



**SCHOOL OF MEDICINE UNIVERSITY OF THESSALY POSTGRADUATE
PROGRAMME (MSC)**

**“Research Methodology in Biomedicine, Biostatistics
and Clinical Bioinformatics”**

Master’s Thesis

**Multicollinearity: diagnostics and PCA as a method
of handling**

**Πολυσυγγραμμικότητα: διαγνωστικές μέθοδοι και
Ανάλυση Κύριων Συνιστωσών ως μέθοδος χειρισμού**

Scientific Committee:

*Apostolos Batsidis, MSc, PhD, Assistant Professor, Probability, Statistics
and Operations Research Unit, Department of Mathematics, University of
Ioannina (Supervisor)*

*Ioannis Stefanidis, MD, PhD, Professor of Internal
Medicine/Nephrology, Faculty of Medicine, University of Thessaly*

*Chrysoula Doxani, MSc, MD, PhD, Research Fellow in Genetic
Pharmacoepidemiology, University of Thessaly*

Δημητρακόπουλος Παναγιώτης

E-mail: p.dimitrakopoulos@gmail.com

Academic year: 2016 – 2017

ΠΕΡΙΛΗΨΗ

Η συγγραμμικότητα (collinearity) ή πολυσυγγραμμικότητα (multicollinearity) είναι εκείνη η ανεπιθύμητη κατάσταση, η οποία εμφανίζεται στην πολυμεταβλητή παλινδρόμηση όταν μία ανεξάρτητη μεταβλητή είναι γραμμική συνάρτηση των υπόλοιπων ή κάποιων ανεξάρτητων μεταβλητών.

Στην παρούσα εργασία πραγματοποιείται ανάλυση όλων των δεικτών διάγνωσης της πολυσυγγραμμικότητας, καθώς και της μεθόδου αντιμετώπισής της, η οποία είναι η Ανάλυση Κύριων Συνιστωσών (PCA). Μάλιστα, εφαρμόζεται η συγκεκριμένη μέθοδος σε ένα σύνολο ιατρικών δεδομένων με τη βοήθεια του SPSS.

Πιο συγκεκριμένα, το πρώτο κεφάλαιο αποτελεί μια εισαγωγή στην έννοια της πολυσυγγραμμικότητας. Παρουσιάζεται ένα πλήθος δεικτών με τη βοήθεια των οποίων μπορεί να εντοπιστεί το συγκεκριμένο πρόβλημα και αναφέρεται η μέθοδος Ανάλυσης Κύριων Μεταβλητών, με την οποία το αντιμετωπίζουμε.

Στο δεύτερο κεφάλαιο γίνεται μια εκτενής ανάλυση της μεθόδου PCA (Principal Component Analysis). Επίσης, παρουσιάζεται ένα παράδειγμα, στο οποίο αναλύεται η συγκεκριμένη μέθοδος.

Στο τρίτο κεφάλαιο εφαρμόζεται η μέθοδος PCA σε ένα σύνολο ιατρικών δεδομένων. Παρουσιάζεται η αποτελεσματικότητα της μεθόδου στη διάγνωση και την αντιμετώπιση του προβλήματος της πολυσυγγραμμικότητας, καθώς κι ο απλός χειρισμός της με τη βοήθεια του SPSS.

Λέξεις κλειδιά

Πολυσυγγραμμικότητα, Πολυμεταβλητή Ανάλυση Παλινδρόμησης, Ανάλυση Κύριων Συνιστωσών, SPSS.

ABSTRACT

Multicollinearity exists when two or more of the predictors in a regression model are moderately or highly correlated. Unfortunately, when it exists, it can wreak havoc on our analysis and thereby limit the research conclusions we can draw.

This dissertation introduces all indices of multicollinearity diagnoses, the basic principle of principal component regression and determination of "best" equation method.

The first chapter is an introduction to the multicollinearity (or collinearity). Afterwards, the procedure of diagnosis of this problem is presented. There are many indices that indicate this problem.

The second chapter deals with the method PCA (Principal Component Analysis). Specifically, the basic concepts are cited and the Principal Component Regression is analyzed. At the end of the chapter, PCA is proposed and illustrated by an example.

In the third chapter, the method PCA is applied to a real medical data set. This demonstrates the utility and versatility of the method.

Key words

Multicollinearity diagnosis, Multiple Regression Analysis, Principal Component Analysis, SPSS.

ΚΕΦΑΛΑΙΟ 1^ο

ΕΙΣΑΓΩΓΗ ΣΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΠΟΛΥΣΥΓΓΡΑΜΜΙΚΟΤΗΤΑΣ

Η ερμηνεία ενός προβλήματος με χρήση της μεθόδου αναλύσεως της πολλαπλής παλινδρόμησης επιτυγχάνεται καλύτερα όταν οι ανεξάρτητες μεταβλητές, οι οποίες αποτελούν το μοντέλο είναι μεταξύ τους ασυσχέτιστες. Όταν υπάρχουν έντονες συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών είναι αρκετά δύσκολο να αξιολογηθεί η πραγματική επίδραση μιας συγκεκριμένης ανεξάρτητης μεταβλητής πάνω στην εξαρτημένη μεταβλητή. Στην περίπτωση κατά την οποία οι ανεξάρτητες μεταβλητές δεν είναι ορθογώνιες μεταξύ τους είναι πιθανό οι εκτιμούμενοι συντελεστές παλινδρόμησης να είναι εξαιρετικά ασταθείς. Τότε οι τιμές τους υφίστανται σημαντικές αλλαγές όταν κάποια νέα μεταβλητή προστίθεται ή απομακρύνεται ή όταν συμβαίνουν μικρές μεταβολές στα δεδομένα του προβλήματος. Η κατάσταση που δημιουργείται όταν υπάρχουν ισχυρές συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών στην πολλαπλή παλινδρόμηση ονομάζεται πολυσυγγραμμικότητα (multicollinearity). Όταν εμφανίζεται αυτό το πρόβλημα χρειάζεται ιδιαίτερη προσοχή στην ερμηνεία των εκτιμητριών που προκύπτουν από το συγκεκριμένο μοντέλο^[1].

Ο δείκτης που χρησιμοποιείται συχνότερα για να δικαιολογηθεί η πολυσυγγραμμικότητα είναι ο απλός συντελεστής συσχέτισης. Όταν ο απλός συντελεστής συσχέτισης μεταξύ δύο ανεξάρτητων μεταβλητών είναι μεγάλος, λαμβάνεται υπόψη η πολυσυγγραμμικότητα. Εκτός από τον απλό συντελεστή συσχέτισης, μπορεί να προσδιοριστεί η πολυσυγγραμμικότητα με τη χρήση του SPSS και πιο συγκεκριμένα με τη βοήθεια των δεικτών *tolerance* και *variance in flation factor (VIF)*.

$$Tolerance = 1 - R_i^2,$$

όπου R_i^2 είναι ο συντελεστής προσδιορισμού της ανεξάρτητης μεταβλητής ως προς τις άλλες ανεξάρτητες μεταβλητές.

Όταν οι τιμές του δείκτη *tolerance* είναι μικρές, δηλαδή κοντά στο 0, τότε υφίσταται πρόβλημα πολυσυγγραμμικότητας. Ο δείκτης *VIF* είναι αντιστρόφως ανάλογος με το δείκτη *tolerance*, $VIF_i = \frac{1}{1-R_i^2}$. Συνεπώς, οι μεταβλητές των οποίων ο δείκτης *tolerance* έχει μικρές τιμές, έχουν μεγάλο *VIF*. Για τις μεταβλητές αυτές υφίσταται το πρόβλημα της πολυσυγγραμμικότητας. Επιπρόσθετοι δείκτες διάγνωσης πολυσυγγραμμικότητας είναι οι δείκτες *eigenvalue* (ιδιοτιμή), *condition index* και *variance proportion*. Οι ιδιοτιμές παρέχουν μια ένδειξη για τις διακριτές διαστάσεις που υπάρχουν μεταξύ των ανεξάρτητων μεταβλητών. Όταν αρκετές ιδιοτιμές είναι κοντά στο 0, τότε οι ανεξάρτητες μεταβλητές είναι αρκετά συσχετισμένες μεταξύ τους και ο πίνακας *X* καλείται *ill-conditioned*. Ο *condition index* είναι η τετραγωνική

ρίζα του πηλίκου της μέγιστης ιδιοτιμής προς κάθε διαδοχική ιδιοτιμή. Όταν ο δείκτης *condition index* παίρνει τιμές από 15 έως 30, τότε υποδεικνύεται μέτριο πρόβλημα πολυσυγγραμμικότητας, ενώ όταν παίρνει τιμές μεγαλύτερες από 30 υπάρχει σοβαρό πρόβλημα πολυσυγγραμμικότητας. Ο δείκτης *variance proportion* είναι αναλογίες διακύμανσης της εκτίμησης που υπολογίζονται από κάθε κύρια συνιστώσα που σχετίζεται με κάθε μία από τις ιδιοτιμές. Οι υψηλές τιμές του δείκτη *condition index* συμβάλλουν στην αύξηση της διακύμανσης των ανεξάρτητων μεταβλητών. Συνεπώς, οι ανεξάρτητες μεταβλητές με μεγάλες διακυμάνσεις είναι εκείνες που είναι αρκετά συσχετισμένες μεταξύ τους. Η κύρια συνιστώσα παλινδρόμησης είναι η μέθοδος συνδυασμού γραμμικής παλινδρόμησης με την ανάλυση κύριων συνιστωσών (*principal component analysis* (PCA)). Η ανάλυση κύριων συνιστωσών είναι μία στατιστική διαδικασία, η οποία μετατρέπει μία ομάδα παρατηρήσεων συσχετιζόμενων μεταβλητών σε μία ομάδα νέων τιμών μη γραμμικά συσχετιζόμενων μεταβλητών οι οποίες καλούνται κύριες συνιστώσες. Κατόπιν, φτιάχνουμε τις εξισώσεις παλινδρόμησης με ένα σύνολο μη συσχετισμένων βασικών μεταβλητών και παίρνουμε την «καλύτερη» εξίσωση σύμφωνα με την αρχή του μέγιστου συντελεστή προσδιορισμού R^2 και του ελάχιστου τυπικού σφάλματος εκτίμησης. Τέλος, η «καλύτερη» εξίσωση μετατρέπεται σε γενική εξίσωση γραμμικής παλινδρόμησης.^[1]

ΚΕΦΑΛΑΙΟ 2^ο

ΑΝΑΛΥΣΗ ΤΗΣ ΜΕΘΟΔΟΥ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΤΩΝ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ

Τα βήματα που ακολουθούμε για τη μέθοδο της παλινδρόμησης των κύριων συνιστωσών είναι τα εξής^[5]:

1. Στην περίπτωση που έχουμε μεγάλο αριθμό ανεξάρτητων μεταβλητών εφαρμόζουμε την μέθοδο *stepwise*, με σκοπό να βρούμε ποιές p ανεξάρτητες μεταβλητές είναι στατιστικά σημαντικές ($p < 0.05$) για το μοντέλο μας.
2. Παίρνουμε αυτό το σετ των p ανεξάρτητων μεταβλητών και εφαρμόζουμε ανάλυση κύριων συνιστωσών, ούτως ώστε να μετατρέψουμε το σύνολο των συσχετιζόμενων μεταβλητών σε ένα σύνολο μη συσχετιζόμενων κύριων συνιστωσών.
3. Υπολογίζουμε την τυποποιημένη (standardized) εξαρτημένη μεταβλητή, τις p τυποποιημένες (standardized) ανεξάρτητες μεταβλητές και τις τιμές των p κύριων συνιστωσών αντίστοιχα σύμφωνα με τις εξισώσεις (1), (2), (3) για την κατασκευή των p τυποποιημένων εξισώσεων παλινδρόμησης.

$$Y' = \frac{Y - \bar{Y}}{S_Y} \quad (1)$$

$$X'_i = \frac{X_i - \bar{X}_i}{S_{X_i}}, i=1,2,\dots,p \quad (2)$$

$$C_i = a_{i1}X'_1 + a_{i2}X'_2 + \dots + a_{ip}X'_p, i = 1,2,\dots,p \quad (3),$$

όπου Y' είναι standardized εξαρτημένη μεταβλητή, Y η εξαρτημένη μεταβλητή, S_Y η τυπική απόκλιση της εξαρτημένης μεταβλητής, \bar{Y} η μέση τιμή της εξαρτημένης μεταβλητής, X'_i είναι η i -οστή standardized ανεξάρτητη μεταβλητή, X_i είναι η i -οστή ανεξάρτητη μεταβλητή και \bar{X}_i είναι η μέση τιμή των i ανεξάρτητων μεταβλητών, S_{X_i} είναι η τυπική απόκλιση των i ανεξάρτητων μεταβλητών, C_i η i -οστή κύρια συνιστώσα και a_{ij} είναι ο συντελεστής του πίνακα των κύριων συνιστωσών.

4. Κατασκευάζουμε την εξίσωση παλινδρόμησης της standardized κύριας συνιστώσας με την πρώτη κύρια συνιστώσα, με τη δεύτερη, κ.ο.κ., με τη m -οστή. Παράλληλα ελέγχουμε αν οι κύριες συνιστώσες είναι ανεξάρτητες μεταξύ τους. Με αυτόν τον τρόπο καθορίζουμε την «καλύτερη» εξίσωση με βάση το μέγιστο

συντελεστή προσδιορισμού R^2 και το μικρότερο τυπικό σφάλμα (S.E.). Έτσι προκύπτει ότι:

$$\hat{y}'_j = \sum B'_i C_i \quad (4),$$

όπου $j = 1, \dots, m \leq p$ και $i = 1, \dots, K \leq p$,
 \hat{y}'_j είναι η εκτιμήτρια των j -οστων standardized κύριων συνιστωσών του μοντέλου και B'_i είναι ο i -στος standardized συντελεστής μερικής παλινδρόμησης του μοντέλου.

5. Από τις εξισώσεις (3) και (4) προκύπτει ότι:

$$\hat{y}' = \sum b'_i X'_i (i = 1, \dots, K \leq p) \quad (5),$$

όπου \hat{y}' είναι η εκτιμήτρια των τυποποιημένων (standardized) κύριων συνιστωσών του μοντέλου και b'_i είναι ο standardized συντελεστής μερικής παλινδρόμησης της τυποποιημένης (standardized) εξίσωσης γραμμικής παλινδρόμησης.

6. Υπολογίζουμε τους συντελεστές μερικής παλινδρόμησης και τη σταθερά, σύμφωνα με τις εξισώσεις (6) και (7). Τέλος, μετασχηματίζουμε την standardized εξίσωση γραμμικής παλινδρόμησης στην εξίσωση γενικής γραμμικής παλινδρόμησης, όπως φαίνεται στην εξίσωση (8).

$$b_i = b'_i (Lyy/Lx_i x_i)^{1/2} (i = 1, \dots, K \leq p) \quad (6)$$

$$b_0 = \bar{Y} - \sum b_i \bar{X}_i (i = 1, \dots, K \leq p) \quad (7)$$

$$\hat{y} = b_0 + \sum b_i X_i (i = 1, \dots, K \leq p) \quad (8)$$

b_i : είναι ο i -στος συντελεστής μερικής παλινδρόμησης της γενικής εξίσωσης γραμμικής παλινδρόμησης,

Lyy : το άθροισμα των τετραγώνων της εξαρτημένης μεταβλητής Y ,

$Lx_i x_i$: το άθροισμα των τετραγώνων της i -στης ανεξάρτητης μεταβλητής X_i ,

b_0 : η σταθερά της γενικής εξίσωσης γραμμικής παλινδρόμησης.

ΠΑΡΑΔΕΙΓΜΑ

Το παρακάτω σύνολο δεδομένων περιέχεται στο βιβλίο «Στατιστικά Μοντέλα Παλινδρόμησης» των Π. Οικονόμου και Χ. Καρώνη^[9].

Στον παρακάτω πίνακα παρουσιάζονται τα δεδομένα για την αντοχή (y) διάφορων τύπων τσιμέντων σε σχέση με τις ποσότητες των διάφορων τύπων συστατικών τους (x_1, x_2, x_3, x_4), όπου y είναι η εξαρτημένη μεταβλητή και x_1, x_2, x_3, x_4 είναι οι ανεξάρτητες μεταβλητές.

Y	X ₁	X ₂	X ₃	X ₄
78,5	7	26	6	60
74,3	1	29	15	52
104,3	11	56	8	20
87,6	11	31	8	47
95,9	7	52	6	33
109,2	11	55	9	22
102,7	3	71	17	6
72,5	1	31	22	44
93,1	2	54	18	22
115,9	21	47	4	26
83,8	1	40	23	34
113,2	11	66	9	12
109,4	10	68	8	12

Ο Πίνακας 2.1 εμφανίζει τη μέση τιμή και την τυπική απόκλιση (S.D.) όλων των μεταβλητών.

ΠΙΝΑΚΑΣ 2.1

	N	Mean	Std. Deviation
y	13	95,4154	15,03384
x1	13	7,4615	5,88239
x2	13	48,1538	15,56088
x3	13	11,7692	6,40513
x4	13	30,0000	16,73818
Valid N (listwise)	13		

Επιλέγουμε τις στατιστικά σημαντικές ανεξάρτητες μεταβλητές ($p < 0.05$) με το SPSS κι ελέγχουμε την πολυσυγγραμμικότητα για κάθε ανεξάρτητη μεταβλητή. Στο παράθυρο διαλόγου γραμμικής παλινδρόμησης SPSS, πληκτρολογούμε 'y' (εξαρτημένη μεταβλητή) στο "dependent box" και 'x₁, x₂, x₃, x₄' (όλες οι ανεξάρτητες μεταβλητές) στο "independent box" κι επιλέγουμε «backward» στον έλεγχο επιλογής

μεθόδου. Στο παράθυρο διαλόγου «linear regression» του SPSS επιλέγουμε: Descriptives → Covariance Matrix → Collinearity diagnostics.

Αφού το SPSS εκτελέσει τη διαδικασία γραμμικής παλινδρόμησης, λαμβάνουμε τα αποτελέσματα των Πινάκων 2.2 και 2.3.

ΠΙΝΑΚΑΣ 2.2

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	63,166	69,934		,903	,393		
	x1	1,543	,743	,604	2,076	,072	,026	38,496
	x2	,502	,722	,520	,695	,507	,004	254,423
	x3	,094	,753	,040	,125	,904	,021	46,868
	x4	-,152	,708	-,169	-,214	,836	,004	282,513

a. Dependent Variable: y

ΠΙΝΑΚΑΣ 2.3

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	x1	x2	x3	x4
1	1	4,120	1,000	,00	,00	,00	,00	,00
	2	,554	2,727	,00	,01	,00	,00	,00
	3	,289	3,778	,00	,00	,00	,00	,00
	4	,038	10,462	,00	,06	,00	,05	,00
	5	6,614E-5	249,578	1,00	,93	1,00	,95	1,00

a. Dependent Variable: y

Κατόπιν, ελέγχουμε αν υφίσταται πολυσυγγραμμικότητα μεταξύ των ανεξάρτητων μεταβλητών. Επίσης, ο πίνακας 2.2 εμφανίζει ότι ο δείκτης *Tolerance* των ανεξάρτητων μεταβλητών X_1 , X_2 , X_3 , X_4 λαμβάνει μικρές τιμές και συγκεκριμένα κάτω από 0,1 (0,026, 0,04, 0,021 και 0,004 αντίστοιχα) κι οι τιμές του δείκτη VIF για τις αντίστοιχες μεταβλητές είναι αρκετά μεγάλες. Εν συνεχεία, από τον πίνακα 2.3 φαίνεται ότι οι ιδιοτιμές των ανεξάρτητων μεταβλητών είναι κοντά στο μηδέν. Από όλα αυτά γίνεται αντιληπτό ότι υφίσταται το πρόβλημα της πολυσυγγραμμικότητας ανάμεσα στις μεταβλητές X_1 , X_2 , X_3 και X_4 . Συνεπώς, το επόμενο βήμα είναι να εφαρμόσουμε τη μέθοδο PCA (Ανάλυση Κύριων Συνιστωσών), ούτως ώστε να ξεπεράσουμε το πρόβλημα της πολυσυγγραμμικότητας.

Αρχικά, ξεκινούμε μετατρέποντας τις μεταβλητές ως “standardized” (Descriptive Statistics → Descriptives → Save standardized values as variables).

Στη συνέχεια χρησιμοποιούμε τη διαδικασία factor analysis του SPSS για να αποκτήσουμε τον πίνακα κύριων συνιστωσών των ανεξάρτητων μεταβλητών X_1 , X_2 , X_3 και X_4 και το δείκτη *cumulative variance proportion* των διαφορετικών κύριων συνιστωσών.

Κατόπιν, επιλέγουμε: Dimension Reduction → Factor, τοποθετούμε τις standardized ανεξάρτητες μεταβλητές και μετά στο “Extraction dialog box” επιλέγουμε Method → Principal Components and Factors to extract → 4, δηλαδή όσες είναι οι ανεξάρτητες μεταβλητές. Έπειτα, στο Factor Scores dialog box επιλέγουμε

“Save as variables”, Method ➔ Regression. Όλα τα αποτελέσματα φαίνονται στους Πίνακες 2.4, 2.5.

ΠΙΝΑΚΑΣ 2.4

Component	Total	Initial Eigenvalues		Total	Extraction Sums of Squared Loadings	
		% of Variance	Cumulative %		% of Variance	Cumulative %
1	2,236	55,893	55,893	2,236	55,893	55,893
2	1,576	39,402	95,294	1,576	39,402	95,294
3	,187	4,665	99,959	,187	4,665	99,959
4	,002	,041	100,000	,002	,041	100,000

Extraction Method: Principal Component Analysis.

ΠΙΝΑΚΑΣ 2.5

Component Matrix^a

	Component			
	1	2	3	4
x1	,712	-,639	,292	,010
x2	,843	,520	-,136	,026
x3	-,589	,759	,275	,011
x4	-,819	-,566	-,084	,027

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Ο πίνακας 2.4 δείχνει ότι η αθροιστική variance proportion της κύριας συνιστώσας C_1 είναι 55,893%, των C_1 και C_2 είναι 95,294% και των C_1 , C_2 και C_3 είναι 99,959% και των C_1 , C_2 και C_3 και C_4 είναι 100%. Επιπροσθέτως στον πίνακα 2.5 φαίνονται οι εκφράσεις των κύριων συνιστωσών:

$$C_1 = 0,712X'_1 + 0,843X'_2 - 0,589X'_3 - 0,819X'_4$$

$$C_2 = -0,639X'_1 + 0,520X'_2 + 0,759X'_3 - 0,566X'_4$$

$$C_3 = 0,292X'_1 - 0,136X'_2 + 0,275X'_3 - 0,084X'_4$$

$$C_4 = 0,010X'_1 + 0,026X'_2 + 0,011X'_3 + 0,027X'_4$$

Κατόπιν, στο παράθυρο διαλόγου descriptives του SPSS, πληκτρολογούμε τις μεταβλητές: y , x_1 , x_3 και x_4 , επιλέγουμε «Save standardized values as variables» και κάνουμε κλικ στο κουμπί OK για να δημιουργήσουμε την standardized

εξαρτημένη μεταβλητή zy και τις standardized ανεξάρτητες μεταβλητές zx_1 , zx_3 και zx_4 στο τρέχον αρχείο δεδομένων εργασίας.

Στο παράθυρο διαλόγου `compute variable` του SPSS πληκτρολογούμε c_1 ως το όνομα μεταβλητής της πρώτης κύριας συνιστώσας C_1 , κατόπιν πληκτρολογούμε $0,712 * zx_1 + 0,843 * zx_2 - 0,589 * zx_3 - 0,819 * zx_4$ στο πλαίσιο `numeric expression` κι επιλέγουμε το κουμπί OK για να δημιουργήσουμε μια νέα μεταβλητή c_1 και την τιμή της στο τρέχον αρχείο δεδομένων εργασίας. Αφού υπολογίσουμε τη δεύτερη C_2 , την τρίτη κύρια συνιστώσα C_3 και την τέταρτη κύρια συνιστώσα C_4 πληκτρολογούμε « c_2 », « c_3 » και « c_4 » στα `target boxes` και " $-0,639 * zx_1 + 0,520zx_2 + 0,759 * zx_3 - 0,566 * zx_4$ ", " $0,292*zx_1-0,136*zx_2+0,275*zx_3 - 0,084zx_4$ " και " $0,010zx_1 + 0,026zx_2 + 0,011zx_3 + 0,027zx_4$ " στο `numeric expression box` αντίστοιχα. Μετά την εκτέλεση της διαδικασίας μεταβλητής υπολογισμών SPSS μία προς μία, δημιουργούμε τις νέες μεταβλητές c_2 , c_3 και c_4 και τις τιμές τους στο τρέχον αρχείο δεδομένων εργασίας.

Εν συνεχεία, χρησιμοποιούμε τη διαδικασία γραμμικής παλινδρόμησης του SPSS για να κάνουμε την ανάλυση παλινδρόμησης των standardized κύριων συνιστωσών, η οποία περιλαμβάνει την κατασκευή της εξίσωσης παλινδρόμησης για κάθε standardized κύρια συνιστώσα, τον έλεγχο αν όλες οι κύριες συνιστώσες είναι ανεξάρτητες μεταξύ τους και τον προσδιορισμό της «καλύτερης» εξίσωσης παλινδρόμησης των τυποποιημένων (standardized) κύριων συνιστωσών.

Στο παράθυρο διαλόγου γραμμικής παλινδρόμησης (linear regression) του SPSS, πληκτρολογούμε " zy " και " c_1 " στα `dependent` και `independent boxes` αντίστοιχα. Στο παράθυρο `linear regression` του SPSS: επιλέγουμε τα `Covariance matrix` και `Collinearity diagnostics`. Έτσι, παράγουμε την εξίσωση παλινδρόμησης της πρώτης τυποποιημένης (standardized) κύριας συνιστώσας: $\hat{y}'_1 = B'_1 C_1$. Ακολουθώντας τα ίδια βήματα, προσαρμόζουμε τις εξισώσεις: $\hat{y}'_2 = B'_1 C_1 + B'_2 C_2$ και $\hat{y}'_3 = B'_1 C_1 + B'_2 C_2 + B'_3 C_3$. Μετά την εκτέλεση της διαδικασίας γραμμικής παλινδρόμησης του SPSS, όλα τα αποτελέσματα παρουσιάζονται στους Πίνακες 2.6 έως 2.8, αντίστοιχα.

ΠΙΝΑΚΑΣ 2.6

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	4,037E-16	,054		,000	1,000		
	c1	,439	,025	,982	17,407	,000	1,000	1,000
2	(Constant)	4,024E-16	,057		,000	1,000		
	c1	,439	,026	,982	16,622	,000	1,000	1,000
	c2	,007	,038	,010	,175	,864	1,000	1,000
3	(Constant)	3,768E-16	,043		,000	1,000		
	c1	,439	,020	,982	22,085	,000	1,000	1,000
	c2	,007	,028	,011	,246	,811	1,000	1,000
	c3	,701	,238	,131	2,941	,016	1,000	1,000
4	(Constant)	3,728E-16	,045		,000	1,000		
	c1	,438	,022	,978	20,159	,000	,933	1,072
	c2	,006	,030	,009	,196	,849	,988	1,012
	c3	,695	,252	,130	2,754	,025	,993	1,007
	c4	9,311	28,851	,016	,323	,755	,916	1,091

a. Dependent Variable: Zscore(y)

ΠΙΝΑΚΑΣ 2.7

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,982 ^a	,965	,962	,19549431
2	,982 ^b	,965	,958	,20472160
3	,991 ^c	,982	,976	,15408815
4	,991 ^d	,982	,974	,16238159

ΠΙΝΑΚΑΣ 2.8

Collinearity Diagnostics ^a								
Model	Dimension	Eigenvalue	Condition Index	(Constant)	c1	c2	c3	c4
1	1	1,000	1,000	1,00	,00			
	2	1,000	1,000	,00	1,00			
2	1	1,000	1,000	,00	,50	,50		
	2	1,000	1,000	1,00	,00	,00		
	3	1,000	1,000	,00	,50	,50		
3	1	1,005	1,000	,00	,00	,50	,50	
	2	1,000	1,002	,00	,99	,01	,00	
	3	1,000	1,002	1,00	,00	,00	,00	
	4	,995	1,005	,00	,00	,50	,50	
4	1	1,289	1,000	,00	,29	,05	,03	,36
	2	1,004	1,133	,00	,00	,46	,52	,00
	3	1,000	1,135	1,00	,00	,00	,00	,00
	4	,997	1,137	,00	,20	,39	,40	,00
	5	,710	1,347	,00	,51	,10	,06	,64

a. Dependent Variable: Zscore(y)

Στον Πίνακα 2.6 παρουσιάζονται όλοι οι τυποποιημένοι (standardized) συντελεστές μερικής παλινδρόμησης B'_i με μεγάλη στατιστική σημαντικότητα ($P < 0.005$) από όλες τις κύριες συνιστώσες C_i από όλα τα μοντέλα (εξισώσεις) για τη δημιουργία τεσσάρων εξισώσεων παλινδρόμησης των τυποποιημένων κύριων μεταβλητών, οι οποίες είναι:

$$\hat{y}'_1 = 0,439 * C_1 ,$$

$$\hat{y}'_2 = 0,439 * C_1 + 0,07 * C_2,$$

$$\hat{y}'_3 = 0,439 * C_1 + 0,07 * C_2 + 0,701 * C_3 \text{ και}$$

$$\hat{y}'_4 = 0,438 * C_1 + 0,06 * C_2 + 0,695 * C_3 + 9,311 * C_4.$$

Ο Πίνακας 2.6 παρουσιάζει ότι οι δείκτες tolerance και VIF είναι κοντά στο 0 και στο 1, αντίστοιχα. Ο Πίνακας 2.8 δείχνει ότι οι ιδιοτιμές τους κι ο δείκτης condition index είναι κοντά στο 1. Αυτά υποδηλώνουν ότι όλες οι κύριες συνιστώσες είναι ανεξάρτητες μεταξύ τους.

Ο συντελεστής R^2 είναι ένα μέτρο της καλής προσαρμογής ενός γραμμικού μοντέλου και παίρνει τιμές από 0 έως 1. Όσο πιο κοντά στο 1 είναι η τιμή του R^2 , τόσο καλύτερο είναι το γραμμικό μοντέλο. Όμως, καθώς το R^2 επηρεάζεται από τον αριθμό των ανεξάρτητων μεταβλητών του μοντέλου και από το μέγεθος του δείγματος, συνήθως χρησιμοποιούμε το προσαρμοσμένο (adjusted) R^2 όταν συγκρίνουμε τις συνθήκες καλής προσαρμογής μεταξύ διαφορετικών γραμμικών μοντέλων. Το προσαρμοσμένο (adjusted) R^2 έχει σχεδιαστεί για να αντισταθμίζει την αισιόδοξη μεροληψία του R^2 . Το τυπικό σφάλμα εκτιμήτριας είναι η τετραγωνική ρίζα του μέσου τετραγώνου του υπολοίπου (residual) και μετρά την εξάπλωση των υπολοίπων γύρω από την προσαρμοσμένη ευθεία (fitted line). Έτσι είναι ένα

επιπλέον μέτρο της καλής προσαρμογής ενός γραμμικού μοντέλου. Όπως φαίνεται από τον Πίνακα 2.7 η καλύτερη εξίσωση είναι η εξής:

$$\hat{y}'_3 = 0,439 * C_1 + 0,07 * C_2 + 0,701 * C_3,$$

διότι έχει το μεγαλύτερο προσαρμοσμένο συντελεστή R^2 και το μικρότερο τυπικό σφάλμα εκτιμήτριας (standard error of estimate).

Χρησιμοποιώντας τη διαδικασία bivariate correlations του SPSS υπολογίζουμε το άθροισμα των τετραγώνων της εξαρτημένης μεταβλητής Y (L_{yy}) και το άθροισμα των τετραγώνων της i -στης ανεξάρτητης μεταβλητής X_i ($L_{x_i x_i}$). Στο παράθυρο διαλόγου bivariate correlations του SPSS, εισάγουμε στο πλαίσιο των μεταβλητών " y, x_1, x_2, x_3 και x_4 " (η εξαρτημένη μεταβλητή Y και οι ανεξάρτητες μεταβλητές X_1, X_3 και X_4). Επιλέγουμε τις εντολές Cross-product deviation και covariance. Έτσι παίρνουμε $L_{yy} = 2712,197$, $L_{x_1 x_1} = 415,231$, $L_{x_2 x_2} = 2905,692$, $L_{x_3 x_3} = 492,308$ και $L_{x_4 x_4} = 3362$.

Κατόπιν, μετασχηματίζουμε την «καλύτερη» εξίσωση παλινδρόμησης της standardized κύριας συνιστώσας στην standardized γραμμική εξίσωση παλινδρόμησης κι έπειτα στη γενική γραμμική εξίσωση παλινδρόμησης. Από τον πίνακα 2.5 παίρνουμε:

$$\begin{aligned} C_1 &= 0,712X'_1 + 0,843X'_2 - 0,589X'_3 - 0,819X'_4, \\ C_2 &= -0,639X'_1 + 0,520X'_2 + 0,759X'_3 - 0,566X'_4, \\ C_3 &= 0,292X'_1 - 0,136X'_2 + 0,275X'_3 - 0,084X'_4 \text{ και} \\ C_4 &= 0,010X'_1 + 0,026X'_2 + 0,011X'_3 + 0,027X'_4 \end{aligned}$$

Αυτές τις εφαρμόζουμε στην «καλύτερη» εξίσωση παλινδρόμησης της standardized κύριας συνιστώσας:

$$\hat{y}'_3 = 0,439 * C_1 + 0,07 * C_2 + 0,701 * C_3.$$

Κατόπιν, παίρνουμε την standardized γραμμική εξίσωση παλινδρόμησης:

$$\hat{y}' = 1,22 * X'_1 + 0,356 * X'_2 - 0,578 * X'_3 + 0,397 * X'_4 + 0,192 * X'_5.$$

Τέλος, υπολογίζουμε τους γενικούς συντελεστές μερικής παλινδρόμησης b_i και τη σταθερά b_0 , όπου $b_0 = \bar{Y} - \sum b_i \bar{X}_i$ και $b_i = b'_i * \sqrt{\frac{L_{yy}}{L_{x_i x_i}}}$ και καταλήγουμε στη γενική γραμμική εξίσωση παλινδρόμησης, η οποία είναι:

$$\hat{y} = 91,12 + 1,208 * X_1 + 0,301 * X_2 - 0,030 * X_3 - 0,411 * X_4.$$

ΚΕΦΑΛΑΙΟ 3^ο ΧΡΗΣΗ ΤΗΣ ΜΕΘΟΔΟΥ PCA ΣΕ ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ

Το παρακάτω σύνολο δεδομένων περιέχεται στο βιβλίο “Data Mining: Concepts, Models and Techniques” των Gorunescu, Florin^[4].

Στον παρακάτω πίνακα παρουσιάζονται τα δεδομένα για την πρόβλεψη του δείκτη PEmax (Y) συναρτήσει του ύψους (X_1), του βάρους (X_2), του BMP (X_3), του FEV₁ (X_4), της ηλικίας (X_5), RV (X_6), του FRC (X_7) (Functional Residual Capacity) και του TLC (X_8) (Total Lung Capacity)

PEmax	HEIGHT	WEIGHT	BMP	FEV ₁	AGE	RV	FRC	TRC
95	109	13,10	68	32	7	258	183	137
85	112	12,90	65	19	7	449	285	134
100	124	14,10	64	22	8	441	268	147
85	125	16,20	67	41	8	234	146	124
95	127	21,50	93	52	8	202	131	104
80	130	17,50	68	44	9	308	155	118
65	139	30,70	89	28	11	305	179	119
110	150	28,40	69	18	12	369	198	103
70	146	25,10	67	24	12	312	194	128
95	155	31,50	68	23	13	413	225	136
110	156	39,90	89	39	13	206	142	95
90	153	42,10	90	26	14	253	191	121
100	160	45,60	93	45	14	174	139	108
80	158	51,20	93	45	15	158	124	90
134	160	35,90	66	31	16	302	133	101
134	153	34,80	70	39	17	204	118	120
165	174	44,70	70	49	17	187	104	103
120	176	60,10	92	29	17	188	129	130
130	171	42,60	69	38	17	172	130	103
85	156	37,20	72	21	19	216	119	81

PEmax: ο δείκτης αντοχής των αναπνευστικών μυών, που εκφράζεται από τη μέγιστη στατική εκπνευστική πίεση.

BMP: Body Mass Percentage

FEV₁: Force Expiratory Volume in 1 second

RV: Residual Volume

FRC: Functional Residual Capacity

TLC: Total Lung Capacity

Ο Πίνακας 3.1 εμφανίζει τη μέση τιμή και την τυπική απόκλιση (S.D.) όλων των μεταβλητών.

ΠΙΝΑΚΑΣ 3.1

Descriptive Statistics			
	N	Mean	Std. Deviation
PEmax	20	101,4000	24,89219
HEIGHT	20	146,7000	19,66322
WEIGHT	20	32,2550	13,58240
BMP	20	76,1000	11,61170
FEV1	20	33,2500	10,63200
AGE	20	12,7000	3,85391
RV	20	267,5500	91,64202
FRC	20	164,6500	50,17157
TLC	20	115,1000	17,54663
Valid N (listwise)	20		

Επιλέγουμε τις στατιστικά σημαντικές ανεξάρτητες μεταβλητές ($p < 0.05$) με το SPSS κι ελέγχουμε την πολυσυγγραμμικότητα για κάθε ανεξάρτητη μεταβλητή εφαρμόζοντας τα βήματα που κάναμε και στο παραπάνω παράδειγμα. Αφού το SPSS εκτελέσει τη διαδικασία γραμμικής παλινδρόμησης, λαμβάνουμε τα αποτελέσματα των Πινάκων 3.2 και 3.3.

ΠΙΝΑΚΑΣ 3.2

Coefficients ^a								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	47,502	182,513		,260	,799		
	HEIGHT	,354	1,078	,280	,329	,749	,052	19,095
	WEIGHT	,892	1,965	,487	,454	,659	,033	30,271
	BMP	-1,356	,977	-,633	-1,389	,192	,183	5,466
	FEV1	1,196	,946	,511	1,265	,232	,233	4,295
	AGE	,540	5,108	,084	,106	,918	,061	16,473
	RV	,032	,178	,118	,180	,861	,088	11,372
	FRC	,008	,344	,016	,023	,982	,079	12,662
	TLC	,173	,431	,122	,401	,696	,412	2,425

ΠΙΝΑΚΑΣ 3.3

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions								
				(Constant)	HEIGHT	WEIGHT	BMP	FEV1	AGE	RV	FRC	TLC
1	1	8,478	1,000	,00	,00	,00	,00	,00	,00	,00	,00	,00
	2	,358	4,869	,00	,00	,00	,00	,00	,00	,01	,00	,00
	3	,111	8,745	,00	,00	,01	,00	,11	,01	,00	,00	,00
	4	,028	17,302	,00	,00	,03	,04	,03	,05	,02	,02	,00
	5	,011	27,302	,00	,00	,00	,03	,04	,00	,15	,00	,56
	6	,008	32,416	,01	,00	,15	,15	,28	,05	,09	,01	,00
	7	,004	45,509	,00	,00	,01	,06	,12	,06	,46	,88	,20
	8	,001	80,613	,08	,35	,00	,15	,41	,73	,21	,00	,12
	9	,000	144,297	,90	,65	,80	,57	,01	,09	,07	,10	,13

Από τον πίνακα 3.2 παρατηρούμε ότι ο δείκτης *tolerance* των ανεξάρτητων μεταβλητών X_1, X_2, X_4, X_5 και X_6 παίρνει τιμές μικρότερες από 0,1 κι οι τιμές του δείκτη VIF για τις αντίστοιχες μεταβλητές είναι αρκετά μεγάλες. Επίσης, από τον πίνακα 3.3 φαίνεται ότι υπάρχουν ιδιοτιμές, οι οποίες είναι κοντά στο μηδέν κι ότι ο δείκτης condition index για κάποιες από αυτές παίρνει τιμές μεγαλύτερες από 30. Βρίσκοντας το συντελεστή συσχέτισης του *Pearson* με τη βοήθεια του SPSS, για όλες τις ανεξάρτητες μεταβλητές, παρατηρούμε, με τη βοήθεια του ΠΙΝΑΚΑ 3.4, ότι υπάρχουν αρκετές στατιστικά συσχετισμένες ανεξάρτητες μεταβλητές ($p\text{-value} < 0,05$). Επομένως, εφαρμόζουμε τη μέθοδο PCA (Ανάλυση Κύριων Συνιστωσών), για να ξεπεράσουμε το πρόβλημα της πολυσυγγραμμικότητας.

ΠΙΝΑΚΑΣ 3.4

Correlations

		HEIGHT	WEIGHT	BMP	FEV1	AGE	RV	FRC	TLC
HEIGHT	Pearson Correlation	1	,917**	,304	,151	,918**	-,518*	-,589**	-,507*
	Sig. (2-tailed)		,000	,192	,524	,000	,019	,006	,023
	N	20	20	20	20	20	20	20	20
WEIGHT	Pearson Correlation	,917**	1	,585**	,213	,848**	-,640**	-,599**	-,493*
	Sig. (2-tailed)	,000		,007	,367	,000	,002	,005	,027
	N	20	20	20	20	20	20	20	20
BMP	Pearson Correlation	,304	,585**	1	,376	,183	-,557*	-,368	-,333
	Sig. (2-tailed)	,192	,007		,102	,440	,011	,111	,152
	N	20	20	20	20	20	20	20	20
FEV1	Pearson Correlation	,151	,213	,376	1	,057	-,705**	-,687**	-,368
	Sig. (2-tailed)	,524	,367	,102		,811	,001	,001	,111
	N	20	20	20	20	20	20	20	20
AGE	Pearson Correlation	,918**	,848**	,183	,057	1	-,536*	-,626**	-,568**
	Sig. (2-tailed)	,000	,000	,440	,811		,015	,003	,009
	N	20	20	20	20	20	20	20	20
RV	Pearson Correlation	-,518*	-,640**	-,557*	-,705**	-,536*	1	,906**	,599**
	Sig. (2-tailed)	,019	,002	,011	,001	,015		,000	,005
	N	20	20	20	20	20	20	20	20
FRC	Pearson Correlation	-,589**	-,599**	-,368	-,687**	-,626**	,906**	1	,697**
	Sig. (2-tailed)	,006	,005	,111	,001	,003	,000		,001
	N	20	20	20	20	20	20	20	20
TLC	Pearson Correlation	-,507*	-,493*	-,333	-,368	-,568**	,599**	,697**	1
	Sig. (2-tailed)	,023	,027	,152	,111	,009	,005	,001	

Αρχικά, ξεκινούμε μετατρέποντας τις μεταβλητές ως “standardized” ακολουθώντας την ίδια πορεία με το προηγούμενο παράδειγμα.

Κατόπιν, χρησιμοποιούμε τη διαδικασία factor analysis του SPSS προκειμένου να αποκτήσουμε τον πίνακα κύριων συνιστωσών των ανεξάρτητων μεταβλητών και το δείκτη *cumulative variance proportion* των διαφορετικών κύριων συνιστωσών. Τα αποτελέσματα φαίνονται στους Πίνακες 3.5, 3.6.

ΠΙΝΑΚΑΣ 3.5

Total Variance Explained						
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,828	60,356	60,356	4,828	60,356	60,356
2	1,506	18,827	79,183	1,506	18,827	79,183
3	,842	10,526	89,708	,842	10,526	89,708
4	,498	6,220	95,928		6,220	95,928
5	,208	2,596	98,524		2,596	98,524
6	,058	,728	99,252	,058	,728	99,252
7	,040	,503	99,756	,040	,503	99,756
8	,020	,244	100,000	,020	,244	100,000

Extraction Method: Principal Component Analysis.

ΠΙΝΑΚΑΣ 3.6

Component Matrix^a

	Component							
	1	2	3	4	5	6	7	8
Zscore(HEIGHT)	,823	,500	,005	,153	,183	-,031	-,097	,068
Zscore(WEIGHT)	,875	,355	,285	,104	,047	,058	-,021	-,105
Zscore(BMP)	,569	-,275	,748	-,187	-,014	-,063	,034	,032
Zscore(FEV1)	,542	-,748	-,120	,237	,269	,029	,055	-,002
Zscore(AGE)	,805	,542	-,163	,060	-,063	,006	,152	,027
Zscore(RV)	-,880	,358	,023	-,105	,255	-,133	,046	-,035
Zscore(FRC)	-,890	,266	,282	-,067	,144	,173	,026	,032
Zscore(TLC)	-,742	,060	,284	,594	-,099	-,038	,013	,006

Extraction Method: Principal Component Analysis.

a. 8 components extracted.

Ο πίνακας 3.5 δείχνει ότι η αθροιστική variance proportion της κύριας συνιστώσας C_1 είναι 60,356%, των C_1 και C_2 είναι 79,183% και των C_1 , C_2 και C_3 είναι 89,708% , των C_1 , C_2 , C_3 και C_4 είναι 95,298%, των C_1 , C_2 , C_3 , C_4 και C_5 είναι 98,524%, των C_1 , C_2 , C_3 , C_4 , C_5 και C_6 είναι 99,252%, των C_1 , C_2 , C_3 , C_4 , C_5 , C_6 και C_7 είναι 99,756% και όλων των ανεξάρτητων μεταβλητών είναι 100%.

Επιπροσθέτως στον πίνακα 3.6 φαίνονται οι εκφράσεις των κύριων συνιστωσών:

$$C_1 = 0,823X'_1 + 0,875X'_2 - 0,569X'_3 - 0,542X'_4 + 0,805X'_5 - 0,88X'_6 - 0,89X'_7 - 0,742X'_8$$

$$C_2 = 0,5X'_1 + 0,355X'_2 - 0,275X'_3 - 0,748X'_4 + 0,542X'_5 + 0,358X'_6 + 0,266X'_7 + 0,060X'_8$$

$$C_3 = 0,005X'_1 + 0,285X'_2 + 0,748X'_3 - 0,12X'_4 - 0,163X'_5 + 0,023X'_6 + 0,2668 + 0,284X'_8$$

$$C_4 = 0,153X'_1 + 0,104X'_2 - 0,187X'_3 + 0,237X'_4 + 0,06X'_5 - 0,105X'_6 - 0,067X'_7 + 0,28X'_8$$

$$C_5 = 0,183X'_1 + 0,047X'_2 - 0,014X'_3 + 0,269X'_4 - 0,063X'_5 + 0,25X'_6 + 0,14X'_7 - 0,099X'_8$$

$$C_6 = -0,03X'_1 + 0,058X'_2 - 0,063X'_3 - 0,029X'_4 + 0,006X'_5 - 0,13X'_6 - 0,17X'_7 - 0,038X'_8$$

$$C_7 = -0,097X'_1 - 0,02X'_2 + 0,03X'_3 + 0,055X'_4 + 0,15X'_5 + 0,046X'_6 + 0,026X'_7 + 0,013X'_8$$

$$C_8 = 0,068X'_1 - 0,105X'_2 + 0,03X'_3 - 0,002X'_4 + 0,027X'_5 - 0,035X'_6 + 0,03X'_7 + 0,006X'_8$$

όπου X'_1, X'_2, \dots, X'_8 είναι οι standardized ανεξάρτητες μεταβλητές.

Κατόπιν, στο παράθυρο διαλόγου descriptives του SPSS, πληκτρολογούμε τις μεταβλητές: (εξαρτημένη μεταβλητή Y και ανεξάρτητες μεταβλητές X_1, \dots, X_8), επιλέγουμε «Save standardized values as variables» και κάνουμε κλικ στο κουμπί OK για να δημιουργήσουμε την τυποποιημένη (standardized) εξαρτημένη μεταβλητή zy και τις τυποποιημένες (standardized) ανεξάρτητες μεταβλητές zx_1, zx_3 και zx_4 στο τρέχον αρχείο δεδομένων εργασίας. Στο παράθυρο διαλόγου compute variable του SPSS πληκτρολογούμε c_1 ως το όνομα μεταβλητής της πρώτης κύριας συνιστώσας C_1 , κατόπιν πληκτρολογούμε " $0,823zx_1 + 0,875zx_2 - 0,569zx_3 - 0,542zx_4 + 0,805zx_5 - 0,88zx_6 - 0,89zx_7 - 0,742zx_8$ " στο πλαίσιο numeric expression κι έτσι δημιουργούμε μια νέα μεταβλητή c_1 . Ομοίως πράττουμε για όλες τις κύριες συνιστώσες $C_i, i=1,2,\dots,8$. Μετά την εκτέλεση της διαδικασίας μεταβλητής υπολογισμών SPSS μία προς μία, δημιουργούμε τις νέες μεταβλητές c_1, c_2, \dots, c_8 και τις τιμές τους στο τρέχον αρχείο δεδομένων εργασίας.

ΠΙΝΑΚΑΣ 3.7

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3,558	1	3,558	4,148	,057 ^b
	Residual	15,442	18	,858		
	Total	19,000	19			
2	Regression	4,346	2	2,173	2,521	,110 ^c
	Residual	14,654	17	,862		
	Total	19,000	19			
3	Regression	6,901	3	2,300	3,042	,059 ^d
	Residual	12,099	16	,756		
	Total	19,000	19			
4	Regression	9,895	4	2,474	4,075	,020 ^e
	Residual	9,105	15	,607		
	Total	19,000	19			
5	Regression	10,946	5	2,189	3,806	,022 ^f
	Residual	8,054	14	,575		
	Total	19,000	19			
6	Regression	11,007	6	1,835	2,984	,047 ^g
	Residual	7,993	13	,615		
	Total	19,000	19			
7	Regression	11,009	7	1,573	2,362	,091 ^h
	Residual	7,991	12	,666		
	Total	19,000	19			
8	Regression	11,064	8	1,383	1,917	,157 ⁱ
	Residual	7,936	11	,721		

ΠΙΝΑΚΑΣ 3.8

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-7,100E-17	,207		,000	1,000		
	c1	,090	,044	,433	2,037	,057	1,000	1,000
2	(Constant)	-2,299E-17	,208		,000	1,000		
	c1	,090	,044	,433	2,032	,058	1,000	1,000
	c2	,135	,141	,204	,956	,352	1,000	1,000
3	(Constant)	-5,981E-16	,194		,000	1,000		
	c1	,090	,041	,433	2,170	,045	1,000	1,000
	c2	,135	,132	,203	1,020	,323	1,000	1,000
	c3	-,435	,237	-,367	-1,838	,085	1,000	1,000
4	(Constant)	-9,805E-16	,174		,000	1,000		
	c1	,090	,037	,434	2,429	,028	1,000	1,000
	c2	,135	,119	,203	1,137	,273	1,000	1,000
	c3	-,435	,212	-,366	-2,050	,058	1,000	1,000
	c4	,798	,359	,397	2,221	,042	1,000	1,000
5	(Constant)	-7,038E-16	,170		,000	1,000		
	c1	,090	,036	,434	2,496	,026	1,000	1,000
	c2	,135	,115	,203	1,165	,263	1,000	1,000
	c3	-,435	,207	-,367	-2,107	,054	1,000	1,000
	c4	,797	,350	,397	2,280	,039	1,000	1,000
	c5	1,132	,838	,235	1,352	,198	1,000	1,000
6	(Constant)	-7,533E-16	,175		,000	1,000		
	c1	,090	,037	,433	2,408	,032	1,000	1,000
	c2	,134	,119	,203	1,126	,280	1,000	1,000
	c3	-,435	,214	-,366	-2,035	,063	1,000	1,000
	c4	,797	,362	,397	2,205	,046	1,000	1,000
	c5	1,132	,866	,235	1,307	,214	1,000	1,000
	c6	,972	3,088	,057	,315	,758	1,000	1,000
7	(Constant)	-7,851E-16	,182		,000	1,000		
	c1	,090	,039	,433	2,314	,039	1,000	1,000
	c2	,134	,124	,203	1,082	,300	1,000	1,000
	c3	-,435	,222	-,366	-1,956	,074	1,000	1,000
	c4	,797	,376	,397	2,118	,056	1,000	1,000
	c5	1,132	,901	,235	1,256	,233	1,000	1,000
	c6	,972	3,214	,057	,302	,768	1,000	1,000
	c7	-,275	4,659	-,011	-,059	,954	1,000	1,000
8	(Constant)	-1,320E-15	,190		,000	1,000		
	c1	,090	,040	,436	2,233	,047	,998	1,002
	c2	,135	,129	,204	1,045	,318	1,000	1,000
	c3	-,434	,231	-,366	-1,877	,087	1,000	1,000
	c4	,796	,392	,396	2,031	,067	1,000	1,000
	c5	1,131	,938	,235	1,206	,253	1,000	1,000
	c6	,967	3,346	,056	,289	,778	1,000	1,000
	c7	-,275	4,850	-,011	-,057	,956	1,000	1,000
	c8	-2,724	9,934	-,054	-,274	,789	,997	1,003

ΠΙΝΑΚΑΣ 3.9

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,433 ^a	,187	,142	,92621332
2	,478 ^b	,229	,138	,92842422
3	,603 ^c	,363	,244	,86959051
4	,722 ^d	,521	,393	,77909777
5	,759 ^e	,576	,425	,75846628
6	,761 ^f	,579	,385	,78411831
7	,761 ^g	,579	,334	,81601753
8	,763 ^h	,582	,279	,84940473

Στον Πίνακα 3.8 παρουσιάζονται όλοι οι τυποποιημένοι (standardized) συντελεστές μερικής παλινδρόμησης B'_i με μεγάλη στατιστική σημαντικότητα, ούτως ώστε να δημιουργήσουμε τις εξισώσεις παλινδρόμησης των τυποποιημένων (standardized) κύριων μεταβλητών, οι οποίες είναι:

$$\hat{y}'_1 = 0,090 * C_1 ,$$

$$\hat{y}'_2 = 0,090 * C_1 + 0,135 * C_2,$$

$$\hat{y}'_3 = 0,090 * C_1 + 0,135 * C_2 - 0,435 * C_3$$

$$\hat{y}'_4 = 0,090 * C_1 + 0,135 * C_2 - 0,435 * C_3 + 0,798 * C_4$$

$$\hat{y}'_5 = 0,090 * C_1 + 0,134 * C_2 - 0,435 * C_3 + 0,797 * C_4 + 1,132 * C_5$$

$$\begin{aligned} \hat{y}'_6 &= 0,090 * C_1 + 0,135 * C_2 - 0,435 * C_3 + 0,797 * C_4 + 1,132 * C_5 + 0,972 \\ &\quad * C_6 \hat{y}'_7 \\ &= 0,090C_1 + 0,135C_2 - 0,435C_3 + 0,797C_4 + 1,132C_5 + 0,972C_6 \\ &\quad - 0,275C_7 \end{aligned}$$

$$\hat{y}'_8 = 0,09C_1 + 0,13C_2 - 0,435C_3 + 0,797C_4 + 1,13C_5 + 0,97C_6 - 0,27C_7 - 2,27C_8$$

Ο Πίνακας 3.8 παρουσιάζει ότι οι δείκτες tolerance και VIF είναι κοντά στο 0 και στο 1 αντίστοιχα. Αυτό σημαίνει ότι όλες οι κύριες συνιστώσες είναι ανεξάρτητες μεταξύ τους.

Όπως φαίνεται από τον Πίνακα 3.9 η καλύτερη εξίσωση είναι η $\hat{y}'_5 = 0,090 * C_1 + 0,135 * C_2 - 0,435 * C_3 + 0,797 * C_4 + 1,132 * C_5$, διότι έχει το μεγαλύτερο προσαρμοσμένο συντελεστή R^2 και το μικρότερο τυπικό σφάλμα εκτιμήτριας.

Κατόπιν, υπολογίζουμε το άθροισμα των τετραγώνων της εξαρτημένης μεταβλητής Y (L_{yy}) και το άθροισμα των τετραγώνων της i -στης ανεξάρτητης μεταβλητής X_i ($L_{x_i x_i}$), τα οποία είναι τα εξής :

$$L_{yy} = 11772,8, \quad L_{x_1 x_1} = 7346,2, \quad L_{x_2 x_2} = 3505,15, \quad L_{x_3 x_3} = 2561,8, \\ L_{x_4 x_4} = 2147,75, \quad L_{x_5 x_5} = 282,2, \quad L_{x_6 x_6} = 159566,95, \quad L_{x_7 x_7} = 47826,55, \\ L_{x_8 x_8} = 5849,8.$$

Εν συνεχεία, μετασχηματίζουμε την «καλύτερη» εξίσωση παλινδρόμησης της standardized κύριας συνιστώσας στην standardized γραμμική εξίσωση παλινδρόμησης κι έπειτα στη γενική γραμμική εξίσωση παλινδρόμησης. Αυτές τις εφαρμόζουμε στην «καλύτερη» εξίσωση παλινδρόμησης της standardized κύριας συνιστώσας:

$$\hat{y}'_5 = 0,090 * C_1 + 0,134 * C_2 - 0,435 * C_3 + 0,797 * C_4 + 1,132 * C_5.$$

Κατόπιν, παίρνουμε την standardized γραμμική εξίσωση παλινδρόμησης:

$$\hat{y}' = 0,4752 * X'_1 + 0,3111 * X'_2 - 0,01267 * X'_3 - 0,45805 * X'_4$$

Τέλος, υπολογίζουμε τους γενικούς συντελεστές μερικής παλινδρόμησης b_i , τη σταθερά b_0 και καταλήγουμε στη γενική γραμμική εξίσωση παλινδρόμησης, η οποία είναι η εξής:

$$\hat{y} = -7,144 + 0,114 * X_1 + 0,247 * X_2 - 0,933 * X_3 + 1,866 * X_4 + 7,308 * X_5.$$

ΣΥΜΠΕΡΑΣΜΑ

Στο παραπάνω σύνολο ιατρικών δεδομένων εντοπίσαμε με τη βοήθεια των διάφορων δεικτών το πρόβλημα της πολυσυγγραμμικότητας και το αντιμετωπίσαμε εφαρμόζοντας τη μέθοδο PCA. Παρόλο που, αρχικά, υπήρχαν κάποιες μεταβλητές αρκετά συσχετισμένες μεταξύ τους, καταλήξαμε σε ένα νέο σύνολο, στο οποίο οι νέες συνιστώσες δεν είναι συσχετισμένες μεταξύ τους.

Από όλα τα παραπάνω διαφαίνεται ότι η μέθοδος PCA είναι μια αποτελεσματική μέθοδος, η οποία όχι μόνο μπορεί να διαγνώσει το πρόβλημα της πολυσυγγραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών, αλλά και να το επιλύσει. Ο χειρισμός της μεθόδου αυτής με τη βοήθεια του SPSS είναι αρκετά εύχρηστος κι απλός.

REFERENCES

1. Chatterjee Samprit and Ali S. Hadi, *Regression Analysis by Examples*, Wiley.(2006)
2. Draper, N. και Smith, H. (1997). *Εφαρμοσμένη Ανάλυση Παλινδρόμησης. 2ηΑγγλική Έκδοση. Μετάφραση-Επιμέλεια: Ε. Χατζηκωνσταντινίδης, Α. Καλαματιανού. Εκδόσεις Παπαζήση.*
3. Edwards Allen L. *Multiple Regression and the Analysis of Variance and Covariance*, (1979)
4. Gorunescu, Florin, *Data Mining: Concepts, Models and Techniques* (2011)
5. Liu R.X., Kuang J., Gong Q., Hou X.L., *Principal component regression analysis with SPSS. Computer Methods and Programs in Biomedicine* 71(2003), 141-14.
6. Rawlings John O., Sastry G. Pantula, David A. Dickey, *Applied Regression Analysis: A Research Tool, Second Edition*, Springer(1998)
7. Seber George A.F. and Lee Alan J., *Linear Regression Analysis*, 2nd ed., Wiley. (2003)
8. Zografos K., *Multivariate Analysis. University of Ioannina* (2007) .
9. Π. Οικονόμου , Χ. Καρώνη, *Στατιστικά Μοντέλα Παλινδρόμησης, Εκδόσεις ΣΥΜΕΩΝ*, (2010)