SCHOOL OF MEDICINE

UNIVERSITY OF THESSALY

POSTGRADUATE PROGRAMME (MSC)

«RESEARCH METHODOLOGY IN BIOMEDICINE, BIOSTATISTICS AND CLINICAL BIOINFORMATICS AT UNIVERSITY OF THESSALY»

Master's Thesis :

'' Multicolinearity : Diagnostics and ridge regression as a method of handling ''

Πολυσυγγραμμικότητα: διαγνωστικές μέθοδοι και η παλινδρόμηση κορυφογραμμής ως μέθοδος χειρισμού

Supervisors : Batsidis Apostolos , Stefanidis Ioannis , Doxani Chrysoula

Φοιτήτρια : Πατσούρα Ιφιγένεια

A.M. : 00095

Ακαδημαϊκό έτος : 2016 – 2017

ΚΕΦΑΛΑΙΟ 1

Περίληψη

Το θέμα που αναλύει η συγκεκριμένη διπλωματική εργασία είναι η πολυσυγγραμικότητα, η οποία δημιουργείται στην πολλαπλή γραμμική παλινδρόμηση όταν υπάρχει ισχυρή συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών. Ως μέθοδος αντιμετώπισης της συγκεκριμένης προβληματικής κατάστασης αναφέρεται και κατ' επέκταση αναλύεται η παλινδρόμηση κορυφογραμμής (Ridge Regression). Η μέθοδος αυτή εφαρμόζεται σε ένα σύνολο δεδομένων που σχετίζεται με την κυστική ίνωση.

Στο κεφάλαιο 2 "Εισαγωγή" αναφέρονται οι ορισμοί της πολλαπλής παλινδρόμησης και της πολυσυγγραμικότητας. Επισημαίνονται οι δείκτες εντοπισμού της πολυσυγγραμικότητας, τα προβλήματα που δημιουργεί στην πολυμεταβλητή ανάλυση παλινδρόμησης καθώς και οι τρόποι αντιμετώπισης του προβλήματος.

Στο κεφάλαιο 3 "Ridge Regression" γίνεται αναφορά στη μέθοδο Ridge Regression, που αποτελεί έναν από τους τρόπους αντιμετώπισης του προβλήματος.

Στο κεφάλαιο 4 "Μέθοδοι – Αποτελέσματα" γίνεται εφαρμογή των διαγνωστικών ελέγχων για τον εντοπισμό του προβλήματος της πολυσυγγραμμικότητας σε ένα σύνολο δεδομένων και έπειτα εφαρμόζεται η μεθόδος Ridge Regression, ως μέθοδος χειρισμού του προβλήματος.

Τέλος στο κεφάλαιο 5 "Συμπεράσματα", συνοψίζονται τα αποτελέσματα και επισημαίνεται η σπουδαιότητα εντοπισμού του προβλήματος της πολυσυγγραμικότητας στην ανάλυση πολλαπλής παλινδρόμησης.

Λέξεις κλειδιά :

Πολυσυγγραμικότητα, Ανάλυση Παλινδρόμησης, Παλινδρόμηση Κορυφογραμμής

Abstract

The present Master's thesis deals with multicollinearity and Rigde Regression. Multicollinearity is appeared in multiple linear regression analysis when there is a strong correlation between independent variables, while Ridge Regression is a method for handling multicollinearity.

In this frame, in chapter 2 "Introduction" the definitions of multiple regression and multicollinearity are given. Methods to diagnose the problem of multicollinearity are given. In chapter 3 "Ridge Regression" is made reference in Ridge Regression method, which is one of the ways to tackle the problem. In chapter 4 "Methods-Results" the diagnostics and the method of handling multicollinearity discussed in the previous chapters are performed in a data set related with cystic fibrosis. Finally, in chapter 5 "Conclusions" the results are summarized and the importance of detecting the problem of the multicollinearity in the multiple regression analysis is mentioned.

Key words: Multicollinearity, Regression Analysis, Ridge Regression

ΚΕΦΑΛΑΙΟ 2

Εισαγωγή :

Η Στατιστική είναι ένας κλάδος των Μαθηματικών , ο οποίος κλάδος αποτελείται από ένα σύνολο αρχών και μεθοδολογιών για τον σχεδιασμό της διαδικασίας συλλογής δεδομένων, τη συνοπτική και αποτελεσματική παρουσίαση τους και την εξαγωγή αντίστοιχων συμπερασμάτων.

Η Στατιστική επιτυγχάνει τη συλλογή, επεξεργασία, παρουσίαση και ανάλυση των στατιστικών στοιχείων (αριθμητικών δεδομένων) με την εφαρμογή κατάλληλων για κάθε περίπτωση στατιστικών μεθόδων, οι οποίες και συνιστούν το περιεχόμενό της.

Σε αρκετές εφαρμογές τα δεδομένα που συλλέγονται αφορούν παραπάνω από μία μεταβλητές, η στατιστική ανάλυση των οποίων είναι υπολογιστικά επίμονη. Για τον λόγο αυτό, η ανάλυση γίνεται με τη χρήση Η/Υ μέσω κατάλληλου λογισμικού. Οι μεθοδολογίες των πολυμεταβλητών δεδομένων έχουν αρκετές εφαρμογές όπως για παράδειγμα στη βιοστατιστική και στη φαρμακολογία, στα οικονομικά, στην εκπαίδευση, στη βιολογία, στη μετεωρολογία κ.α.

Ένα εκ των αντικειμένων της Στατιστικής είναι η εξέταση και ο προσδιορισμός της σχέσης μεταξύ δύο ή περισσότερων μεταβλητών με απώτερο στόχο την πρόβλεψη της τιμής μίας απ'αυτές μέσω των άλλων. Το γνωστικό αυτό αντικείμενο είναι ευρύτερα γνωστό ως **ανάλυση παλινδρόμησης .**

Κάνουμε λόγο για το μοντέλο της απλής παλινδρόμησης όταν εξετάζουμε τη σχέση μεταξύ δύο μεταβλητών και για το μοντέλο της πολλαπλής παλινδρόμησης όταν εξετάζουμε τη σχέση μίας μεταβλητής, της εξαρτημένης, με πλήθος άλλων μεταβλητών, των ανεξάρτητων. Ακολουθούν παραδείγματα στα οποία συναντάμε απλή και πολλαπλή παλινδρόμηση.

Παραδείγματα :

- Ένα παράδειγμα απλής παλινδρόμησης αποτελεί ο προσδιορισμός της σχέσης είτε ανάμεσα στην ηλικία ενός ατόμου και στην πίεση αίματος είτε στην ηλικία ενός παιδιού και το ύψος του.
- Παραδείγματα πολλαπλής παλινδρόμησης είναι η εύρεση της σχέσης μεταξύ της παραγωγής ενός αγρού και της ποσότητας λιπάσματος που χρησιμοποιήθηκε, της θερμοκρασίας της περιοχής και της υγρασίας της περιοχής.

Συνήθως η εξαρτημένη μεταβλητή δεν επηρεάζεται μόνο από μία μεταβλητή αλλά από δύο ή περισσότερες ανεξάρτητες μεταβλητές. Όταν η σχέση αυτή μπορεί να εκφραστεί σε γραμμική μορφή τότε λέμε ότι πρόκειται για πολλαπλή γραμμική παλινδρόμηση και η εξίσωση έχει την ακόλουθη μορφή:

$$\boldsymbol{Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k + e}$$

όπου *k* είναι ο αριθμός των εισαγόμενων ανεξάρτητων μεταβλητών, και e το σφάλμα.

Στην πολλαπλή παλινδρόμηση προσδιορίζονται περισσότερες παράμετροι με τρόπο ανάλογο όπως και για την απλή παλινδρόμηση, δηλαδή με την ελαχιστοποίηση του αθροίσματος των τετραγώνων των καταλοίπων. Για παράδειγμα εφόσον έχουμε δύο ανεξάρτητες μεταβλητές, πρέπει να προσδιοριστούν τρεις παράμετροι, οι *a, b1* και *b2*. Η περίπτωση αυτή της πολλαπλής παλινδρόμησης αντιστοιχεί σε προσαρμογή επιπέδου (αντί ευθείας) και υπάρχει δυνατότητα παράστασης σε τρεις διαστάσεις. Με περισσότερες ανεξάρτητες μεταβλητές δεν υπάρχει δυνατότητα γεωμετρικής απεικόνισης του μοντέλου της παλινδρόμησης .

Ειδικότερα ο εκτιμητής ελαχίστων τετραγώνων έχει την μορφή :

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

όπου X είναι ο πίνακας σχεδιασμού ή δεδομένων διάστασης $n \times (p + 1)$ η πρώτη στήλη του οποίου αναφέρεται στο σταθερό όρο και έχει όλα της τα στοιχεία ίσα με ένα ενώ οι υπόλοιπες έχουν τα δεδομένα κάθε μεταβλητής, ενώ Y είναι ένα διάνυσμα διάστασης $n \times 1$ με τα δεδομένα της μεταβλητής απόκρισης.

Η συγγραμικότητα (**collinearity**) ή η πολυσυγγραμικότητα (**multicollinearity**) είναι εκείνη η ανεπιθύμητη κατάσταση (που εμφανίζεται στην πολυμεταβλητή παλινδρόμηση) όπου μία ανεξάρτητη μεταβλητή είναι γραμμική συνάρτηση των υπολοίπων ή κάποιων ανεξάρτητων μεταβλητών. Μαθηματικά το πρόβλημα έγκειται στο ότι αν μία μεταβλητή (στήλη του X) είναι γραμμικός συνδυασμός των υπολοίπων τότε δεν υπάρχει ο αντίστροφος $(X^T X)^{-1}$. Στην πράξη σπάνια έχουμε τέλεια γραμμική σχέση. Αν όμως μία μεταβλητή σχετίζεται υψηλά με τις υπόλοιπες, δηλαδή όταν διενεργούμε παλινδρόμηση μεταξύ τους προκύψει μεγάλο R^2 (βλέπε και ορισμούς των δεικτών VIF και Tolerance παρακάτω) τότε έχουμε ασταθείς εκτιμήσεις και μεγάλα τυπικά σφάλματα.

Στη συνέχεια ακολουθούν οι ενδείξεις που μαρτυρούν την ύπαρξη του προβλήματος της πολυσυγγραμικότητας (βλέπε μεταξύ άλλων Chattejee and Hadi, 2012).

Ενδείξεις

- Προσθήκη ή αφαίρεση μίας ανεξάρτητης μεταβλητής (ή ακόμα και παρατήρησης) επιφέρει μεγάλες αλλαγές στους εκτιμητές των συντελεστών της γραμμικής παλινδρόμησης.
- Το F – test στον πίνακα ANOVA απορρίπτει την μηδενική υπόθεση, ενώ τα επιμέρους t – test για τους συντελεστές παλινδρόμησης δεν απορρίπτονται.
- Μεγάλοι συντελεστές συσχέτισης μεταξύ ζευγαριών ανεξάρτητων μεταβλητών.
- Τα πρόσημα των εκτιμητών των παραμέτρων είναι μη αναμενόμενα σε σχέση με την εκ των προτέρων γνώση, εμπειρία ή γνωστή θεωρία.
- Τα διαστήματα εμπιστοσύνης για τους συντελεστές παλινδρόμησης σημαντικών μεταβλητών είναι πολύ μεγάλα.

Επιπλέον, έχουν προταθεί στη βιβλιογραφία διάφοροι τρόποι ελέγχου της ύπαρξης πολυσυγγραμμικότητας, πέραν των παραπάνω ενδείξεων. Στη συνέχεια παρατίθενται εν συντομία οι κυριότεροι από αυτούς.

(**Chattejee and Hadi, 2012).**

ΤΡΟΠΟΙ ΕΛΕΓΧΟΥ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ ΤΗΣ ΣΥΓΓΡΑΜΙΚΟΤΗΤΑΣ (Collinearity Diagnostics)

- **Συντελεστής συσχέτισης.**
  Όταν απλός συντελεστής συσχέτισης μεταξύ δύο ανεξάρτητων μεταβλητών είναι κατά απόλυτη τιμή μεγαλύτερος από 0.7 τότε υπάρχει ένδειξη πολυσυγγραμμικότητας.
- Tolerance.
  $Tolerance = 1 - R_i^2$, 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
- **Variance Inflation Factor.**
  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. Οι τιμές του δείκτη VIF που ξεπερνούν το 10 θεωρούνται γενικά απόδειξη ύπαρξης πολυσυγγραμικότητας. Αυτό συμβαίνει όταν $R^2_i > 0.9$.

- **Condition Number.**
  Ο condition number είναι το πηλίκο της μεγαλύτερης προς τη μικρότερη ιδιοτιμή του πίνακα X'X. Τιμές μεγαλύτερες του 1000. υποδεικνύουν πρόβλημα.
- **Condition Index**
  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.
- **Variance – decomposition proportion.**
  Ο δείκτης variance proportion μας υποδηλώνει τις αναλογίες διακύμανσης της εκτίμησης που υπολογίζονται από κάθε κύρια συνιστώσα που σχετίζεται με κάθε μία από τις ιδιοτιμές. Συνεπώς, οι ανεξάρτητες μεταβλητές με μεγάλες διακυμάνσεις είναι εκείνες που είναι αρκετά συσχετισμένες μεταξύ τους.

Οι τρόποι αντιμετώπισης του προβλήματος της πολυσυγγραμικότητας είναι οι εξής :

ΤΡΟΠΟΙ ΑΝΤΙΜΕΤΩΠΙΣΗΣ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ ΤΗΣ ΠΟΛΥΣΥΓΓΡΑΜΜΙΚΟΤΗΤΑΣ

- Διεξαγωγή Παλινδρόμησης με κύριες συνιστώσες .
- Εφαρμογή της μεθόδου Ridge Regression.
- Απαλοιφή ανεξάρτητων μεταβλητών ή προσθήκη περιορισμών για τις παραμέτρους ή προσθήκη μερικών παρατηρήσεων.

(http://ncss.wpengine.netdnacdn.com/wpcontent/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf και Chattejee and Hadi, 2012).

Στη διπλωματική αυτή διατριβή θα ασχοληθούμε με την μέθοδο ***Ridge Regression*** ως μέθοδο επίλυσης του προβλήματος της πολυσυγγραμικότητας. Αρχικά θα παρουσιαστεί η μέθοδος καθώς και η μαθηματική της υπόσταση στο επόμενο κεφάλαιο και έπειτα θα ακολουθήσει η εφαρμογή της μεθόδου με ένα παράδειγμα.

ΚΕΦΑΛΑΙΟ 3

RIDGE REGRESSION

Ένας τρόπος για να διαφύγουμε από το πρόβλημα της πολυσυγγραμικότητας είναι η εγκατάσταση ενός μεροληπτικού εκτιμητή με μικρότερη διασπορά από αυτές των εκτιμητριών των ελαχίστων τετραγώνων. Αυτή ήταν η βασική ιδέα της κορυφογραμμής και της συρρίκνωσης, εκτιμητές που εισήγαγαν οι Hoerl & Kennard (1970) στο μοντέλο παλινδρόμησης. Η διαδικασία που προτείνεται με τη μέθοδο αυτή έχει ως στόχο να ξεπεράσει τη δυσκολία που δημιουργείται από την ύπαρξη συσχέτισης μεταξύ μεταβλητών.

Ο παραπάνω εκτιμητής έχει την μορφή (βλέπε μεταξύ άλλων Chattejee and Hadi, 2012) :

$$\widehat{\boldsymbol{\beta}}_{\mathbf{ridge}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X} + \lambda\,\mathbf{I_p})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

όπου , $\hat{\beta} = (\widehat{\beta_0}, \ldots, \widehat{\beta_p})$ είναι το διάνυσμα των εκτιμητών μέγιστης πιθανοφάνειας διάστασης $(p+1)\times 1$, X είναι ο πίνακας σχεδιασμού ή δεδομένων διάστασης $n \times (p+1)$ η πρώτη στήλη του οποίου αναφέρεται στο σταθερό όρο και έχει όλα της τα στοιχεία ίσα με ένα ενώ οι υπόλοιπες έχουν τα δεδομένα κάθε μεταβλητής. Τέλος Y είναι ένα διάνυσμα διάστασης $n \times 1$ με τα δεδομένα της μεταβλητής απόκρισης.

Η επιλογή της σταθεράς λ γίνεται μεταξύ άλλων με την **Ridge trace** μέθοδο στο διάστημα (0 ,1) για το οποίο παρατηρούμε ότι οι εκτιμητές των συντελεστών της παλινδρόμησης γίνονται πιο σταθεροί .

Για να αναφερθούμε στις ιδιότητες του εκτιμητή της κορυφογραμμής μετατρέπουμε το γραμμικό μοντέλο της παλινδρόμησης σε μία κανονική μορφή , όπου $X^T X$ διαγώνιος πίνακας.

P : πίνακας που έχει στήλες τα ιδιοδιανύσματα του $X^T X$

Λ : διαγώνιος πίνακας που έχει ως κύρια διαγώνιο τις ιδιοτιμές του $X^T X$

Ορίζουμε τα εξής : $\mathbf{X^T X = P\Lambda P^T}$, $\boldsymbol{\alpha} = \mathbf{P^T\beta}$, $\mathbf{X^* = XP}$, $\mathbf{c = X^{*T} y}$

Έτσι έχουμε το νέο μοντέλο : $\boldsymbol{y = X^* \alpha + \varepsilon}$, με αντίστοιχο εκτιμητή OLS για το α : $\hat{\alpha} = (X^{*\prime}X)^{-1} X^{*\prime} y = (P^T X^T X P)^{-1} c = \Lambda^{-1} c$.

Κλιμακωτά παίρνουμε τα εξής :

(1) $\boldsymbol{\alpha_i = \frac{c_i}{\lambda_i}}$ **, i = 1 , ……, p** , εύκολα προκύπτει εάν προσθέσουμε στον παρονομαστή μια σταθερά κ η σχέση :

(2) $\boldsymbol{\alpha_i^R = \frac{c_i}{\lambda_i + k}}$

Ο Groß (2003) παρεμβαίνει και υποστηρίζει ότι λόγω της σταθεροποίησης θα ήταν πιο λογικό ίσως, να προσθέσουμε σε μικρές ιδιοτιμές μεγάλες τιμές και όχι μικρές τιμές σε μεγάλες ιδιοτιμές . Έτσι προκύπτει η παρακάτω σχέση :

(3) $\boldsymbol{\alpha_i^{GR} = \frac{c_i}{\lambda_i + k_i}}$

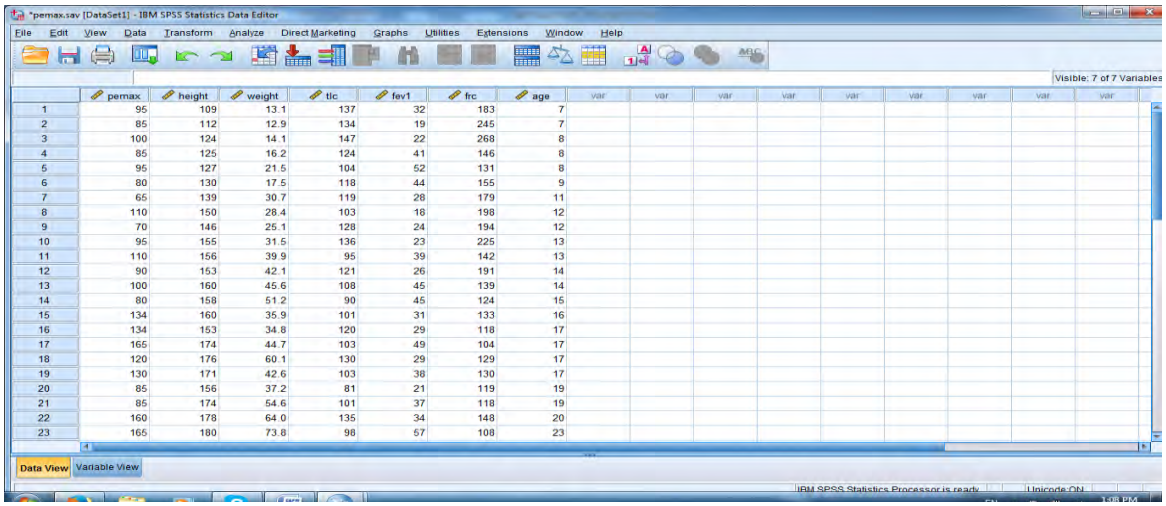
Και οι δύο τύποι εκτιμητών τόσο η κορυφογραμμή όσο και η συρρίκνωση είναι ειδικές περιπτώσεις αυτού του γενικού εκτιμητή κορυφογραμμής. Ο Groß (2003) δηλώνει ότι το μειονέκτημα αυτής της προσέγγισης είναι ότι απαιτείται ο προσδιορισμός όχι μόνο μίας παραμέτρου κορυφογραμμής αλλά p – παραμέτρων (για περισσότερες λεπτομέρειες παραπέμπουμε στους **Paul M. C. de Boer and Christian M. Hafner, 2005**).

ΚΕΦΑΛΑΙΟ 4

ΜΕΘΟΔΟΣ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

Στην συνέχεια χρησιμοποιώντας το λογισμικό SPSS θα εφαρμόσουμε όσα αναφέρθηκαν στα προηγούμενα κεφάλαια σε ένα σύνολο δεδομένων στο οποίο καταγράφονται στοιχεία 25 ασθενών (n = 25) που πάσχουν από κυστική ίνωση. Το σύνολο των δεδομένων του δείγματος λήφθηκε από το σύγγραμμα Julien Hoffman (2015) και είναι διαθέσιμο αν ζητηθεί.

Εισάγουμε τα δεδομένα μας στο SPSS .



Ειδικότερα στο σύνολο των δεδομένων περιέχεται ένα σύνολο μεταβλητών σχετικό με την κυστική ίνωση. Ως ανεξάρτητη μεταβλητή (target) Y ορίζουμε τη μέγιστη πίεση PEmax ενώ ως ανεξάρτητες – προβλεπτικές μεταβλητές ορίζουμε τις εξής: το ύψος (X_1 : height), το βάρος (X_2 : weight), τη συνολική χωρητικότητα του πνεύμονα (X_3 : tlc), την αναγκαστική εκπνοή (X_4: fev1), τη λειτουργική υπολειπόμενη χωρητικότητα (X_5 : frc) και την ηλικία (X_6 : age).

Εξετάζουμε την ύπαρξη προβλήματος πολυσυγγραμμικότητας ακολουθώντας τα εξής βήματα :

Analyze → Reggression → Linear

τοποθετώντας στην εξαρτημένη μεταβλητή (target) τη συνολική πίεση, ενώ στις ανεξάρτητες μεταβλητές τοποθετούμε τις υπόλοιπες : ηλικία , ύψος , βάρος , fev1 , frc , tlc .

Στο πλαίσιο statistics θα επιλέξουμε collinearity diagnostics για να ελέγξουμε την ύπαρξη πολυσυγγραμμικότητας και descriptives (για τον πίνακα συσχετίσεων)

Λαμβάνουμε τα παρακάτω αποτελέσματα στο Output του SPSS :

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,727[a]	,528	,371	26,51507

a. Predictors: (Constant), total lung capacity, weight (kg), forced expiratory volume, functional residual capacity, age in years, height (cm)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14177,755	6	2362,959	3,361	,021[b]
	Residual	12654,885	18	703,049		
	Total	26832,640	24			

a. Dependent Variable: maximum expiratory pressure

b. Predictors: (Constant), total lung capacity, weight (kg), forced expiratory volume, functional residual capacity, age in years, height (cm)

Από τον πίνακα **Model Summary** παρατηρούμε ότι $R^2 = 0.528$ ή διαφορετικά 52,8 % δηλαδή πάνω από το 50% που σημαίνει ότι η μελέτη μας, μας δίνει ικανοποιητικά αποτελέσματα. Επίσης το F – test είναι στατιστικά σημαντικό εφόσον sig = 0.021 < 0.05 συνεπώς μπορούμε να υποθέσουμε ότι το μοντέλο εξηγεί μία σημαντική ποσότητα της διακύμανσης της μέγιστης πίεσης εκπνοής

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-121,420	127,225		-,954	,353		
	age in years	4,082	3,340	,618	1,222	,237	,103	9,745
	height (cm)	,344	,805	,221	,428	,674	,098	10,214
	weight (kg)	-,239	,995	-,128	-,241	,813	,092	10,821
	forced expiratory volume	1,486	,790	,497	1,882	,076	,375	2,668
	functional residual capacity	,156	,253	,204	,616	,545	,240	4,166
	total lung capacity	,459	,454	,233	1,011	,325	,494	2,023

a. Dependent Variable: maximum expiratory pressure

Στον πίνακα **Coefficients** παρατηρούμε ότι καμία από τις προβλεπτικές μεταβλητές δεν είναι στατιστικά σημαντικές ($P > 0.05$) ωστόσο το μοντέλο παλίνδρομησης είναι στατιστικά σημαντικό ($P = 0,021 < 0,05$) γεγονός που αποτελεί ένδειξη ότι υπάρχει πρόβλημα πολυσυγγραμικότητας. Ακόμα οι δείκτες ανοχής (Collinearity Tollerance) καθώς και οι συντελεστές διόγκωσης (VIF) μας δείχνουν το μέγεθος της πολυσυγγραμικότητας. Συγκεκριμένα παρατηρούμε ότι υπάρχει υψηλή συσχέτιση για τις προβλεπτικές μεταβλητές βάρος και ύψος αφού VIF > 10 .

Στον πίνακα **Collinearity Diagnostics** παρατηρούμε ότι οι ιδιοτιμές (Eigenvalue), για τις ανεξάρτητες μεταβλητές X_5 , X_6 και X_7 , είναι πολύ κοντά στο μηδέν ενώ οι δείκτες κατάστασης (Condition Index) έχουν πολύ υψηλές τιμές. Συγκεκριμένα οι τιμές των δεικτών κατάστασης για τις X_6 , X_7 ξεπερνούν την τιμή 30 γεγονός που αποκαλύπτει σοβαρό πρόβλημα πολυσυγγραμικότητας .

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions						
				(Constant)	age in years	height (cm)	weight (kg)	forced expiratory volume	functional residual capacity	total lung capacity
1	1	6,605	1,000	,00	,00	,00	,00	,00	,00	,00
	2	,282	4,838	,00	,01	,00	,02	,00	,02	,00
	3	,082	8,977	,00	,02	,00	,01	,31	,01	,00
	4	,017	19,888	,01	,16	,01	,40	,02	,14	,00
	5	,008	28,398	,00	,02	,00	,02	,07	,54	,86
	6	,005	37,909	,07	,78	,12	,13	,35	,18	,08
	7	,001	83,236	,91	,02	,87	,43	,25	,11	,05

a. Dependent Variable: maximum expiratory pressure

Στην συνέχεια θα ελέγξουμε εάν υπάρχει συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών ($P < 0.05$).

Correlations

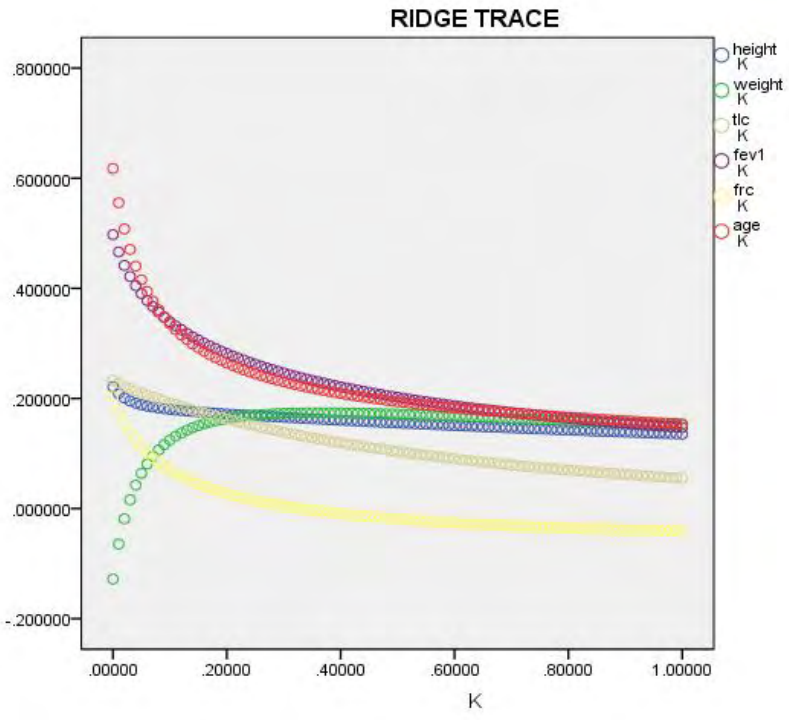
		maximum expiratory pressure	age in years	height (cm)	weight (kg)	forced expiratory volume	functional residual capacity	total lung capacity
Pearson Correlation	maximum expiratory pressure	1,000	,613	,599	,635	,453	-,417	-,182
	age in years	,613	1,000	,926	,906	,294	-,639	-,469
	height (cm)	,599	,926	1,000	,921	,317	-,624	-,457
	weight (kg)	,635	,906	,921	1,000	,449	-,617	-,418
	forced expiratory volume	,453	,294	,317	,449	1,000	-,665	-,443
	functional residual capacity	-,417	-,639	-,624	-,617	-,665	1,000	,704
	total lung capacity	-,182	-,469	-,457	-,418	-,443	,704	1,000
Sig. (1-tailed)	maximum expiratory pressure	.	,001	,001	,000	,011	,019	,192
	age in years	,001	.	,000	,000	,077	,000	,009
	height (cm)	,001	,000	.	,000	,062	,000	,011
	weight (kg)	,000	,000	,000	.	,012	,001	,019
	forced expiratory volume	,011	,077	,062	,012	.	,000	,013
	functional residual capacity	,019	,000	,000	,001	,000	.	,000
	total lung capacity	,192	,009	,011	,019	,013	,000	.

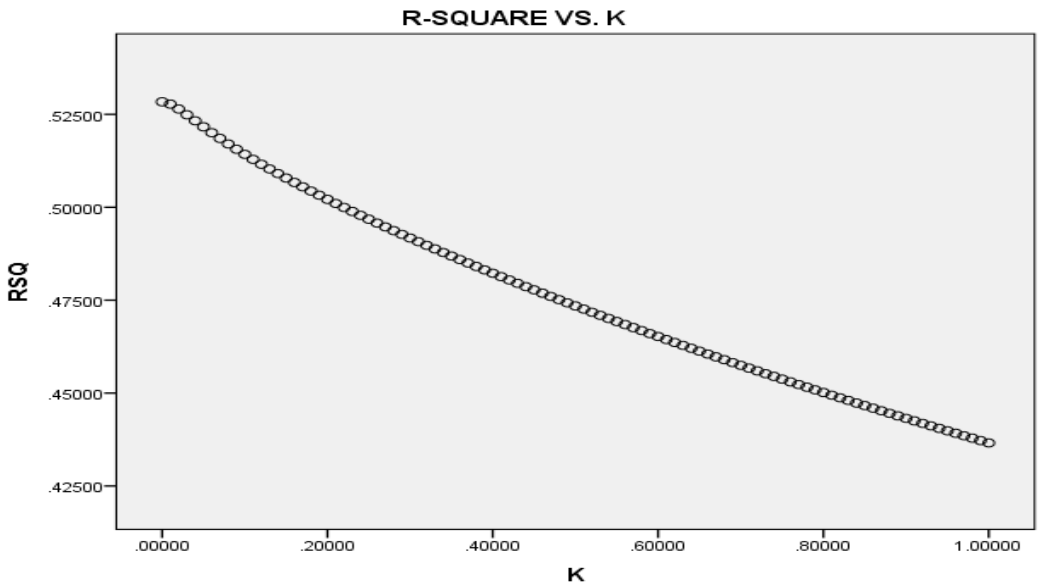
Από τους παραπάνω ελέγχους επιβεβαιώνεται η ύπαρξη πολυσυγγραμικότητας, συνεπώς θα συνεχίσουμε στην επίλυση του προβλήματος χρησιμοποιώντας την μέθοδο Ridge Regression.

Στο SPSS δεν υπάρχει παραθυρικό περιβάλλον για τη διεξαγωγή της μεθόδου αλλά γίνεται μέσω του παρακάτω αρχείου σύνταξης (File→New→Syntax)

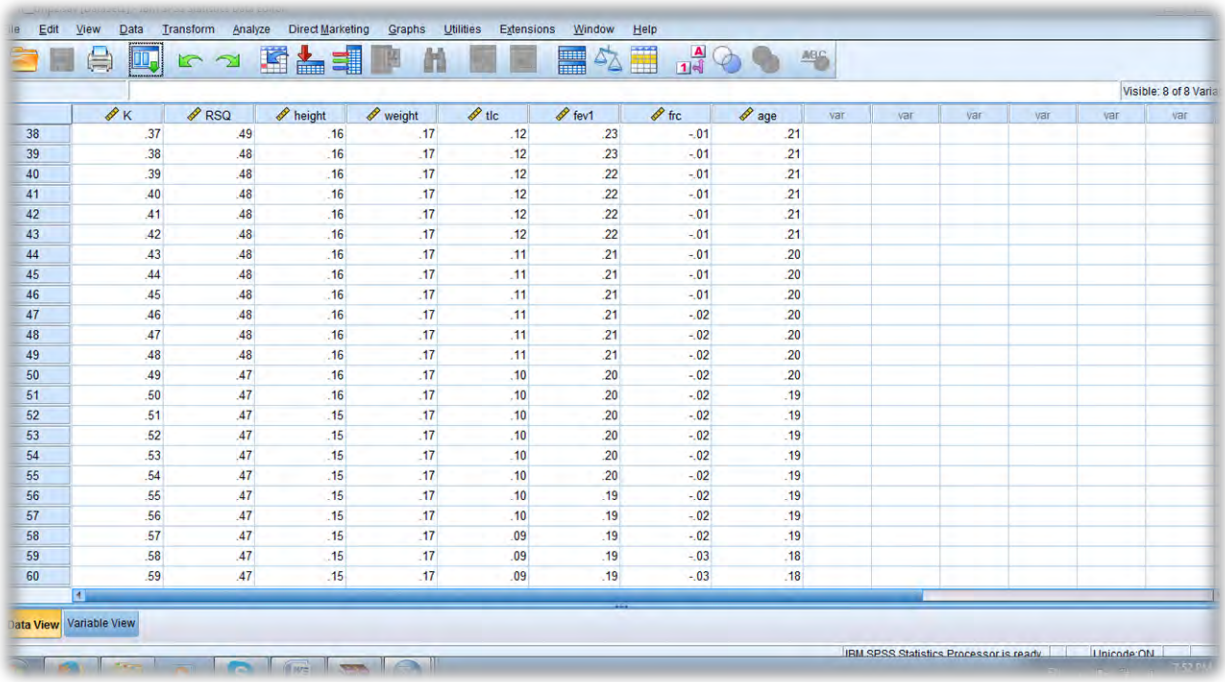
```
cd "C:\Users\User\Desktop\ifigeneia".

INCLUDE 'C:\Program Files\IBM\SPSS\Statistics\23\Samples\English\Ridge regression.sps'.

RIDGEREG DEP=pemax /ENTER = height to age

/START = 0

/STOP = 1

/INC = 0.01

/DEBUG = 'Y'.
```

Επιλέγουμε το run All και λαμβάνουμε όλες τις πιθανές τιμές του k μέσα στο διάστημα (0,1) με βήμα 0,01 στο output του spss.





Μας δίνεται επίσης από το παραπάνω plot η σχέση του R^2 και της σταθεράς k , παρατηρούμε ότι όσο αυξάνονται οι τιμές του k μειώνονται οι τιμές του R^2 .



Μέσω τις μεθόδου Ridge trace θα βρούμε την τιμή του k ∈ (0,1) όπως έχει αναφερθεί και στην θεωρία μας, για την οποία οι συντελεστές παλινδρόμησης σταθεροποιούνται .

Στον πίνακα Data View του spss εμφανίζονται όλες οι τιμές της σταθεράς k ∈ (0,1) με βήμα 0.01 καθώς και οι τιμές του R^2. Παρατηρούμε ότι οι συντελεστές παλινδρόμησης σταθεροποιούνται για k = 0.55, οπότε οι εκτιμητές Ridge για κάθε μεταβλητή είναι αντίστοιχα :

0.15∗height , 0.17∗weight ,0.10 ∗ tlc , 0.20 ∗fev1, – 0.02 ∗ frc ,0.19 ∗ age.

ΚΕΦΑΛΑΙΟ 5

‘’Συμπεράσματα ‘’

Σ’αυτή την διπλωματική εργασία καταφέραμε να εντοπίσουμε και να αντιμετωπίσουμε το πρόβλημα της πούγυγγραμικότητας εφαρμόζοντας τη μέθοδο παλινδρόμησης κορυφογραμμής (Ridge Regression) σ΄ένα σύνολο δεδομένων που σχετίζεται με την ιατρική και συγκεκριμένα με ασθενείς που πάσχουν από κυστική ίνωση. Παρατηρήσαμε στο συγκεκριμένο δείγμα ότι οι ανεξάρτητες μεταβλητές fev1, frc, age συσχετίζονται σε μεγάλο βαθμό. Στόχος της μεθόδου ήταν να βρούμε μία κατάλληλη τιμή της σταθεράς k ∈ (0 , 1) έτσι ώστε να σταθεροποιηθούν οι συντελεστές των ανεξάρτητων μεταβλητών στο μοντέλο παλινδρόμησης.

Η βασική ανάλυση στοιχείων πραγματοποιήθηκε εύκολα και αποτελεσματικά με τη βοήθεια του στατιστικού πακέτου SPSS.

Είναι σημαντικό να εξεταστούν τα διαγνωστικά πολυσυγγραμικότητας χρησιμοποιώντας μοντέλα παλινδρόμησης, διαφορετικά θα μπορούσαμε να οδηγηθούμε σε παραπλανητικές ερμηνείες των αποτελεσμάτων με την παρουσίαση ασταθών και προκατειλημμένων τυπικών σφαλμάτων, οδηγώντας σε πολύ ασταθείς p – value για την εκτίμηση της στατιστικής σημασίας του προγνωστικού.

Βιβλιογραφία

Samprit Chatterjee and Ali S. Hadi (2012). Regression Analysis by Example. Wiley.

E. Hoerl and Robert W. Kennard (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics, Vol. 12, No. 1, pp. 55-67.

Paul M. C. de Boer and Christian M. Hafner (2005). Ridge regression revisited Statistica Neerlandica (2005) Vol. 59, nr. 4, pp. 498–505

Groß, J. (2003), Linear regression, Lecture Notes in Statistics, Springer Verlag.

Julien Hoffman (2015). Biostatistics for Medical and Biomedical Practitioners. Academic Press.

Ιστότοποι

http://ncss.wpengine.netdnacdn.com/wpcontent/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf 3.