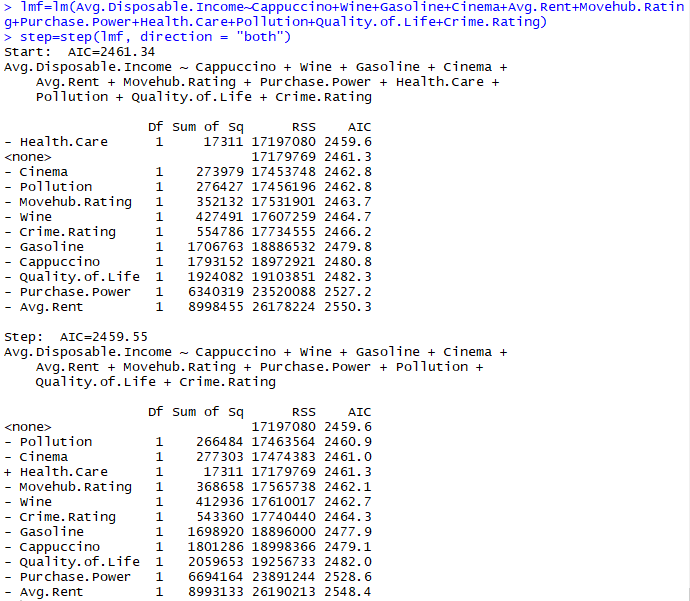
**Επιλογή του καλύτερου μοντέλου για πολλαπλή γραμμική παλινδρόμηση**

Στόχος μας στην πολλαπλή γραμμική παλινδρόμηση είναι να καθορίσουμε ποιες μεταβλητές συνεισφέρουν σημαντικότερα στο να εξηγήσουν διακύμανση της εξαρτώμενης μεταβλητής. Ένας στόχος στο να καθορίσουμε το καλύτερο μοντέλο είναι να ελαχιστοποιήσουμε το μέσο τετράγωνο των υπολοίπων, το οποίο θα οδηγήσει στη μεγιστοποίηση του R-τετράγωνο.

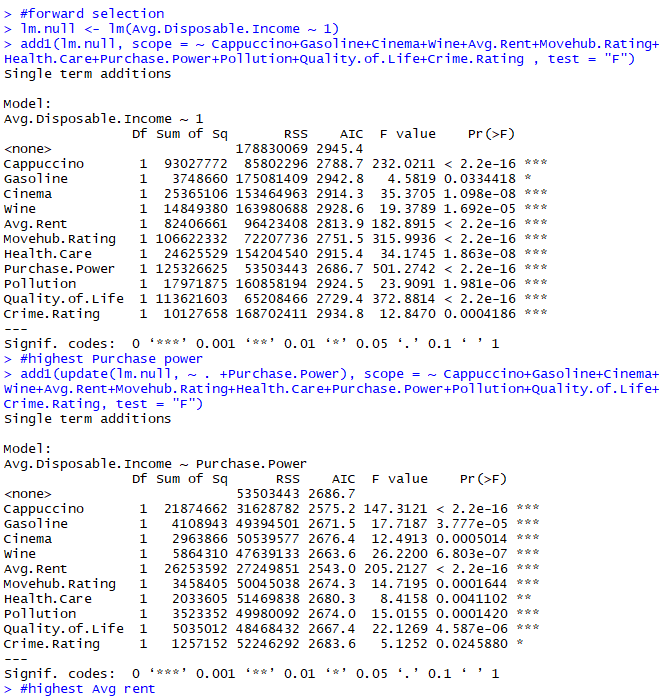
Ένα άλλο στατιστικό το οποίο μπορεί να καθορίσει το καλύτερο μοντέλο είναι το κριτήριο Mallow's Cp. Οι τιμές του Cp μειώνονται όσο ο αριθμός των ανεξάρτητων μεταβλητών του μοντέλου αυξάνεται.

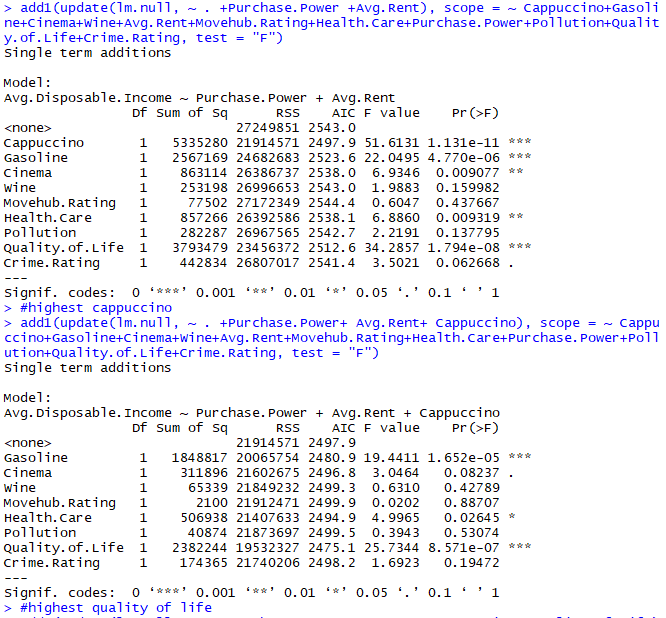
Πλέον, το κριτήριο πληροφορίας Akaike (AIC) χρησιμοποιείται ως επί το πλείστον στον καθορισμό του βέλτιστου μοντέλου.

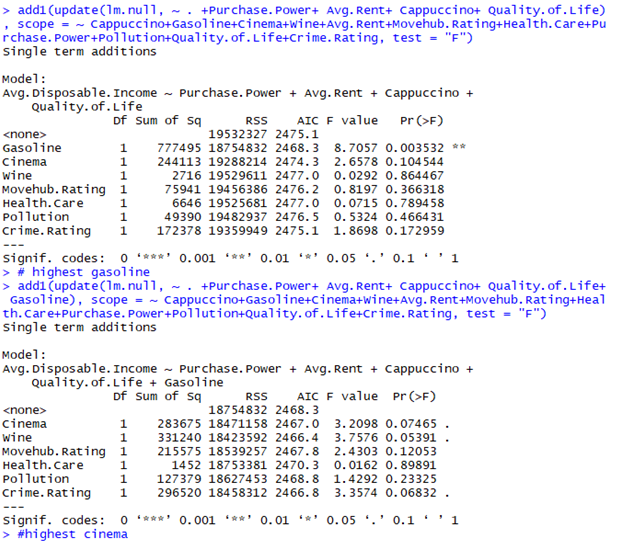
Όσο μικρότερο το AIC τόσο καλύτερο το μοντέλο. Η εντολή step επαναλαμβάνει συγκρίσεις μοντέλων με διαφορετικές μεταβλητές και σταματά όταν βρει το βέλτιστο.

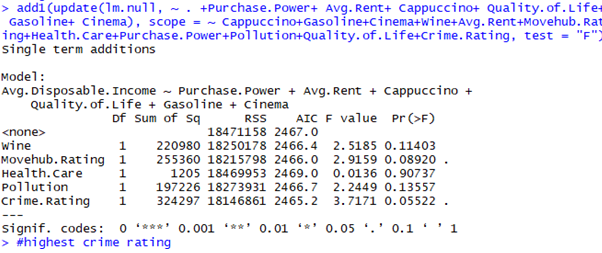
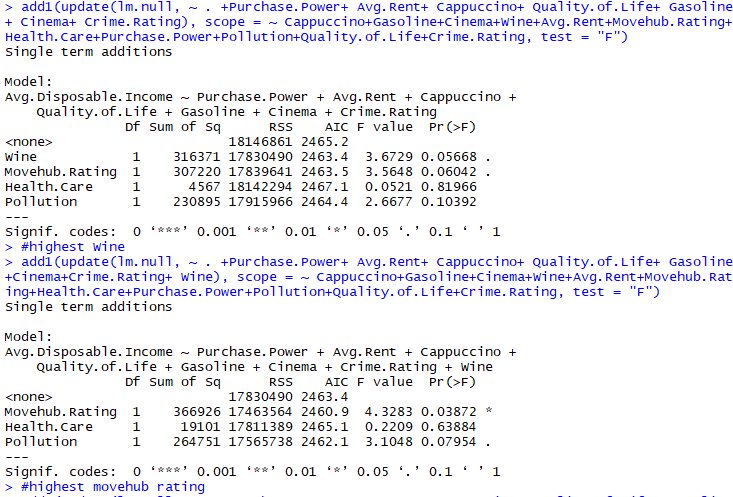
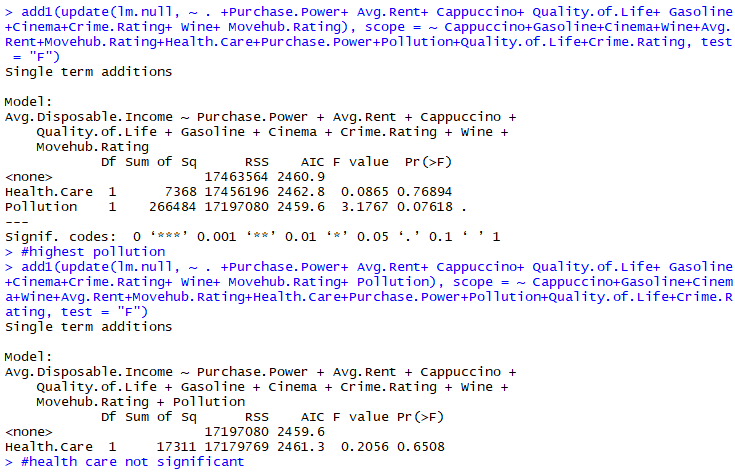
Αν όμως θέλαμε να χρησιμοποιήσουμε κάποιον άλλον τρόπο πολύ πιθανό να επιλέγαμε τη μέθοδο Stepwise regression. Είναι μια μέθοδος επιλογής μεταβλητών στην οποία πολλοί συνδυασμοί μεταβλητών τεστάρονται μαζί. Αρχικά, βρίσκουμε το καλύτερο μοντέλο με μία μόνο μεταβλητή, στη συνέχεια το καλύτερο μοντέλο με δύο μεταβλητές κλπ. Σε κάθε βήμα, πραγματοποιείται ένα F-test σε κάθε μεταβλητή, με μηδενική υπόθεση ότι ο συντελεστής της είναι 0. Επομένως, αν η p-value είναι μικρότερη από κάποιο a, π.χ. a=0.1, σημαίνει ότι η μεταβλητή είναι σημαντική για το μοντέλο. Σε κάθε βήμα επιλέγουμε τη "σημαντικότερη" μεταβλητή λοιπόν και επαναλαμβάνουμε έως ότου όλες οι εναπομείναντες p-values να ξεπερνούν το a.

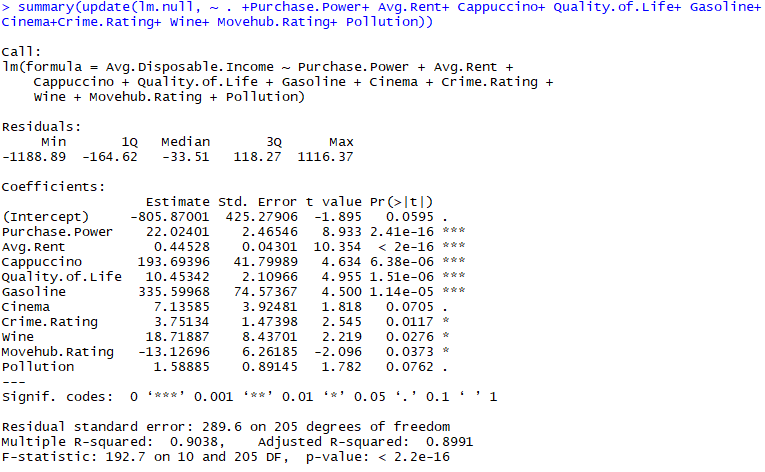
Γενικά, τα καλύτερα μοντέλα είναι αυτά που μεγιστοποιούν το R- τετράγωνο, το Cp ή και τα δύο. Παραλλαγές της Stepwise regression είναι η μέθοδος Forward selection και Backward elimination. Εμείς θα παρουσιάσουμε τη μέθοδο Forward Selection.

Ξεκινάμε με ένα μοντέλο που περιέχει μόνο μια σταθερά και συνεχίζουμε προσθέτοντας τη μεταβλητή που αυξάνει το R-τετράγωνο, δηλαδή τη μεταβλητή που συνεισφέρει περισσότερο στο μοντέλο.







Επομένως, το βέλτιστο μοντέλο είναι αυτό με όλες τις μεταβλητές εκτός της Health Care, που είναι το ίδιο με αυτό που μας έδωσε το κριτήριο Akaike.

Αυτή η μέθοδος μπορεί να χρησιμοποιηθεί και για μοντέλα λογιστικής παλινδρόμησης.

Οι μέθοδοι Stepwise ελέγχουν ένα περιορισμένο αριθμό από τα συνολικά πιθανά μοντέλα και χρησιμοποιούν το F-test που είναι μια αβέβαιη μέθοδος επιλογής μοντέλων. Οι μέθοδοι επιλογής μοντέλων που βασίζονται σε κάποιο κριτήριο πληροφορίας συγκρίνουν περισσότερα μοντέλα και με καλύτερο τρόπο.

**Μέθοδοι επιλογής μοντέλων, βασισμένοι σε κριτήρια**

Αν υπάρχουν ν μεταβλητές υπάρχουν 2ν διαφορετικά μοντέλα, γι' αυτό χρησιμοποιούμε κάποιους αλγορίθμους για την εύρεση του καλύτερου μοντέλου. Κάποια από τα κριτήρια που χρησιμοποιούνται είναι:

1. The Akaike Information Criterion (AIC) και το Bayes Information Criterion (BIC).

2. Adjusted R2. Βέβαια με την προσθήκη μιας μεταβλητής στο μοντέλο το R2 πάντα αυξάνεται, οπότε αυτό δεν είναι πολύ καλό κριτήριο καθώς θα επιλέγεται πάντα το μεγαλύτερο μοντέλο.

3. Predicted Residual Sum of Squares (PRESS)

4. Mallow’s Cp Statistic. Ένα καλό μοντέλο θα πρέπει να κάνει και καλή πρόβλεψη, οπότε η μέση τιμή του MSE ίσως είναι ένα καλό κριτήριο (p οι παράμετροι του μοντέλου).

Το Cp είναι εύκολο να υπολογιστεί. Είναι άμεσα συνδεδεμένο με το R2 και το AIC. Για το μοντέλο με όλες τις μεταβλητές το Cp είναι πάντα ίσο με p. Ένα μοντέλο με κακό fit θα έχει Cp πολύ μεγαλύτερο του p. Επιθυμούμε μοντέλα με μικρό p και Cp κοντά ή μικρότερο από το p.

Σε κάθε μοντέλο παλινδρόμησης υπάρχουν δύο βασικά προβλήματα: η μεροληψία και η διακύμανση στις προβλέψεις.

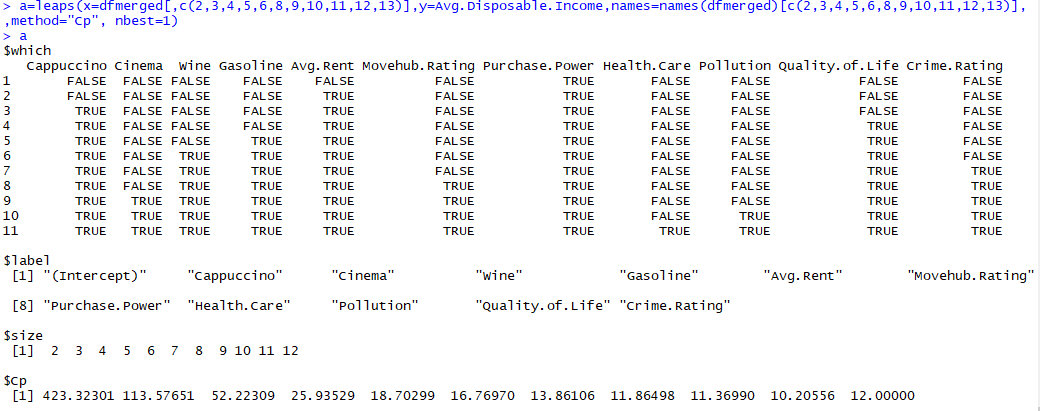
Προκειμένου να ποσοτικοποιήσουμε τη συνολική διακύμανση στις προβλέψεις μας χρησιμοποιούμε το **(standardized) measure of the total variation in the predicted responses Γ***p*

Το Cp μπορεί να χρησιμοποιηθεί ως εκτιμητής του Γp.

Επομένως, σε ένα μοντέλο με p παραμέτρους, δηλαδή p-1 μεταβλητές και το intercept, το μικρό Cp σημαίνει μικρό (standardized) variance of prediction. Όταν η τιμή του Cp είναι κοντά στο p τότε έχουμε μικρή μεροληψία, όταν εείναι πολύ μεγαλύτερο έχουμε σημαντική μεροληψία, ενώ όταν είναι μικρότερο του p το ερμηνεύουμε ως αμεροληψία.

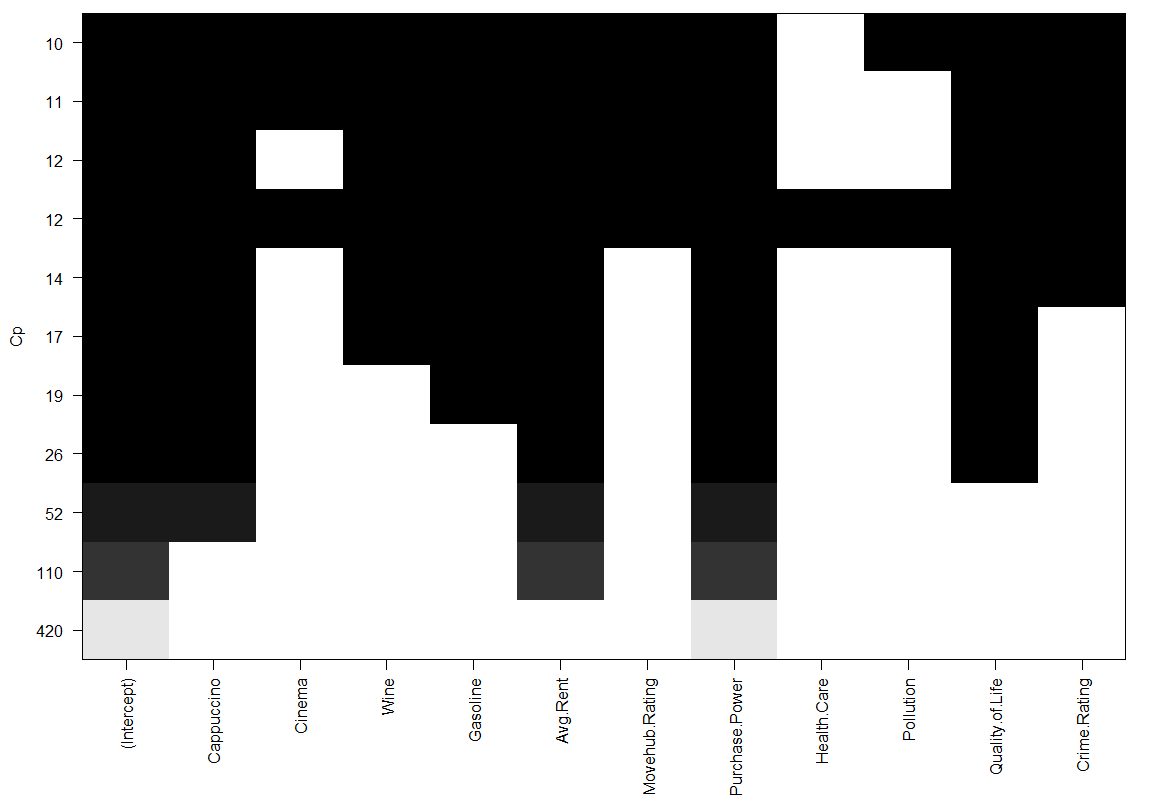
Το μοντέλο με όλες τις μεταβλητές θα έχει πάντα Cp=p, οπότε δεν πρέπει να χρησιμοποιούμε το Cp για να το κρίνουμε.

Μια καλή στρατηγική χρήσης του Cp είναι να βρούμε τα υποσύνολα που το Cp είναι κοντά στο p, χωρίς φυσικά το πλήρες μοντέλο. Αν όλα τα μοντέλα έχουμε μεγάλο Cp ίσως να μην έχουμε πάρει κάποια σημαντική μεταβλητή πρόβλεψης στην ανάλυση μας. Αν μερικά μοντέλα έχουν μικρό Cp παίρνουμε το μοντέλο με το μικρότερο, ώστε να διασφαλίσουμε μικρή μεροληψία και διακύμανση. Σε περίπτωση που δύο ή περισσότερα μοντέλα έχουν μικρό Cp διαλέγουμε το απλούστερο ή αυτό που εξυπηρετεί τις ανάγκες της έρευνας μας.

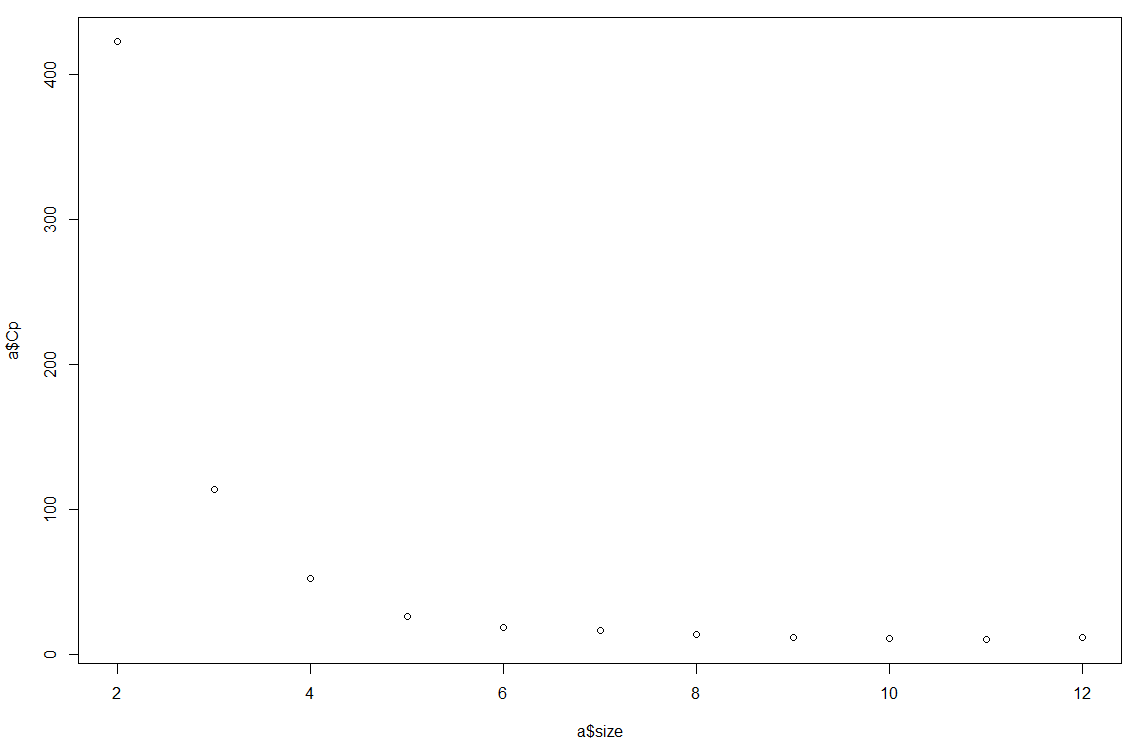


Με την εντολή leaps από το πακέτο leaps και με παραμέτρους τις μεταβλητές του dataset, το κριτήριο Cp και το nbest=1, δηλαδή εμφάνιση του καλύτερου μοντέλου για 1,2,…, 11 μεταβλητές, έχουμε το παραπάνω output.

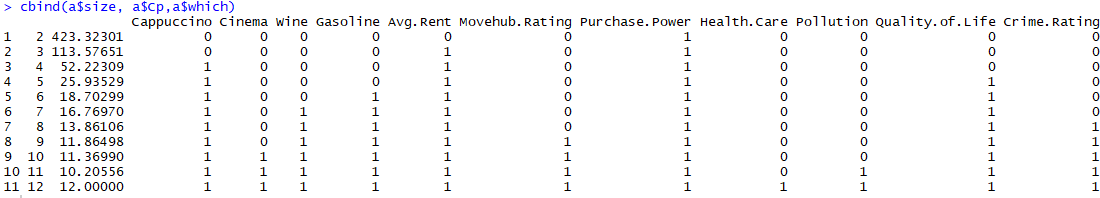
Με true/false παριστάνεται η συμμετοχή ή όχι στο μοντέλο για την κάθε μεταβλητή. Ως size έχουμε το μέγεθος του μοντέλου και ως Cp την τιμή Cp του.



Στο παραπάνω γράφημα έχουμε σε κάθε σειρά το βέλτιστο μοντέλο με 1,2,…,11 παραμέτρους και με μαύρο χρώμα αν η εκάστοτε μεταβλητή συμμετέχει στο μοντέλο, ενώ στον κατακόρυφο άξονα έχουμε την τιμή του Cp.

Στη συνέχεια, έχουμε plot του Cp ανάλογα με το μέγεθος του μοντέλου (για κάθε μέγεθος έχει χρησιμοποιηθεί το βέλτιστο μοντέλο).

Και τέλος, στον πίνακα φαίνονται κατά στήλες: A/A, μέγεθος του μοντέλου, Cp του μοντέλου και στη συνέχεια μια στήλη για κάθε μεταβλητή (με 0/1 σηματοδοτείται η απουσία/ παρουσία της μεταβλητής στο μοντέλο).



Όπως είναι εμφανές με όλα τα παραπάνω, το βέλτιστο μοντέλο είναι αυτό με μέγεθος 11. Το Cp του συγκεκριμένου είναι 10.2 που είναι μικρότερο από το μέγεθος του και σε αυτό χρησιμοποιήθηκαν όλες οι μεταβλητές εκτός της Health Care.

(Το μοντέλο με μέγεθος δεν λαμβάνεται υπόψιν)