Γενικευμένα Γραμμικά μοντέλα
Τρίτη εργασία

Καρατζάς Παντελής

Μάιος 8, 2017

1 Άσκηση 1

2 Άσκηση 2

	country	GDP	LFG	EQP	NEQ	GAP
1	Argentin	0.01	0.01	0.02	0.23	0.61
2	Austria	0.03	0.00	0.10	0.13	0.58
3	Belgium	0.03	0.01	0.07	0.17	0.41
4	Bolivia	0.01	0.02	0.02	0.11	0.86
5	Botswana	0.07	0.02	0.13	0.15	0.95
6	Brazil	0.04	0.03	0.06	0.16	0.85
7	Cameroon	0.05	0.02	0.04	0.09	0.93
8	Canada	0.02	0.03	0.08	0.15	0.18
9	Chile	0.00	0.02	0.02	0.28	0.54
10	Colombia	0.02	0.03	0.02	0.16	0.77
11	CostaRic	0.01	0.04	0.04	0.11	0.70
12	Denmark	0.02	0.01	0.07	0.18	0.41
13	Dominica	0.02	0.03	0.03	0.14	0.83
14	Ecuador	0.03	0.03	0.03	0.21	0.82
15	ElSalvad	0.00	0.03	0.02	0.06	0.84
16	Ethiopia	0.01	0.02	0.02	0.03	0.98
17	Finland	0.03	0.01	0.12	0.25	0.56
18	France	0.03	0.01	0.09	0.18	0.47
19	Germany	0.03	0.00	0.09	0.19	0.46
20	Greece	0.04	0.00	0.07	0.22	0.79
21	Guatemal	0.01	0.02	0.04	0.05	0.79
22	Honduras	0.01	0.03	0.04	0.10	0.89
23	HongKong	0.05	0.04	0.08	0.12	0.75
24	India	0.01	0.02	0.03	0.14	0.94
25	Indonesi	0.03	0.02	0.02	0.12	0.92
26	Ireland	0.03	0.01	0.08	0.19	0.65
27	Israel	0.05	0.03	0.11	0.18	0.68
28	Italy	0.04	0.00	0.07	0.18	0.54
29	IvoryCoa	0.03	0.03	0.02	0.10	0.92
30	Jamaica	0.01	0.02	0.06	0.15	0.82

31	Japan	0.05	0.01	0.12	0.25	0.75
32	Kenya	0.01	0.03	0.05	0.13	0.94
33	Korea	0.05	0.03	0.06	0.18	0.88
34	Luxembou	0.02	0.01	0.07	0.19	0.29
35	Madagasc	-0.01	0.02	0.02	0.05	0.92
36	Malawi	0.02	0.02	0.04	0.09	0.96
37	Malaysia	0.03	0.03	0.04	0.19	0.79
38	Mali	0.00	0.02	0.04	0.03	0.95
39	Mexico	0.02	0.03	0.03	0.17	0.59
40	Morocco	0.02	0.03	0.03	0.05	0.84
41	Netherla	0.02	0.01	0.08	0.18	0.36
42	Nigeria	-0.00	0.03	0.04	0.08	0.86
43	Norway	0.03	0.01	0.07	0.22	0.38
44	Pakistan	0.03	0.03	0.03	0.09	0.92
45	Panama	0.03	0.03	0.04	0.22	0.80
46	Paraguay	0.03	0.03	0.02	0.10	0.85
47	Peru	0.01	0.03	0.03	0.09	0.74
48	Philippi	0.02	0.03	0.04	0.10	0.87
49	Portugal	0.03	0.01	0.07	0.16	0.80
50	Senegal	-0.00	0.03	0.02	0.08	0.89
51	Spain	0.04	0.01	0.04	0.13	0.66
52	SriLanka	0.01	0.02	0.01	0.14	0.86
53	Tanzania	0.02	0.03	0.09	0.09	0.98
54	Thailand	0.03	0.03	0.04	0.14	0.92
55	Tunisia	0.03	0.03	0.04	0.10	0.78
56	U.K.	0.02	0.00	0.07	0.11	0.43
57	U.S.	0.01	0.02	0.08	0.14	0.00
58	Uruguay	0.00	0.01	0.02	0.12	0.58
59	Venezuel	0.01	0.04	0.03	0.08	0.50
60	Zambia	-0.01	0.03	0.07	0.20	0.87
61	Zimbabwe	0.01	0.03	0.08	0.13	0.89

Προσαρμόζοντας ένα γραμμικό μοντέλο στα δεδομένα προκύπτει το παρακάτω:

$$\hat{y} = -0.0143 - 0.0298x_1 + 0.2654x_2 + 0.0624x_3 + 0.0203x_4$$

Κάνοντας anova τεστ βλέπουμε τις σημαντικές παρατηρήσεις για το μοντέλο. Από αυτές προκύπτει η μόνη σημαντική να είναι αυτή του EQP (Επένδυση σε εξοπλισμό) για εισαγωγή της στο μοντέλο. Παρακάτω θα δούμε τεχνικές επιλογής αυτών.

Παρακάτω δίνεται και το Correlations plot το οποίο είναι ένα διάγραμμα που θα μας δώσει τις όποιες πιθανές συσχετίσεις υπάρχουν στα δεδομένα.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0143	0.0103	-1.39	0.1697
LFG	-0.0298	0.1984	-0.15	0.8811
EQP	0.2654	0.0653	4.06	0.0002
NEQ	0.0624	0.0348	1.79	0.0787
GAP	0.0203	0.0092	2.21	0.0313

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
LFG	1	0.00	0.00	1.74	0.1930
EQP	1	0.00	0.00	20.47	0.0000
NEQ	1	0.00	0.00	1.61	0.2092
GAP	1	0.00	0.00	4.88	0.0313
Residuals	56	0.01	0.00		

Για να δούμε αν υπάρχει πολυσυγγραμικότητα κάνουμε το αντίστοιχο τεστ στην R, $vif()$. Οι τιμές που προκύπτουν είναι είναι

$$1.328094 \gamma\iota\alpha\tau o LFG,$$
$$1.316094 \gamma\iota\alpha\tau o EQP,$$
$$1.384458 \gamma\iota\alpha\tau o NEQ,$$
$$1.408114 \gamma\iota\alpha\tau o GAP.$$

Η τετραγωνική ρίζα των τιμών του test είναι η :

$$1.152429 1.147212 1.176630 1.186640 < 2$$

Μπορούμε να πούμε λοιπόν ότι δεν υπάρχει πολυσυγγραμμικότητα Για να δούμε αν το μοντέλο προσαρμόζεται καλά στα δεδομένα προχοράμε σε διαγνωστικούς ελέγχους των υπολοίπων. Παρακάτω βλέπουμε τα διαγράμματα ελέγχου προσαρμογής καθώς και κάποια άλλα όπως αποστάσεις σημείων επιρροής κλπ..

Από τα παρακάτω διαγράμματα βλέπουμε πως το γραμμικό μοντέλο προσαρμόζεται σχετικά καλά στα δεδομένα. Παρόλα αυτά υπάρχουν κάποια σημεία όπως η παρατήρηση 60 που φαίνεται να δημιουργούν κάποια προβλήματα. Παρακάτω θα δούμε παραπάνω ελέγχους για αυτές τις περιπτώσεις.

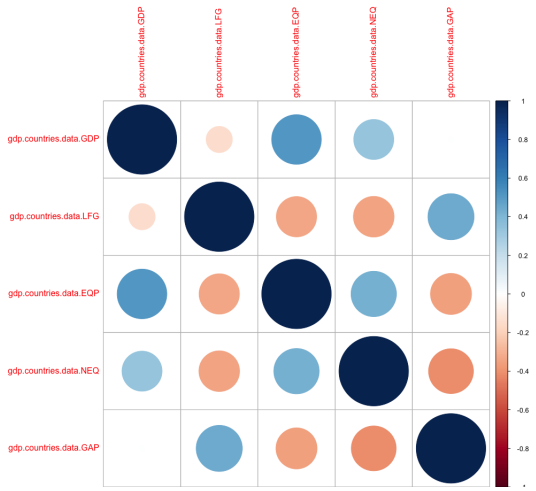

Figure 1: Correlations Plot

2.1 Ερώτημα 2

Καταρχήν χρησιμοποιούμε την stepwise τεχνική επιλογής καλύτερου μοντέλου. Ξεκινάμε είτε με ολόκληρο το μοντέλο και αφαιρόντας μεταβλητές με κάποιο τεστ είτε F-test ή το κριτήριο AIC.
Ξεκινόντας με το AIC test. Γενική αρχή είναι να διαλέγουμε το μοντέλο με την μικρότερη τιμή AIC.

Ο πίνακας Stepwise Backward Elimination AIC μας δείχνει το μοντέλο που επιλέγεται αν ξεκινήσουμε με όλες τις παραμέτρους και αφαιρούμε με βάση την τιμή AIC. Η τιμή AIC για το μοντέλο που καταλήγουμε με Backward Elimination είναι -526.46.

Ξεκινόντς με Forward Ellimination καταλήγουμε στο επόμενο μοντέλο:

Start: AIC=-524.48 GDP LFG + EQP + NEQ + GAP

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0148	0.0097	-1.53	0.1308
EQP	0.2670	0.0638	4.18	0.0001
NEQ	0.0631	0.0342	1.84	0.0705
GAP	0.0198	0.0086	2.30	0.0253

Table 1: Stepwise Backward Elimination AIC

Αντίστοιχα μπορούμε να κάνουμε και πρως τις δύο κετευθύνσεις Elimination: Καταλήγουμε στο ίδιο μοντέλο με την τεχνική Backward Ellimination.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0148	0.0097	-1.53	0.1308
EQP	0.2670	0.0638	4.18	0.0001
NEQ	0.0631	0.0342	1.84	0.0705
GAP	0.0198	0.0086	2.30	0.0253

Πέρα από το AIC test μπορούμε να χρησιμοποιήσουμε F-test για να πραγματοποιήσουμε την αντίστοιχη διαδικασία. Παρακάτω εμφανίζονται τα μοντέλα που επιλέγονται.
Για both direction το μοντέλο που προκύπτει είναι το παρακάτω :

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0148	0.0097	-1.53	0.1308
EQP	0.2670	0.0638	4.18	0.0001
GAP	0.0198	0.0086	2.30	0.0253
NEQ	0.0631	0.0342	1.84	0.0705

Table 2: Stepwise Both Elimination with f-test

Με την επιλογή Forward elimination προκύπτει το παρακάτω:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0148	0.0097	-1.53	0.1308
EQP	0.2670	0.0638	4.18	0.0001
GAP	0.0198	0.0086	2.30	0.0253
NEQ	0.0631	0.0342	1.84	0.0705

Table 3: Stepwise Forward Elimination with f-test

Με την επιλογή Backward elimination προκύπτει το παρακάτω:

Για να χρησιμοποιήσουμε το R^2 τεστ προσαρμόζουμε τα διάφορα μοντέλα και κάνουμε anova σύγκριση των όσων μοντέλων προκύπτουν.

Έχουμε ήδη δημιουργήσει το πρώτο μοντέλο $first.model$. Προσαρμόζουμε άλλα τέσσερα μοντέλα με τρείς από τις τέσσερις επεξηγηματικές μεταβλητές και πραγματοποιούμε τα τεστ.

Για το μοντέλο χωρίς EQP (όπου 2 το μοντέλο χωρίς το EQP).

Για το μοντέλο χωρίς το GAP (όπου 2 το μοντέλο χωρίς το GAP).

Για το μοντέλο χωρίς το LFG (όπου 2 το μοντέλο χωρίς το LFG).

Για το μοντέλο χωρίς το LFG (όπου 2 το μοντέλο χωρίς το NEQ).

Προκύπτει ότι οποιοδήποτε μοντλέλο με τρείς επεξηγηματικές προσαρμόζεται καλύτερα στα δεδομένα. Κάνουμε αντίστοιχο τέστ για να επιλέξουμε ένα από τα τέσσερα αυτά. Με βάση τις τιμές RSS και Sum of squares καλλίτερη επιλογή μοντέλου με βάση το R^2 test είναι το τρίτο. Αυτό χωρίς το LFG .

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0148	0.0097	-1.53	0.1308
EQP	0.2670	0.0638	4.18	0.0001
GAP	0.0198	0.0086	2.30	0.0253
NEQ	0.0631	0.0342	1.84	0.0705

Table 4: Stepwise Backward Elimination with f-test

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	56	0.01				
2	57	0.01	-1	-0.00	16.52	0.0002

Table 5: Anova σύγκριση με πλήρες και χωρίς EQP μοντέλο

R^2 value model without EQP 0.143819 R^2 value model without GAP 0.2812658 R^2 value model without LFG 0.338576 R^2 value model without NEQ 0.300978

Δεν χρειάζεται να συνεχίσουμε σε αφαίρεση μεταβλητών καθότι οι R^2 τιμές χωρίς 2 μεταβλητές είναι $0.1119343, 0.277298, 0.2622098$ κοκ..

Άλλο ένα διαγνωστικό τεστ είναι το να βρεθούν ο Cp values. Παρακάτω δίνεται το διάγραμμα αυτών των τιμών. Καλύτερο μοντέλο θεωρείται αυτό που οι Cp value ειναι μικρότερη από την P-value του αντίστοιχου μοντέλου. Η $p-value$ του "Πρώτου μοντέλου με όλες τις μεταβλητές " είναι 0.1930475. Από την σύγκριση με τις Cp -mallows τιμές βλέπουμε ότι το μοντέλο με αριθμό 11 έχει καλλίτερη προβλεπτική ικανότητα. Αυτο περιέχει τις επεξηγηματικές μεταβλητές

Για να δούμε τυχών σημεία επιρροής στο μοντέλο προχοράμε σε αντίστοιχα διαγράμματα όπως αυτών των Hat values, Coocks distance, standarized errors, df-bettas τιμές. Τα διαγράμματα ακολουθούν παρακάτω. Σε όλα τα διαγράμματα μπορούμε να δούμε ότι υπάρχουν τιμές που μπορούν να θεωρηθούν σημεία επιρροής. Ειδικά για την 60ή παρατήρηση. Αν θέλουμε να έχουμε ένα μοντέλο με καλλίτερη προβλεπτική ικανότητα θα πρέπει να αφαιρέσουμε την συγκεκριμενη μεταβλητή. Πιθανά και τις 5 και 8.

Τέλος πραγματοποιούμε και τα Added variable , Component residuals plots τα οποία επίσης μας δείνουν πληροφορία για το αν θα πρέπει ή όχι να μπούνε μεταβλητές στο μοντέλο. Αυτό γίνεται με την γραμμικότητα ή μή των διαγραμμάτων όπως φαινεται σε αυτά.

Ένας συγκεντρωτικός πίνακας των κριτηρίων επιλογής μοντέλου δίνεται παρακάτω:

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	56	0.01				
2	57	0.01	-1	-0.00	4.88	0.0313

Table 6: Anova σύγκριση με πλήρες και χωρίς GAP μοντέλο

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	56	0.01				
2	57	0.01	-1	-0.00	0.02	0.8811

Table 7: Anova σύγκριση με πλήρες και χωρίς LFG μοντέλο

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	56	0.01				
2	57	0.01	-1	-0.00	3.21	0.0787

Table 8: Anova σύγκριση με πλήρες και χωρίς NEQ μοντέλο

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	57	0.0123676	0			
2	57	0.0103821	0	0.00198543		
3	57	0.0095543	0	0.00082785		
4	57	0.0100974	0	-0.00054310		

Table 9: Σύγκριση μοντέλων τριών μεταβλητών

	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik
1	0.34	0.29	0.01	7.17	0.00	5	180.69
	AIC	BIC	deviance	df.residual			
	-349.37	-336.71	0.01	56			

Table 10: Συγκεντρωτικός πίνακας τιμών κριτηρίων επιλογής μοντέλου για το πρώτο μοντέλο

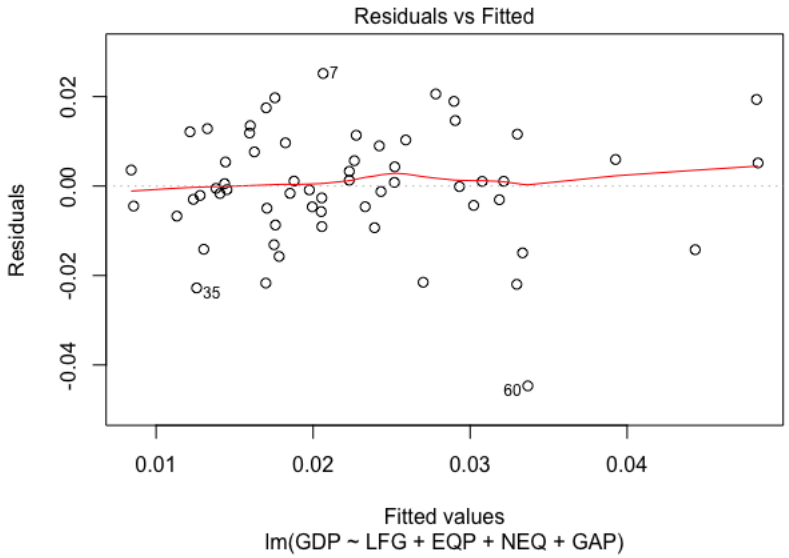

Figure 2: Residuals of the initial model

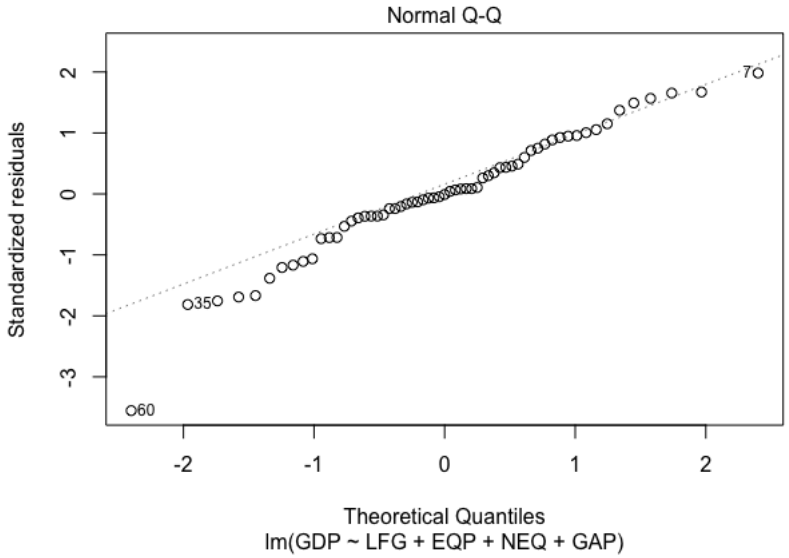

Figure 3: Normal Q-Q test

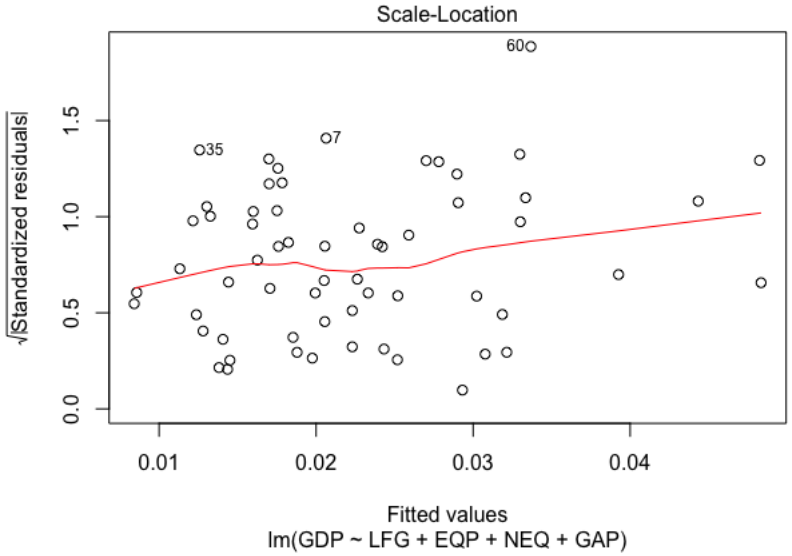

Figure 4: Standarized Residuals plot

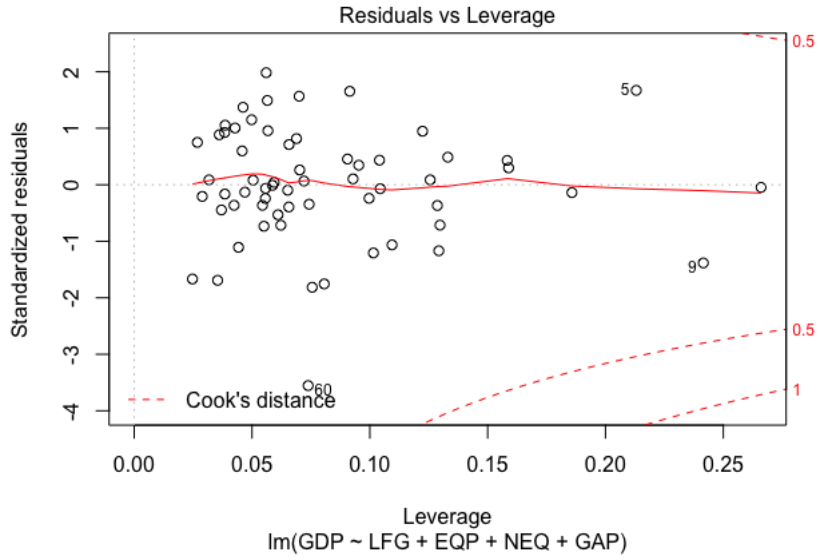

Figure 5: Cooks Distance

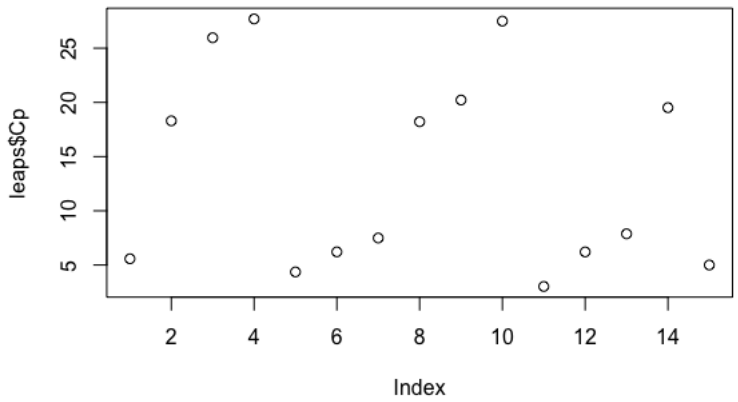

Figure 6: Cp values plot

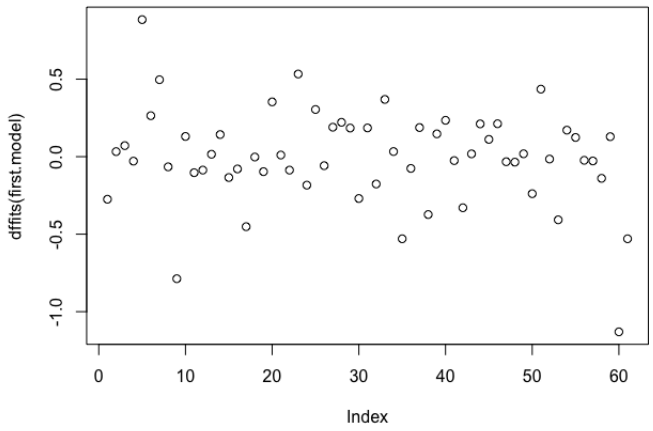

Figure 7: Df fits values plot

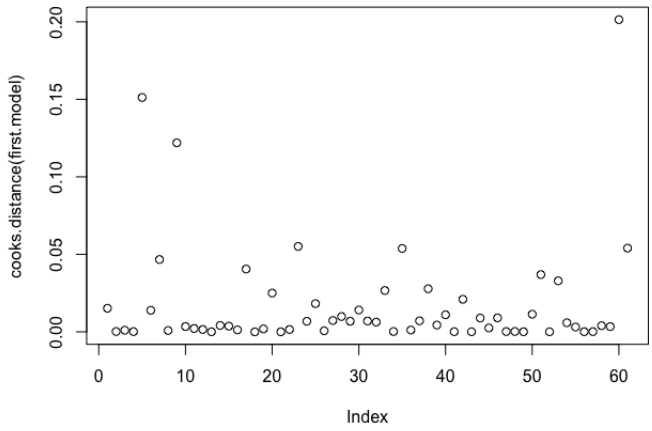

Figure 8: Cooks Distance plot

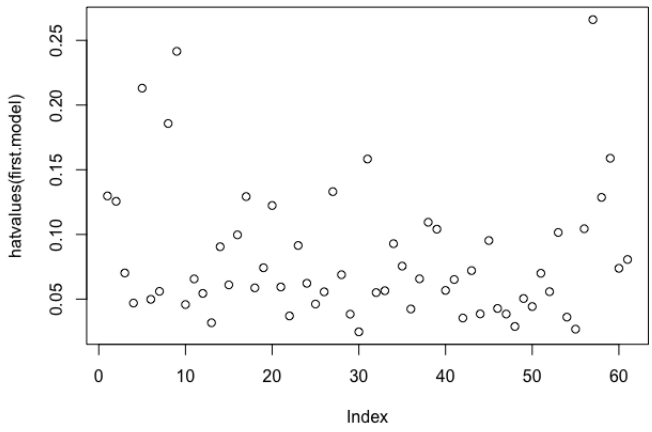

Figure 9: Hat Values plot

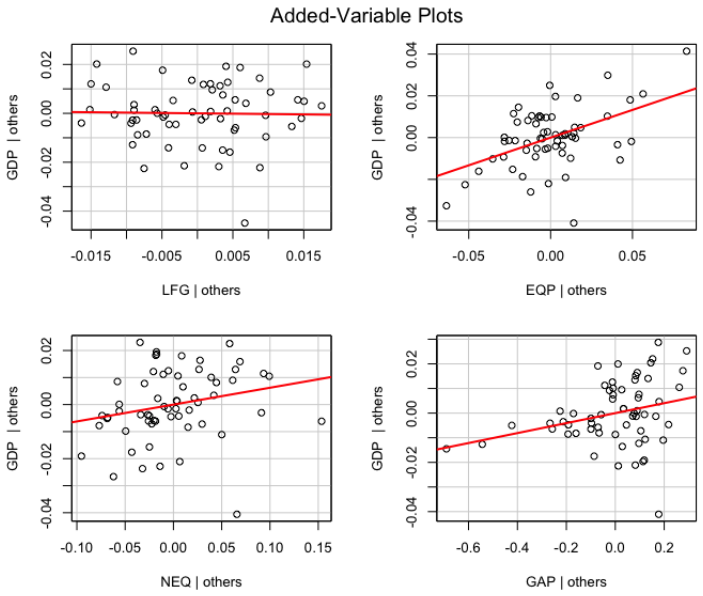

Figure 10: Added variables plot

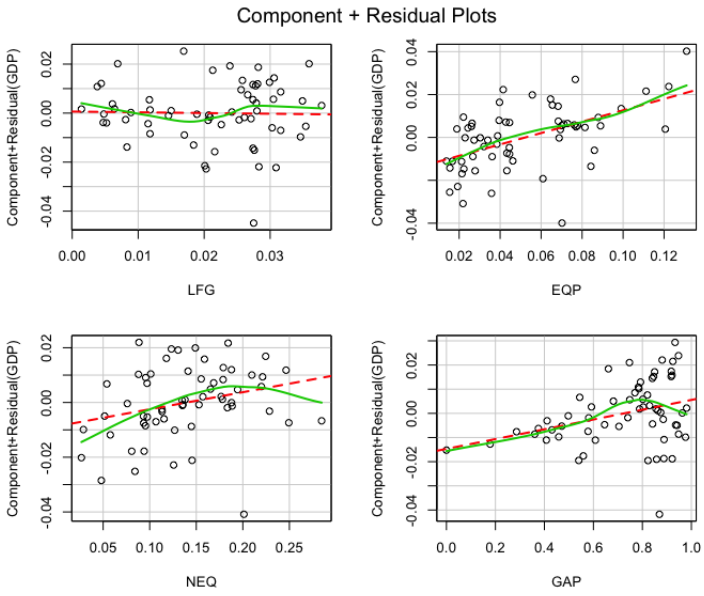

Figure 11: Component residuals plots

R-Squared	Higher the better	(> 0.70)
Adj R-Squared	Higher the better	
F-Statistic	Higher the better	
Std. Error	Closer to zero the better	
t-statistic	Should be greater 1.96	p-value < 0.05
AIC	Lower the better	
BIC	Lower the better	
Mallows cp	Should be close to the number of predictors in model	
MAPE	Lower the better	
MSE (Mean sq. error)	Lower the better	
Min Max Accuracy	mean(min(actual, predicted)/ max(actual, predicted))	Higher better

Table 11: Κριτήρια επιλογής μοντέλου