Detecting Outliers In Multiple Linear Regression

م. إيهاب عبد السلام محمود

By: Ehab A. Mahmood
Babylon University / Administration and Economic College

Abstract :

It is well-known that the existence of outliers in the data will adversely affect the efficiency of estimation and results of the current study. In this paper four methods will be studied to detect outliers for the multiple linear regression model in two cases : first, in real data; and secondly, after adding the outliers to data and the attempt to detect it. The study is conducted for samples with different sizes, and uses three measures for comparing between these methods . These three measures are : the mask, dumping and standard error of the estimate.

اكتشاف القيم الشاذة في الانحدار الخطي المتعدد

من المعروف ان وجود القيم الشاذة في البيانات يؤثر سلبا على كفاءة التقديرات والنتائج للدراسة الموضوعة، وفي هذا البحث سيتم دراسة (4) طرائق لاكتشاف القيم الشاذة لنموذج الانحدار الخطي المتعدد ولحالتين: لبيانات حقيقية والحالة الثانية بعد اقحام قيم شاذة للبيانات ومحاولة كشفها، وقد تمت الدراسة باحجام عينات مختلفة واعتماد (3) مقاييس للمقارنة بين هذه الطرائق هي: القناع، الاغراق والخطأ المعياري للتقدير .

1. Introduction :

There are many methods for detecting outliers in linear regression model as:

Elashoff (1972) studied the linear regression model. She illustrated the existing outliers cause the bias in estimator and the high variance . Draper and John (1981) illustrated the benefit of using the Cook Distance.

Pena and Yohai (1999) suggested fast procedure to estimate linear regression parameters in case of existing outliers and how to detect it . Chen (2003) detected outliers in multiple linear regression model. He depended on many robust estimate methods such as (LTS) . Gal (2005) presented several methods for the detection of outliers in univariate and multivariate . Karpinski (2007) illustrated in his book the outliers and how to detect them by using several methods .

Mishra (2008) studied several robust and non-robust methods for detecting outliers in multiple linear regression ; he used a Monte Carlo method for comparison between real data and theoretical data .

Asikgil and Erar (2009) tried to determine multiple outliers by using various methods in the presence of masking and swamping effects for the linear regression model .

The multiple linear regression model is as the following equation :

$$Y = X\beta + € \qquad \text{…….(1.1)}$$

where :

Y : vertical vector (n*1) of observed response values.

X : matrix (n*p) of (p) regressors .

β : vertical vector (p*1) of regression coefficients .

€ : vertical vector (n*1) of error terms .

n : sample size .

The method of ordinary least squares (OLS) is the most widely used technique to find the best estimates of (β) which minimizes the sum of squared distance for actual observations to the regression surface under the assumption (€~NID(0,σ^2I)); but if the data has outliers the assumption is not satisfied and the estimate dose not minimize the sum of squared distance and will not be optimal . In this case, we must firstly detect outliers and treat them and then apply (OLS) method or we can estimate (β) by robust methods of estimate instead of (OLS) method.

Outlier : we can define the outlier as; the observation (or subset of observations) that appear inconsistent (extreme) with the remainder of the data set and has a profound destructive influence on the statistical analysis ; and in linear regression model is not necessarily be extreme (Barnett & Lewis 1994) .

There are several types of outliers in linear regression :

i. **In X-Space : If one or more of the observation values lie far away from the group observations at the (X) axis .**
ii. **In Y-Space : If one or more of the observation values lie far away from the group observations at the (Y) axis .**
iii. **In (XY) - Space : If one or more of the observation values lie far away from the group observations at the (X) and (Y) axis.**

Care should be taken in detecting the outlier in set data to prevent masking and swamping problems ; where :
masking ; the unable of the procedure to detect the outliers , swamping; consider the clean observations as outliers (Adnan and others 2003) .

2. Methods Of Detecting Outliers :

There are various methods to detect the influential observations in linear regression model. Some of these methods is to detect a single outlier and the other is to detect multiple outliers (single – row diagnostics). The single – row diagnostics can be extended to include subset of observations rather than a single observation (Belsley & Welsch 1980). In this paper, we will use some of widely-used measures that depend on the single – row diagnostics as the following :

2.1 Mahalanobis Distance : (McLachlan 1999)-(Mishra 1994)

Mahalanobis proposed this measure in (1936) to detect contaminated or outlier data points in linear regression model . His measure has played an important role in statistics and data analysis .

The generalized distance can defined as follows :

$$D_i = \sqrt{(y_i - E(y_i))' S^{-1} (y_i - E(y_i))} \quad \text{……. (2.1)}$$

Where :

S : covariance matrix .

We will reject the null hypothesis; and the observation will be outlier when :

$$D_i^2 > \chi^2_{(n-p,\alpha)}$$

In the linear regression model can compute the distance as the following equation :

$$D_i = \sqrt{(\hat{y}_i - \bar{\hat{y}})' S^{-1} (\hat{y}_i - \bar{\hat{y}})} \quad \text{……. (2.2)}$$

Where :

$\hat{y}_i$ **: forecasting value .**

$\bar{\hat{y}}$ **: forecasting values mean .**

We can compute Mahalanobis distance in many statistical packages like SPSS .

2.2 Cook‘s Distance : (Cook 1979)

In (1979) Cook presented a method to detect the influential observation in multiple linear regression which is based on the measure of the distance between ($\hat{\beta}$) and ($\hat{\beta}_i$) as follows :

$$D_i = \frac{(\hat{\beta}_i - \hat{\beta})' x' x (\hat{\beta}_i - \hat{\beta})}{p S^2} \leq F_{(p,n-p,1-\alpha)} \quad \text{…….. (2.3)}$$

Where :

$\hat{\beta}$ **: denotes the least square estimate of (*β*) .**

$\hat{\beta}_i$ **: denotes the least squares estimate of (*β*) with the (ith) point deleted.**

$$S^2 = \frac{\hat{e}'\hat{e}}{n-p} \quad \text{…….. (2.4)}$$

If the (D_i) > $F_{(p,n-p,1-\alpha)}$; then the (ith) single – row is an outliers .

2.3 Serbert, Montgomery and Rollier Procedure (1998) :

(Adnan et al, 2003)

They considered a procedure to identify the outliers in multiple linear regression by using the (OLS) method and the single linkage clustering method , where :

The cluster analysis is a method for detecting a natural groupings of items or variables where the items show a high internal homogeneity and low external homogeneity. It includes two groups: hierarchical and non-hierarchical,where the hierarchical method divided into two types :

i. Agglomerative hierarchical method :

It starts with (n) clusters and ends with one cluster which contains all of the data points . (This was conducted by Serbert and et al) .

ii. Divisive hierarchical method :

It starts with one cluster and ends up with (n) clusters with each cluster contains one data point .

The single linkage clustering method : It is a method that depends on the smallest distance between a data point in the first cluster and a point in the second cluster .

Serbert et al method depends on the following steps :

i. **Find the standardized predicted values (depending on the OLS) .**
ii. **Grouping the data set by using the single linkage clustering algorithm (Agglomerative hierarchical method) with Euclidean distance between pairs of standardized predicted values , and this can be graphically shown in the form of a dendogram or tree diagram .**
iii. **Number of the clusters depend on the height of the cut (stopping rule) ; which determine as the following equation :**

$$ch = \bar{h} + ks_h \quad \text{…… (2.5)}$$

Where :

$\bar{h}$ **: Average height of the tree .**

k : constant .

s_h **: The standard deviation of the heights .**

iv. **The clean data set is the largest cluster formed. It includes the median , and the other clusters contains the outliers .**

2.4 Adnan , Mohamad and Setan Procedure : (Adnan et al 2003)

Adnan et al (2003) proposed a modified procedure of Serbert et al, where they used the robust fit (least trimmed of squares (LTS) instead of the ordinary least squares (OLS) fit; then they applied the backward steps of (Serbert and others) procedure, depending on the standardized predicted values or the residual values .

The (LTS) : It is a method of robust regression estimate proposed by Rousseeuw (1984). It minimizes the sum of squared residuals by selecting smallest (m) of residual and (n-m) residuals are deleted ; and then find the estimators, depending on the (m) observations which satisfy the objective function as the following :

$Min._{\hat{\beta}} \sum_{i=1}^{m} e_i^2$ **(Rousseeuw 1984)**

Where :

m = (n/2) + ((p+1)/2) ……. (2.6)

The (LTS) has a high breakdown point of up to (50%) ; which is the highest possible value . (Georgiev 2008)

Breakdown point : It is the smallest part of unusual data that can cause to false the estimator .

3. Application :

We will apply a aforementioned methods to detecting the outliers : Mahalanobis , Cook , Serbert and Adnan ; and we will use three measures for comparison : masking ; swamping and standard error estimate (which calculate after delete the outliers), we study one of an important disease that infects the people; which is called (Hepatitis Disease). It will represent the dependent variable . This disease depends on four tests to detecting it :

i. **Glutamate Oxaloacetate Transaminase (G.O.T) .**
ii. **Glutamate Pyruvate Transaminase (G.P.T) .**
iii. **Total Serum Bilirubin (T.S.B) .**
iv. **Alkaline Phosphatase (Alk) .**

They will represent the independent variables .

In this paper, we will study several sample sizes , small (n=25) ; medium (n=50) ; large (n=150) and for two cases : firstly real values of observations; and secondly, after adding (10%) of outliers to this observations .

i. (n = 25) :

We will apply the four detecting methods to the real observations which appear in (table 3.1), and to the observations after adding the (3) outliers in the No. of (5,10,15) . The results for Mahalanobis and Cook distances are shown in the table (3.2) .

The dendrograms for Serbert and Adnan for the real observations are shown in figures (3.1),(3.2), respectively; and the dendrograms for the observations after adding the outliers are shown in figures (3.3),(3.4), respectively .

Table (3.1) Independent variables (n=25)

No.	Diseased	G.O.T.	G.P.T	Alk	T.S.B
1	no	13	12	63	6.8
2	yes	40	44	510	73.5
3	no	16	14	108	8.9
4	yes	200	360	189	153.0
5	yes	100	356	183	82.0
6	no	18	20	99	6.8
7	yes	75	65	243	98.0
8	no	19	18	78	5.1
9	yes	100	232	279	202.0
10	yes	90	95	510	61.0
11	yes	96	272	252	342.0
12	no	15	14	60	5.3
13	no	6	8	90	8.0
14	yes	80	216	234	136.0
15	yes	39	280	246	34.2
16	no	18	16	137	8.5
17	no	5	7	151	3.4
18	yes	148	328	267	51.3
19	yes	76	280	495	76.9
20	no	4	7	40	6.8
21	no	17	16	144	5.1
22	yes	100	172	270	342.0
23	yes	133	312	228	35.0
24	yes	80	344	288	49.5
25	yes	35	35	245	117.9

Reference : Educational Babylon Hospital for Women and Children

Table (3.2) Mahalanobis and Cook distance (n=25)

No.	Before		After	
	Mahalanobis	Cook	Mahalanobis	Cook
1	11.99857	0.03062	12.00222	0.00410
2	1.27141	0.08668	1.27052	0.00449
3	1.01633	0.05353	1.00514	0.00116
4	6.00028	0.06635	6.03832	0.00020
5	4.91981	0.01475	5.14593	0.33897
6	0.70824	0.04492	0.70024	0.00015
7	2.76970	0.00999	2.77873	0.00007
8	1.14788	0.02159	1.17660	0.00963
9	3.69325	0.00180	3.72104	0.00387
10	7.48166	0.09509	7.27161	0.43499
11	0.72748	0.03141	0.71056	0.00117
12	2.77173	0.00134	2.76441	0.00130
13	2.68825	0.01193	2.69780	0.01876
14	4.33185	0.02453	4.23407	0.04126
15	1.20830	0.01586	1.16285	0.15982
16	10.68513	0.14125	10.73735	0.02747
17	13.48812	0.00001	13.53214	0.00193
18	1.22275	0.01605	1.24265	0.00481
19	1.21766	0.01517	1.23113	0.00389
20	2.09132	0.01000	2.09130	0.00032
21	1.25078	0.01701	1.25262	0.00349
22	0.83532	0.00856	0.87174	0.00143
23	1.40295	0.01481	1.40620	0.00195
24	9.80173	0.01490	9.67376	0.90935
25	1.26947	0.01459	1.28106	0.00347

Figure (3.1) Serbert before (n=25)

Rescaled Distance Cluster Combine

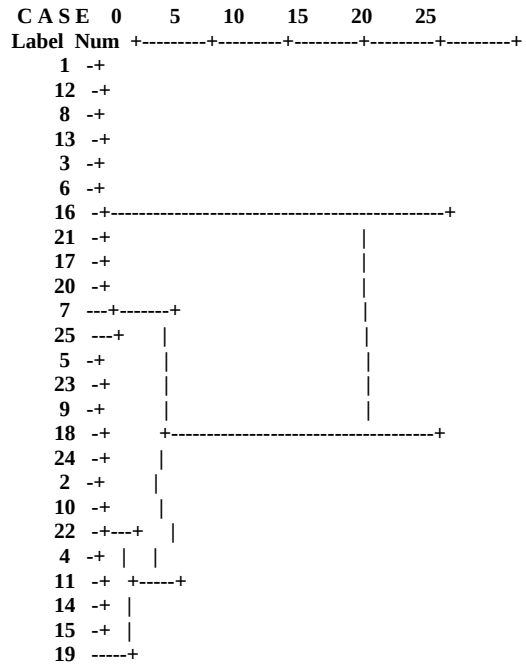


Figure (3.2) Adnan before (n=25)

Rescaled Distance Cluster Combine

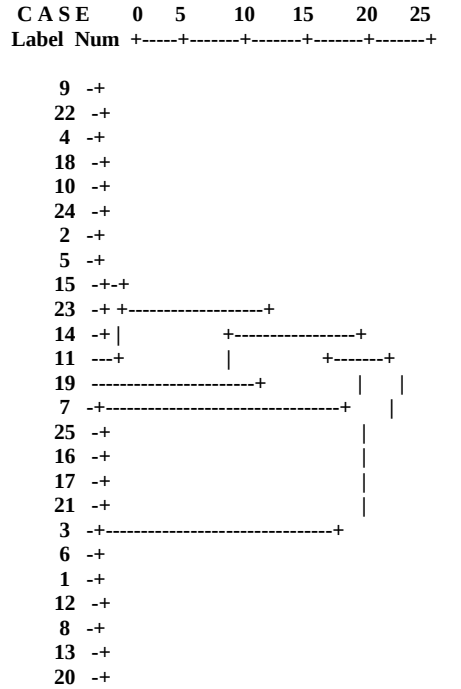


Figure (3.3) Serbert after (n=25)

Rescaled Distance Cluster Combine

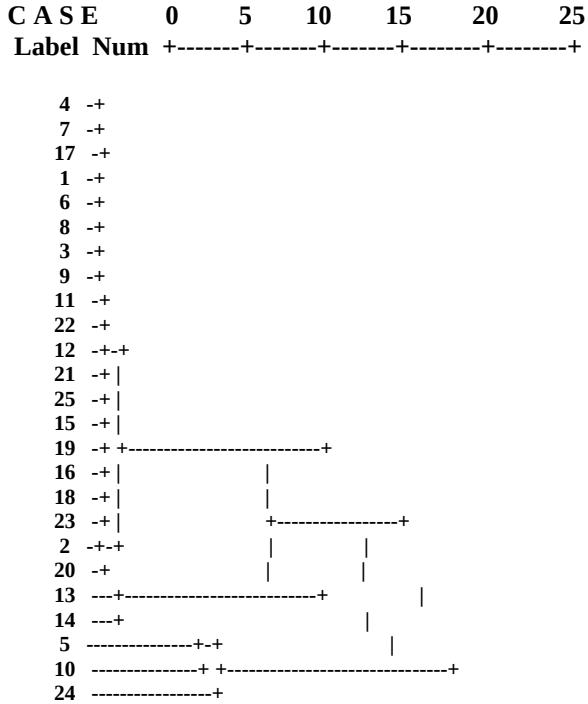


Figure (3.4) Adnan after (n=25)

Rescaled Distance Cluster Combine

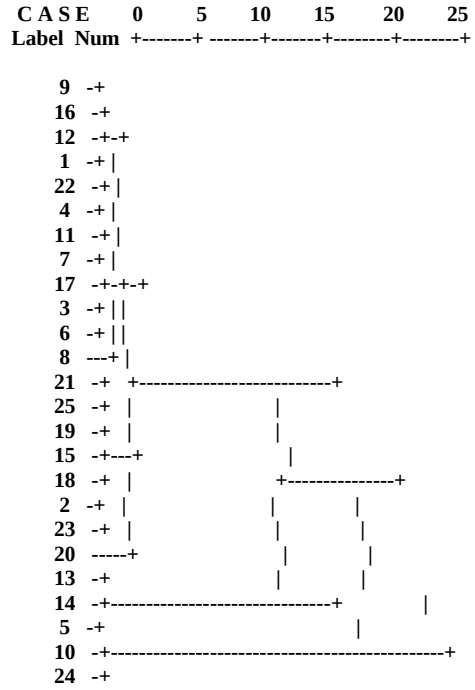


Table (3.3) shows the summary for the methods and contains (7) columns :

i. (Case) : Represents before and after adding the outliers .

ii. (Method) : includes (non) i.e without depending on any method for detecting outliers , and the methods of detect .

iii. (ch) : calculate the (ch) value by the equation (2.4) for Serbert and Adnan methods .

iv. (Outliers) : The outliers detected by the methods ; where they are all compared with Mahalanobis distance ($\chi^2_{(20,0.05)}$=31.4) , and Cook distance ($F_{(5,20,0.05)}$=2.71) . Both methods can not detect any outliers in the two cases . Outliers are detected by Serbert and Adnan , depending on (ch) value . Differences are found between them for (before) case, and same results for the (after) case .

v. (Masking) : Mahalanobis and Cook have masking for all the adding outliers ; but Serbert and Adnan have only No. (15) .

vi. (Swamping) : Mahalanobis and Cook did not have swamping ; but Serbert and Adnan have the same swamping in the No. (13,14,24) .

vii. (Std. Error Est.) : Adnan has (0.09) ; which is less than others for (before) case , and for the (after) case Serbert and Adnan have (1.2) which is less than others .

Table (3.3) Summary (n=25)

case	method	ch	outliers	masking	swamping	Std. Error Est.
before	none	-	-	-	-	0.2
	Mah.	-	-	-	-	0.2
	Cook	-	-	-	-	0.2
	Serbert	8.6	(16,17,20,21)	-	-	0.2
	Adnan	11.1	(3,7,16,17,19, 21,25)	-	-	0.09
after	none	-	-	-	-	1.58
	Mah.	-	-	5,10,15	-	1.58
	Cook	-	-	5,10,15	-	1.58
	Serbert	9.9	(5,10,13,14,24)	15	(13,14,24)	1.2
	Adnan	9.9	(5,10,13,14,24)	15	(13,14,24)	1.2

ii. (n = 50) :

We will apply the detecting methods to the real observations which are shown (table 3.4) , and to the observations after adding the (5) outliers in the No. of (1,10,20,35,45) . The results for Mahalanobis and Cook distances are shown in the table (3.5) ; the dendrograms for Serbert and Adnan for the real observations are shown in figures (3.5),(3.6), respectively. The dendrograms for the observations after adding the outliers will are shown in figures (3.7),(3.8), respectively .

Table (3.4) Independent variables (n=50)

No.	Diseased	G.O.T.	G.P.T	Alk	T.S.B
1	yes	55	129	238	342.0
2	yes	29	80	72	20.5
3	yes	24	25	180	120.0
4	yes	22	248	198	47.8
5	yes	50	89	370	31.0
6	yes	27	37	160	32.5
7	yes	51	270	207	71.5
8	no	5	4	136	5.8
9	yes	62	310	254	24.0
10	yes	37	262	481	81.0
11	yes	31	50	199	51.3
12	yes	96	332	258	83.3
13	yes	78	331	348	85.5
14	yes	110	318	351	63.3
15	no	15	11	90	3.4
16	yes	104	312	248	342.0
17	yes	152	310	199	65.0
18	no	6	8	100	3.4
19	no	9	8	92	5.1
20	no	5	5	36	8.5
21	no	19	13	88	6.8
22	yes	33	168	243	85.5
23	no	17	16	69	5.3
24	yes	37	80	490	106.0
25	no	9	5	87	8.3
26	no	13	12	63	6.8
27	yes	40	44	510	73.5
28	no	16	14	108	8.9
29	yes	200	360	189	153.0
30	yes	100	356	183	82.0
31	no	18	20	99	6.8
32	yes	40	35	243	98.0
33	no	19	18	78	5.1
34	yes	100	232	279	202.0
35	yes	90	95	510	61.0

36	yes	96	272	252	342.0
37	no	15	14	60	5.3
38	no	6	8	90	8.0
39	yes	80	216	234	136.0
40	yes	39	280	246	34.2
41	no	18	16	137	8.5
42	no	5	7	151	3.4
43	yes	148	328	267	51.3
44	yes	76	280	495	76.9
45	no	4	7	40	6.8
46	no	17	16	144	5.1
47	yes	100	172	270	342.0
48	yes	133	312	228	35.0
49	yes	80	344	288	49.5
50	yes	35	35	245	117.9

Reference : Educational Babylon Hospital for Women and Children .

Table (3.5) Mahalanobis and Cook distance (n=50)

No.	Before		After	
	Mahalanobis	Cook	Mahalanobis	Cook
1	11.55246	0.00804	22.60231	0.82894
2	1.26605	0.07605	1.03690	0.00000
3	0.83997	0.04503	0.83827	0.00015
4	6.04798	0.03984	4.79790	0.00119
5	3.34130	0.01218	3.12448	0.00305
6	0.61491	0.03962	0.62250	0.00006
7	3.18641	0.00968	2.52039	0.00006
8	1.16377	0.00749	1.13489	0.00074
9	4.25875	0.00285	4.46694	0.00069
10	8.50483	0.05691	8.36947	0.94038
11	0.50593	0.02585	0.53584	0.00005
12	2.47368	0.00015	2.41421	0.00078
13	3.49111	0.00842	3.85138	0.00056
14	3.02260	0.00621	3.07207	0.00456
15	1.30307	0.00432	1.25226	0.00059
16	9.91979	0.05116	9.35716	0.01477
17	7.90796	0.00559	6.05256	0.01822
18	1.29367	0.00465	1.21089	0.00088
19	1.30803	0.00418	1.23517	0.00083
20	2.12532	0.00109	2.49348	0.11222
21	1.32687	0.00462	1.28224	0.00052
22	1.26710	0.00989	0.89731	0.00008
23	1.50046	0.00346	1.42229	0.00071
24	7.46761	0.00148	7.27531	0.00157
25	1.36244	0.00382	1.28912	0.00091
26	1.58730	0.00283	1.48905	0.00095
27	9.76717	0.00000	9.09608	0.00826
28	1.09685	0.00589	1.06456	0.00052
29	15.72338	0.01314	13.62994	0.06395
30	4.14689	0.00154	3.60324	0.00063
31	1.13425	0.00560	1.08577	0.00050
32	1.52601	0.03101	1.58316	0.00011
33	1.39312	0.00419	1.32679	0.00057
34	2.30459	0.00023	2.68846	0.00002
35	11.99321	0.01148	9.53270	0.58342
36	9.62294	0.02815	9.30624	0.01547
37	1.62746	0.00278	1.53233	0.00087
38	1.36153	0.00416	1.25187	0.00104
39	0.66540	0.00191	0.70630	0.00002
40	5.16338	0.00994	4.90688	0.00002
41	0.95809	0.00845	0.94648	0.00034
42	1.15217	0.00924	1.14591	0.00066
43	6.84444	0.00089	5.13656	0.01728

44	5.79342	0.05676	6.73713	0.00521
45	2.07649	0.00130	2.48259	0.11139
46	0.96551	0.00900	0.95832	0.00032
47	10.31945	0.00421	11.91207	0.00994
48	5.56087	0.00272	4.24201	0.01116
49	4.01669	0.00093	4.28585	0.00118
50	2.14730	0.02702	2.19439	0.00000

Figure (3.5) Serbert before (n=50)

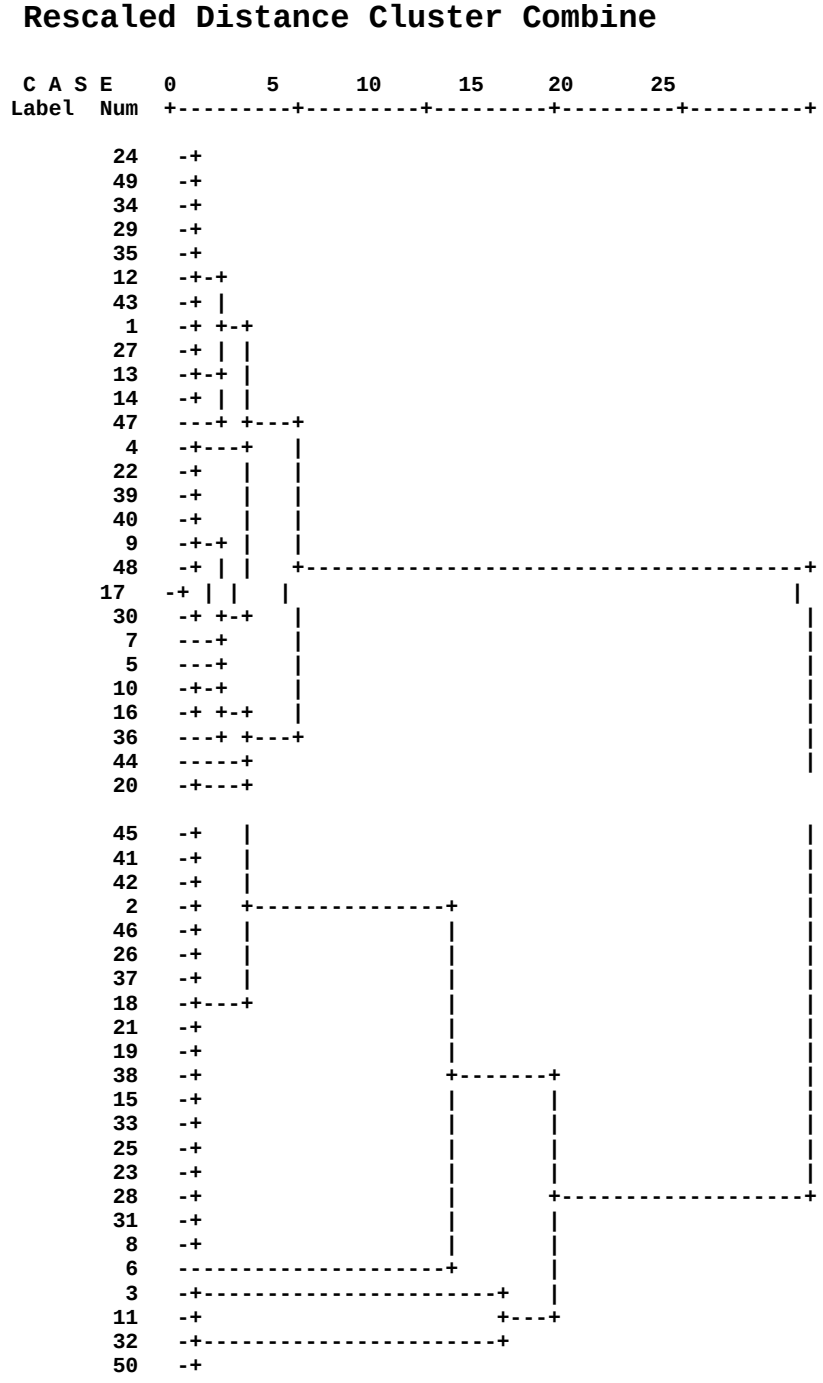


Figure (3.6) Adnan before (n=50)

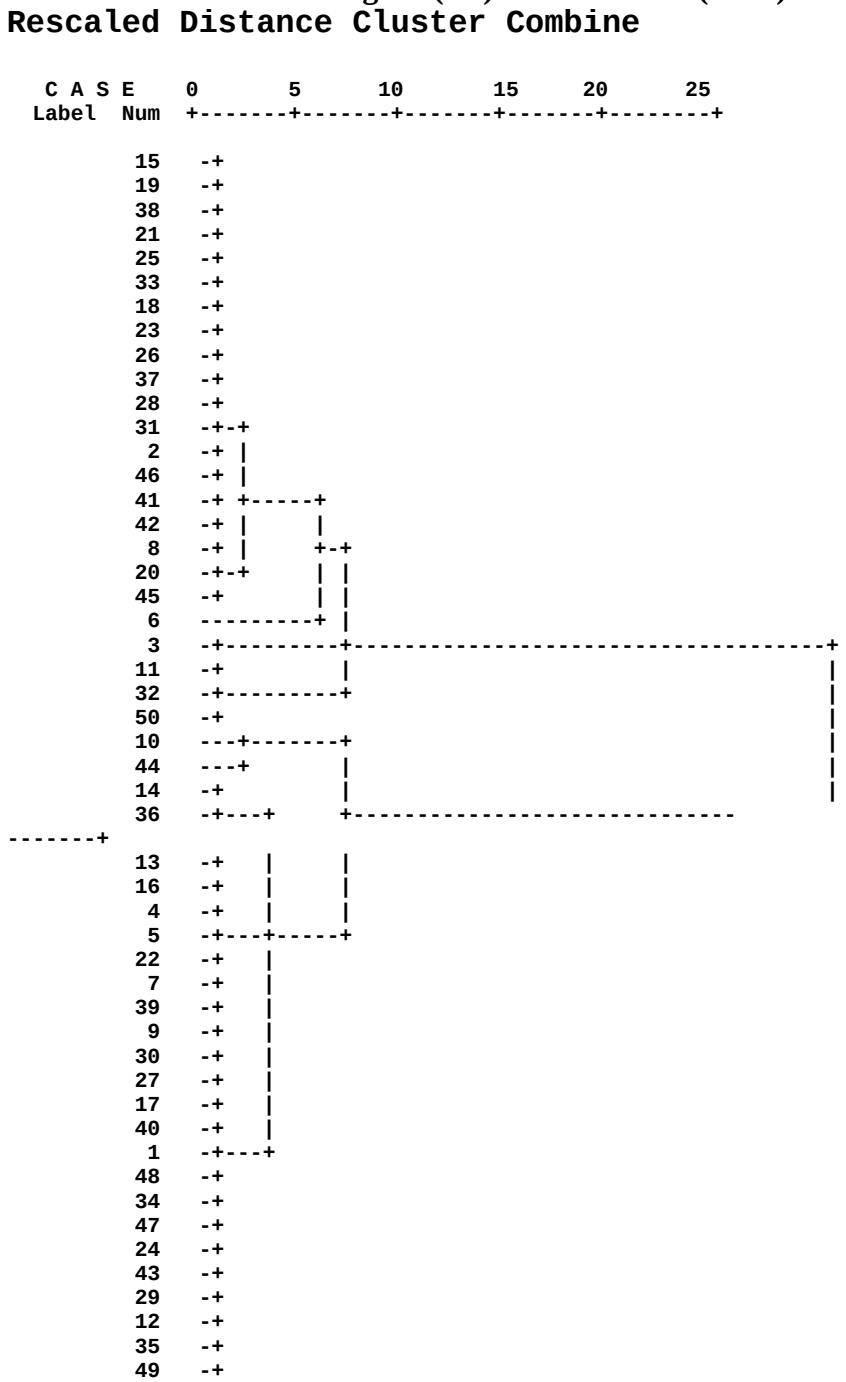


Figure (3.7) Serbert after (n=50)

```
Rescaled Distance Cluster Combine

  C A S E    0       5       10      15      20      25
 Label  Num  +-------+-------+-------+-------+-------+
         25   -+
         50   -+
         18   -+
         19   -+
         26   -+
          2   -+
          8   -+
         37   -+
         34   -+
         42   -+
         23   -+
         31   -+
         33   -+
         15   -+
         28   -+
         39   -+
         21   -+
          6   -+
         32   -+
         41   -+
         11   -+
         46   -+
         38   -+
         40   -+
          9   -+
         13   -+
         30   -+
         24   -+
          3   -+
         12   -+
         49   -+
         20   -+
         22   -+
         45   -+
          7   -+
          4   -+
         27   -+
         44   -+
          5   -+
         14   -+
         16   -+
         36   -+
         10   -+-----------------------------------------------+
         47   -+                                               |
         29   -+                                               |
         43   -+                                               |
         35   -+                                               |
         17   -+                                               |
         48   -+                                               |
          1   -------------------------------------------------+
```

Figure (3.8) Adnan after (n=50)

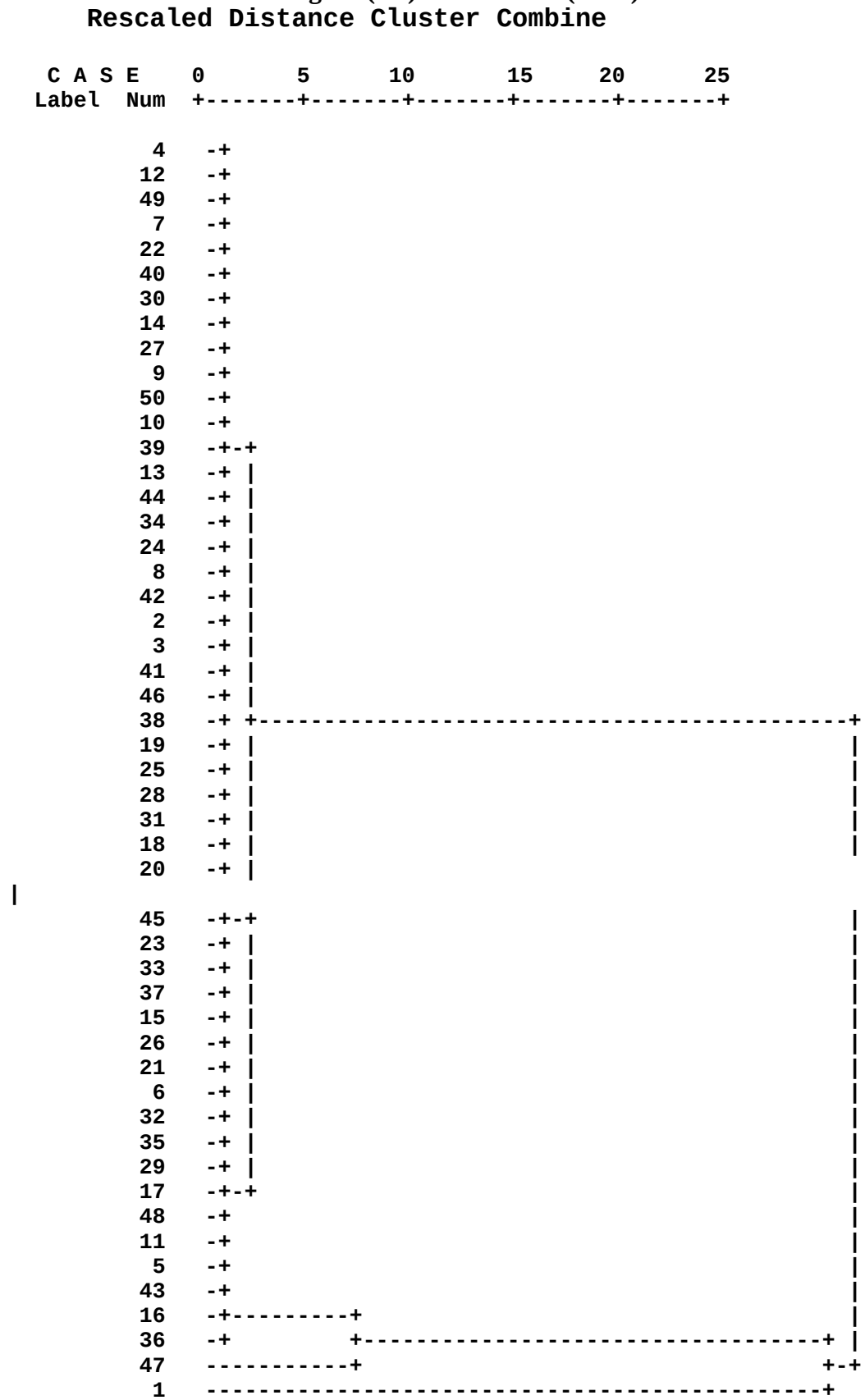


Table (3.6) below shows the summary for the methods where they are all compared with Mahalanobis distance ($\chi^2_{(45,0.05)}$=61.4) , and Cook distance ($F_{(5,45,0.05)}$=2.42) . Both methods can not detect any outliers in the two cases . Outliers are detected by Serbert and Adnan , depending on (ch) value . Differences are found between them for (before) case, and same results for the (after) case .

Mahalanobis and Cook have masking for all the adding outliers ; but Serbert has only two in the No. (20,45) ; and Adnan has in the No. (10,20,35,45)

Mahalanobis and Cook did not have swamping ; but Serbert has in the No. (17,29,43,47,48) ; and Adnan has in the No. (5,11,16,36,43,47,48) .

Serbert has (0.23),(2.14) Std. Error Est. before and after cases, respectively ; which are less than others .

Table (3.6) Summary (n=50)

case	method	ch	outliers	masking	swamping	Std. Error Est.
Before	none	-	-	-	-	0.26
	Mah.	-	none	-	-	0.26
	Cook	-	none	-	-	0.26
	Serbert	7.7	(3,6,8,11,15,19,21,23,25,28, 31,32,33, 38)	-	-	0.23
	Adnan	6.3	(2,3,6,8,11,15,18,19, 20,21,23,25,26,28, 31, 32,33,37, 38,41,42,45, 46)	-	-	-
After	none	-	-	-	-	3.32
	Mah.	-	-	1,10,20,35,45	-	3.32
	Cook	-	-	1,10,20,35,45	-	3.32
	Serbert	11.6	(1,10,17,29,35, 43,47, 48)	20,45	17,29,43, 47,48	2.14
	Adnan	8.4	(1,5,11,16,36, 43,47, 48)	10,20,35, 45	5,11,16,36,43,47 ,48	3.46

iii. (n = 150) :

The observations appear in (table 3.7) ; and adding (15) outliers for the No. of (10,20,35,48,65,70,85,90,100,108,115,125,133,140,148) .

The results for Mahalanobis and Cook distances are shown in table (3.8) ; the dendrograms for Serbert and Adnan for the real observations are shown in figures (3.9),(3.10), respectively ; and the dendrograms for the observations after adding the outliers are shown in figures (3.11),(3.12), respectively .

Table (3.7) Independent variables (n=150)

No.	Diseased	G.O.T.	G.P.T	Alk	T.S.B
1	yes	55	129	238	342.0
2	yes	29	80	72	20.5
3	yes	24	25	180	21.0
4	yes	22	248	198	47.8
5	yes	50	89	370	31.0
6	yes	27	37	160	32.5
7	yes	51	270	207	71.5
8	no	5	4	136	5.8
9	yes	62	310	254	24.0
10	yes	37	262	481	81.0
11	yes	31	50	199	51.3
12	yes	96	332	258	83.3
13	yes	78	331	348	85.5
14	yes	110	318	351	63.3
15	no	15	11	90	3.4
16	yes	104	312	248	342.0
17	yes	152	310	199	65.0
18	no	6	8	100	3.4
19	no	9	8	92	5.1
20	no	5	5	36	8.5
21	no	19	13	88	6.8
22	yes	33	168	243	85.5
23	no	17	16	69	5.3
24	yes	37	80	490	106.0
25	no	9	5	87	8.3
26	no	13	12	63	6.8
27	yes	40	44	510	73.5
28	no	16	14	108	8.9
29	yes	200	360	189	153.0
30	yes	100	356	183	82.0
31	no	18	20	99	6.8
32	yes	40	35	243	90.0
33	no	19	18	78	5.1
34	yes	100	232	279	202.0
35	yes	90	95	510	61.0
36	yes	96	272	252	342.0
37	no	15	14	60	5.3
38	no	6	8	90	8.0
39	yes	80	216	234	136.0
40	yes	39	280	246	34.2
41	no	18	16	137	8.5
42	no	5	7	151	3.4
43	yes	148	328	267	51.3
44	yes	76	280	495	76.9
45	no	4	7	40	6.8
46	no	17	16	144	5.1
47	yes	100	172	270	342.0
48	yes	133	312	228	35.0
49	yes	80	344	288	49.5
50	yes	35	35	276	117.9
51	yes	96	306	287	37.6
52	yes	31	39	191	28.0
53	yes	25	20	147	48.0
54	yes	29	94	367	25.0
55	no	5	4	115	3.5
56	no	17	10	122	4.0
57	no	11	4	96	3.3
58	no	5	5	114	3.4
59	yes	65	50	398	109.4
60	yes	25	176	297	77.0
61	no	110	4	100	3.3
62	no	6	9	117	8.5
63	yes	50	48	196	203.0
64	yes	48	48	186	203.0

65	yes	70	88	900	60.0
66	no	13	4	127	8.5
67	yes	124	368	198	83.0
68	yes	51	264	219	107.0
69	yes	29	26	150	58.0
70	no	9	7	94	8.3
71	yes	25	34	525	98.0
72	yes	74	251	517	205.5
73	yes	50	350	496	225.7
74	yes	121	282	405	87.5
75	no	8	8	94	0.5
76	yes	250	350	496	200.0
77	yes	224	304	291	205.2
78	no	9	8	81	0.4
79	no	11	7	69	0.8
80	yes	44	136	306	13.7
81	no	8	16	114	7.7
82	no	25	9	55	0.5
83	no	8	10	156	0.6
84	no	27	9	126	0.9
85	yes	144	184	105	138.0
86	no	15	4	132	5.9
87	no	9	9	105	8.5
88	yes	135	240	280	51.3
89	no	8	12	123	6.8
90	yes	128	44	200	29.0
91	yes	148	25	189	20.5
92	no	10	14	114	136.0
93	yes	34	34	42	18.1
94	yes	250	428	246	27.3
95	yes	114	264	222	48.5
96	no	17	8	58	5.2
97	no	15	15	78	5.3
98	no	15	4	120	3.5
99	yes	50	176	291	38.5
100	no	5	5	105	3.5
101	no	17	74	93	8.6
102	no	15	5	129	5.8
103	no	18	19	105	3.8
104	no	13	7	120	8.5
105	no	8	9	84	8.5
106	no	19	5	102	5.1
107	no	10	20	117	8.5
108	yes	60	72	96	32.5
109	yes	82	132	112	38.7
110	yes	41	72	105	19.3
111	yes	104	270	300	71.8
112	no	15	5	120	8.5
113	no	8	10	129	5.8
114	no	9	4	140	15.4
115	no	6	5	72	14.3
116	no	13	18	120	11.0
117	yes	120	334	250	51.5
118	yes	68	88	264	20.1
119	No	17	18	90	5.3
120	yes	125	256	420	28.8
121	no	8	8	123	9.8
122	no	5	12	72	13.3
123	yes	112	304	359	25.5
124	yes	70	312	360	45.5
125	yes	47	352	276	170.5
126	no	8	8	51	3.5
127	yes	41	416	228	222.0
128	no	60	26	298	8.5
129	no	5	5	129	8.5

130	yes	85	424	390	117.0
131	yes	35	120	290	38.5
132	no	10	18	120	8.5
133	yes	35	88	190	38.5
134	yes	104	400	210	36.8
135	yes	125	432	207	53.5
136	yes	33	328	264	18.5
137	no	4	4	90	8.5
138	no	19	9	150	10.2
139	no	10	10	50	13.3
140	no	8	6	85	13.6
141	no	9	5	66	6.8
142	no	8	7	102	5.1
143	yes	35	34	120	29.1
144	no	11	5	90	5.1
145	yes	80	232	225	66.3
146	no	8	9	111	6.8
147	yes	35	304	150	29.4
148	yes	52	30	90	53.0
149	yes	100	280	419	18.9
150	yes	47	200	430	20.5

Reference : Educational Babylon Hospital for Women and Children .

Table (3.8) Mahalanobis and Cook distance (n=150)

No.	Before		After	
	Mahalanobis	Cook	Mahalanobis	Cook
1	21.89768	0.00106	17.20784	0.00003
2	1.00962	0.01609	0.96210	0.00009
3	0.55319	0.01068	0.57237	0.00002
4	4.51209	0.01019	4.23912	0.00001
5	3.03155	0.00609	3.56580	0.00000
6	0.35026	0.00898	0.32347	0.00002
7	2.87809	0.00306	3.16342	0.00005
8	0.88865	0.00119	0.81017	0.00013
9	4.50912	0.00209	4.71983	0.00009
10	7.54123	0.00091	21.73428	0.50364
11	0.38916	0.00687	0.43290	0.00001
12	2.89401	0.00000	3.36871	0.00016
13	3.50713	0.00049	4.02198	0.00001
14	2.95976	0.00060	3.33918	0.00011
15	0.88644	0.00085	0.78069	0.00007
16	18.77024	0.03090	15.14304	0.00014
17	5.75783	0.00005	3.67041	0.00087
18	0.93799	0.00080	0.85119	0.00010
19	0.92546	0.00077	0.83772	0.00008
20	1.59615	0.00032	25.67554	0.36890
21	0.86578	0.00095	0.73350	0.00006
22	1.16241	0.00351	1.14163	0.00000
23	1.04121	0.00079	0.91223	0.00006
24	7.88081	0.00192	8.91247	0.00039
25	0.95792	0.00073	0.86338	0.00008
26	1.13082	0.00065	1.01687	0.00006
27	9.66911	0.00521	11.12950	0.00043
28	0.72274	0.00110	0.63619	0.00007
29	12.09855	0.01074	6.83061	0.00236
30	4.79013	0.00022	5.27816	0.00047
31	0.74817	0.00110	0.65028	0.00006
32	1.63944	0.00932	1.68268	0.00001
33	0.94000	0.00091	0.80751	0.00005
34	4.62171	0.00020	3.77549	0.00008
35	9.31505	0.00020	8.30275	0.19081

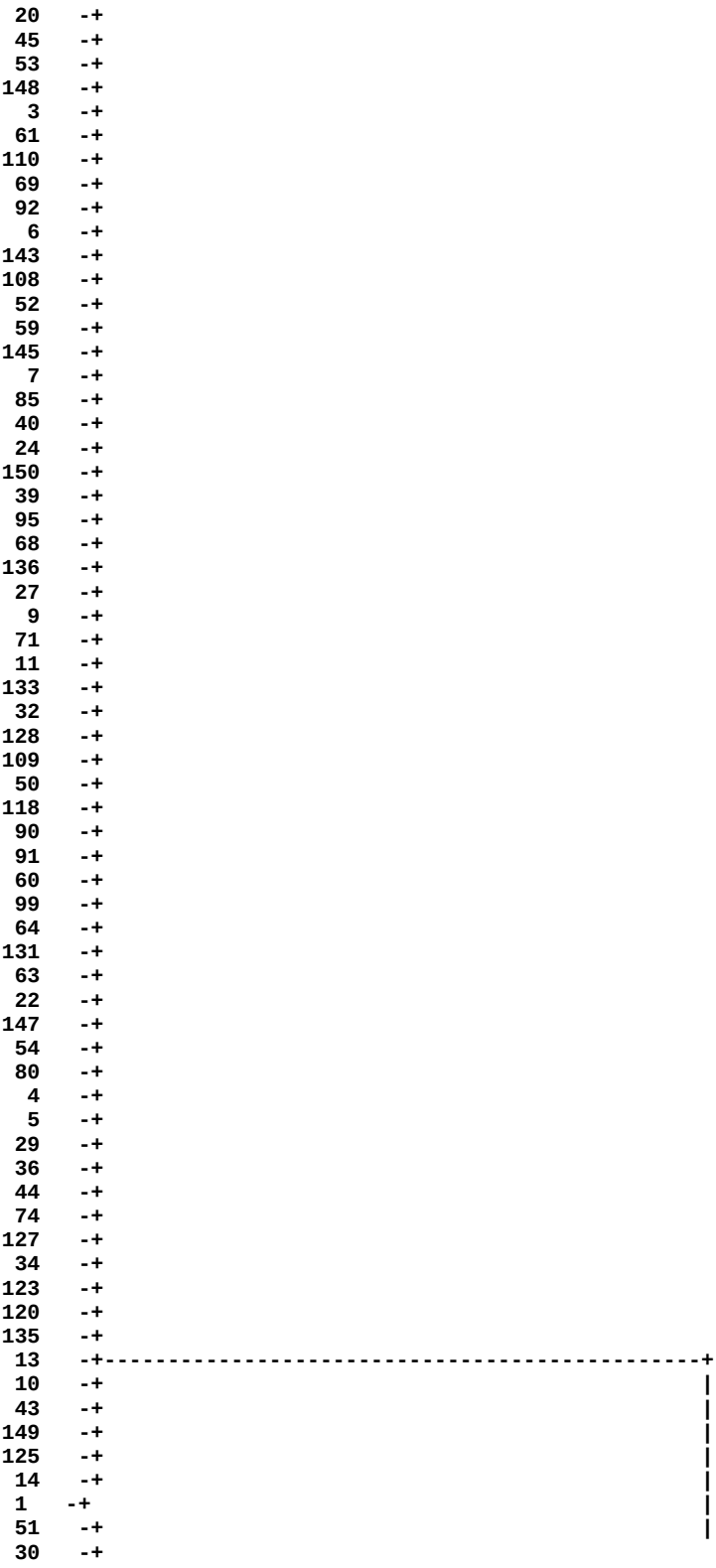
36	18.78468	0.01916	14.94916	0.00007
37	1.15778	0.00067	1.03283	0.00006
38	0.97442	0.00075	0.88498	0.00009
39	1.56940	0.00075	1.41531	0.00005
40	4.48634	0.00460	4.38572	0.00002
41	0.62944	0.00149	0.57361	0.00008
42	0.92664	0.00146	0.84697	0.00014
43	5.13886	0.00037	3.87321	0.00064
44	5.73356	0.00356	6.43259	0.00005
45	1.55509	0.00034	1.46427	0.00008
46	0.66761	0.00156	0.61854	0.00009
47	20.19349	0.00634	15.63289	0.00010
48	4.64247	0.00023	28.74191	0.60236
49	4.32331	0.00001	4.78133	0.00010
50	2.91290	0.01060	2.94997	0.00000
51	2.94884	0.00014	3.65493	0.00061
52	0.43268	0.00840	4.78133	0.00010
53	0.64364	0.01175	0.53953	0.00002
54	3.25776	0.00820	3.45144	0.00002
55	0.91620	0.00091	0.82968	0.00011
56	0.73475	0.00119	0.65354	0.00008
57	0.90847	0.00078	0.81520	0.00008
58	0.91534	0.00090	0.82826	0.00011
59	5.47401	0.00466	6.32913	0.00000
60	2.15278	0.00369	1.92852	0.00001
61	8.73021	0.01853	4.53206	0.00004
62	0.84329	0.00104	0.76139	0.00011
63	7.71011	0.02007	5.98991	0.00004
64	7.75655	0.02189	5.97920	0.00004
65	40.80106	0.18087	1.66924	0.04099
66	0.75931	0.00118	0.69443	0.00009
67	5.04574	0.00007	0.69443	0.00009
68	2.85930	0.00189	3.01226	0.00003
69	0.70678	0.01118	0.55986	0.00003
70	0.90192	0.00081	18.43944	0.43982
71	11.28084	0.00555	12.51310	0.00096
72	8.48097	0.01597	9.11455	0.00033
73	12.68861	0.03947	11.50662	0.00088
74	3.17954	0.00175	3.58178	0.00008
75	0.95754	0.00074	0.87013	0.00009
76	20.52392	0.21812	14.78093	0.00273
77	16.99284	0.04615	9.25301	0.00272
78	1.03482	0.00064	0.94221	0.00008
79	1.12324	0.00057	1.01661	0.00007
80	1.60379	0.00509	1.72286	0.00001
81	0.79536	0.00109	0.71887	0.00010
82	1.34373	0.00072	1.06899	0.00004
83	0.89650	0.00159	0.83064	0.00014
84	0.84678	0.00148	0.68613	0.00006
85	8.54698	0.00804	30.55678	0.49236
86	0.76439	0.00125	0.70062	0.00009
87	0.83277	0.00093	0.75220	0.00009
88	3.74085	0.00025	2.36393	0.00034
89	0.78979	0.00114	0.71625	0.00010
90	8.90657	0.03301	13.92218	0.19178
91	14.07242	0.05758	7.94613	0.00223
92	4.47718	0.01075	3.46591	0.00042
93	1.49146	0.02533	1.11762	0.00015
94	22.77034	0.04725	13.82283	0.00840
95	2.48560	0.00099	1.90963	0.00027
96	1.20455	0.00062	1.04198	0.00005
97	0.95800	0.00081	0.84950	0.00006
98	0.79600	0.00108	0.71435	0.00008
99	0.99718	0.00246	1.20381	0.00001

100	0.94306	0.00081	2.14175	0.03925
101	0.97331	0.00212	0.96129	0.00008
102	0.75715	0.00122	0.68978	0.00009
103	0.74044	0.00112	0.64554	0.00007
104	0.74720	0.00112	0.67578	0.00009
105	0.98061	0.00074	0.89016	0.00009
106	0.84475	0.00098	0.71601	0.00006
107	0.73314	0.00120	0.66344	0.00009
108	1.20358	0.01381	0.89126	0.03212
109	1.67454	0.01031	0.72270	0.00018
110	0.58362	0.01098	0.39011	0.00007
111	1.60151	0.00002	1.75001	0.00010
112	0.75200	0.00113	0.67435	0.00008
113	0.79830	0.00118	0.72732	0.00011
114	0.77502	0.00138	0.72080	0.00012
115	1.12714	0.00064	11.13448	0.15719
116	0.66972	0.00125	0.60482	0.00009
117	3.77522	0.00000	0.60482	0.00009
118	1.43610	0.00660	1.33528	0.00006
119	0.83499	0.00097	0.73024	0.00006
120	5.52114	0.00079	5.39468	0.00019
121	0.79095	0.00112	0.71896	0.00011
122	1.13646	0.00070	1.04117	0.00009
123	4.16246	0.00020	1.04117	0.00009
124	4.17867	0.00001	4.19683	0.00017
125	8.48951	0.00081	3.31529	0.06807
126	1.34577	0.00043	1.24786	0.00007
127	16.89933	0.00724	1.24786	0.00007
128	3.41820	0.01892	3.49507	0.00029
129	0.86458	0.00113	0.78396	0.00012
130	7.28863	0.01249	7.85709	0.00000
131	1.00666	0.00448	1.16450	0.00000
132	0.72922	0.00120	0.66042	0.00010
133	0.08831	0.00481	7.46867	0.04780
134	7.31182	0.00000	7.90126	0.00073
135	8.31259	0.00117	8.79451	0.00117
136	8.21648	0.00472	7.60679	0.00001
137	1.02229	0.00070	0.92698	0.00010
138	0.69234	0.00168	0.65219	0.00009
139	1.32216	0.00057	1.20481	0.00007
140	0.97435	0.00077	5.77285	0.02917
141	1.15122	0.00055	1.04582	0.00008
142	0.88715	0.00084	0.80354	0.00009
143	0.55986	0.01177	0.35165	0.00005
144	0.92967	0.00075	0.83316	0.00008
145	0.84879	0.00123	1.00322	0.00007
146	0.83261	0.00097	0.75397	0.00010
147	7.05740	0.01413	7.27369	0.00016
148	1.63475	0.01934	0.83031	0.03103
149	5.05390	0.00038	5.28602	0.00006
150	5.16450	0.00129	5.14554	0.00003

Figure (3.9) Serbert before (n=150)

```
Rescaled Distance Cluster Combine
    C A S E    0      5       10       15       20      25
  Label  Num  +-------+-------+--------+--------+-------+

         55   -+
         58   -+
         33   -+
         15   -+
        142   -+
         25   -+
        105   -+
        140   -+
        144   -+
         23   -+
         38   -+
         75   -+
         18   -+
         97   -+
        100   -+
         19   -+
         57   -+
         70   -+
         79   -+
        141   -+
         96   -+
        139   -+
         26   -+
         82   -+
         37   -+
        115   -+
         78   -+
        122   -+
        137   -+
         42   -+
        116   -+
         84   -+
        114   -+
         83   -+
         56   -+
        107   -+
        102   -+
        132   -+
         86   -+
         81   -+
        129   -+
         98   -+
         28   -+
        104   -+
        103   -+
        112   -+
         89   -+
         31   -+
        121   -+
          8   -+
         66   -+
        113   -+
         62   -+
         93   -+
         21   -+
         87   -+
        119   -+
        146   -+
        106   -+
         41   -+
         46   -+
        138   -+
        101   -+
        126   -+
          2   -+
```


20
45
53
148
3
61
110
69
92
6
143
108
52
59
145
7
85
40
24
150
39
95
68
136
27
9
71
11
133
32
128
109
50
118
90
91
60
99
64
131
63
22
147
54
80
4
5
29
36
44
74
127
34
123
120
135
13
10
43
149
125
14
1
51
30

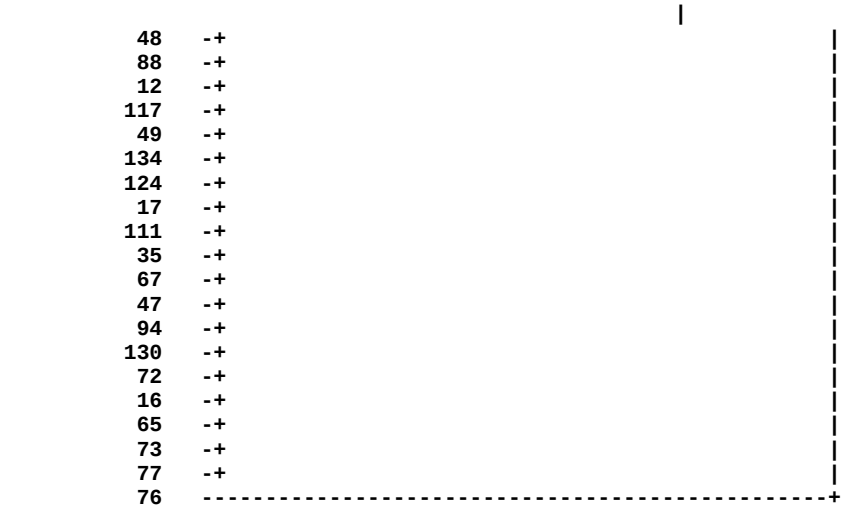
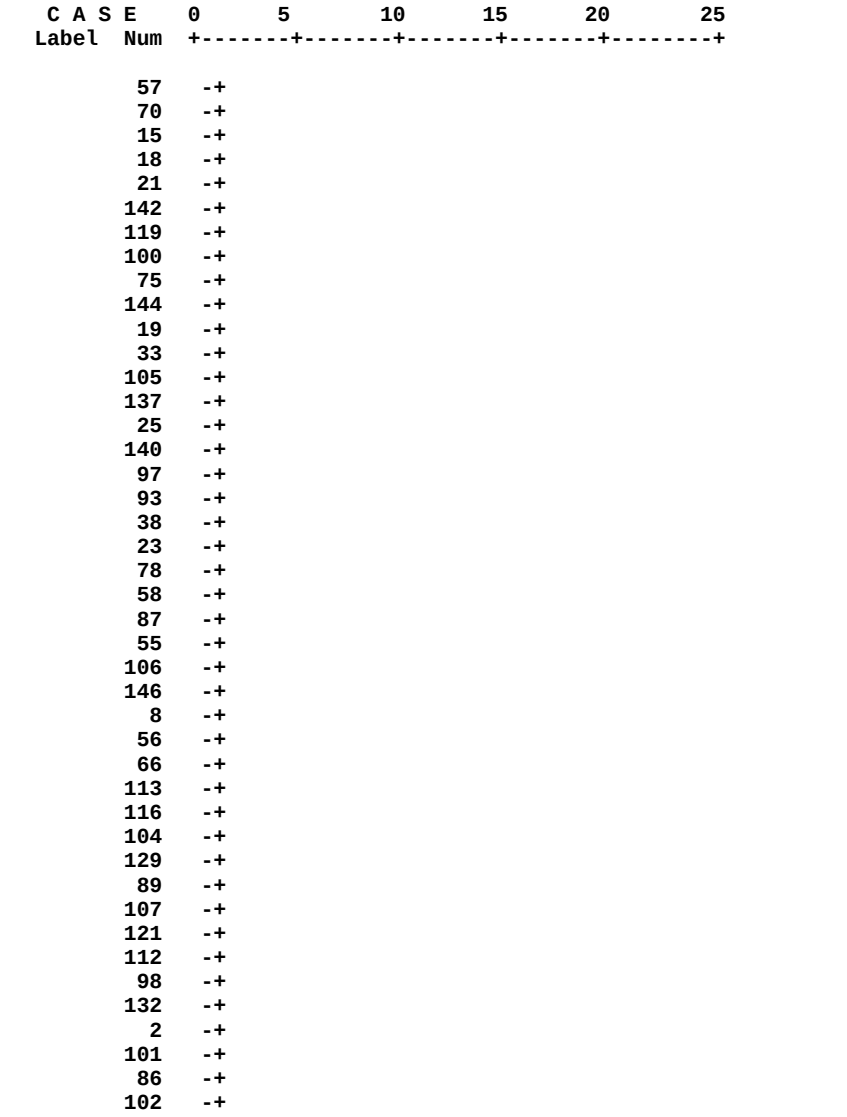
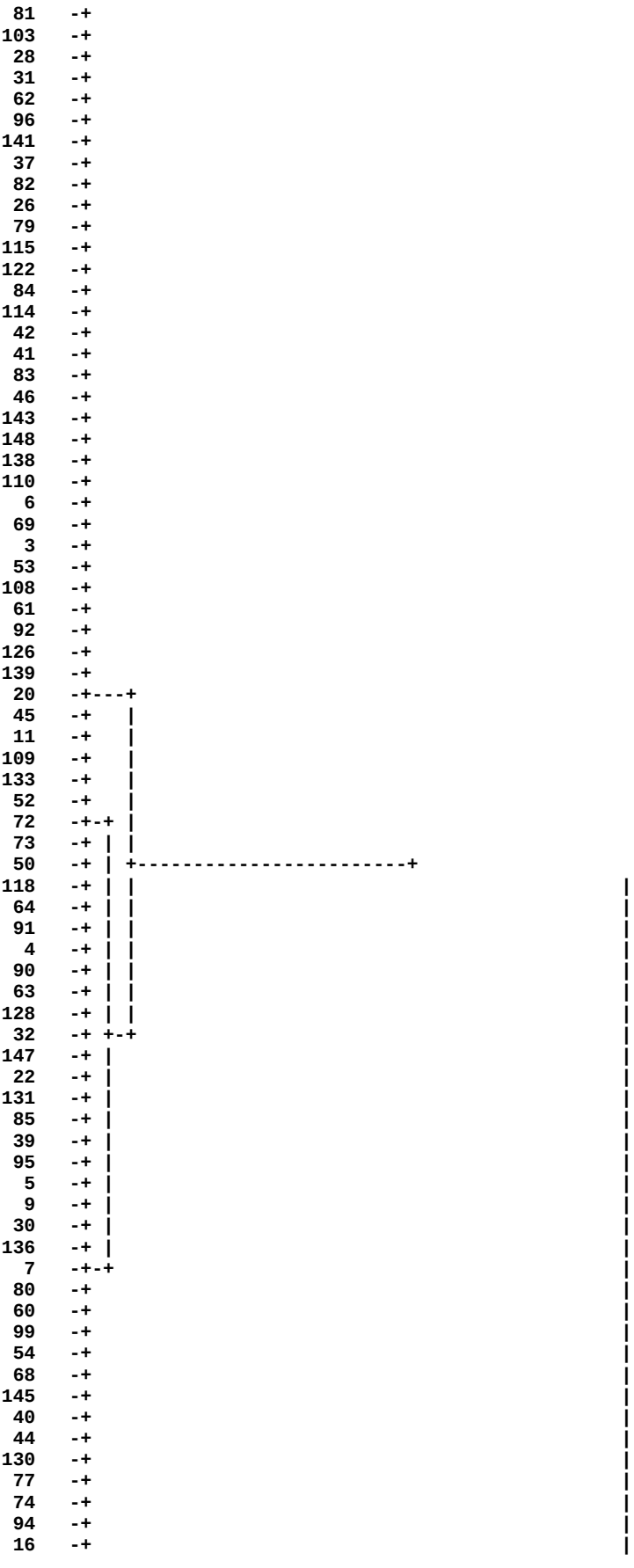


Figure (3.10) Adnan before (n=150)

Rescaled Distance Cluster Combine




81
103
28
31
62
96
141
37
82
26
79
115
122
84
114
42
41
83
46
143
148
138
110
6
69
3
53
108
61
92
126
139
20
45
11
109
133
52
72
73
50
118
64
91
4
90
63
128
32
147
22
131
85
39
95
5
9
30
136
7
80
60
99
54
68
145
40
44
130
77
74
94
16

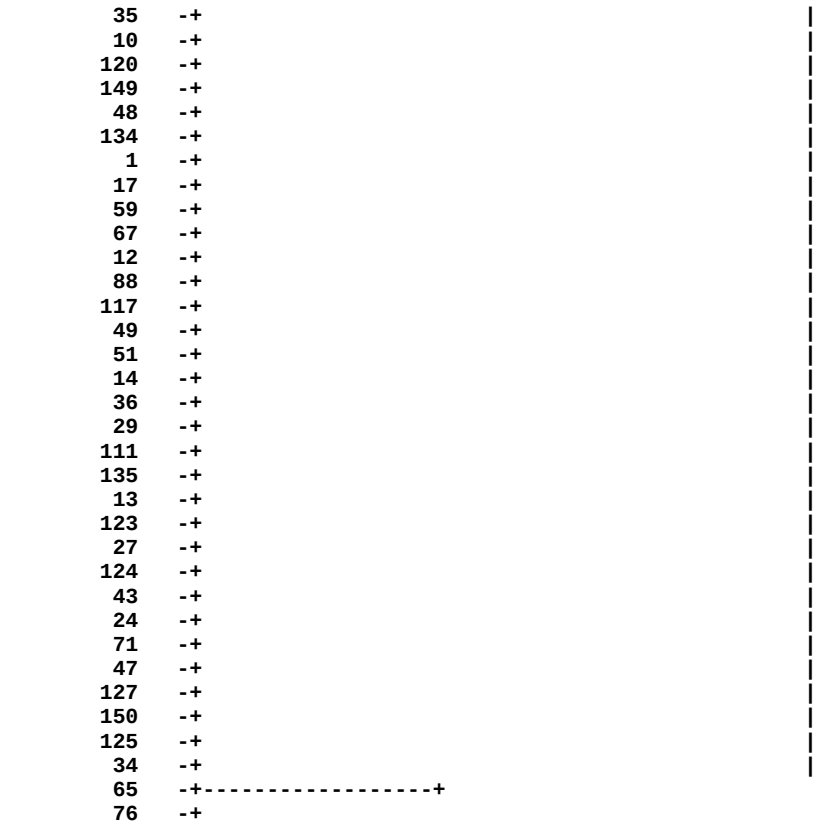
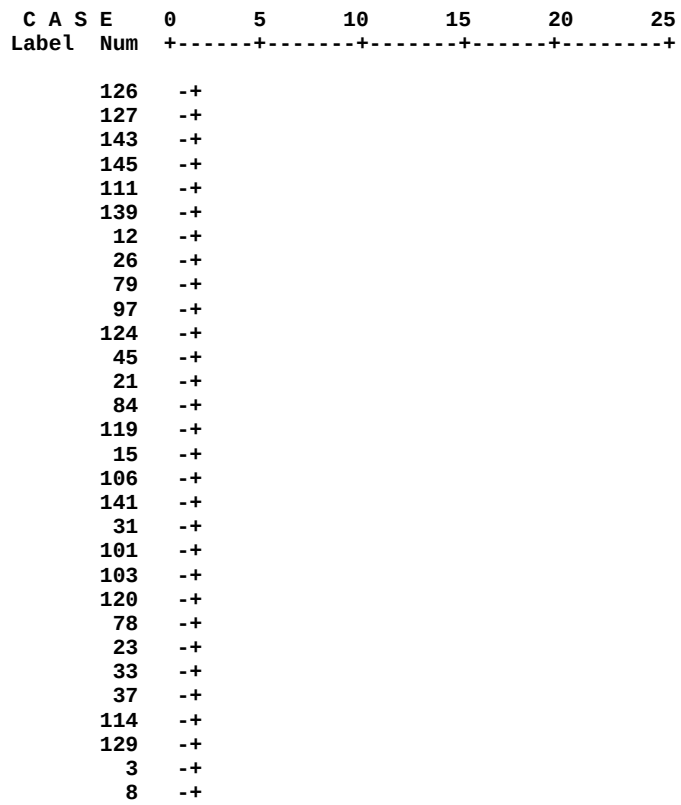


Figure (3.11) Serbert after (n=150)

Rescaled Distance Cluster Combine



```
 81   -+
132   -+
 53   -+
138   -+
  6   -+
 66   -+
 67   -+
 49   -+
 52   -+
 34   -+
 69   -+
146   -+
107   -+
 46   -+
116   -+
117   -+
 86   -+
 41   -+
 74   -+
104   -+
137   -+
 18   -+
102   -+
147   -+
 87   -+
142   -+
 14   -+
144   -+
118   -+
 38   -+
112   -+
 39   -+
 98   -+
 56   -+
 75   -+
 28   -+
 57   -+
 25   -+
105   -+
122   -+
123   -+
 19   -+
 62   -+
121   -+
 55   -+
 89   -+
 58   -+
128   -+
113   -+
  9   -+
  7   -+
 11   -+
 83   -+
149   -+
 16   -+
 42   -+
 68   -+
 92   -+
115   -+
 36   -+
 32   -+
 99   -+
 40   -+
 47   -+
 63   -+
 64   -+
  4   -+
 80   -+
 13   -+
100   -+
110   -+
  2   -+
 96   -+
```

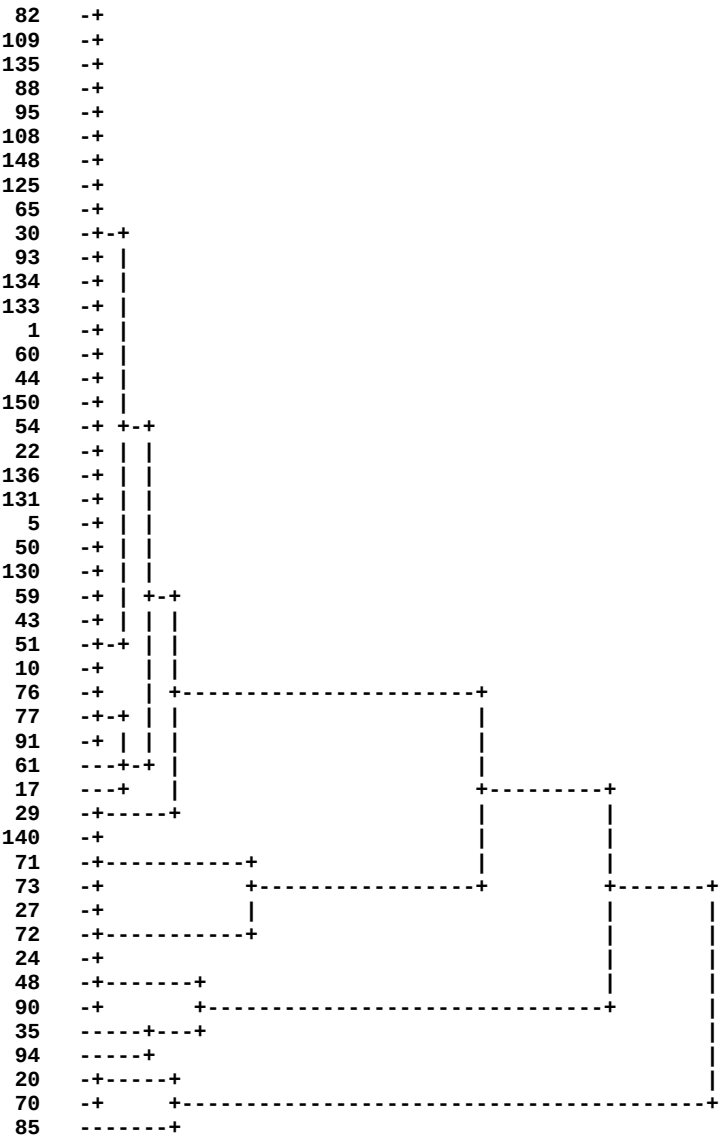
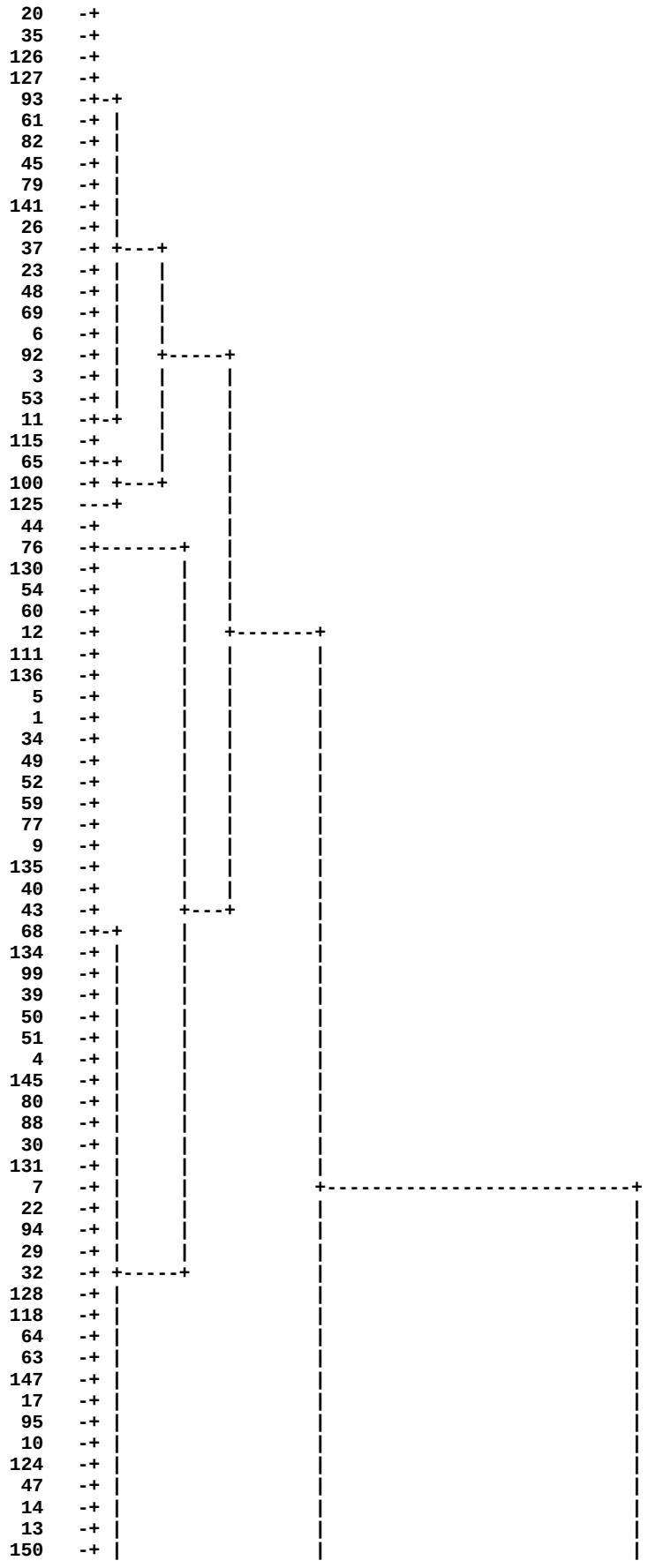

82
109
135
88
95
108
148
125
65
30
93
134
133
1
60
44
150
54
22
136
131
5
50
130
59
43
51
10
76
77
91
61
17
29
140
71
73
27
72
24
48
90
35
94
20
70
85

Figure (3.12) Adnan after (n=150)

Rescaled Distance Cluster Combine

```
 C A S E 0     5      10    15   20      25
Label  Num  +-------+-------+------+-------+-------+

       75   -+
      137   -+
       57   -+
       19   -+
       38   -+
       90   -+
      119   -+
      106   -+
       21   -+
      105   -+
       15   -+
      144   -+
       25   -+
        8   -+
       91   -+
       41   -+
       46   -+
      114   -+
      108   -+
       31   -+
      142   -+
       18   -+
       85   -+
      148   -+
        2   -+
      103   -+
       87   -+
       28   -+
       55   -+
       98   -+
       58   -+
      146   -+
       84   -+
       62   -+
      133   -+
       81   -+
      104   -+
       56   -+
      112   -+
      113   -+
      129   -+
      143   -+
       86   -+
      110   -+
      116   -+
      117   -+
      121   -+
       89   -+
      132   -+
       66   -+
       67   -+
      102   -+
      101   -+
      107   -+
       83   -+
      109   -+
       42   -+
      138   -+
      122   -+
      123   -+
       78   -+
       97   -+
       33   -+
       96   -+
      139   -+
```


20
35
126
127
93
61
82
45
79
141
26
37
23
48
69
6
92
3
53
11
115
65
100
125
44
76
130
54
60
12
111
136
5
1
34
49
52
59
77
9
135
40
43
68
134
99
39
50
51
4
145
80
88
30
131
7
22
94
29
32
128
118
64
63
147
17
95
10
124
47
14
13
150

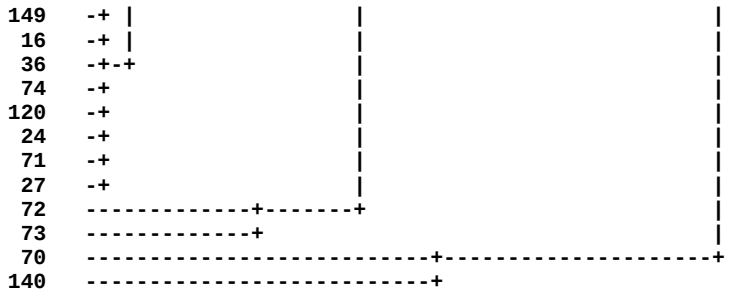


Table (3.9) below shows the summary for the methods , where they are all compared with Mahalanobis distance ($\chi^2_{(145,0.05)}$=67.2), and Cook distance ($F_{(5,45,0.05)}$=2.21) . Both of them can not detect any outliers in the two cases. Outliers are detected by Serbert and Adnan , depending on (ch) value . Differences are found between them for before and after cases .

Mahalanobis and Cook have masking for all the adding outliers ; but Serbert has in the No. (10,65,100,108,115,125,133,148) ; and Adnan has in the No. (20,35,48,65,85,90,100,108,115,125,133,148)

Mahalanobis and Cook didn't have swamping ; but Serbert has in the No. (24,27,71,72,73,94) ; and Adnan has in the No.(13,14,16,17,24,27,36,47,63,64,71,72,73,74,95,118,120,124,128,147,149,150) .

Mahalanobis , Cook and Serbert have (0.3) Std. Error Est. before case , and Serbert has (2.51) after case , which are less than others .

Table (3.9) Summary (n=150)

case	method	ch	outliers	masking	swamping	Std. Error Est.
Before	none	-	-	-	-	0.3
	Mah.	-	none	-	-	0.3
	Cook	-	none	-	-	0.3
	Serbert	12.5	(1,10,12,13,14,16,17, 30,35,43,47,48,49,51,65,67,72,73,76,77,88,94,111,117,124,125, 130,134,149)	-	-	0.3
	Adnan	3.6	(1,10,12,13,14,16,17, 24,27,29,34,35,36,40,43,44,47,48,49,51,54,59,60,65,67,68, 71,74,77,80,88,94,99,111, 117,120,123,124,125,127,130,134,135,145,149,150)	-	-	0.33
After	none	-	-	-	-	3.21
	Mah.	-	-	10,20,35, 48,65,70, 85,90,100,108,115, 125,133, 140,148	-	3.21
	Cook	-	-	10,20,35, 48,65,70, 85,90,100,108,115, 125,133, 140,148	-	3.21
	Serbert	5	(20,24,27,35,48,70,71,72,73,85,90,94,140)	10,65,100,108,115, 125,133, 148	24,27,71, 72,73,94	2.51
	Adnan	4.3	(10,13,14,16,17,24,27,36,47,63,64,70,71,72,73,74,95,118,120,124,128,140,147,149,150)	20,35,48, 65,85,90, 100,108, 115,125, 133,148	13,14,16, 17,24,27, 36,47,63, 64,71,72, 73,74,95, 118,120, 124,128, 147,149, 150	3.05

4. Conclusions :

- **Mahalanobis and Cook Methods could not detect any outlier for all cases, although many of outliers are inserted in the observations. This is because both of them depend on detecting a single outlier, and if the outliers are grouping, they may have detected them .**
- **When (n=25), Adnan's method has the smallest (Std. Error Est.) for before case . It has the same results of Serbert method's after adding the outliers and both of them reduced (Std. Error Est.) .**
- **When (n=50), Serbert method's has the smallest (Std. Error Est.) , masking and swamping .**
- **When (n=150), Serbert method's has the smallest (Std. Error Est.) , masking and swamping .**

References :

- **Adnan,R.; Mohamad,M.N. and Setan,H. (2003) "Multiple Outliers Detection Procedures in Linear Regression" , Matematika ,2003,Jilid 19 ,bil. 1,hlm.(29-45) .**
- **Asikgil, Baris and Erar, Aydin (2009) "Research Into Multiple Outliers in Linear Regression Analysis" . Journal of mathematics and Statistics Volume 38 (2) , (185-198) .**
- **Barnett, V. and Lewis, T. (1994) "Outliers in Statistical Data ", 3rd ed. John Wiley .**
- **Chen, Colin . (2003) . "Robust Regression and Outlier Detection with the ROBUSTREG Procedure" . http://www2.sas.com/proceedings/sugi27/p265-27.pdf**
- **Cook, R.D. (1979) "Influential Observations in Linear Regression" ,JASA. 74,(169-174) .**
- **Draper, N.R. & John, J.A.(1981) "Influential Observations and Outliers in Regression " Technometrics ,23, (21-26) .**
- **Elashoff, J.D. (1972) "A Model for Quadratic Outliers in Linear Regression " ,JASA,67,(478-485) .**
- **Gal, I. B. (2005)" Outlier detection" http://www.eng.tau.ac.il/~bengal/outlier.pdf**
- **Georgiev, T.B. (2008) " An introduction to the method of the robust regression of Rousseeuw " , Astrophys. Invest 10,(93-116) .**
- **Karpinski, A. (2007) " Simple Linear Regression " .**
- **McCane, B. (2009) "Mahalanobis Distance" . http://www.mail-archive.com/morphmet@morphometrics.org/msg01466/mahal.pdf**
- **McLachlan, G.J. (1999) "Mahalanobis Distance" . http://www.ias.ac.in/resonance/June1999/pdf/June1999p20-26.pdf**
- **Mishra, SK. (2008) "A New Method of Robust Linear Regression Analysis : Some Monte Carlo Experiments". http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1155135**
- **Pena, D. & Yohai, V. (1999) " A Fast Procedure for Outlier Diagnostics in Large Regression Problems " JASA ,94,(434-445) .**
- **Rousseeuw, P.J.(1984)" Least Median of Squares Regression " , JASA . 79,(871-880) .**
- **Serbert, D.M. (1998) " A clustering algorithm for identifying multiple outliers in linear regression " , Computational Statistics & Data Analysis , 27,(461-484) .**