Τεχνικές Μηχανικής Μάθησης

Εργασία 1- Παλινδρόμηση

Πασιοπούλου Ιωάννα (396)
Ταμπάκη Ειρήνη- Μαρία (401)

29/4/2018

A. Εισαγωγή

Η παρούσα εργασία πραγματεύεται τη δημιουργία μοντέλων παλινδρόμησης για δεδομένο data set, πιο συγκεκριμένα για το Computer Hardware Data Set όπως δόθηκε από την εκφώνηση της εργασίας.

B. Δεδομένα

Το data set πραγματεύεται τη σχετική απόδοση της CPU διαφόρων υπολογιστών δεδομένων του κύκλου μηχανής, της μνήμης, της κρυφής μνήμης, των καναλιών και της εκτιμώμενης σχετικής απόδοσης. Έχει 209 δείγματα, 9 ανεξάρτητες μεταβλητές και μία εξαρτημένη.

Αρχικά κάναμε μια επισκόπηση των δεδομένων:

```
> summary(cpu)
      Vendor           Model          MYCT            MMIN            MMAX
 ibm      : 32   100       :  1   Min.   :  17.0   Min.   :   64   Min.   :   64
 nas      : 19   1100/61-h1:  1   1st Qu.:  50.0   1st Qu.:  768   1st Qu.: 4000
 honeywell: 13   1100/81    :  1   Median : 110.0   Median : 2000   Median : 8000
 ncr      : 13   1100/82    :  1   Mean   : 204.2   Mean   : 2881   Mean   :11824
 sperry   : 13   1100/83    :  1   3rd Qu.: 225.0   3rd Qu.: 4000   3rd Qu.:16000
 siemens  : 12   1100/84    :  1   Max.   :1500.0   Max.   :32000   Max.   :64000
 (other)  :106   (other)    :202
      CACH            CHMIN            CHMAX            PRP              ERP
 Min.   :  0.0   Min.   : 0.000   Min.   :  0.00   Min.   :   6.0   Min.   :  15.00
 1st Qu.:  0.0   1st Qu.: 1.000   1st Qu.:  5.00   1st Qu.:  27.0   1st Qu.:  28.00
 Median :  8.0   Median : 2.000   Median :  8.00   Median :  49.5   Median :  45.00
 Mean   : 24.1   Mean   : 4.644   Mean   : 17.74   Mean   : 105.2   Mean   :  98.85
 3rd Qu.: 32.0   3rd Qu.: 6.000   3rd Qu.: 24.00   3rd Qu.: 111.5   3rd Qu.:  99.50
 Max.   :256.0   Max.   :52.000   Max.   :176.00   Max.   :1150.0   Max.   :1238.00
```

Διαγράψαμε τις στήλες Vendor και Model, οι οποίες είναι αναγνωριστικά των μοντέλων CPU και όχι χαρακτηριστικά αυτών. Επίσης, διαγράψαμε την στήλη ERP, επειδή το υπολογισμένο performance είναι κατά προσέγγιση και δεν είναι αντιπροσωπευτικό κάθε CPU.

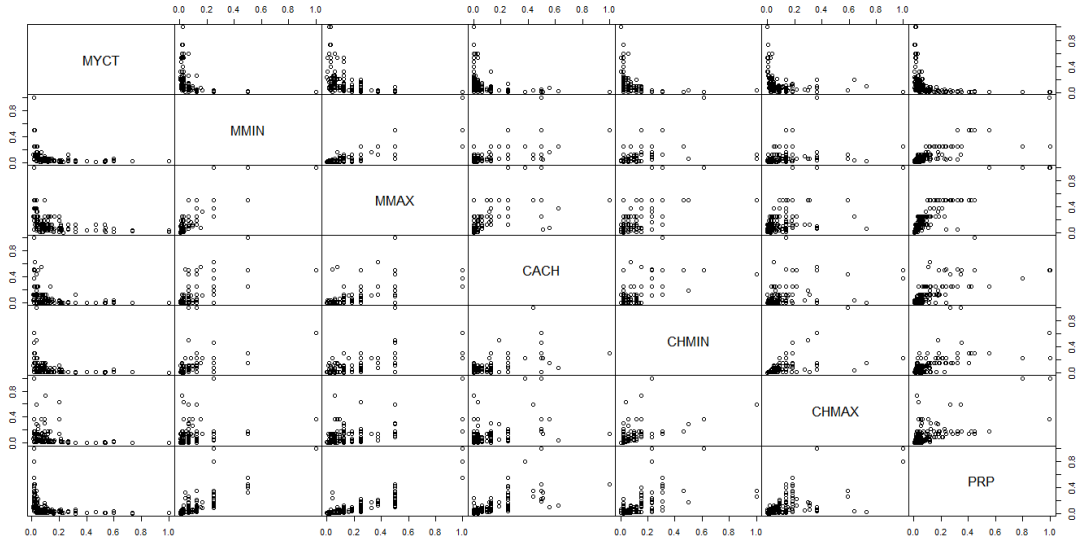
Πριν προχωρήσουμε στην διαγραφή στηλών παρατηρήσαμε ότι υπήρχαν πολλαπλές εγγραφές ίδιου κατασκευαστή- Vendor, για τις οποίες τα πεδία “MYCT", “MMIN”, “MMAX”, “CACH”, “CHMIN, “CHMAX” και”ERP” ταυτίζονται. Πιο συγκεκριμένα για κάθε εταιρεία ίσχυαν τα εξής:

Id	Vendors	Total CPUs	Unique Specified CPUs
1	amdahl	9	6
2	apollo	2	2

3	basf	2	2
4	bti	2	2
5	burroughs	8	8
6	c.r.d	6	6
7	bambex	5	4
8	bdc	9	7
9	dec	6	6
10	dg	7	7
11	formation	5	1
12	four-phase	1	1
13	gould	3	3
14	harris	7	7
15	honeywell	13	11
16	hp	7	7
17	ibm	32	31
18	ipl	6	5
19	magnuson	6	5
20	microdata	1	1
21	nas	19	18
22	ncr	13	12
23	nixdorf	3	3
24	perkin-elmer	3	3
25	prime	5	5
26	siemens	12	12
27	sperry	13	11
28	sratus	1	1
29	wang	2	2

Ωστόσο, επειδή αποφασίσαμε ότι το πεδίο ERP δεν ήταν χρήσιμο για το μοντέλο μας, δεν προχωρήσαμε σε διαγραφή των πολλαπλών εγγραφών.

Επίσης, διαπιστώθηκε ότι χρειαζότανε να κάνουμε scaling για να είναι όλα στο ίδιο εύρος. Έτσι διαιρέσαμε με το μέγιστο για να είναι στο [0,1].




```
> summary(cpu)
      MYCT              MMIN              MMAX              CACH        
 Min.   :0.01133   Min.   :0.00200   Min.   :0.0010   Min.   :0.00000  
 1st Qu.:0.03333   1st Qu.:0.02400   1st Qu.:0.0625   1st Qu.:0.00000  
 Median :0.07333   Median :0.06250   Median :0.1250   Median :0.03125  
 Mean   :0.13613   Mean   :0.09002   Mean   :0.1848   Mean   :0.09413  
 3rd Qu.:0.15000   3rd Qu.:0.12500   3rd Qu.:0.2500   3rd Qu.:0.12500  
 Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000  
     CHMIN             CHMAX              PRP         
 Min.   :0.00000   Min.   :0.00000   Min.   :0.005217  
 1st Qu.:0.01923   1st Qu.:0.02841   1st Qu.:0.023478  
 Median :0.03846   Median :0.04545   Median :0.043043  
 Mean   :0.08931   Mean   :0.10080   Mean   :0.091459  
 3rd Qu.:0.11538   3rd Qu.:0.13636   3rd Qu.:0.096957  
 Max.   :1.00000   Max.   :1.00000   Max.   :1.000000
```

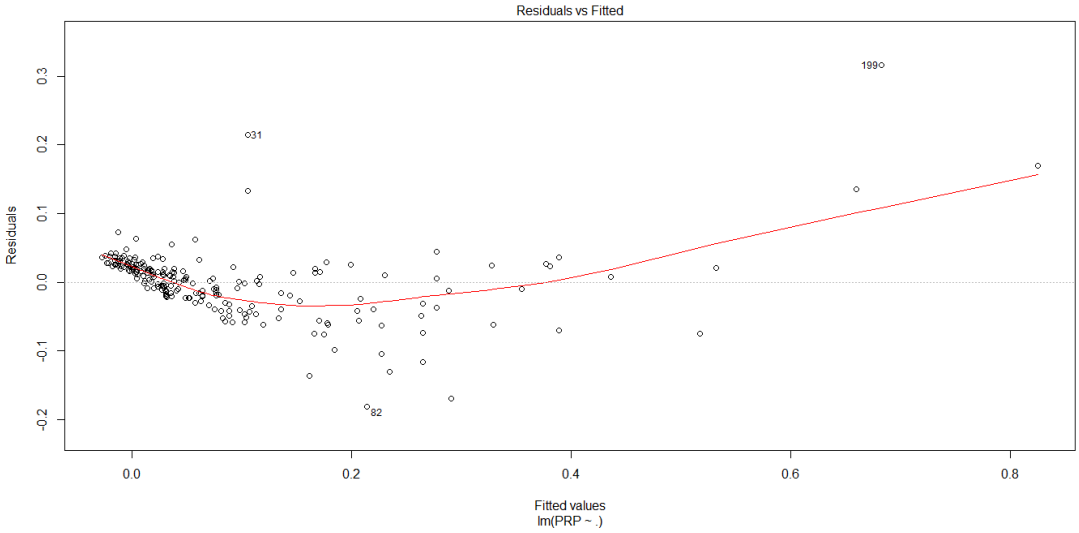
Τέλος, διαπιστώθηκε ότι το δείγμα μας είχε πολλά outliers, ωστόσο αποφασίσαμε να τα κρατήσουμε ως έχουν, λόγω μικρού αριθμού εγγραφών.

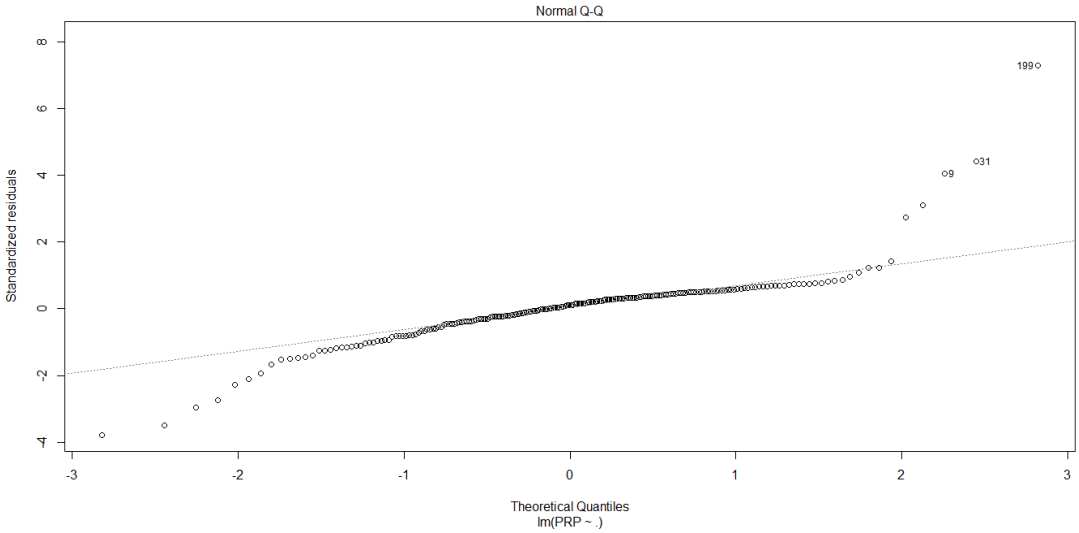
Γ. Μοντέλα

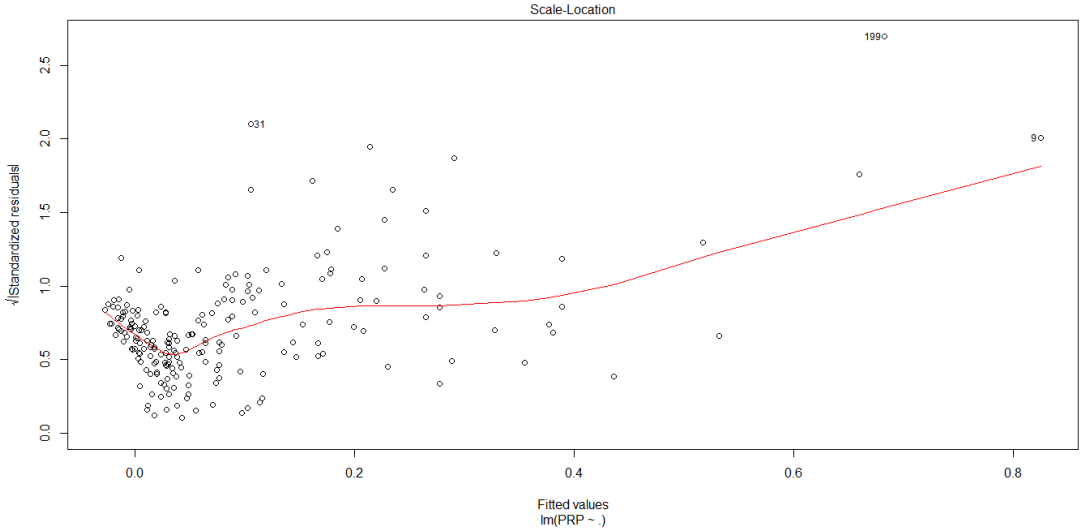
1. Πρώτο απλό μοντέλο χωρίς επεξεργασία δεδομένων

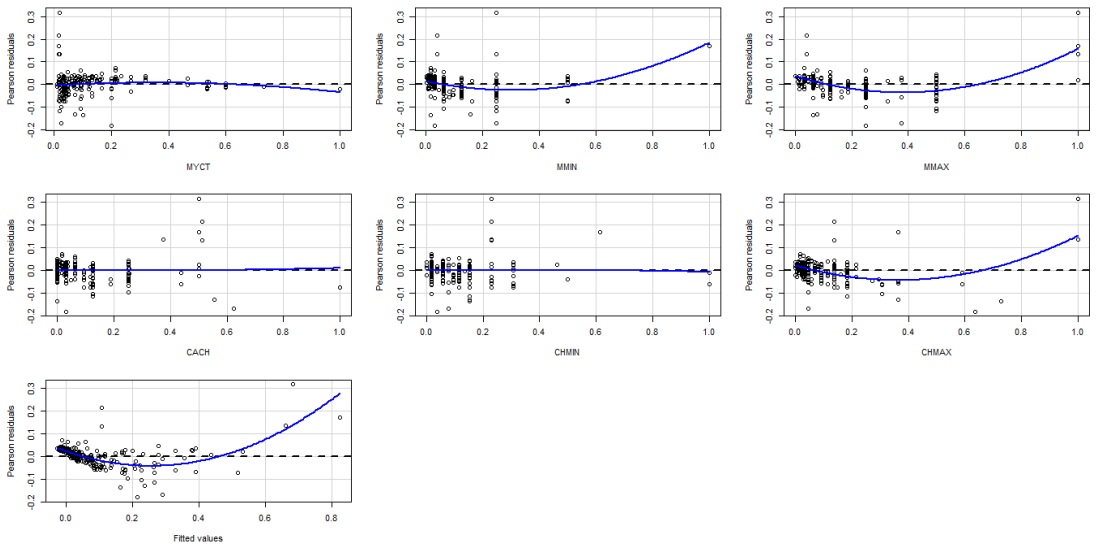
Πείραμα 1. model_simple <- lm(PRP ~ ., data = cpu)

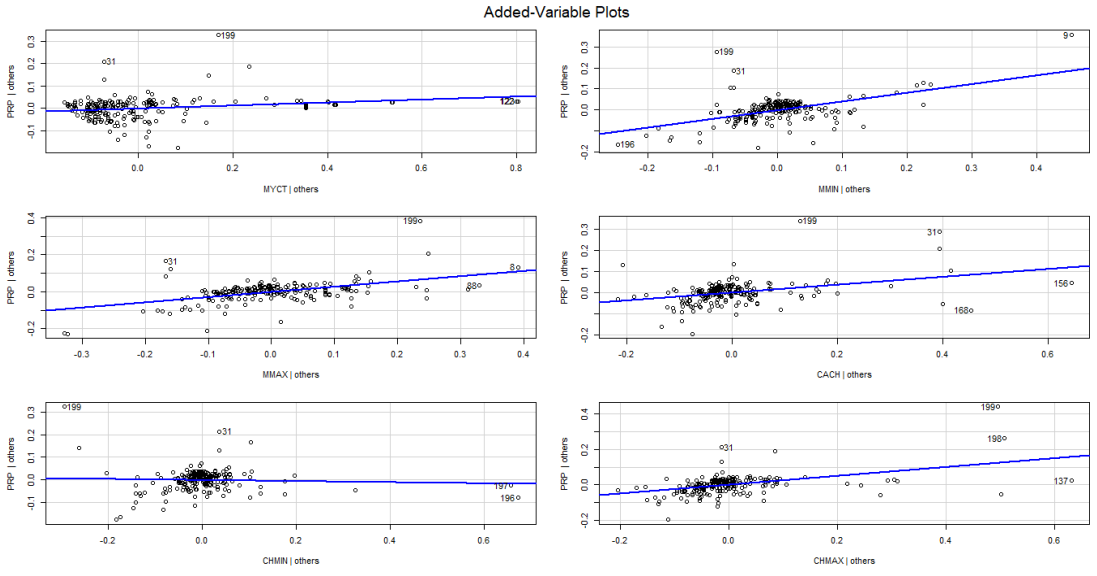
```
Call:
lm(formula = PRP ~ ., data = cpu)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18136 -0.02069  0.00556  0.02385  0.31669

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.048880   0.006876  -7.109 2.00e-11 ***
MYCT         0.066362   0.022476   2.953  0.00353 **
MMIN         0.416279   0.050060   8.316 1.36e-14 ***
MMAX         0.287131   0.036008   7.974 1.14e-13 ***
CACH         0.186721   0.034188   5.462 1.38e-07 ***
CHMIN       -0.024471   0.038264  -0.640  0.52320
CHMAX        0.251946   0.034238   7.359 4.64e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05126 on 201 degrees of freedom
Multiple R-squared:   0.87,     Adjusted R-squared:  0.8661
F-statistic: 224.1 on 6 and 201 DF,  p-value: < 2.2e-16
```

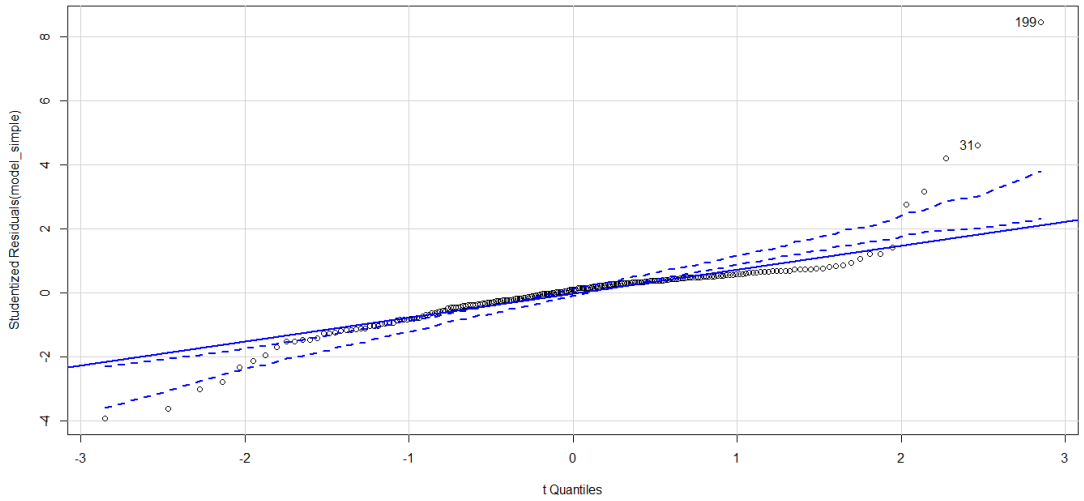





Scale-Location
√|Standardized residuals|
Fitted values
lm(PRP ~ .)
199
31
9


Pearson residuals
MYCT
MMIN
MMAX
CACH
CHMIN
CHMAX
Fitted values


Added-Variable Plots
PRP | others
MYCT | others
MMIN | others
MMAX | others
CACH | others
CHMIN | others
CHMAX | others



Από το summary και μετά από cfs αποφασίσαμε να κρατήσουμε μόνο MMIN, MMAX, CACH, CHMAX.

2. Δεύτερο μοντέλο μετά από αφαίρεση μεταβλητών

Πείραμα 2. model2 <- lm(PRP ~ MMIN + MMAX + CACH + CHMAX, data = cpu)

Πείραμα 3. model3 <- lm(PRP ~ MMIN * MMAX * CACH * CHMAX, data = cpu)

Πέιραμα 4. model4 <- lm(PRP ~ (MMIN + MMAX + CHMAX)*CACH , data = cpu)

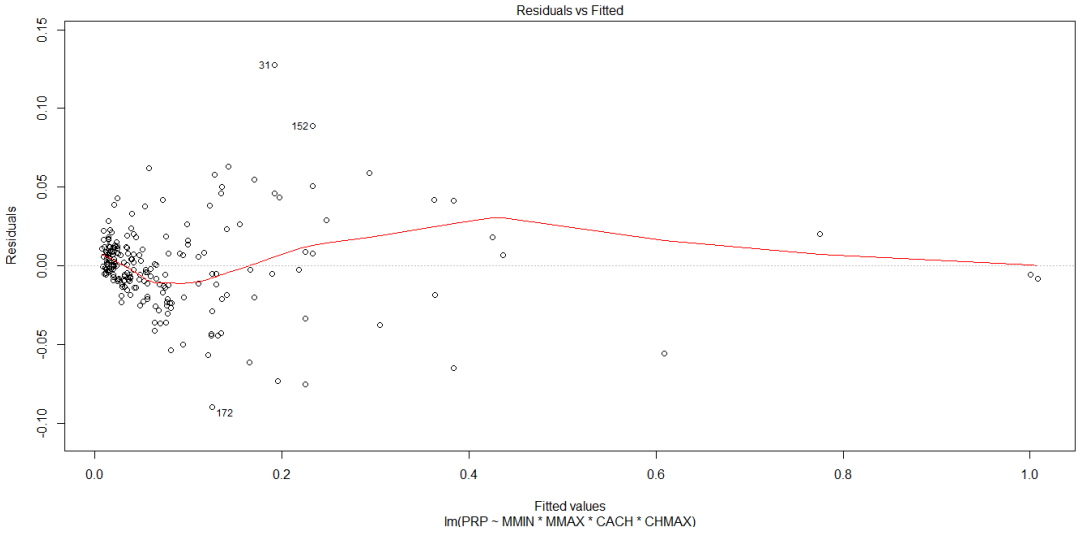
Πείραμα 5. model5 <- lm(PRP ~ (log2(MMIN) + log2(MMAX))*CACH , data = cpu)

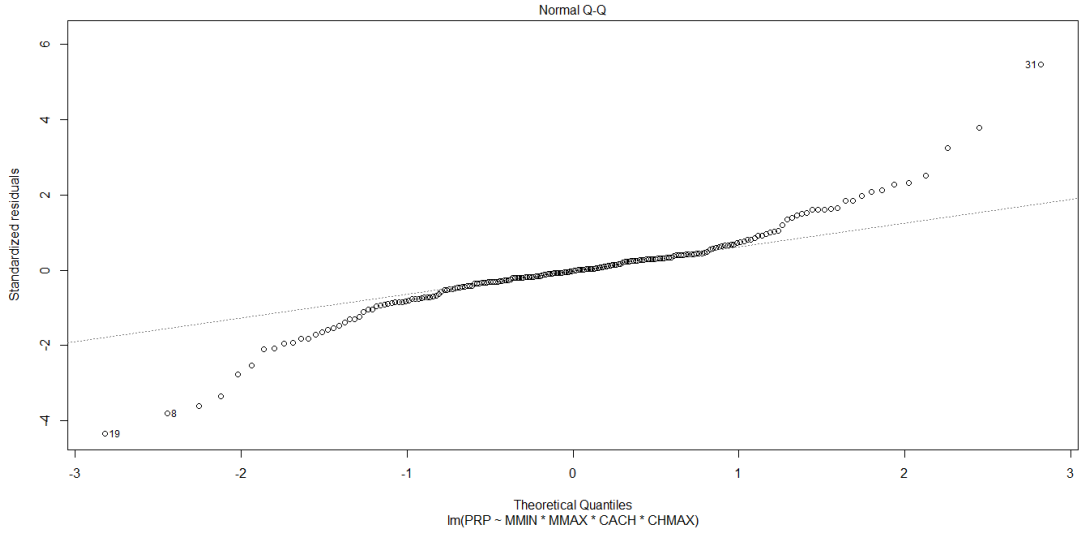
Πείραμα 6. model6 <- lm(PRP ~ log2(MMIN) + log2(MMAX) + CACH , data = cpu)

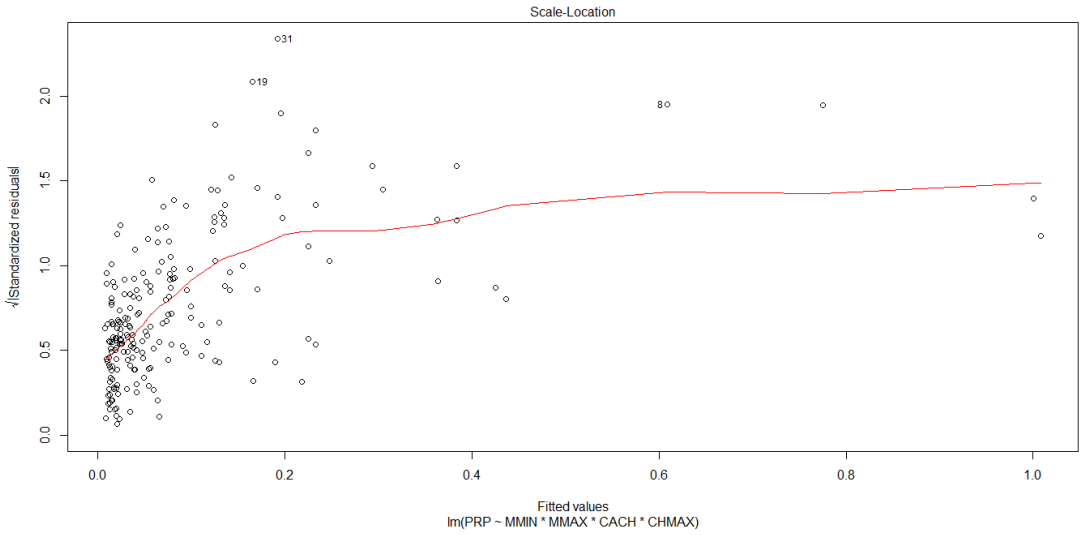
	R^2	R-adjusted
1	0.87	0.8661
2	0.8639	0.8613
3	0.9628	0.9598
4	0.9265	0.924
5	0.6885	0.6807
6	0.5864	0.5803

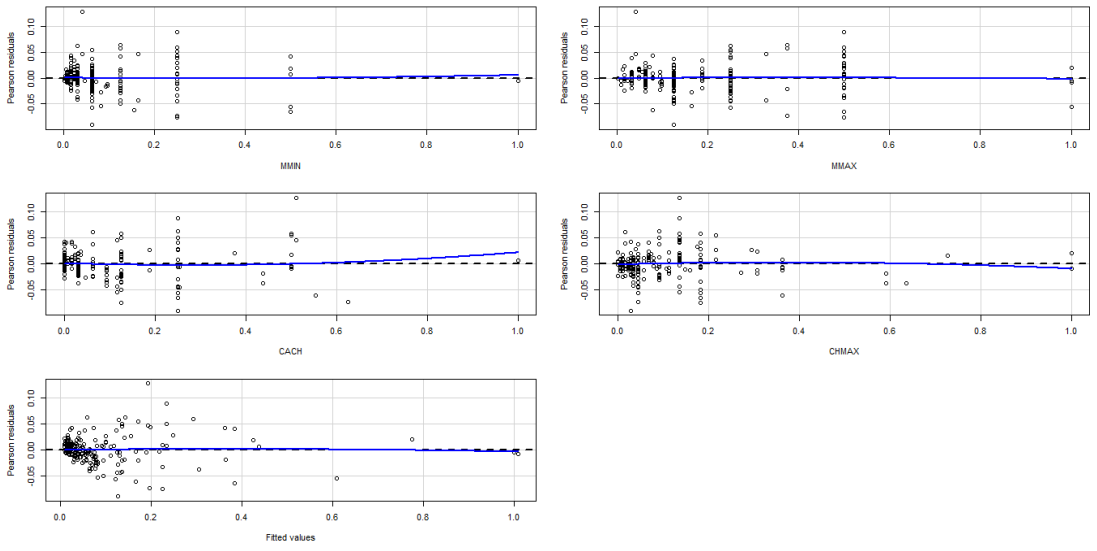
Από τα παραπάνω πειράματα προέκυψε ότι το model3 είναι το καλύτερο. Για αυτό ισχύουν τα εξής:

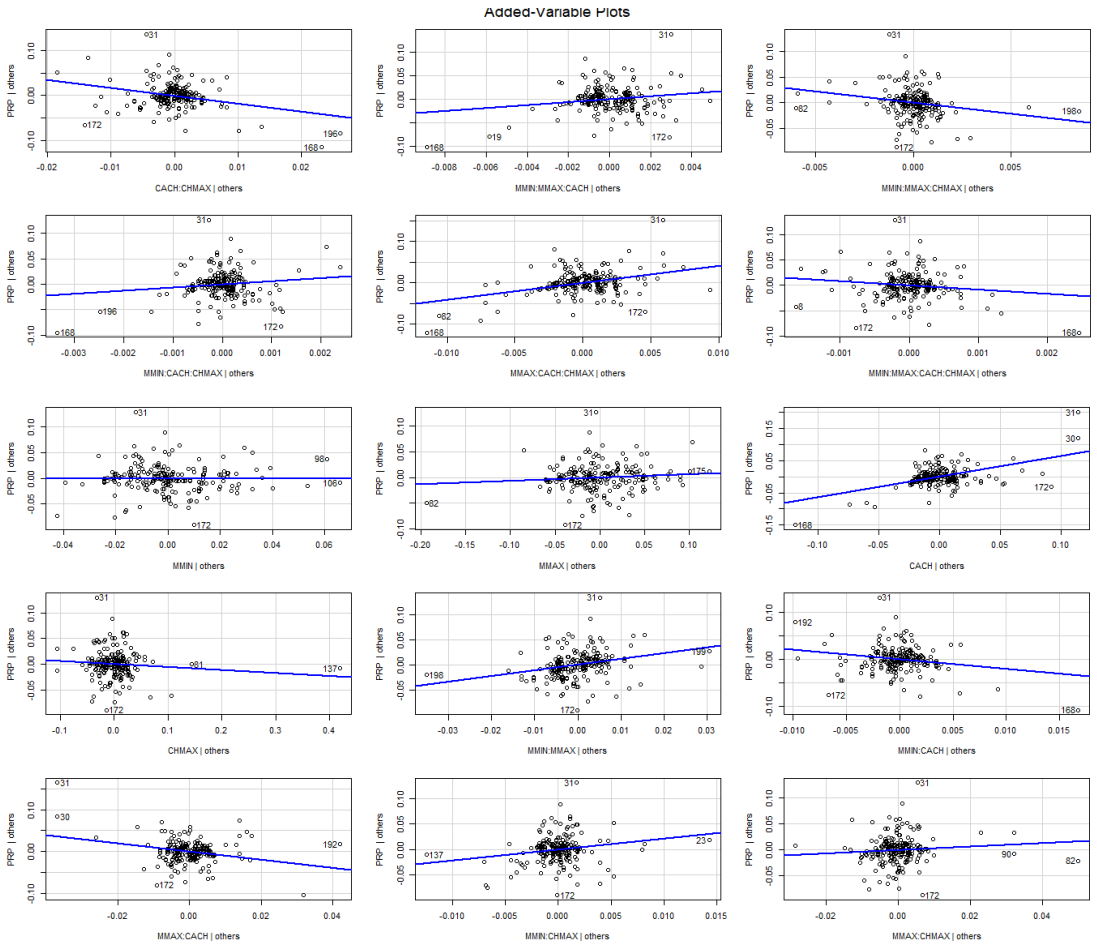
```
Call:
lm(formula = PRP ~ MMIN * MMAX * CACH * CHMAX, data = cpu)

Residuals:
     Min       1Q   Median       3Q      Max
-0.09002 -0.01172 -0.00052  0.01116  0.12781

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.010134   0.005845   1.734 0.084548 .
MMIN               -0.008085   0.119238  -0.068 0.946010
MMAX                0.067416   0.049892   1.351 0.178209
CACH                0.640589   0.080818   7.926 1.80e-13 ***
CHMAX              -0.058380   0.047221  -1.236 0.217847
MMIN:MMAX           1.136465   0.296025   3.839 0.000168 ***
MMIN:CACH          -2.070182   0.724330  -2.858 0.004733 **
MMAX:CACH          -0.998465   0.258086  -3.869 0.000150 ***
MMIN:CHMAX          2.177622   0.824825   2.640 0.008970 **
MMAX:CHMAX          0.317024   0.279851   1.133 0.258699
CACH:CHMAX         -1.755461   0.431723  -4.066 6.97e-05 ***
MMIN:MMAX:CACH      3.170063   1.271544   2.493 0.013510 *
MMIN:MMAX:CHMAX    -4.343970   1.507025  -2.882 0.004395 **
MMIN:CACH:CHMAX     6.349930   3.493364   1.818 0.070667 .
MMAX:CACH:CHMAX     4.135516   0.763833   5.414 1.83e-07 ***
MMIN:MMAX:CACH:CHMAX -8.085443   4.440025  -1.821 0.070158 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02807 on 192 degrees of freedom
Multiple R-squared:  0.9628,    Adjusted R-squared:  0.9598
F-statistic: 330.9 on 15 and 192 DF,  p-value: < 2.2e-16
```

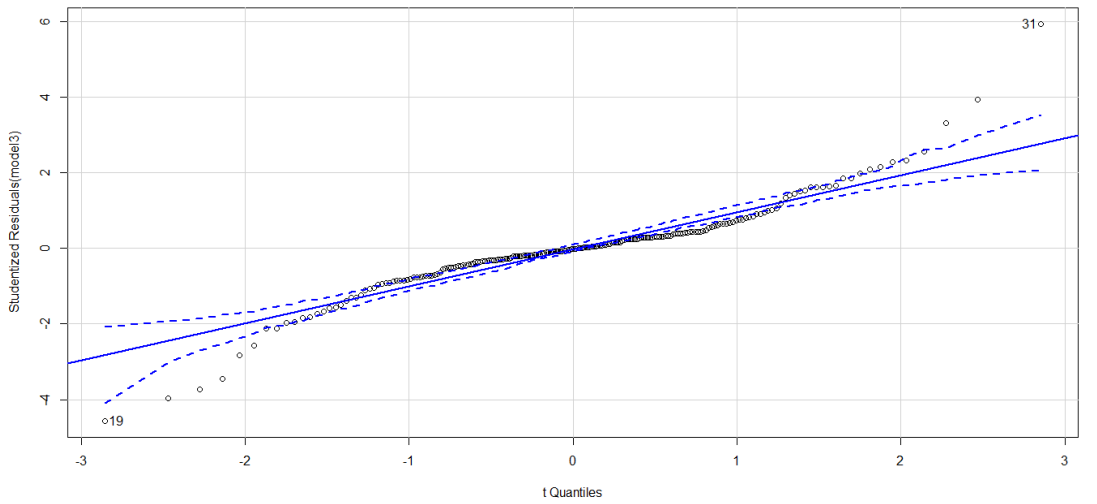



Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(PRP ~ MMIN * MMAX * CACH * CHMAX)
31
8
19


Scale-Location
√|Standardized residuals|
Fitted values
lm(PRP ~ MMIN * MMAX * CACH * CHMAX)
31
19
8


Pearson residuals
MMIN
MMAX
CACH
CHMAX
Fitted values


Added-Variable Plots
PRP | others
CACH:CHMAX | others
MMIN:MMAX:CACH | others
MMIN:MMAX:CHMAX | others
MMIN:CACH:CHMAX | others
MMAX:CACH:CHMAX | others
MMIN:MMAX:CACH:CHMAX | others
MMIN | others
MMAX | others
CACH | others
CHMAX | others
MMIN:MMAX | others
MMIN:CACH | others
MMAX:CACH | others
MMIN:CHMAX | others
MMAX:CHMAX | others


Studentized Residuals(model3)
t Quantiles