Στατιστική Μοντελοποίηση

- **ΔΠΜΣ Επιστήμη Δεδομένων & Μηχανική Μάθηση**
- **Ορφανουδάκης Φίλιππος Σκόβελεφ ΑΜ:03400107**
- **ΣΕΙΡΑ 2**

Α)

1)

Πριν ξεκινήσουμε την αναλυτική μελέτη του μοντέλου μας θέλουμε να έχουμε μια πρώτη εντύπωση για την αποτελεσματικότητα του και για αυτό κοιτάμε την R^2 και την p-value

```
S = 2.65020   R-Sq = 86.9%   R-Sq(adj) = 80.7%

Analysis of Variance

Source          DF        SS      MS      F      P
Regression      10   978.553  97.855  13.93  0.000
Residual Error  21   147.494   7.024
Total           31  1126.047
```

Βλέπουμε αρκετά καλές τιμές , επομένως αναμένουμε το μοντέλο να περιγράφει σε μεγάλο βαθμό το πείραμα και να μην προκύψουν έντονα προβλήματα στους επιμέρους ελέγχους.

Δουλεύοντας στο MINITAB παίρνουμε τα εξής αποτελέσματα για τις **συσχετίσεις**:

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear
cyl	-0.852									
disp	-0.848	0.902								
hp	-0.776	0.832	0.791							
drat	0.681	-0.700	-0.710	-0.449						
wt	-0.868	0.782	0.888	0.659	-0.712					
qsec	0.419	-0.591	-0.434	-0.708	0.091	-0.175				
vs	0.664	-0.811	-0.710	-0.723	0.440	-0.555	0.745			
am	0.600	-0.523	-0.591	-0.243	0.713	-0.692	-0.230	0.168		
gear	0.480	-0.493	-0.556	-0.126	0.700	-0.583	-0.213	0.206	0.794	
carb	-0.551	0.527	0.395	0.750	-0.091	0.428	-0.656	-0.570	0.058	0.274

Cell Contents: Pearson correlation

Η μετρική Pearson Correlation μας υποδεικνύει τις συσχετίσεις μεταξύ των μεταβλητών και παίρνει τιμές [-1,+1]. Συνεπώς μπορούμε να πούμε πως έχουμε συσχετίσεις μεταξύ των μεταβλητών και μερικές είναι αρκετά έντονες.

Οι μεταβλητές :

- disp - cycl : 0.902
- hp - cycl : 0.832
- hp - disp : 0.791
- wt - disp : 0.888

- vs - cycl : - 0.8111

Έχουν μεγάλη συσχέτιση (Θετική-Αρνητική) .Όπως επίσης η εξαρτημένη μεταβλητή έχει μεγάλη συσχέτιση με τις ανεξάρτητες μεταβλητές cycl, disp, wt.

Στη συνέχεια για την **πολυσυγγραμμικότητα** παίρνουμε τα εξής αποτελέσματα:

Predictor	Coef	SE Coef	T	P	VIF
Constant	12.30	18.72	0.66	0.518	
cyl	-0.111	1.045	-0.11	0.916	15.4
disp	0.01334	0.01786	0.75	0.463	21.6
hp	-0.02148	0.02177	-0.99	0.335	9.8
drat	0.787	1.635	0.48	0.635	3.4
wt	-3.715	1.894	-1.96	0.063	15.2
qsec	0.8210	0.7308	1.12	0.274	7.5
vs	0.318	2.105	0.15	0.881	5.0
am	2.520	2.057	1.23	0.234	4.6
gear	0.655	1.493	0.44	0.665	5.4
carb	-0.1994	0.8288	-0.24	0.812	7.9

Από τον παραπάνω πίνακα θα μελετήσουμε την στήλη VIF.

Στην περίπτωση της πολυσυγγραμμικότητας το κριτήριο VIF μας υποδεικνύει το πόσο κάθε μεταβλητή έχει κάποια γραμμική συσχέτιση με όλες τις υπόλοιπες μεταβλητές μαζί. Όταν έχουμε τιμή >5 τότε λέμε ότι έχουμε έντονη πολυσυγγραμμικότητα. Στη συγκεκριμένη περίπτωση αυτό συμβαίνει στις μεταβλητές :

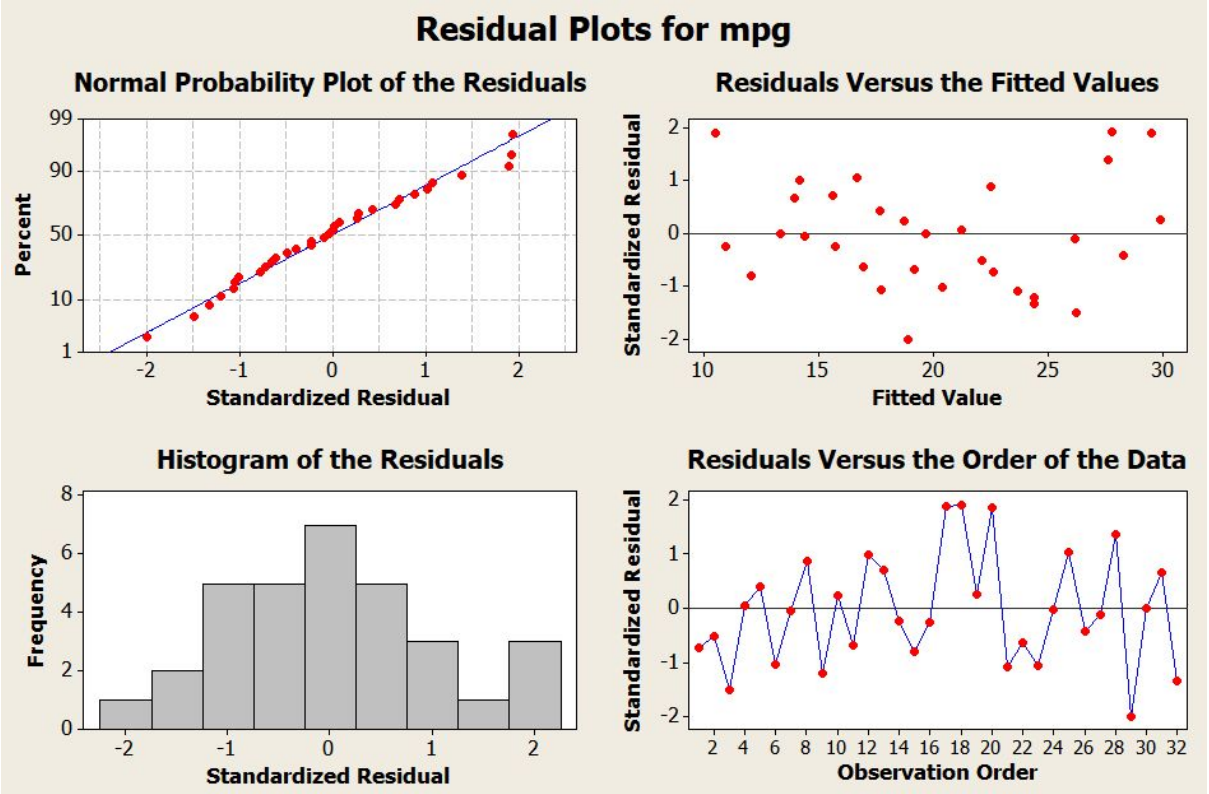
1. cycl
2. disp
3. hp
4. wt
5. qsec
6. gear
7. carb

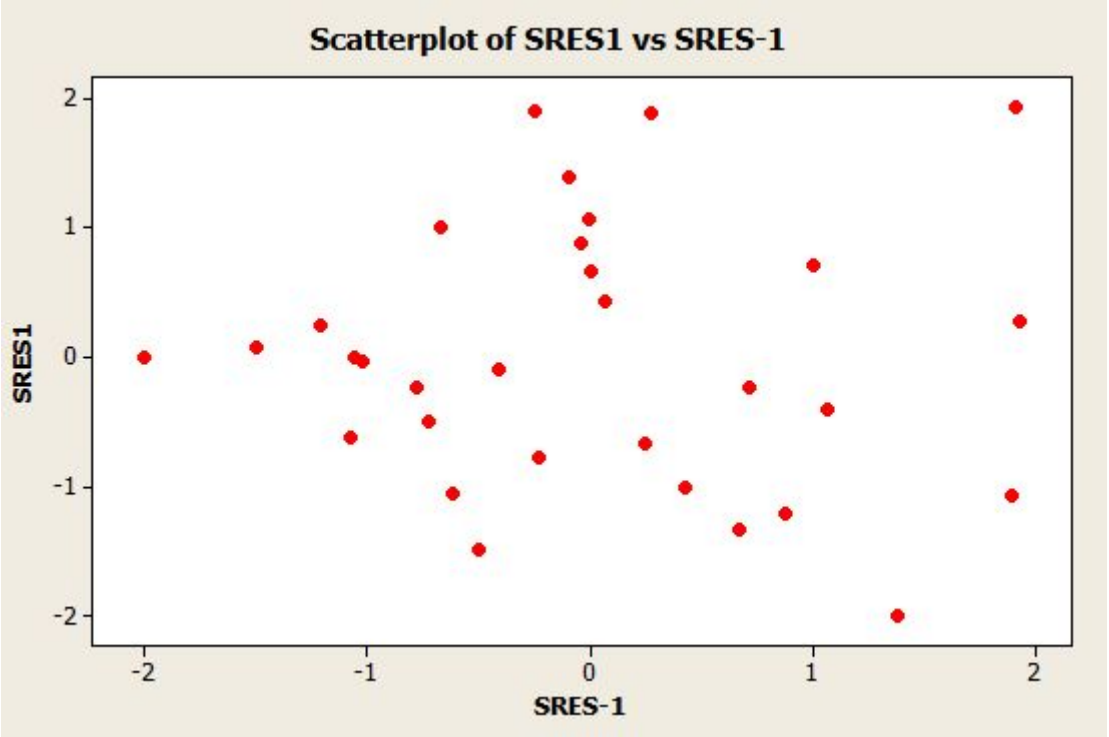
Μέχρι στιγμή με τα παραπάνω αποτελέσματα έχουμε ενδείξεις ότι οι επεξηγηματικές μεταβλητές μας έχουν έντονες συσχετίσεις, γεγονός που ίσως δυσκολέψει στο να βγάλουμε συμπεράσματα για τον ρόλο κάθε μιας από αυτές.

Στη συνέχεια πηγαίνουμε στον έλεγχο των υπολοίπων:

Δεν θα χρησιμοποιήσουμε τα Regular Residuals λόγω της ετεροσκεδαστικότητας που εμφανίζουν , για αυτό το λόγο ο έλεγχος θα γίνει με τα Standardized και τα Deleted.

Για τα Standardized Residuals παίρνουμε τα εξής διαγράμματα (αγνοούμε το Histogram) :



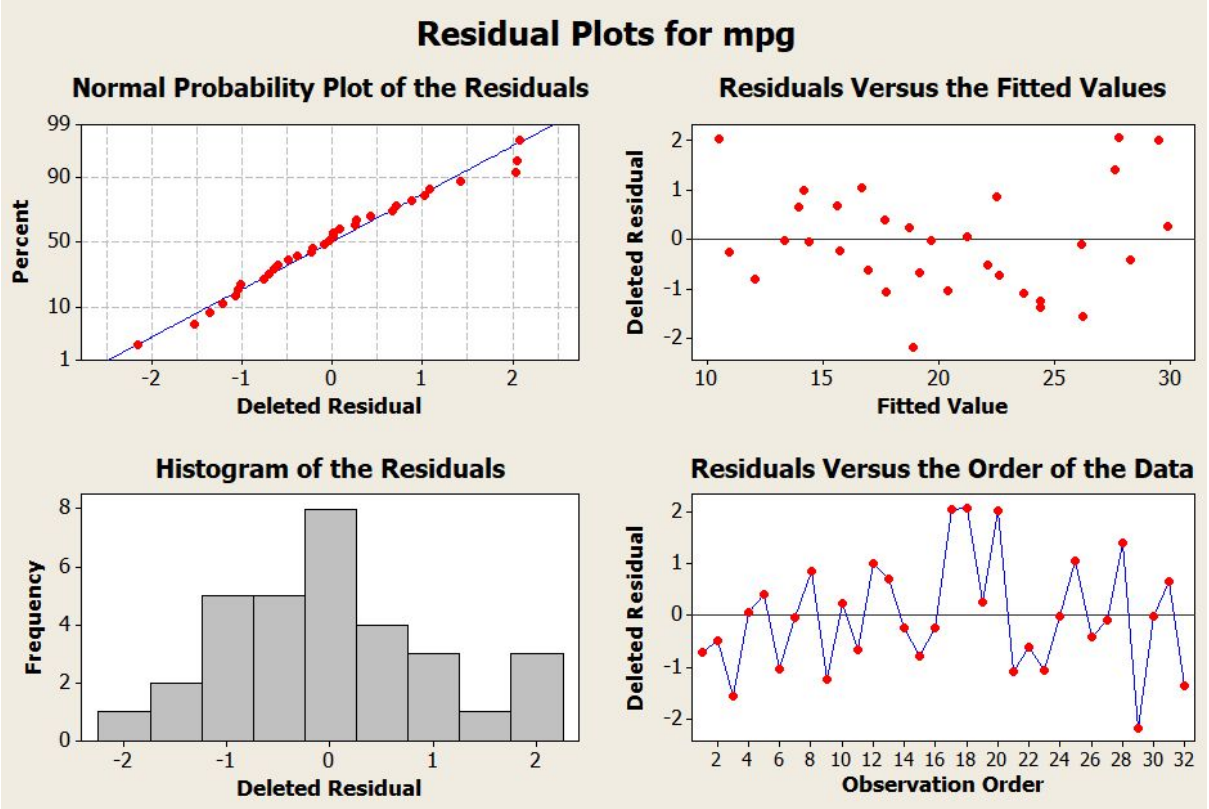


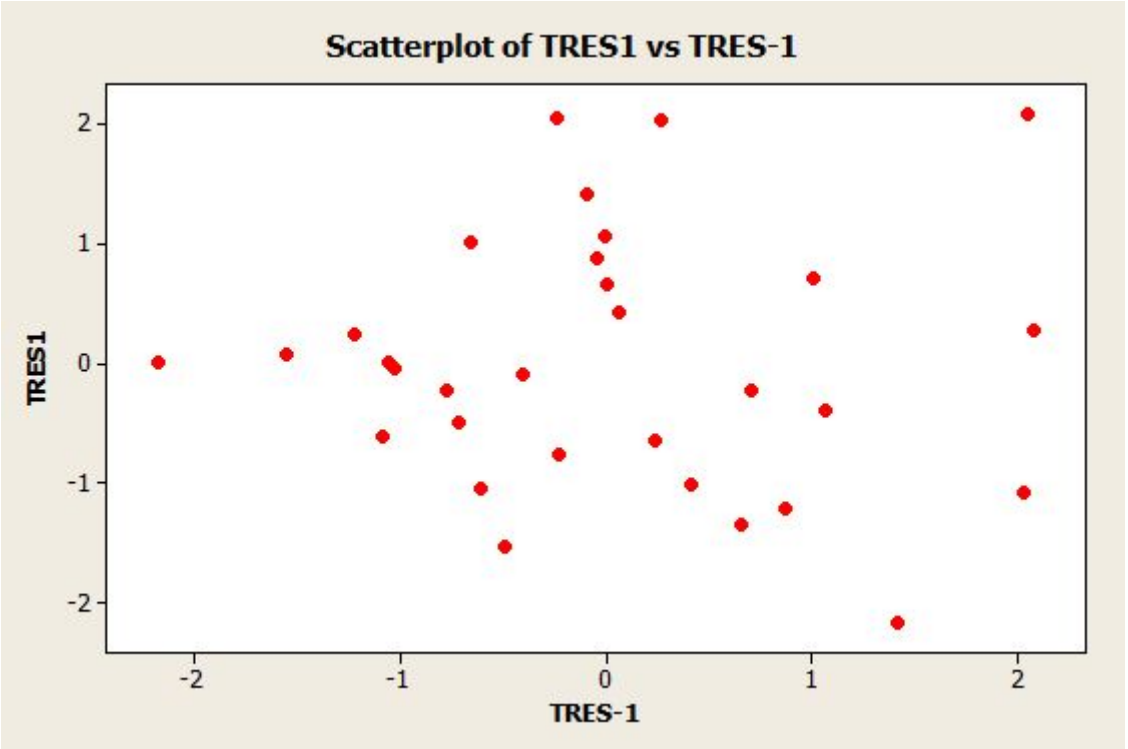
1. -0.72267
2. -0.49799
3. -1.49237
4. 0.06981
5. 0.42451
6. -1.01685
7. -0.03964
8. 0.87786
9. -1.20344
10. 0.25023
11. -0.66415
12. 1.00711
13. 0.71385
14. -0.23221
15. -0.77957
16. -0.24351
17. 1.90620
18. 1.92684
19. 0.27182
20. 1.89024
21. -1.07432
22. -0.61575
23. -1.05158

24. -0.00295
25. 1.06171
26. -0.40474
27. -0.09402
28. 1.38261
29. -1.99624 (outlier)
30. 0.00298
31. 0.66848
32. -1.32969

Θα μπορούσαμε να πούμε ότι το σημείο 29 είναι **άτυπο** καθώς εμφανίζει μεγάλη τιμή στο τυποποιημένο υπόλοιπο.

Για τα Deleted παίρνουμε τα εξής διαγράμματα (αγνοούμε το Histogram) :





1. -0.71419
2. -0.48888
3. -1.54038
4. 0.06814
5. 0.41607
6. -1.01772
7. -0.03869
8. 0.87287
9. -1.21716
10. 0.24456
11. -0.65506
12. 1.00747
13. 0.70525
14. -0.22690
15. -0.77204
16. -0.23798
17. 2.04563
18. 2.07251
19. 0.26573
20. 2.02499
21. -1.07848
22. -0.60641
23. -1.05438
24. -0.00288
25. 1.06510
26. -0.39653
27. -0.09178

28. 1.41524
29. -2.16427
30. 0.00291
31. 0.65942
32. -1.35598

Τα τελικά μας συμπεράσματα είναι τα εξής :

1. Διάγραμμα Κανονικής κατανομής :

   Και στις 2 περιπτώσεις παρατηρούμε αρκετά καλή γραμμικότητα των υπολοίπων μας , γεγονός που μας οδηγεί στο να επιβεβαιώσουμε την υπόθεση της κανονικότητας.

2. Διάγραμμα Versus Fits :

   Και στις 2 περιπτώσεις μπορούμε να πούμε ότι έχουμε τυχαιότητα στην κατανομή των υπολοίπων γύρω από το 0 και συνεπώς επιβεβαιώνεται η υπόθεση της ομοσκεδαστικότητας.

3. Διάγραμμα Versus Order :

   Στα διαγράμματα αυτά παρατηρούμε αν υπάρχει κάποια εξάρτηση των υπολοίπων με την χρονική σειρά των δεδομένων. Δεν εντοπίζουμε κάποια συσχέτιση.

4. Διάγραμμα Res vs Res-1 :

   Και στις 2 περιπτώσεις δεν εμφανίζεται κάποια εξάρτηση των υπολοίπων από από τις προηγούμενες τιμές τους.

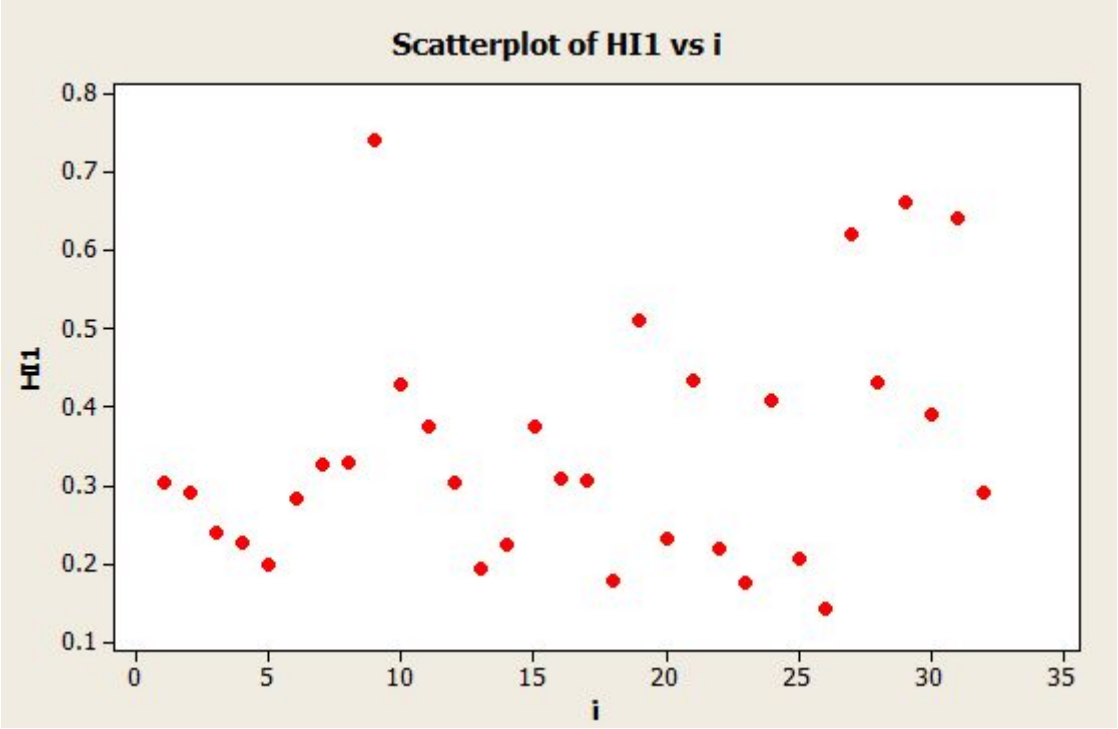
5. Τιμές υπολοίπων :

   Δεν έχουμε κάποια μεγάλη απόκλιση στις τιμές και για αυτό το λόγο το MINITAB δεν μας βγάζει κάποια ένδειξη για unusual τιμές, όπως παρατηρήσαμε η παρατήρηση 29 είναι υποψήφια για άτυπο σημείο .

Τέλος θα εξετάσουμε για σημεία επιρροής με τα κριτήρια h_{ii} (leverage), απόσταση Cook, DFFiTS και DFBETAS .

1. h_{ii} **(leverage)**

Με τη βοήθεια του MINITAB υπολογίζουμε τις τιμές h_{ii} και θεωρούμε υποψήφιο σημείο επιρροής , όποιο ξεπερνάει την τιμή $\frac{2*p}{n} = \frac{22}{32} = 0.6875$.

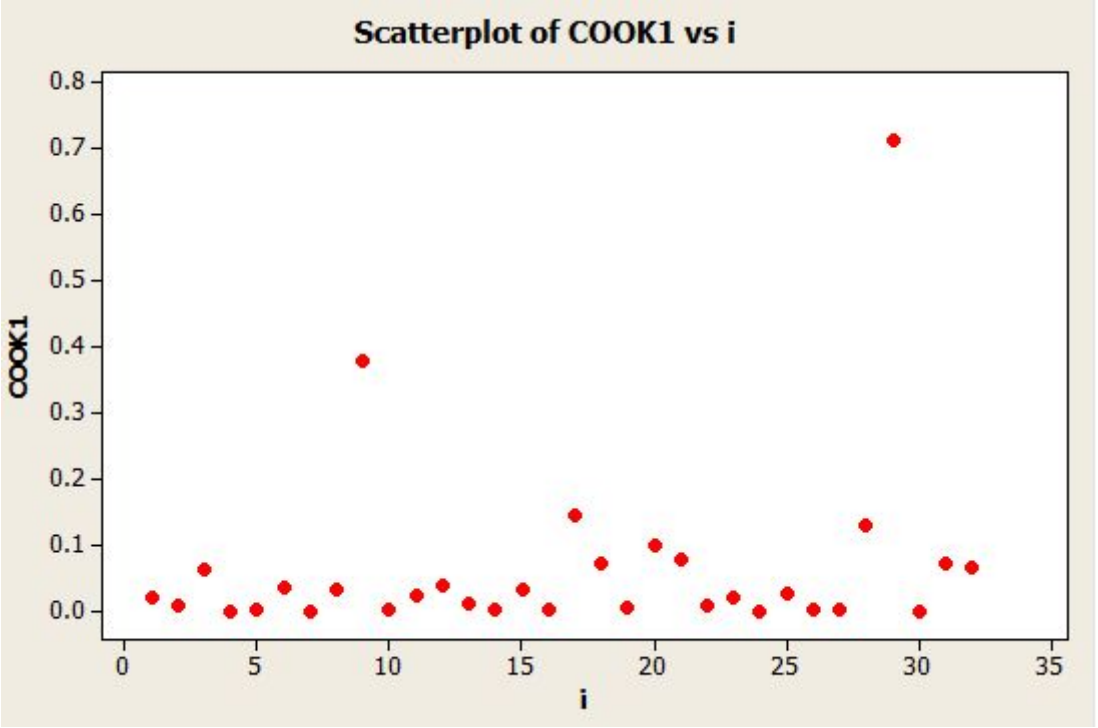


1. 0.302506
2. 0.290221
3. 0.238817
4. 0.227739
5. 0.199512
6. 0.282284
7. 0.325918
8. 0.330231
9. 0.742287 (υποψήφιο)
10. 0.429326
11. 0.374857
12. 0.303281
13. 0.192115
14. 0.223659
15. 0.374480
16. 0.309044
17. 0.306696
18. 0.178951
19. 0.511932
20. 0.232872
21. 0.433414
22. 0.218010
23. 0.174445
24. 0.408073

25. 0.205305
26. 0.142164
27. 0.623226
28. 0.431098
29. 0.663252 (οριακό)
30. 0.391019
31. 0.642757
32. 0.290508

2. Απόσταση Cook

Με την βοήθεια του MINITAB υπολογίζουμε τις τιμές και εστιάζουμε την προσοχή μας σε αυτές όπου $D_i > 1$.



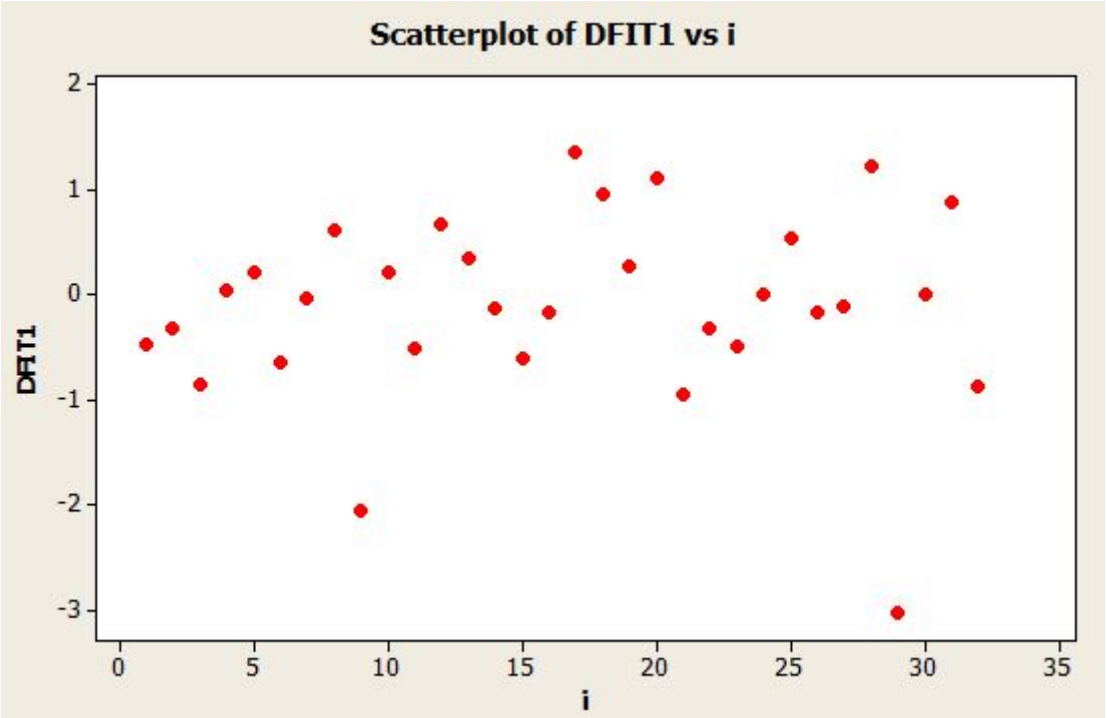
1. 0.020591
2. 0.009218
3. 0.063524
4. 0.000131
5. 0.004083
6. 0.036971
7. 0.000069
8. 0.034542
9. 0.379221
10. 0.004282
11. 0.024045
12. 0.040137

13. 0.011016
14. 0.001412
15. 0.033076
16. 0.002411
17. 0.146126
18. 0.073564
19. 0.007045
20. 0.098603
21. 0.080262
22. 0.009609
23. 0.021243
24. 0.000001
25. 0.026474
26. 0.002468
27. 0.001329
28. 0.131687
29. 0.713521
30. 0.000001
31. 0.073091
32. 0.065814

Βλέπουμε ότι δεν υπάρχει καμία τιμή με $D_i > 1$, μετά την ολοκλήρωση όλων των ελέγχων θα έχουμε πιο ολοκληρωμένη εικόνα για να βγάλουμε τα συμπεράσματα μας.

3. DFFiTS

Σε αυτόν τον έλεγχο κοιτάμε ποιά απόλυτη τιμή είναι πάνω από $2 * \sqrt{\frac{p}{n}} = 2 * 0.58630196997 = 1,173$.



1. -0.47034
2. -0.31261
3. -0.86281
4. 0.03700
5. 0.20772
6. -0.63826
7. -0.02690
8. 0.61291
9. -2.06569 (υποψήφιο)
10. 0.21213
11. -0.50725
12. 0.66470
13. 0.34391
14. -0.12179
15. -0.59736
16. -0.15915
17. 1.36057 (υποψήφιο)
18. 0.96756
19. 0.27215
20. 1.11570
21. -0.94326
22. -0.32019
23. -0.48468
24. -0.00239
25. 0.54137
26. -0.16143
27. -0.11804
28. 1.23196 (υποψήφιο)

29. -3.03737 (υποψήφιο)
30. 0.00233
31. 0.88452
32. -0.86768

4. DFBETAS

Με τη βοήθεια της R παίρνουμε τα εξής αποτελέσματα και ψάχνουμε απόλυτη τιμή η οποία να είναι μεγαλύτερη από $\frac{2}{\sqrt{n}} = 0.35355339059$

	(Intercept)	cyl	disp	hp	drat
1	-0.0802281838	0.0062625932	-0.0966338918	0.2600486555	0.0354849408
2	0.0052328159	-0.0322931507	-0.0097810980	0.1413196427	0.0030085566
3	-0.2478182269	0.1380197891	0.2994530767	-0.3054268783	0.1870032533
4	0.0055880788	0.0014548382	0.0162420317	-0.0107229679	-0.0146515287
5	-0.0148865231	0.0218880243	0.1017261117	-0.0411448924	0.0001760854
6	-0.0143219808	-0.1319932444	0.0501885266	0.0172930412	0.3927172853
7	-0.0116472383	0.0127220434	-0.0080168469	-0.0069470392	0.0017836523
8	0.1809706701	-0.2413738635	0.0610506722	-0.1747670868	-0.0869272866
9	1.2621677975	0.0241003847	-0.2615139174	-0.5280186930	-0.2705710277
10	-0.0018443256	0.0714213801	-0.0455635431	-0.0668762141	0.0612215256
11	0.1104050493	-0.2204173928	0.0774670476	0.1562832948	-0.1616551197
12	-0.0095080493	0.2234813540	-0.5673552154	0.2426556337	0.0427879697
13	-0.0565173271	0.1356455995	-0.2097605554	0.1165897350	-0.0027277286
14	0.0396300560	-0.0533580267	0.0678154540	-0.0435573241	-0.0016872618
15	0.0192842107	0.1567043100	-0.3532330924	0.2338593996	0.1005074681
16	-0.0008974285	0.0401614475	-0.0356717210	0.0246545620	0.0068456750
17	0.0429492736	-0.3060891928	-0.1948915128	0.2324904149	0.3214121655
18	-0.2107550564	0.2509539824	-0.3050736908	0.1652926417	0.0772650900
19	-0.0398080182	0.0343816888	0.1155667827	-0.1065518357	0.1544897497
20	-0.5592002881	0.3509602416	0.1023978375	0.1122960701	0.1982794347
21	-0.4234594478	0.5942808502	0.0584352076	-0.4427638538	0.0143209064
22	-0.0858086282	-0.0609476021	-0.0107072656	0.1076241384	0.1489518699
23	0.0832923658	-0.2391054064	0.0301861488	0.1024188732	-0.0330241207
24	-0.0008290439	0.0006531818	0.0004374677	-0.0009149880	-0.0011951733
25	0.0010437621	-0.0019604462	0.3260796338	-0.1949922314	-0.0298638528
26	-0.0052000786	-0.0191132919	0.0130579934	-0.0036923500	0.0017418874
27	-0.0175584836	0.0426272994	0.0125342546	-0.0085660023	-0.0174960048
28	0.5102015963	-0.3117084737	0.2604600955	-0.0493228786	-0.6604033145
29	1.5031369368	-1.3250467886	-0.0083468216	-0.5335536869	-1.2628586930
30	0.0007154665	-0.0004883266	-0.0001073003	-0.0004285907	-0.0011904630
31	-0.1181378553	-0.0259328258	0.0435358245	0.3084974640	-0.2209181204
32	-0.2099417251	0.1874010237	0.3419571679	-0.3001747963	-0.0749213726

	wt	qsec	vs	am	gear
1	0.1090686613	-0.0157343673	9.862100e-02	-0.1397359321	0.1668142181

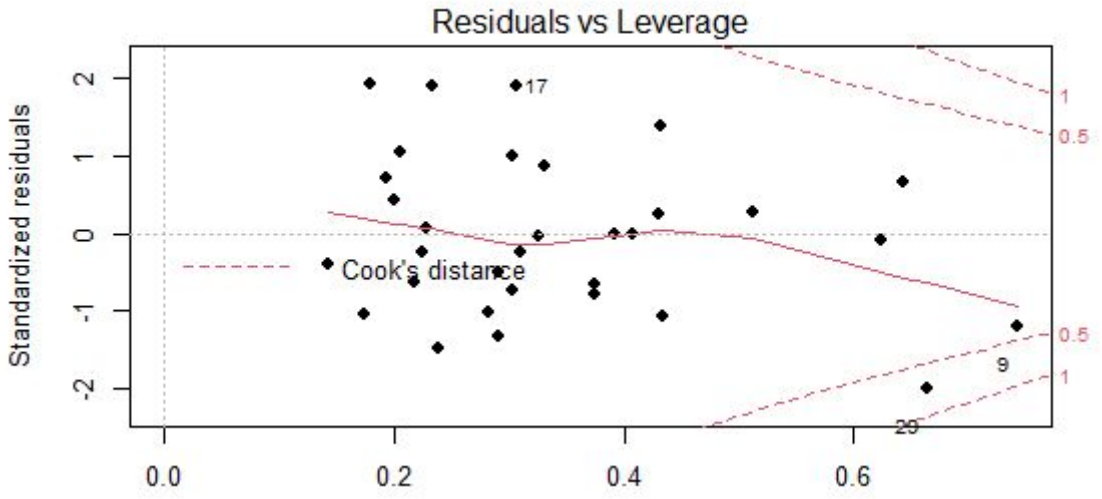
2	0.0136763524	-0.0466591388	9.068081e-02	-0.1293494728	0.0867605925	
3	-0.3859113573	0.2733576071	-1.906396e-01	-0.3712309391	0.1027036364	
4	-0.0130389264	-0.0004311089	2.099255e-02	0.0026894351	-0.0016927101	
5	-0.1198671265	0.0426644008	-2.877391e-02	-0.0409196316	0.0131446799	
6	-0.0310496244	-0.0548181304	-2.424401e-01	-0.1645406990	-0.0760120651	
7	0.0136687266	0.0023130583	1.641978e-03	0.0076646570	0.0141758222	
8	0.0822111534	-0.1836150430	-1.563950e-02	-0.3713108072	0.2198323189	
9	0.5690921612	-1.5950527663	9.378371e-01	0.4025619443	-0.5608331367	
10	0.0595024864	-0.0800701708	1.152810e-01	-0.0833337299	0.0474909343	
11	-0.0827656630	0.0678245575	-2.449552e-01	0.1757084416	-0.1326324102	
12	0.4346811830	-0.1292769487	-1.201149e-01	-0.0083615610	0.0365739106	
13	0.0699370271	0.0559280879	-1.100568e-01	-0.0115710423	-0.0073029170	
14	-0.0177541345	-0.0420954706	4.571447e-02	-0.0048521942	-0.0015719420	
15	0.0752603707	-0.1031804149	-2.113319e-02	-0.1710874994	0.0509740475	
16	-0.0454868094	0.0040360128	-4.740677e-03	-0.0464016613	0.0107269720	
17	0.7752259274	-0.2843233814	1.930443e-03	0.2330302771	-0.1663461545	
18	0.3088224890	0.1156942135	1.154706e-01	0.5721444850	-0.0681930279	
19	-0.1108498547	0.0325095023	5.833917e-02	0.0308896834	-0.1091580202	
20	-0.3099084456	0.6683675458	-4.717503e-02	0.5630018916	-0.1600944519	
21	0.1709862018	0.0802762574	2.299156e-01	0.3701412076	0.4735304181	
22	0.0355037155	0.0620033316	4.932487e-03	0.0371467677	-0.0644015845	
23	0.0691085979	-0.0473316838	8.385001e-02	0.0784474972	-0.0950109401	
24	-0.0002582488	0.0010231151	-3.195125e-05	0.0008576534	0.0011020800	
25	-0.2235259345	0.0341735385	-1.441958e-02	-0.0530393093	0.0758721542	
26	-0.0101633974	0.0042732041	-3.940118e-02	-0.0810327574	0.0329877789	
27	-0.0249762189	0.0241920466	7.447629e-02	0.0483032571	-0.0462534029	
28	-0.2557366267	-0.3998012490	3.447797e-01	-0.2498187278	0.4233467519	
29	-0.2912018391	-0.1955388198	-4.565789e-01	-0.0935737950	-1.6594052827	
30	-0.0001552195	-0.0002400735	-4.720437e-04	-0.0001419433	0.0002766015	
31	-0.2103408144	0.2458690372	7.996855e-02	0.1832247195	-0.0287468353	
32	-0.5194358361	0.3142646375	-2.158247e-01	-0.3717999505	0.2609872777	

	carb
1	-0.2585584508
2	-0.1260086701
3	0.4066847273
4	0.0080458630
5	0.0210984708
6	0.0573968931
7	-0.0098181791
8	0.0376607501
9	-0.1706675580
10	0.0105186294
11	-0.0637457212
12	-0.3873200989
13	-0.1076173848
14	0.0340863144
15	-0.2439605438
16	-0.0128565290

17 -0.3207287566
18 -0.3492065828
19 0.1213327686
20 0.0288609544
21 -0.0703297039
22 0.0264518337
23 0.0902594054
24 0.0002596754
25 0.0443070575
26 0.0264918850
27 0.0436178419
28 -0.0326844062
29 1.3946783541
30 0.0008559461
31 0.2623689542
32 0.3061125751

Φυσικά όλα τα παραπάνω κριτήρια ελέγχουν τα σημεία επιρροής με διαφορετικό τρόπο και για αυτό έχουμε διαφορετικά αποτελέσματα. Η απόσταση Cook η οποία κρίνει τα σημεία επιρροής όχι μόνο από τις τιμές των ανεξάρτητων μεταβλητών αλλά και από την τιμή της εξαρτημένης, δεν επιστρέφει κάποιο αποτέλεσμα. Αν συνδυάσουμε όμως τις υπόλοιπες ενδείξεις μπορούμε να πούμε ότι η παρατήρηση 9 και 29 είναι υποψήφια σημεία επιρροής.

Αυτό φαίνεται και από το διάγραμμα στην R όπου οι 2 αυτές οι παρατηρήσεις βρίσκονται πιο κοντά στη Cook distance παρόλο που δεν βρίσκονται εκτός από τις διακεκομμένες .



2)

Έχοντας πραγματοποιήσει μια πρώτη ανάλυση του μοντέλου μας και έχοντας αποκτήσει μια εικόνα για την σημαντικότητα του και τις μεταβλητές του, θα βρούμε το βέλτιστο μοντέλο

σύμφωνα με ορισμένες μετρικές. Αρχικά θα κοιτάξουμε μερικές μετρικές για το μοντέλο που έχουμε τώρα με τις 10 μεταβλητές .

Predictor	Coef	SE Coef	T	P
Constant	12.30	18.72	0.66	0.518
cyl	-0.111	1.045	-0.11	0.916
disp	0.01334	0.01786	0.75	0.463
hp	-0.02148	0.02177	-0.99	0.335
drat	0.787	1.635	0.48	0.635
wt	-3.715	1.894	-1.96	0.063
qsec	0.8210	0.7308	1.12	0.274
vs	0.318	2.105	0.15	0.881
am	2.520	2.057	1.23	0.234
gear	0.655	1.493	0.44	0.665
carb	-0.1994	0.8288	-0.24	0.812

S = 2.65020 R-Sq = 86.9% R-Sq(adj) = 80.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	10	978.553	97.855	13.93	0.000
Residual Error	21	147.494	7.024		
Total	31	1126.047			

- Αρχικά βλέπουμε την τιμή του R-Sq = 86,9% μια αρκετά καλή τιμή, αλλά δεν μπορούμε να στηριχθούμε πλήρως σε αυτό καθώς το πλήθος των παρατηρήσεων δεν είναι πολύ μεγαλύτερο από το πλήθος των μεταβλητών , περίπτωση που μπορεί η υψηλή τιμή του R-Sq να μην σημαίνει και καλό μοντέλο.

- Στη συνέχεια κοιτάμε τις T-value κάθε μεταβλητής. Μια υψηλή τιμή της T-value μας υποδεικνύει ότι με την προϋπόθεση ότι όλες οι υπόλοιπες εξαρτημένες μεταβλητές συμπεριλαμβάνονται στο μοντέλο , η συγκεκριμένη μεταβλητή είναι σημαντική για το μοντέλο. Λόγω αυτής της προυπόθεσης δεν μπορούμε να βγάλουμε τόσο χρήσιμα αποτελέσματα, παρόλα αυτά βλέπουμε αρκετά χαμηλή τιμή για ορισμένες μεταβλητές , όπως την cycl και την vs.

- Ο έλεγχος F με τις αντίστοιχες τιμές P μας δείχνει αν μπορούμε να απορρίψουμε τις μηδενικές υποθέσεις για κάθε μεταβλητή (απορρίπτεται σε περίπτωση χαμηλής τιμής). Βλέπουμε ότι μόνο για την wt μπορούμε να την απορρίψουμε, ενώ όλες οι άλλες μεταβλητές θα πρέπει να μελετηθεί αν είναι απαραίτητες στο μοντέλο μας.

- Για πληρότητα κατασκευάζουμε και τα διαστήματα εμπιστοσύνης 95% με την βοήθεια της R :

```
> confint(model)
                    2.5 %      97.5 %
(Intercept) -26.62259745 51.22934576
cyl          -2.28468553  2.06180457
disp         -0.02380146  0.05047194
hp           -0.06675236  0.02378812
drat         -2.61383350  4.18805545
wt           -7.65495413  0.22434628
qsec         -0.69883421  2.34091571
vs           -4.05880242  4.69432805
am           -1.75681208  6.79726585
gear         -2.44999107  3.76081711
carb         -1.92290442  1.52406591
```

Βλέπουμε ότι παντού περιέχεται το 0.

Στη συνέχεια θα πραγματοποιήσουμε τεχνικές με βήματα. Η R μας δίνει τα εξής τελικά αποτελέσματα (θα παρουσιάσουμε μόνο την τελική επανάληψη):

- **Forward**

```
Step:  AIC=62.66
cars$mpg ~ cars$wt + cars$cyl + cars$hp

            Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>                  176.62 62.665
+ cars$am    1    6.6228 170.00 63.442  1.0519 0.3142
+ cars$disp  1    6.1762 170.44 63.526  0.9784 0.3314
+ cars$carb  1    2.5187 174.10 64.205  0.3906 0.5372
+ cars$drat  1    2.2453 174.38 64.255  0.3477 0.5603
+ cars$qsec  1    1.4010 175.22 64.410  0.2159 0.6459
+ cars$gear  1    0.8558 175.76 64.509  0.1315 0.7197
+ cars$vs    1    0.0599 176.56 64.654  0.0092 0.9245
```

- **Backward**

```
Step:  AIC=61.31
mpg ~ wt + qsec + am

       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              169.29 61.307
- am    1    26.178 195.46 63.908  4.3298 0.0467155 *
- qsec  1   109.034 278.32 75.217 18.0343 0.0002162 ***
- wt    1   183.347 352.63 82.790 30.3258 6.953e-06 ***
```

- **Both**

```
Step:  AIC=62.66
cars$mpg ~ cars$wt + cars$cyl + cars$hp

            Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                    176.62 62.665
- cars$hp    1    14.551 191.17 63.198  2.3069 0.1400152
+ cars$am    1     6.623 170.00 63.442  1.0519 0.3141799
+ cars$disp  1     6.176 170.44 63.526  0.9784 0.3313856
- cars$cyl   1    18.427 195.05 63.840  2.9213 0.0984801 .
+ cars$carb  1     2.519 174.10 64.205  0.3906 0.5372269
+ cars$drat  1     2.245 174.38 64.255  0.3477 0.5603447
+ cars$qsec  1     1.401 175.22 64.410  0.2159 0.6459140
+ cars$gear  1     0.856 175.76 64.509  0.1315 0.7197354
+ cars$vs    1     0.060 176.56 64.654  0.0092 0.9244706
- cars$wt    1   115.354 291.98 76.750 18.2873 0.0001995 ***
```

Με την τεχνική both και forward παίρνω το ίδιο μοντέλο :

mpg ~ wt + cyl + hp (1)

ενώ με την τεχνική backward παίρνω το μοντέλο :

mpg ~ wt + qsec + am (2)

Για το μοντέλο (1) οι μετρικές είναι οι εξής :

rsquare	adjr	predrsqr	cp	aic
0.8431500	0.8263446	0.7956775	4.000000	62.66

Για το μοντέλο (2) οι μετρικές είναι οι εξής:

rsquare	adjr	predrsqr	cp	aic
0.8496636	0.8335561	0.79458810	4.000000	61.36

!! Ομοιώς εκτελούμε αυτές τις 3 μεθόδους στο MINITAB για πληρότητα καθώς το MINITAB πραγματοποιεί έλεγχο F σε κάθε επανάληψη ενώ η R έχει σαν κριτήριο την αυξομείωση της μετρικής AIC.

- **Forward**

mpg ~ wt + cyl + hp , το ίδιο με την R

- **Backward**

mpg ~ wt + qsec + am, το ίδιο με την R

- **Both**

mpg ~ wt + cyl + hp , το ίδιο με την R

Επομένως καλούμαστε να επιλέξουμε ποιό απο 2 μοντέλα θα επιλέξουμε. Το μοντέλο που θα επιλέξουμε και θα κάνουμε τους τελικούς ελεγχους στο επόμενο βήμα είναι το (2) λόγω του μικρότερου aic και του μεγαλύτερου adjr , παρόλο που έχει μια μικρότερη τιμή στο predrsqr.

3)

Στο βήμα αυτό θα πραγματοποιήσουμε ελέγχους για να εξετάσουμε την καταλληλότητα του μοντέλου μας αλλά και να σχολιάσουμε αν οι αμφιβολίες που είχαμε στο βήμα 1) για το μοντέλο 10 μεταβλητών έχουν εξαλειφθεί .

```
The regression equation is
mpg = 9.62 - 3.92 wt + 1.23 qsec + 2.94 am
```

Predictor	Coef	SE Coef	T	P
Constant	9.618	6.960	1.38	0.178
wt	-3.9165	0.7112	-5.51	0.000
qsec	1.2259	0.2887	4.25	0.000
am	2.936	1.411	2.08	0.047

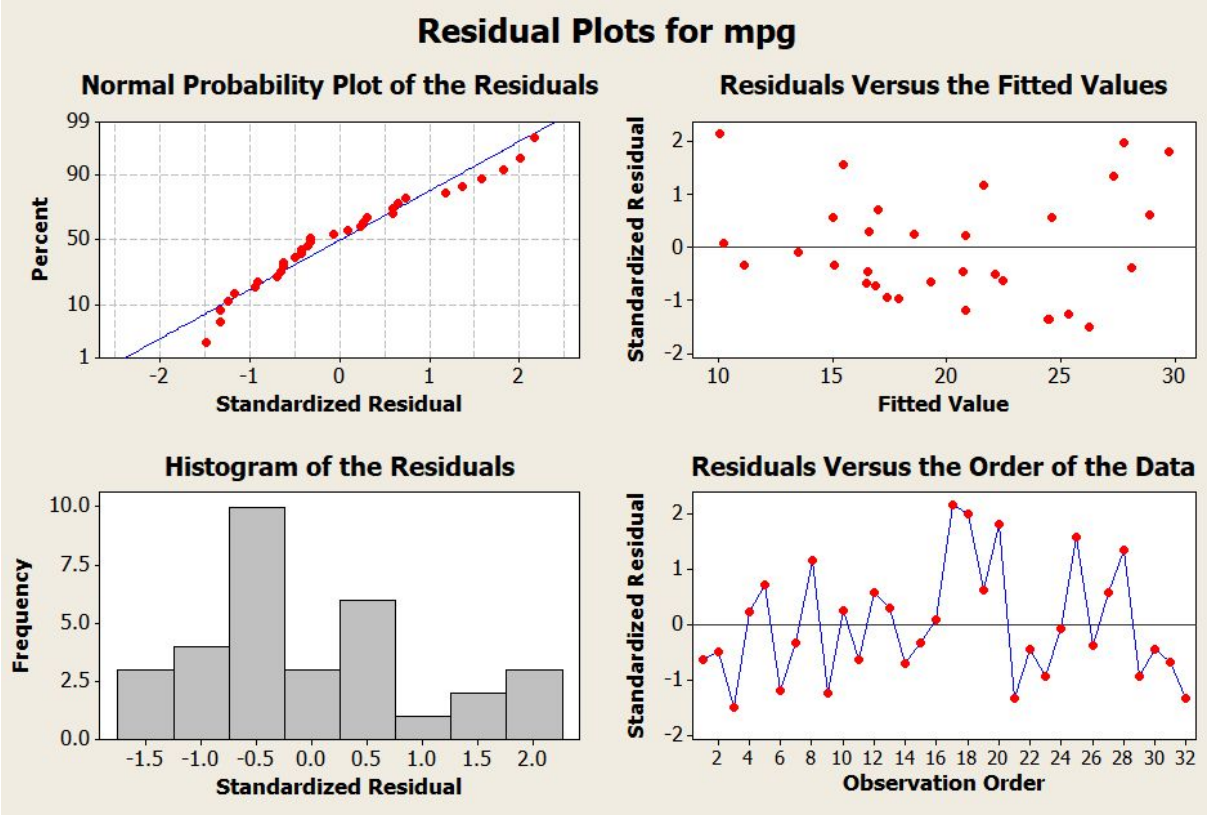
Τα p-values μας δείχνουν την σημαντικότητα των μεταβλητών, με λιγότερο σημαντική την am για την οποία δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση και θα εξετάσουμε αν είναι απαραίτητη ή όχι στη συνέχεια.

Predictor	Coef	SE Coef	T	P	VIF
Constant	9.618	6.960	1.38	0.178	
wt	-3.9165	0.7112	-5.51	0.000	2.5
qsec	1.2259	0.2887	4.25	0.000	1.4
am	2.936	1.411	2.08	0.047	2.5

Παρατηρούμε από το VIF ότι δεν έχουμε πολυσυγγραμμικότητα.

Θα παραθέσουμε πάλι τα διαγράμματα υπολοίπων για να εξετάσουμε κανονικότητα αλλά και πιθανά άτυπα σημεία και στη συνέχεια θα εξετάσουμε για σημεία επιρροής.

Για τα Standardized Residuals παίρνουμε τα εξής διαγράμματα (αγνοούμε το Histogram) :



1. -0.62542
2. -0.49419
3. -1.48858
4. 0.22975
5. 0.72174
6. -1.17901
7. -0.33191
8. 1.17731
9. -1.23866
10. 0.26070
11. -0.63558
12. 0.58422
13. 0.29971
14. -0.70185
15. -0.32268
16. 0.08537
17. 2.15973
18. 2.00067

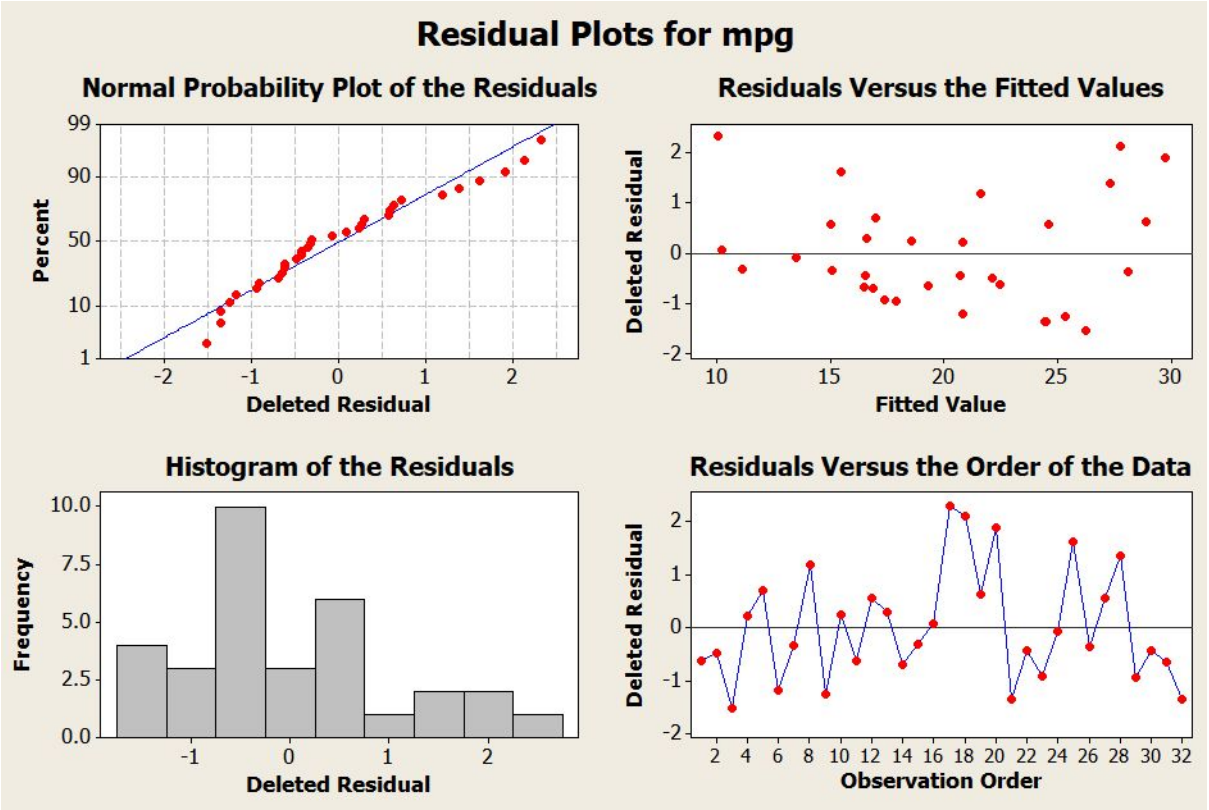
19. 0.63600
20. 1.82147
21. -1.33149
22. -0.43212
23. -0.92224
24. -0.07480
25. 1.57550
26. -0.36299
27. 0.57940
28. 1.35596
29. -0.94227
30. -0.43467
31. -0.66451
32. -1.33300

Οι παρατηρήσεις 17 και 18 είναι καλές υποψήφιες για άτυπα σημεία , καθώς έχουν μεγάλο τυποποιημένο υπόλοιπο. Αυτό επιβεβαιώνεται και από την ένδειξη του Minitab

Unusual Observations

Obs	wt	mpg	Fit	SE Fit	Residual	St Resid
17	5.35	14.700	10.039	1.178	4.661	2.16R
18	2.20	32.400	27.805	0.878	4.595	2.00R

Για τα Deleted παίρνουμε τα εξής διαγράμματα (αγνοούμε το Histogram) :



1. -0.61848
2. -0.48741
3. -1.52327
4. 0.22582
5. 0.71542
6. -1.18762
7. -0.32657
8. 1.18581
9. -1.25111
10. 0.25631
11. -0.62868
12. 0.57723
13. 0.29478
14. -0.69534
15. -0.31745
16. 0.08384
17. 2.32312
18. 2.12215
19. 0.62910
20. 1.90508
21. -1.35097
22. -0.42575
23. -0.91970
24. -0.07346
25. 1.62061

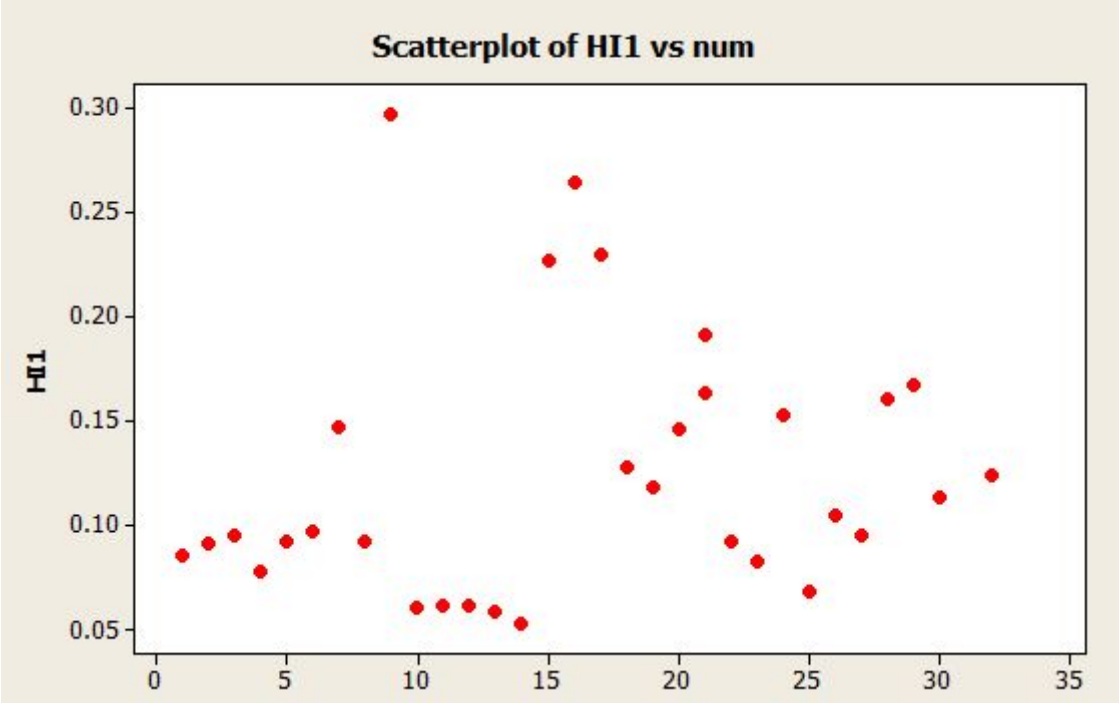
26. -0.35729
27. 0.57240
28. 1.37752
29. -0.94032
30. -0.42829
31. -0.65774
32. -1.35261

Από τα υπόλοιπα παρατηρούμε ότι η υπόθεση της κανονικότητας και η υπόθεση της ομοσκεδαστικότητας δεν παραβιάζονται

Συνεχίζουμε με το leverage, την απόσταση Cook και τα DFFiTS.

1. h_{ii} **(leverage)**

Με τη βοήθεια του MINITAB υπολογίζουμε τις τιμές h_{ii} και θεωρούμε υποψήφιο σημείο επιρροής , όποιο ξεπερνάει την τιμή $\frac{2*p}{n}$ = $\frac{8}{32}$ = 0.25.

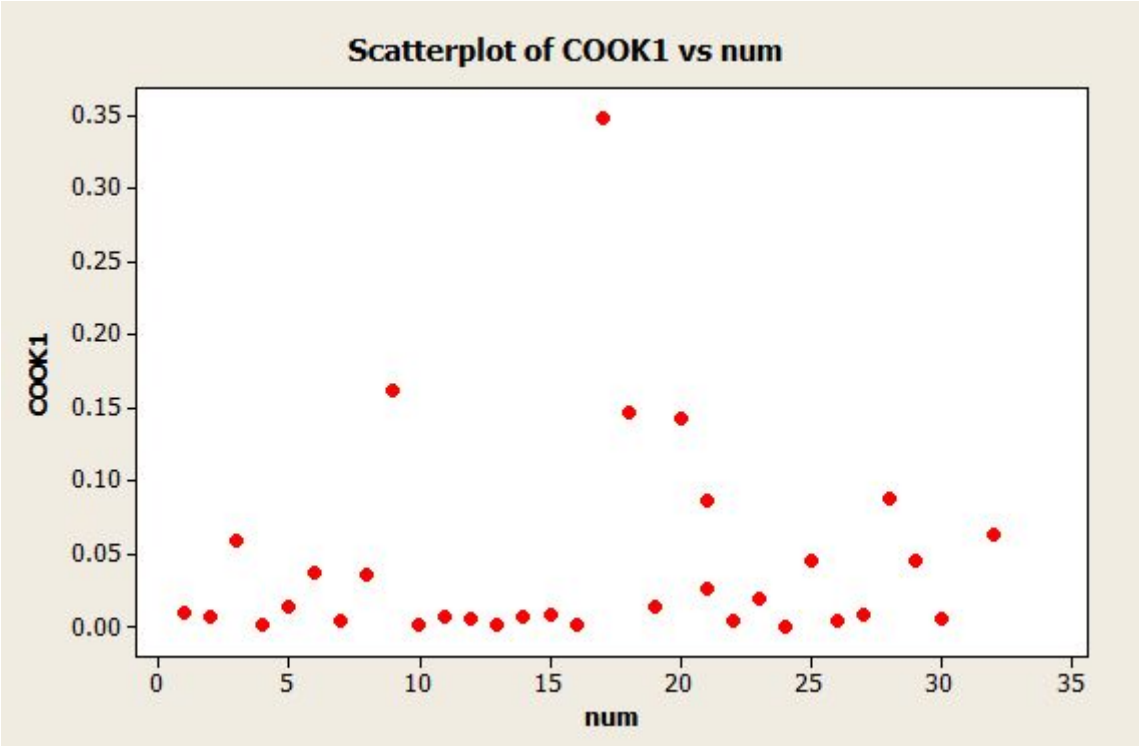


1. 0.085666
2. 0.091433

3. 0.095478
4. 0.077587
5. 0.092684
6. 0.097475
7. 0.146666
8. 0.092195
9. 0.297042
10. 0.060628
11. 0.061149
12. 0.061054
13. 0.058178
14. 0.053039
15. 0.227007
16. 0.264215
17. 0.229634
18. 0.127631
19. 0.118656
20. 0.146349
21. 0.163933
22. 0.092137
23. 0.082233
24. 0.152682
25. 0.068029
26. 0.104891
27. 0.094848
28. 0.160645
29. 0.167748
30. 0.113822
31. 0.190981
32. 0.124285

2. Απόσταση Cook

Με την βοήθεια του MINITAB υπολογίζουμε τις τιμές και εστιάζουμε την προσοχή μας σε αυτές όπου $D_i > 1$.

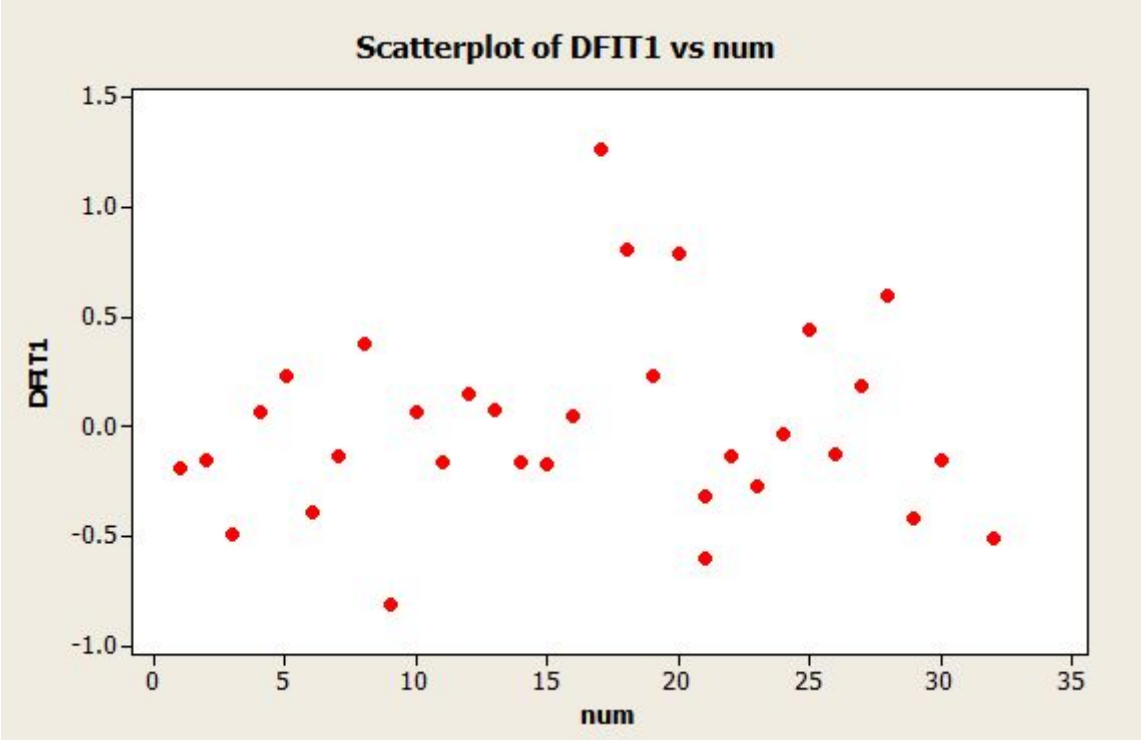


1. 0.009162
2. 0.006144
3. 0.058474
4. 0.001110
5. 0.013303
6. 0.037533
7. 0.004734
8. 0.035191
9. 0.162083
10. 0.001097
11. 0.006578
12. 0.005548
13. 0.001387
14. 0.006897
15. 0.007644
16. 0.000654
17. 0.347597
18. 0.146402
19. 0.013615
20. 0.142198
21. 0.086904
22. 0.004738
23. 0.019052
24. 0.000252
25. 0.045297
26. 0.003860
27. 0.008794

28. 0.087975
29. 0.044739
30. 0.006067
31. 0.026060
32. 0.063046

3. DFFiTS

Σε αυτόν τον έλεγχο κοιτάμε ποιά απόλυτη τιμή είναι πάνω από $2 * \sqrt{\frac{p}{n}} = 0.70710678118$

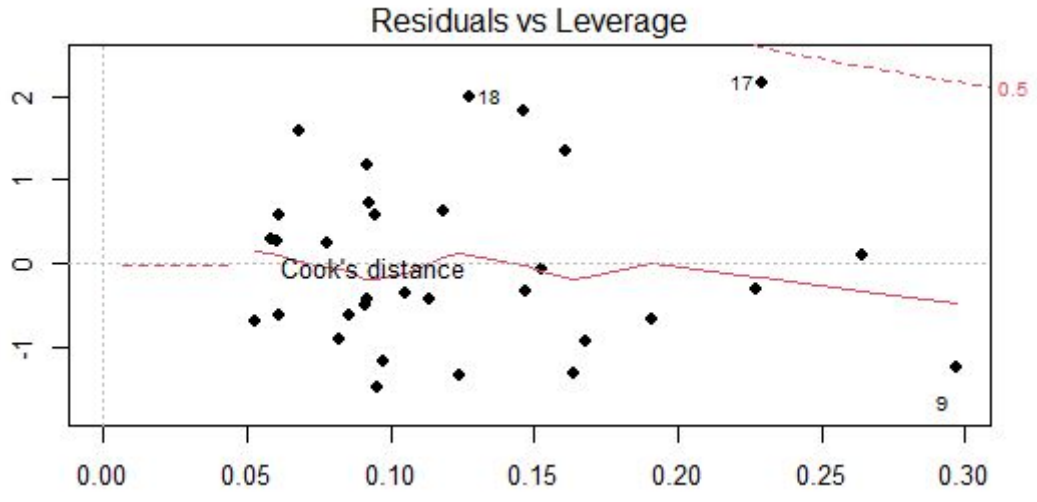


1. -0.18931
2. -0.15462
3. -0.49490
4. 0.06549
5. 0.22866
6. -0.39030
7. -0.13539
8. 0.37790
9. -0.81328
10. 0.06512
11. -0.16044
12. 0.14719
13. 0.07327
14. -0.16456
15. -0.17203
16. 0.05024

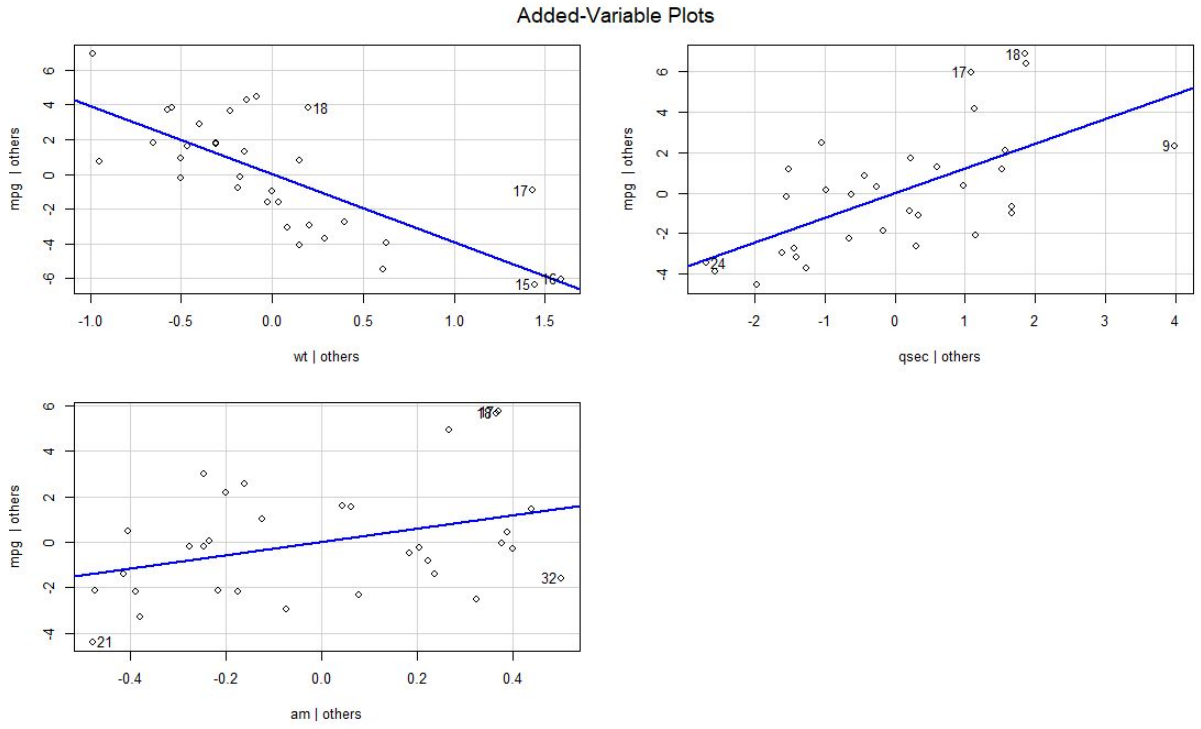
17. 1.26836
18. 0.81172
19. 0.23083
20. 0.78880
21. -0.59821
22. -0.13563
23. -0.27530
24. -0.03118
25. 0.43785
26. -0.12231
27. 0.18529
28. 0.60264
29. -0.42216
30. -0.15349
31. -0.31957
32. -0.50956

Συγκεντρωτικά παρατηρούμε μια μεγάλη βελτίωση στις αντίστοιχες τιμές των παρατηρήσεων, χωρίς όμως να μπορούμε να πούμε με σιγουριά ότι δεν υπάρχουν σημεία επιρροής. Η παρατήρηση 29 που παρουσίασε πρόβλημα στο προηγούμενο μοντέλο δεν είναι πλεον υποψήφια για σημείο επιρροής.

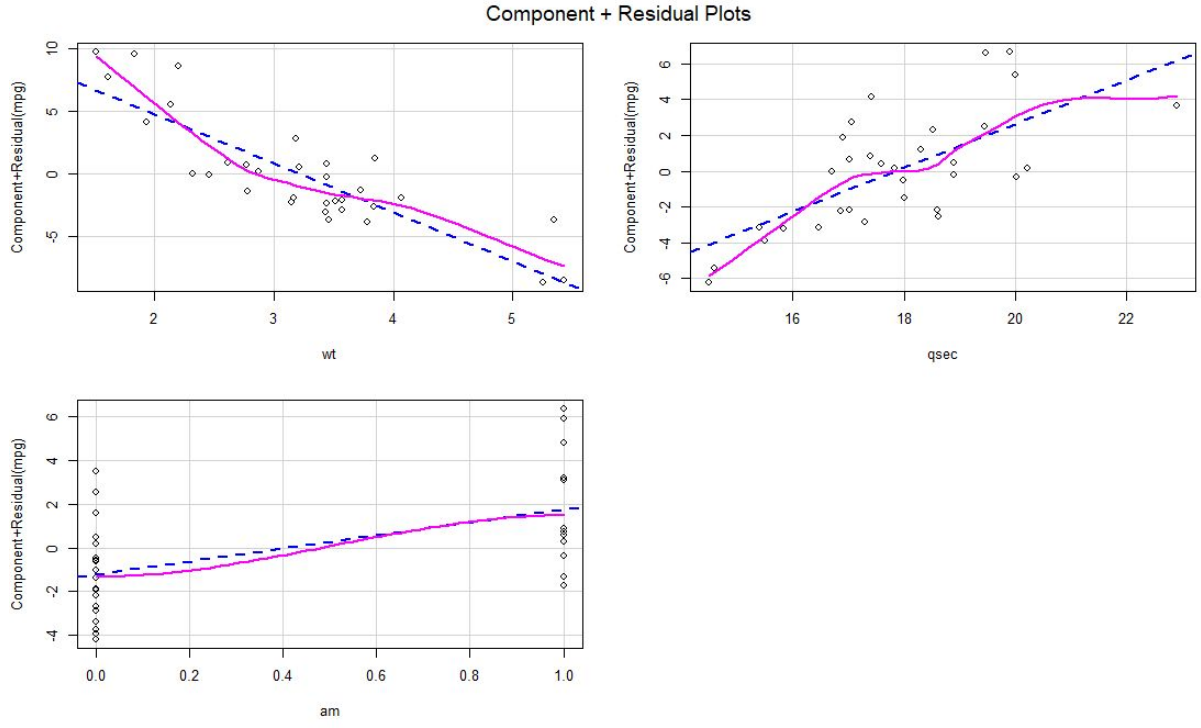
Για πληρότητα παραθέτουμε και το αντίστοιχο διάγραμμα της R με την απόσταση Cook που βοηθάει στην οπτικοποίηση και μας βοηθάει στο να πούμε ότι πιθανότατα δεν υπάρχουν σημεία επιρροής.



Γραφικές παραστάσεις πρόσθετων μεταβλητών



Γραφικές παραστάσεις των μερικών υπολοίπων (partial residual plots)



Από τα παραπάνω διαγράμματα μπορούμε να πούμε ότι μας προβληματίζει αρκετά η μη γραμμικότητα και επομενως η άμεση συσχέτιση μεταξύ των μεταβλητών και της εξαρτημένης μεταβλητής , συνεπώς θα εξετάσουμε περισσότερο αν κάποιο άλλο μοντέλο ανταποκρίνεται καλύτερα.

Με την εντολή best subsets του Minitab, παίρνουμε τα εξής αποτελέσματα:

Response is mpg

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
1	75.3	74.5	11.6	3.0459					X					
1	72.6	71.7	15.9	3.2059	X									
2	83.0	81.9	1.2	2.5675	X				X					
2	82.7	81.5	1.8	2.5934			X		X					
3	85.0	83.4	0.1	2.4588					X	X		X		
3	84.3	82.6	1.1	2.5115	X		X		X					
4	85.8	83.7	0.8	2.4348			X		X	X		X		
4	85.7	83.6	1.0	2.4438					X	X		X		X
5	86.4	83.8	1.8	2.4293		X	X		X	X		X		
5	86.1	83.4	2.3	2.4554				X	X	X		X		X
6	86.7	83.5	3.4	2.4503		X	X	X	X	X		X		
6	86.6	83.4	3.4	2.4532		X	X		X	X		X	X	
7	86.8	83.0	5.1	2.4877		X	X	X	X	X		X	X	
7	86.8	82.9	5.2	2.4924	X	X	X	X	X	X		X		
8	86.9	82.3	7.0	2.5353		X	X	X	X	X		X	X	X
8	86.8	82.3	7.1	2.5375		X	X	X	X	X	X	X	X	

Τα αποτελέσματα ευνοούν το μοντέλο που έχουμε επιλέξει.

Στη συνεχεια θα ελέγξουμε τα εμφωλευμένα μοντέλα :

Analysis of Variance Table

Model 1: mpg ~ wt + qsec + am
Model 2: mpg ~ wt + qsec

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	28	169.29				
2	29	195.46	-1	-26.178	4.3298	0.04672 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Το p-value μας οδηγεί στο ΄να επιλέξουμε το πιο σύνθετο μοντέλο, δηλαδή αυτό που είχαμε επιλέξει.

Analysis of Variance Table

Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + am + gear + carb

Model 2: mpg ~ wt + qsec + am
  Res.Df   RSS Df Sum of Sq   F Pr(>F)
1   22 147.66
2   28 169.29 -6   -21.631 0.5372 0.7742

Το p-value μας οδηγεί στο να επιλέξουμε το πιο απλό μοντέλο, δηλαδή αυτό που έχουμε επιλέξει.

Συνεπώς από όλα τα διαθέσιμα μοντέλα καταλήγουμε ότι έχουμε το βέλτιστο και από τα διαγράμματα δεν παρατηρούμε κάποια μη γραμμική σχέση που να απαιτεί κάποιον μετασχηματισμό.

Το Δ.Ε. 95% των συντελεστών του μοντέλου είναι το εξής και βλέπουμε ότι δεν υπάρχει το 0.

```
                2.5 %    97.5 %
(Intercept) -4.63829946 23.873860
wt          -5.37333423 -2.459673
qsec         0.63457320  1.817199
am           0.04573031  5.825944
```

Στη συνέχεια παίρνουμε τα διαστήματα εμπιστοσύνης της πρόβλεψης για τιμές [2.5,18,1]:

```
predict(mod,new = pred_grid , interval = "confidence")
     fit      lwr      upr
24.82831 23.35925 26.29736
```

```
predict(mod,new = pred_grid , interval = "predict")
     fit      lwr      upr
24.82831 19.58172 30.07489
```

Η τελική ερμηνεία θα προέλθει από την παρακάτω εξίσωση παλινδρόμησης

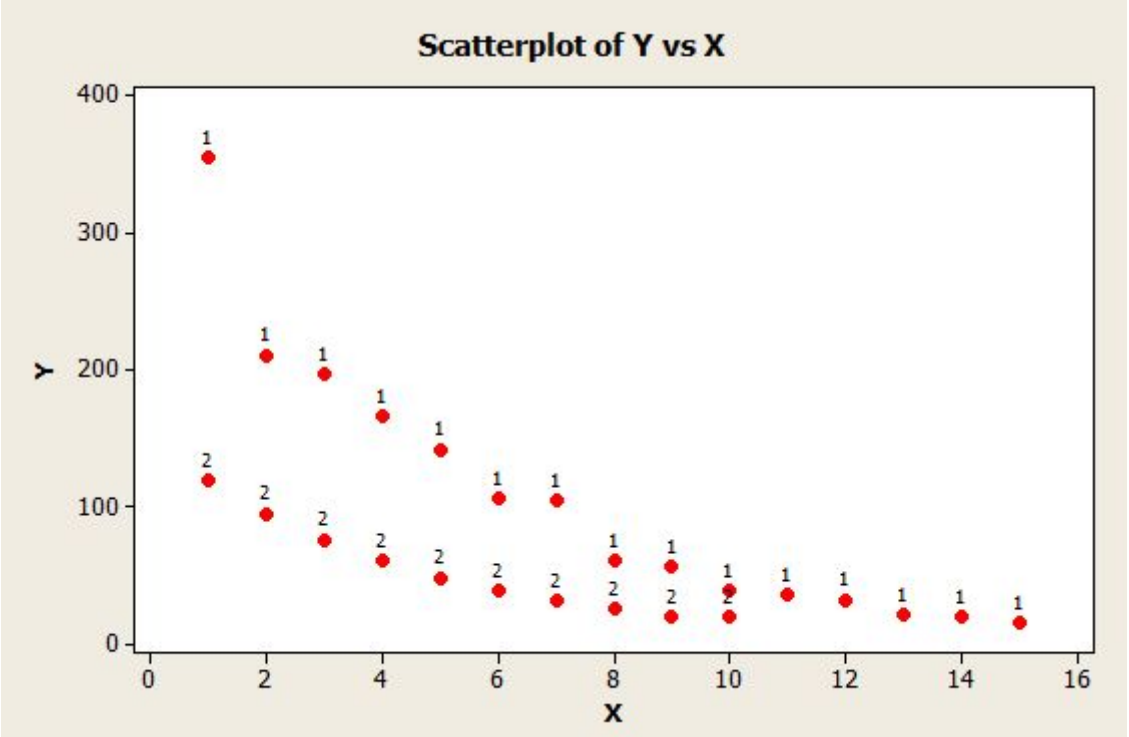
The regression equation is
mpg = 9.62 - 3.92 wt + 1.23 qsec + 2.94 am

Παρατηρούμε την αρνητική κλίση της wt ανεξάρτητης μεταβλητής καθώς και της σημαντικότητας της απο τον συντελεστή, με την παραπάνω εξίσωση μπορούμε να δούμε την αναμενόμενη μεταβολή της mpg για μια μονάδα μεταβολής κάθε μιας ανεξάρτητης μεταβλητης αν οι υπόλοιπες είναι σταθερές.

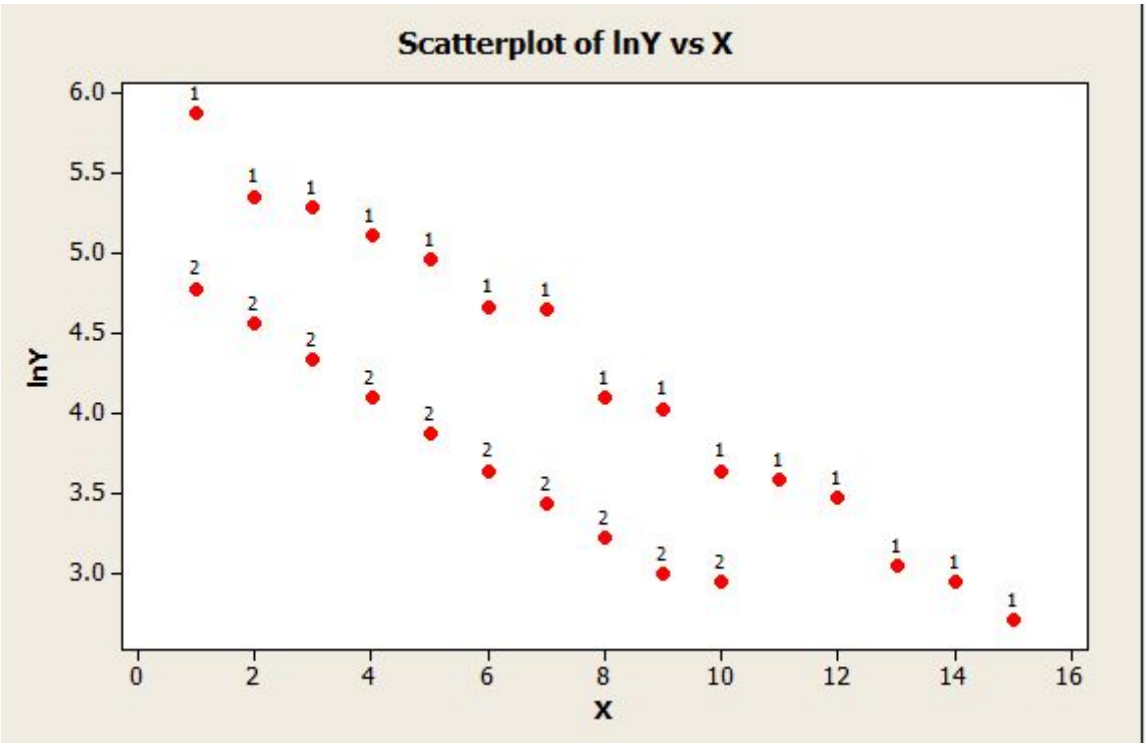
Β)

1)

Πριν:



Μετά:



2)

Κατασκευάζουμε ένα γενικό γραμμικό μοντέλο της μορφής:

$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \varepsilon$,

x_1 : ανεξάρτητη μεταβλητή
x_2 : 1 για την ομάδα 2 και 0 για την ομάδα 1
x_3 : $x_1 * x_2$, εκφράζει την αλληλεπίδραση

Και προκύπτουν οι 3 διαφορετικές περιπτώσεις :

1. $b_3 \neq 0$: δύο διαφορετικές ευθείες για τις 2 ομάδες
2. Αν $b_3 = 0$ και $b_2 \neq 0$, τότε έχουμε 2 παράλληλες ευθείες
3. Αν $b_3 = 0$ και $b_2 = 0$, τότε έχουμε 1 ευθεία.

Για να αποφασίσουμε σε ποιά κατάσταση βρισκόμαστε δημιουργούμε πρώτα το γενικό μοντέλο και βλέπουμε την p-value του x_3. Αν p-value<0.001 απορρίπτουμε την μηδενική υπόθεση ότι δηλαδή $b_3 = 0$. Αν όμως την αποδεχτούμε δημιουργούμε ένα δεύτερο μοντέλο,

$y = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon$, και παρατηρούμε την p-value του x_2 για να αποφασίσουμε για την μηδενική υπόθεση της b_2.

3)

Πραγματοποιούμε τους παραπάνω ελέγχους και έχουμε τα εξής :

```
The regression equation is
lnY = 5.97 - 0.218 X - 1.01 label2 + 0.0051 X3


Predictor       Coef   SE Coef       T      P
Constant     5.97316   0.05023  118.92  0.000
X          -0.218425  0.005524  -39.54  0.000
label2      -1.01407   0.08069  -12.57  0.000
X3           0.00513   0.01158    0.44  0.662
```

Επομένως αφού p-value=0.662 $\geq$ 0.001 δεν απορρίπτουμε την μηδενική υπόθεση , δηλαδή την $b_3 = 0$.

Στη συνέχεια στο επόμενο μοντέλο μας έχουμε το $y = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon$ και έχουμε το εξής αποτέλεσμα :

```
The regression equation is
lnY = 5.96 - 0.217 X - 0.983 label2


Predictor       Coef   SE Coef       T      P
Constant     5.96381   0.04475  133.27  0.000
X          -0.217257  0.004766  -45.59  0.000
label2      -0.98291   0.03891  -25.26  0.000
```

Επομένως απορρίπτουμε την μηδενική υπόθεση και βρισκόμαστε την **2η περίπτωση** , ότι δηλαδή έχουμε δύο διαφορετικά γκρουπ βακτηρίων (δύο ευθείες) με τον ίδιο ρυθμό μείωσης επιζώντων (παράλληλες ευθείες).