ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

Όνομα: Δημήτρης

Επίθετο: Μανωλάκης

Α.Μ.: it1423

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

1η Εργασία

ΠΑΛΙΝΔΡΟΜΗΣΗ

Μέθοδος 1: Πολλαπλή Γραμμική Παλινδρόμηση

Στο μοντέλο χρησιμοποιούνται όλες οι μεταβλητές του dataset. Αρχικά γίνεται Κανονικοποίηση των δεδομένων μέσω της εντολής ***scale***. Στην συνέχεια με τυχαίο τρόπο τα δεδομένα χωρίζονται σε training και testing datasets, και δημιουργείται το μοντέλο:

Εικόνα 1: Summary της Π/πλης Γραμμικής Παλ/σης

```
Call:
lm(formula = effort ~ ., data = DataFrame, subset = train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5210 -0.2018 -0.0848  0.1099  3.6568

Coefficients: (1 not defined because of singularities)
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            -0.05072    0.03140  -1.615 0.107624
AdjustedFunctionPoints  0.13730    0.33335   0.412 0.680779
Inputcount              0.04442    0.22091   0.201 0.840821
Outputcount             0.08514    0.08191   1.039 0.299609
Enquirycount            0.18706    0.04528   4.131 4.96e-05 ***
Filecount              -0.42537    0.12151  -3.501 0.000552 ***
Interfacecount         -0.01519    0.03958  -0.384 0.701422
Addedcount              0.62883    0.18442   3.410 0.000761 ***
Changedcount            0.08981    0.03849   2.333 0.020449 *
Deletedcount                 NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.496 on 244 degrees of freedom
Multiple R-squared:  0.6806,    Adjusted R-squared:  0.6701
F-statistic: 64.98 on 8 and 244 DF,  p-value: < 2.2e-16
```

Στην εικόνα φαίνονται οι 4 μεταβλητές οι οποίες μέσω του p-value χαρακτηρίζονται στατιστικά σημαντικές για το μοντέλο.

Ο συντελεστής της μεταβλητής Deletedcount δεν ορίζεται καθώς είναι ισχυρά συσχετισμένος με άλλες ανεξάρτητες μεταβλητές. Κάνοντας χρήση της εντολής ***alias*** βλέπουμε ότι:

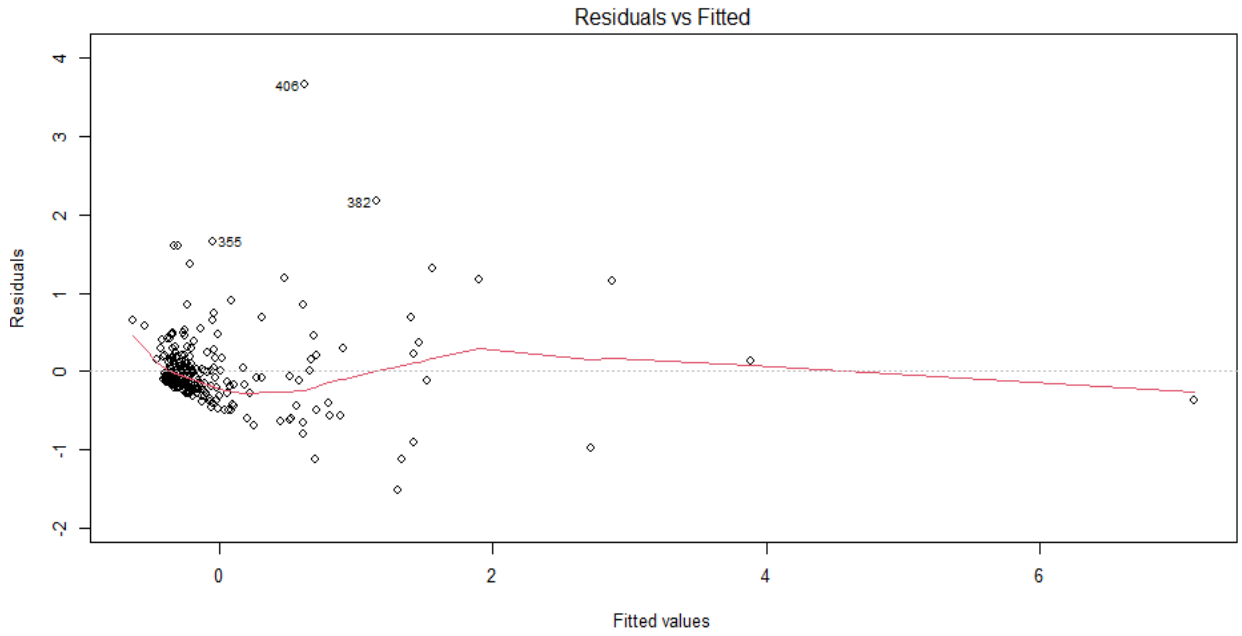
Εικόνα 2: Χρήση της εντολής alias στο μοντέλο.

```
Model :
effort ~ AdjustedFunctionPoints + Inputcount + Outputcount +
    Enquirycount + Filecount + Interfacecount + Addedcount +
    Changedcount + Deletedcount

Complete :
             (Intercept)        AdjustedFunctionPoints Inputcount
Deletedcount                   0                      0           156770/39377
             Outputcount        Enquirycount       Filecount
Deletedcount           8889/5872    2612285/2935752         40032/24451
             Interfacecount     Addedcount          Changedcount
Deletedcount      358963/904496      -303528/43301 -31399447/33595028
```

συσχετίζεται με όλες τις υπόλοιπες μεταβλητές εκτός της AdjustedFunctionPoint.

Εικόνα 3:Residual Plot.



Στο plot των κατάλοιπων φαίνεται ένα U-shape το οποίο υποδεικνύει μια μη γραμμικότητα στα δεδομένα της άσκησης.

Το μοντέλο έχει test Mean Squared Error ίσο με: 0.7718952.

Κάνοντας αναδειγματοληψία με επανάθεση μέσω της μεθόδου Bootstrap στα δεδομένα ελέγχου, εκτιμάται το accuracy του μοντέλου δημιουργώντας τα εξής διαστήματα εμπιστοσύνης για το Mean Squared Error:

(0.5852927, 0.9584976)

μέσω του τύπου (MSE – 2*standard error, MSE + 2* standard error) κάνοντας χρήση των τιμών που υπολογίστηκαν από το bootstrap.

Μέθοδος 2: Subset Selection

Εικόνα 4: Summary της εκτέλεσης του Best Subset Selection.

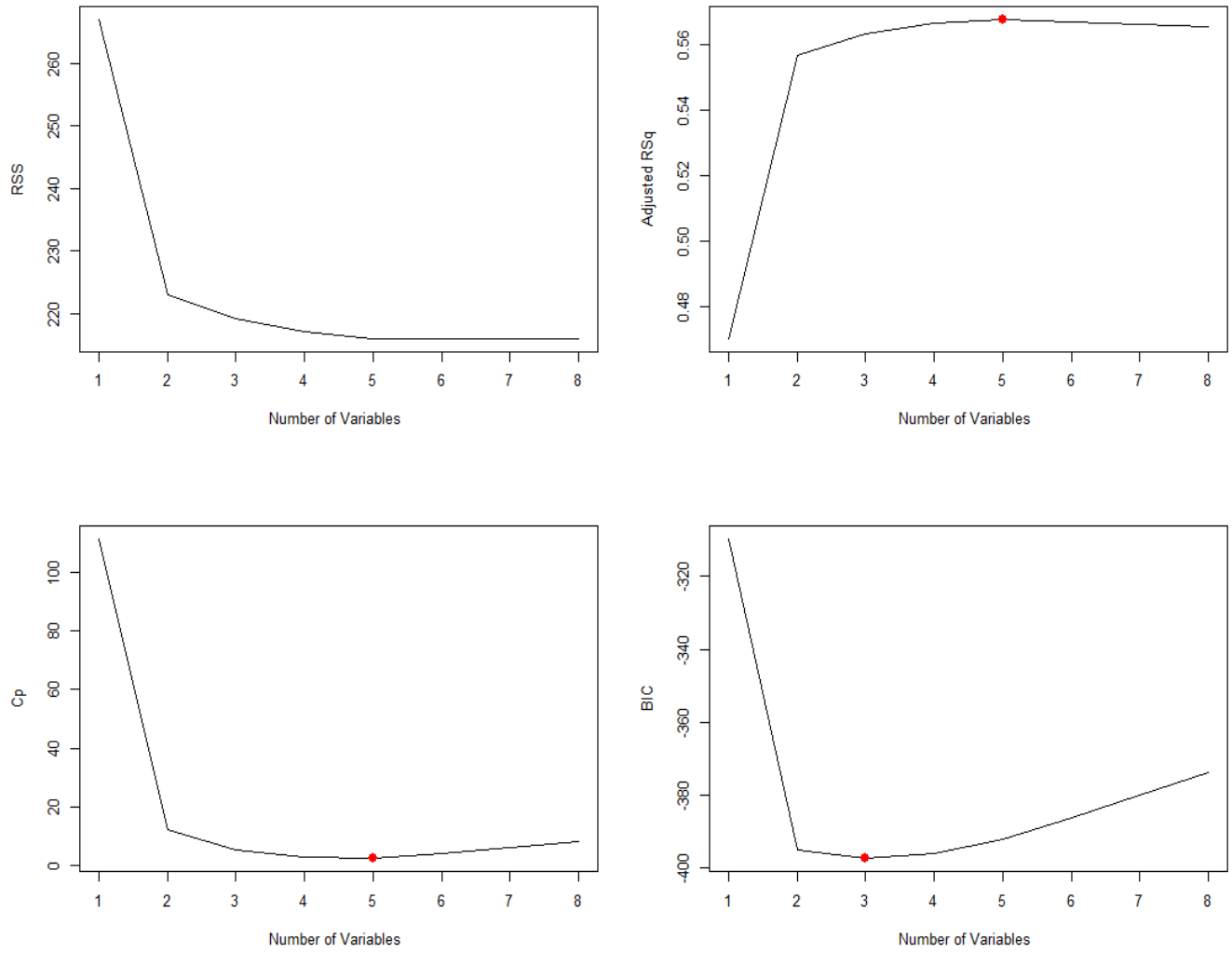
```
Subset selection object
Call: regsubsets.formula(effort ~ ., DataFrame)
9 Variables  (and intercept)
                       Forced in Forced out
AdjustedFunctionPoints     FALSE      FALSE
Inputcount                 FALSE      FALSE
Outputcount                FALSE      FALSE
Enquirycount               FALSE      FALSE
Filecount                  FALSE      FALSE
Interfacecount             FALSE      FALSE
Addedcount                 FALSE      FALSE
Changedcount               FALSE      FALSE
Deletedcount               FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         AdjustedFunctionPoints Inputcount Outputcount Enquirycount Filecount Interfacecount Addedcount Changedcount
1  ( 1 ) " "                    " "        " "         " "          " "       " "            "*"         " "
2  ( 1 ) "*"                    "*"        " "         " "          " "       " "            " "         " "
3  ( 1 ) "*"                    "*"        " "         "*"          " "       " "            " "         " "
4  ( 1 ) "*"                    "*"        " "         "*"          "*"       " "            " "         " "
5  ( 1 ) "*"                    "*"        " "         "*"          "*"       "*"            " "         " "
6  ( 1 ) "*"                    "*"        " "         "*"          "*"       "*"            " "         "*"
7  ( 1 ) "*"                    "*"        "*"         "*"          "*"       "*"            " "         "*"
8  ( 1 ) "*"                    "*"        "*"         "*"          "*"       "*"            "*"         "*"
         Deletedcount
1  ( 1 ) " "
2  ( 1 ) " "
3  ( 1 ) " "
4  ( 1 ) " "
5  ( 1 ) " "
6  ( 1 ) " "
7  ( 1 ) " "
8  ( 1 ) " "
```

Στην παραπάνω εικόνα φαίνεται ποιες μεταβλητές επιλέγονται από το best subset selection για μοντέλα με καθορισμένο πλήθος μεταβλητών από 1 έως 8. Η επιλογή αυτή γίνεται σύμφωνα με το RSS.

Κάνοντας χρήση του Adjusted R-Squared, Cp και του Bayesian Information Criterion βλέπουμε στην επόμενη εικόνα το πλήθος μεταβλητών που επιλέγεται ως

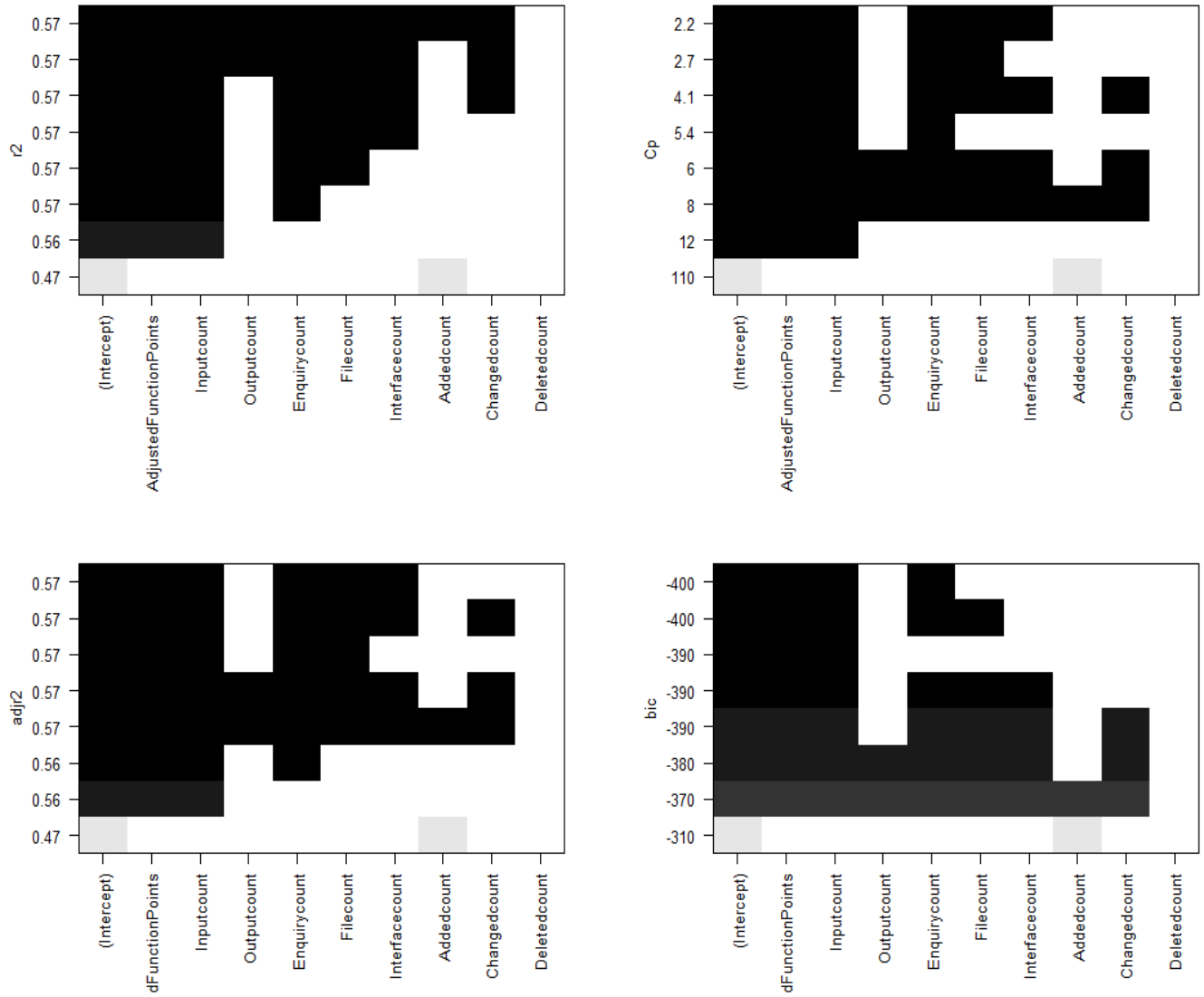
optimal από τα διαφορά κριτήρια. Τα δυο από τα τέσσερα επιλέγουν ως ιδανικό πλήθος μεταβλητών τις 5.

Εικόνα 5: Optimal πλήθος μεταβλητών σύμφωνα με RSS, Adjusted R-Squared, Cp και BIC.



Προκειμένου να βρούμε ποιες 5 μεταβλητές προτείνουν τα κριτήρια αυτά χρησιμοποιείται το παρακάτω plot στην κορυφαία γραμμή του οποίου υπάρχει ένα μαύρο κουτί για κάθε μεταβλητή του μοντέλου που απέδωσε καλυτέρα ανάλογα με το κριτήριο.

Εικόνα 6: Optimal μεταβλητές σύμφωνα με RSS, Adjusted R-Squared, Cp και BIC.



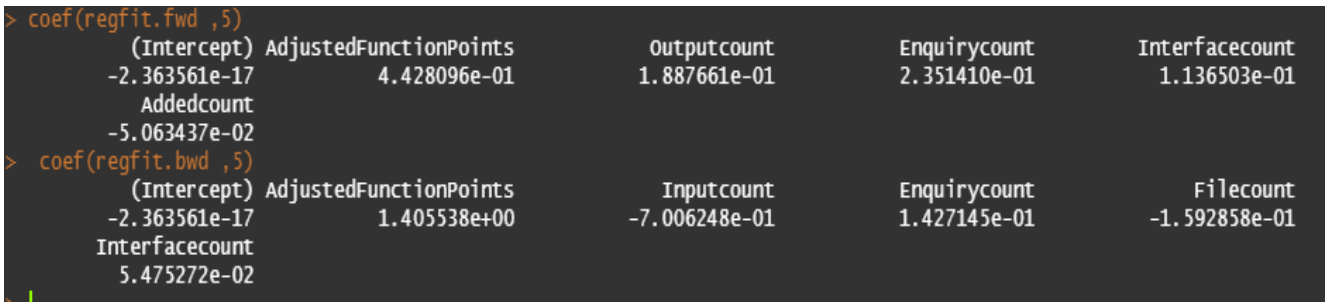
Ο παρακάτω πίνακας περιέχει τις 5 μεταβλητές που επιλέχθηκαν από τα κριτήρια Cp και Adjusted R-Squared

Πίνακας 1: Optimal μεταβλητές σύμφωνα με Adjusted R-Squared και Cp.

AdjustedFunctionPoints	Filecount
Inputcount	Interfacecount
Enquirycount	

Για περαιτέρω επαλήθευση των αποτελεσμάτων εξετάζουμε τα optimal μοντέλα 5 μεταβλητών που προτείνουν οι μέθοδοι Forward και Backward selection.

Εικόνα 7: Optimal μοντέλα 5 μεταβλητών με FWD και BWD selection.



Το Forward selection επιλεγεί διαφορετικές μεταβλητές ενώ το Backward selection επιλεγεί ακριβώς τις ίδιες.

Τρέχουμε επομένως πολλαπλή γραμμική παλινδρόμηση με τις 5 αυτές μεταβλητές και αξιολογούμε τα αποτελέσματα.

Εικόνα 8: Summary της Π/πλης Γραμμικής Παλ/σης

```
Call:
lm(formula = effort ~ AdjustedFunctionPoints + Inputcount + Enquirycount +
    Filecount + Interfacecount, data = DataFrame, subset = train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4725 -0.2078 -0.0976  0.1050  3.8147

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.04499    0.03229  -1.394    0.165
AdjustedFunctionPoints  0.99749    0.18722   5.328 2.24e-07 ***
Inputcount            -0.22568    0.15083  -1.496    0.136
Enquirycount           0.20544    0.04321   4.754 3.38e-06 ***
Filecount             -0.34987    0.07499  -4.665 5.05e-06 ***
Interfacecount        -0.01523    0.03910  -0.390    0.697
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5106 on 247 degrees of freedom
Multiple R-squared:  0.6573,    Adjusted R-squared:  0.6503
F-statistic: 94.74 on 5 and 247 DF,  p-value: < 2.2e-16
```

Από το summary βλέπουμε τις 3 μεταβλητές που είναι στατιστικά σημαντικές σύμφωνα με το p-value.

Το test MSE του μοντέλου είναι ίσο με: 0.705943, σχετικά κοντά με το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης με όλες τις μεταβλητές αλλά κατά 7 μονάδες μικρότερο και επομένως ελαφρά καλύτερο.

Κάνοντας χρήση της μεθόδου Bootstrap υπολογίζεται το εξής διάστημα εμπιστοσύνης για το test MSE:

(0.500185, 0.911701)

Παρατηρούμε πως το διάστημα εμπιστοσύνης βρίσκεται γύρω από μικρότερη τιμή, ίση με αυτή που υπολογίστηκε παραπάνω και επίσης είναι πιο ‘στενό’, γεγονός που υποδηλώνει ότι υπάρχει μεγαλύτερη βεβαιότητα για τις τιμές του MSE και γενικότερα μικρότερο τυπικό λάθος.

Μέθοδος 3: Non Linear Regression

Χρησιμοποιώντας τις πληροφορίες που συλλέχθηκαν με τις 2 προηγούμενες μεθόδους γίνεται προσπάθεια βελτίωσης της απόδοσης του μοντέλου. Αρχικά τα κατάλοιπα στην πρώτη μέθοδο έδειξαν μη γραμμικότητα στα δεδομένα. Επομένως επιλέγεται μια μεταβλητή και εξετάζουμε την πρόσθεση στο μοντέλο, πολυωνυμικού ορού μεγαλύτερης τάξης της μεταβλητής αυτής. Στο μοντέλο διατηρούνται οι 5 μεταβλητές που επιλέχθηκαν προηγουμένως.

Σε πρώτο βήμα, κάνοντας χρήση της εντολής ***poly*** προστίθενται στο μοντέλο πολυωνυμικοί οροί της μεταβλητής AdjustedFunctionPoints μέχρι τον 5^{ο} βαθμό. Στην εικόνα δίνεται το summary του μοντέλου αυτού:

Εικόνα 9: Summary της Π/πλης Μη Γραμμικής Παλ/σης.

```
Call:
lm(formula = effort ~ poly(AdjustedFunctionPoints, 5) + Inputcount +
    Enquirycount + Filecount + Interfacecount, data = DataFrame,
    subset = train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6487 -0.1796 -0.0614  0.1264  3.3900

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -0.03896    0.03299  -1.181 0.238737
poly(AdjustedFunctionPoints, 5)1 20.44201    4.18081   4.889 1.84e-06 ***
poly(AdjustedFunctionPoints, 5)2 -4.59308    1.54390  -2.975 0.003226 **
poly(AdjustedFunctionPoints, 5)3 -2.14580    2.74441  -0.782 0.435047
poly(AdjustedFunctionPoints, 5)4 -3.83939    3.12071  -1.230 0.219777
poly(AdjustedFunctionPoints, 5)5 -2.79153    1.98752  -1.405 0.161439
Inputcount                      -0.04337    0.16897  -0.257 0.797667
Enquirycount                     0.10336    0.05687   1.818 0.070366 .
Filecount                       -0.29921    0.07913  -3.781 0.000196 ***
Interfacecount                  -0.04821    0.04107  -1.174 0.241528
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5003 on 243 degrees of freedom
Multiple R-squared:  0.6763,    Adjusted R-squared:  0.6643
F-statistic: 56.42 on 9 and 243 DF,  p-value: < 2.2e-16
```

Βλέπουμε ότι το πολυώνυμο δευτέρου βαθμού της μεταβλητής αυτής έχει συντελεστή στατιστικά σημαντικό. Ξανατρέχουμε την παλινδρόμηση προσθέτοντας το πολυώνυμο δευτέρου βαθμού μόνο. Εξετάζοντας το test MSE βρίσκουμε ότι βελτιώθηκε σε 0.653612 από 0.705943.

Για περαιτέρω επαλήθευση των αποτελεσμάτων γίνεται χρήση της εντολής ***anova*** προκειμένου τα δυο μοντέλα αυτά να συγκριθούν. Στην παρακάτω εικόνα φαίνεται πως το σχετικό με το F-statistic p-value είναι μικρότερο του 0.05.

Εικόνα 10: Χρήση της εντολής anova στα μοντέλα lm.fit, lm.fit2.

```
> anova(lm.fit ,lm.fit2)
Analysis of Variance Table

Model 1: effort ~ AdjustedFunctionPoints + Inputcount + Enquirycount +
    Filecount + Interfacecount
Model 2: effort ~ poly(AdjustedFunctionPoints, 2) + Inputcount + Enquirycount +
    Filecount + Interfacecount
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    247 64.405
2    246 61.722  1    2.6835 10.695 0.001228 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Επομένως η μηδενική υπόθεση πως τα δυο μοντέλα μοντελοποιούν εξίσου καλά τα δεδομένα απορρίπτεται, πράγμα που σημαίνει ότι η προσθήκη του πολυωνυμικού ορού, πράγματι βελτίωσε το μοντέλο.

Τέλος κάνοντας χρήση του bootstrap δημιουργούμε ένα διάστημα εμπιστοσύνης για το test MSE:

(0.4133736, 0.8939488)

Το διάστημα εμπιστοσύνης βρίσκεται γύρω από μικρότερη τιμή σε σχέση με τα υπόλοιπα μοντέλα με αποτέλεσμα να επιβεβαιώνεται η βελτίωση του μη γραμμικού μοντέλου.