CHAPTER SEVEN

UMMC BREAST CANCER DATA: DETECTION OF OUTLIERS AND INFLUENTIAL OBSERVATIONS

In this Chapter, we apply the methods discussed in Chapter 5 and 6 to detect the presence of outliers and influential observations in two cohorts of local breast cancer data.

7.1 The First Cohort of Local Breast Cancer Data

The first cohort of local breast cancer data has about 53% censored observations in 423 patients with the median survival time 55 months and standard deviation 29.1 months as mentioned in Chapter 4. The 'best' fitted Cox PHM is

$$h_i(t) = (0.282x_{stageII} + 0.737x_{stageIII} + 1.502x_{stageIV} + 0.832x_{LN0} + 0.189x_{LN2} + 0.497x_{LN3} + 0.052x_{size1} + 0.639x_{size2} - 0.246x_{GD1} + 0.365x_{GD2})h_0(t)$$

with significant overall proportional hazard assumption 0.0508. That is, the significant prognostic factors for the breast cancer data are *stg*, *LN*, *size* and *GD*. The interaction term of prognostic factors is insignificant.

7.1.1 Outliers Detection

In detecting the outliers, the cut points for deviance r_{Di} and normal deviate r_{Ni} residual are ±1.96 while for log-odds r_{Li} residual is ±3.66. Figure 7.1 gives the r_{Di} versus prognostic index plot for patients in the first cohort. It is found that 16 points are identified as outliers; 13 points exceed 1.96 while 3 points are smaller than -1.96.

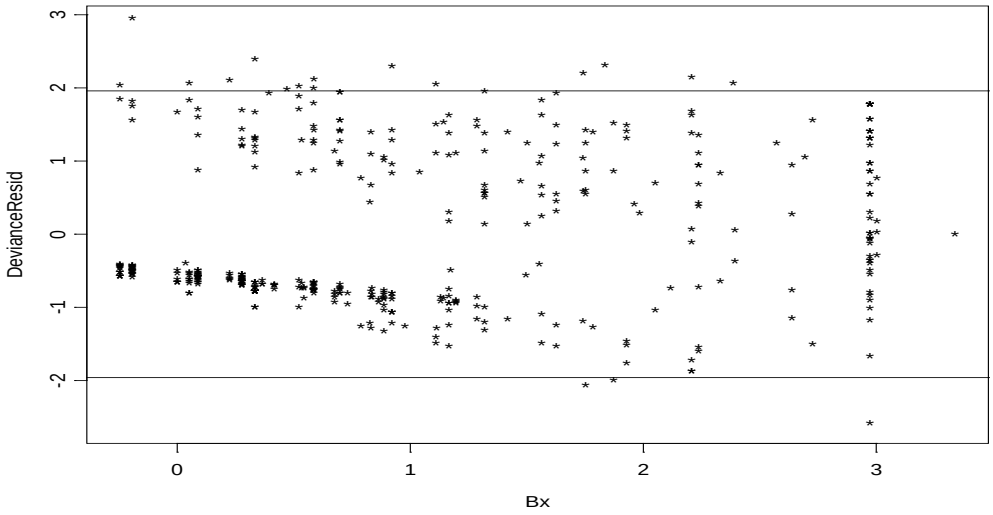

Figure 7.1
r_{Di} vs. prognostic index plot for first cohort

Table 7.1
Profile of outliers in first cohort screening by deviance residuals

patient (*i*)	**time**	**status**	**prognostic factors**				$S(t_i)$	r_{D_i}
			Stg	***LN***	***size***	***GD***		
82	1	1	4	0	1	0	0.9502	2.0125
163	1	1	1	0	2	2	0.9710	2.2601
367	1	1	3	0	2	0	0.9582	2.0953
20	2	1	4	2	1	0	0.9622	2.1427
400	2	1	1	1	1	1	0.9945	2.8979
3	5	1	2	1	2	0	0.9700	2.2456
27	8	1	2	2	2	0	0.9489	2.0001
33	8	1	2	1	1	0	0.9761	2.3441
326	11	1	1	0	0	1	0.9555	2.0663
28	13	1	2	2	1	0	0.9464	1.9764
275	14	1	2	2	0	1	0.9541	2.0513
207	18	1	1	1	1	0	0.9498	2.0088
251	26	1	1	1	0	1	0.9481	1.9930
274	68	0	4	0	2	0	0.0308	-2.6381
6	103	0	3	3	2	0	0.1218	-2.0521
19	106	0	2	0	2	0	0.1050	-2.1231

Table 7.1 gives the profile of the 16 patients detected as outliers using r_{Di}. The information indicates 13 patients die due to breast cancer while the rest are still alive by

the time the study ends. Note that some of them do not follow the definition of outliers as given by Nardi and Schemper (1999). For example, patients 251 and 207 have survival time not far from the median survival times (39.5 months) with favorable prognostic factors, but still detected as outlier. In fact, their residuals are also quite small.

Figure 7.2 gives the r_{Ni} versus prognostic index plot for patients in the first cohort. Note that two points exceed cut points 1. 96 and another two points are smaller than cut points -1. 96. They are patients 33, 400, 221 and 274, respectively. Table 7.2 gives the profile of patients detected as outliers using r_{N_i}. Note that two patients with survival times 2 and 8 months are unexpectedly die too early with favorable early stage of prognostic factors. While another two patients, with survival times 68 and 77 months, survive longer with unfavorable advanced stage of cancer. Therefore, we conclude that the outliers in this data set are patients who are mostly poorly predicted by the model. An investigation of unknown prognostic factors that have important effect on survival of patients can be initiated by looking at their medical record.

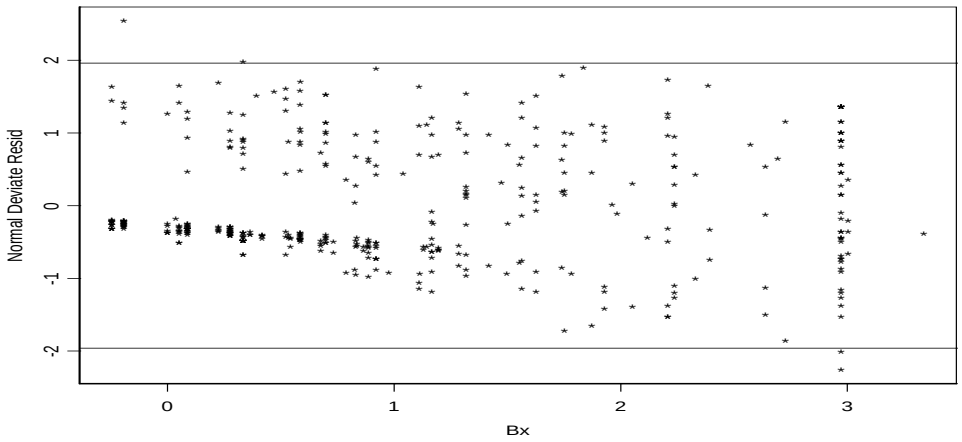

Figure 7.2
r_{N_i} vs. prognostic index plot for the first cohort

Table 7.2
Profile of outliers in first cohort screening by normal deviate residual

patient (*i*)	time	status	prognostic factors				$S(t_i)$	r_{N_i}
			stg	*LN*	*size*	*GD*		
400	2	1	1	1	1	1	0.9945	2.5403
33	8	1	2	1	1	0	0.9761	1.9798
274	68	0	4	0	2	0	0.0308	-2.2577
221	77	1	4	0	2	0	0.0221	-2.0119

Figure 7.3 gives the r_{L_i} versus prognostic index plot for the first cohort. The resulting r_{L_i} are larger in magnitude compared to that of r_{N_i}. There are four points exceeding the cut points. After screening the profile of outliers as listed in Table 7.3, we note that outliers detected by r_{L_i} approach are similar to that detected by r_{N_i} approach.

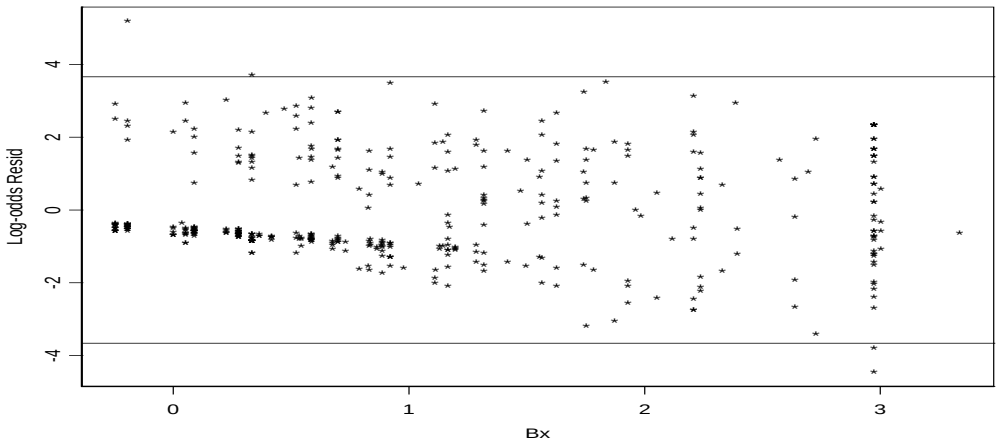

Figure 7.3
r_{L_i} vs. prognostic index plot for the first cohort

7.1.2 Influential Observation Detection

Further, we extend the study to identify influential observations that affect the inferences based on the 'best' fitted Cox PHM.

Delete-case method

In this section, the influential observations are detected based on delete-case method. The resulting plot is given in Figure 7.4. Any points far from the rest are considered to be influential patients and the points will be recorded for the profile screening purposes.

Table 7.3
Profile of outliers in first cohort screening by log-odds residual

patient (*i*)	time	status	prognostic factors				$S(t_i)$	r_{L_i}
			stg	*LN*	*size*	*GD*		
400	2	1	1	1	1	1	0.9945	5.1906
33	8	1	2	1	1	0	0.9761	3.7112
274	68	0	4	0	2	0	0.0308	-4.4643
221	77	1	4	0	2	0	0.0221	-3.7891

Table 7.4 gives a profile of nine patients identified as influential patients using the delete-case method. Three patients die due to the breast cancer while the rest are still alive until the end of study. None of them are identified as outliers in the previous section. Omitting these patients from the data set may change the parameter estimate and hazard for each prognostic factor as shown in Table 7.5. The parameter and hazard values are decreasing, while the global test on PHA also increases from 0.0508 to 0.3071.

Table 7.4
Profile of influential observation using delete case method

patient (*i*)	time	status	prognostic factors			
			stg	*LN*	*size*	*GD*
163	1	1	1	0	2	2
22	19	1	4	1	0	0
13	64	0	1	0	1	0
211	83	1	4	0	2	1
113	86	0	3	2	2	2
61	104	0	1	2	2	0
21	99	0	2	0	0	0
6	103	0	3	3	2	0
19	106	0	2	0	2	0

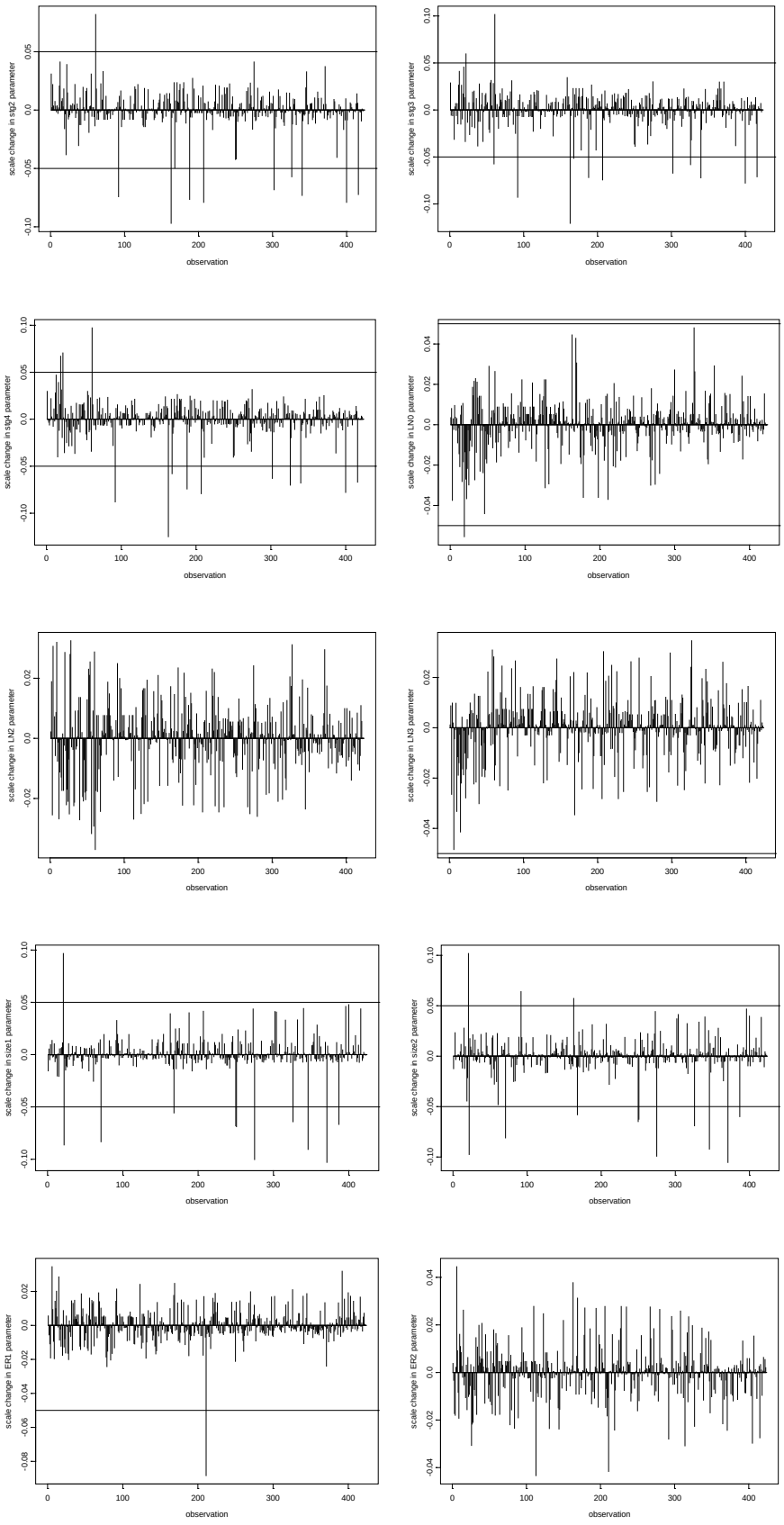

Figure 7.4
dfbetas residuals plot for first cohort

Table 7.5
Cox PHM result by omitting the influential observation using delete case method

Patient exclude from the fitted model	Variables	Parameter estimate	Hazard	C.I of hazard	Global test (PHA)
None	stg2	0.2822	1.326	(0.702, 2.51)	0.0508
	stg3	0.7367	2.089	(0.996, 4.38)	
	stg4	1.5022	4.492	(2.159, 9.35)	
	LN0	0.8316	2.297	(1.419, 3.72)	
	LN2	0.1885	1.207	(0.760, 1.92)	
	LN3	0.4971	1.644	(1.020, 2.65)	
	size1	0.0518	1.053	(0.533, 2.08)	
	size2	0.6389	1.894	(0.923, 3.89)	
	GD1	-0.2455	0.782	(0.534, 1.15)	
	GD2	0.3647	1.440	(0.942, 2.20)	
Influential patients (listed in Table 8.4)	stg2	0.2497	1.284	(0.647, 2.55)	0.3071
	stg3	0.6571	1.929	(0.870, 4.28)	
	stg4	1.3449	3.838	(1.729, 8.52)	
	LN0	0.9942	2.703	(1.645, 4.44)	
	LN2	0.2598	1.297	(0.810, 2.08)	
	LN3	0.6012	1.824	(1.121, 2.97)	
	size1	0.0341	1.035	(0.498, 2.15)	
	size2	0.7404	2.097	(0.964, 4.56)	
	GD1	-0.2133	0.808	(0.536, 1.22)	
	GD2	0.3578	1.430	(0.920, 2.22)	

Cox PHM FS method

We consider the Cox PHM FS method proposed in Chapter 6. Due to large percentage of censoring observation exist in the data set, we set the percentage of initial subset for techniques 1, 2 and 3 to be 90% of the data size and the values of γ for technique 4 to be 3, 6 and 9. For techniques 5, 6 and 7, we set the lower cut point equals -1 while the upper cut point equals 3 for r_{Di} and r_{Ni}, while -1 and 6 respectively for r_{Li}.

Table 7.6 gives the proportion of similar and non-similar patients selected by seven different techniques of the initial subset. We observe that the proportions for all techniques are more than 0.5 as shown in Table 7.6(a). The proportion of similarity for techniques 2 and 3 are almost equal. The initial subset for the Cox PHM FS methods

Table 7.6
Proportion of patients selected by seven different initial subset techniques a) similar proportion b) non-similar proportion

		Techniques								
	(a)	**1**	**2**	**3**	**4**			**5**	**6**	**7**
					γ=3	**γ=6**	**γ= 9**			
Techniques	**1**	1	0.97	0.97	0.63	0.90	0.91	0.84	0.90	0.78
	2		1	0.995	0.65	0.93	0.93	0.85	0.91	0.79
	3			1	0.95	0.91	0.93	0.85	0.91	0.79
	γ=3				1	0.68	0.67	0.65	0.65	0.61
4	**γ=6**					1	0.97	0.84	0.88	0.78
	γ=9						1	0.84	0.89	0.78
	5							1	0.91	0.93
	6								1	0.87
	7									1
	(b)									
Techniques	**1**	0	0.02	0.02	0.02	0.03	0.03	0.04	0.04	0.04
	2	0.02	0	0.002	0.005	0.03	0.03	0.04	0.04	0.04
	3	0.02	0.002	0	0.004	0.03	0.03	0.04	0.04	0.04
	γ =3	0.35	0.34	0.50	0	0.32	0.33	0.31	0.34	0.30
4	**γ =6**	0.07	0.05	0.06	0	0	0.03	0.06	0.07	0.06
	γ =9	0.06	0.04	0.04	0	0	0	0.05	0.05	0.05
	5	0.12	0.11	0.11	0.04	0.10	0.11	0	0.08	0
	6	0.06	0.05	0.05	0.01	0.05	0.06	0.02	0	0
	7	0.18	0.17	0.17	0.09	0.16	0.17	0.07	0.13	0

on this data gives the same pattern of initial subset that has been discussed in Chapter 6. In contrast, the proportion of non-similar patients is shown in Table 7.6(b). The proportions of non-similar are less than 0.34. We note that initial subset based on technique 4 with $\gamma = 3$ give the smallest similar proportion and biggest non-similar proportion compared to the others. Hence, the Cox PHM FS methods FS4, FS5 and FS6 based on technique 4 with $\gamma = 3$ are not considered for further analysis.

Figure 7.5 shows the progression plot for the FS1 procedure. It is found that the parameter estimates are constantly changing for one point to another in each prognostic factor. Thus, to identify which patients may affect the parameter estimates, we use the *IM* plots as given in Figure 7.6. It can be seen that the influential patients exist at step $m+i$ equal to 14, 19, 21, 26, 27, 29, 30, 32 and 33. The $m+i$ steps above correspond to patients number 113, 371, 275, 6, 19, 168, 302, 33, 188 and 416 respectively. We also obtain the progression and *IM* plots for the other Cox PHM FS methods and are given in Appendix B.

Table 7.7 gives the full list of influential patients identified by 9 different types of Cox PHM FS method. It is found that the number of influential patients detected by Cox PHM FS methods is less than 5% of the sample size. Four main results are observed from this result. Firstly, we find that patients 19, 168, 340 and 387 are identified as influential patients by every Cox PHM FS methods. It is interesting to note that none of these patients are identified as outliers. While, patients 33, 221 and 274 are selected by FS9 are identified as influential observations and also as outliers. Secondly, FS7 and FS9 which use r_{Di} and r_{Li} respectively in step 2 identify large number of influential patients compared to (FS1, FS3) and (FS4, FS6) procedures respectively, even though similar type of residual are employed. Meanwhile, we observe that FS8 gives the smallest number of selected influential patients compared to the other types of the Cox PHM FS methods. Thirdly, another eight patients are selected by at least five Cox PHM FS procedures. Lastly, we observed that FS6 gives similar influential patients as the FS2. Similar influential observations are identified by FS5 and FS3 method.

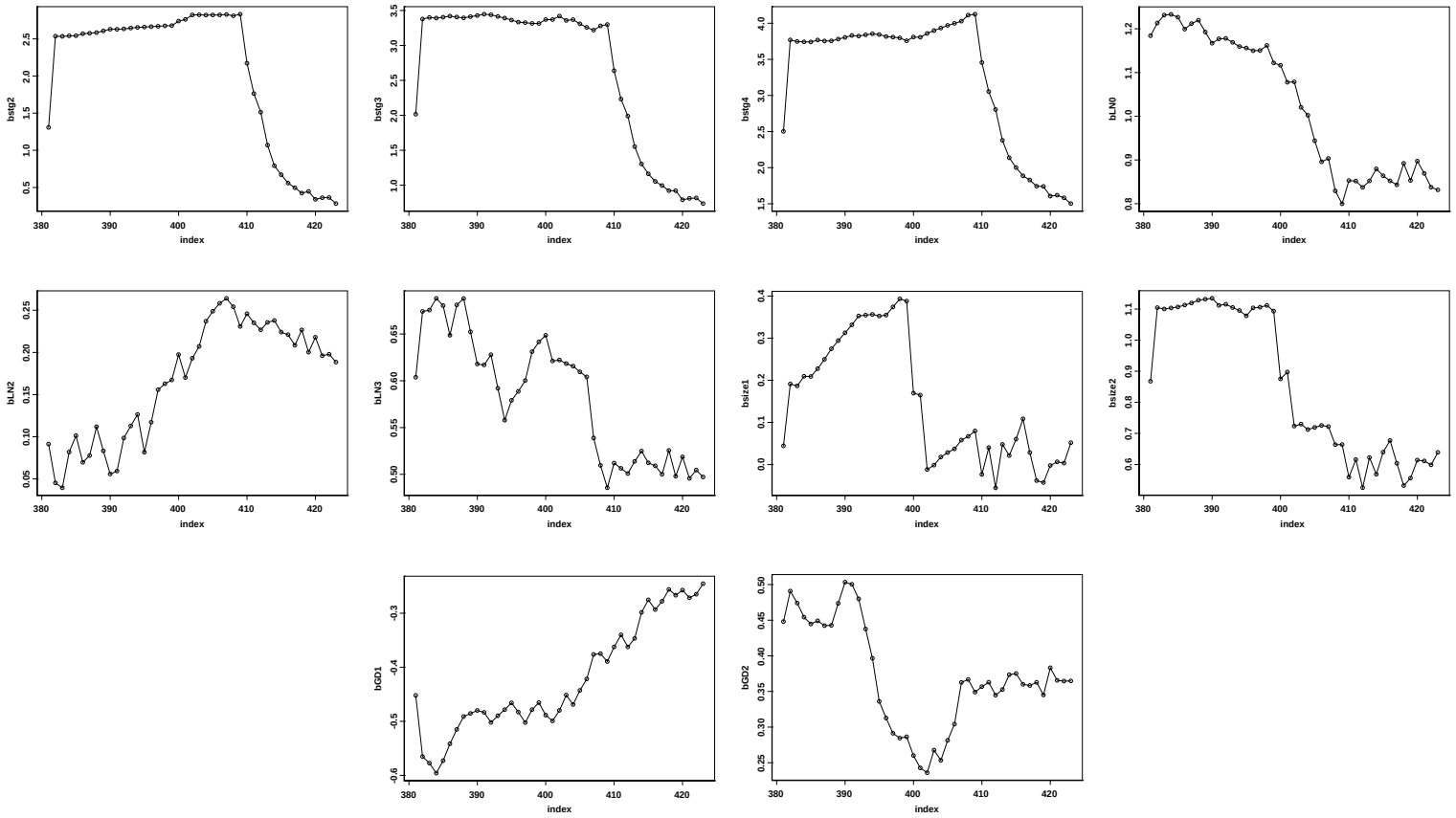

Figure 7.5
The progression plots for FS1 procedures on the first cohort of local breast cancer data

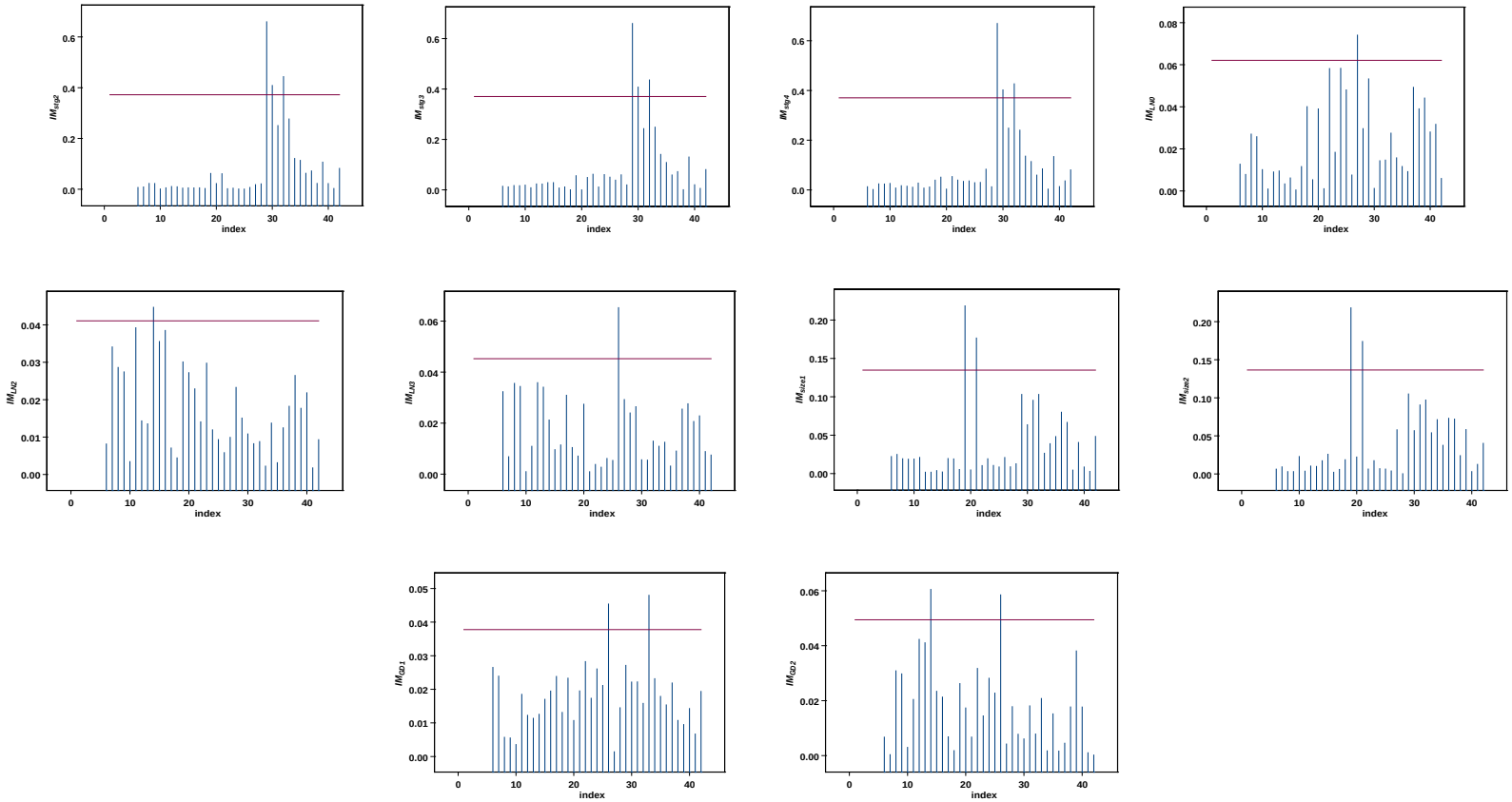

Figure 7.6
The *IM* plots for FS1 procedures on the first local breast cancer data

Table 7.7
Influential observation selection by Cox PHM FS methods

Patient number	**FS1**	**FS2**	**FS3**	**FS4** γ =6	**FS5** γ =6	**FS6** γ =6	**FS7**	**FS8**	**FS9**
19	/	/	/	/	/	/	/	/	/
168	/	/	/	/	/	/	/	/	/
340	/	/	/	/	/	/	/	/	/
387	/	/	/	/	/	/	/	/	/
33	/	/	/	/	/	/	/		/
188	/	/	/	/	/	/	/		/
275	/	/	/	/	/	/	/	/	
302	/	/	/	/	/	/	/		/
371	/	/	/	/	/	/	/		
416	/	/	/	/	/	/			/
6	/		/	/	/		/		/
211		/	/		/	/		/	/
3	/			/			/		
113	/						/		/
274							/	/	/
total	13	11	12	12	12	11	13	7	12

Note: observations 12, 22 ,36, 71, 101,169, 221, 269, 345 detected by FS9 only
observations 27, 292, 314 are detected by FS7 only
observation 28 is detected byFS4 only

Table 7.8 gives the estimate of the parameter, hazard and the p-value of the global test based on PHA for Cox PHM before and after omitting the influential patients selected by the Cox PHM FS method. It is found that the influential patients in the first cohort do affect the statistics considered. However, the changes on the estimate and hazard ratio of each prognostic factor are not too large except *stg* and *size* factors. The changes are greater than 0.2. The p-value of the global test results also shows some changes when influential patients are removed. We note that by omitting the influential patients detected by FS9 give the largest changes to *stg* and *size* factors with the *p*-value of the global test on PHA changes from 0.0508 to 0.2183.

Table 7.8
Cox Model result when omit influential observation using Cox PHM FS methods

	Prognostic Factors	Patients omit based on patients in							
		None	FS1	FS2 & FS6	FS3 & FS5	FS4	FS7	FS8	FS9
β -parameter	stg2	0.2822	0.664	0.705	0.698	0.662	0.528	0.451	0.553
	stg3	0.7367	1.152	1.076	1.114	1.128	1.100	0.820	1.002
	stg4	1.5022	1.817	1.876	1.847	1.812	1.783	1.661	1.818
	LN0	0.8316	0.934	0.959	0.946	0.934	0.961	0.947	1.029
	LN2	0.1885	0.241	0.169	0.162	0.172	0.257	0.151	0.269
	LN3	0.4971	0.609	0.519	0.577	0.618	0.547	0.479	0.634
	size1	0.0518	0.233	0.263	0.245	0.220	0.295	0.281	0.552
	size2	0.6389	0.921	0.969	0.969	0.915	0.979	0.962	1.355
	GD1	-0.2455	-0.330	-0.172	-0.221	-0.302	-0.315	-0.196	-0.356
	GD2	0.3647	0.422	0.466	0.408	0.393	0.532	0.419	0.332
Hazard	stg2	1.326	1.942	2.024	2.010	1.939	1.70	1.570	1.739
	stg3	2.089	3.164	2.934	3.045	3.090	3.00	2.269	2.724
	stg4	4.492	6.152	6.526	6.343	6.124	5.95	5.262	6.161
	LN0	2.297	2.544	2.608	2.576	2.546	2.61	2.577	2.799
	LN2	1.207	1.272	1.185	1.176	1.187	1.29	1.163	1.308
	LN3	1.644	1.838	1.680	1.781	1.855	1.73	1.614	1.885
	size1	1.053	1.262	1.301	1.277	1.246	1.34	1.325	1.737
	size2	1.894	2.512	2.634	2.635	2.497	2.66	2.616	3.877
	GD1	0.782	0.719	0.842	0.802	0.739	0.73	0.822	0.701
	GD2	1.440	1.525	1.593	1.504	1.481	1.70	1.520	1.394
Overall PHA *p*-value	Global Test	0.0508	0.0696	0.0688	0.0827	0.06283	0.1105	0.1183	0.2183

Table 7.9 gives the full profile list of influential patients in the first cohort of local breast cancer data. Let us focus on four patient numbers 19, 168, 340 and 387 that are identified as influential patients by every types of Cox PHM FS method. Patient 19 who has a long survival times is predicted to have quite short survival times. Meanwhile, for the other three patients; 168, 340 and 387 who have short survival times are predicted to survive longer. Here, further study can be carried out to identify other factors that might affect the survival of their patients.

Table 7.9
Profile of influential patients in first cohort of local breast cancer data

patient (*i*)	**time**	**status**	**prognostic factors**				$S(t_i)$
			stg	***LN***	***size***	***GD***	
3	5	1	2	1	2	0	0.9700
27	8	1	2	2	2	0	0.9489
33	8	1	2	1	1	0	0.9761
28	13	1	2	2	1	0	0.9464
275	14	1	2	2	0	1	0.9541
371	14	1	2	2	0	0	0.9416
346	17	1	2	3	0	2	0.8661
22	19	1	4	1	0	0	0.7971
188	29	1	1	1	1	0	0.9211
168	34	1	1	0	0	1	0.8435
340	36	1	1	1	1	1	0.9205
387	36	1	1	1	0	1	0.9243
416	39	1	1	1	1	1	0.9101
101	41	1	4	0	2	0	0.1016
71	42	1	2	3	0	1	0.8077
36	54	1	4	3	2	0	0.1277
302	60	1	1	1	1	1	0.8739
292	63	0	3	2	2	2	0.3193
268	65	0	3	0	2	0	0.2075
274	68	0	4	0	2	0	0.0308
314	71	0	3	2	2	2	0.2878
169	77	1	4	3	2	0	0.0654
221	77	1	4	0	2	0	0.0221
211	83	1	4	0	2	1	0.0317
113	86	0	3	2	2	2	0.1920
6	103	0	3	3	2	0	0.1218
12	104	1	2	1	1	1	0.6804
19	106	0	2	0	2	0	0.1050

7.2 The Second Cohort of Local Breast Cancer Data

The second cohort of local breast cancer data cohort has about 76% of censored patients out of 965 patients as discussed in Chapter 4. Moreover, the median of survival times is 58 months and the standard deviation is 18.3 months. The 'best' fitted Cox PHM for this data is

$$h_i(t) = (0.592x_{stageII} + 1.625x_{stageIII} + 2.614x_{stageIV} + 1.679x_{LN0} + 0.744x_{LN2} + 1.183x_{\mathrm{LN3}} - 0.146x_{ER1} + 0.553x_{ER2} + 0.101x_{race1} + 0.382x_{race2})h_0(t)$$

and the interaction term of prognostic factors is insignificant.

7.2.1 Outliers Detection

In this section we use the same cut points considered in the first cohort. The cut points for r_{Di} and r_{Ni} are ±1.96 while for r_{Li} are ±3.66. Figure 7.7 gives the deviance residuals vs. prognostic index plot for patients in the second cohort. There are 41 points that exceed the cut points; 28 points with large positive r_{Di} and 13 points with small negative r_{Di}.

Table 7.10 gives the profiles of the 41 patients detected as outliers. There are 28 patients died due to breast cancer while the rest are still alive by the end of study. Note that some of them do not follow the definition of outlier. The patients selected are well predicted; for example breast cancer patient number 111 with empirical survival time 58 months. The survival time exactly equals the median survival time. This patient has prognostic factors that well explain her survival time.

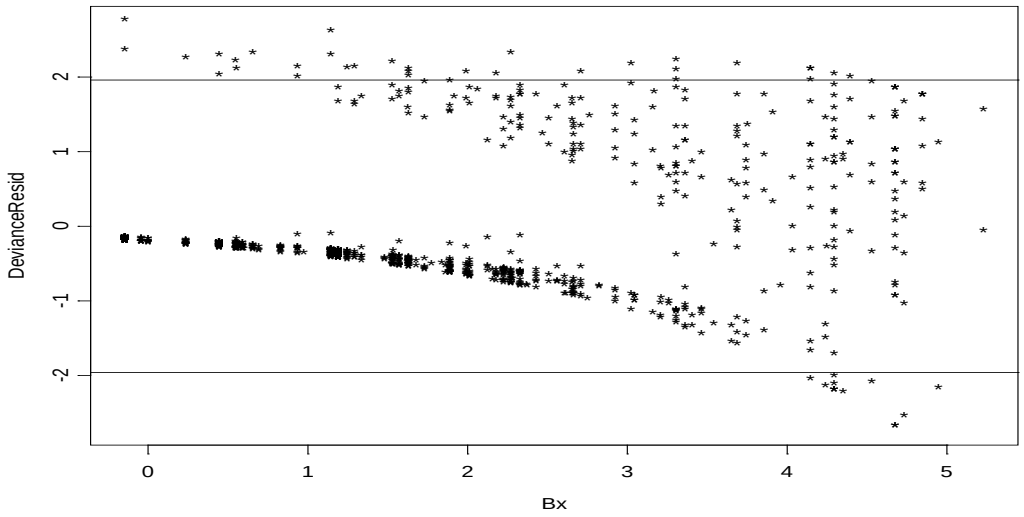

Figure 7.7
r_{Di} vs. prognostic index plot for second cohort

Figure 7.8 gives the normal deviate residuals vs. prognostic index plot for patients in the second cohort. Patients number 146, 405, 453, 876 and 905 are identified as outlier. Table 7.11 gives the profiles of these patients. Note that patients 146 and 453 have r_{Ni} exceeding 1.96. These patients are not expected to die early due to favorable early stage of prognosis. In additional, patients 405, 876 and 905 have r_{Ni} smaller than -1. 96. These patients are unexpectedly surviving longer with unfavorable late stage of prognosis. Therefore, we conclude that the outliers in this data set are patients who are mostly poorly predicted by the model. Further investigation should be carried out to determine other factors that contribute to the survival of these patients.

Table 7.10
Profile of outliers in second cohort screening by deviance residuals

patient (*i*)	**time**	**status**	**prognostic factors**				$S(t_i)$	r_{D_i}
			stg	***LN***	***ER***	***race***		
770	1	1	4	0	0	0	0.9489	2.0002
830	1	1	4	0	1	0	0.9557	2.0683
919	1	1	4	0	1	0	0.9557	2.0683
867	2	1	3	0	0	2	0.9618	2.1375
848	3	1	3	0	0	0	0.9651	2.1775
661	5	1	3	2	2	2	0.9541	2.0523
720	6	1	3	2	2	1	0.9670	2.1276
949	8	1	2	0	0	0	0.9724	2.2819
424	9	1	2	3	2	2	0.9508	2.0189
453	11	1	2	0	2	0	0.9862	2.5658
608	14	1	3	0	2	0	0.9486	1.9975
293	16	1	2	2	2	1	0.9517	2.0271
561	18	1	2	1	2	2	0.9632	2.1550
640	19	1	2	3	1	0	0.9551	2.0625
371	20	1	2	1	2	0	0.9700	2.2460
459	20	1	2	3	1	0	0.9518	2.0285
146	21	1	0	1	1	0	0.9911	2.7286
665	22	1	2	3	1	0	0.9451	1.9650
499	24	1	2	2	1	1	0.9576	2.0893
509	25	1	2	1	2	1	0.9562	2.0736
274	28	1	1	1	2	1	0.9725	2.2844
285	33	1	1	1	2	2	0.9573	2.0856
778	34	0	4	0	2	1	0.0876	-2.2067
815	38	1	2	1	1	0	0.9703	2.2505
402	42	1	2	1	1	1	0.9647	2.1731
548	53	0	4	0	1	2	0.1027	-2.1337
206	57	1	1	1	1	2	0.9676	2.2118
111	58	1	1	1	2	0	0.9553	2.0640
832	64	0	4	0	0	0	0.1240	-2.0434
216	67	1	1	1	1	0	0.9749	2.3234
876	67	0	4	3	2	2	0.0356	-2.5831
939	75	0	4	0	0	0	0.0970	-2.1600
771	88	0	4	3	2	0	0.0765	-2.2677
405	89	0	4	0	0	2	0.0251	-2.7148
520	89	1	2	1	1	0	0.9478	1.9896
908	90	0	4	0	0	0	0.0809	-2.2424
811	91	0	4	0	1	0	0.1138	-2.0848
905	93	0	4	0	0	2	0.0251	-2.7148
847	94	0	4	0	0	0	0.0809	-2.2424
872	94	0	4	0	0	0	0.0809	-2.2424

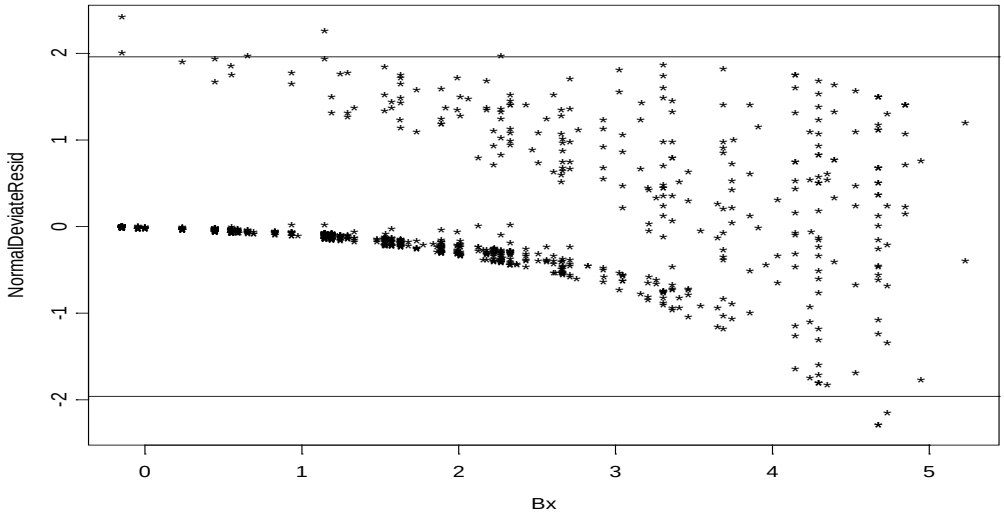

Figure 7.8
r_{Ni} vs. prognostic index plot for second cohort

Table 7.11
Profile of outliers in second cohort screening by normal deviate residuals

patient (*i*)	**time**	**Statu s**	**prognostic factors**				$S(t_i)$	r_{N_i}
			stg	***LN***	***ER***	***race***		
453	11	1	2	0	2	0	0.9862	2.2035
146	21	1	0	1	1	0	0.9911	2.3684
876	67	0	4	3	2	2	0.0356	-2.2013
405	89	0	4	0	0	2	0.0251	-2.3364
905	93	0	4	0	0	2	0.0251	-2.3364

Figure 7.9 gives the log-odds residuals vs. prognostic index plot for patients in the second cohort. The detection based on r_{Li} gives a similar result as the r_{Ni} case. Three of them exceed the cut point 3.66 while the rest are less than -3.66. They are patients' number 146, 405, 453, 876 and 905. Table 7.12 gives the profiles of these patients with the estimated r_{Li} given in the last column. It is clear that their survival probabilities are not well predicted with respect to their prognosis.

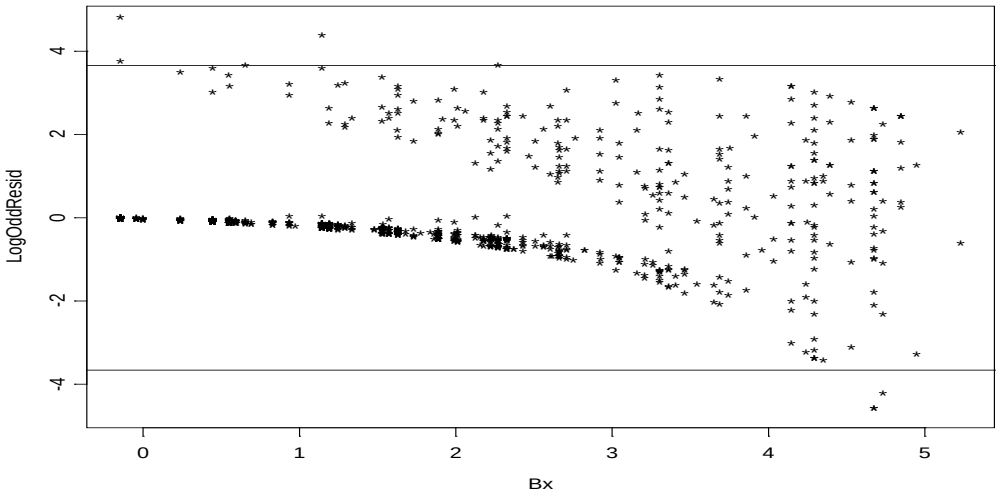

Figure 7.9
r_{Li} vs. prognostic index plot for second cohort

Table 7.12
Profile of outliers in second cohort screening by log-odds residuals

patient (*i*)	time	Status	prognostic factors				$S(t_i)$	r_{L_i}
			stg	*LN*	*ER*	*race*		
453	11	1	2	0	2	0	0.9862	4.2707
146	21	1	0	1	1	0	0.9911	4.7092
876	67	0	4	3	2	2	0.0356	-4.3182
405	89	0	4	0	0	2	0.0251	-4.6724
905	93	0	4	0	0	2	0.0251	-4.6724

7.2.2 Influential Observation Detection

We now perform the analysis on detecting influential observations for the second cohort of local breast cancer data.

Delete-case method

Figure 7.10 gives the influential patients identified using the delete-case method. This plot illustrates that the influential patients do exist in each prognostic factor. Table 7.13

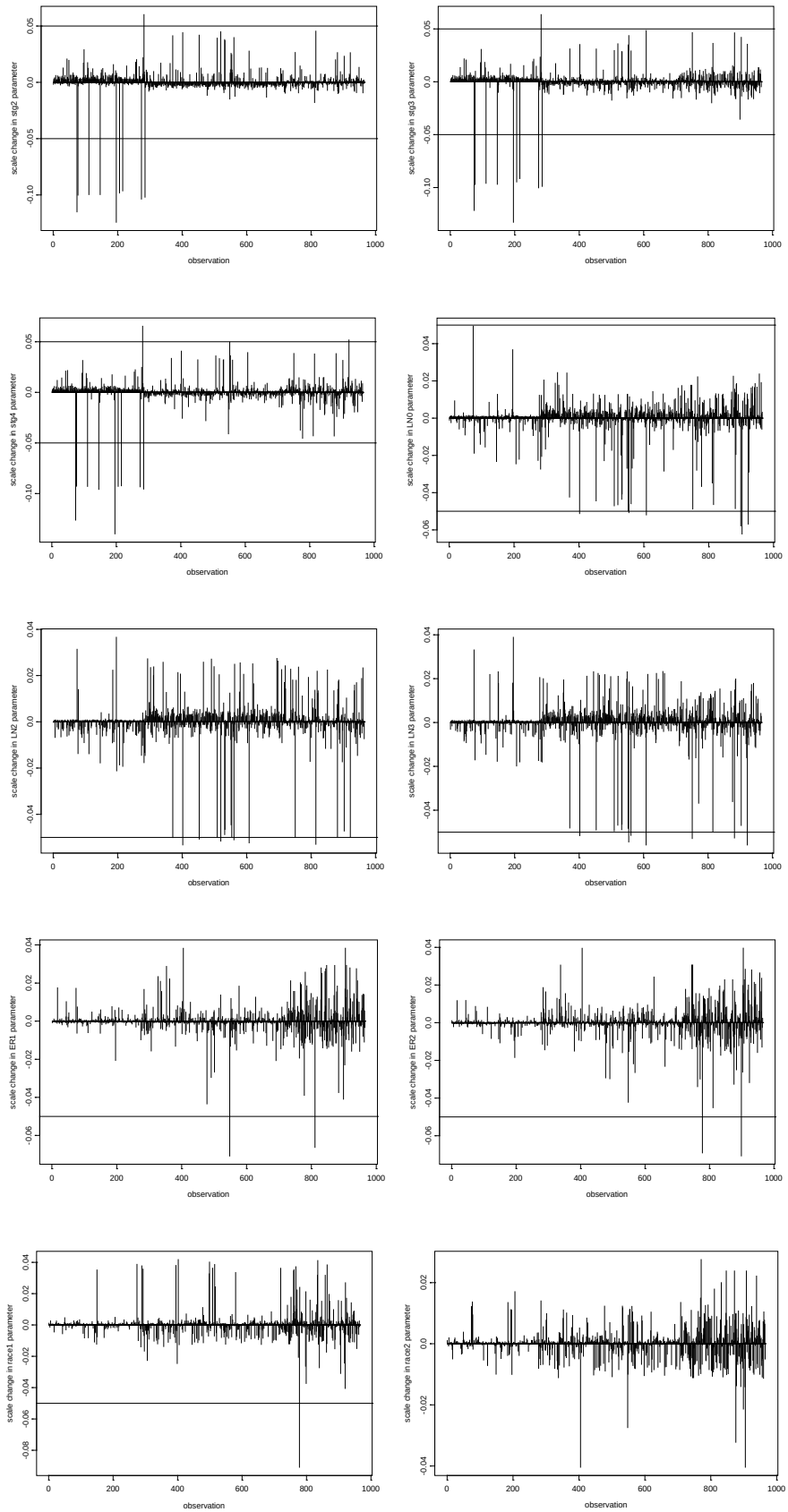

Figure 7.10
dfbetas residuals plot for second cohort

gives the profiles of 9 patients detected as influential patients by delete-case method. The influential patients are 74, 196, 282, 405, 548, 771, 778, 876 and 899. Note that two patients are identified earlier as outliers by r_{Ni}. They are patients 405 and 876. Omitting these patients from the data set also change the parameter estimate and hazard as shown in Table 7.14. However, the changes are not too large except for *stg*2, *stg*3, *stg*4 and ER2 in parameter estimate values and *stg*2, *stg*3 and *stg*4, *ER*2, *race*1 and *race*2 in hazard values. The changes are more than 0.2 for each level of the prognostic factors involved. Therefore, further investigation should be carried out to determine other factors that contribute to the survival of these patients.

Table 7.13
Profile of influential observation using delete case method

patient (*i*)	**time**	**status**	**prognostic factors**			
			stg	***LN***	***ER***	***race***
196	21	1	1	0	0	2
74	28	1	1	0	1	2
778	34	0	4	0	2	1
548	53	0	4	0	1	2
876	67	0	4	3	2	2
282	81	0	1	0	2	2
771	88	0	4	3	2	0
405	89	0	4	0	0	2
899	95	0	3	0	2	2

Cox PHM FS method

The Cox PHM FS method proposed in Chapter 6 is considered. Here, we set the percentage of initial subset for techniques 1, 2 and 3 to be 97% due to high percentage of censoring in the data set and the values of γ is set to be 3, 6 and 9. For techniques 5, 6 and 7, we set the lower cut point equals -1 while the upper cut point

equals 3 for deviance and normal deviate residual, while -1 and 6 respectively for log-odds residual.

Table 7.14
Cox Model result when omit influential observation using delete case method

Patient exclude from the fitted model	Variables	Parameter estimate	Hazard	C.I of hazard
None	stg2	0.592	1.807	(0.851, 3.84)
	stg3	1.625	5.079	(2.355, 10.95)
	stg4	2.614	13.656	(6.257, 29.81)
	LN0	1.679	5.361	(3.061, 9.39)
	LN2	0.744	2.103	(1.205, 3.67)
	LN3	1. 183	3.264	(1.908, 5.58)
	ER1	-0.146	0.864	(0.579, 1.29)
	ER2	0.553	1.739	(1.152, 2.63)
	race1	0.101	1.107	(0.716, 1.71)
	race2	0.382	1.466	(1.109, 1.94)
Delete-case (patients in Table 7.14)	stg2	0.773	2.167	(0.938, 5.00)
	stg3	1.868	6.475	(2.758, 15.20)
	stg4	3.078	21.706	(9.090, 51.83)
	LN0	1.732	5.650	(3.209, 9.95)
	LN2	0.664	1.942	(1.116, 3.38)
	LN3	1.157	3.182	(1.867, 5.42)
	ER1	-0.003	0.997	(0.664, 1.50)
	ER2	0.850	2.341	(1.524, 3.60)
	race1	0.206	1.230	(0.795, 1.90)
	race2	0.474	1.607	(1.212, 2.13)

Table 7.15 gives the proportion of similar and non-similar patients selected by seven different techniques of the initial subset. We observe that similar proportions for all techniques are more than 0.5 as shown in Table 7.15(a) while the non-similar proportion is not more than 0.37 as shown in Table 7.15(b). It is found that the initial subset for techniques 5 and 7 are subsets to the initial subset obtained using technique 6. We also find that FS4, FS5, GS6, FS7 and FS9 do not perform well when we have high percentage of censoring in the data set. Thus, these procedures are not considered for further analysis.

Table 7.15
Proportion of patients selected by seven different initial subset techniques a) similar proportion b) non-similar proportion

		Techniques								
	(a)	**1**	**2**	**3**	**4**			**5**	**6**	**7**
					γ=3	**γ =6**	**γ =9**			
Techniques	**1**	1	0.99	0.99	0.64	0.77	0.88	0.91	0.95	0.89
	2		1	1	0.64	0.77	0.88	0.91	0.95	0.89
	3			1	0.64	0.77	0.88	0.91	0.95	0.89
	4 **γ =3**				1	0.83	0.71	0.65	0.63	0.65
	γ =6					1	0.86	0.77	0.76	0.76
	γ =9						1	0.87	0.87	0.85
	5							1	0.96	0.97
	6								1	0.94
	7									1
	b)									
Techniques	**1**	0	0.007	0.007	0	0	0.009	0.04	0.04	0.04
	2	0.007	0	0	0	0	0.008	0.04	0.04	0.04
	3	0.007	0	0	0	0	0.008	0.04	0.04	0.04
	4 **γ =3**	0.36	0.36	0.36	0	0.17	0.29	0.35	0.37	0.34
	γ =6	0.23	0.23	0.23	0	0	0.14	0.22	0.24	0.21
	γ =9	0.11	0.11	0.11	0	0	0	0.11	0.13	0.11
	5	0.05	0.05	0.05	0	0.007	0.02	0	0.04	0
	6	0.02	0.02	0.02	0	0	0.003	0.003	0	0
	7	0.07	0.07	0.07	0.007	0.03	0.04	0.03	0.06	0

Figure 7.11 shows the progression plot for the FS1 method. To assist in identifying patients that affect the parameter estimates, we obtain the *IM* plots as given in Figure 7.12. It can be seen that influential patients exist at step $m+i$ equal to 6, 14, 19, and 28 corresponding to patients 520, 778, 77 and 405 respectively. We also obtain the progression and *IM* plots for FS2, FS3 and FS8 methods and are given in Appendix C.

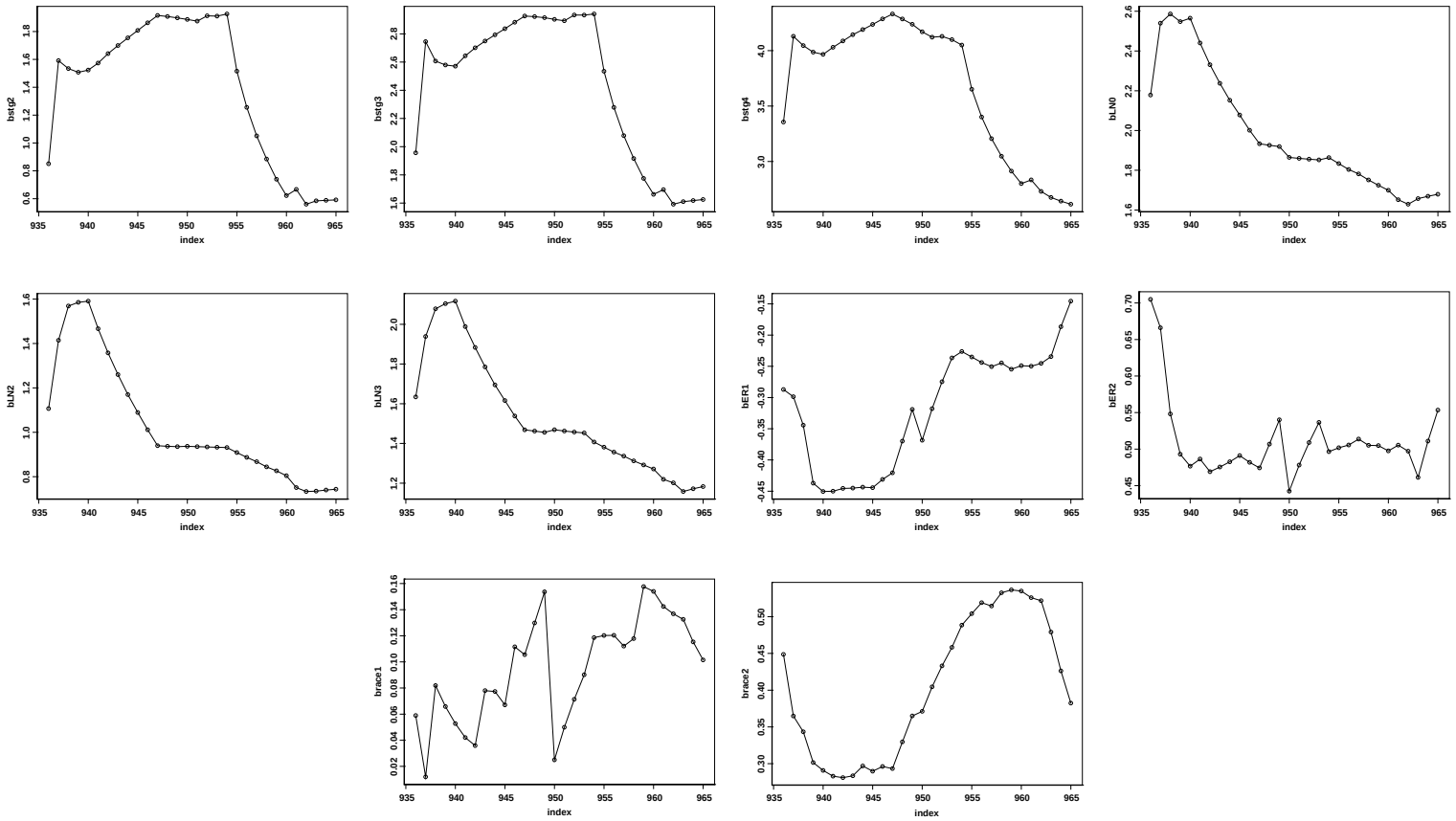

Figure 7.11
The progression plots for FS1 procedures on the second local breast cancer data

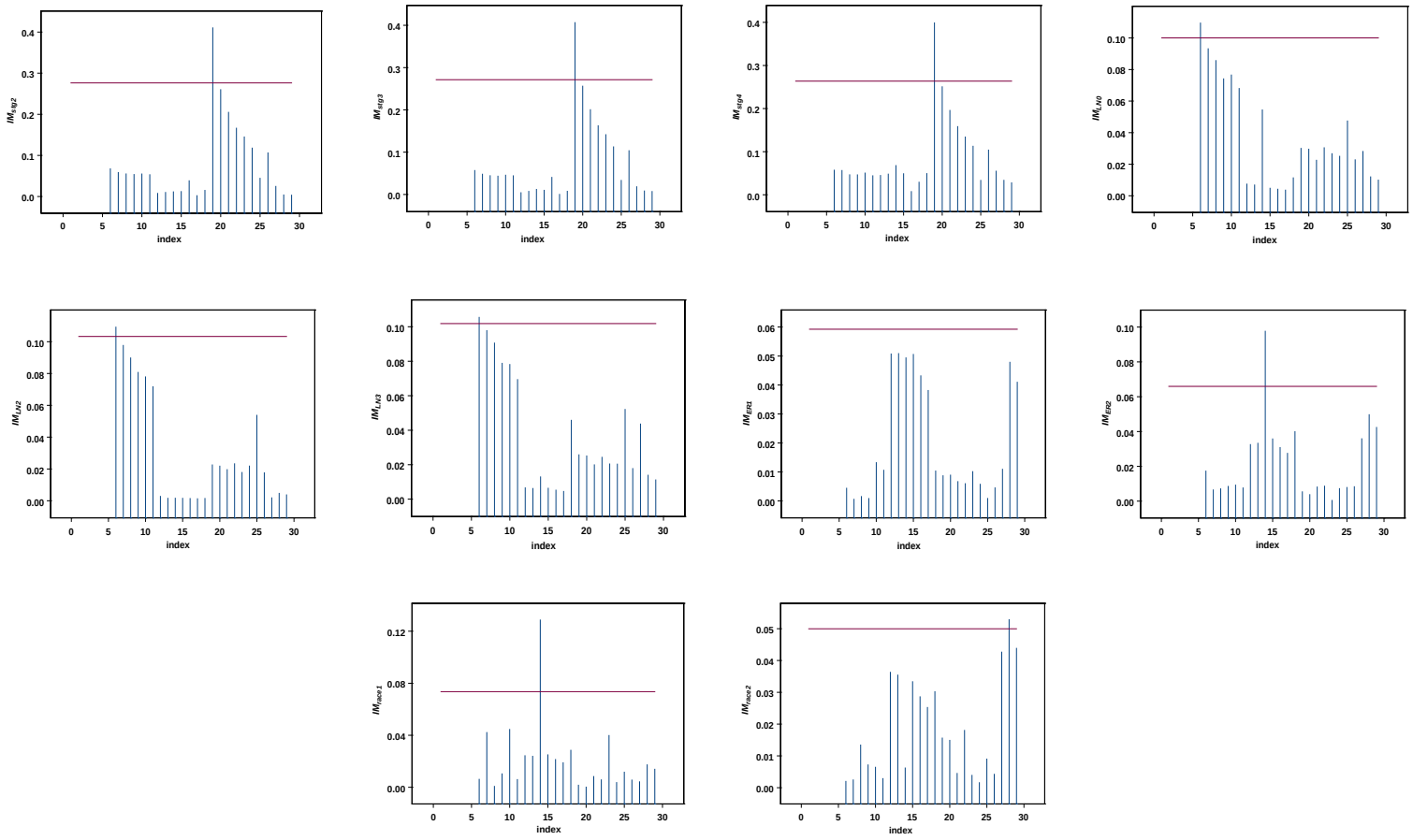

Figure 7.12
The *IM* plots for FS1 procedures on the second local breast cancer data

Table 7.16 gives the full list of influential patients identified by Cox PHM FS methods. It is found that the total number of influential patients identified by Cox PHM FS methods is 0.4% of the size of data set. Three main results are observed from this result. Firstly, we find that patients' number 77, 405 and 778 are identified as influential patients by every types of the Cox PHM FS method. Note that patient 405 has also been identified as an outlier in section 7.2.1. Secondly, the FS1, FS2 and FS3 methods identify similar set of influential patients. Lastly, only one patient identified by FS8 is different from that identified by FS1, FS2 and FS3. Note that the influential patients identified by every type of the Cox PHM FS method are almost the same.

Table 7.16
Influential observation selection by Cox PHM FS methods

Patient	**FS1**	**FS2**	**FS3**	**FS8**
77	/	/	/	/
405	/	/	/	/
778	/	/	/	/
520	/	/	/	
509				/
total	4	4	4	4

Table 7.17 gives the estimate of the parameter estimates, hazard for the Cox PHM before and after omitting the influential patients selected by each type of the Cox PHM FS method. It is found that the influential patients do affect the statistic considered. However, the changes on the statistics of each prognostic factor are not too large except for the hazard ratio of *stg*3, *stg*4 and *LN*0 only, where the changes are more than 0.2.

Table 7.17
Cox Model result when omit influential observation using Cox PHM FS method

Patient exclude from the fitted model	Variables	Parameter estimate	Hazard	Standard error	C.I of hazard
Null	stg2	0.592	1.807	0.384	(0.851, 3.84)
	stg3	1.625	5.079	0.392	(2.355, 10.95)
	stg4	2.614	13.656	0.398	(6.257, 29.81)
	LN0	1.679	5.361	0.286	(3.061, 9.39)
	LN2	0.744	2.103	0.284	(1.205, 3.67)
	LN3	1.183	3.264	0.274	(1.908, 5.58)
	ER1	-0.146	0.864	0.204	(0.579, 1.29)
	ER2	0.553	1.739	0.210	(1.152, 2.63)
	race1	0.101	1.107	0.222	(0.716, 1.71)
	race2	0.382	1.466	0.142	(1.109, 1.94)
FS1, FS2 and FS3	stg2	0.654	1.924	0.403	(0.874, 4.23)
	stg3	1.691	5.427	0.410	(2.431, 12.11)
	stg4	2.761	15.822	0.416	(7.003, 35.75)
	LN0	1.784	5.951	0.293	(3.354, 10.56)
	LN2	0.805	2.237	0.291	(1.264, 5.96)
	LN3	1.222	3.394	0.282	(1.955, 5.89)
	ER1	-0.146	0.864	0.204	(0.580, 1.29)
	ER2	0.594	1.814	0.213	(1.192, 2.75)
	race1	0.228	1.256	0.223	(0.812, 1.94)
	race2	0.410	1.506	0.143	(1.138, 1.99)
FS8	stg2	0.660	1.934	0.403	(0.878, 4.26)
	stg3	1.697	5.460	0.410	(2.445, 12.19)
	stg4	2.756	15.736	0.416	(6.964, 35.56)
	LN0	1.781	5.934	0.293	(3.345, 10.53)
	LN2	0.803	2.233	0.291	(1.262, 3.95)
	LN3	1.225	3.405	0.282	(1.961, 5.91)
	ER1	-0.133	0.875	0.203	(0.588, 1.30)
	ER2	0.583	1.791	0.213	(1.178, 2.72)
	race1	0.184	1.202	0.226	(0.772, 1.87)
	race2	0.405	1.500	0.143	(1.133, 1.98)

Table 7.18 gives the profiles of influential patients in the second cohort of local breast cancer data. We observe that there are three patients; 77, 405 and 509 that are identified as influential patients by every Cox PHM FS methods. Patients 405 who have longer survival times are predicted to have quite short survival times. For patient 77 and 509, she has short survival time though is predicted to survive longer. Therefore, a research of unknown prognostic factors can be originated from the

thorough investigation of the medical records of these patients. For example, the study may need to include more prognostic factors such as human epidermal growth factor receptor 2 (HER2 or c-erbB-2).

Table 7.18
Profile of influential patients in the second cohort of local breast cancer data

patient (*i*)	**time**	**status**	**prognostic factors**				$S(t_i)$
			stg	***LN***	***ER***	***race***	
509	25	1	2	1	2	1	0.9562
778	34	0	4	0	2	1	0.0876
77	46	1	1	1	2	2	0.9440
405	89	0	4	0	0	2	0.0251
520	89	1	2	1	1	0	0.9478

7.3 Summary

The outlier and influential observation detection methods developed in this thesis have been applied on two cohorts of Malaysian breast cancer data. In general, methods based on the r_{N_i} and r_{L_i} give smaller number of outlier and influential observations which agree to the definition of outlier in survival model. That is, they are patients who unexpectedly die too early with early stage of cancer prognosis, and patients who survive longer with advanced stage of cancer prognosis.