ανάλυση ιδιοτιμών του πίνακα $X'X$</u>. Η ύπαρξη ή μη της πολυσυγγραμμικότητας στα δεδομένα μας μπορεί να ελεγχθεί αξιόπιστα και με την χρήση των ιδιοτιμών $\lambda_1, \lambda_2, \dots, \lambda_k$ του πίνακα $X'X$. Για τον σκοπό αυτό στηριζόμαστε στην παραδοχή ότι αν οι τιμές μιας ή περισσότερων ιδιοτιμών είναι μικρές τότε υπάρχει γραμμική εξάρτηση μεταξύ των στηλών του πίνακα X.

Ορισμένοι αναλυτές, για τον σκοπό αυτό, εξετάζουν τον αριθμό κατάστασης *(condition number)* του πίνακα $X'X$ που ορίζεται ως εξής:

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}$$

όπου υποδεικνύει το μέγεθος της περιοχής μέσα στην οποία ανήκουν όλες οι ιδιοτιμές του $X'X$. Θεωρούμε ότι αν $100 \leq \kappa < 1000$ τότε η πολυσυγγραμμικότητα στο μοντέλο κρίνεται ως μέτρια προς ισχυρή, ενώ αν $\kappa \geq 1000$ τότε η πολυσυγγραμμικότητα κρίνεται ως αρκετά σοβαρή.

Επιπρόσθετα μπορούμε να υπολογίσουμε τις τιμές

$$\kappa_j = \frac{\lambda_{max}}{\lambda_j}$$

Το πλήθος των κ_j που είναι μεγαλύτερα ή ίσα του 1000 μας υποδεικνύει το πλήθος των γραμμικών εξαρτήσεων στον πίνακα $X'X$.

Στο παράδειγμα με τους βαθμούς των μαθητών οι ιδιοτιμές του πίνακα $X'X$ είναι:

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
5.28045021	0.51234927	0.42429232	0.33120982	0.20745022	0.17215607	0.07209134

Παρατηρούμε ότι υπάρχει μια ιδιοτιμή που είναι πολύ μικρή. Το γεγονός αυτό αποτελεί ένδειξη ύπαρξης γραμμικής σχέσης μεταξύ των μεταβλητών. Στην συνέχεια υπολογίζουμε την τιμή του λόγου κ άλλα και των υπόλοιπων λόγων κ_j.

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}} = \frac{5{,}28045021}{0{,}07209134} = 73{,}2466 < 100$$

Εφόσον ισχύει ότι $\kappa < 100$ τότε και οι υπόλοιποι λόγοι κ_j θα είναι μικρότεροι του 100. Γενικά η μέθοδος αυτή δεν μας υποδεικνύει την ύπαρξη πολυσυγγραμμικότητας

Δ. <u>Υπολογισμός της ορίζουσας του πίνακα $X'X$</u> . Εφόσον ο $X'X$ είναι ο πίνακας συσχέτισης τότε για την ορίζουσα του θα ισχύει:

$$0 \leq |X'X| \leq 1$$

Αν ισχύει $|X'X| = 1$ τότε οι μεταβλητές είναι ανεξάρτητες ενώ αν $|X'X| = 0$ οι μεταβλητές θα είναι γραμμικά εξαρτημένες. Όσο περισσότερο η παραπάνω ορίζουσα πλησιάζει στην τιμή του μηδενός τόσο πιο σημαντική είναι η πολυσυγγραμμικότητα. Αν και η μέθοδος αυτή είναι αρκετά απλή για να ελέγχουμε την ύπαρξη της πολυσυγγραμμικότητας στο μοντέλο, δεν μπορεί να μας υποδείξει την πηγή του προβλήματος.

Για το παράδειγμα μας, η ορίζουσα του πίνακα $X'X$ είναι:

$$|X'X| = 9.7887e - 004$$

Η τιμή αυτή είναι πολύ κοντά στο μηδέν, κάτι που σημαίνει ότι στο μοντέλο υπάρχει σημαντική πολυσυγγραμμικότητα.

Ε. <u>Οι έλεγχοι F και t</u>. Ξέρουμε ότι με τον έλεγχο F διαπιστώνουμε αν το μοντέλο παλινδρόμησης προσαρμόζεται καλά (είναι στατιστικά σημαντικό). Ενώ ο

έλεγχος t υποδεικνύει αν μια μεταβλητή συμβάλει στις διαμόρφωση των τιμών της εξαρτημένης μεταβλητής. Ειδικότερα, αν η τιμή της ελεγχοσυνάρτησης F είναι στατιστικά σημαντική ενώ για κάθε μεταβλητή η τιμή του t δεν είναι σημαντική, τότε λέμε ότι εμφανίζεται η πολυσυγγραμμικότητα. Έχει όμως διαπιστωθεί ότι η μέθοδος αυτή δεν είναι αξιόπιστη αφού σε μοντέλα με σοβαρή πολυσυγγραμμικότητα η τιμές των F και t δεν υποδείκνυαν κάτι ανάλογο.

Τέλος, πιστεύεται *(Montgomery et al., 2006)*, ότι οι τιμές του VIF_j καθώς και οι διαδικασίες που σχετίζονται με τις ιδιοτιμές του πίνακα $X'X$ (αν και στο παράδειγμα δεν προέκυψε) είναι οι πλέον αξιόπιστες για την διάγνωση της πολυσυγγραμμικότητας σε ένα μοντέλο. Θεωρείται ότι υπολογίζονται και ερμηνεύονται εύκολα καθώς επίσης και ότι συμβάλουν στον εντοπισμό των πηγών της πολυσυγγραμμικότητας.

1.3.5 Μέθοδοι για την αντιμετώπιση της πολυσυγγραμμικότητας

Α. Συλλογή επιπλέον δεδομένων. Αν και από πολλούς είχε θεωρηθεί η σημαντικότερη μέθοδος για την αντιμετώπιση της πολυσυγγραμμικότητας ωστόσο παρουσιάζει σημαντικά μειονεκτήματα. Το πιο σημαντικό από αυτά είναι οι οικονομικοί περιορισμοί που ενδεχομένως ισχύουν κατά την διάρκεια μιας έρευνας. Είναι γνωστό ότι η συλλογή δεδομένων γενικά, απαιτεί μέρος τον οικονομικών πόρων που διατίθενται για την έρευνα. Επίσης, η συλλογή παραπάνω δεδομένων μπορεί να μην είναι δυνατή επειδή το θέμα το οποίο μελετάμε ίσως δεν προσφέρεται για επιπλέον δειγματοληψία. Τέλος η μέθοδος αυτή δεν ενδείκνυται όταν η πολυσυγγραμμικότητα οφείλεται σε περιορισμούς του μοντέλου ή του πληθυσμού.

Β. Επανακαθορισμός του μοντέλου. Πολλές φορές μια λανθασμένη προσαρμογή του μοντέλου μπορεί να είναι η αιτία για την εμφάνιση της πολυσυγγραμμικότητας. Σε μια τέτοια περίπτωση εφαρμόζοντας τις κατάλληλες τεχνικές είναι δυνατό να αποφύγουμε την πολυσυγγραμμικότητα. Ας δούμε λοιπόν, ποιες είναι αυτές οι τεχνικές.

1. Εάν γνωρίζουμε ποιες είναι οι ανεξάρτητες μεταβλητές που δημιουργούν την πολυσυγγραμμικότητα στο μοντέλο μας, μπορούμε να τις αντικαταστήσουμε με μια μεταβλητή η οποία θα εξαρτάται, μέσω μιας σχέσης, από τις παραπάνω μεταβλητές χωρίς ουσιαστικά να μεταβάλλεται το σύνολο των πληροφοριών που μας παρέχει το αρχικό μοντέλο. Για παράδειγμα, αν στο μοντέλο οι μεταβλητές X_1, X_2 εμφανίζουν μια γραμμική εξάρτηση, μπορούμε να τις αντικαταστήσουμε με την μεταβλητή X ώστε $X = \frac{X_1}{X_2}$ ή $X = X_1 \cdot X_2$ ή κάποια άλλη σχέση.

2. Μπορούμε να παραλείψουμε κάποια από τις μεταβλητές που προκαλούν την πολυσυγγραμμικότητα. Η τεχνική αυτή είναι αρκετά διαδεδομένη και πολλές φορές αποτελεσματική. Δεν αποτελεί, όμως, καλή λύση στην περίπτωση που το X_j το οποίο παραλείπουμε έχει σημαντική επίδραση στην μεταβλητή Y. Αν συμβεί κάτι τέτοιο τότε το μοντέλο θα χάσει σημαντικό μέρος της προβλεπτικής του ικανότητας παρόλο που θα βελτιωθεί στο θέμα της πολυσυγγραμμικότητας.

Επίσης, οι δύο από τις τρεις πιο σημαντικές μέθοδοι για την αντιμετώπιση της πολυσυγγραμμικότητας είναι η παλινδρόμηση κορυφογραμμής *(Ridge Regression,*

RR) και η παλινδρόμηση κύριων συνιστωσών *(Principal Component Regression, PCR)*. Λόγω της πολυπλοκότητας και σημαντικότητας αυτών των δύο μεθόδων, θα τις μελετήσουμε αναλυτικότερα στις παραγράφους που ακολουθούν. Η τρίτη μέθοδος *(Partial Least Square Regression, PLSR)* αποτελεί το θέμα του δεύτερου κεφαλαίου.

1.3.6 Παλινδρόμηση Κορυφογραμμής *(Ridge Regression)*

Με αυτή την μέθοδο αλλά και τις υπόλοιπες θα αντιμετωπίσουμε το πρόβλημα της πολυσυγγραμμικότητας εφαρμόζοντας μεθόδους παλινδρόμησης οι οποίες δεν στηρίζονται στην μεταβολή του πλήθους των παρατηρήσεων ούτε στην αντικατάσταση ή εξάλειψη των μεταβλητών που παρουσιάζουν κάποιας μορφή γραμμική εξάρτηση. Στόχος αποτελεί η όσο το δυνατό καλύτερη εκτίμηση των συντελεστών των ανεξάρτητων μεταβλητών.

Όπως έχουμε ήδη αναφέρει (παράγραφος 1.3.3) μία από της επιδράσεις της πολυσυγγραμμικότητας στο μοντέλο είναι ότι κάποιες από τις εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_i$ είναι αρκετά μεγάλες κατά απόλυτη τιμή. Αυτό συμβαίνει διότι οι διασπορές των εκτιμητριών $\hat{\beta}_i$ που προκαλούν την πολυσυγγραμμικότητα είναι αρκετά μεγάλες όπως και το μέτρο του διανύσματος $\hat{\beta}$. Όλα αυτά μας οδηγούν σε εκτιμήτριες αρκετά ασταθείς κάτι που σημαίνει ότι μια μικρή αλλαγή στο δείγμα μπορεί να προκαλέσει σημαντική μεταβολή στις τιμές τους. Όλα αυτά δημιουργούνται επειδή απαιτούμε η εκτιμήτρια $\hat{\beta}$ να είναι αμερόληπτη εκτιμήτρια του β.

Από το θεώρημα των *Gauss – Markov* γνωρίζουμε ότι η εκτιμήτρια ελαχίστων τετραγώνων $\hat{\beta}$ είναι *BLUE* δηλαδή από τις αμερόληπτες εκτιμήτριες είναι αυτή με την μικρότερη διασπορά. Όμως δεν γνωρίζουμε αν μεταξύ όλων των εκτιμητριών η διασπορά αυτή είναι η μικρότερη. Για τον λόγο αυτό καταργούμε την απαίτηση η εκτιμήτρια του β να είναι αμερόληπτη και αναζητούμε μια εκτιμήτρια $\hat{\beta}^*$ με μικρό ποσοστό μεροληψίας και μικρότερη διασπορά από την αμερόληπτη εκτιμήτρια ελαχίστων τετραγώνων $\hat{\beta}$. Αυτό σημαίνει ότι μπορούμε να πετύχουμε η καινούρια εκτιμήτρια (μεροληπτική) να έχει μέσο τετραγωνικό σφάλμα *(ΜΤΣ)* μικρότερο από την διασπορά της αρχικής (αμερόληπτης). Αυτό αποδεικνύεται και από την παρακάτω ανάλυση:

$$MT\Sigma(\hat{\beta}^*) = E(\hat{\beta}^* - \beta)^2 = E[\hat{\beta}^* - E(\hat{\beta}^*) + E(\hat{\beta}^*) - \beta]^2 =$$

$$= E\left(\hat{\beta}^* - E(\hat{\beta}^*)\right)^2 + 2E\left(\hat{\beta}^* - E(\hat{\beta}^*)\right)E(E(\hat{\beta}^*) - \beta) + E(E(\hat{\beta}^*) - \beta)^2 =$$

$$= V(\hat{\beta}^*) + (E(\hat{\beta}^*) - \beta)^2 = V(\hat{\beta}^*) + (bias_{\hat{\beta}^*})^2$$

Από το παραπάνω αποτέλεσμα συμπεραίνουμε ότι αν η τιμή της μεροληψίας είναι αρκετά μικρή και με δεδομένο ότι η διασπορά της μεροληπτικής εκτιμήτριας είναι

μικρότερη από αυτή της αμερόληπτης, μπορούμε να πετύχουμε το $MT\Sigma(\hat{\beta}^*)$ να είναι μικρότερο από την διασπορά της αμερόληπτης εκτιμήτριας.

Μια από τις τεχνικές που έχουν αναπτυχθεί για την επίτευξη του παραπάνω σκοπού είναι και η *Ridge Regression, (Hoerl και Kennard, 1970)*. Ο υπολογισμός της κατάλληλης εκτιμήτριας $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}$ πραγματοποιείται με τον υπολογισμό κατάλληλης σταθεράς $\theta \geq 0$ που αντικαθιστά, ουσιαστικά, τον πίνακα $X'X$ με τον πίνακα $X'X + \theta I$ έτσι ώστε:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}} = (X'X + \theta I)^{-1}X'y$$

Η σταθερά θ ονομάζεται παράμετρος μεροληψίας και αν $\theta = 0$ τότε εύκολα καταλαβαίνουμε ότι $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}} = \widehat{\boldsymbol{\beta}}$.

Η σχέση της εκτιμήτριας $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}$ με την αντίστοιχη των ελαχίστων τετραγώνων $\widehat{\boldsymbol{\beta}}$ είναι ότι πρόκειται για ένα γραμμικό μετασχηματισμό της εκτιμήτριας $\widehat{\boldsymbol{\beta}}$ αφού:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}} = (X'X + \theta I)^{-1}X'y = (X'X + \theta I)^{-1}(X'X)\widehat{\boldsymbol{\beta}} = Z_\theta \cdot \widehat{\boldsymbol{\beta}}$$

Εύκολα αποδεικνύουμε ότι η εκτιμήτρια $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}$ είναι μεροληπτική εκτιμήτρια, αφού:

$$E(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}) = E(Z_\theta \cdot \widehat{\boldsymbol{\beta}}) = Z_\theta \cdot E(\widehat{\boldsymbol{\beta}}) = Z_\theta \cdot \boldsymbol{\beta} \neq \boldsymbol{\beta}$$

Και το μέσο τετραγωνικό σφάλμα είναι:

$$MT\Sigma(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}) = V(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}) + \left(bias_{\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}}\right)^2 = \sigma^2 \cdot \sum_{j=1}^{k} \frac{\lambda_j}{(\lambda_j + \theta)^2} + \theta^2 \boldsymbol{\beta}'(X'X + \theta I)^{-2}\boldsymbol{\beta}$$

όπου $\lambda_1, \lambda_2, \ldots, \lambda_k$ οι ιδιοτιμές του πίνακα $X'X$ και

$$V(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}) = \sigma^2 \cdot \sum_{j=1}^{k} \frac{\lambda_j}{(\lambda_j + \theta)^2}$$

$$\left(bias_{\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}}\right)^2 = \theta^2 \boldsymbol{\beta}'(X'X + \theta I)^{-2}\boldsymbol{\beta}$$

Παρατηρούμε ότι καθώς αυξάνεται η τιμή του θ η διασπορά $V(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}})$ μειώνεται ενώ η τιμή της μεροληψίας αυξάνεται. Θα πρέπει να βρούμε την κατάλληλη τιμή του θ (αν υπάρχει) ώστε τελικά, το μέσο τετραγωνικό σφάλμα $MT\Sigma(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}})$ να είναι μικρότερο της διασποράς της εκτιμήτριας ελαχίστων τετραγώνων. Για την ύπαρξη ή μη της κατάλληλης τιμής του θ οι *Hoerl* και *Kennard* απέδειξαν το ακόλουθο θεώρημα.

Θεώρημα ύπαρξης του θ

Υπάρχει πάντοτε ένα $\theta > 0$ τέτοιο ώστε: $MT\Sigma(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}) < MT\Sigma(\widehat{\boldsymbol{\beta}}) = V(\widehat{\boldsymbol{\beta}})$

με την προϋπόθεση ότι η τιμή $\boldsymbol{\beta}'\boldsymbol{\beta}$ είναι ορισμένη.

Για την απόδειξη αυτού του θεωρήματος αρκεί να δείξουμε ότι θα υπάρχει πάντα μια τουλάχιστον τιμή του θ ώστε η συνάρτηση $MT\Sigma(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}})$ ως προς θ να είναι φθίνουσα. Δηλαδή αρκεί να δείξουμε ότι:

$$\exists\, \theta > 0\colon \frac{\partial MT\Sigma(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}})}{\partial\theta} = \frac{\partial V(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}})}{\partial\theta} + \frac{\partial\left[\left(bias_{\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}}\right)^2\right]}{\partial\theta} < 0$$

Συγκρίνοντας τώρα, τον συντελεστή προσδιορισμού του μοντέλου που προκύπτει από την *Ridge* με αυτόν που προκύπτει από την μέθοδο των ελαχίστων τετραγώνων μπορούμε να έχουμε μια εικόνα για το κατά πόσο επηρεάζεται η προσαρμογή του μοντέλου με την εφαρμογή της *Ridge* παλινδρόμησης.

Γνωρίζουμε ότι:

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

και για την παλινδρόμηση Ridge έχουμε:

$$SSE_{Ridge} = SSE_{OLS} + \left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}} - \widehat{\boldsymbol{\beta}}\right)' X'X\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}} - \widehat{\boldsymbol{\beta}}\right)$$

$$SSE_{Ridge} > SSE_{OLS}\,, \qquad \theta > 0$$

Παρατηρούμε ότι καθώς το θ αυξάνει θα αυξάνεται και το SSE_{Ridge} σε σχέση με το SSE_{OLS}. Από τον τύπο λοιπόν που δίνει το R^2 και επειδή το SST είναι σταθερό, προκύπτει ότι $R^2_{Ridge} < R^2_{OLS}$. Καταλήγουμε τελικά στο συμπέρασμα ότι εφαρμόζοντας τη παλινδρόμηση Ridge έχουμε χειρότερη προσαρμογή του μοντέλου από ότι με την μέθοδο των ελαχίστων τετραγώνων. Κάτι τέτοιο δεν θα πρέπει να μας ανησυχεί διότι με την διαδικασία αυτή εμείς επιθυμούμε να βρούμε ένα σταθερό σύνολο από εκτιμημένες παραμέτρους.

Το τελευταίο στο οποίο θα πρέπει να αναφερθούμε είναι οι τρόποι με τους οποίους υπολογίζουμε την καταλληλότερη τιμή της παραμέτρου θ. Για τον σκοπό αυτό, έχει αναπτυχθεί ένας γραφικός τρόπος υπολογισμού του θ αλλά και μαθηματικοί τύποι ή ακόμα και διαδικασίες.

Η γραφική μέθοδος, γνωστή ως ***Ridge Trace***, προτάθηκε από τους *Hoerl* και *Kennard* και παρουσιάζει την εξέλιξή των συντελεστών $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}$ σε συνάρτηση με το θ, για $\theta \in (0,1)$. Σε αυτό το γράφημα η μεγάλη μεταβλητότητα των παραμέτρων για χαμηλές τιμές του θ είναι δείκτης ύπαρξης πολυσυγγραμμικότητας. Σε μια τέτοια περίπτωση, δεχόμαστε το διάστημα των τιμών του θ στο οποίο το σύνολο των καμπυλών σταθεροποιείται. Όπως καταλαβαίνουμε η μέθοδος αυτή είναι υποκειμενική και εξαρτάται από την πείρα του μελετητή.

Οι *Hoerl, Kennar* και *Baldwin (1975)* επίσης, έχουν προτείνει και τον ακόλουθο τύπο για την επιλογή του θ

$$\theta = \frac{ks^2}{\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}}}$$

όπου k ο αριθμός των στηλών του πίνακα X και $\widehat{\boldsymbol{\beta}}$, s^2 είναι οι εκτιμήτριες που προκύπτουν με την μέθοδο των ελαχίστων τετραγώνων. Οι *MacDonald* και *Galarneau (1975)* επίσης υποστήριξαν ότι η καταλληλότερη τιμή του θ προκύπτει από την λύση της εξίσωσης:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}'\widehat{\boldsymbol{\beta}}_{\boldsymbol{\theta}} = \widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}} - s^2\sum_{j=1}^{k}\frac{1}{\lambda_j}$$

Το μειονέκτημα όμως του τύπου αυτού είναι ότι μπορεί να δώσει και αρνητικές τιμές για το θ.

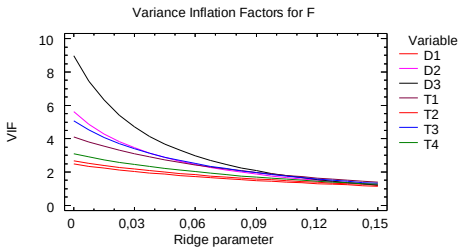
Γενικά, το πλεονέκτημα της μεθόδου αυτής είναι ότι μας επιτρέπει να ανακαλύψουμε την δομή των μεταβλητών, να αναλύσουμε την ευαισθησία των εκτιμώμενων συντελεστών των παραμέτρων και να βρούμε ένα σύνολο από εκτιμητές πιο κοντά στις πραγματικές τιμές των παραμέτρων, λαμβάνοντας υπόψη πάντα το κριτήριο που έχουμε θέσει από την αρχή, που είναι η σταθεροποίηση του Μέσου Τετραγωνικού Σφάλματος. Στο σημείο αυτό αξίζει να αναφέρουμε και την ευκολία των υπολογισμών, αφού χρησιμοποιώντας ένα μικρό αριθμό από τιμές της παραμέτρου θ μπορούμε να καταλάβουμε που σταθεροποιείται η καμπύλη *Ridge Trace*. Ωστόσο το μειονέκτημα της είναι η εκλογή της βέλτιστης τιμής του θ, διότι η εκλογή αυτή είναι υποκειμενική. Το γεγονός αυτό έκανε την παλινδρόμηση *Ridge* λιγότερο δημοφιλή από τις άλλες μεθόδους.

1.3.7 Εφαρμογή της *Ridge regression* στο παράδειγμα με τους βαθμούς των μαθητών.

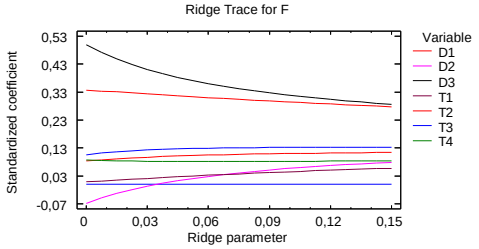
Θα προσπαθήσουμε να αντιμετωπίσουμε την πολυσυγγραμμικότητα με την προσαρμογή στα δεδομένα ενός μοντέλου της *Ridge* παλινδρόμησης. Θα πρέπει να αναφέρουμε ότι όλες οι μεταβλητές έχουν τυποποιηθεί, δηλαδή σύμφωνα με τους τύπους (1.3.5) και (1.3.6). Από τον Πίνακα Ι και το Σχήμα 1 μπορούμε να συμπεράνουμε ότι οι τιμές του *VIF* είναι μικρότερες του 5 για $\theta \geq 0.03$. Επίσης όπως έχει ειπωθεί, καθώς αυξάνονται οι τιμές του θ μειώνεται το R^2 και η τυπική απόκλιση των υπολοίπων αυξάνεται. Η μεταβολές αυτές όμως, είναι μικρές σε σχέση με τις αντίστοιχες τιμές στο μοντέλο ελαχίστων τετραγώνων ($\theta = 0$). Μπορούμε να πούμε λοιπόν ότι τα μοντέλα με τιμές του θ αυτές στον Πίνακα Ι, φαίνεται να προσαρμόζονται καλά. Αν χρησιμοποιήσουμε και την καμπύλη *Ridge Trace* (Σχήμα 2) οι τιμές των συντελεστών σταθεροποιούνται περίπου για την τιμή $\theta = 0.1$. Στον Πίνακα ΙΙ λοιπόν, παρουσιάζονται οι συντελεστές που αντιστοιχούν στο μοντέλο για $\theta = 0.1$ και τα αντίστοιχα $p - values$ που προκύπτουν από τους ελέγχους t. Τελικά καταλήξαμε σε ένα μοντέλο το οποίο δεν διαφέρει πολύ από το αρχικό των ελαχίστων τετραγώνων αλλά δεν γνωρίζουμε κατά πόσο έχει επηρεαστεί η προβλεπτική του αξία.

ΠΙΝΑΚΑΣ Ι (Variance Inflation Factors)

Ridge Parameter	D1	D2	D3	T1	T2	T3	T4	R-Squared	residual s.d
0,0	2,67549	5,61685	8,98085	4,09111	2,47622	5,07578	3,10133	87,88	0,404506
0,025	2,24823	3,70104	5,16811	3,23474	2,09296	3,60021	2,54472	87,01	0,405933
0,03	2,17954	3,45675	4,71566	3,0982	2,03045	3,39642	2,45519	86,85	0,406398
0,035	2,11488	3,24063	4,3237	2,97061	1,97151	3,21244	2,37115	86,70	0,406887
0,1	1,51105	1,73646	1,87613	1,85427	1,42218	1,81707	1,60957	85,03	0,413628
0,11	1,44416	1,61389	1,70407	1,74154	1,3617	1,6934	1,52874	84,80	0,414617
0,12	1,382	1,50592	1,55673	1,63907	1,30555	1,58318	1,45436	84,58	0,415584



Σχήμα 1



Σχήμα 2

ΠΙΝΑΚΑΣ ΙΙ (θ = 0.1)

	D1	D2	D3	T1	T2	T3	T4
Συντελεστές	0,0821634	0,0571901	0,31869	0,0451104	0,294207	0,132258	0,110367
p - values	0.24009	0.39681	0.00233 **	0.66470	0.00125 **	0.19922	0.57014

1.3.8 Παλινδρόμηση Κύριων Συνιστωσών *(Principal Component Regression, PCR)*

Η παλινδρόμηση κύριων συνιστωσών είναι μια μέθοδος η οποία αν χρησιμοποιηθεί σωστά, μπορεί να αντιμετωπίσει την πολυσυγγραμμικότητα και να οδηγήσει σε εκτιμήσεις και προβλέψεις καλύτερες από ότι η μέθοδος ελαχίστων τετραγώνων. Σύμφωνα με τη μέθοδο αυτή, οι αρχικές ανεξάρτητες μεταβλητές μετασχηματίζονται σε ένα σύνολο από ορθογώνιες και ασυσχέτιστες μεταβλητές που ονομάζονται κύριες συνιστώσες *(Principal Components)* αυτών. Ο μετασχηματισμός αυτός ταξινομεί τις ασυσχέτιστες μεταβλητές (συνιστώσες) σύμφωνα με τη σημαντικότητα τους και η διαδικασία συνεχίζεται με την εξάλειψη των λιγότερο σημαντικών, έτσι ώστε να μειωθεί η τελική διασπορά. Μετά την αφαίρεση αυτών των μεταβλητών (συνιστωσών) εφαρμόζουμε την πολλαπλή γραμμική παλινδρόμηση, με εξαρτημένη μεταβλητή αυτή που είχαμε αρχικά και ανεξάρτητες μεταβλητές τις υπόλοιπες κύριες συνιστώσες. Αυτή η προσαρμογή του μοντέλου γίνεται με τη μέθοδο των ελαχίστων τετραγώνων επειδή οι ορθογώνιες κύριες συνιστώσες είναι ανά δύο ασυσχέτιστες. Τέλος, όταν εκτιμηθούν οι συντελεστές των ορθογώνιων μεταβλητών, τους μετασχηματίζουμε σε συντελεστές που αντιστοιχούν στις αρχικές συσχετισμένες μεταβλητές.

Στην συνέχεια ακολουθεί η διαδικασία σύμφωνα με την οποία πραγματοποιούνται όλα όσα περιγράψαμε παραπάνω. Γνωρίζουμε ότι το αρχικό μοντέλο παλινδρόμησης είναι:

$$\boldsymbol{y} = X \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Όπως ήδη έχουμε πει, σκοπός της μεθόδου αυτής είναι το αρχικό μοντέλο να αντικατασταθεί από το

$$\boldsymbol{y} = Z \cdot \boldsymbol{\alpha} + \boldsymbol{\varepsilon} \qquad (1.3.9)$$

Όπου $Z = [Z_1, Z_2, \dots, Z_k]$ είναι ο $n \times k$ πίνακας του οποίου οι k στήλες είναι πλέον οι ορθογώνιες ανεξάρτητες μεταβλητές που αντικαθιστούν τις αντίστοιχες του πίνακα X, ονομάζονται κύριες συνιστώσες *(Principal Components)* των παρατηρήσεων και ορίζονται από την σχέση

$$Z = XT \qquad (1.3.10)$$

Ο πίνακας T είναι ο $k \times k$ πίνακας του οποίου οι στήλες είναι τα ιδιοδιανύσματα που αντιστοιχούν στις ιδιοτιμές του πίνακα $X'X$. Έτσι λοιπόν, επειδή ο πίνακας T είναι ορθογώνιος ($T' = T^{-1}$) το διάνυσμα $X\boldsymbol{\beta}$ γράφεται:

$$X\boldsymbol{\beta} = XTT'\boldsymbol{\beta} = Z\boldsymbol{\alpha} \qquad (1.3.11)$$

Με αυτόν τον τρόπο καταλήγουμε στη σχέση (1.3.9). Έπειτα, για την αντιμετώπιση της πολυσυγγραμμικότητας, αφαιρούμε τις μη σημαντικές συνιστώσες και η σχέση (1.3.9) γίνεται:

$$\boldsymbol{y} = Z_{(m)} \cdot \boldsymbol{\alpha_m} + \boldsymbol{\varepsilon_m}$$

όπου $\boldsymbol{\alpha_m}$ είναι το διάνυσμα που περιέχει m από τα k στοιχεία του $\boldsymbol{\alpha}$ και $Z_{(m)}$ ο $n \times m$ πίνακας που οι στήλες του αντιστοιχούν στις m σημαντικές συνιστώσες. Από την σχέση (1.3.11) προκύπτει ότι $\boldsymbol{\alpha} = T'\boldsymbol{\beta}$ και επομένως $\boldsymbol{\beta} = T\boldsymbol{\alpha}$. Τέλος, εφαρμόζοντας την μέθοδο ελαχίστων τετραγώνων μπορούμε να εκτιμήσουμε το διάνυσμα $\boldsymbol{\alpha}$ και στην συνέχεια το $\boldsymbol{\beta}$ από την ισότητα $\widehat{\boldsymbol{\beta}} = T\widehat{\boldsymbol{\alpha}}$.

Έχουμε αναφέρει ότι οι κύριες συνιστώσες, δηλαδή οι στήλες του πίνακα Z, είναι ορθογώνιες μεταξύ τους. Αυτό σημαίνει ότι:

$$Z'Z = (XT)'XT = T'X'XT = T'CT = \Lambda = diag(\lambda_1, \lambda_2, \ldots., \lambda_k)$$

Από την ισότητα αυτή προκύπτει ότι $\boldsymbol{z}_i'\boldsymbol{z}_i = \lambda_i$ και $\boldsymbol{z}_j'\boldsymbol{z}_i = 0$ για $i \neq j$. Επίσης από την σχέση (1.3.10) καταλαβαίνουμε ότι κάθε κύρια συνιστώσα του πίνακα Z είναι γραμμικός συνδυασμός της αντίστοιχης ανεξάρτητης μεταβλητής του πίνακα X με συντελεστές τις συντεταγμένες του αντίστοιχου ιδιοδιανύσματος.

Για την εκτιμήτρια ελαχίστων τετραγώνων $\widehat{\boldsymbol{\alpha}}$ έχουμε:

$$\widehat{\boldsymbol{\alpha}} = (Z'Z)^{-1}Z'\boldsymbol{y} = \Lambda^{-1}Z'\boldsymbol{y}$$

και για τον πίνακα διασποράς συνδιασποράς του $\widehat{\boldsymbol{\alpha}}$:

$$V(\widehat{\boldsymbol{\alpha}}) = \sigma^2(Z'Z)^{-1} = \sigma^2\Lambda^{-1} \qquad (1.3.12)$$

Από την σχέση αυτή προκύπτει ότι εάν μια ιδιοτιμή του πίνακα $X'X$ είναι μικρή τότε η διασπορά του αντίστοιχου συντελεστή μιας συνιστώσας θα είναι μεγάλη. Και επειδή $Z'Z = \Lambda$ συχνά αναφέρουμε σαν διασπορά της Z_j συνιστώσας την ιδιοτιμή λ_j. Εάν στο μοντέλο, επομένως, εμφανίζεται σοβαρή πολυσυγγραμμικότητα τότε όπως έχουμε ήδη πει, μια τουλάχιστον ιδιοτιμή θα είναι πολύ μικρή. Έτσι αν διαγράψουμε την συνιστώσα που σχετίζεται με την ιδιοτιμή αυτή θα πετύχουμε μείωση της συνολικής διασποράς με αποτέλεσμα να οδηγηθούμε σε καταλληλότερο και πιο αξιόπιστο μοντέλο.

Το θέμα με το οποίο πρέπει να απασχοληθούμε στην συνέχεια είναι με ποιο κριτήριο χαρακτηρίζουμε μια ιδιοτιμή πολύ μικρή. Ή γενικότερα με ποιο κριτήριο επιλέγουμε τις κατάλληλες συνιστώσες. Αρχικά, ταξινομούμε τις μεταβλητές έτσι ώστε οι ιδιοτιμές τους να βρίσκονται σε φθίνουσα σειρά, δηλαδή

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$$

Σημειώνουμε εδώ ότι η συνιστώσα με την μεγαλύτερη ιδιοτιμή συμμετέχει με το μεγαλύτερο ποσοστό στην συνολική διασπορά των ανεξάρτητων μεταβλητών και όσο προχωράμε σε μικρότερες ιδιοτιμές οι αντίστοιχες συνιστώσες συμμετέχουν λιγότερο στην συγκεκριμένη διασπορά. Πρέπει να γνωρίζουμε επίσης ότι η ιδιότητα που καθιστά την *PCR* μοναδική και πιο σύνθετη είναι ότι δεν υπάρχει μια κοινά αποδεκτή διαδικασία επιλογής των κύριων συνιστωσών που θα μετέχουν στο τελικό μοντέλο. Οι σημαντικότερες μέθοδοι που χρησιμοποιούνται για να υποδείξουν ποιες και πόσες συνιστώσες πρέπει να παραλείψουμε ώστε να πετύχουμε ουσιαστική μείωση της διασποράς είναι οι ακόλουθες:

Α. Επιλέγουμε τις συνιστώσες εκείνες που εξηγούν την συνολική διασπορά των δεδομένων κατά ένα ποσοστό που έχουμε ορίσει εμείς. Αν για παράδειγμα το ποσοστό αυτό είναι 85%, επιλέγουμε εκείνες τις συνιστώσες που το συνολικό ποσοστό συμμετοχής τους ξεπερνά για πρώτη φορά το 85%. Δηλαδή:

$$\frac{\sum_{i=1}^{s} \lambda_i}{\sum_{i=1}^{k} \lambda_i} < 0{,}85 \qquad , s < k$$

Β. Ορισμένοι ερευνητές επιλέγουν τις συνιστώσες με ιδιοτιμές μεγαλύτερες της μονάδας. Η μέθοδος αυτή είναι γνωστή ως *"Kaiser Gutman rule"*.

Γ. Επιλέγονται οι συνιστώσες που το γινόμενο των αντίστοιχων ιδιοτιμών τους είναι για τελευταία φορά μεγαλύτερο της μονάδας.

Δ. Η πιο σωστή, στατιστικά, μέθοδος είναι να γίνει ο γνωστός έλεγχος t *(t-test)* για τους συντελεστές $\hat{a}_j$ με μηδενική υπόθεση την H_0: $\hat{a}_j = 0$ και εναλλακτική την H_1: $\hat{a}_j \neq 0$. Αν $S^2 = \hat{\sigma}^2$ τότε από την σχέση (1.3.12) έχουμε

$$\hat{V}(\hat{\boldsymbol{a}}) = S^2 \Lambda^{-1} \Leftrightarrow \hat{V}(\hat{a}_j) = S^2 \lambda_j \Leftrightarrow s.e(\hat{a}_j) = S\left(\sqrt{\lambda_j}\right)^{-1}$$

Επομένως η ελεγχοσυνάρτηση t θα είναι

$$t = \frac{\hat{a}_j}{s.e(\hat{a}_j)} = \frac{\hat{a}_j \sqrt{\lambda_j}}{S}$$

Το μειονέκτημα αυτής της μεθόδου είναι ότι συνήθως μένουν παραπάνω συνιστώσες από αυτές που χρειαζόμαστε πραγματικά.

Ε. Χρήση της μεθόδου *Cross – Validation*. Η μέθοδος αυτή ενδιαφέρεται πρωτίστως, για την βελτίωση της προβλεπτικής ικανότητας του μοντέλου. Επειδή αποτελεί μια από τις βασικότερες μεθόδους επιλογής του βέλτιστου μοντέλου, η περιγραφή της πραγματοποιείται με λεπτομερή τρόπο στην παράγραφο 2.6.1.

Επιλέγοντας βέβαια, να εξαλείψουμε τις συνιστώσες με την μικρότερη διασπορά (μέθοδοι Α, Β, Γ), ουσιαστικά δεχόμαστε ότι αυτές δεν είναι σημαντικές για το μοντέλο παλινδρόμησης. Κάτι τέτοιο έχει αποδειχτεί πειραματικά ότι δεν ισχύει. Διότι ορισμένες από αυτές τις συνιστώσες είναι πιθανόν να έχουν σημαντική προβλεπτική αξία. Μπορούμε να πούμε λοιπόν, ότι οι δύο στόχοι που είναι η εξάλειψη συνιστωσών με μικρές διασπορές και η διατήρηση συνιστωσών με σημαντική προβλεπτική αξία, είναι πολύ πιθανό να μην μπορούν να επιτευχθούν ταυτόχρονα *(Jolliffe, 2002)*.

Για τον σκοπό αυτό έχουν αναπτυχθεί πιο εξελιγμένες μέθοδοι επιλογής των κατάλληλων συνιστωσών. Οι περισσότερες από αυτές τις μεθόδους, από την στιγμή που δεν καταλήγουμε σε αμερόληπτες εκτιμήτριες του $\boldsymbol{\beta}$, λαμβάνουν υπόψη τους όχι μόνο την διασπορά αλλά και την μεροληψία των εκτιμητριών. Για παράδειγμα, ο *Lott (1973)* διατύπωσε ένα κριτήριο στο οποίο χρησιμοποιείται ο διορθωμένος συντελεστής προσδιορισμού R^2_{adj} για όλα τα υποσύνολα του συνόλου των συνιστωσών. Όπου το καλύτερο υποσύνολο θεωρείται αυτό που μεγιστοποιεί το R^2_{adj}.

Αν υποθέσουμε, τώρα, ότι το πλήθος των συνιστωσών που επιλέγονται είναι $p < k$ και $\widehat{\boldsymbol{a}}_{\boldsymbol{p}}$ το διάνυσμα με τις αντίστοιχες εκτιμήτριες ελαχίστων τετραγώνων των συνιστωσών αυτών. Επίσης, ονομάζουμε T_p τον πίνακα $k \times p$ με στήλες τα ιδιοδιανύσματα που αντιστοιχούν σε αυτές τις συνιστώσες. Τότε μπορούμε να υπολογίσουμε τις εκτιμήτριες των ανεξάρτητων μεταβλητών χρησιμοποιώντας τον τύπο

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{p}} = T_p \cdot \widehat{\boldsymbol{a}}_{\boldsymbol{p}} = \sum_{j=1}^{p} \lambda_j^{-1} t_j' X' y t_j \qquad (1.3.13)$$

Αν θεωρήσουμε ότι η εκτιμήτρια $\widehat{\boldsymbol{\beta}}_{\boldsymbol{p}}$ προέρχεται από τους μετασχηματισμούς (1.3.1) και (1.3.2), τότε χρησιμοποιώντας τους τύπους (1.3.3) και (1.3.4) μπορούμε να βρούμε τις εκτιμήτριες των πραγματικών (αρχικών) ανεξάρτητων μεταβλητών.

Αξίζει να σημειώσουμε ότι μπορούμε να χρησιμοποιήσουμε και την μέθοδο *SVD (Singular Value Decomposition)* του $(n \times k)$ πίνακα X η οποία μπορεί να μας δώσει μια εναλλακτική διατύπωση του τύπου (1.3.13) και να μας βοηθήσει στην ερμηνεία των αποτελεσμάτων της παλινδρόμησης κύριων συνιστωσών. Σύμφωνα με την μέθοδο αυτή ο πίνακας X αναλύεται σε γινόμενο τριών πινάκων ως εξής:

$$X = U\Delta V'$$

Ο πίνακας U είναι ο $(n \times l)$ πίνακας με τα αριστερά κανονικοποιημένα ιδιάζοντα διανύσματα του X, ο V είναι ο $(l \times k)$ πίνακας με τα δεξιά κανονικοποιημένα ιδιάζοντα διανύσματα του X και ο Δ είναι ο $(l \times l)$ πίνακας με τις ιδιάζουσες τιμές του X (l είναι η τάξη του πίνακα X, δηλαδή $l = rank(X) \leq k$). Επιπλέον οι πίνακες U και V είναι ορθογώνιοι ($U'U = V'V = I$). Η μέθοδος *SVD* σχετίζεται άμεσα με την μέθοδο που χρησιμοποιήσαμε παραπάνω αφού ο πίνακας U περιέχει τα κανονικοποιημένα ιδιοδιανύσματα του XX', ο πίνακας V τα κανονικοποιημένα ιδιοδιανύσματα του $X'X$ και οι ιδιάζουσες τιμές είναι οι τετραγωνικές ρίζες των ιδιοτιμών του XX' και του $X'X$ (οι πίνακες αυτοί έχουν τις ίδιες ιδιοτιμές). Η χρήση της *SVD* αποτελεί πλεονέκτημα διότι μας οδηγεί σε πιο αποδοτικές υπολογιστικές διαδικασίες σε σχέση με κάποια άλλη μέθοδο.

Συγκρίνοντας τώρα, την Παλινδρόμηση Κύριων Συνιστωσών *(PCR)* με την Παλινδρόμηση *Ridge (RR)* καταλήγουμε στο ότι η βασικότερη διαφορά τους είναι πως για την *RR* η μείωση της διασποράς των εξαρτημένων μεταβλητών επιτυγχάνεται με την κατάλληλη επιλογή της τιμής της παραμέτρου θ ενώ για την *PCR* με την επιλογή των κατάλληλων συνιστωσών. Όπως ήδη έχουμε αναφέρει, η επιλογή του θ γίνεται με καθαρά εμπειρικό τρόπο ενώ για την επιλογή των κατάλληλων συνιστωσών ακολουθούμε κάποιους κανόνες. Οι κανόνες αυτοί όμως μπορεί να μας οδηγήσουν στην προσαρμογή ενός μοντέλου με χαμηλή προβλεπτικότητα. Επιπλέον, αυτό που μπορεί να κάνει η *PCR* σε σχέση με την μέθοδο των ελαχίστων τετραγώνων είναι ότι υποδεικνύει προβλήματα που σχετίζονται με την αντιμετώπιση της πολυσυγγραμμικότητας. Όπως για παράδειγμα, αν η προσπάθεια που γίνεται για να αντιμετωπιστεί η αστάθεια των συντελεστών παλινδρόμησης οδηγεί σε ταυτόχρονη απώλεια της προβλεπτικής ικανότητας του μοντέλου.

1.3.9 Εφαρμογή της *PCR* στο παράδειγμα με τους βαθμούς των μαθητών.

Αρχικά, θα πρέπει να βρούμε τις συνιστώσες που συμμετέχουν περισσότερο στην συνολική διασπορά. Έτσι, ο Πίνακας Ι παρουσιάζει μεταξύ άλλων το αθροιστικό ποσοστό συμμετοχής των συνιστωσών στην συνολική διασπορά του μοντέλου. Οι συνιστώσες είναι ταξινομημένες κατά φθίνουσα σειρά με βάση την διασπορά τους. Στην τέταρτη γραμμή του Πίνακα Ι παρουσιάζονται οι αντίστοιχοι συντελεστές προσδιορισμού, ανάλογα με το πόσες συνιστώσες πρόκειται να χρησιμοποιηθούν για την προσαρμογή του μοντέλου. Για την επιλογή του πλήθους των συνιστωσών που θα χρησιμοποιήσουμε τελικά, θα εφαρμόσουμε τις μεθόδους που περιγράψαμε στην προηγούμενη παράγραφο. Σύμφωνα με την μέθοδο Α, μια καλή επιλογή θα ήταν να χρησιμοποιηθούν τρεις συνιστώσες οι οποίες συμμετέχουν κατά 88.82% στην συνολική διασπορά. Αν ακολουθήσουμε την μέθοδο Β θα πρέπει να χρησιμοποιήσουμε μόνο την πρώτη συνιστώσα. Η μέθοδος Γ συμφωνεί με την Α αφού $\lambda_1 \cdot \lambda_2 \cdot \lambda_3 > 1$ και $\lambda_1 \cdot \lambda_2 \cdot \lambda_3 \cdot \lambda_4 < 1$. Αν ακολουθήσουμε την μέθοδο Δ θα πρέπει να προσαρμόσουμε ένα μοντέλο παλινδρόμησης με ανεξάρτητες μεταβλητές τις επτά συνιστώσες.

ΠΙΝΑΚΑΣ Ι

Comp	1	2	3	4	5	6	7
eigenval	5.28045	0.51234	0.42429	0.33120	0.20745	0.17215	0.07209
%	75.44	82.75	88.82	93.55	96.51	98.97	100.00
R^2	83.92	85.35	85.42	85.68	86.12	86.97	87.88

Φαίνεται λοιπόν ότι ένα μοντέλο με τρεις συνιστώσες έχει καλή προσαρμογή. Εκτός από αυτό όμως, θα πρέπει να επιλέξουμε και το καλύτερο μοντέλο λαμβάνοντας υπόψη μας και την προβλεπτική του ικανότητα. Είναι συνηθισμένο το φαινόμενο, ένα μοντέλο να παρουσιάζει καλή προσαρμογή και η προβλεπτική του ικανότητα να είναι περιορισμένη. Χρησιμοποιώντας την μέθοδο *cross - validation* (περιγράφεται αναλυτικά στην παράγραφο 2.6.1) θα εκτιμήσουμε την τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος των προβλέψεων *(RMSEP)*. Σκοπός μας είναι να επιλέξουμε το μοντέλο με το ελάχιστο *RMSEP*. Από τον Πίνακα ΙΙ και με την βοήθεια του Σχήματος 1 συναντάμε την ελάχιστη τιμή του *RMSEP* στο μοντέλο με τρεις συνιστώσες. Επίσης από το Σχήμα 2 φαίνεται μια καλή γραμμική σχέση μεταξύ των εκτιμημένων τιμών και των αντίστοιχων τιμών του δείγματος πράγμα που

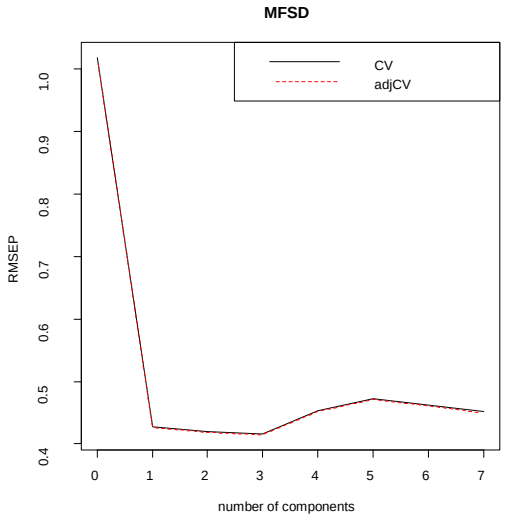
σημαίνει ότι το μοντέλο έχει καλή προσαρμογή. Προσαρμόζοντας λοιπόν στα δεδομένα το μοντέλο με τις τρεις συνιστώσες και με την χρήση του *Jackknife test*, μπορούμε να εντοπίσουμε τις σημαντικές μεταβλητές όπως φαίνεται στον Πίνακα ΙΙΙ.

Γενικά τα μοντέλα *PCR* και *Ridge* παλινδρόμησης που προσαρμόσαμε στα δεδομένα του παραδείγματος φαίνεται ότι προσαρμόζονται καλά. Δεν μπορούμε να πούμε όμως το ίδιο και για την προβλεπτική τους αξία (ιδιαίτερα για το μοντέλο της *Ridge*). Όπως θα δούμε στο επόμενο κεφάλαιο (παράγραφος 2.6) η προβλεπτική ικανότητα ενός μοντέλου θα μπορούσε να ελεγχθεί αν αυτό χρησιμοποιούνταν για προβλέψεις στα δεδομένα ενός νέου και ανεξάρτητου δείγματος.

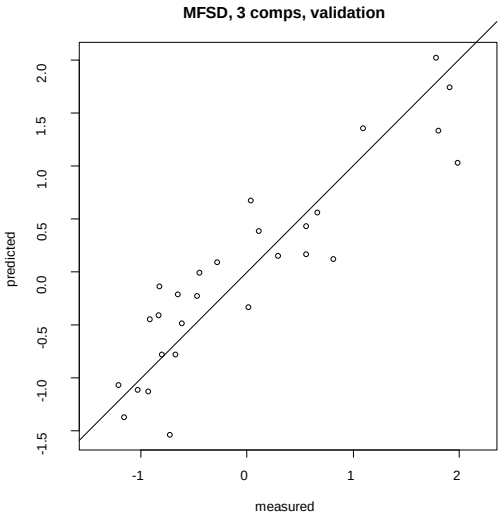
ΠΙΝΑΚΑΣ ΙΙ ($\widehat{RMSEP}$)

Μέθοδος: *Leave – one – out cross – validation*

	Intercept	Comps 1	Comps 2	Comps 3	Comps 4	Comps 5	Comps 6	Comps 7
RMSEP	1.018	0.4271	0.4199	0.4158	0.4531	0.4730	0.4625	0.4516



Σχήμα 1



Σχήμα 2

ΠΙΝΑΚΑΣ ΙΙΙ (Συντελεστών)

	D1	D2	D3	T1	T2	T3	T4
Συντελεστές	0.090157	0.195241	0.170612	0.159018	0.267939	0.088051	0.090526
p - values	0.458852	0.005721 **	0.000440 ***	0.003331 **	0.005802 **	0.151289	0.172151

ΚΕΦΑΛΑΙΟ 2

Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων

(Partial Least Square Regression, PLSR)

2.1 Ιστορική Αναδρομή

Η προσέγγιση της μεθόδου *PLS* χρονολογείται γύρω στα 1965 από τον *Herman Wold* στην προσπάθεια του να μοντελοποιήσει πολύπλοκα σύνολα δεδομένων σε μορφή αλυσίδων πινάκων τα οποία είναι γνωστά ως μοντέλα διαδρομών. Η προσέγγιση αυτή πραγματοποιήθηκε με μια απλή και αποτελεσματική διαδικασία εκτίμησης των παραμέτρων των συγκεκριμένων μοντέλων, η οποία ονομάστηκε *NIPALS (Non – linear Iterative Partial Least Squares)*. Από αυτή την ονομασία προήλθε στην συνέχεια το ακρωνύμιο *PLS (Partial Least Squares)*. Με την παλινδρόμηση PLS ασχολήθηκαν στην συνέχεια και άλλοι ερευνητές όπως οι *Lindberg et al. (1983), Wiklund et al., (2007), Abdi (2010), Krishnan et al. (2010)* και άλλοι.

Η ονομασία *PLS* προέρχεται από την βασική ιδέα του τρόπου εκτίμησης των παραμέτρων, δηλαδή ότι κάθε παράμετρος του μοντέλου υπολογίζεται με μια επαναληπτική διαδικασία, θεωρώντας ότι είναι η κλίση $\boldsymbol{\beta}$ ενός απλού γραμμικού μοντέλου παλινδρόμησης *(least squares)* μεταξύ μιας στήλης ή γραμμής ενός πίνακα που αντιστοιχεί στην εξαρτημένη μεταβλητή $\boldsymbol{y}$ και ενός διανύσματος από τις παραμέτρους που αντιστοιχεί στην ανεξάρτητη μεταβλητή $\boldsymbol{x}$. Ο όρος *"partial"* (μερικός) υποδεικνύει ότι αναφερόμαστε σε μια μερική παλινδρόμηση αφού το διάνυσμα $\boldsymbol{x}$ θεωρείται σταθερό στην εκτίμηση. Αυτό μας δίνει την δυνατότητα να προσαρμόζουμε ένα απλό μοντέλο παλινδρόμησης και στις περιπτώσεις που στην θέση των διανυσμάτων $\boldsymbol{x}$ και $\boldsymbol{y}$ έχουμε πίνακες.

Στις αρχές της δεκαετίας του 1980, το απλό μοντέλο *PLS* με δύο πίνακες τροποποιήθηκε ελαφρώς από τους *Svante Wold* (γιος του *Herman Wold*) και *Harald Martens* ώστε να είναι καταλληλότερο για τα δεδομένα της επιστήμης και της τεχνολογίας. Δείχνει επίσης, ότι είναι ικανό να ανταπεξέρχεται σε περιπτώσεις με πολύπλοκα σύνολα δεδομένων που τα άλλα μοντέλα παλινδρόμησης αντιμετωπίζουν δυσκολία ή και αδυναμία να εφαρμόσουν.

2.2 Εισαγωγή στην *PLSR*

Στο προηγούμενο κεφάλαιο είδαμε ότι μπορούμε να αντιμετωπίσουμε το φαινόμενο της πολυσυγγραμμικότητας με δύο κυρίως μεθόδους που είναι η *Ridge* Παλινδρόμηση *(RR)* και η Παλινδρόμηση Κύριων Συνιστωσών *(PCR)*.

Η προσαρμογή ενός μοντέλου με την χρήση της *PCR* αντιμετωπίζει αρκετά καλά την πολυσυγγραμμικότητα λόγω της ορθογωνιότητας των συνιστωσών. Ωστόσο το πρόβλημα της επιλογής του βέλτιστου αριθμού συνιστωσών παραμένει. Για το σκοπό αυτό αναφέραμε ορισμένους τρόπους που μας βοηθάνε να επιλέξουμε κάποιες από τις αρχικές συνιστώσες. Όμως οι συνιστώσες επιλέγονται με κριτήριο να εξηγούν όσο το δυνατό καλύτερα τις ανεξάρτητες μεταβλητές και όχι την εξαρτημένη μεταβλητή. Έτσι, τίποτα δεν μας εξασφαλίζει ότι η προβλεπτική ικανότητα ενός τέτοιου μοντέλου θα είναι η βέλτιστη.

Μια μέθοδος παλινδρόμησης που σχετίζεται αρκετά με την *PCR* είναι η Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων *(Partial Least Square Regression, PLSR)*. Ας υποθέσουμε ότι έχουμε ένα δείγμα από n παρατηρήσεις με m εξαρτημένες και k ανεξάρτητες μεταβλητές. Μπορούμε να ορίσουμε τον $n \times m$ πίνακα Y των εξαρτημένων μεταβλητών και τον $n \times k$ πίνακα X των ανεξάρτητων μεταβλητών.

Σε αντίθεση με την *PCR* η παλινδρόμηση *PLS* επιλέγει συνιστώσες από τον X οι οποίες δίνουν την καλύτερη πρόβλεψη για τον Y. Συγκεκριμένα, η παλινδρόμηση *PLS* αφού δημιουργήσει συνιστώσες ταυτόχρονα από τους X και Y, στην συνέχεια επιλέγει τις πιο κατάλληλες *(latent vectors)* στηριζόμενη στο κριτήριο ότι αυτές θα πρέπει να εξηγούν όσο το δυνατόν καλύτερα την συνδιακύμανση μεταξύ των X και Y. Έπειτα οι συνιστώσες που προέρχονται από τον X χρησιμοποιούνται για να προβλέψουν τον Y.

Γενικά, βασικός σκοπός της *PLSR* είναι να προβλέπει τον Y από τον X και να περιγράφει την κοινή τους δομή όσο το δυνατό πιο αποτελεσματικά. Αν ο Y είναι ένα διάνυσμα και η τάξη *(rank)* του X είναι η μέγιστη, τότε ο παραπάνω σκοπός μπορεί να επιτευχθεί και από την παλινδρόμηση με την μέθοδο των ελαχίστων τετραγώνων. Στην πραγματικότητα όμως οι ανάγκες της έρευνας οδηγούν στο να χρησιμοποιούμε όλο και περισσότερες ανεξάρτητες μεταβλητές. Έτσι με δεδομένο ότι το πλήθος των παρατηρήσεων τις περισσότερες φορές είναι αδύνατο να αυξηθεί, οδηγούμαστε σε ένα σύνολο δεδομένων όπου οι ανεξάρτητες μεταβλητές εμφανίζουν έντονες

συσχετίσεις (πολυσυγγραμμικότητα). Το πρόβλημα αυτό είναι γνωστό ως *"small N large P problem"*, όπου *N* το πλήθος των παρατηρήσεων και *P* το πλήθος των ανεξάρτητων μεταβλητών. Το πρόβλημα αυτό το συναντάμε σε αρκετούς τομείς της επιστήμης όπως στην βιοπληροφορική ,στην χημειομετρία ,στην γονιδιοματική, στην νευροαπεικόνιση κ.ά. Η *PLSR* λοιπόν, δημιουργήθηκε για να επιλύει σύνθετα προβλήματα σαν το προηγούμενο και να αναλύει τα δεδομένα με πιο ρεαλιστικό τρόπο.

2.3 Το μοντέλο *PLSR*

Η διαδικασία για την εφαρμογή ενός μοντέλου παλινδρόμησης *PLS* στηρίζεται ουσιαστικά στην αποσύνθεση των δύο πινάκων X και Y, με στόχο να δημιουργηθεί ένα κοινό σύνολο από ορθογώνιους παράγοντες και ένα σύνολο από συγκεκριμένα *loadings*. Έτσι λοιπόν, ορίζεται ο πίνακας T με στήλες τις λίγες συνιστώσες που δημιουργούνται. Οι συνιστώσες αυτές ονομάζονται *latent vectors* ενώ ο πίνακας T ονομάζεται *Score Matrix* και είναι ορθογώνιος ($T'T = I$).

Υποθέτουμε ότι το πλήθος των *latent vectors* είναι l και συμβολίζουμε το καθένα από αυτά με $\boldsymbol{t_a}$, όπου $\alpha = 1,2,\dots\dots, l$. Τα $\boldsymbol{t_a}$ είναι ορθογώνια διανύσματα και γραμμικοί συνδυασμοί των αρχικών X_j $(j = 1,2,\dots\dots, k)$ με συντελεστές που προκύπτουν από τα διανύσματα $\boldsymbol{w_a}$ τα οποία ονομάζονται συντελεστές βαρύτητας *(weights)*. Δηλαδή:

$$\boldsymbol{t_a} = X\boldsymbol{w_a} \qquad (2.3.1)$$

Τα *latent vectors* $\boldsymbol{t_a}$έχουν δύο βασικές ιδιότητες οι οποίες ισχύουν και για τον αντίστοιχο $(n \times l)$ πίνακα T.

α. Αν ονομάσουμε με P τον $(k \times l)$ πίνακα που περιέχει τα *loadings (loading matrix)* τότε για τον αρχικό πίνακα X ισχύει

$$X = TP' \qquad (2.3.2)$$

Αντίστοιχη σχέση με την (2.3.2) προκύπτει για τον πίνακα Y (όταν $m > 1$), αν ονομάσουμε με U τον $(n \times l)$ πίνακα *(Score Matrix)* των συνιστωσών *(latent vectors)* $\boldsymbol{u_a}$ και με C τον $(m \times l)$ πίνακα των αντίστοιχων διανυσμάτων βαρύτητας $\boldsymbol{c_\alpha}$. Τότε θα έχουμε:

$$\boldsymbol{u_a} = Y\boldsymbol{c_\alpha} \qquad (2.3.3)$$

β. Τα διανύσματα $\boldsymbol{t_a}$ μπορούν να δώσουν αρκετά καλές προβλέψεις για τον Y σύμφωνα με τον τύπο:

$$\hat{Y} = TC'$$

$$\text{ή αλλιώς} \qquad Y = TC' + F \qquad (2.3.4)$$

όπου F ο $(n \times m)$ πίνακας των καταλοίπων που προκύπτουν από την προσαρμογή του παραπάνω μοντέλου.

Από τους τύπους (2.3.2) και (2.3.4) προκύπτει η εξής ισότητα:

$$Y = XWC' + F = XB + F$$

Καταλήγουμε λοιπόν, σε μια ισότητα που μοιάζει με το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης. Ο $(k \times m)$ πίνακας B για τον οποίο ισχύει ότι $B = WC'$ ονομάζεται πίνακας συντελεστών της *PLSR*.

Ενδιαφέρον παρουσιάζει η περίπτωση κατά την οποία ο πίνακας Y έχει μια στήλη ($m = 1$) και ο πίνακας $X'X$ είναι διαγώνιος. Στην περίπτωση αυτή δεν υπάρχει συσχέτιση μεταξύ των στηλών του πίνακα X και έτσι η *PLSR* χρησιμοποιώντας μια συνιστώσα ταυτίζεται με το κλασικό μοντέλο της πολλαπλής γραμμικής παλινδρόμησης με τους συντελεστές παλινδρόμησης να είναι ίσοι με

$$\boldsymbol{w_1} \cdot \boldsymbol{c_1}'$$

Από τους τύπους (2.3.1) και (2.3.3) προκύπτει ότι το σημαντικό για την *PLS* παλινδρόμηση είναι να βρεθούν τα κατάλληλα διανύσματα βαρύτητας $\boldsymbol{w_a}$ και $\boldsymbol{c_a}$ προκειμένου να ορισθούν γραμμικοί συνδυασμοί για τις στήλες των πινάκων X και Y αντίστοιχα. Επιπλέον, τα $\boldsymbol{w_a}$ και $\boldsymbol{c_a}$ θα πρέπει να έχουν ορισθεί έτσι ώστε κάθε ζευγάρι γραμμικών συνδυασμών να έχει μέγιστη συνδιασπορά *(Abdi, 2010)*.

Από τον αλγόριθμο της *PLSR*, που θα δούμε αναλυτικά παρακάτω, προκύπτει η χρήση ιδιοδιανυσμάτων. Έτσι μπορούμε να υποθέσουμε ότι η *PLSR* σχετίζεται με την *SVD*. Αυτό είναι κάτι που ισχύει πραγματικά, αφού η *SVD* θεωρείται το βασικό εργαλείο ανάλυσης στην *PLSR*. Πράγματι το πρώτο διάνυσμα βαρύτητας $\boldsymbol{w_1}$ θα αποδείξουμε αργότερα ότι είναι το πρώτο τυποποιημένο ιδιοδιάνυσμα του πίνακα $X'YY'X$ ενώ τα υπόλοιπα διανύσματα προκύπτουν με αντίστοιχο τρόπο αν ο αρχικός πίνακας X αντικατασταθεί από τον νέο πίνακα X_a μετά τον υπολογισμό της συνιστώσας $\boldsymbol{t_a}$, δηλαδή το $\boldsymbol{w_a}$ είναι το πρώτο τυποποιημένο ιδιοδιάνυσμα του $X_\alpha' YY' X_\alpha$. Παρατηρούμε επομένως ότι το $\boldsymbol{w_1}$, για παράδειγμα, είναι το πρώτο δεξιό ιδιάζον διάνυσμα του πίνακα $X'Y$. Με όμοιο τρόπο αποδεικνύεται ότι το $\boldsymbol{c_1}$ είναι το πρώτο τυποποιημένο ιδιοδιάνυσμα του $Y'XX'Y$, δηλαδή είναι το πρώτο αριστερό ιδιάζον διάνυσμα του $X'Y$. Με ανάλογους τρόπους αποδεικνύεται ότι το $\boldsymbol{t_1}$ είναι το πρώτο ιδιοδιάνυσμα του πίνακα $XX'YY'$ και το $\boldsymbol{u_1}$ το πρώτο ιδιοδιάνυσμα του $YY'XX'$ (*Wold et al., 2001)*.

Από τα παραπάνω ιδιοδιανύσματα προκύπτει ότι τα διανύσματα $\boldsymbol{w_a}$ και $\boldsymbol{t_a}$ είναι ορθογώνια μεταξύ τους. Τα διανύσματα $\boldsymbol{p_a}$ του πίνακα P δεν είναι ορθογώνια μεταξύ τους όπως επίσης και οι συνιστώσες $\boldsymbol{u_a}$ του Y. Όμως τα $\boldsymbol{u}$ και $\boldsymbol{p}$ είναι ορθογώνια με τα $\boldsymbol{t}$ και $\boldsymbol{c}$ αντίστοιχα, όταν τα $\boldsymbol{t}$ και $\boldsymbol{c}$ αντιστοιχούν σε προηγούμενες συνιστώσες από τα $\boldsymbol{u}$ και $\boldsymbol{p}$. Δηλαδή, $\boldsymbol{u_b' t_a} = 0$ και $\boldsymbol{p_b' w_a} = 0$ για $b > \alpha$. Επίσης ισχύει ότι $\boldsymbol{w_\alpha' p_a} = 1$.

2.4 Ο αλγόριθμος *NIPALS* για την *PLSR*

Η μέθοδος *PLSR*, όπως και άλλα μοντέλα παλινδρόμησης, μπορεί να διαχειριστεί περιπτώσεις όπου ένα μεσαίου μεγέθους πλήθος από δεδομένα έχει παραληφθεί από τους πίνακες X και Y (ο Y πρέπει να έχει περισσότερες από μια στήλες). Όσο πιο πολλές είναι οι παρατηρήσεις και οι μεταβλητές, τόσο περισσότερα «χαμένα» δεδομένα μπορεί να «αντέξει» η μέθοδος. Όλα αυτά βέβαια ισχύουν με την προϋπόθεση ότι η παράληψη των δεδομένων έχει γίνει με τυχαίο τρόπο και δεν έχει ακολουθηθεί κάποια διαδικασία.

Οι περισσότεροι αλγόριθμοι που έχουν αναπτυχθεί για τον υπολογισμό του μοντέλου της *PLSR* έχουν σαν σκοπό να μπορούν να διαχειρίζονται περιπτώσεις όπου δεδομένα έχουν παραληφθεί. Ο αλγόριθμος *NIPALS* που θα περιγράψουμε αναλυτικά παρακάτω, εντοπίζει αυτόματα τα δεδομένα που λείπουν και έπειτα τα αντικαθιστά με επαναληπτικό τρόπο, χρησιμοποιώντας τις προβλέψεις που έχουν γίνει στο μοντέλο. Αυτό έχει σαν αποτέλεσμα, τα δεδομένα αυτά να έχουν μηδενικά κατάλοιπα (σε κάθε συνιστώσα) και έτσι να μην επηρεάζουν τις τιμές των $\boldsymbol{t_\alpha}$ και $\boldsymbol{p_\alpha}$. Επίσης χρησιμοποιεί τους αυθεντικούς πίνακες X και Y που έχουν τυποποιηθεί, ενώ άλλοι αλγόριθμοι χρησιμοποιούν διαφορετικούς πίνακες. Όπως, για παράδειγμα, ο αλγόριθμος *Kernel* που χρησιμοποιεί τους πίνακες διασποράς – συνδιασποράς $X'X$ και $Y'Y$. Εκτός από τις μεθόδους *NIPALS* και *Kernelpls* χρησιμοποιούνται και άλλες με πιο γνωστές τις *SIMPLS (Straightforward Implementation of a statistically inspired Modification of the PLS model)* και *WAPLS (Weighted Averaging PLS)*.

Σύμφωνα με τον αλγόριθμο *NIPALS* λοιπόν, ακολουθούμε τα παρακάτω βήματα:

Πριν ξεκινήσουμε μετασχηματίζουμε τους πίνακες X και Y με μια από τις μεθόδους που περιγράψαμε στην παράγραφο (1.3.1) στους πίνακες E και D αντίστοιχα.

Βήμα 1ο: Επιλέγω ένα διάνυσμα $\boldsymbol{u}$ για να ξεκινήσω. Συνήθως αυτό είναι μια από τις στήλες του Y.

Βήμα 2ο: Υπολογίζω $\boldsymbol{w} = E'\boldsymbol{u}$

Βήμα 3ο: Υπολογίζω $\boldsymbol{t} = E\boldsymbol{w}$

Βήμα 4ο: Υπολογίζω $\boldsymbol{c} = D'\boldsymbol{t}$

Βήμα 5ο: Υπολογίζω $\boldsymbol{u} = D\boldsymbol{c}$

(Κάθε νέο διάνυσμα που προκύπτει από τα παραπάνω βήματα το κανονικοποιώ (normalize) και το χρησιμοποιώ στο επόμενο βήμα)

Βήμα 6^{ο}: Ελέγχω εάν το διάνυσμα $\boldsymbol{t}$ συγκλίνει σύμφωνα με την σχέση

$$\frac{\|\boldsymbol{t_{old}} - \boldsymbol{t_{new}}\|}{\|\boldsymbol{t_{new}}\|} < \varepsilon$$

όπου ε ένας πολύ μικρός αριθμός της τάξης του 10^{-6} ή 10^{-8}, $\boldsymbol{t_{old}}$ η συνιστώσα του προηγούμενου βήματος και $\boldsymbol{t_{new}}$ η νέα συνιστώσα. Αν το $\boldsymbol{t}$ συγκλίνει συνεχίζω στο επόμενο Βήμα 7, διαφορετικά επιστρέφω στο Βήμα 2.

Βήμα 7^{ο}: Υπολογίζω τους νέους *(deflated)* πίνακες X και Y σύμφωνα με τους τύπους:

$$\boldsymbol{p} = E'\boldsymbol{t}$$
$$E = E - \boldsymbol{t}\boldsymbol{p}'$$
$$D = D - \boldsymbol{t}\boldsymbol{c}'$$

Βήμα 8^{ο}: Υπολογίζω την επόμενη συνιστώσα (επιστροφή στο Βήμα1) έως ότου η μέθοδος *cross validation* υποδείξει ότι ο E δεν μου δίνει σημαντικές πληροφορίες για τον D.

Από τον αλγόριθμο αυτό προκύπτει ότι, το άθροισμα των τετραγώνων του X που εξηγείται από μια συνιστώσα *(latent vector)* $\boldsymbol{t}$ υπολογίζεται χρησιμοποιώντας το αντίστοιχο διάνυσμα $\boldsymbol{p}$ από τον τύπο $\boldsymbol{p}'\boldsymbol{p}$, ενώ για τον πίνακα Y το συγκεκριμένο άθροισμα τετραγώνων αντιστοιχεί στην τιμή b^2 όπου $b = \boldsymbol{t}'\boldsymbol{u}$. Διαιρώντας τώρα, τα αποτελέσματα αυτά με τα αντίστοιχα συνολικά αθροίσματα τετραγώνων των X και Y υπολογίζουμε το ποσοστό που η κάθε συνιστώσα συμμετέχει στην διασπορά των πινάκων αυτών.

Επίσης, με την βοήθεια του παραπάνω αλγορίθμου παρατηρούμε ότι για κάθε διάνυσμα $\boldsymbol{w}$ ισχύει ότι:

$$\boldsymbol{w} = E'\boldsymbol{u} = E'D\boldsymbol{c} = E'DD'\boldsymbol{t} = E'DD'E\boldsymbol{w}$$

Αποδεικνύεται δηλαδή ο ισχυρισμός μας στην παράγραφο 2.3, όπου το $\boldsymbol{w}$ είναι το πρώτο ιδιοδιάνυσμα του $E'DD'E$ ή το πρώτο δεξιό ιδιάζων διάνυσμα του $E'D$. Όμοια αποδεικνύεται και ότι το διάνυσμα $\boldsymbol{c}$ είναι το πρώτο αριστερό ιδιάζον διάνυσμα του $E'D$. Γίνεται κατανοητή λοιπόν, η σχέση που υπάρχει μεταξύ της μεθόδου *SVD* και του αλγόριθμου *NIPALS*.

2.5 Ερμηνεία και καλή προσαρμογή ενός μοντέλου *PLSR*

Απ' όσα είπαμε στις προηγούμενες παραγράφους καταλαβαίνουμε ότι η *PLS* παλινδρόμηση δημιουργεί νέες ανεξάρτητες μεταβλητές (τις συνιστώσες $\boldsymbol{t_\alpha}$) οι οποίες είναι γραμμικοί συνδυασμοί των αρχικών (X_j) που στην συνέχεια χρησιμοποιούνται για την πρόβλεψη του Y. Το πλήθος των συνιστωσών $\boldsymbol{t_\alpha}$ είναι τέτοιο ώστε το μοντέλο να έχει καλή προβλεπτική ικανότητα.

Επιχειρώντας μια πιο λεπτομερή ερμηνεία της *PLSR* μπορούμε να πούμε ότι τα διανύσματα *(scores)* $\boldsymbol{t}$ και $\boldsymbol{u}$ μας παρέχουν πληροφορίες για τις μεταβλητές καθώς και τις ομοιότητες ή διαφορές τους. Τα διανύσματα βαρύτητας $\boldsymbol{w}$ και $\boldsymbol{c}$, μας πληροφορούν για το αν οι μεταβλητές συνδυάζονται ώστε να αναδείξουν την ποσοτική σχέση μεταξύ των X και Y. Έτσι, μας βοηθάνε να εντοπίσουμε τις σημαντικές ανεξάρτητες μεταβλητές, αλλά και τις ανεξάρτητες μεταβλητές που μας δίνουν ίδιες πληροφορίες. Ένα μοντέλο *PLSR* όμως δεν ασχολείται με τα κατάλοιπα *(residuals)* που προκύπτουν κατά την προσαρμογή του. Γνωρίζουμε ότι οι υψηλές τιμές καταλοίπων υποδεικνύουν ότι το μοντέλο μας δεν είναι το βέλτιστο. Σε μια τέτοια περίπτωση το διάγραμμα της κανονικής πιθανότητας, για κάθε εξαρτημένη μεταβλητή ξεχωριστά, θα μπορούσε να μας υποδείξει ακραίες τιμές που προκύπτουν από την σχέση μεταξύ των T και Y.

Η παλινδρόμηση *PLS* έχει επίσης την δυνατότητα να αναλύει ταυτόχρονα περισσότερες από μία εξαρτημένες μεταβλητές και έτσι να παρουσιάζει μια ολοκληρωμένη εικόνα στην έρευνα που κάνουμε. Από την άλλη όμως, προσαρμόζοντας ένα μοντέλο για όλες τις μεταβλητές υπάρχει ο κίνδυνος να χρησιμοποιήσουμε πολλές συνιστώσες κάτι που θα μας δυσκολέψει στην ερμηνεία του. Η απάντηση στο δίλλημα αυτό είναι ότι εάν οι εξαρτημένες μεταβλητές συσχετίζονται τότε είναι καλύτερα να τις αναλύουμε όλες μαζί στο ίδιο μοντέλο. Στην αντίθετη περίπτωση προτιμάμε να αναλύουμε κάθε μια ξεχωριστά. Όπως αναφέραμε στο πρώτο κεφάλαιο, υπάρχουν αρκετοί τρόποι για να ελέγξουμε αν οι μεταβλητές αυτές συσχετίζονται. Για παράδειγμα, θα μπορούσαμε να υπολογίσουμε τον πίνακα συσχέτισης $Y'Y$ ή να υπολογίσουμε τις αντίστοιχες τιμές *VIF* των εξαρτημένων μεταβλητών.

Ένα ακόμη πιο σημαντικό πρόβλημα που αντιμετωπίζουμε στην παλινδρόμηση *PLS*, όπως και στην *PCR*, είναι η επιλογή του κατάλληλου πλήθους συνιστωσών ώστε το μοντέλο που θα προσαρμόσουμε να έχει την βέλτιστη

προβλεπτική ικανότητα. Σε αντίθετη περίπτωση υπάρχει ο κίνδυνος να καταλήξουμε σε ένα πολύ καλό μοντέλο που όμως θα έχει μικρή ή καθόλου προβλεπτική αξία *(over fitting)*. Το μέτρο, σύμφωνα με το οποίο μπορούμε να ελέγξουμε την ποιότητα των προβλέψεων που μπορεί να δώσει ένα μοντέλο για οποιοδήποτε πλήθος συνιστωσών, είναι το Μέσο Τετραγωνικό Σφάλμα των Προβλέψεων (*Mean Square Error of Prediction, MSEP*). Οι τεχνικές που χρησιμοποιούνται για την εκτίμηση του *MSEP* απαιτούν, λόγω της πολυπλοκότητας τους, την χρήση υπολογιστών και είναι, κυρίως, η *cross – validation* και η *bootstrap*. Η πολυπλοκότητα των τεχνικών αυτών οφείλεται στο γεγονός ότι βασίζονται στην συνεχή δειγματοληψία *(resampling)*. Οι τρόποι εκτίμησης του *MSEP* αναφέρονται αναλυτικά στην επόμενη παράγραφο.

2.6 Εκτιμήτριες του Μέσου Τετραγωνικού Σφάλματος των Προβλέψεων *(MSEP)*

Για την εκτίμηση του *MSEP* το σύνολο των δεδομένων χωρίζεται σε δύο υποσύνολα. Το ένα υποσύνολο το ονομάζουμε *training set* και χρησιμοποιείται για την προσαρμογή ενός μοντέλου *PLSR* ή *PCR* και με το άλλο υποσύνολο, γνωστό ως *testing set*, εκτιμούμε το *MSEP*. Οι εκτιμήτριες αυτές μπορούν να συγκριθούν μεταξύ τους ως προς την μεροληψία τους, την τυπική τους απόκλιση και το τετραγωνικό τους σφάλμα.

Υποθέτουμε λοιπόν, ότι ένα σύνολο από n παρατηρήσεις χωρίζεται σε δύο υποσύνολα. Το σύνολο $L = \{(\boldsymbol{x}_i, y_i)\}$ που αντιστοιχεί στο *training set* και αποτελείται από n_L παρατηρήσεις και το σύνολο $T = \{(\boldsymbol{x}_{T,i}, y_{T,i})\}$ που αντιστοιχεί στο *testing set* και αποτελείται από n_T παρατηρήσεις. Όπως έχουμε πει, το υποσύνολο L χρησιμοποιείται για την προσαρμογή ενός μοντέλου *PLSR* ή *PCR*.

Οι περισσότερες μέθοδοι που ελέγχουν τη προβλεπτική αξία ενός μοντέλου στηρίζονται στην ακόλουθη λογική. Εφόσον έχουμε εκτιμήσει το μοντέλο από το *training set*, στην συνέχεια το χρησιμοποιούμε για να υπολογίσουμε τις τιμές $\hat{Y}$ που προκύπτουν από το *testing set*. Έπειτα υπολογίζουμε το μέσο τετραγωνικό σφάλμα των προβλέψεων *(Mean Square Error of Prediction, MSEP)* που αντιστοιχεί στο *testing set*. Είναι προφανές ότι το *MSEP* θα είναι μεγαλύτερο από το μέσο τετραγωνικό σφάλμα *(MSE)* του μοντέλου. Η μελέτη και σύγκριση των τιμών του *MSEP*, όπως θα δούμε παρακάτω, μας βοηθά στην επιλογή του κατάλληλου αριθμού συνιστωσών. Για το σκοπό αυτό, μπορούν να υπολογισθούν και άλλα στατιστικά μεγέθη όπως για παράδειγμα το *PRESS* το οποίο είναι ισοδύναμο με το *MSEP* αφού:

$$PRESS = n_L \times MSEP$$

Οι κυριότεροι τρόποι εκτίμησης του *MSEP* είναι η *cross – validation* και η *bootstrap* στις οποίες θα αναφερθούμε αναλυτικά παρακάτω. Εκτός όμως από αυτές τις τεχνικές, υπάρχουν και άλλες *(Mevik και Cederkvisk, 2004)*, όπως :

1. <u>Εκτίμηση του *MSEP* από το *testing set*.</u>

$$MSEP_{test} = \frac{1}{n_T}\sum_{i=1}^{n_T}\left(\hat{y}_{T,i}^{L} - y_{T,i}\right)^2 \qquad (2.6.1)$$

Όπου $\hat{y}_{T,i}^{L}$ οι εκτιμήσεις (προβλέψεις) των μεταβλητών του *testing set* οι οποίες προκύπτουν από το μοντέλο που έχουμε προσαρμόσει στα δεδομένα του *training set*. Αυτή η εκτιμήτρια είναι αμερόληπτη.

2. Εκτίμηση του MSEP από το *training set (Apparent MSEP)*

$$MSEP_{app} = \frac{1}{n_L}\sum_{i=1}^{n_L}(\hat{y}_i - y_i)^2 \qquad (2.6.2)$$

Η εκτιμήτρια αυτή είναι μεροληπτική και η μεροληψία της αυξάνεται με την προσθήκη περισσότερων μεταβλητών ή συνιστωσών στο μοντέλο. Λόγω της υψηλής μεροληψίας δεν χρησιμοποιείται ως εκτιμήτρια αλλά για τον υπολογισμό άλλων εκτιμητριών.

2.6.1 Η τεχνική *Cross – Validation (CV)*

Η τεχνική *Cross – Validation (CV)* είναι μια μέθοδος επαναληπτικής δειγματοληψιας *(resampling method)* και θεωρείται ένας πρακτικός και αξιόπιστος τρόπος μέτρησης της προβλεπτικής ικανότητας ενός στατιστικού μοντέλου. Σύμφωνα με την τεχνική αυτή η προβλεπτική ικανότητα ενός μοντέλου ελέγχεται με την χρήση δεδομένων τα οποία δεν έχουν χρησιμοποιηθεί για την εκτίμηση του. Όπως ήδη έχουμε αναφέρει, το σύνολο των δεδομένων αυτών ονομάζεται *testing set* και το αντίστοιχο σύνολο δεδομένων που χρησιμοποιείται για την εκτίμηση του μοντέλου καλείται *training set*. Θα πρέπει να σημειώσουμε εδώ ότι τα *training set* και *testing set* προκύπτουν από το διαχωρισμό του αρχικού *training set* $L = \{(\boldsymbol{x}_i, y_i)\}$ και τις περισσότερες φορές επιλέγονται τυχαία από τον υπολογιστή. Στην παράγραφο αυτή με τους όρους *training set* και *testing set* **δεν** θα εννοούμε τα υποσύνολα Lκαι T. Με αυτό τον τρόπο έχουμε την δυνατότητα να ελέγξουμε δύο φορές την προβλεπτική ικανότητα των μοντέλων. Μια φορά από την εκτιμήτρια $MSEP_{test}$ και μια από την εκτιμήτρια $MSEP_{CV}$. Ωστόσο, επειδή πολλές φορές το πλήθος των δεδομένων είναι αρκετά μικρό και επειδή τα σύνολα Lκαι T συνήθως δεν είναι ανεξάρτητα αφού προέρχονται από το ίδιο δείγμα, επιλέγουμε να μην διαχωρίσουμε τα δεδομένα στα σύνολα Lκαι T διότι μπορεί καταλήξουμε σε διαφορετικό μοντέλο από αυτό που θα καταλήγαμε αν δεν τα είχαμε διαχωρίσει. Έτσι η εκτίμηση του *MSEP* πραγματοποιείται μόνο από την *CV*. Η ιδανική περίπτωση θα ήταν να διαθέταμε δύο διαφορετικά δείγματα όπως στο Κεφάλαιο 3 στην εφαρμογή 3 *(Fearn, 1983)*.

Θα περιγράψουμε τώρα τις τρεις βασικές διαδικασίες που έχουν αναπτυχθεί για την εφαρμογή της τεχνικής *Cross – Validation*, καθώς επίσης και μια γραφική απεικόνιση των διαδικασιών αυτών. Στα γραφήματα που ακολουθούν το σχήμα 1 παριστάνει το σύνολο των παρατηρήσεων, το σχήμα 2 ένα από τα Κ ισομεγέθη ανεξάρτητα υποσύνολα, και το σχήμα 3 μια παρατήρηση.

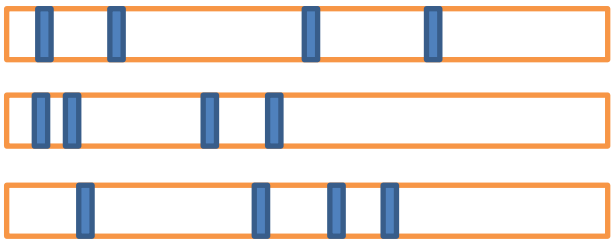
(Σχήμα 1)

(Σχήμα 2)

(Σχήμα 3)

1. *Random Subsampling ή leave – k – out cross – validation*

   Σύμφωνα με την διαδικασία αυτή εκτιμούμαι το *MSEP* επιλέγοντας, τυχαία, *k* παρατηρήσεις που χρησιμοποιούνται ως *testing set*. Η *leave – k – out cross – validation* είναι η γενίκευση της *LOOCV (k=1)* που περιγράφεται παρακάτω και θεωρείται αρκετά επίπονη και δαπανηρή, σε υπολογιστικό επίπεδο, αφού χρησιμοποιεί όλα τα δυνατά υποσύνολα με *k* στοιχεία ως *testing set*.



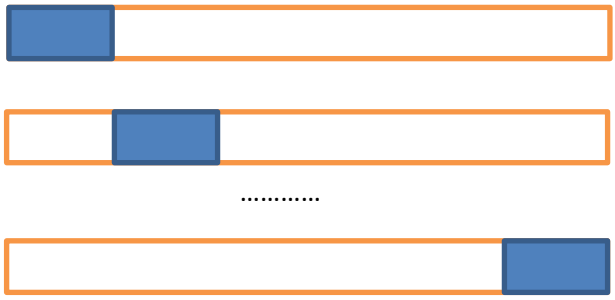
2. *K – fold cross – validation*

   Το σύνολο L χωρίζεται σε Κ ισομεγέθη υποσύνολα L_k με $k = 1,2,\dots,K$. Σε κάθε βήμα εκτιμούμε το *MSEP* όπου το *training set* αποτελείται από τα $K-1$ υποσύνολα και το *testing set* από το υποσύνολο που απομένει. Η διαδικασία αυτή ολοκληρώνεται όταν όλα τα υποσύνολα χρησιμοποιηθούν σαν *testing set*. Άρα η διαδικασία αυτή αποτελείται από K βήματα.

Η εκτίμηση του *MSEP* δίνεται από τον τύπο

$$MSEP_{CV} = \frac{1}{n_L}\sum_{k=1}^{K}\sum_{i\epsilon L_k}(\hat{y}_i - y_i)^2 \qquad (2.6.3)$$

όπου, τα $\hat{y}_i$ προκύπτουν από την εξίσωση παλινδρόμησης η οποία έχει υπολογισθεί από το *training set* (χωρίς το k-οστό υποσύνολο) χρησιμοποιώντας τα δεδομένα του k-οστού υποσυνόλου, και οι τιμές y_i προέρχονται επίσης από τις παρατηρήσεις του k-οστού υποσυνόλου.



3. *Leave – one – out cross – validation (LOOCV)* ή *full cross – validation*

Είναι συχνό το φαινόμενο το πλήθος των παρατηρήσεων να είναι αρκετά μικρό με αποτέλεσμα να μην είναι εφικτός ο διαχωρισμός των δεδομένων σε *training set* και *testing set*. Έτσι λοιπόν, επιλέγουμε την διαδικασία αυτή όπου, ουσιαστικά, είναι η προηγούμενη διαδικασία (*K – fold cross – validation*) για $K = n_L$. Δηλαδή κάθε *testing set* θα αποτελείται από ένα στοιχείο.

Ακολουθεί μια σύντομη περιγραφή αυτής της διαδικασίας η οποία θα μας βοηθήσει να κατανοήσουμε και τις προηγούμενες.

Υποθέτουμε ότι $y_1, y_2, \dots, y_{n_L}$ οι τιμές της εξαρτημένης μεταβλητής.

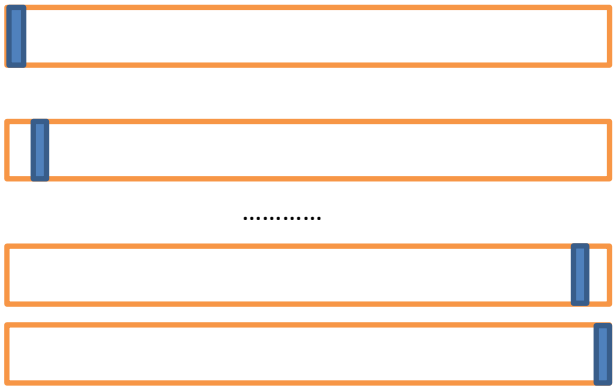
α. Έστω ότι η i-οστή παρατήρηση αποτελεί το *testing set* τότε υπολογίζουμε το υπόλοιπο $e_i^* = \hat{y}_i - y_i$.

β. Επαναλαμβάνουμε το βήμα (α) για $i = 1,2, \dots, n_L$.

γ. Υπολογίζουμε το Μέσο Τετραγωνικό Σφάλμα των Προβλέψεων *(MSEP)* των $e_1^*, e_2^*, \dots, e_{n_L}^*$ σύμφωνα με τον τύπο (2.6.3) ο οποίος απλοποιείται και γίνεται

$$MSEP_{LOOCV} = \frac{1}{n_L}\sum_{i=1}^{n_L}(\hat{y}_i - y_i)^2$$

Η εκτιμήτρια του *MSEP* που προκύπτει με αυτή την διαδικασία είναι σχεδόν αμερόληπτη.



Η ελαχιστοποίηση του *MSEP* θεωρείται ένας αξιόπιστος τρόπος για την επιλογή του κατάλληλου μοντέλου. Θεωρείται, επίσης, καλύτερος από διαδικασίες που βασίζονται σε στατιστικούς ελέγχους και μας παρέχει σχεδόν αμερόληπτες εκτιμήτριες του *MSE* των νέων παρατηρήσεων. Ωστόσο, όπως συμβαίνει σε κάθε διαδικασία επιλογής μεταβλητών ή συνιστωσών, μπορεί να χρησιμοποιηθεί καταχρηστικά. Γι' αυτό θα πρέπει να είμαστε προσεκτικοί με τους στατιστικούς ελέγχους, μετά την επιλογή μεταβλητών με την μέθοδο *cross – validation*. Σε μια τέτοια περίπτωση οι στατιστικοί έλεγχοι δεν λαμβάνουν υπόψη τους την επιλογή των μεταβλητών που έχει προηγηθεί και οι τιμές *p – value* ενδέχεται να μας παραπλανήσουν. Σε αυτό το σημείο θα πρέπει να σημειώσουμε ότι τις περισσότερες φορές χρησιμοποιείται η τετραγωνική ρίζα του *MSEP*. Δηλαδή

$$RMSEP_{CV} = \sqrt{MSEP_{CV}} = \sqrt{\frac{PRESS}{n_L}}$$

Για την δεύτερη διαδικασία που περιγράψαμε (*K – fold cross – validation*), η εξίσωση παλινδρόμησης εκτιμάται από τα $K-1$ υποσύνολα του συνόλου L για κάθε ένα από τα K βήματα. Είναι φυσικό λοιπόν, οι εξισώσεις αυτές να διαφέρουν από την εξίσωση που θα προέκυπτε αν χρησιμοποιούσαμε όλα τα στοιχεία του συνόλου L. Κάτι τέτοιο μπορεί να μας οδηγήσει σε υπερεκτίμηση του *MSEP*, ιδιαίτερα όταν $K \ll n_L$. Για τον σκοπό αυτό ορίζουμε την διορθωμένη εκτιμήτρια του $MSEP_{CV}$ (*Adjusted K – fold cross – validation*) όπως φαίνεται αμέσως παρακάτω.

Αρχικά διορθώνουμε το $MSEP_{app}$ και η διόρθωση που προκύπτει είναι:

$$MSEP_{adj} = MSEP_{app} - \frac{1}{n_L}\sum_{k=1}^{K}\frac{n_k}{n_L}\sum_{i\notin L_k}(\hat{y}_i - y_i)^2 \qquad (2.6.4)$$

Χρησιμοποιώντας την διόρθωση αυτή (2.6.4) ορίζουμε την διορθωμένη εκτιμήτρια του $MSEP_{CV}$ (*Adjusted K – fold cross – validation*) που είναι:

$$MSEP_{adjCV} = MSEP_{CV} + MSEP_{adj} \qquad (2.6.5)$$

Παρατηρούμε από τον τύπο (2.6.4) ότι αν $n_k = n_L$ τότε $MSEP_{adj} = 0$ που σημαίνει ότι δεν υπάρχει διόρθωση αφού η εξίσωση παλινδρόμησης θα είναι η ίδια. Μπορούμε να διαπιστώσουμε, λοιπόν, ότι κατά την διαδικασία *LOOCV* επειδή $n_k = n_L - 1$ τότε

$$\frac{n_k}{n_L} \to 1$$

Επομένως $MSEP_{adj} \cong 0$ και η σχέση (2.6.5) γίνεται $MSEP_{adjCV} \cong MSEP_{CV}$. Αυτό σημαίνει ότι όταν χρησιμοποιούμε την *LOOCV* οι εκτιμήτριες $MSEP_{CV}$ και $MSEP_{adjCV}$ θα είναι σχεδόν ίσες.

Γενικά η τεχνική *cross – validation* χαρακτηρίζεται ως αργή διαδικασία εξαιτίας των πολλών υπολογισμών που απαιτούνται, χωρίς να σημαίνει όμως ότι είναι και η πιο αργή. Για τα γραμμικά μοντέλα υπάρχει ένας απλός και γρήγορος τρόπος εφαρμογής της *LOOCV*.

Πράγματι, αποδεικνύεται ότι η εκτίμηση του *MSEP* με την χρήση της διαδικασίας *LOOCV* μας οδηγεί στον τύπο:

$$MSEP_{LOOCV} = \frac{1}{n_L}\sum_{i=1}^{n_L}\left(\frac{\hat{y}_i - y_i}{1 - h_i}\right)^2$$

όπου $\hat{y}_i$ είναι οι εκτιμήσεις του μοντέλου που έχει προκύψει από όλα τα δεδομένα και τα h_i είναι τα διαγώνια στοιχεία του πίνακα $H = X(X'X)^{-1}X'$ με τον πίνακα X να είναι ο πίνακας σχεδιασμού που ορίσαμε στην παράγραφο 1.2 (βλ. απλή απόδειξη στους *Seber* και *Lee*, 2003).

Επιπλέον, η εκτιμήτρια $MSEP_{CV}$ σχετίζεται με τα μέτρα *AIC* και *BIC* που ορίσαμε στην παράγραφο (1.2.4). Πιο συγκεκριμένα ισχύει ασυμπτωτικά ότι η ελαχιστοποίηση του *AIC* ισοδυναμεί με την ελαχιστοποίηση του $MSEP_{CV}$ και μάλιστα η ιδιότητα αυτή δεν ισχύει μόνο για γραμμικά μοντέλα *(Stone, 1977)*. Από την ιδιότητα αυτή καταλαβαίνουμε την χρησιμότητα του AIC στην επιλογή μοντέλου

όταν μας ενδιαφέρει να βελτιώσουμε την προβλεπτική του ικανότητα. Επίσης ισχύει ασυμπτωτικά, μόνο για γραμμικά μοντέλα, ότι η ελαχιστοποίηση της τιμής του *BIC* είναι ισοδύναμη με την εκτιμήτρια του $MSEP$ που προκύπτει από την *Leave – k – out cross – validation (Shao 1997)* όπου

$$k = n\left(1 - \frac{1}{\log(n) - 1}\right)$$

Αρκετοί ερευνητές χρησιμοποιούν το *BIC* επειδή θεωρείται πιο «συνεπές» στην επιλογή του «τέλειου» μοντέλου. Κάτι τέτοιο θα ήταν σωστό αν υπήρχε το «τέλειο» μοντέλο και αν το πλήθος των δεδομένων ήταν πολύ μεγάλο. Όμως, στην πραγματικότητα σπάνια υπάρχει το «τέλειο» μοντέλο αλλά και αν υπήρχε, αυτό δεν θα σήμαινε ότι η προβλεπτική του ικανότητα θα ήταν η καλύτερη.

Επίσης είναι αρκετά συνηθισμένο η διαδικασία *LOOCV* να ταυτίζεται με την διαδικασία *Jackknife*. Και στις δύο περιπτώσεις λειτουργούμε ακριβώς με τον ίδιο τρόπο (παραλείπουμε μια παρατήρηση). Όμως η *LOOCV* χρησιμοποιείται για την εκτίμηση του $MSEP$, όπως περιγράψαμε παραπάνω, ενώ κατά την διαδικασία *Jackknife* υπολογίζουμε την μεροληψία των εκτιμητριών που μας ενδιαφέρουν. Τα βήματα που ακολουθούμε είναι τα εξής. Αρχικά υπολογίζουμε την εκτιμήτρια που μας ενδιαφέρει σε κάθε *training set* και στην συνέχεια συγκρίνουμε την μέση τιμή αυτών των εκτιμητριών με την αντίστοιχη εκτιμήτρια που βρίσκουμε, χρησιμοποιώντας ολόκληρο το δείγμα, με σκοπό τον υπολογισμό της μεροληψίας της τελευταίας. Με αντίστοιχο τρόπο μπορούμε να υπολογίσουμε και το τυπικό σφάλμα αλλά και το $MSEP$ της παραπάνω εκτιμήτριας. Η εκτίμηση όμως του $MSEP$ με αυτή την διαδικασία κρίνεται πιο πολύπλοκη από ότι με την *LOOCV (Efron, 1982; Ripley, 1996)* και για αυτό δεν την προτιμάμε. Η διαδικασία *Jackknife* όμως, θεωρείται αρκετά χρήσιμη στις περιπτώσεις που επιθυμούμε να προχωρήσουμε σε στατιστικούς ελέγχους κυρίως για τις εκτιμήτριες των συντελεστών ενός μοντέλου. Εφόσον σύμφωνα με αυτά που είπαμε μπορεί να εκτιμήσει το τυπικό τους σφάλμα. Κάτι τέτοιο θα δούμε στις εφαρμογές του επόμενου κεφαλαίου.

2.6.2 Η τεχνική *Bootstrap*

Υποθέτουμε ότι από το σύνολο L δημιουργούμε R υποσύνολα *(bootstrap samples)* που τα συμβολίζουμε $L_r^* = (x_{r,i}^*, y_{r,i}^*),\ \ r = 1,2, \dots, R$ και $i = 1,2, \dots, n_L$ και έστω f_r^* η εξίσωση παλινδρόμησης που προκύπτει από το δείγμα r. Γενικά, η τεχνική *bootstrap* κρίνεται τις περισσότερες φορές επιτυχημένη όταν για την εκτίμηση χρησιμοποιείται η μεροληψία της εκτιμήτριας. Έτσι αποδεικνύεται ότι η εκτιμήτρια της μεροληψίας (με την τεχνική *bootstrap)* του $MSEP_{app}$ είναι

$$Bias_{app} = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{n_L}\sum_{i=1}^{n_L}(f_r^*(x_i) - y_i)^2 - \frac{1}{n_L}\sum_{i=1}^{n_L}\left(f_r^*\left(x_{r,i}^*\right) - y_{r,i}^*\right)^2\right)$$

Και τελικά το εκτιμημένο $MSEP$ με την τεχνική *bootstrap* είναι

$$MSEP_{boot} = MSEP_{app} + Bias_{app}$$

Υπάρχουν, επίσης, και άλλες εκτιμήτριες του $MSEP$ που προκύπτουν από την μέθοδο *bootstrap*. Μια από αυτές επειδή θυμίζει την *LOOCV* ονομάζεται *leave – one – out bootstrap estimate* ή *bootstrap smoothed cross – validation estimate*, συμβολίζεται $MSEP_{BCV}$ και ορίζεται από τον τύπο

$$MSEP_{BCV} = \frac{1}{n_L}\sum_{i=1}^{n_L}\frac{1}{R_{-i}}\sum_{r,i\notin L_r^*}(f_r^*(x_i) - y_i)^2$$

όπου η τιμή R_{-i} το πλήθος των υποσυνόλων που δεν περιέχουν την παρατήρηση i.

Μια επίσης εκτιμήτρια του $MSEP$ που προκύπτει με την βοήθεια της *bootstrap* είναι η εκτιμήτρια $0{,}632$, η οποία είναι από τις πιο σημαντικές και ορίζεται από τον τύπο

$$MSEP_{0,632} = 0{,}632\ MSEP_{BVC} + (1 - 0{,}632)MSEP_{app}$$

Σε αυτή την παράγραφο λοιπόν, παρουσιάστηκαν αρκετές εκτιμήτριες του $MSEP$. Για την επιλογή της καταλληλότερης θα πρέπει να λαμβάνουμε υπόψη και το υπολογιστικό κόστος, τον χρόνο (βήματα) δηλαδή, που απαιτείται για να υπολογισθεί η κάθε εκτιμήτρια. Ο χρόνος αυτός δεν θεωρείται αμελητέος, ιδίως αν ο όγκος των δεδομένων είναι μεγάλος. Για τον σκοπό αυτό στον παρακάτω πίνακα καταγράφονται το πλήθος των εξισώσεων παλινδρόμησης και το πλήθος των προβλέψεων που απαιτούνται για κάθε μια από τις παραπάνω εκτιμήτριες.

Εκτιμήτριες	Πλήθος εξισώσεων	Πλήθος προβλέψεων
$\boldsymbol{MSEP_{test}}$	1	n_T
$\boldsymbol{MSEP_{app}}$	1	n_L
$\boldsymbol{MSEP_{CV}}$	K	n_L
$\boldsymbol{MSEP_{adjCV}}$	$K+1$	$2n_L$
$\boldsymbol{MSEP_{boot}}$	$R+1$	$(R+1)n_L$
$\boldsymbol{MSEP_{BCV}}$	R	$\approx 0{,}368Rn_L$
$\boldsymbol{MSEP_{0{,}632}}$	$R+1$	$\approx (0{,}368R+1)n_L$

Έχει αποδειχθεί πειραματικά *(Mevik και Cederkvist, 2004)* ότι στις περιπτώσεις των *PCR* και *PLSR*, για την εκτίμηση του $MSEP$ θα έπρεπε να χρησιμοποιείται η εκτιμήτρια $0{,}632$ ή η εκτιμήτρια *LOOCV*. Εάν όμως ο υπολογιστικός χρόνος που απαιτείται για τις παραπάνω εκτιμήτριες μας δημιουργεί πρόβλημα τότε μια καλή επιλογή θα ήταν η $MSEP_{adjCV}$ για $K = 10$ ή 5.

2.7 Μέθοδοι επιλογής του κατάλληλου πλήθους συνιστωσών για την *PLSR*

Σε αυτή την παράγραφο θα περιγράψουμε τους τρόπους που χρησιμοποιούνται για την εκτίμηση του βέλτιστου μοντέλου στην παλινδρόμηση *PLS*. Αν και υπάρχουν αρκετοί τρόποι, καλύτερος θεωρείται εκείνος που για την συγκεκριμένη εκτίμηση χρησιμοποιεί ένα μεγάλο μέρος των δεδομένων το οποίο δεν συμμετέχει στην δημιουργία του μοντέλου, όπως ακριβώς συμβαίνει και για την εκτίμηση του *MSEP*. Κάτι τέτοιο όπως έχουμε πει απαιτεί μεγάλο όγκο δεδομένων που στην πράξη δύσκολα συμβαίνει. Άλλωστε, η *PLSR* έχει αναπτυχθεί για να αντιμετωπίσει τις περιπτώσεις που το σύνολο των δεδομένων είναι μικρό. Παρακάτω θα αναφερθούμε στους τρόπους οι οποίοι μας επιτρέπουν να εκτιμούμε το κατάλληλο πλήθος συνιστωσών και συνεπώς να καταλήγουμε στο βέλτιστο μοντέλο σε μια παλινδρόμηση *PLS*.

A. *Cross – Validation (CV)*

Όπως είδαμε στην προηγούμενη παράγραφο ο λόγος που χρησιμοποιούμε την τεχνική *CV* είναι για την εκτίμηση του *MSEP*. Στην συνέχεια χρησιμοποιώντας κάποιες μεθόδους έχουμε την δυνατότητα να επιλέξουμε και το κατάλληλο μοντέλο. Η *CV* είναι η πιο διαδεδομένη μέθοδος για τον σκοπό αυτό και ένα βασικό της πλεονέκτημα είναι ότι δεν καταναλώνει βαθμούς ελευθερίας. Παρ' όλα αυτά όμως είναι μια μέθοδος επαναληπτικής δειγματοληψιας *(resampling method)*. Αυτό σημαίνει ότι η σωστή εφαρμογή της περιορίζεται σε δεδομένα τα οποία μπορεί να θεωρηθεί ότι ανήκουν σε ένα τυχαίο δείγμα του συνόλου των δεδομένων. Κάτι τέτοιο μπορεί να δημιουργήσει πρόβλημα αν η συλλογή των δεδομένων έγινε ακολουθώντας κάποιο στατιστικό σχεδιασμό. Μια άλλη περίπτωση που η μέθοδος αυτή μπορεί να αποτύχει είναι όταν τα δεδομένα θα πρέπει να ενημερώνονται από κάποιο άλλο πληθυσμό.

Μια επιπλέον αδυναμία που παρατηρείται κατά την εφαρμογή της *CV* είναι η ακόλουθη. Η ιδανική περίπτωση ,όπως ήδη έχουμε αναφέρει, είναι να επιλέξουμε το μοντέλο με το ελάχιστο $RMSEP_{CV}$. Στην πράξη όμως, σπάνια καταλήγουμε σε μια ξεκάθαρη ελάχιστη τιμή του $RMSEP_{CV}$. Κάτι τέτοιο μας αναγκάζει να στηριχθούμε στην οπτική επιθεώρηση διαφόρων γραφημάτων και να χρησιμοποιήσουμε εμπειρικούς κανόνες όπως «επιλέγουμε την συνιστώσα με την πρώτη χαμηλή τιμή του $RMSEP_{CV}$» ή «επιλέγουμε την συνιστώσα από την οποία ξεκινά η

σταθεροποίηση της τιμής του $RMSEP_{CV}$ σε χαμηλά επίπεδα». Το ποιον κανόνα θα εφαρμόσουμε εξαρτάται από τα δεδομένα που έχουμε κάθε φορά. Το συμπέρασμα στο οποίο καταλήγουμε λοιπόν, είναι ότι η χρήση της *CV* για την επιλογή του βέλτιστου μοντέλου είναι μια μέθοδος που από την φύση της είναι υποκειμενική.

Μια προσπάθεια για να γίνει πιο αντικειμενική η παραπάνω διαδικασία είναι, εκτός από την ελαχιστοποίηση του $RMSEP$, να χρησιμοποιούμε το κλάσμα Q_a^2 *(Abdi, 2010)* το οποίο ορίζεται ως εξής:

$$Q_a^2 = 1 - \frac{PRESS_a}{\sum_{j=1}^{m} SS_{\boldsymbol{y}_j, a-1}}\ , \quad j = 1,2,\dots,m \quad a = 1,2,\dots,l$$

όπου

$$SS_{\boldsymbol{y}_j, a-1} = \sum_{i=1}^{n} \left(y_{ij} - \hat{y}_{ij}\right)^2$$

δηλαδή το $SS_{\boldsymbol{y}_j, a-1}$ είναι το άθροισμα τετραγώνων των υπολοίπων της μεταβλητής $\boldsymbol{y_j}$ που αντιστοιχεί στην προηγούμενη συνιστώσα. Η τιμή του Q_a^2 αντιστοιχεί στην συνιστώσα α κάτι που σημαίνει ότι θα πρέπει να υπολογίζουμε το συγκεκριμένο κλάσμα για κάθε συνιστώσα ξεχωριστά. Μια συνιστώσα θα θεωρείται σημαντική αν ικανοποιείται η σχέση $Q_a^2 > 0.0975$ ενώ ένα εναλλακτικό κριτήριο είναι $Q_a^2 > h$ με $h = 0$ για $n > 100$ και $h = 0.05$ για $n \leq 100$.

Β. Επιλογή του βέλτιστου μοντέλου με την χρήση των ιδιοτιμών του $\boldsymbol{X'X}$.

Την μέθοδο αυτή την περιγράψαμε στο προηγούμενο κεφάλαιο ως μέθοδο επιλογής των κύριων συνιστωσών για την *PCR*. Όπου κάθε ιδιοτιμή αντιστοιχεί σε μια συνιστώσα και δείχνει σε ποιό ποσοστό η συνιστώσα αυτή επηρεάζει το μοντέλο. Για την επιλογή του κατάλληλου μοντέλου στην *PLSR* όμως, δεν χρησιμοποιείται όσο η *CV*, αν και αποτελεί την βάση για την δημιουργία τρόπων επιλογής συνιστωσών στην *PCR*.

Γ. Τυχαιοποιημένος Έλεγχος (*Randomization test)*

Έχει καθιερωθεί, η επιλογή συνιστωσών στην *PLSR* να βασίζεται αποκλειστικά στην τεχνική *CV*. Όπως είπαμε όμως, το βασικό μειονέκτημα της τεχνικής αυτής είναι ότι τις περισσότερες φορές η επιλογή του βέλτιστου μοντέλου γίνεται με υποκειμενικό τρόπο. Αυτό συμβαίνει επειδή το *testing set* που

χρησιμοποιεί είναι αρκετά μικρό και όχι ανεξάρτητο από τα υπόλοιπα δεδομένα, με αποτέλεσμα να οδηγούμαστε σε μοντέλα υπερπροσαρμοσμένα. Με άλλα λόγια, το ελάχιστο *MSEP* είναι πιθανό να εμφανίζεται σε μεγαλύτερης ή μικρότερης τάξης μοντέλο, από εκείνο που θα μας έδινε ένα ανεξάρτητο και μεγαλύτερο *testing set*. Η ανάγκη ύπαρξης ενός πιο αντικειμενικού τρόπου επιλογής του βέλτιστου μοντέλου μας οδηγεί στη χρήση του τυχαιοποιημένου στατιστικού ελέγχου, γνωστού ως *randomization test*.

Στην παλινδρόμηση *PLS* το *randomization test* χρησιμοποιείται με δύο τρόπους. Στον ένα συγκρίνονται ως προς την σημαντικότητα τους δύο μοντέλα με διαφορετικό πλήθος συνιστωσών *(Van Der Voet, 1994)* και στον άλλο γίνεται έλεγχος σημαντικότητας καθεμιάς από τις συνιστώσες ξεχωριστά *(Wiklund et al., 2007)*. Στο παράδειγμα των βαθμών των μαθητών αλλά και στις εφαρμογές του τρίτου κεφαλαίου χρησιμοποιείται ο δεύτερος τρόπος από τους δύο παραπάνω. Για το λόγο αυτό θα αναφερθούμε αναλυτικότερα στον συγκεκριμένο τρόπο.

Το βασικό πλεονέκτημα του τρόπου αυτού είναι ότι χρησιμοποιεί το σύνολο των δεδομένων, με αποτέλεσμα να αποφεύγονται μη ρεαλιστικές προϋποθέσεις σαν αυτές της *CV*, όπου για παράδειγμα αποκλείονται δεδομένα, αφού διαχωρίζεται το σύνολο του δείγματος σε *training set* και *testing set*. Στον συγκεκριμένο έλεγχο η μόνη προϋπόθεση είναι ότι τα δεδομένα κατανέμονται τυχαία κάτω από την μηδενική υπόθεση.

Σύμφωνα με την διαδικασία που ακολουθείται χρησιμοποιούνται όλες οι δυνατές μεταθέσεις ($n!$) των στοιχείων του διανύσματος $\boldsymbol{y}$ *(permutations)*, σε κάθε μια από τις οποίες προσαρμόζεται ένα μοντέλο *PLSR* ενώ ο πίνακας X των ανεξάρτητων μεταβλητών παραμένει αμετάβλητος. Για τον λόγο αυτό ο συγκεκριμένος έλεγχος ονομάζεται και *permutation test*. Στα συγκεκριμένα μοντέλα *PLSR* που δημιουργούνται αντικατοπτρίζεται η απουσία μιας πραγματικής σχέσης μεταξύ των $\boldsymbol{y}$ και X (μη σημαντικές συνιστώσες). Αυτό αποτελεί και την μηδενική υπόθεση του συγκεκριμένου ελέγχου. Με αυτό τον τρόπο μπορούμε να διερευνήσουμε την σημαντικότητα κάθε συνιστώσας ξεχωριστά χωρίς την ύπαρξη συγκεκριμένων προϋποθέσεων για τα δεδομένα.

Κατά την εφαρμογή του συγκεκριμένου ελέγχου θα πρέπει οι σημαντικές συνιστώσες να ακολουθούνται από τις μη σημαντικές. Στην πράξη όμως δεν συμβαίνει πάντα κάτι τέτοιο. Στις περιπτώσεις αυτές ο έλεγχος μας υποδεικνύει ότι τα

δεδομένα χρειάζονται κάποιο μετασχηματισμό ή κάποια διόρθωση. Γενικά αποδεικνύεται ότι το *randomization test* αποτελεί ένα πολύ χρήσιμο στατιστικό εργαλείο, αλλά θα πρέπει να χρησιμοποιείται σε συνδυασμό με άλλες μεθόδους επιλογής συνιστωσών *(cross – validation)*.

2.8 Προσαρμογή μοντέλου *PLSR* στο παράδειγμα των μαθητών

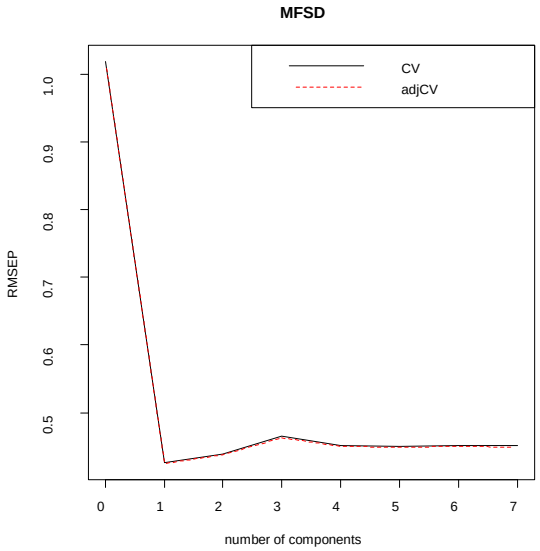
Προσπαθώντας να εντοπίσουμε το βέλτιστο μοντέλο PLSR στο παράδειγμα των μαθητών υπολογίζουμε αρκετές τιμές που αναφέραμε στις προηγούμενες παραγράφους η οποίες παρουσιάζονται στον Πίνακα Α που ακολουθεί για κάθε συνιστώσα ξεχωριστά. Επίσης στην τελευταία στήλη του πίνακα γίνεται εκτίμηση του βέλτιστου μοντέλου που προκύπτει από την κάθε μια τιμή. Πιο συγκεκριμένα υπολογίσαμε, τα ποσοστά εξήγησης των διασπορών των μεταβλητών X και Y, εκτίμηση του RMSEP με τις μεθόδους $LOOCV$, *K – fold cross – validation* για *K=10*, και *bootstrap smoothed cross – validation estimate*. Ακόμα έχουμε υπολογίσει, το κλάσμα Q_a^2 που προκύπτει από την $LOOCV$ και έχει γίνει το *randomization test* για τα $RMSEP_{LOOCV}$ και $RMSEP_{BCV}$. Επίσης χρησιμοποιούμε τα Σχήματα 1 έως και 4 για επιπλέον βοήθεια. Στα Σχήματα 1 και 2 παριστάνονται γραφικά οι τιμές των $RMSEP_{LOOCV}$ και $RMSEP_{CV}$ $(K = 10)$ αντίστοιχα. Στο Σχήμα 3 παριστάνονται οι εκτιμημένοι συντελεστές όλων των μοντέλων και στο Σχήμα 4 συγκρίνουμε γραφικά τις εκτιμήσεις των τιμών του Y για τα μοντέλα με μια και δύο συνιστώσες με τις πραγματικές τιμές του δείγματος για κάθε μια παρατήρηση ξεχωριστά. Επίσης στο Σχήμα 5 παριστάνονται οι τιμές των *loadings* κάθε συνιστώσας σε σχέση με τις μεταβλητές, στο Σχήμα 6 παριστάνονται οι τιμές των *scores* μιας συνιστώσας σε σχέση με τις αντίστοιχες τιμές μιας άλλης και στο Σχήμα 7 οι διασπορές των συντελεστών για τα μοντέλα με μια και δύο συνιστώσες (μέθοδος *jackknife*). Επιπλέον, επειδή το πλήθος των παρατηρήσεων είναι αρκετά μικρό επιλέξαμε να μην το χωρίσουμε σε *training set* και *testing set*.

Σχολιάζοντας τα αποτελέσματα του Πίνακα Α μπορούμε να πούμε ότι η εξήγηση της διασποράς του X μπορεί να κριθεί ικανοποιητική από την πρώτη συνιστώσα ενώ ο συντελεστής προσδιορισμού είναι σχεδόν σταθερός και σε υψηλά επίπεδα. Η εκτίμηση του *RMSEP* έχει γίνει με τέσσερεις τρόπους και σε κάθε περίπτωση η ελάχιστη τιμή του αντιστοιχεί στο μοντέλο με τη μία συνιστώσα (Σχήματα 1, 2). Οι τιμές όμως του $RMSEP_{CV}(K = 10)$ και $RMSEP_{adjCV}(K = 10)$ για τα μοντέλα με μια και δύο συνιστώσες είναι αρκετά κοντά. Το μοντέλο με μια συνιστώσα υποδεικνύουν η τιμή του κλάσματος Q_a^2 καθώς και το *randomization test* που προκύπτει από την μέθοδο *Bootstrap CV*. Ενώ από το αντίστοιχο *test* για την $LOOCV$ δεν προκύπτει ξεκάθαρο αποτέλεσμα.

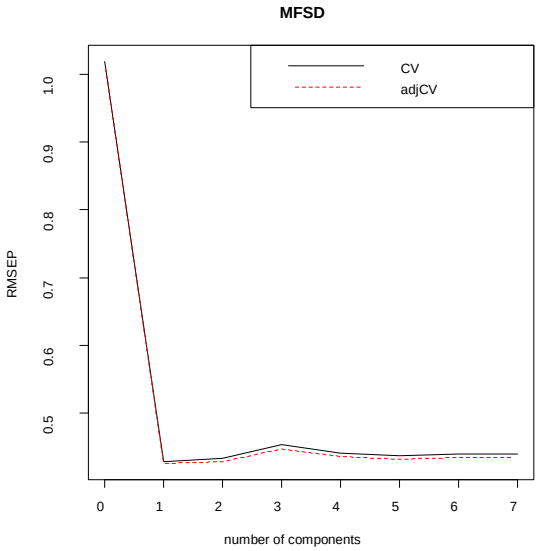
Από το γράφημα των συντελεστών των επτά μοντέλων (Σχήμα 3) παρατηρούμε ότι οι γραμμές σχεδόν ταυτίζονται από το μοντέλο με τις δύο συνιστώσες. Δηλαδή υποδεικνύεται ως βέλτιστο το μοντέλο με δύο συνιστώσες. Στο Σχήμα 4 όμως παρατηρούμε ότι οι εκτιμήσεις της εξαρτημένης μεταβλητής που προκύπτουν από το μοντέλο με μια συνιστώσα είναι πάρα πολύ κοντά με τις αντίστοιχες τιμές των δεδομένων. Από το Σχήμα 5 προκύπτει ότι η πρώτη συνιστώσα επηρεάζει θετικά στον ίδιο βαθμό όλες τις μεταβλητές ενώ η δεύτερη συνιστώσα επηρεάζει είτε θετικά είτε αρνητικά σε υψηλό βαθμό μόνο τις τρεις τελευταίες μεταβλητές. Στο Σχήμα 6 τα σημεία που αντιστοιχούν στις τιμές των scores της πρώτης και δεύτερης συνιστώσας κατανέμονται τυχαία χωρίς να σχηματίζονται ομάδες σημείων. Τέλος στο Σχήμα 7 είναι φανερό ότι οι διασπορές των συντελεστών του μοντέλου με μια συνιστώσα είναι πολύ μικρές σε αντίθεση με το μοντέλο των δύο συνιστωσών. Μπορούμε λοιπόν να καταλήξουμε στο συμπέρασμα ότι το βέλτιστο μοντέλο προσεγγίζεται σε ικανοποιητικό βαθμό από εκείνο με τη μια συνιστώσα. Προσαρμόζοντας λοιπόν στα δεδομένα το συγκεκριμένο μοντέλο έχουμε τα αποτελέσματα του Πίνακα Β όπου με την βοήθεια του *Jackknife Test* όλες οι μεταβλητές εμφανίζονται σημαντικές.

ΠΙΝΑΚΑΣ Α

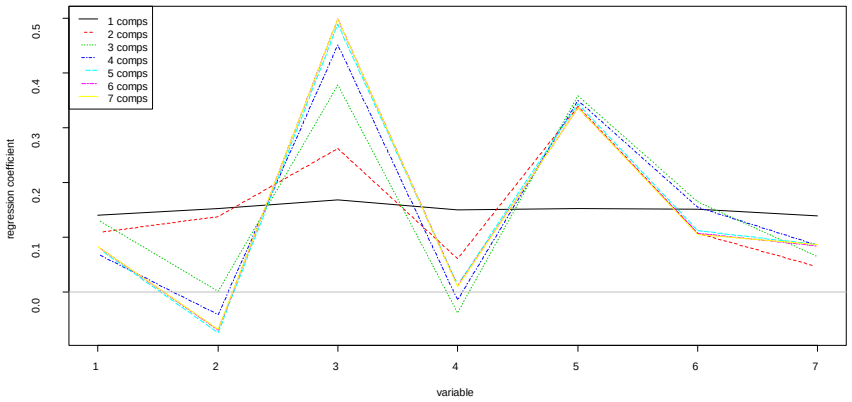
	Comps 1	Comps 2	Comps 3	Comps 4	Comps 5	Comps 6	Comps 7	Επιλογή Συνιστωσών
X Variance explained	75.42	81.98	85.41	88.92	91.93	96.87	100.00	1 – 4
Y Variance explained (R^2)	84.34	86.83	87.63	87.81	87.87	87.88	87.88	1 – 7
$RMSEP_{LOOCV}$	0.4260	0.4390	0.4658	0.4519	0.4507	0.4520	0.4516	1
$RMSEP_{CV}$ ($K = 10$)	0.4283	0.4330	0.4533	0.4415	0.4378	0.4403	0.4397	1 – 2
$RMSEP_{adjCV}$ ($K = 10$)	0.4263	0.4292	0.4470	0.4363	0.4327	0.4350	0.4345	1 – 2
$RMSEP_{BCV}$	0.4419	0.4731	0.5621	0.5445	0.5576	0.5627	0.5174	1
Q_a^2 $(LOOCV)$	0,8139	-0,2766	-0,7163	-0,7024	-0,7278	-0,7459	-0,7446	1
LOOCV p –value (rand test)	NA	0.846	0.999	0.000	0.378	0.914	0.064	1 – 4
Bootstrap p –value (rand test)	NA	0.793	0.986	0.981	0.501	0.836	0.849	1



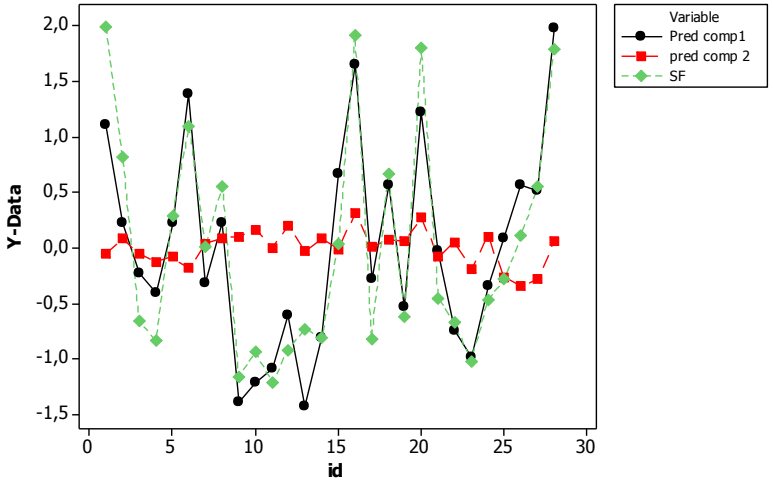
Σχήμα 1



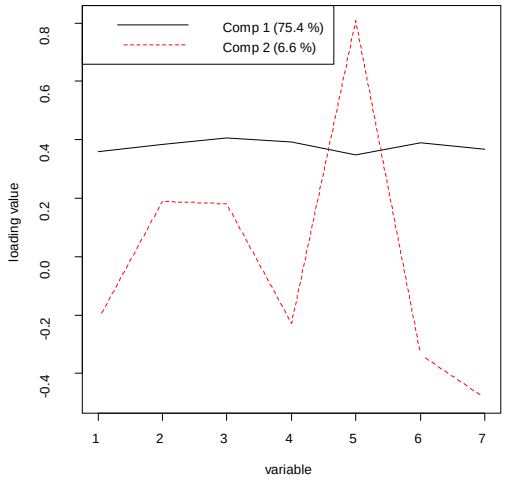
Σχήμα 2



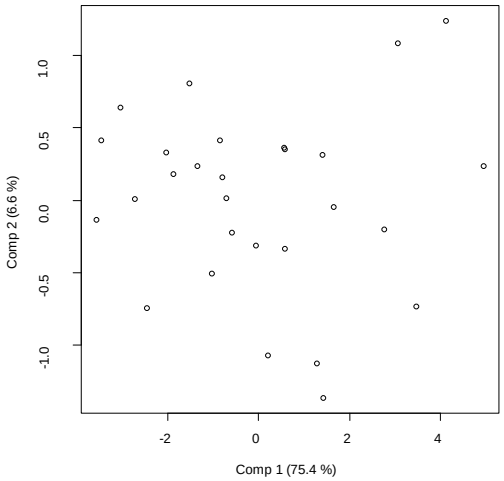
Σχήμα 3



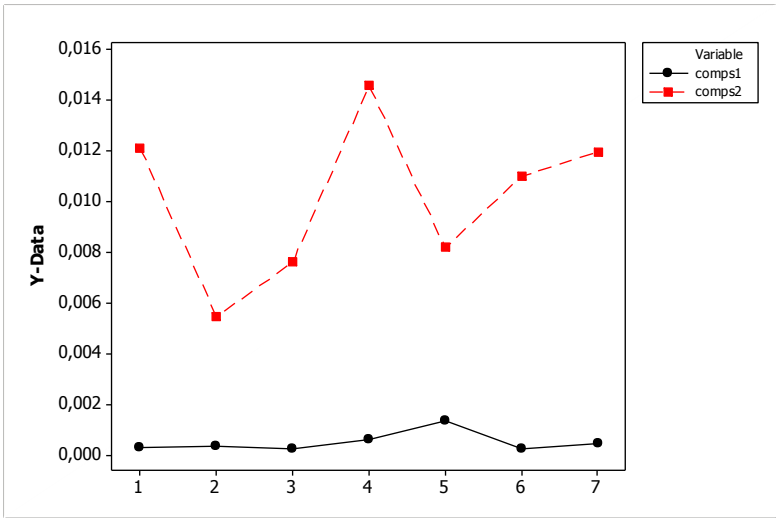
Σχήμα 4



Σχήμα 5



Σχήμα 6



Σχήμα 7

ΠΙΝΑΚΑΣ Β

	D1	D2	D3	T1	T2	T3	T4
Συντελεστές	0.140934	0.152491	0.168896	0.150168	0.152917	0.151773	0.139637
p - values	2.9e-08***	3.8e-08***	2.5e-10***	2.54e-06***	0.0003***	2.47e-09***	6.9e-07***

2.9 Σύγκριση της *PLSR* με την Παλινδρόμηση Ελαχίστων Τετραγώνων *(OLSR)*, την Παλινδρόμηση *RIDGE*, και την Παλινδρόμηση Κύριων Συνιστωσών *(PCR)*

Σε αυτή την παράγραφο θα αναφερθούμε και στις τέσσερεις μεθόδους παλινδρόμησης, αν και έχουν περιγραφεί αναλυτικά, έτσι ώστε από την σύγκριση αυτή να αναδειχθεί το πλεονέκτημα της PLSR έναντι των υπολοίπων.

Όπως έχουμε ήδη αναφέρει, στην προσπάθεια μας να προσαρμόσουμε σε δεδομένα ένα μοντέλο *OLSR*, είναι πολύ πιθανό να βρεθούμε αντιμέτωποι με το φαινόμενο της πολυσυγγραμμικότητας. Η εμφάνιση του συγκεκριμένου φαινομένου καθιστά την προσαρμογή ενός τέτοιου μοντέλου προβληματική. Αφού, οι εκτιμήτριες των συντελεστών παλινδρόμησης εμφανίζονται στατιστικά μη σημαντικές, οι εκτιμήτριες των εξαρτημένων μεταβλητών απέχουν πολύ από την πραγματικότητα αν και ο συντελεστής προσδιορισμού μπορεί να κυμαίνεται σε υψηλά επίπεδα. Για την αντιμετώπιση του συγκεκριμένου φαινομένου έχουν αναπτυχθεί διάφορες μέθοδοι, με σημαντικότερες τις *RR, PCR* και *PLSR* που ήδη έχουμε περιγράψει.

Η μέθοδος *PCR* αντιμετωπίζει την πολυσυγγραμμικότητα με την ανάλυση των ανεξάρτητων μεταβλητών σε κύριες συνιστώσες. Σκοπός της μεθόδου είναι να δημιουργήσει την καλύτερη δυνατή παραλλαγή του πίνακα X που θα βελτιστοποιεί την προβλεπτική ικανότητα του μοντέλου.

Η αντιμετώπιση της πολυσυγγραμμικότητας από την μέθοδο *RR* πραγματοποιείται με την προσθήκη μιας σταθεράς θ στα διαγώνια στοιχεία του πίνακα $X'X$. Εξαιτίας της πολυσυγγραμμικότητας ο πίνακας $X'X$ είναι σχεδόν κανονικός. Η προσθήκη της παραπάνω σταθεράς μετατρέπει τον $X'X$ σε ένα, καθαρά, μη κανονικό πίνακα.

Η *PLSR*, όπως έχουμε αναφέρει, αναπτύχθηκε από τον *Herman Wald* για την δημιουργία μοντέλων πρόβλεψης όταν οι επεξηγηματικές μεταβλητές είναι πολλές και ισχυρά συγγραμμικές. Μπορεί να εφαρμοσθεί για οσοδήποτε πλήθος επεξηγηματικών μεταβλητών, ακόμα και όταν είναι περισσότερες από το πλήθος των παρατηρήσεων. Αν και η μέθοδος αυτή χρησιμοποιείται αρκετά από τους χημικούς ερευνητές, παραμένει σε μεγάλο βαθμό άγνωστη στους κύκλους της Στατιστικής. Η *PLSR* εμφανίζει ορισμένες ομοιότητες με την *PCR*. Η βασικότερη είναι ότι και οι δύο μέθοδοι στοχεύουν στην δημιουργία μεταβλητών (συνιστωσών) οι οποίες θα επεξηγούν όσο το δυνατό καλύτερα τις εξαρτημένες μεταβλητές. Η βασική τους

διαφορά, όπως έχουμε αναφέρει, είναι ότι ενώ η *PCR* δημιουργεί αυτές της νέες μεταβλητές χρησιμοποιώντας μόνο τον πίνακα *X*, η *PLSR* χρησιμοποιεί και τους δύο πίνακες *X* και *Y* για την δημιουργία μεταβλητών που θα αναλάβουν τον ρόλο επεξηγηματικών μεταβλητών.

Οι μέθοδοι που χρησιμοποιούνται περισσότερο είναι η *PCR* και η *RR*, οι οποίες απαιτούν αρκετούς υπολογισμούς, ιδιαίτερα όταν το πλήθος των μεταβλητών είναι αρκετά μεγάλο *(Helland, 1988)*. Πειραματικά *(Yeniay και Goctas, 2002)* έχει δειχθεί (με την χρήση πραγματικών δεδομένων) ότι όταν θεωρείται σημαντική η καλή προσαρμογή ενός μοντέλου τότε η *OLSR* και η *RR* φαίνεται να είναι οι καλύτερες μέθοδοι. Η μέθοδος *PLSR* όμως, αντιμετωπίζει το πρόβλημα της πολυσυγγραμμικότητας εμφανίζοντας ταυτόχρονα καλύτερη προβλεπτική ικανότητα με την χρήση λιγότερων συνιστωσών από την *PCR*. Αυτό αποτελεί και το βασικότερο πλεονέκτημα της *PLSR* έναντι των άλλων. Αφού ένα μοντέλο με λιγότερες μεταβλητές είναι πιο εύκολο να ερμηνευθεί. Ένα ακόμα σημαντικό πλεονέκτημα της *PLSR* είναι ότι ο μοναδικός αυτός τρόπος επιλογής μεταβλητών απαιτεί λιγότερους υπολογισμούς απ’ ότι η *PCR* και η *RR*.

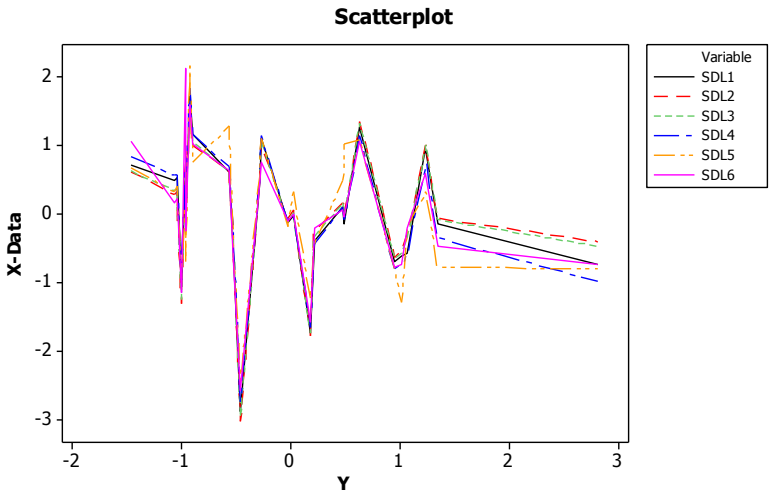
ΚΕΦΑΛΑΙΟ 3

ΕΦΑΡΜΟΓΗ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ *PLS*

3.1 1η Εφαρμογή (*Fearn, 1983)*

Στην εφαρμογή αυτή θα χρησιμοποιήσουμε δύο ομάδες δεδομένων. Η πρώτη ομάδα θα χρησιμοποιηθεί ως *training set*, για την προσαρμογή του μοντέλου και η δεύτερη ομάδα ως *testing set*. Όταν έγινε η μελέτη των παραπάνω δεδομένων *(T. Fearn, 1983)* οι ομάδες αυτές χρησιμοποιήθηκαν ακριβώς με τον αντίθετο τρόπο. Σκοπός της μελέτης αυτής ήταν να παρουσιαστεί ένα παράδειγμα που να αποδεικνύει ότι αντιμετωπίζοντας το πρόβλημα της πολυσυγγραμμικότητας με την χρήση της παλινδρόμησης *Ridge (RR)* ενδέχεται να καταλήξουμε σε ένα μοντέλο με πολύ χειρότερη προβλεπτική ικανότητα από το αντίστοιχο της πολλαπλής γραμμικής παλινδρόμησης που προσαρμόστηκε επίσης στα δεδομένα, με την μέθοδο των ελαχίστων τετραγώνων *(OLSR)*. Τα δύο αυτά μοντέλα συγκρίθηκαν ως προς το τυπικό σφάλμα των υπολοίπων τους και ως προς την τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος των προβλέψεων *(RMSEP)* που εκτιμήθηκε από το *testing set*. Παρατηρήθηκε ότι ενώ το τυπικό σφάλμα των υπολοίπων δεν παρουσίαζε σημαντική μεταβολή, η τιμή του *RMSEP* μεταβάλλονταν αισθητά προς το χειρότερο, για τιμές του k πολύ κοντά στο 0 που ουσιαστικά δεν έλυναν το πρόβλημα της πολυσυγγραμμικότητας. Αυτό σημαίνει ότι αντιμετωπίζοντας την πολυσυγγραμμικότητα με την κατάλληλη επιλογή της τιμής της παραμέτρου k, το μοντέλο μας ενδέχεται να χάσει ένα σημαντικό μέρος της προβλεπτικής του αξίας.

Ο λόγος που συμβαίνει αυτό είναι ότι αν αναλύσουμε τα δεδομένα σε κύριες συνιστώσες θα δούμε ότι η πρώτη συνιστώσα των ανεξάρτητων μεταβλητών αποτελείται από τιμές σχεδόν ίσες μεταξύ τους. Αυτό οφείλεται στο ότι οι τιμές των ανεξάρτητων μεταβλητών αυξάνονται και μειώνονται μαζί, από παρατήρηση σε παρατήρηση. Το ίδιο συμβαίνει αν χρησιμοποιήσουμε τα δεδομένα με τον αντίστροφο τρόπο, όπως φαίνεται στο Σχήμα 1 και στις τιμές της πρώτης συνιστώσας $(-0.415, -0.414, -0.415, -0.414, -0.389, -0.402)$.



Σχήμα 1

Χρησιμοποιώντας λοιπόν το συγκεκριμένο παράδειγμα αντίστροφα, θα ελέγξουμε την προβλεπτική αξία ενός μοντέλου *RR* συγκρίνοντας το με μοντέλα *OLSR*, *PCR* και *PLSR*, αντιμετωπίζοντας ταυτόχρονα και την πολυσυγγραμμικότητα. Για μεγαλύτερη ευκολία οι τιμές του *RMSEP*, του τυπικού σφάλματος και του συντελεστή προσδιορισμού όλων των μοντέλων που θα συγκρίνουμε υπάρχουν και στον Πίνακα VII στο τέλος της εφαρμογής. Αρχικά τυποποιούμε όλες τις μεταβλητές. Προσαρμόζοντας ένα μοντέλο *OLSR* στα δεδομένα (Πίνακας I) παρατηρούμε ότι το πρόβλημα της πολυσυγγραμμικότητας είναι εμφανές. Επίσης οι τιμές που μας ενδιαφέρουν είναι το *Residual standard error: 0.1075* και το *RMSEP = 0.486596*.

ΠΙΝΑΚΑΣ I (OLSR)

	L1	L2	L3	L4	L5	L6
Συντελεστές	-3.597	-2.6387	9.861	-2.9212	-0.2504	-0.7587
p - values	0.0132*	0.0067 **	0.0000 ***	0.00248 **	0.0804	0.00014 ***
VIF	3753.5	1630.6	3962.4	1521.3	39.8	55.7

Residual standard error: 0.1075 on 19 degrees of freedom

Multiple R-squared: 0.9912, Adjusted R-squared: 0.9885

F-statistic: 357.6 on 6 and 19 DF, p-value: < 2.2e-16

RMSEP =0.486596

Προσαρμόζοντας ένα μοντέλο *RR* για μια τιμή του k πολύ κοντά στο 0 ($k = 0.001$) έχουμε αποτελέσματα του Πίνακα II.

ΠΙΝΑΚΑΣ ΙΙ (RR, $k = 0.001$)

	L1	L2	L3	L4	L5	L6
Συντελεστές	-0.828073	1.753	3.19716	-3.58666	0.0120387	-0.822275
p - values	0.00529 **	9.76e-09 ***	< 2e-16 ***	< 2e-16 ***	0.88125	8.06e-14 ***
VIF	84.6128	89.6405	83.5215	64,8991	6.21215	11.6265

Residual standard error: 0,171046 on 19 degrees of freedom

Multiple R-squared: 0.9502, Adjusted R-squared: 0.9345

Ridge parameter: 0.001

RMSEP = 0.7479373

Παρατηρούμε ότι η τιμή του *RMSEP* και το τυπικό σφάλμα των υπολοίπων αυξήθηκαν σημαντικά. Κάτι τέτοιο σημαίνει ότι η προβλεπτική αξία του μοντέλου μειώθηκε σε σχέση με το αντίστοιχο της *(OLSR)* χωρίς ακόμα να έχει αντιμετωπιστεί η πολυσυγγραμμικότητα. Αν προσπαθήσουμε να επιλέξουμε την κατάλληλη τιμή του k ώστε να εξαλείψουμε την πολυσυγγραμμικότητα, τότε αυτή θα είναι $k = 0.035$ (περίπου). Προσαρμόζοντας ένα μοντέλο για αυτή την τιμή, όπως μπορούμε να δούμε (Πίνακας ΙΙΙ) η πολυσυγγραμμικότητα αντιμετωπίστηκε αλλά η προβλεπτική αξία του μοντέλου μειώθηκε. Επίσης, το συγκεκριμένο μοντέλο φαίνεται πως δεν προσαρμόζεται καλά στα δεδομένα αφού ο συντελεστής προσδιορισμού είναι πολύ μικρός. Γίνεται εύκολα αντιληπτό λοιπόν, ότι κανένα μοντέλο *RR* δεν είναι κατάλληλο για να προσαρμοστεί στα δεδομένα μας.

ΠΙΝΑΚΑΣ ΙΙΙ (RR, $k = 0.035$)

	L1	L2	L3	L4	L5	L6
Συντελεστές	-0.383356	1.43532	1.0688	-1.60087	-0.148436	-0.645566
p - values	0.00341 **	2.22e-16 ***	2.22e-16 ***	6.44e-15 ***	0.46221	0.00564 **
VIF	1.51418	2.71471	1.49854	3.72768	3.60095	4.80401

Residual standard error: 0.5772 on 19 degrees of freedom

Multiple R-squared: 0.5245, Adjusted R-squared: 0.3743

Ridge parameter: 0.035

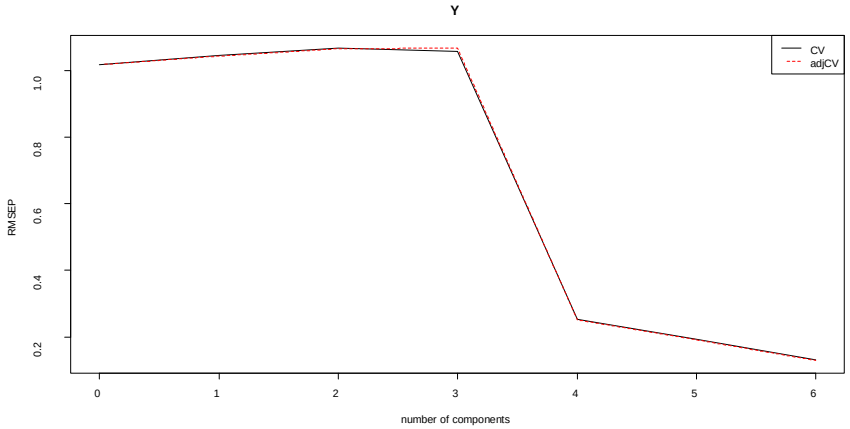
RMSEP = 0.869019

Θα ερευνήσουμε τώρα, αν όλα αυτά τα προβλήματα αντιμετωπίζονται προσαρμόζοντας στα δεδομένα ένα μοντέλο *PCR* αρχικά, και στην συνέχεια ένα μοντέλο *PLSR*.

ΠΙΝΑΚΑΣ IV (PCR)

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	Επιλογή Συνιστωσών
X Variance expl. %	95.380	98.106	99.46	99.99	100.00	100.00	1
Y Variance expl. %	5.974	7.597	18.08	97.16	98.26	99.12	4 – 6
$RMSEP_{LOOCV}$	1.048	1.069	1.058	0.2522	0.1930	0.1302	6
$RMSEP$ (testing set)	1.1123	1.0910	1.0810	0.7536	0.7190	0.7607	5

Residual standard error: **0.1511** on 19 degrees of freedom (5 comps)



Σχήμα 2

Εφαρμόζοντας την μέθοδο της PCR (Πίνακας IV), αρχικά παρατηρούμε ότι η ιδιοτιμή που αντιστοιχεί στην πρώτη συνιστώσα αντιστοιχεί στο 95.38% του συνόλου των ιδιοτιμών. Αυτό μας οδηγεί στην επιλογή μιας συνιστώσας για την προσαρμογή του κατάλληλου μοντέλου. Από τις τιμές όμως, του *RMSEP*, του *RMSEP* του *testing set*, του συντελεστή προσδιορισμού της εξαρτημένης μεταβλητής αλλά και από το παραπάνω γράφημα (Σχήμα 2) καταλαβαίνουμε ότι μια συνιστώσα δεν είναι αρκετή και θα πρέπει να επιλέξουμε μεταξύ τεσσάρων, πέντε ή έξι συνιστωσών.

Για πέντε συνιστώσες η τιμή του *RMSEP* του *testing set* είναι η πιο χαμηλή. Επίσης η τιμή αυτή είναι μικρότερη από τις αντίστοιχες των μοντέλων *OLSR* και *RR* ($k = 0.001$). Αυτό που θα πρέπει να μας προβληματίσει στην συγκεκριμένη περίπτωση είναι η τιμή του τυπικού σφάλματος των υπολοίπων που ισούται με 0.1511 αν και βελτιώνεται σε σχέση με την αντίστοιχη του μοντέλου *RR* (0,171046),η οποία εξακολουθεί να απέχει αρκετά από την αρχική που ήταν 0.1075. Σε ανάλογους

προβληματισμούς καταλήγουμε αν επιλέξουμε τέσσερεις ή έξι συνιστώσες. Παρόλα αυτά όμως, ένα μοντέλο *PCR* με πέντε συνιστώσες θεωρείται αυτή τη στιγμή ως η καλύτερη επιλογή.

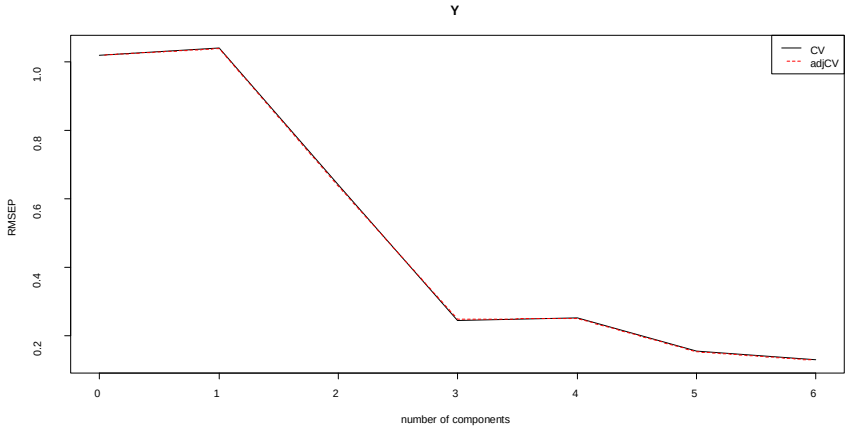
Συνεχίζοντας, προσαρμόζουμε στα δεδομένα ένα μοντέλο *PLSR*. Συγκρίνοντας τα αποτελέσματα του Πίνακα V με τα αντίστοιχα του μοντέλου *PCR* μπορούμε να πούμε ότι τα ποσοστά των ιδιοτιμών δεν παρουσιάζουν κάποια αξιόλογη μεταβολή. Ο συντελεστής προσδιορισμού βελτιώνεται αισθητά στις δύο συνιστώσες έναντι των τεσσάρων που ήταν πριν. Από την εκτίμηση του *RMSEP* και με τη βοήθεια του Σχήματος 3 προκύπτει ότι η τιμή του αρχίζει να σταθεροποιείται σε χαμηλές τιμές από τις τρεις συνιστώσες, ενώ στο μοντέλο *PCR* αυτό συνέβη από τις τέσσερεις συνιστώσες. Ωστόσο, τη χαμηλότερη τιμή του *RMSEP* και στις δύο μεθόδους την παρατηρούμε στις έξι συνιστώσες. Θυμίζουμε ότι ένα μοντέλο *PLSR* με έξι συνιστώσες (μέγιστος αριθμός συνιστωσών) και ένα μοντέλο *PCR* επίσης με έξι συνιστώσες, είναι ακριβώς τα ίδια. Επίσης η τιμή του *RMSEP* στις πέντε συνιστώσες είναι αρκετά βελτιωμένη σε σχέση με την αντίστοιχη τιμή στο μοντέλο *PCR* (0.156 και 0.1930 αντίστοιχα). Επίσης το *RMSEP* που προκύπτει από το *testıng set* δεν εμφανίζει κάποια ουσιαστική μεταβολή (ελάχιστά βελτιωμένο στο μοντέλο της *PCR* (0.7190) έναντι του μοντέλου *PLSR* (0.7288)).

Το τυπικό σφάλμα των υπολοίπων σε ένα μοντέλο *PLSR* με πέντε συνιστώσες εμφανίζεται εντυπωσιακά βελτιωμένο σε σχέση με όλα τα προηγούμενα και ισούται με 0.1267. Μπορούμε να πούμε λοιπόν ότι επιλέγοντας ένα μοντέλο *PLSR* με πέντε συνιστώσες η τιμή του *RMSEP* βελτιώνεται συγκριτικά με τα μοντέλα της παλινδρόμησης *Ridge* ενώ το τυπικό σφάλμα των υπολοίπων και ο συντελεστής προσδιορισμού πλησιάζει αρκετά τις τιμές που προκύπτουν από το αντίστοιχο μοντέλο της μεθόδου *OLS* σε αντίθεση με όλες τις προηγούμενες μεθόδους παλινδρόμησης. Φαίνεται λοιπόν ότι το μοντέλο *PLSR* με πέντε συνιστώσες προσαρμόζεται καλύτερα στα δεδομένα από τα υπόλοιπα ενώ παρουσιάζει και την καλύτερη προβλεπτική ικανότητα. Εναλλακτικά βέβαια, θα μπορούσαμε να χρησιμοποιήσουμε και το μοντέλο *PCR* με πέντε συνιστώσες το οποίο είναι ελαφρώς χειρότερο από το αντίστοιχο της *PLSR*.

ΠΙΝΑΚΑΣ V (PLSR)

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	Επιλογή Συνιστωσών
X Variance expl. %	95.287	96.68	98.61	99.99	100.00	100.00	1 – 6
Y Variance expl. %	7.312	75.52	93.32	97.21	98.78	99.12	3 – 6
$RMSEP_{LOOCV}$	1.041	0.6413	0.2461	0.2528	0.156	0.1302	6
$RMSEP$ (testing set)	1.1241	0.8753	0.7768	0.7529	0.7288	0.7607	5

Residual standard error: **0.1267** on 19 degrees of freedom (5 comps)



Σχήμα 3

Στον Πίνακα VI εμφανίζονται οι εκτιμήτριες των συντελεστών των μοντέλων με πέντε συνιστώσες για τις μεθόδους *PCR* και *PLSR*, όπου βλέπουμε ότι το μοντέλο με την υψηλότερη προβλεπτική αξία (*PLSR* μοντέλο) δίνει τις μεταβλητές L3 και L4 ως στατιστικά σημαντικές, ενώ το αντίστοιχο της *PCR* δίνει επιπλέον και τις L1, L5.

ΠΙΝΑΚΑΣ VI (coefficients)

	L1	L2	L3	L4	L5	L6
Συντελεστές (PCR model, comps=5)	1.59491	-0.37053	4.44450	-5.79482	0.25018	-0.37764
p - values	0.018551 *	0.792414	0.001969 **	1.531e-08 ***	0.033333 *	0.336534
Συντελεστές (PLSR model, comps=5)	-0.036860	-1.801988	6.969237	-5.054480	0.090523	-0.436902
p - values	0.968759	0.242382	0.001175 **	1.401e-08 ***	0.467909	0.085916

ΠΙΝΑΚΑΣ VII (Συγκριτικός Πίνακας)

	OLSR	RR	RR	PCR (5 comps)		PLSR(4 comps)		PLSR(5 comps)	
				CV	TEST	CV	TEST	CV	TEST
RMSEP	0.486596	0.7479373	0.869019	0.1930	0.7190	0.2528	0.7529	0.156	0.7288
St. error	0.1075	0.171046	0.5772	0.1511		0.1914833		0.1267	
R^2	99.12	95.02	52.45	98.26		97.21		98.78	

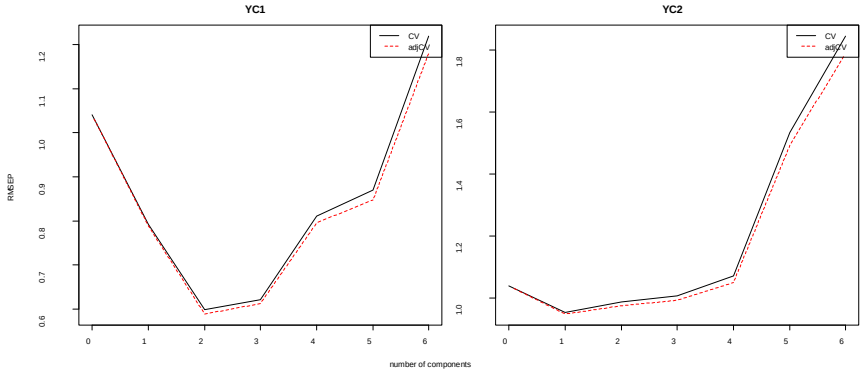
3.2 2η Εφαρμογή (*Pietrogrante et al., 1989)*

Τα δεδομένα σε αυτή την εφαρμογή αποτελούνται από δεκατρείς παρατηρήσεις με έξι ανεξάρτητες μεταβλητές και δύο εξαρτημένες.

Αρχικά επιλέγουμε να προσαρμόσουμε ένα μοντέλο παλινδρόμησης και για τις δύο εξαρτημένες μεταβλητές. Επειδή το μέγεθος του δείγματος είναι πολύ μικρό επιλέγουμε να μην το χωρίσουμε σε *training set* και *testng set*. Επίσης για την εκτίμηση του *RMSEP* χρησιμοποιείται η μέθοδος *leave – one – out Cross Validation*. Το συγκεκριμένο μοντέλο ακολουθεί παρακάτω:

ΠΙΝΑΚΑΣ Ι

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	Επιλογή Συνιστωσών
X Variance expl.	84.73	92.18	96.96	99.13	99.87	100.00	1 – 3
YC1 Variance expl.	52.88	81.55	83.28	83.58	86.73	87.16	2 – 3
YC2 Variance expl.	32.06	56.49	60.67	64.83	65.64	68.01	2 – 4
YC1 $RMSEP_{LOOCV}$	0.7925	0.5983	0.6208	0.8105	0.8697	1.219	2
YC2 $RMSEP_{LOOCV}$	0.9548	0.9888	1.0084	1.071	1.535	1.845	1

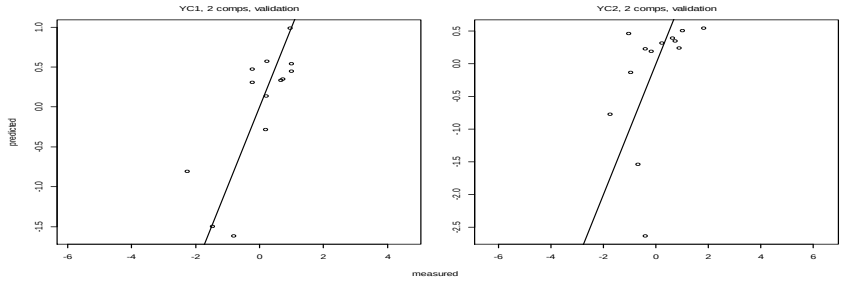


Σχήμα 1

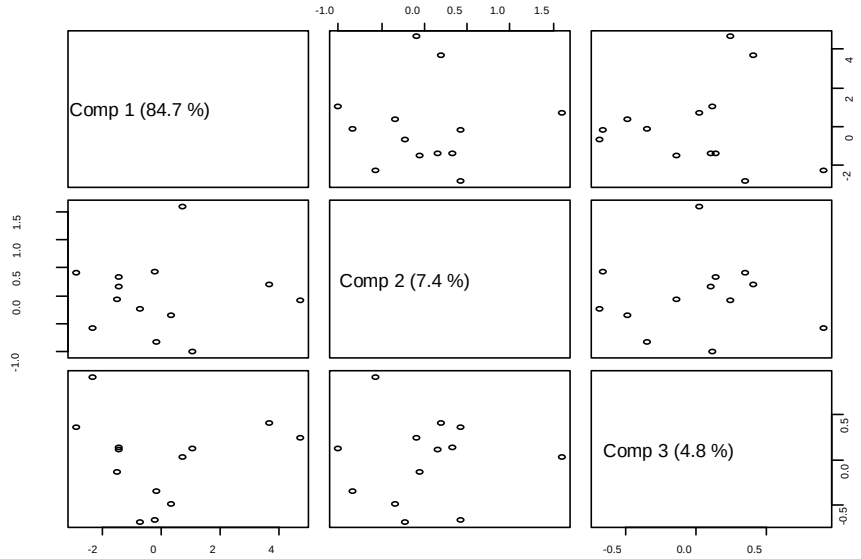
Παρατηρούμε ότι η πρώτη συνιστώσα συμμετέχει στην διασπορά των δεδομένων σε ποσοστό 84.73% και οι δύο συνιστώσες συμμετέχουν σε ποσοστό 92.18%. Στην συνέχεια η μεταβολή του ποσοστού αυτού είναι αρκετά μικρή. Επίσης ο συντελεστής προσδιορισμού της μεταβλητής $YC1$ φαίνεται να σταθεροποιείται μεταξύ της δεύτερης και της τρίτης συνιστώσας (81.55% και 83.28% αντίστοιχα). Το ίδιο παρατηρούμε και για την μεταβλητή $YC2$ (56.49% και 60.67% αντίστοιχα). Ένα γεγονός που θα πρέπει να μας προβληματίσει είναι ότι οι τιμές του συντελεστή προσδιορισμού για την $YC2$ είναι μάλλον μικρές. Για ελέγξουμε το μοντέλο ως προς

την προβλεπτική του ικανότητα χρησιμοποιούμε τις εκτιμήσεις του RMSEP του Πίνακα Ι όπως και τα αντίστοιχα γραφήματα (Σχήμα 1). Έτσι για την $YC1$, το RMSEP εμφανίζει χαμηλές τιμές στην δεύτερη και τρίτη συνιστώσα (0.5983 και 0.6208 αντίστοιχα) ενώ για την $YC2$ οι χαμηλές τιμές του RMSEP εμφανίζονται στην πρώτη και δεύτερη συνιστώσα (0.9548 και 0.9888 αντίστοιχα). Από όλες αυτές τις παρατηρήσεις υποψιαζόμαστε ότι η χρήση δύο συνιστωσών αποτελεί την καλύτερη επιλογή. Ωστόσο θα πρέπει να προχωρήσουμε και σε άλλους ελέγχους μέχρι να καταλήξουμε στην συγκεκριμένη απόφαση οι οποίοι πραγματοποιούνται κυρίως με την χρήση γραφημάτων.

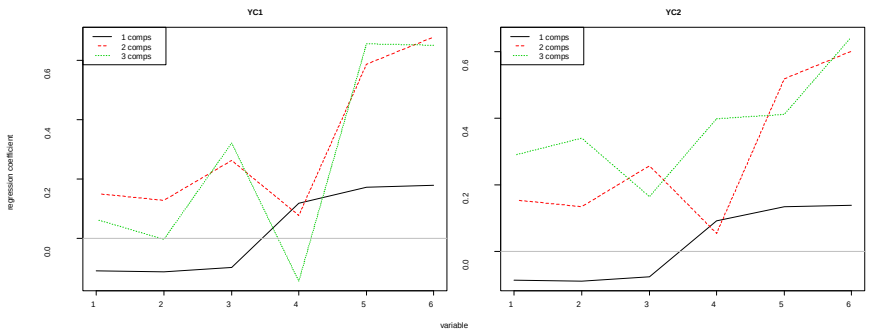
Στο γράφημα που ακολουθεί (Σχήμα 2) παρουσιάζεται η σχέση μεταξύ των εκτιμημένων τιμών των $YC1$ και $YC2$ για δύο συνιστώσες με τις αντίστοιχες μετρημένες τιμές. Για την μεταβλητή $YC1$ υπάρχει μια ασθενής γραμμική σχέση σε αντίθεση με την $YC2$ που εμφανίζεται μια ασθενής καμπυλότητα. Η αρνητική αυτή εξέλιξη δείχνει ότι το μοντέλο μας μάλλον δεν είναι το κατάλληλο τουλάχιστον για την $YC2$. Το επόμενο ζευγαρωτό γράφημα (Σχήμα 3) παριστάνει τις τιμές των *scores* για τις τρεις πρώτες συνιστώσες όπου ελέγχουμε για την ύπαρξη ομάδων τιμών ή ακραίων τιμών. Όπως παρατηρούμε δεν υπάρχουν τέτοιες ενδείξεις. Στο γράφημα των συντελεστών των ανεξάρτητων μεταβλητών (Σχήμα 4) παρατηρούμε ότι για την $YC1$ οι συντελεστές των αντίστοιχων μεταβλητών που προκύπτουν από δύο και τρεις συνιστώσες εμφανίζουν μια τάση ταύτισης εκτός από τους συντελεστές της $1^{ης}$ της $2^{ης}$ και της $4^{ης}$ μεταβλητής ενώ για την $YC2$ μόνο οι συντελεστές της $6^{ης}$ μεταβλητής είναι πολύ κοντά. Η τελευταία παρατήρηση μας οδηγεί στην σκέψη για την εισαγωγή τουλάχιστον μιας ακόμα συνιστώσας.



Σχήμα 2



Σχήμα 3



Σχήμα 4

Τέλος, εφαρμόζουμε χρησιμοποιούμε το *Jackknife Test*, ώστε να ελέγξουμε την σημαντικότητα των συντελεστών του μοντέλου για καθεμία από τις εξαρτημένες μεταβλητές. Παρατηρούμε ότι για την μεταβλητή $YC1$ στατιστικά σημαντικές θεωρούνται οι μεταβλητές $CR3, CR5, CR6$. Ενώ για την μεταβλητή $YC2$ καμιά από τις εξαρτημένες μεταβλητές δεν θεωρείται σημαντική.

	CR1	CR2	CR3	CR4	CR5	CR6
Συντελεστές YC1	0.150087	0.127634	0.262218	0.076878	0.588072	0.679405
p - values	0.1275527	0.3705697	0.0055832 **	0.7751561	0.0019457 **	0.0002666 ***
Συντελεστές YC2	0.154424	0.134274	0.601341	0.053919	0.518002	0.256180
p - values	0.5053	0.7031	0.2252	0.8159	0.1706	0.3108

Προσαρμόζοντας λοιπόν, ένα κοινό μοντέλο και στις δύο μεταβλητές καταλήξαμε στο συμπέρασμα ότι ένα μοντέλο με δύο συνιστώσες προσεγγίζει το βέλτιστο. Όπως έχουμε ήδη αναφέρει στην παράγραφο 2.4, προσαρμόζουμε ξεχωριστά μοντέλα σε κάθε μεταβλητή όταν αυτές είναι ασυσχέτιστες. Διαφορετικά, η απόφαση για ένα κοινό μοντέλο μπορεί να μας οδηγήσει στην επιλογή πολλών συνιστωσών το οποίο δύσκολα θα ερμηνεύεται. Στην συγκεκριμένη εφαρμογή το πλήθος των συνιστωσών είναι πραγματικά μικρό. Επειδή, όμως, υπάρχουν αμφιβολίες για την προβλεπτική του ικανότητα (κυρίως για την μεταβλητή $YC2$), θα προσαρμόσουμε ξεχωριστά μοντέλα για την κάθε μεταβλητή.

Ξεκινώντας θα ελέγξουμε τον βαθμό συσχέτισης των δύο μεταβλητών χρησιμοποιώντας τον πίνακα συσχέτισης $Y'Y$ και την τιμή του *VIF*. Τα αποτελέσματα είναι τα ακόλουθα:

Ο πίνακας $Y'Y$

	YC1	YC2
YC1	1.0000000	0.6743356
YC2	0.6743356	1.0000000

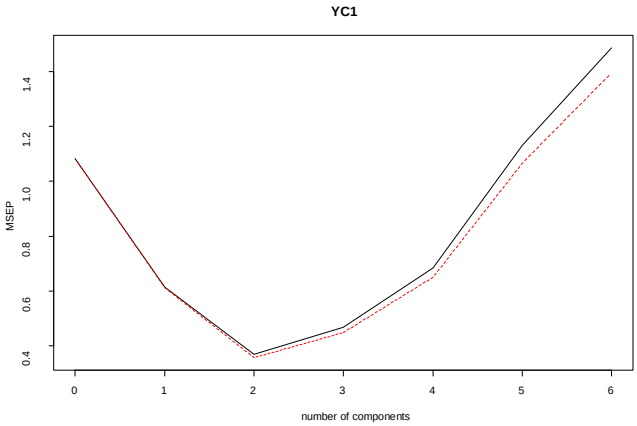
Η τιμή του *VIF*

	YC1	YC2
YC1	1.833949	-1.236697
YC2	-1.236697	1.833949

Όπως προκύπτει από τους δύο πίνακες οι δύο μεταβλητές δεν εμφανίζουν γραμμική εξάρτηση (*VIF*=1.8339) ενώ δεν εμφανίζουν και υψηλό βαθμό συσχέτισης (0.6743). Έτσι λοιπόν, προσαρμόζοντας αρχικά ένα μοντέλο PLSR για την μεταβλητή $YC1$, χρησιμοποιώντας την ίδια μεθοδολογία με το προηγούμενο μοντέλο, έχουμε τα αποτελέσματα του Πίνακα ΙΙ.

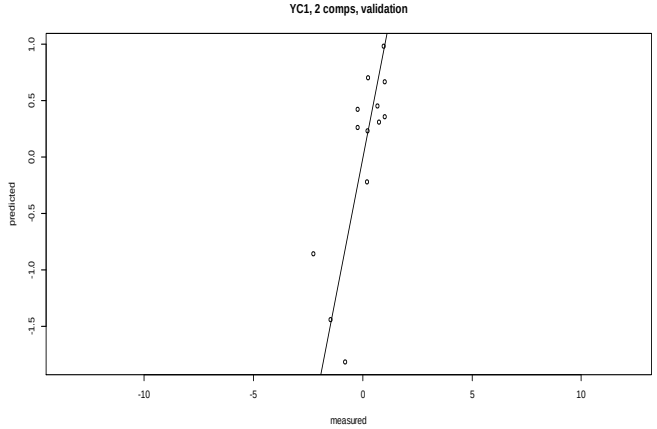
ΠΙΝΑΚΑΣ ΙΙ (μεταβλητή $YC1$)

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	Επιλογή Συνιστωσών
X Variance expl.	84.77	92.16	95.62	98.14	99.78	100.00	1 – 3
YC1 Variance expl.	52.71	83.11	85.13	87.02	87.06	87.16	2 – 3
$RMSEP_{LOOCV}$	0.7836	0.6069	0.6832	0.8267	1.063	1.219	2
p –value (rand test)	NA	0.001	0.999	0.999	0.992	0.999	2

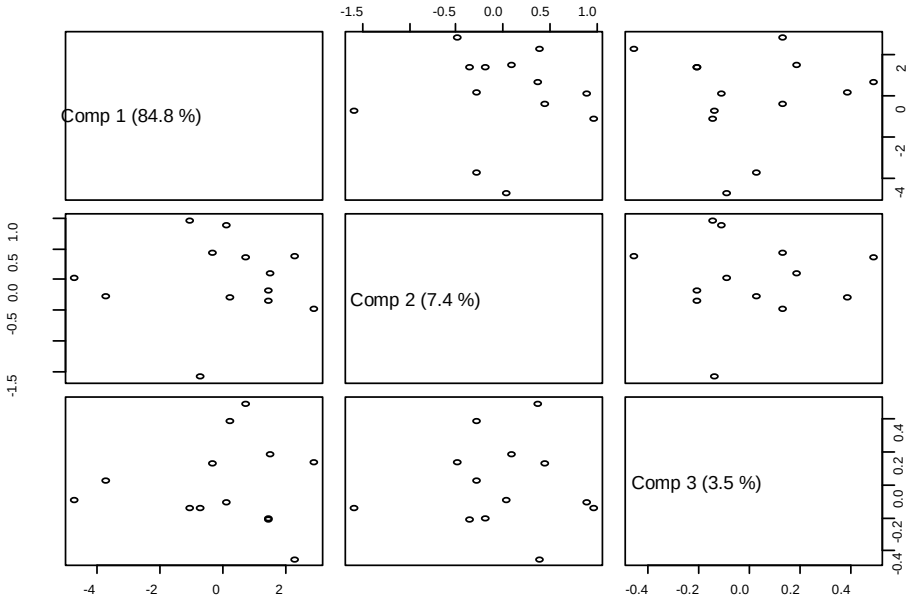


Σχήμα 5

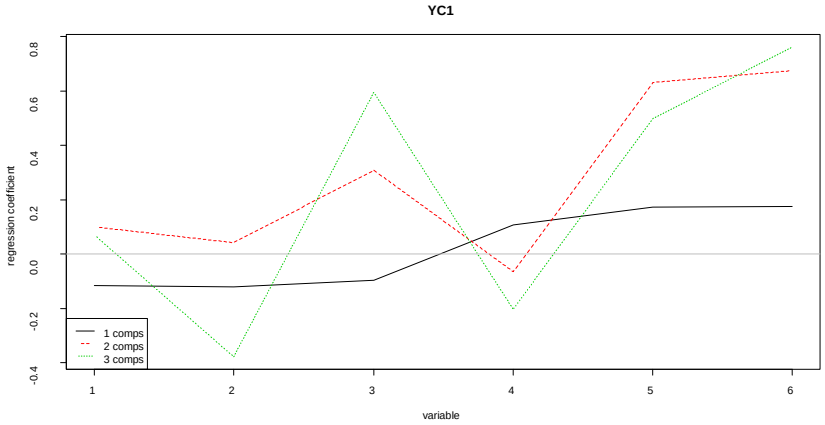
Παρατηρούμε ότι τα αποτελέσματα ελάχιστα μεταβάλλονται σε σχέση με το κοινό μοντέλο. Για παράδειγμα στην δεύτερη συνιστώσα ο συντελεστής προσδιορισμού της $YC1$ μετατρέπεται από 81.55% σε 83.11% και το *RMSEP* από 0.5983 σε 0.6069. Από το παραπάνω γράφημα (Σχήμα 5) μπορούμε να πούμε ότι είναι πιο ξεκάθαρη η επιλογή των δύο συνιστωσών για την προσέγγιση του βέλτιστου μοντέλου. Επίσης, με την χρήση των γραφημάτων που ακολουθούν, όπως και στο προηγούμενο μοντέλο, επαληθεύεται η αρχική επιλογή μας. Πράγματι στο Σχήμα 6 υπάρχει μια ικανοποιητική γραμμική σχέση χωρίς να παρατηρείται κάποιο είδος καμπυλότητας. Μπορούμε να συμπεράνουμε μάλιστα, ότι είναι ελαφρώς βελτιωμένη αν συγκρίνουμε το γράφημα αυτό με το αντίστοιχο του Σχήματος 2. Στο γράφημα με τις τιμές των *scores* (Σχήμα 7) για τις πρώτες τρεις συνιστώσες είναι εμφανής η τυχαία κατανομή των σημείων και η απουσία ακραίων τιμών. Στο διάγραμμα των συντελεστών (Σχήμα 8) για τις τρεις πρώτες συνιστώσες παρατηρούμε ότι η κατάσταση βελτιώνεται συγκριτικά με το αντίστοιχο διάγραμμα του Σχήματος 4 αφού μόνο οι συντελεστές της $2^{ης}$ και $3^{ης}$ μεταβλητής φαίνονται αρκετά διαφορετικοί. Την απόφαση μας για την χρήση δύο συνιστωσών στο μοντέλο ενισχύει το *randomization t test* καθώς και το αντίστοιχο γράφημα (Σχήμα 9). Όπως φαίνεται από τις *p – values* στατιστικά σημαντική φαίνεται να είναι η χρήση δύο συνιστωσών στο μοντέλο *p – value=0.001*.



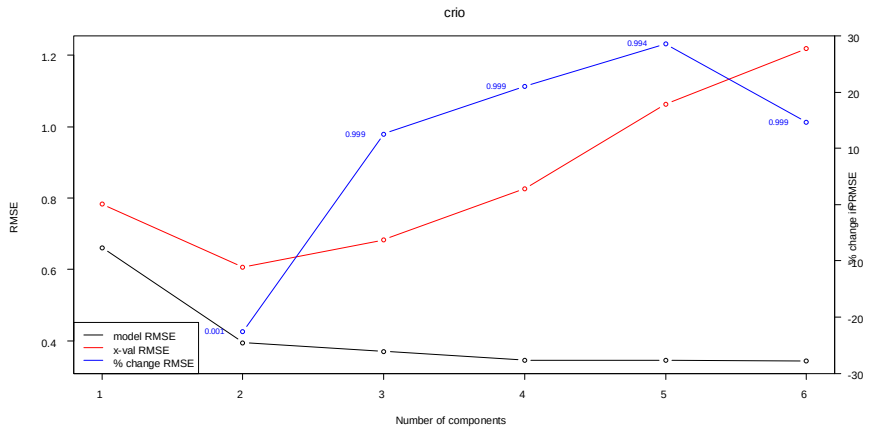
Σχήμα 6



Σχήμα 7



Σχήμα 8



Σχήμα 9

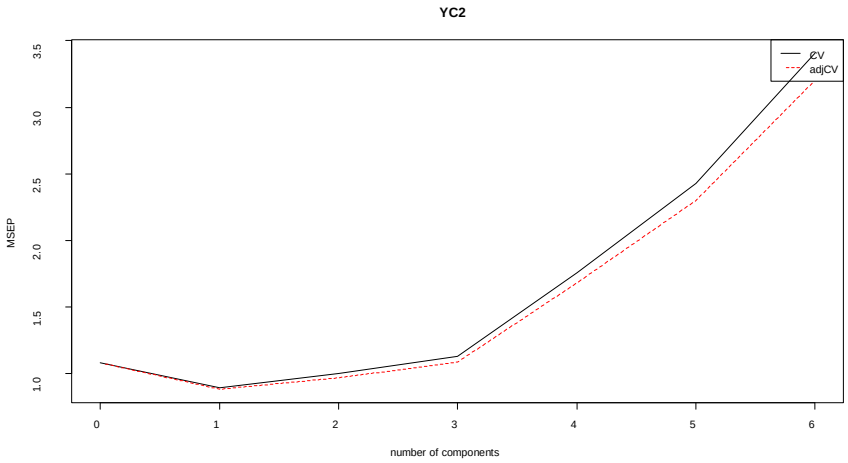
Εφαρμόζοντας και σε αυτό το μοντέλο το *Jackknife test* παρατηρούμε ότι σημαντικές μεταβλητές είναι οι ίδιες με το αρχικό μοντέλο.

	CR1	CR2	CR3	CR4	CR5	CR6
Συντελεστές YC1	0.098921	0.041451	0.308251	-0.065287	0.631672	0.673914
p - values	0.5705957	0.7113434	0.0017415 **	0.7408219	0.0006757 ***	0.0002781 ***

Συνεχίζουμε με την προσαρμογή ενός μοντέλου PLSR για την μεταβλητή $YC2$ ακολουθώντας για μια ακόμα φορά την ίδια διαδικασία όπως στην προσαρμογή των δύο προηγούμενων μοντέλων και τα αποτελέσματα είναι αυτά του Πίνακα ΙΙΙ.

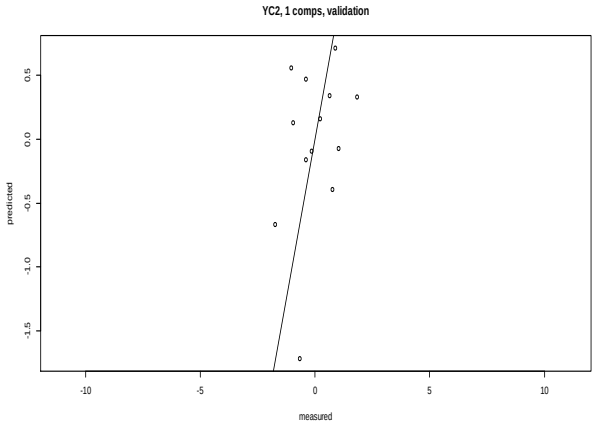
ΠΙΝΑΚΑΣ ΙΙΙ (μεταβλητή $YC2$)

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	Επιλογή Συνιστωσών
X Variance expl.	84.63	92.09	94.29	98.94	99.82	100.00	1 – 3
YC2 Variance expl.	32.79	59.43	63.93	64.81	66.78	68.01	2 – 3
$RMSEP_{LOOCV}$	0.9457	1.000	1.064	1.327	1.558	1.845	1
p –value (rand test)	NA	0.728	0.998	0.999	0.999	0.999	1

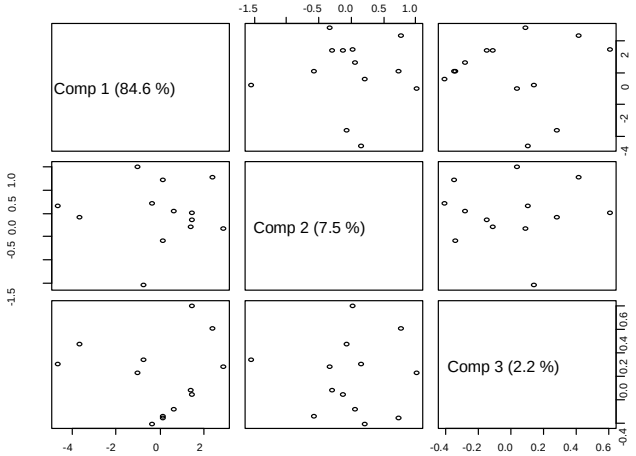


Σχήμα 10

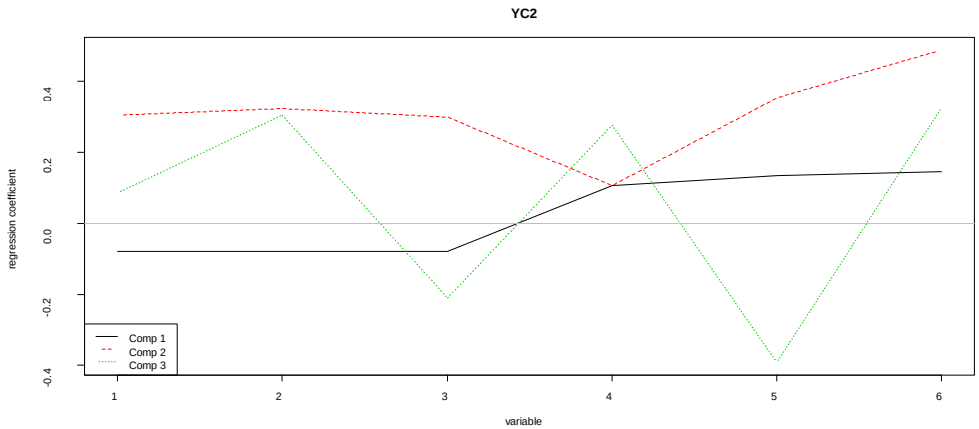
Παρατηρούμε ότι και σε αυτή την περίπτωση τα αποτελέσματα ελάχιστα μεταβάλλονται σε σχέση με το κοινό μοντέλο. Το παραπάνω γράφημα (Σχήμα 10) για το RMSEP μας υποδεικνύει την χρήση μιας συνιστώσας για την προσέγγιση του βέλτιστου μοντέλου. Βέβαια ο συντελεστής προσδιορισμού της εξαρτημένης μεταβλητής για το μοντέλο με μια συνιστώσα είναι αρκετά μικρός (32.79%) και για αυτό είναι απαραίτητος ο γραφικός έλεγχος καλής προσαρμογής του μοντέλου. Στο Σχήμα 11 παρατηρούμε μια βελτιωμένη γραμμική σχέση συγκριτικά με το αντίστοιχο γράφημα του Σχήματος 2, ενώ η καμπυλότητα η οποία ήταν εμφανής στο σχήμα 2 τώρα έχει εξαλειφθεί.. Στο γράφημα με τις τιμές των *scores* (Σχήμα 12) για τις πρώτες τρεις συνιστώσες είναι εμφανής η τυχαία κατανομή των σημείων και η απουσία ακραίων τιμών. Στο διάγραμμα των συντελεστών (Σχήμα 13) για τις τρεις πρώτες συνιστώσες παρατηρούμε ότι η κατάσταση δεν βελτιώνεται συγκριτικά με το αντίστοιχο διάγραμμα του Σχήματος 4. Με το *randomization t test* καθώς και το αντίστοιχο γράφημα (Σχήμα 9) που ακολουθούν επιβεβαιώνεται η επιλογή μας για την χρήση μιας συνιστώσας. Όπως φαίνεται από τις *p – values* στατιστικά σημαντική φαίνεται να είναι η χρήση μόνο μιας συνιστώσας στο μοντέλο.



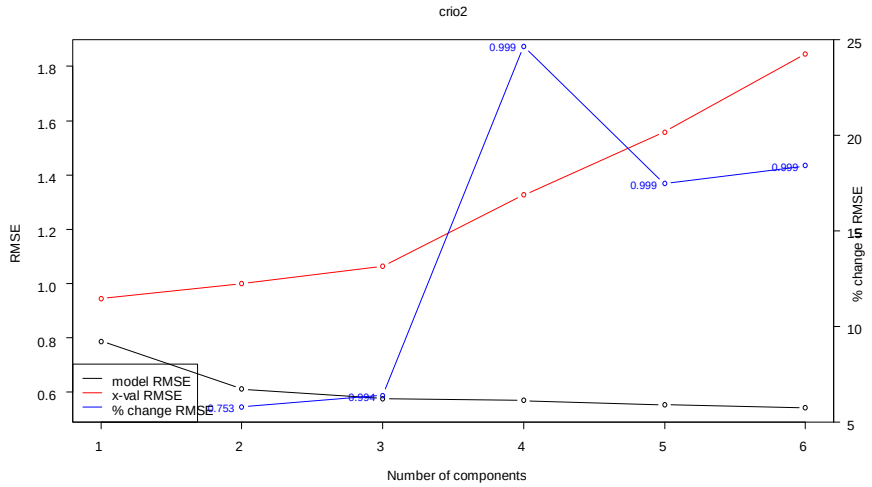
Σχήμα 11



Σχήμα 12



Σχήμα 13



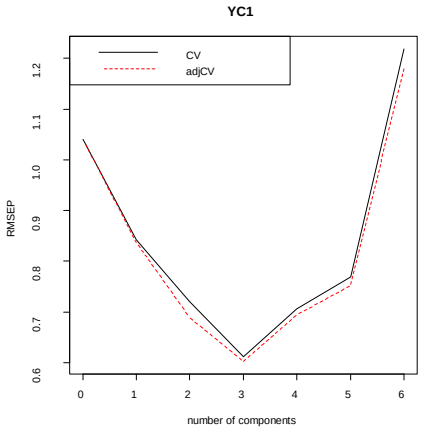
Σχήμα 14

Αν χρησιμοποιήσουμε και σε αυτή την περίπτωση το *Jackknife Test* παρατηρούμε ότι το συγκεκριμένο μοντέλο προσαρμόζεται καλύτερα από το αρχικό κοινό μοντέλο. Έτσι σημαντικές εμφανίζονται οι μεταβλητές $CR2, CR5, CR6$ σε επίπεδο 10%.

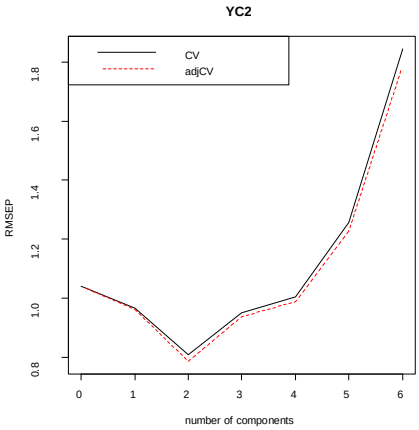
	CR1	CR2	CR3	CR4	CR5	CR6
Συντελεστές YC2	0.144931	-0.079583	-0.078678	0.106878	0.134520	-0.079487
p - values	0.15796	0.08449	0.06735	0.18406	0.09189	0.10761

Λαμβάνοντας υπόψη και τα τρία μοντέλα που προσαρμόσαμε σε αυτή την εφαρμογή συμπεραίνουμε ότι η χρήση του κοινού μοντέλου δεν επηρεάζει την μεταβλητή $YC1$. Η προβλεπτική του αξία σε σχέση με την $YC1$ δεν μεταβάλλεται συγκριτικά με το μοντέλο που προσαρμόσαμε μόνο για την μεταβλητή αυτή. Για την μεταβλητή $YC2$ όμως, φαίνεται ότι η χρήση δύο συνιστωσών μειώνει την προβλεπτική αξία του μοντέλου ενώ από το *Jackknife test* υποψιαζόμαστε ότι η προσαρμογή αυτού του μοντέλου για την $YC2$ είναι μάλλον προβληματική. Άρα καταλήγουμε στο συμπέρασμα ότι είναι προτιμότερο να χρησιμοποιηθούν δύο ξεχωριστά μοντέλα. Αν τώρα, προσαρμόσουμε στα δεδομένα ένα μοντέλο PCR για κάθε μία από τις μεταβλητές $YC1$ και $YC2$, παρατηρούμε ότι για την $YC1$ προτείνεται ως βέλτιστο το μοντέλο με τρεις συνιστώσες, ενώ για την $YC2$ προτείνεται το μοντέλο με τις δύο συνιστώσες. Στο ίδιο συμπέρασμα καταλήγουμε μελετώντας τα παρακάτω γραφήματα του RMSEP, για την $YC1$ (Σχήμα 15) και την $YC2$ (Σχήμα 16).

Με αυτό τον τρόπο άλλωστε έχουν μελετηθεί και τα συγκεκριμένα δεδομένα *(Pietrogrande et al.,1989).*



Σχήμα 15



Σχήμα 16

3.3 3η Εφαρμογή *(Lindberg et al., 1983)*

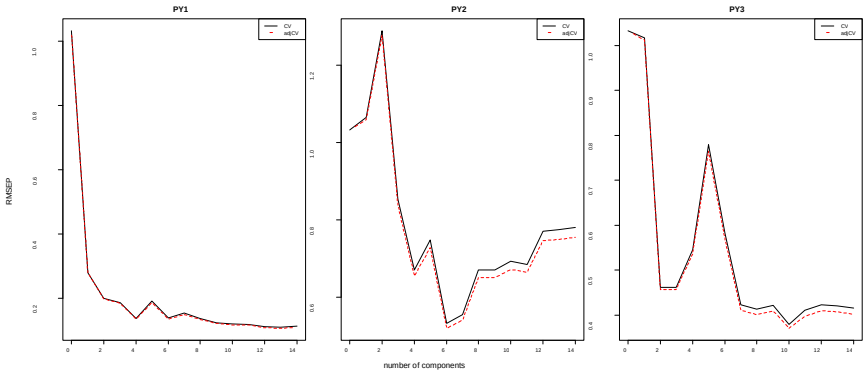
Στην εφαρμογή αυτή τα δεδομένα αποτελούνται από δεκαέξι παρατηρήσεις με είκοσι επτά ανεξάρτητες μεταβλητές και τρεις εξαρτημένες. Η εκτίμηση του μέσου τετραγωνικού σφάλματος των προβλέψεων, σε όλες τις περιπτώσεις, θα γίνει με την μέθοδο *cross – validation* και συγκεκριμένα με την *leave one – out cross validation* η οποία κάθε φορά θα εφαρμόζεται 16 φορές. Οι Πίνακες I, II, III, IV που αναφέρονται παρακάτω βρίσκονται στο τέλος της εφαρμογής (σελ. 101).

Από τα αποτελέσματα του Πίνακα I, η πρώτη συνιστώσα εξηγεί τα δεδομένα κατά 97.46% και επομένως οι επόμενες δεκατρείς συμμετέχουν με ελάχιστα ποσοστά στην εξήγηση των δεδομένων. Επίσης οι τιμές του συντελεστή προσδιορισμού των εξαρτημένων μεταβλητών κρίνονται ικανοποιητικές από την επιλογή τριών συνιστωσών και έπειτα. Λαμβάνοντας υπόψη τα παραπάνω, η επιλογή ενός μοντέλου με τέσσερεις συνιστώσες φαντάζει ικανοποιητική. Πρέπει όμως να ελέγξουμε και την προβλεπτική ικανότητα του μοντέλου από τις τιμές του εκτιμημένου *RMSEP* χρησιμοποιώντας και την βοήθεια του σχετικού γραφήματος (Σχήμα 1).

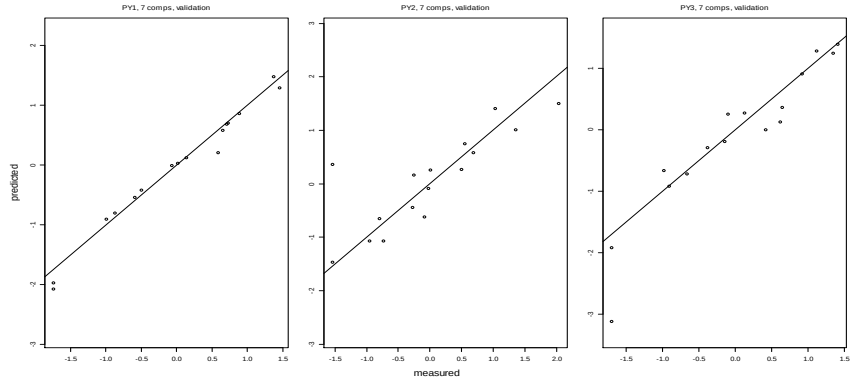
Έτσι, για την μεταβλητή $PY1$ το *RMSEP* σταθεροποιείται σε χαμηλές τιμές, από το μοντέλο με έξι συνιστώσες και έπειτα. Ενώ την χαμηλότερη τιμή την συναντάμε στο μοντέλο με τις δεκατρείς συνιστώσες. Για την $PY2$ το *RMSEP* δεν εμφανίζει σταθερά χαμηλές τιμές παρά μόνο για τα μοντέλα με έξι και επτά συνιστώσες, με την χαμηλότερη τιμή να εμφανίζεται στο μοντέλο με έξι συνιστώσες. Τέλος, για την $PY3$ το *RMSEP* σταθεροποιείται σε χαμηλές τιμές από το μοντέλο με επτά συνιστώσες. Η χαμηλότερη τιμή του *RMSEP* εμφανίζεται στο μοντέλο με τις δέκα συνιστώσες. Καταλήγουμε στο συμπέρασμα ότι πρέπει να αναθεωρήσουμε την αρχική μας εκτίμηση που ήταν ένα μοντέλο με τέσσερεις συνιστώσες, σε ένα μοντέλο με επτά ή οκτώ συνιστώσες. Η τελική απόφαση θα παρθεί μετά την μελέτη και επιπλέον γραφημάτων.

Πράγματι, από τα Σχήματα 2 και 3 που παριστάνουν τις εκτιμημένες τιμές των τριών εξαρτημένων μεταβλητών σε συνάρτηση με τις αντίστοιχες τιμές του δείγματος μετά την προσαρμογή των μοντέλων με επτά και οκτώ συνιστώσες αντίστοιχα, χωρίς να εμφανίζουν κάποια διαφορά. Τα δύο μοντέλα, φαίνεται να προσαρμόζονται αρκετά καλά για τις δύο πρώτες μεταβλητές. Ενώ για την μεταβλητή $PY3$ υπάρχουν ενδείξεις καμπυλότητας. Επίσης για κάθε μια από τις μεταβλητές $PY2$ και $PY3$ υπάρχει σημείο το οποίο απέχει αρκετά από την ευθεία. Στο Σχήμα 4 παριστάνονται

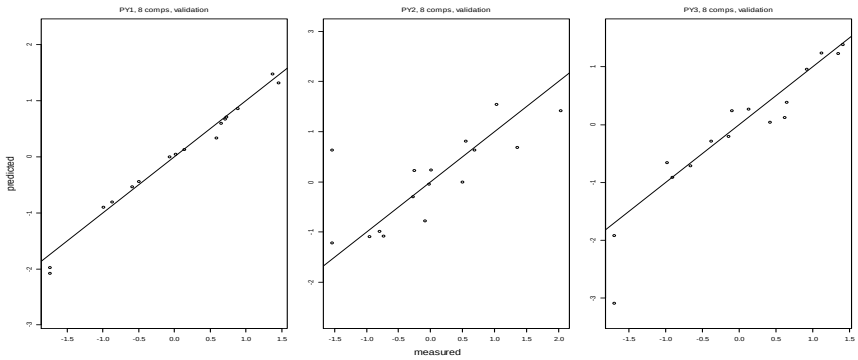
οι τιμές των *scores* που προκύπτουν από την έκτη, την έβδομη και την όγδοη συνιστώσα. Παρατηρούμε ότι δεν εμφανίζονται ομάδες δεδομένων ή ακραία σημεία. Στο γράφημα με τους εκτιμημένους συντελεστές των είκοσι επτά ανεξάρτητων μεταβλητών (Σχήμα 5), συμπεραίνουμε ότι για την μεταβλητή $PY1$ καλή επιλογή θα ήταν ένα μοντέλο με έξι συνιστώσες, για την μεταβλητή $PY2$ ένα μοντέλο με τις έξι ή επτά συνιστώσες και για την μεταβλητή $PY3$ ένα μοντέλο με επτά συνιστώσες. Λαμβάνοντας υπόψη όλα τα παραπάνω καταλήγουμε στο συμπέρασμα ότι στην περίπτωση που θέλουμε να προσαρμόσουμε ένα κοινό μοντέλο η καλύτερη λύση φαίνεται ότι είναι εκείνη του μοντέλου με τις επτά συνιστώσες.



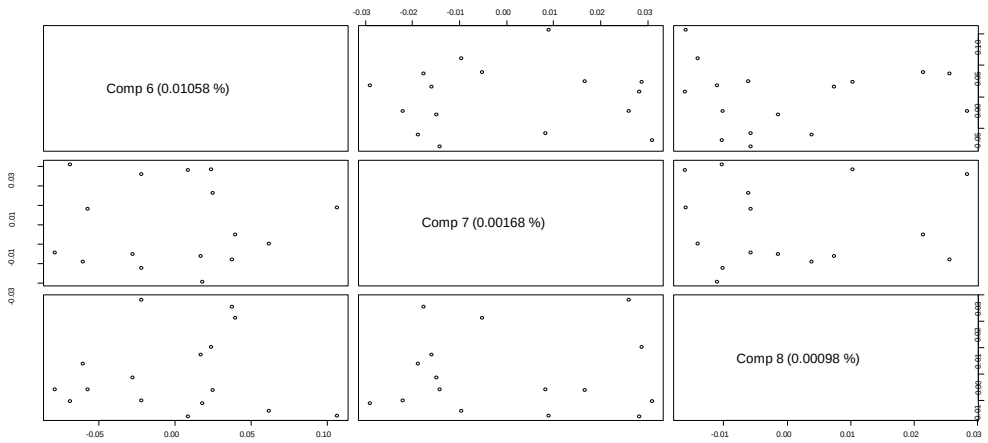
Σχήμα 1



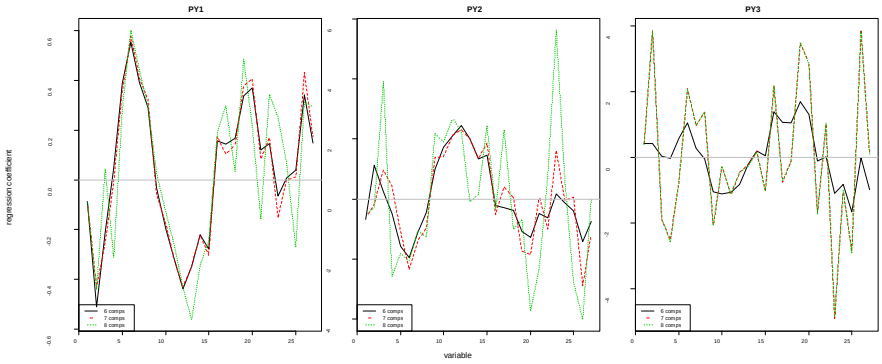
Σχήμα 2



Σχήμα 3



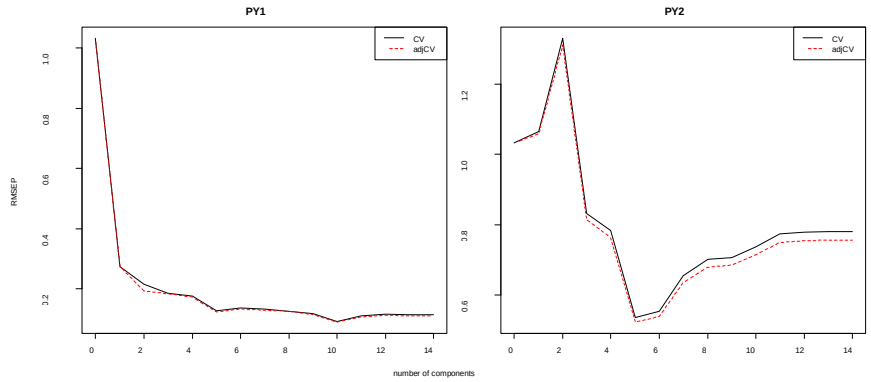
Σχήμα 4



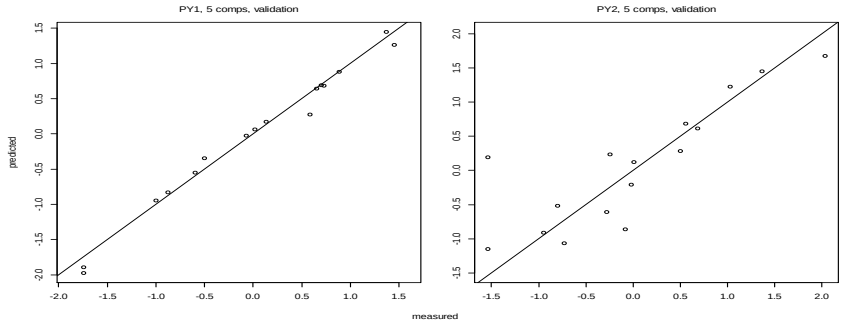
Σχήμα 5

Προσαρμόζοντας λοιπόν, ένα μοντέλο και για τις τρεις μεταβλητές καταλήξαμε στο συμπέρασμα ότι η καλύτερη περίπτωση είναι εκείνη με τις οκτώ συνιστώσες. Επειδή αυτό το πλήθος των συνιστωσών ενδέχεται να θεωρηθεί μεγάλο είμαστε αναγκασμένοι να εξετάσουμε και άλλες περιπτώσεις. Έχοντας σαν δεδομένο ότι σε περιπτώσεις που είναι αντίστοιχες με την συγκεκριμένη εφαρμογή (παραπάνω από μια εξαρτημένες μεταβλητές) επιθυμούμε να προσαρμόσουμε μοντέλα *PLSR* με όσο το δυνατό χαμηλό πλήθος συνιστωσών, θα εξετάσουμε την περίπτωση του μοντέλου με δύο μεταβλητές. Όπως είναι φανερό μπορούμε να προσαρμόσουμε τρία τέτοια διαφορετικά μοντέλα. Από την μελέτη που έγινε, το μοντέλο που προσαρμόστηκε για τις μεταβλητές $PY1$ και $PY3$ απαιτεί την χρήση οκτώ μεταβλητών ενώ το μοντέλο για τις $PY2$ και $PY3$ απαιτεί την χρήση επτά μεταβλητών. Συγκριτικά με το αρχικό μοντέλο, στο μοντέλο των $PY1$ και $PY3$ οι τιμές του *RMSEP* βελτιώνονται ελάχιστα. Αντίστοιχα, στο μοντέλο των $PY2$ και $PY3$ όπου χρησιμοποιείται το ίδιο πλήθος συνιστωσών η βελτίωση των τιμών του *RMSEP* είναι σχεδόν μηδενική. Κατά συνέπεια θα προτιμήσουμε το αρχικό μας μοντέλο και

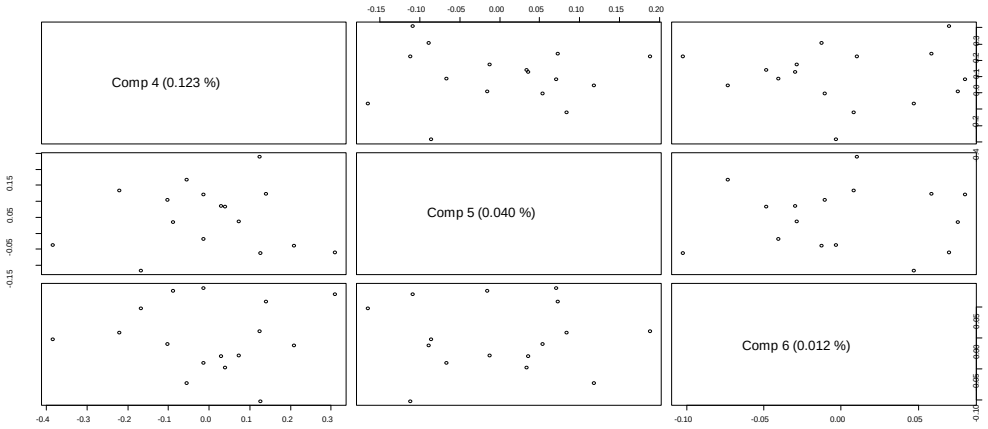
όχι τα δύο προηγούμενα. Η περίπτωση της προσαρμογής μοντέλου στις μεταβλητές $PY1$ και $PY2$ παρουσιάζει ξεχωριστό ενδιαφέρον και θα την μελετήσουμε αναλυτικότερα. Πράγματι, προσαρμόζοντας ένα μοντέλο *PLSR* στις παραπάνω δύο μεταβλητές παίρνουμε τα αποτελέσματα του Πίνακα ΙΙ. Σύμφωνα με τα αποτελέσματα αυτά και με την βοήθεια του γραφήματος για το *RMSEP* (Σχήμα 6) παρατηρούμε ότι για την $PY1$ το *RMSEP* σταθεροποιείται σε χαμηλές τιμές, από το μοντέλο με τις πέντε συνιστώσες, ενώ παίρνει την ελάχιστη τιμή του στο μοντέλο με τις δέκα συνιστώσες. Για την $PY2$ δεν υπάρχει σταθεροποίηση του *RMSEP* σε χαμηλές τιμές αλλά την ελάχιστη τιμή του την συναντάμε στο μοντέλο με πέντε συνιστώσες. Επίσης οι συντελεστές προσδιορισμού των δύο αυτών μεταβλητών για το συγκεκριμένο μοντέλο βρίσκονται σε αρκετά υψηλό επίπεδο (99.74% και 96.17% αντίστοιχα). Υποψιαζόμαστε λοιπόν, ότι το καταλληλότερο μοντέλο για αυτές τις δύο μεταβλητές είναι εκείνο με τις πέντε συνιστώσες. Επίσης οι τιμές του *RMSEP* σε αυτό το μοντέλο είναι μικρότερες από εκείνες του αρχικού με τις επτά συνιστώσες (0.1266 έναντι 0.1533 για την $PY1$ και 0.5344 έναντι 0.5555 για την $PY2$). Συνεχίζοντας με τον γραφικό έλεγχο από τα Σχήματα 7, 8, φαίνεται ότι το μοντέλο προσαρμόζεται αρκετά καλά. Τέλος, στο Σχήμα 9 παρατηρούμε ότι και για τις δύο μεταβλητές οι γραμμές που αντιστοιχούν στα μοντέλα με πέντε και έξι συνιστώσες κινούνται παράλληλα και πολύ κοντά, ενώ σε ορισμένα διαστήματα σχεδόν ταυτίζονται. Βλέπουμε λοιπόν, ότι και οι γραφικοί έλεγχοι επιβεβαιώνουν την αρχική μας εκτίμηση, ότι το μοντέλο με τις πέντε συνιστώσες είναι το καλύτερο. Επιλέγοντας αυτό το μοντέλο για τις μεταβλητές $PY1$ και $PY2$ θα πρέπει να προσαρμόσουμε και ένα επιπλέον για την μεταβλητή $PY3$. Κάτι τέτοιο ακολουθεί αμέσως παρακάτω, αφού ούτως ή άλλως θα μελετήσουμε την περίπτωση προσαρμογής μοντέλων PLSR για καθεμία μεταβλητή ξεχωριστά.



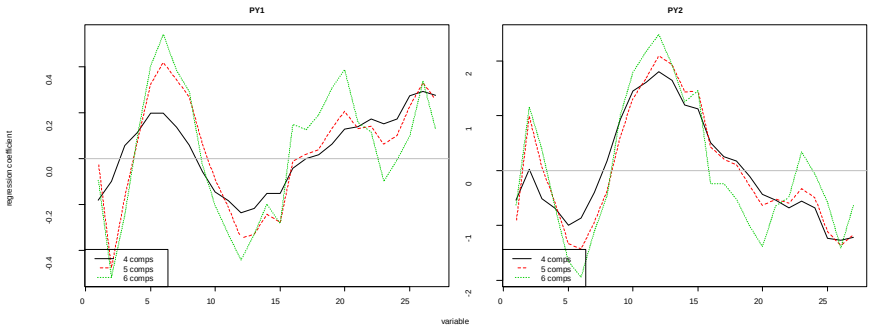
Σχήμα 6



Σχήμα 7



Σχήμα 8



Σχήμα 9

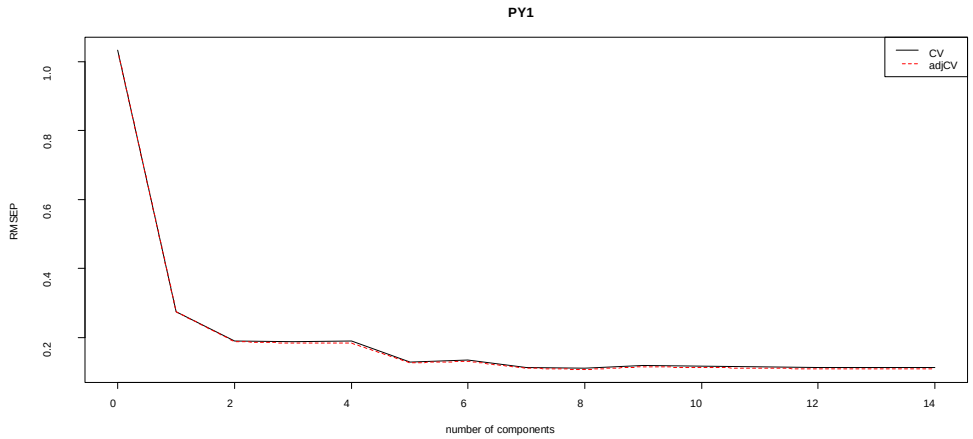
Πριν προχωρήσουμε στην προσαρμογή μοντέλων για κάθε μεταβλητή ξεχωριστά θα πρέπει να ελέγξουμε αν υπάρχει εξάρτηση μεταξύ των εξαρτημένων μεταβλητών. Κάτι τέτοιο ,όπως έχουμε πει, θα μας βοηθήσει να αποφασίσουμε αν θα προσαρμόσουμε ένα μοντέλο *PLSR* για όλες τις εξαρτημένες μεταβλητές ή ξεχωριστά μοντέλα για κάθε μεταβλητή. Ο έλεγχος αυτός μπορεί να γίνει μέσω του πίνακα συσχέτισης $Y'Y$ αλλά και μέσω του υπολογισμού της τιμής του *VIF* των εξαρτημένων μεταβλητών $PY1$, $PY2$ και $PY3$.

$$Y'Y = \begin{bmatrix} 1.0000000 & 0.22691112 & 0.26322550 \\ 0.2269111 & 1.00000000 & 0.02018386 \\ 0.2632255 & 0.02018386 & 1.00000000 \end{bmatrix}$$

$$(Y'Y)^{-1} = \begin{bmatrix} 1.1343186 & -0.25146541 & -0.2935060 \\ -0.2514654 & 1.05615455 & 0.04487483 \\ -0.2935060 & 0.04487483 & 1.07635252 \end{bmatrix}$$

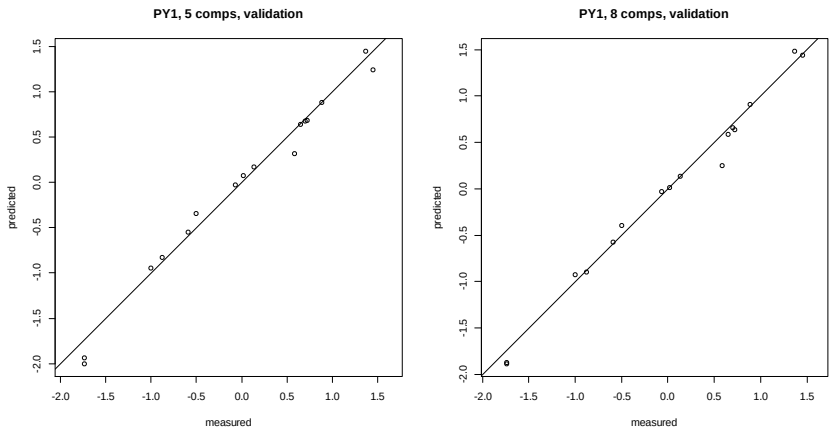
Παρατηρούμε ότι κανένα από τα διαγώνια στοιχεία *(VIF)* του πίνακα $(Y'Y)^{-1}$ δεν είναι μεγαλύτερο του 5. Και από τον πίνακα $Y'Y$ οι συσχετίσεις των εξαρτημένων μεταβλητών είναι πολύ μικρές. Επομένως η προσαρμογή τριών διαφορετικών μοντέλων ενδεχομένως να μας οδηγήσει σε καλύτερα αποτελέσματα.

Προσαρμογή μοντέλου για την PY_1

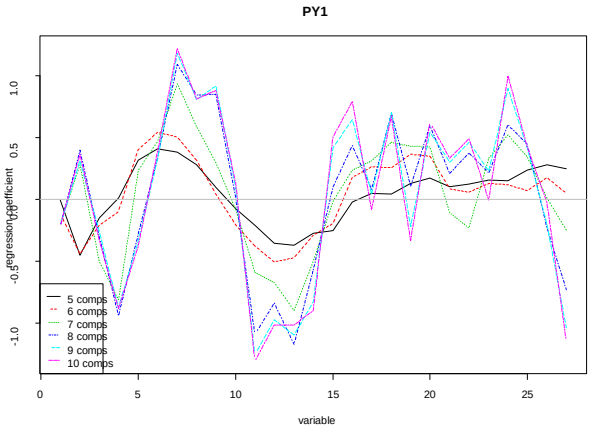


Σχήμα 10

Μελετώντας τα αποτελέσματα του Πίνακα ΙΙΙ, αλλά και με την βοήθεια του Σχήματος 10 διαπιστώνουμε ότι οι τιμές του *RMSEP* σταθεροποιούνται σε χαμηλά επίπεδα, από το μοντέλο με τις πέντε συνιστώσες (0,1295), ενώ η ελάχιστη τιμή του βρίσκεται στο μοντέλο με τις οκτώ συνιστώσες (0.1111). Από την παρατήρηση αυτή, φαίνεται ότι το καλύτερο μοντέλο θα είναι κάποιο από τα παραπάνω δύο. Ξεκινώντας τους γραφικούς ελέγχους στο Σχήμα 11 φαίνεται ότι τα μοντέλα με τις πέντε και οκτώ συνιστώσες προσαρμόζονται το ίδιο καλά και είναι σχεδόν όμοια με τα αντίστοιχα γραφήματα των προηγούμενων μοντέλων. Από το γράφημα όμως, των εκτιμημένων συντελεστών (Σχήμα 12), φαίνεται καθαρά ότι οι συντελεστές αρχίζουν να ταυτίζονται από το μοντέλο με τις οκτώ συνιστώσες και έπειτα. Η παρατήρηση αυτή μας επιτρέπει να επιλέξουμε το μοντέλο με τις οκτώ συνιστώσες ως το καλύτερο. Επίσης το *randomization test* μας υποδεικνύει ως καλά μοντέλα εκείνα με τις 8,9 ή 10 συνιστώσες.

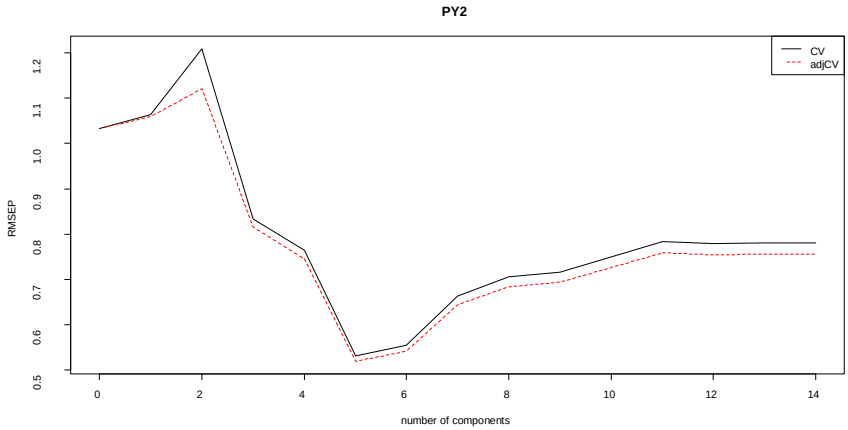


Σχήμα 11


PY1
regression coefficient
variable
5 comps
6 comps
7 comps
8 comps
9 comps
10 comps

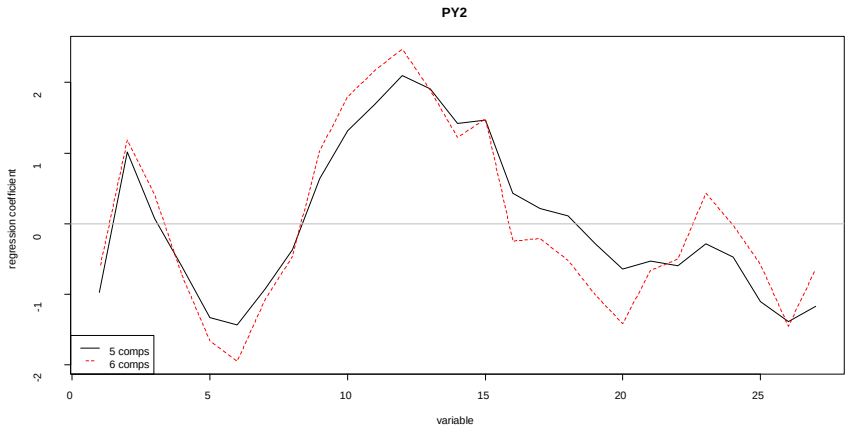
Σχήμα 12

Προσαρμογή μοντέλου για την PY_2



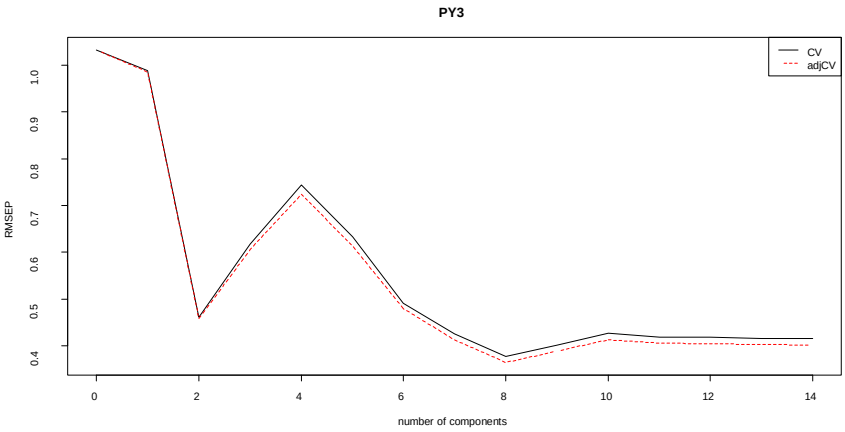
Σχήμα13

Από τα αποτελέσματα του Πίνακα IV και το Σχήμα 13 μπορούμε να πούμε ότι η περίπτωση της μεταβλητής $PY2$ δείχνει να είναι πιο ξεκάθαρη. Παρατηρούμε μια μικρή σταθεροποίηση της εκτιμήτριας του *RMSEP* σε χαμηλές τιμές στα μοντέλα με πέντε και έξι συνιστώσες, με την ελάχιστη τιμή να ανήκει στο μοντέλο με τις πέντε συνιστώσες. Είναι φανερό ότι αναμένουμε το βέλτιστο μοντέλο να είναι εκείνο με τις πέντε συνιστώσες. Από το γράφημα των εκτιμημένων συντελεστών (Σχήμα 14) για τα παραπάνω δύο μοντέλα παρατηρείται μια ισχυρή τάση ταύτισης των γραμμών. Επιπλέον, το *randomization test* για την συγκεκριμένη μεταβλητή επιβεβαιώνει την αρχική μας υποψία. Το συγκεκριμένο *test* μας δίνει την δυνατότητα απόρριψης των μοντέλων από έξι συνιστώσες και πάνω. Άρα θα δεχθούμε ως βέλτιστο μοντέλο για την μεταβλητή $PY2$ εκείνο με τις πέντε συνιστώσες.



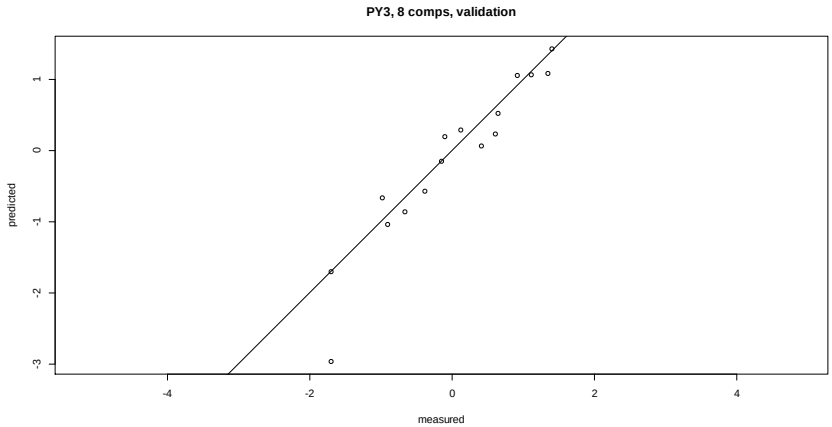
Σχήμα14

Προσαρμογή μοντέλου για την PY_3

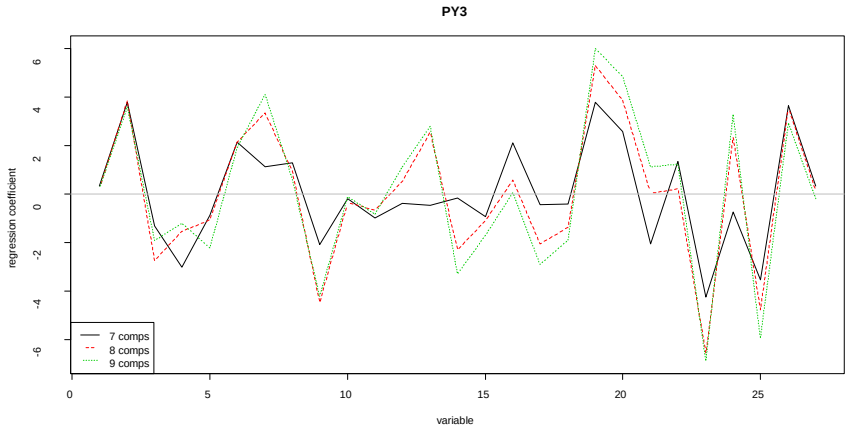


Σχήμα 15

Με την βοήθεια του Πίνακα V και του Σχήματος 15 παρατηρούμε ότι η ελάχιστη τιμή του RMSEP εμφανίζεται στο μοντέλο με τις οκτώ συνιστώσες, ενώ αρχίζει να σταθεροποιείται σε χαμηλές τιμές από το μοντέλο με τις εφτά συνιστώσες. Σε αυτή την περίπτωση φαίνεται ότι το μοντέλο με τις οκτώ συνιστώσες είναι το καλύτερο. Με δεδομένο ότι το *randomization test* δεν μας βοηθά ώστε να επιβεβαιώσουμε την προηγούμενη διαπίστωση, συνεχίζουμε με τους γραφικούς ελέγχους. Από το Σχήμα 16 διαπιστώνουμε την καλή προσαρμογή του μοντέλου. Τέλος, από το Σχήμα 17 παρατηρούμε ότι οι συντελεστές των μοντέλων με οκτώ και εννιά συνιστώσες σχεδόν ταυτίζονται. Αυτό, αποτελεί ισχυρή ένδειξη ότι το μοντέλο με τις οκτώ συνιστώσες είναι το καλύτερο.



Σχήμα 16



Σχήμα 17

Συνοψίζοντας όλα τα παραπάνω, προσαρμόσαμε στα δεδομένα πέντε διαφορετικά μοντέλα με τρεις διαφορετικούς τρόπους. Σύμφωνα με τον πρώτο τρόπο προσαρμόσαμε στα δεδομένα ένα μοντέλο και για τις τρεις μεταβλητές καταλήγοντας στο συμπέρασμα ότι το βέλτιστο είναι αυτό με τις επτά συνιστώσες. Στον δεύτερο τρόπο προσαρμόσαμε ένα μοντέλο για τις μεταβλητές $PY1$ και $PY2$,αφού πρώτα απορρίψαμε τους άλλους συνδυασμούς. Το βέλτιστο μοντέλο στο οποίο καταλήξαμε είναι αυτό με τις πέντε συνιστώσες. Και στον τρίτο τρόπο προσαρμόσαμε ένα μοντέλο για κάθε μεταβλητή ξεχωριστά. Καταλήξαμε στο συμπέρασμα ότι το βέλτιστο μοντέλο για την μεταβλητή $PY1$ είναι εκείνο με τις οκτώ συνιστώσες, για την $PY2$ εκείνο με τις πέντε και για την $PY3$ εκείνο με τις οκτώ.

Στην συνέχεια, θα πρέπει να αποφασίσουμε ποίο ή ποια από τα παραπάνω μοντέλα, περιγράφουν καλύτερα τα δεδομένα μας. Όπως έχουμε αναφέρει (Κεφάλαιο 2), η χρήση ενός μοντέλου και για τις τρεις μεταβλητές θεωρείται η καλύτερη επιλογή αφού μας δίνει την δυνατότητα να αποκτήσουμε μια συνολική εικόνα για τα δεδομένα. Υπάρχει όμως ο κίνδυνος να οδηγηθούμε σε ένα μοντέλο με πολλές συνιστώσες, με αποτέλεσμα η ερμηνεία του να γίνει αρκετά περίπλοκη. Επίσης ένας άλλος κίνδυνος που υπάρχει είναι να επιλέξουμε ένα μοντέλο που προσαρμόζεται αρκετά καλά αλλά με μικρή προβλεπτική αξία ή το αντίθετο.

Λαμβάνοντας υπόψη τις παραπάνω παρατηρήσεις και με την βοήθεια του συνοπτικού Πίνακα VI, παρατηρούμε ότι οι συντελεστές προσδιορισμού παραμένουν σταθεροί και σε υψηλά επίπεδα. Επίσης, για το μοντέλο με τις δύο μεταβλητές $PY1$ και $PY2$ οι τιμές του $RMSEP$ είναι, σε ικανοποιητικό επίπεδο, καλύτερες από τις αντίστοιχες του μοντέλου με τις τρεις μεταβλητές και ελάχιστα χειρότερες από εκείνες των ξεχωριστών μοντέλων των συγκεκριμένων μεταβλητών. Ένα επιπλέον

σημαντικό στοιχείο είναι ότι το συγκεκριμένο μοντέλο χρησιμοποιεί λιγότερες συνιστώσες (πέντε) από τα υπόλοιπα μοντέλα, εκτός από το μοντέλο της $PY2$ που χρησιμοποιεί το ίδιο πλήθος συνιστωσών. Μπορούμε να συμπεράνουμε λοιπόν, ότι η χρήση του μοντέλου για τις μεταβλητές $PY1$ και $PY2$ και του μοντέλου για την $PY3$ αποτελούν την καλύτερη επιλογή.

ΠΙΝΑΚΑΣ VI

	Ένα μοντέλο και για τις τρεις μεταβλητές			Μοντέλο για δύο μεταβλητές		Μοντέλα για κάθε μεταβλητή ξεχωριστά		
Βέλτιστο μοντέλο	7 comps			5 comps		8 comps	5 comps	8 comps
	$PY1$	$PY2$	$PY3$	$PY1$	$PY2$	$PY1$	$PY2$	$PY3$
RMSEP	0.1533	0.5555	0.4228	0.126	0.5344	0.1111	0.5312	0.377
R^2	99.82	97.73	99.27	99.74	96.17	99.99	96.20	99.81

ΠΙΝΑΚΑΣ Ι

comps	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Good models
X Var.(%)	97.46	99.64	99.82	99.94	99.98	99.99	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	1 – 14
PY1 (R^2)	92.96	97.64	98.66	99.31	99.66	99.82	99.82	99.83	99.88	99.92	99.93	99.99	100.00	100.00	1 – 14
PY2 (R^2)	12.97	13.00	83.44	93.22	95.46	97.30	97.73	99.22	99.48	99.59	99.59	99.90	99.98	99.98	3 – 14
PY3 (R^2)	19.82	87.83	89.98	90.91	91.34	96.19	99.27	99.27	99.33	99.51	99.96	99.98	99.98	100.00	2 – 14
$RMSEP_{PY1}$	0.278	0.199	0.1857	0.1370	0.1914	0.1389	0.1533	0.1368	0.1237	0.1201	0.1185	0.1113	0.1093	0.1136	13
$RMSEP_{PY2}$	1.065	1.289	0.8539	0.6702	0.7473	0.5326	0.5555	0.6701	0.6711	0.6931	0.6851	0.7699	0.7742	0.7802	6
$RMSEP_{PY3}$	1.017	0.460	0.4617	0.5449	0.7789	0.5812	0.4228	0.4125	0.4217	0.3794	0.4099	0.4220	0.4200	0.4152	10

ΠΙΝΑΚΑΣ ΙΙ

comps	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Good model
X Var.(%)	97.46	99.64	99.82	99.94	99.98	99.99	100	100	100	100	100	100	100	100	1 – 14
PY1 (R^2)	93.25	98.57	98.67	99.34	99.74	99.81	99.82	99.82	99.94	99.96	99.99	100	100	100	1 – 14
PY2 (R^2)	12.98	34.64	85.73	93.50	96.17	97.42	99.05	99.53	99.56	99.82	99.93	99.99	100	100	3 – 14
$RMSEP_{PY1}$	0.274	0.215	0.1849	0.176	0.126	0.1364	0.132	0.1255	0.1177	0.0917	0.110	0.1149	0.1146	0.1136	10
$RMSEP_{PY2}$	1.064	1.331	0.8321	0.7833	0.5344	0.5528	0.6546	0.7007	0.7061	0.7369	0.7742	0.7786	0.7802	0.7802	5

ΠΙΝΑΚΑΣ ΙΙΙ

comps	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Good model
X Var.(%)	97.46	99.64	99.82	99.94	99.98	99.99	100	100	100	100	100	100	100	100	1 – 14
PY1 (R^2)	93.28	97.85	99.21	99.53	99.75	99.86	99.96	99.99	100	100	100	100	100	100	1 – 14
$RMSEP_{PY1}$	0.275	0.190	0.188	0.1893	0.1295	0.1348	0.1142	0.1111	0.1187	0.1170	0.1148	0.1136	0.1136	0.1136	8
Rand.t.test	NA	0.369	0.999	0.998	0.000	0.790	0.991	0.000	0.002	0.001	0.141	0.077	0.862	0.855	8 – 10

ΠΙΝΑΚΑΣ ΙV

comps	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Good model
X Var.(%)	97.46	97.66	99.74	99.94	99.98	99.99	99.99	100.00	100.00	100.00	100.00	100	100	100	1 – 14
PY2 (R^2)	13.20	83.79	85.54	93.36	96.20	97.46	99.07	99.54	99.77	99.93	99.96	100	100	100	3 – 14
$RMSEP_{PY2}$	1.064	1.209	0.8330	0.7642	0.5312	0.5547	0.6638	0.7050	0.7163	0.7496	0.7828	0.7784	0.7799	0.7802	5
Rand.t.test	NA	0.908	0.000	0.017	0.003	0.877	0.999	0.999	0.970	0.999	0.999	0.053	0.998	0.999	5

ΠΙΝΑΚΑΣ V

comps	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Good model
X Var.(%)	97.30	99.64	99.84	99.87	99.89	99.98	100.00	100.00	100.00	100	100	100	100	100	1 – 14
PY3 (R^2)	22.74	87.89	91.49	94.68	96.72	97.04	99.28	99.81	99.94	100	100	100	100	100	2 – 14
$RMSEP_{PY3}$	0.989	0.4617	0.618	0.7435	0.6336	0.4908	0.425	0.377	0.4009	0.4265	0.4187	0.4179	0.4157	0.4152	8

ΒΙΒΛΙΟΓΡΑΦΙΑ

Abdi, H. (2010) Partial Least Squares Regression and Projection on Latent Structure Regression (PLS Regression), *WIREs Computational Statistics*, **2**: 97 – 106.

Draper, N., Smith, H. (1998), *Applied Regression Analysis*, 3rd ed., Wiley.

Efron, B. (1982) *The Jackknife, the Bootstrap and Other Resampling Plans*, Philadelphia: SIAM

Fakedulegn, B. Desta; Colbert, J.J.; Hicks R.R., Jr; Schuckers, Mickael E. (2002), Coping with Multicollinearity: An Example on Application of Principal Components Regression in Dendroecology. Res. Pap. NE-721. Newton Square, PA: U. S. Department of Agriculture, Forest Service, *Northeastern Research Station*. 43p.

Fearn, T. (1983) A Misuse of Ridge Regression in the Calibration of a Near Infrared Reflectance Instrument, *Journal of the Royal Statistical Society*, Series C (Applied statistics), **32**: 73 – 79

Helland, I.S. (1988) On the Structure of Partial Least Squares Regression, *Communications in Statistics – Simulations and Computation*, **17**: 581 - 607

Hoerl, E.A., Kennard, W.R (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, **12**: 55 – 67

Hoerl, E.A., Kennard, W.R, Baldwin, K.F. (1975) Ridge Regression: some simulations, *Communications in Statistics – Theory and Methods*, **4**: 105 – 123

Jolliffe, T.(2002) *Principal Component Analysis*, Series: Springer Series in Statistics, 2nd ed., Springer, NY

Krishnan, A., Williams, L., McIntosh, A.R., Abdi, H. (2010) Partial Least Squares (PLS) Methods for Neuroimaging: A Tutorial and Review, *NeuroImage*, **56**: 455 – 475

Lindberg, W., Persson, J.-A., Wold, S. (1983). Partial Least Squares Method for Spectrofluorimetric Analysis of Mixtures of Humic Acid and Ligninsulfonate. *Analytical Chemistry* **55**, 643-648.

Lott, W.F. (2003), The Optimal Set of Regression Components Restrictions on a Least Squares Regression, *Communications in Statistics – Theory and Methods*, **2**: 449 – 464

Mason,R.L., Gunst,R.F., Webster, J.T. (1975) Regression Analysis and Problems of Multicollinearity, *Communications in Statistics*, **4**: 297 – 292

McDonald, G.C., Garleneau, D.I. (1975) A Monte Carlo Evaluation of Some Ridge – type Estimators, *Journal of the American Statistical Association*, **70(350)**: 407 – 412

Mevik, B-H., Cederkvisk, H.R. (2004), Mean Square Error of Prediction (MSEP) Estimates for Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR), *Journal of Chemometrics*, **18**: 422 – 429

Mevik, B.H., Wehrens, R (2007) The pls Package: Principal Component and Partial Least Squares Regression in R, *Journal of Statistical Software*, **18**(2): 1 – 24.

Montgomery, D., Peck, E., Vining, G. (2006), *Introduction to Linear Regression Analysis*.4th ed. Wiley.

Pietrogrande, M.C., Dondi, F., Borea, P.A., Bighi, C. (1989) Principal Component Analysis in Structure – Retention and Retention – Activity Studies of Benzodiazepines, *Chemometrics and Intelligent Laboratory Systems*, **5**: 257 – 262

Ripley, B.D. (1996) *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press

Seber, G, Lee, A. (2003) *Linear Regression Analysis*, 2nd ed. Wiley

Shao, J. (1997) An Asymptotic Theory for Linear Model Selection, *Statistica Sinica*, **7**: 221 – 264

Stone, M. (1977) An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion, *Journal of the Royal Statistical Society*. Series B (Methodological), **39**: 44-47

Van der Voet, H. (1994) Comparing the Predictive Accuracy of Models Using a Simple Randomization Test, *Chemometrics and Intelligent Laboratory Systems*, **25**: 313 – 323

Wiklund, S., Nilsson, D., Eriksson, L., Sjostrom, L., Wald, S. ,Faber, K. (2007) A randomization test for PLS component selection, *Journal of Chemometrics*, **21**:427 – 439

Wold, H. (1966) Estimation of Principal Components and Related Models by Iterative Least Squares, In Krishnaiaah, P.R. *Multivariate Analysis*. New York: Academic Press. pp. 391–420.

Wold, H. (1985). Partial least squares. In Kotz, Samuel; Johnson, Norman L. *Encyclopedia of Statistical Sciences* 6. New York: Wiley. pp. 581–591

Wold, S., Sjostrom, M., Eriksson, L. (2001), PLS – Regression: a Basic Tool of Chemometrics, *Chemometrics and Intelligent Laboratory Systems*, **58**: 109 – 130

Yeniay, O., Goktas, A. (2002) A comparison of Partial Least Squares Regression with other prediction methods, *Hacettepe Journal of Mathematics*, 31: 99 – 111

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

http://robjhyndman.com/hyndsight/crossvalidation/
http://robjhyndman.com/hyndsight/loocv-linear-models/
www.faqs.org/faqs/ai-faq/neural-nets/part3/section-12.html
http://www.chemometry.com/Research/COM.html
http://statmaster.sdu.dk/courses/ST02/data/index.html