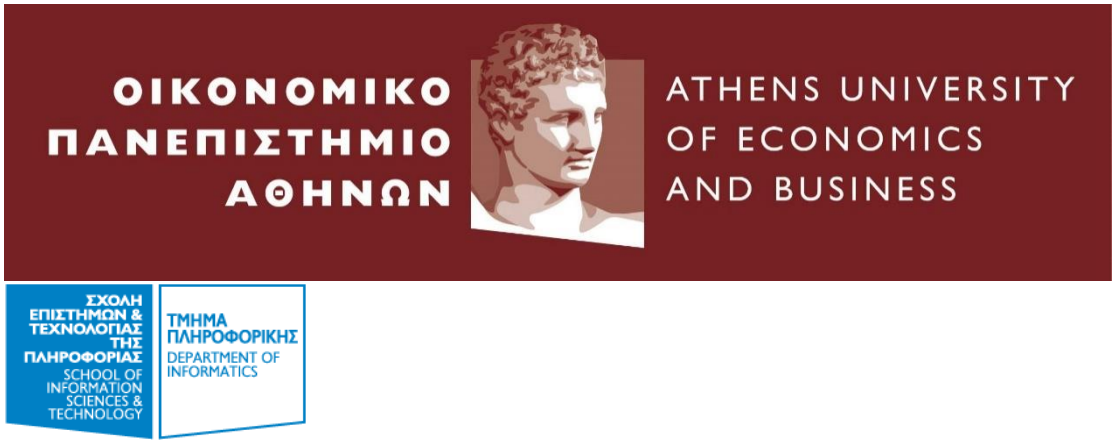


Νίκος Χριστοδούλου – ΑΜ: p3190223

2η Εργασία

Μέθοδοι Στατιστικής και Μηχανικής Μάθησης

Χειμερινό Εξάμηνο 2022-2023

Προεπεξεργασία:

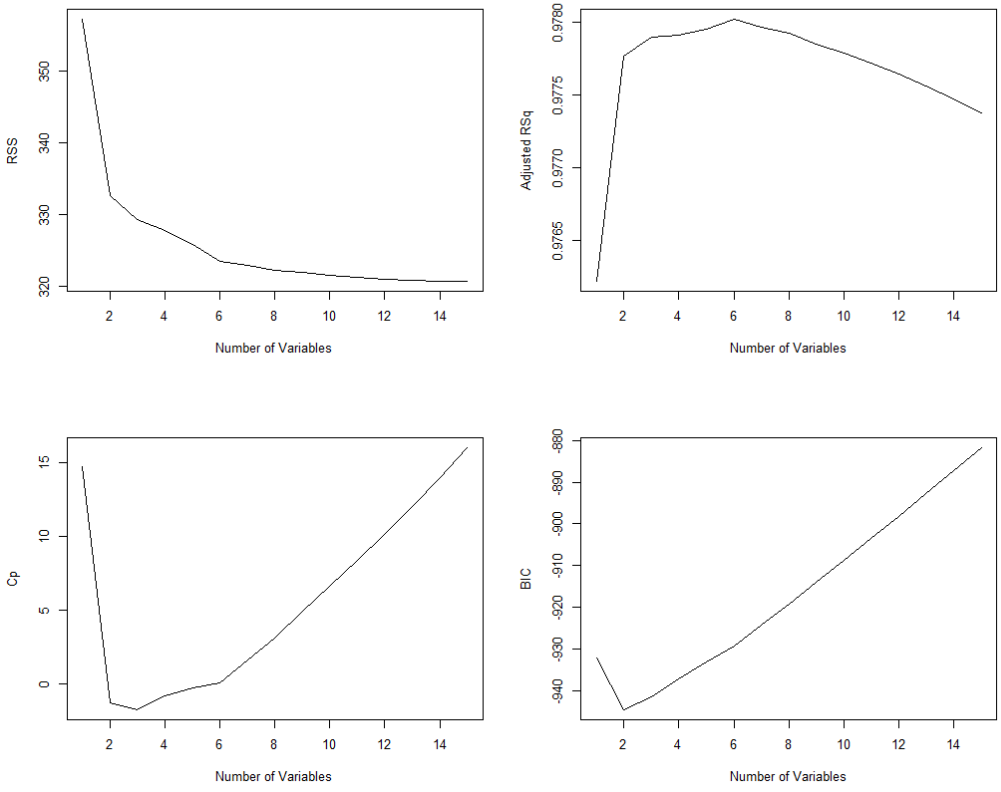
Αρχικά, φορτώνουμε το dataset μας με την εντολή read.csv και ελέγχουμε αν περιέχει na στοιχεία. Με την εντολή sum(is.na(body)), παρατηρούμε ότι δεν υπάρχουν na's, οπότε μπορούμε να αρχίσουμε να δουλεύουμε με τα δεδομένα.

Ερώτημα 1

Για την επιλογή του βέλτιστου υποσυνόλου μεταβλητών με σκοπό την πρόβλεψη του δείκτη σωματικού λίπους, μπορούμε να χρησιμοποιήσουμε ποικίλες τεχνικές. Επιλέγουμε τις μεθόδους:

- Best Subset Selection
- Forward – Stepwise Selection
- Backward – Stepwise Selection

i) Εφαρμόζοντας την πρώτη τεχνική με την εντολή regsubsets(), οπτικοποιούμε τα αποτελέσματα με τη βοήθεια ορισμένων μετρικών (rss, bic, cp, adjusted rsq) για να καταλήξουμε σε ένα μόνο υποσύνολο μεταβλητών. Να σημειωθεί ότι περνάμε ως παράμετρο στην συνάρτηση regsubsets() το nvmax = 15, ώστε να λάβουμε κάποιο αποτέλεσμα για κάθε αριθμό μεταβλητών μέχρι και n – 1, όπου n το πλήθος των μεταβλητών του συνόλου δεδομένων. Προκύπτουν τα παρακάτω γραφήματα:

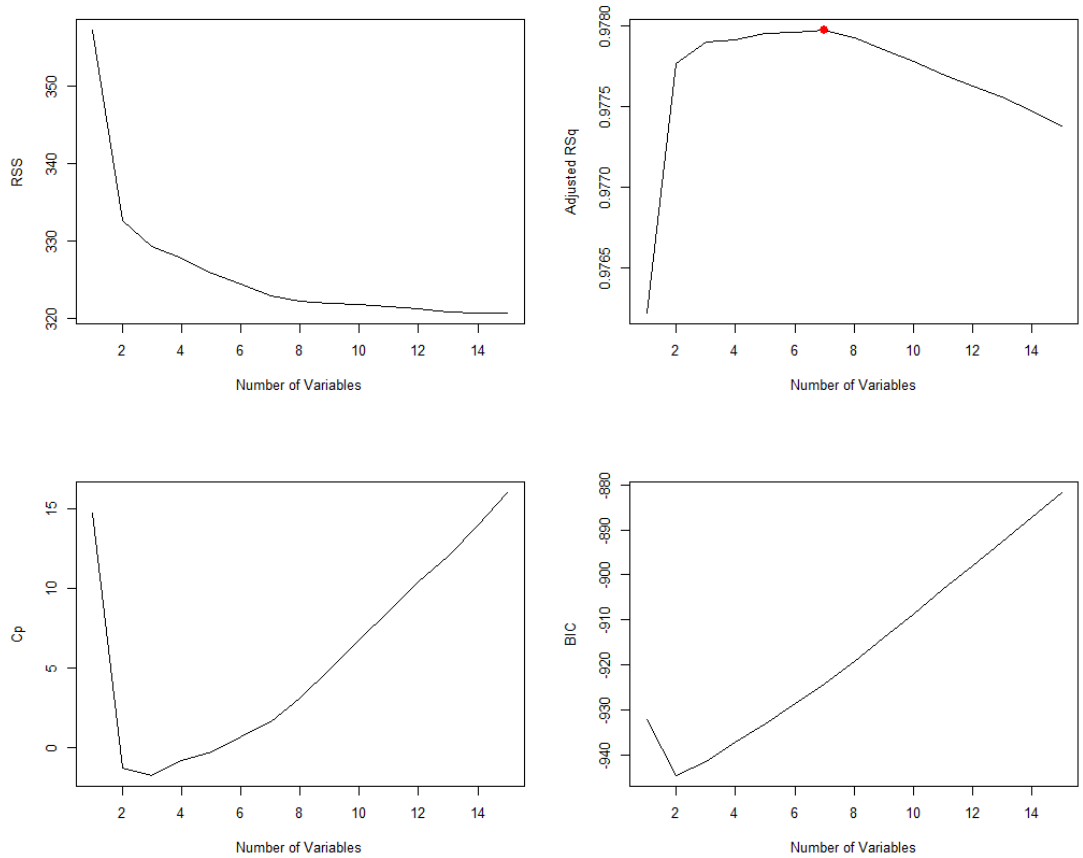


Γνωρίζουμε ότι η τιμή του rss θα μειώνεται με την αύξηση των μεταβλητών στο μοντέλο. Ωστόσο, παρατηρείται μία κύρτωση στο n = 6 και στην συνέχεια το rss μειώνεται με αργό ρυθμό όσο αυξάνεται το n. Το adjusted rss βλέπουμε ότι λαμβάνει τη μέγιστη τιμή του για n = 6. Από την άλλη, τα cp και bic δε μας δίνουν κάποιο σταθερό αποτέλεσμα για την επιλογή του υποσυνόλου. Συνεπώς, θα επαναλάβουμε την διαδικασία αυτή τη φορά με forward-stepwise και backward-stepwise selection, ώστε να συγκρίνουμε τα αποτελέσματα και να καταλήξουμε στο subset.

Forward και Backward Stepwise Selection

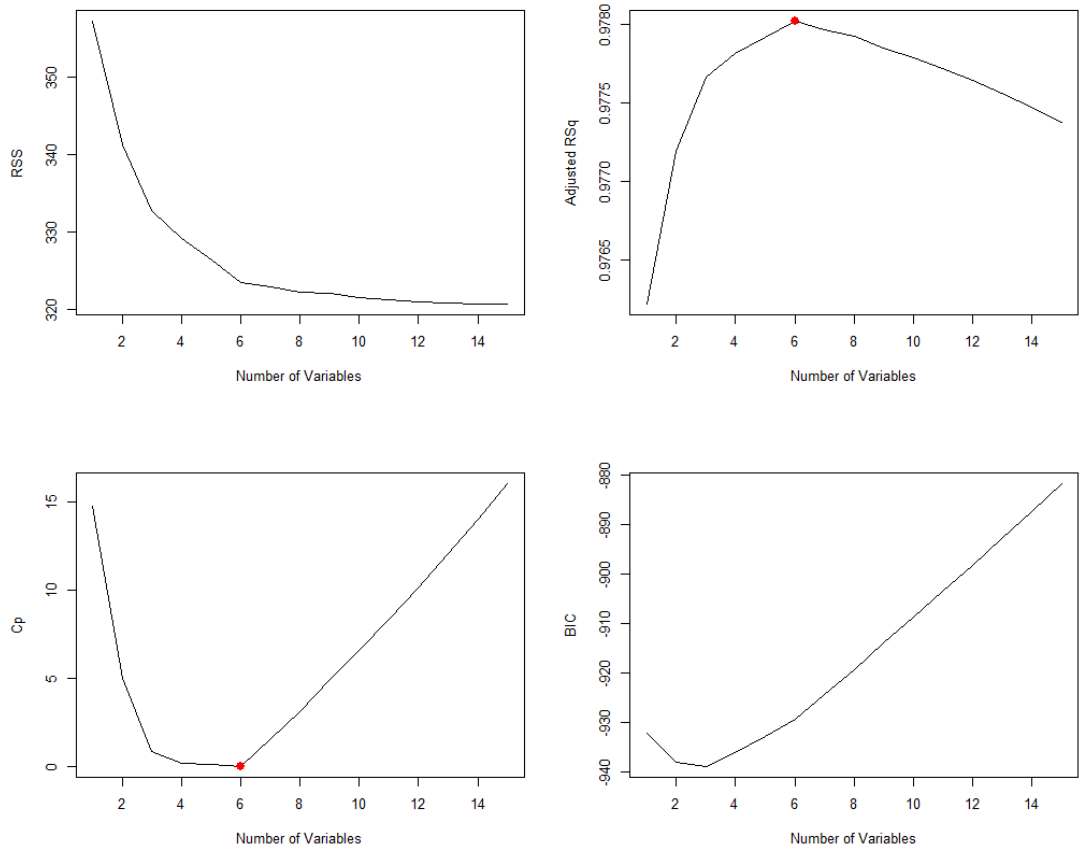
Επαναλαμβάνουμε την διαδικασία με την ίδια λογική. Τα γραφήματα που προκύπτουν είναι τα εξής:

- Για το forward:



Παρατηρούμε ότι παίρνουμε παρόμοια αποτελέσματα σε όλες τις μετρικές πλην της adjusted rsq που αυτή τη φορά επιλέγει υποσύνολο με 7 μεταβλητές.

- Για το backward:



Εδώ όμοια με την πρώτη μέθοδο, το adjusted rsq δίνει βέλτιστο υποσύνολο με 6 μεταβλητές, όπως και η μετρική Cp που δίνει βέλτιστο το μοντέλο που λαμβάνει την ελάχιστη τιμή, δηλαδή για n = 6.

Έχοντας εφαρμόσει και τις 3 τεχνικές, χρησιμοποιούμε το coef() για να δούμε αν λαμβάνουν τους ίδιους συντελεστές για n = 6 και n = 7. Παρατηρούμε ότι το best subset μαζί με το backward έχουν επιλέξει το ίδιο υποσύνολο μεταβλητών για n = 6, ενώ το backward αντί για τη μεταβλητή weight έχει επιλέξει την abdom. Από την άλλη, για n = 7 και τα τρία μοντέλα επέλεξαν τις ίδιες μεταβλητές.

Είναι προφανές ότι για να λάβουμε την τελική απόφαση δεν αρκούν μόνο τα παραπάνω. Αν σταματούσαμε, βέβαια, σε αυτό το σημείο θα καταλήγαμε στο μοντέλο με n = 6 που έδωσαν τα best subset και forward με τις μεταβλητές density, age, weight, chest, ankle, biceps.

Ερώτημα 2

Για την αξιολόγηση του μοντέλου στο οποίο καταλήξαμε στο προηγούμενο ερώτημα, θα χρησιμοποιήσουμε την προσέγγιση του validation set και θα εφαρμόσουμε k fold cross validation.

Θέτουμε ένα σταθερό seed και χωρίζουμε με τυχαίο τρόπο το σύνολο δεδομένων σε train και test δεδομένα. Έπειτα, εφαρμόσουμε το best subset selection στο training set και πάμε να προβλέψουμε την τιμή του brozek στο test set με n μέχρι και 15. Κάθε φορά υπολογίζουμε το MSE του κάθε μοντέλου και κρατάμε το ελάχιστο. Βλέπουμε ότι λαμβάνει ελάχιστη τιμή για n = 1, ωστόσο δε μπορούμε να επιλέξουμε προφανώς αυτό το μοντέλο.

Σε αυτό το σημείο θα εφαρμόσουμε το k fold validation για k = 10. Με τη βοήθεια μίας μεθόδου predict που δημιουργήσαμε για την πρόβλεψη της τιμής του brozek διασπάμε κάθε φορά το σύνολο δεδομένων σε train και test ανάλογα με την τιμή του k fold. Προβλέπουμε την τιμή του brozek για τον καλύτερο αριθμό υποσυνόλων και χρησιμοποιούμε πάλι το MSE για την επιλογή του καλύτερου subset.

Παρατηρώντας τα MSE για κάθε n, βλέπουμε ότι πάλι λαμβάνει ελάχιστη τιμή για n =1. Ωστόσο, μέχρι και για n = 5 το MSE αυξάνεται και για n = 6 μειώνεται και στην συνέχεια αυξάνεται συνεχώς μέχρι και το n = 15. Συνεπώς, το subset που επιλέξαμε είναι το n = 6 και υπολογίζοντας το coef() του βέλτιστου μοντέλου, παρατηρούμε ότι ταυτίζονται με τα coefficients που πήραμε και στο πρώτο ερώτημα. Άρα, η αρχική μας επιλογή φαίνεται να ήταν η βέλτιση. Παρακάτω παρατίθενται τα mean cross validation errors από n = 1 μέχρι και n = 6.

```
1.509696 1.553842 1.632771 1.649859 1.673810 1.640654 1.655405 1.665
561 1.675770 1.675271 1.677888 1.693643 1.689179 1.691028 1.688224
```

Ερώτημα 3

Για την εφαρμογή των εν λόγω μοντέλων επιβλεπόμενης μάθησης πρέπει να δημιουργήσουμε μία δίτιμη μεταβλητή Boolean ανάλογα με το αν ο δείκτης brozek είναι μεγαλύτερος του 24.0 ή όχι. Έπειτα, θα διασπάσουμε το σύνολο δεδομένων μας σε train και test data με train size = 70% και θα εφαρμόσουμε το κάθε μοντέλο. Για την τελική επιλογή του καλύτερου μοντέλου πρέπει να βεβαιωθούμε ότι εκπαιδεύουμε όλα τα μοντέλο στο ίδιο σύνολο εκπαίδευσης και τα αξιολογούμε στο ίδιο test σύνολο. Επίσης, εκτός από το accuracy καλό είναι να ληφθούν υπόψιν και παραπάνω μετρικές όπως τα Sensitivity, Specificity, Precision.

Και για τις 4 μεθόδους χρησιμοποιήθηκε κώδικας από τα εργαστήρια του μαθήματος. Παρατίθενται τα αποτελέσματα στο σύνολο επικύρωσης:

- **Logistic Regression**
    - Accuracy = 0.9733333
    - Error = 0.02666667
    - True positive rate = Sensitivity = 0.95
    - False positive rate = 1 – Specificity = 0.01818182

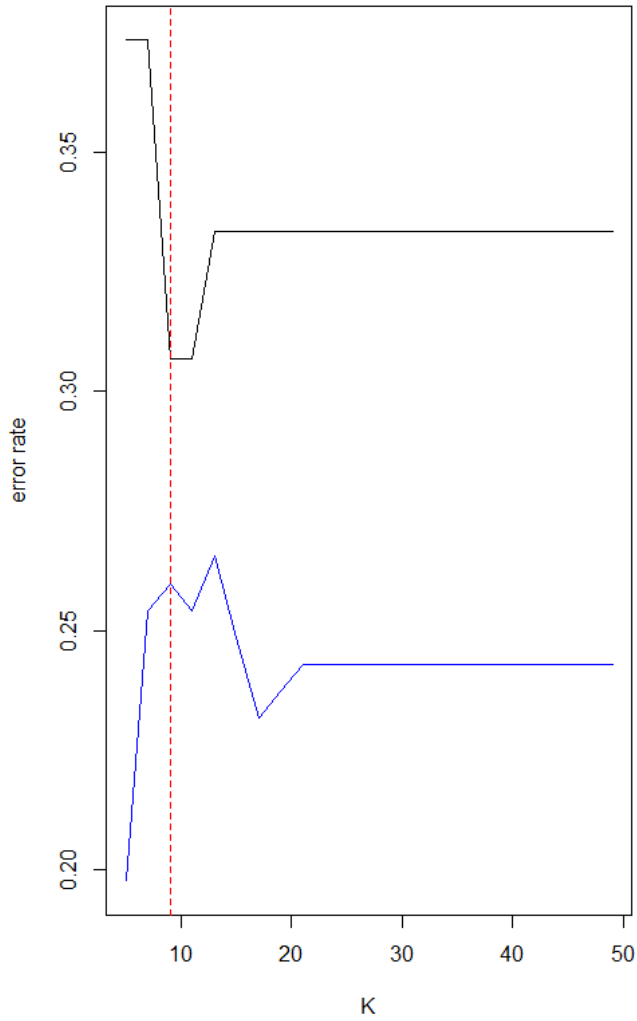
- Specificity = 0.9818182
- Precision = 0.95
- True Negative predicted value = 0.9818182

- **Linear Discriminant Analysis**
  - Accuracy = 0.9866667
  - Error = 0.01333333
  - True positive rate = Sensitivity = 1
  - False positive rate = 1 – Specificity = 0.01818182
  - Specificity = 0.9818182
  - Precision = 0.952381
  - True Negative predicted value = 1

- **Quadratic Discriminant Analysis**
  - Accuracy = 0.9466667
  - Error = 0.05333333
  - True positive rate = Sensitivity = 0.8
  - False positive rate = 1 – Specificity = 0
  - Specificity = 1
  - Precision = 1
  - True Negative predicted value = 0.9322034

- **K Nearest Neighbors**

  Για το συγκεκριμένο μοντέλο είναι πολύ σημαντική η επιλογή της υπερπαραμέτρου k. Συνεπώς, θα τρέξουμε το μοντέλο για διάφορες τιμές του k από k = 5 μέχρι 50 ώστε να βρούμε το βέλτιστο. Όπως βλέπουμε στο παρακάτω σχήμα την καλύτερη αναλογία error σε train και test δεδομένα την λαμβάνουμε για k = 9:



Τέλος, παρατίθεται η επίδοση του knn για k = 9:

- Accuracy = 0.6933333
- Error = 0.3066667
- True positive rate = Sensitivity = 0.08
- False positive rate = 1 – Specificity = 0
- Specificity = 1
- Precision = 1
- True Negative predicted value = 0.6849315

Με βάση, λοιπόν, τις παραπάνω μετρικές ο **LDA** φαίνεται πως δίνει τα καλύτερα αποτελέσματα στο συγκεκριμένο σύνολο δεδομένων.