CHAPTER FIVE

OUTLIERS AND INFLUENTIAL OBSERVATIONS IN SURVIVAL DATA

Data collected in survival studies may contain outliers. An outlier is usually identified as an individual whose survival time is not well fitted by the model. This outlier may have a large positive or negative residual value according to whether the individuals die "too early" or "too late" with respect to the model prediction. The individual may have a long survival time but the explanatory variables values suggest that she/he should have died earlier.

Outliers may exist due to various reasons (see Hampel *et al.* (1986)). For example, it might be due to the gross error. The error refers to copying or punching errors, in particular wrong decimal point, wrong scale of measurement taken over a period of time, interchange two values with different meaning, inadvertent observation of a member of a different population, transient effect, or equipment failure. Gross error is rarely found in highly quality data which are obtained with special care under good conditions. In general, Hampel *et al.* (1986) stated that data typically has 1-10% gross error. They further showed some examples for the occurrence and frequency of gross errors. In one of the examples, data on patients in medical area contain around 8-12% gross error.

Therneau *et al.* (1990) labeled outliers as observations or individuals who die "too early" or "too late". Meanwhile, Maller and Zhou (1994) associated the occurrence of outlier in survival studies to the possible existence of "immune" or "cured" individuals in the population. Nardi and Schemper (1999) further specified the meaning of outliers in survival study as individual who "lived far too long" or "die far too early". The study showed that patients who "lived far too long" mostly come from uncensored survivors while patients who "die far too early" come from censored survivors.

Outliers may give undue effect to the inferences made based on the model. These outliers are called influential observations. They are often found among long survivors. In the Cox PHM, influential observations may affect the parameter estimation, and consequently change the hazard ratios. In some cases, omitting the influential observation may lead to different Cox PHM. Thus, it is important to investigate the occurrence of outliers and influential observations in the data set. Procedures in detecting outliers and influential observations will be discussed in the next section, while the illustration of the procedures will be given in section 5.3.

5.1 Outliers Detection Procedure

In data analysis, outlier detection is commonly carried out based on the residuals analysis. In the Cox PHM, a type of residual that can be used is martingale residual. It refers to the distance between the observed number of deaths in the interval (0, t_i) and the corresponding estimated expected number on the basis of the fitted model, where t_i is the survival time of the *i*th individual. This type of residual is not symmetrically distributed about zero, even when the fitted model is correct. For illustration, the martingale residual plot of a simulated data set as given in Figure 5.1 clearly shows this property. This skewness makes the residual plot difficult to interpret. The individuals who die "too early" will have the residual values squeezed toward +1, while individuals who die “too late” will have large residual values scaling toward - ∞ (see Collet (2003)).

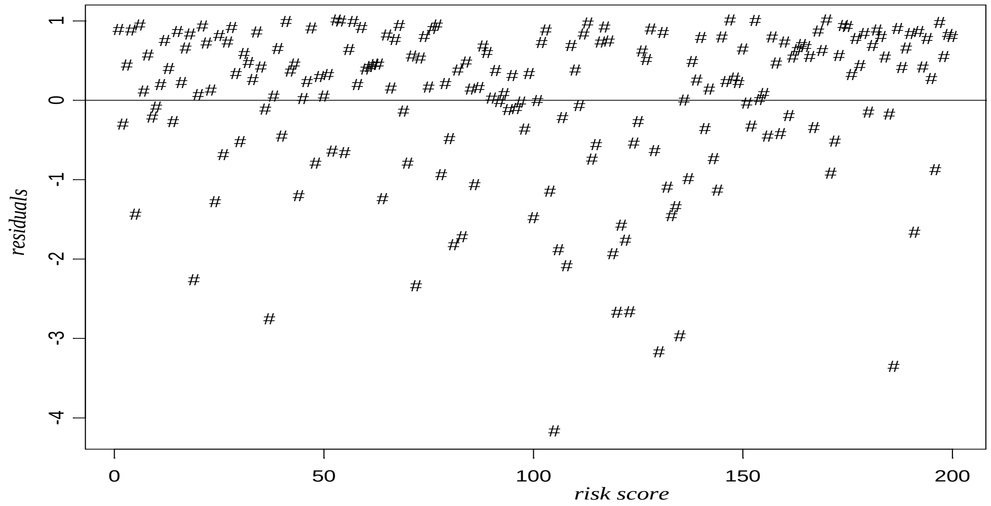


Figure 5.1
Martingale residuals vs. risk score plot on generated sample of size 200 with no censoring

Therneau *et al.* (1990) proposed deviance residuals to overcome the skewness problem. This residual is more symmetrically distributed about zero as shown in Figure 5.2 for the simulated data and is approximately normally distributed. The index plot of the deviance residuals is now easier to interpret. As a result, this residual can be used with graphical tools to identify individuals whose survival times are extremely different to other residuals. A number of authors have used deviance residual plot for detecting the outliers (see Collet (1994), Marubini and Valsecchi (1995) and Therneau and Grambsch (2000)). They used subjective reasoning without any formal hypothesis testing to identify outliers.

Recently, Nardi and Schemper (1999) proposed two new types of residuals with the main purpose of detecting outliers in the Cox PHM. Further they extended the

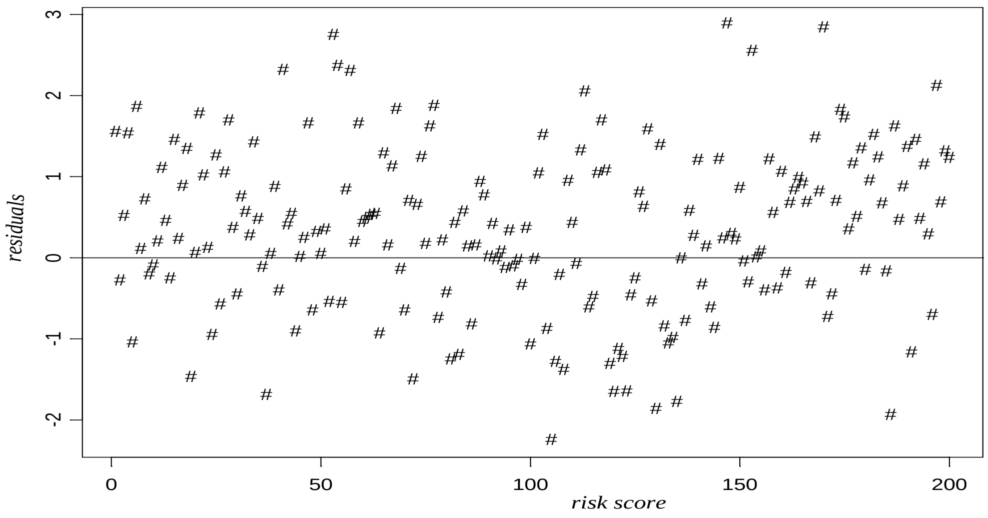


Figure 5.2
Deviance residuals vs. risk score plot on generated sample of size 200 with no censoring

theory to several parametric models (see Nardi and Schemper (2003)). These residuals are known as log-odds (r_{L_i}) and normal deviate (r_{N_i}) residuals. They are obtained by transforming the survival probability using the logit and probit transformations, respectively. As we can see in Figures 5.3 and 5.4 for the generated sample, these two residuals are symmetrically distributed about zero. Since these two residuals have their reference distribution, formal tests on detection of outliers can be performed.

Consequently, the formal graphical and numerical test on detecting outliers can be obtained based on these three types of residuals; deviance, normal deviate, and log-odds residuals. The graphical technique is by plotting the residual values versus the risk score or prognostic index ($\hat{\boldsymbol{\beta}}\boldsymbol{x}_i$) for all individuals as shown in Figures 5.2, 5.3, and 5.4 for the generated sample. On the other hand, numerical techniques are performed by introducing cut points. The cut point is determined based on the reference distributions of the residuals and to be discussed in section 5.3.

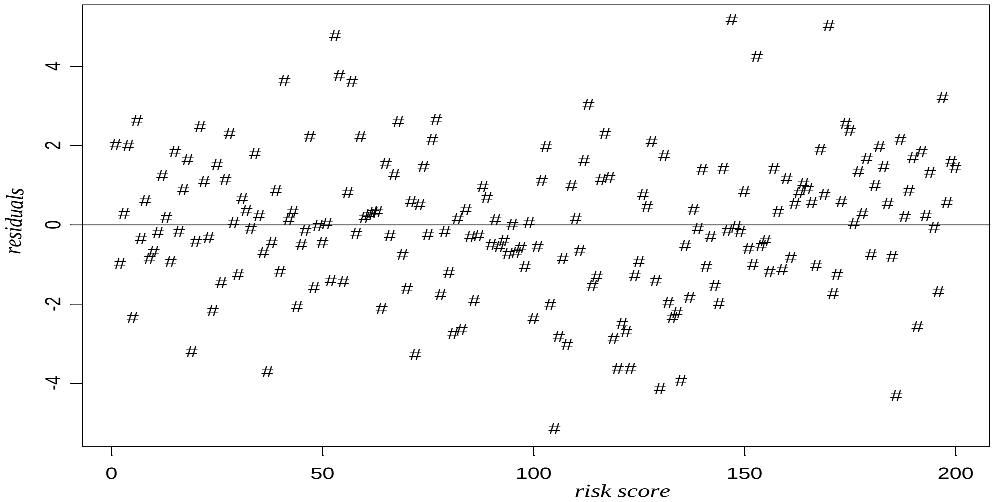


Figure 5.3
Log-odds residual vs. risk score plot on generated sample of size 200 with no censoring

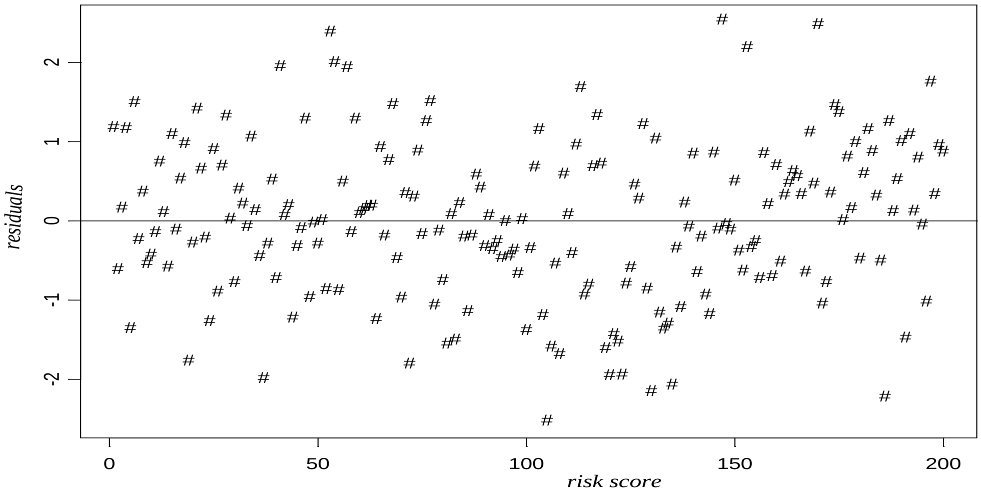


Figure 5.4
Normal deviate residual vs. risk score plot on generated sample of size 200 with no censoring

5.1.1 Deviance Residuals

The deviance residual is introduced by Therneau et al. (1990) and is denoted as r_{Di}. The equation is defined as follows:

$$r_{D_i} = \operatorname{sgn}\left(r_{M_i}\right)\left[-2\left\{r_{M_i} + \delta_i \log\left(\delta_i - r_{M_i}\right)\right\}\right]^{\frac{1}{2}} \tag{5.1}$$

where $r_{M_i} = \delta_i - \exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_i)\hat{H}_o(t_i)$ is the martingale residuals, $\delta_i = 0$ for the censored observations and $\delta_i = 1$ for the uncensored observations, $i = 1,2...,n$ and n are number of individuals in sample. The second term in the formula for r_{M_i} requires the estimate of $\hat{H}_i(t_i)$, which is the cumulative hazard or cumulative probability of death of the *i*th observation over the interval $(0, t_i)$. On the other hand, $\hat{H}_o(t_i)$ is known as the cumulative baseline hazard function given by $\hat{H}_0(t) = -\sum_{j=1}^{k} \log \xi_j$ where $\tilde{\xi}_j = 1 - \frac{d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}\mathbf{x}_l)}$. The function $\operatorname{sgn}(r_{M_i})$ in equation (5.1) is the sign of r_{M_i} function. This is the function that takes the value +1 if its argument is positive and -1 if negative. Therefore, $\operatorname{sgn}(r_{M_i})$ ensures that the deviance residuals have the same sign as the r_{M_i}.

Martingale residuals take values between -∞ to unity. The values may approach -∞ when the observation is censored. The sum of residuals may be equal to zero when the sample size is too large. Since the deviance residual overcomes the skewness of martingale residuals, therefore the deviance residual have range from -∞ to +∞ and symmetrically distributed about zero. In this study the residuals approximately follow the standard normal distribution with mean zero and variance unity.

5.1.2 Normal Deviate residuals

The normal deviate residual is introduced by Nardi and Schemper (1999) and denoted as r_{Ni}. The reference sampling distribution of the residual is standard normal distribution. If the unknown survival function is replaced by its estimate and assuming a correctly specified model, $\hat{N}_i$ converge in probability to N_i. Therefore, we have

$$N_i = \Phi^{-1}\{S_i(T_i)\}$$

and residual values $\hat{n}_i = \Phi^{-1}\{\hat{S}_i(t_i)\}$.

Note that the probability of survival for censored time $S_i(t_i^c)$ is unknown. There are various options for accommodating the residuals of censored survival times (Crowley and Hu (1977)). True survival time is larger than the observed censored one and distribution of the unknown true residuals is related to the uniform distribution of $S_i(T_i)$ in [0, $\hat{S}_i(t_i^c)$]. Thus, the $\hat{S}_i(t_i^c)$ will be replaced with the conditional time median value $\frac{\hat{S}_i(t_i^c)}{2}$. Consequently the r_{N_i} for time censored is

$$n_i^c = \Phi^{-1}\left\{\frac{\hat{S}_i(t_i^c)}{2}\right\}$$

where n_i^c is mean of r_{N_i} time censored or can be replaced with

$$n_i^m = -\frac{\exp\left(0.5\left(n_i^c\right)^2\right)}{\sqrt{2\pi}\hat{S}_i\left(t_i^c\right)}.$$

where n_i^m is median of r_{N_i} time censored.

5.1.3 Log-odds Residuals

Log-odds residual is introduced by Nardi and Schemper (1999) and denoted as r_{Li}. The reference sampling distribution of the residuals is logistic distribution with mean $E(L_i) = 0$ and variance $\text{var}(L_i) = (\pi / 3)^2$. By having the correctly specified model and the unknown survival function is replaced by its estimate, we have the $\hat{L}_i$ converge in probability to L_i. Therefore, we have

$$L_i = \log[S_i(T_i)/\{1 - S_i(T_i)\}]$$

$$\text{and } \hat{l}_i = \log[\hat{S}_i(t_i)/\{1 - \hat{S}_i(t_i)\}]$$

The probability of survival for censored time $S_i(t_i^c)$ is unknown. Therefore, to accommodate the residuals of censored survival time values, Nardi and Schemper (1999) suggested to replaced $\hat{S}_i(t_i^c)$ with the conditional time median value $\frac{\hat{S}_i(t_i^c)}{2}$. The median l_i^c and mean l_i^m of log-odds residual values are then given by

$$l_i^c = \log[\hat{S}_i(t_i^c) / \{2 - \hat{S}_i(t_i^c)\}]$$

$$l_i^m = l_i^c - \frac{1 + \exp(l_i^c)}{\exp(l_i^c)} \log\{1 + \exp(l_i^c)\},$$

respectively.

5.1.4 Cut Point for Detecting Outliers

In earlier discussion r_{N_i} and r_{L_i} are said to follow certain distributions. Since the normal deviate residuals are normally distributed and the log-odds residuals follow the logistic distribution, outliers can be identified by the cut point based on these distributions.

The cut points are categorized as "Too Early Died" = TED and "Too Long Lived" = TLL. The cut points for normal deviate residual are given by:

$$R_{TED,N} = \{n_i : n_i > z_{1-\alpha}\}$$

$$R_{TLL,N} = \{n_i : n_i < z_{\alpha}\}$$

where z_α is the αth percentage point of standard normal distribution. Similar cut points can be used for deviance residual since both have the same reference distribution. While, the cut points for log-odds residual are given by

$$R_{TED,L} = \{l_i : l_i > w_{1-\alpha}\}$$

$$R_{TLL,L} = \{l_i : l_i < w_{\alpha}\}$$

where w_α is the αth percentage point of logistic distribution. If the residual values of individuals exceed the cut points they are identified as outliers.

5.2 Simulation Study

The simulation study in this chapter is designed to study the empirical distribution of r_{D_i}, r_{N_i} and r_{L_i}. We will then find the sampling behavior of the minimum/maximum values of residuals for the purpose of identifying outlier in Cox PHM.

5.2.1 Sampling Scheme

As mentioned earlier, the r_{D_i}, r_{N_i} and r_{L_i} approximately follow certain reference sampling distributions. We are interested to investigate whether the residuals do follow their respective reference sampling distributions in the presence of censoring. For this purpose, we perform 10,000 trials with sample size *n* of 20, 40, 80, 200, 400 and 800. Values of single covariate X are assumed to be binomially distributed with $P(X = 1) = P(X = 2) = 0.5$, and the corresponding survival times *T* are assumed to be exponentially distributed with hazard $h(t|X=1)=1$ and $h(t|X=2)=2$.

In this study, the censored observations are introduced in the sample based on Gehan and Thomas (1969). Assume that the observations time of entry t_i is taken from uniform distribution $(0,D)$ while random value x_i is taken from the exponential distribution. Let the assumed survival time be u_i for each observation i=1, 2,…, *n*. That is,

$$u_i = \begin{cases} x_i & \text{if } \delta = 1 \\ D - t_i & \text{if } \delta = 0 \end{cases} \quad \text{and} \quad \delta = \begin{cases} 1 & \text{if } \; D - t_i \geq x_i \\ 0 & \text{if } \; D - t_i < x_i \end{cases}.$$

When $\delta=1$, the corresponding observation is uncensored observations while when $\delta=0$, we have censored observation.

For each experimental condition, a value of *D* which is the time of analysis, is chosen based on the approach used by Lininger *et al.* (1979) to achieve an expected percentage of censoring ρ in data set. The formulation is given below:

$$\rho = \frac{100}{D}\int_0^D \left\{ 1 - \sum_m \sum_j \Pi_{mj} \int_0^t f_{mj}(s)ds \right\} dt \tag{5.2}$$

where Π_{mj} is the proportion of the patients in the treatment *m* and stratum *j* and f_{mj} is the probability density of the failure time. In this simulation study, we use *m* = 0 or 1 while *j* = 1. Using the Newton - Raphson technique, the value of *D* in equation (5.2) can be generated when the percentage of censoring ρ is known.

5.2.2 Effect of Percentage of Censoring on The Residuals Distribution

A Cox PHM can now be fitted on the generated data of each trial. Thus, the sampling distribution of r_{Di}, r_{Ni} and r_{Li} residual of each trial are obtained. Our main interest is to see how well the residuals follow the reference distributions when censoring are introduced in the generated data set. From the theory, we know that r_{Di} and r_{Ni} follow the standard normal distribution, while the r_{Li} follow the logistic distribution.

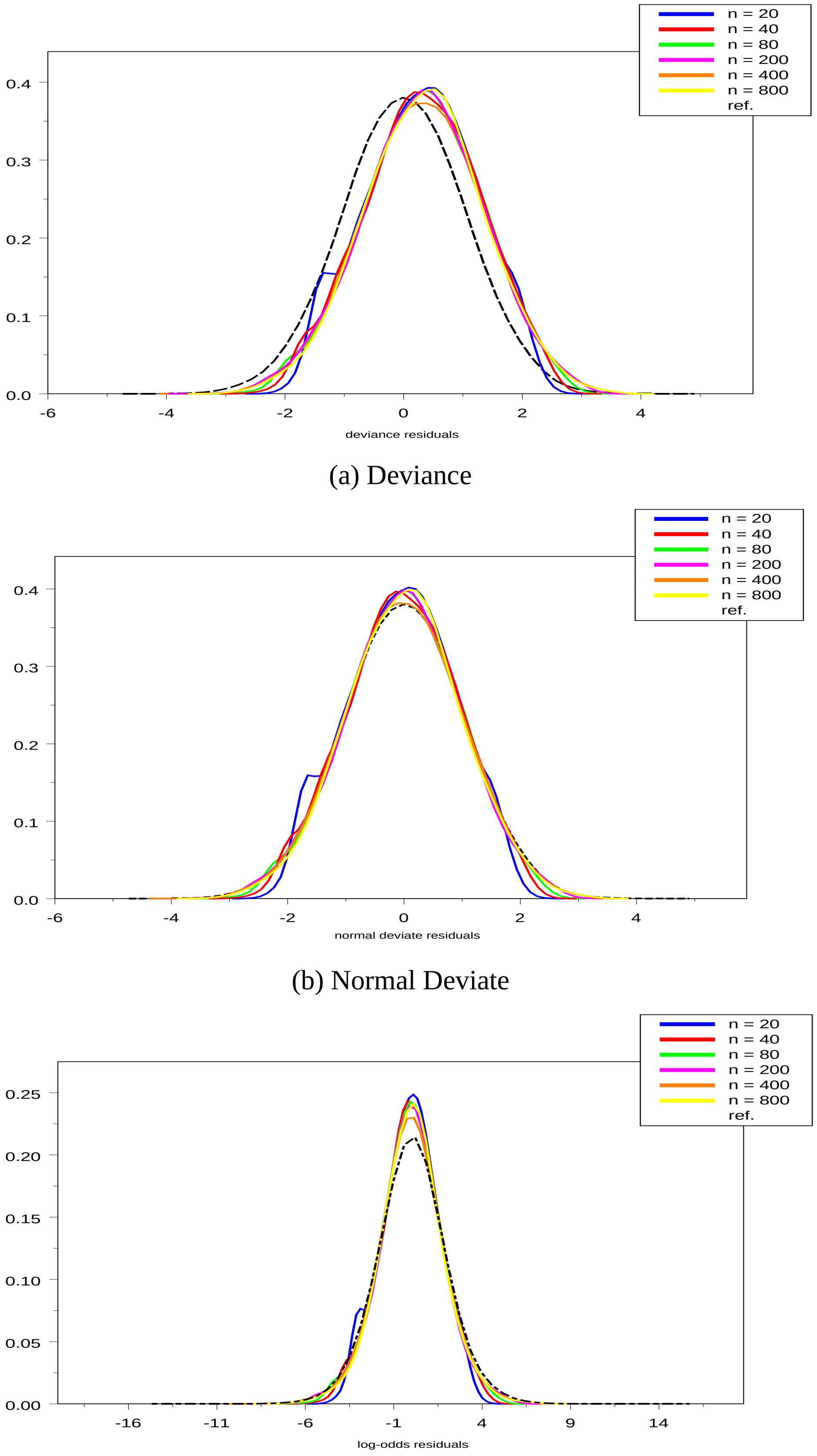
Figure 5.5 gives the empirical distributions of the residuals from the last observation in each sample size *n* in 10,000 trials with no censored observations introduced in the generated data set. The r_{Di} with small sample size $n = 20$ as given in Figure 5.5(a) shows the empirical sampling distribution already departs from the standard normal distribution. However, the empirical distribution of this residual follows the bell-shape of a normal distribution. We conclude that r_{Di} is normally distributed but not necessarily standard normal distribution. Similar finding are observed to increasing number of *n*.

In addition, from Figure 5.5(b), we observe that the empirical sampling distribution of r_{Ni} agrees quite well to the standard normal distribution when $n = 20$. It can be seen that the shape of the empirical distribution of residuals follows almost exactly the shape as the reference sampling distribution with mean zero and variance unity. The same finding is observed for large *n*.

Whilst, in Figure 5.5(c), we observe that the empirical sampling distribution of r_{Li} gives similar finding as r_{Ni}. The empirical sampling distribution of r_{Li} agree quite well to the logistic distribution when n = 20. For large n, similar findings are obtained. That is, with 0% censoring, the empirical sampling distribution of r_{Ni} and r_{Li} agree quite well with their theoretical reference distributions while r_{Di} is assumed to be normally distributed. In the meantime, by taking sample size to be 200 as an example, we find that the Kolmogorof-Smirnov test shows that the empirical distribution of normal deviate (p-value = 0.9343) and log-odds (p-value = 0.9807) residual accepting the hypothesis that the residuals follow their respective reference distributions.

Next, we explore the empirical distributions of r_{Di}, r_{Ni} and r_{Li} when censored observations are introduced into the simulated data. Figure 5.6 gives the empirical distribution of r_{Di} for different n and percentages of censoring, denoted by ρ. We observe that, when ρ are small and n are increasing, the empirical distribution departs from the reference sampling distribution. The curve of empirical distribution also shows a bimodal shape. Similar finding is observed for larger ρ as given in Figures 5.6(b) and 5.6(c). It becomes worse when the degree of censoring is increasing for all n. The bimodal shape becomes clear for 20% censoring and above.

Figure 5.7 shows the empirical distribution of r_{Ni} for different sizes of n and percentage of censoring ρ in the simulated data. When n increases and ρ = 20%, the empirical distribution agrees quite well to the reference sampling distribution. However, the empirical distribution starts to shows bimodal shape and depart from the reference sampling distribution at ρ = 40% censoring. Similar results are observed for r_{Li} as shown in Figure 5.8. The bimodal shape is attributed to the mixture of uncensored and censored observations in the generated data set.

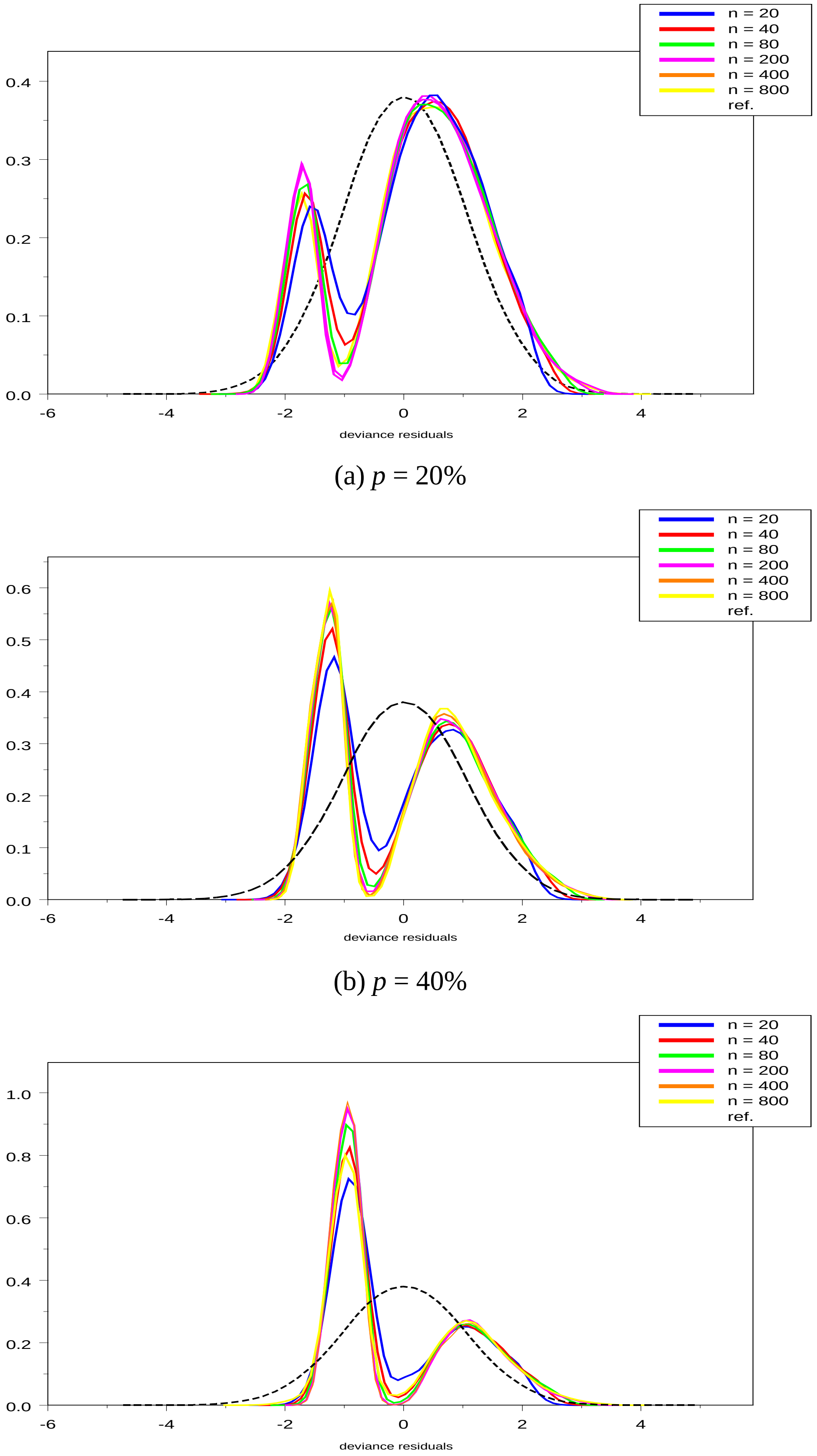


(a) Deviance

(b) Normal Deviate

(c) Log –odds

Figure 5.5
Empirical distribution of three suggested residuals from the last observation in each sample size in 10,000 trials without censoring

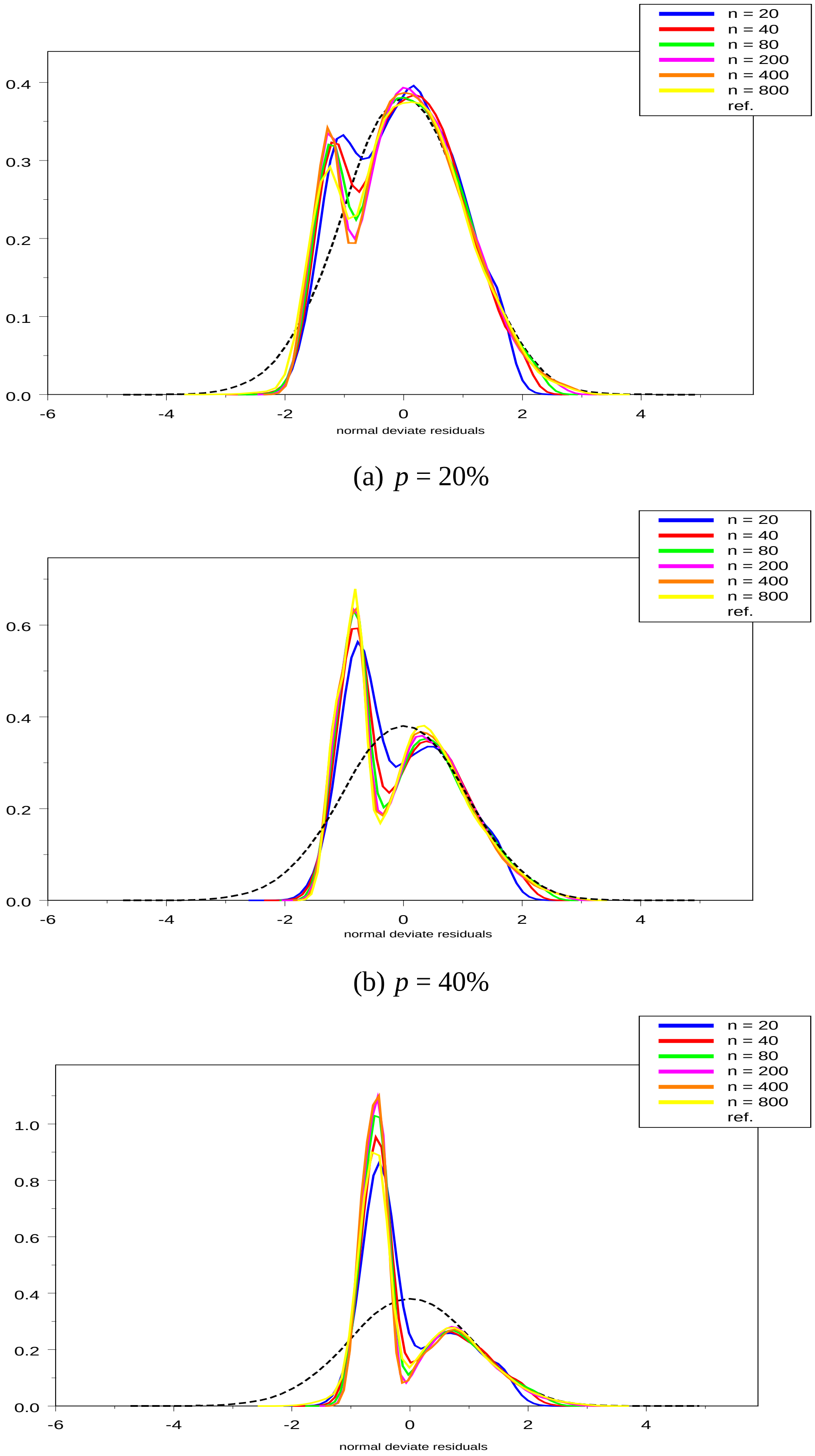


(a) *p* = 20%

(b) *p* = 40%

(c) *p* = 60%

Figure 5.6
Empirical distribution of deviance residual from the last observation in difference ρ and n in 10,000 trials

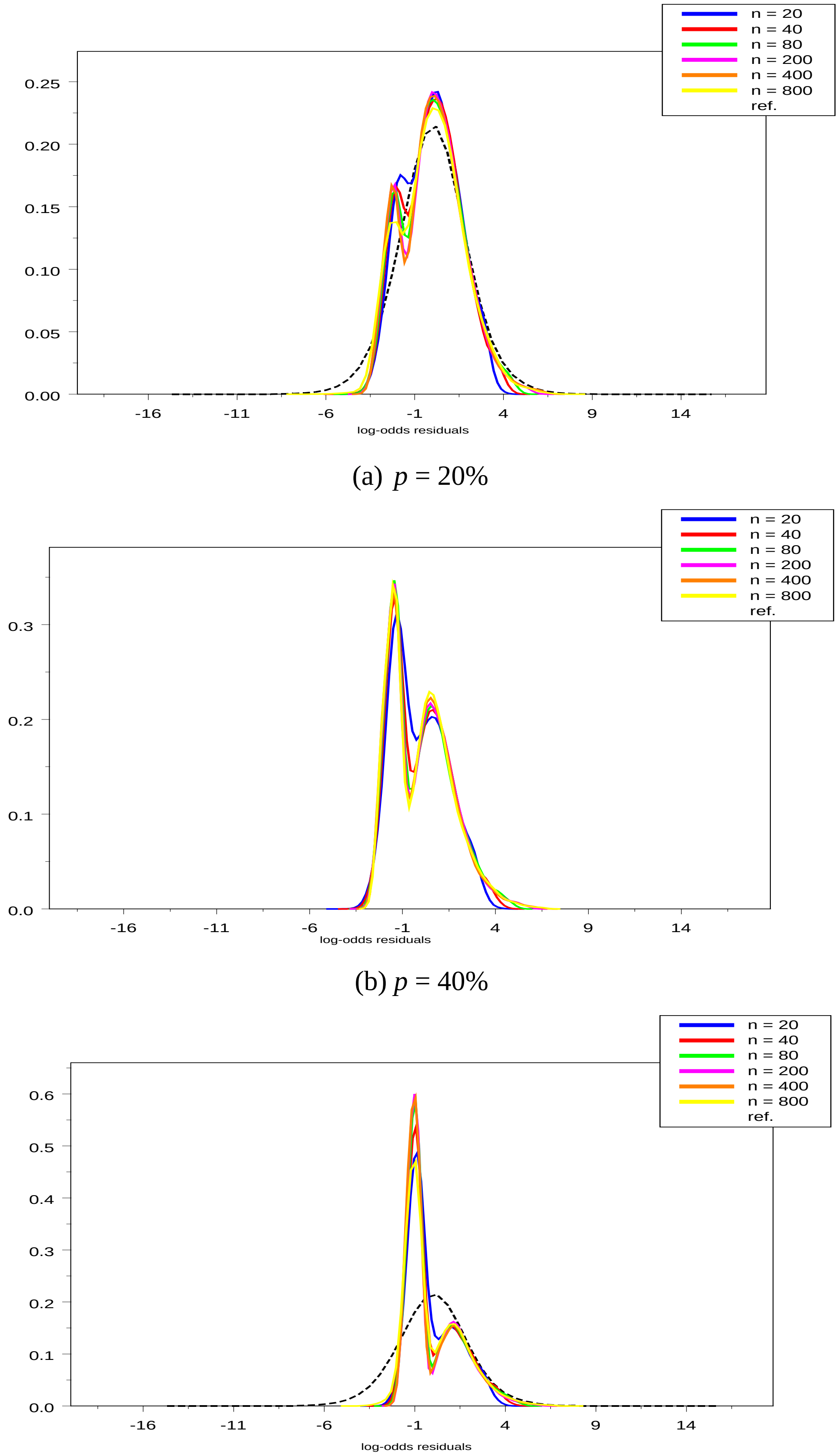


(a) p = 20%

(b) p = 40%

(c) p = 60%

Figure 5.7
Empirical distribution of normal deviate residual from the last observation in difference ρ and n in 10,000 trials



(a) p = 20%

(b) p = 40%

(c) p = 60%

Figure 5.8
Empirical distribution of log-odds residual from the last observation in difference ρ and n in 10,000 trials

5.2.3 Sampling Behavior of the Maximum/Minimum Values of Residuals

Outliers may exist in survival data at the beginning and end of the range of survival time. Therefore, two cut points should be considered. The cut points are obtained by taking the 90th, 95th and 99th percentiles of both the minimum and maximum values of residual values for 10,000 simulated data sets.

Tables 5.1 to 5.4 give the cut points for different percentages of censoring ρ and sample size n. Let take a look at n = 400, when the log-odds residual is being used, the lower and upper cut points are (-5.8946, 6.4673) for the 90th percentile, (-5.8701, 6.5473) for the 95th percentile and (-5.8271, 6.7766) for the 99th percentile. For normal deviate residual, the cut points are (-2.7686, 2.9459) for the 90th percentile, (-2.7686, 2.9574) for the 95th percentile and (-2.7546, 2.9798) for the 99th percentile. For the deviance residual, the cut points are (-2.4991, 3.2961) for the 90th percentile, (-2.4910, 3.3074) for the 95th percentile and (-2.4767, 3.3294) for the 99th percentile. We observe that the cut points are different for each type of residual with various sample size and percentage of censoring in the data set. When the percentage of censoring ρ increases, the percentiles of minimum and maximum values of r_{Di}, r_{Ni} and r_{Li} also increases. We suggest that the cut points for r_{Di}, r_{Ni} and r_{Li} are taken to be (-1, 3), (-1,3) and (-1, 6), respectively. This cut points will further use in section 6.2.3 and Chapter 6.

5.3 Delete-case Method for Detecting Influential Observation

Outliers do exist in survival data and some of them may give an undue impact to the inference made by the model. They might change the parameter estimates and hazard

Table 5.1
Percentile of min and max of residuals in 10,000 trials without censoring

Residuals		**Percentile of min residual values**			**Percentile of max residual values**		
	N	**90**	**95**	**99**	**90**	**95**	**99**
Log-odds	20	-2.9365	-2.8568	-2.7275	3.4654	3.6510	4.0392
	40	-3.6172	-3.5578	-3.4493	4.1547	4.2924	4.5421
	80	-4.3026	-4.2549	-4.1702	4.8387	4.9247	5.0795
	200	-5.2075	-5.1761	-5.1183	5.7465	5.8025	5.9005
	400	-5.8946	-5.8701	-5.8271	6.4673	6.5404	6.7766
	800	-6.5824	-6.5680	-6.5324	7.1111	7.1509	7.1937
Normal Deviate	20	-1.6412	-1.6043	-1.5434	1.8762	1.9547	2.1129
	40	-1.9406	-1.9155	-1.8693	2.1584	2.2119	2.3066
	80	-2.2158	-2.1974	-2.1645	2.4156	2.4465	2.5014
	200	-2.5462	-2.5352	-2.5151	2.7282	2.7466	2.7785
	400	-2.7686	-2.7686	-2.7546	2.9459	2.9574	2.9798
	800	-2.9915	-2.9883	-2.9774	3.1504	3.1620	3.1744
Deviance	20	-1.3368	-1.2989	-1.2365	2.2413	2.3193	2.4761
	40	-1.6436	-1.6179	-1.5706	2.5211	2.5740	2.6675
	80	-1.9255	-1.9066	-1.8728	2.7751	2.8055	2.8597
	200	-2.2635	-2.2524	-2.2317	3.0826	3.1007	3.1320
	400	-2.4991	-2.4910	-2.4767	3.2961	3.3074	3.3294
	800	-2.7188	-2.7135	-2.7029	3.4978	3.5064	3.5170

Table 5.2
Percentile of min and max of residuals in 10,000 trials with 20% censoring

Residuals		**Percentile of min residual values**			**Percentile of max residual values**		
	N	**90**	**95**	**99**	**90**	**95**	**99**
Log-odds	20	-2.1255	-1.9936	-1.7268	3.3045	3.4661	3.7926
	40	-2.3396	-2.2234	-2.0081	3.9621	4.0766	4.3115
	80	-2.4644	-2.3604	-2.1913	4.6421	4.7183	4.8735
	200	-2.5965	-2.5186	-2.4015	5.5477	5.5997	5.6859
	400	-2.6775	-2.6193	-2.5209	6.2267	6.2626	6.3241
	800	-2.7465	-2.7076	-2.6355	6.9137	6.9407	6.9822
Normal Deviate	20	-1.2167	-1.1409	-0.9969	1.8066	1.8765	2.0133
	40	-1.3193	-1.2580	-1.1449	2.0821	2.1277	2.2192
	80	-1.3810	-1.3288	-1.2409	2.3438	2.3718	2.4281
	200	-1.4468	-1.4081	-1.3492	2.6622	2.6796	2.7082
	400	-1.4865	-1.4580	-1.4092	2.8826	2.8939	2.9131
	800	-1.5200	-1.5011	-1.4660	3.0922	3.1003	3.1126
Deviance	20	-1.4940	-1.3936	-1.2304	2.1719	2.2416	2.3774
	40	-1.6789	-1.6069	-1.4642	2.4456	2.4907	2.5813
	80	-1.7735	-1.7203	-1.6282	2.7043	2.7319	2.7874
	200	-1.8410	-1.8024	-1.7435	3.0178	3.0347	3.0630
	400	-1.8805	-1.8522	-1.8035	3.2341	3.2451	3.2640
	800	-1.9138	-1.8950	-1.8601	3.4395	3.4474	3.4594

Table 5.3
Percentile of min and max of residuals in 10,000 trials with 40% censoring

Residuals		**Percentile of min residual values**			**Percentile of max residual values**		
	N	**90**	**95**	**99**	**90**	**95**	**99**
Log-odds	20	-1.4910	-1.3562	-1.1252	3.3847	3.5660	3.9568
	40	-1.5991	-1.4961	-1.3197	4.0697	4.2045	4.4929
	80	-1.6961	-1.6148	-1.4728	4.7483	4.8496	5.0347
	200	-1.8112	-1.7576	-1.6580	5.6539	5.7218	5.8393
	400	-1.8718	-1.8341	-1.7620	6.3345	6.3761	6.4564
	800	-1.9235	-1.8990	-1.8447	7.0176	7.0465	7.0976
Normal Deviate	20	-0.8582	-0.7805	-0.6459	1.8415	1.9190	2.0800
	40	-0.9191	-0.8608	-0.7594	2.1249	2.1778	2.2881
	80	-0.9734	-0.9279	-0.8475	2.3828	2.4195	2.4856
	200	-1.0371	-1.0076	-0.9522	2.6977	2.7201	2.7586
	400	-1.0703	-1.0497	-1.0200	2.9163	2.9293	2.9541
	800	-1.0983	-1.0851	-1.0555	3.1230	3.1315	3.1464
Deviance	20	-1.2136	-1.1274	-0.9916	2.2067	2.2838	2.4435
	40	-1.3046	-1.2434	-1.1360	2.4880	2.5404	2.6494
	80	-1.3616	-1.3143	-1.2299	2.7427	2.7789	2.8441
	200	-2.4272	-1.3969	-1.3395	3.0526	3.0746	3.1124
	400	-2.4613	-1.4402	-1.3994	3.2672	3.2798	3.3042
	800	-1.4900	-1.4764	-1.4461	3.4696	3.4779	3.4925

Table 5.4
Percentile of min and max of residuals in 10,000 trials with 60% censoring

Residuals		**Percentile of min residual values**			**Percentile of max residual values**		
	N	**90**	**95**	**99**	**90**	**95**	**99**
Log-odds	20	-0.9256	-0.8229	-0.6389	3.4042	3.6230	4.1163
	40	-1.0394	-0.9518	-0.8031	4.132	4.3124	4.6982
	80	-1.1044	-1.0396	-0.9293	4.8164	4.9529	5.2164
	200	-1.1993	-1.1539	-1.0727	5.7104	5.7954	5.9438
	400	-1.2475	-1.2145	-1.1500	6.3925	6.4502	6.5499
	800	-1.2845	-1.2594	-1.2202	7.0809	7.1229	7.1825
Normal Deviate	20	-0.5281	-0.4672	-0.3579	1.8499	1.9430	2.1433
	40	-0.5955	-0.5436	-0.4554	2.1498	2.2196	2.3644
	80	-0.6338	-0.5956	-0.5303	2.4075	2.4566	2.5493
	200	-0.6894	-0.6628	-0.6151	2.7163	2.7443	2.7925
	400	-0.7175	-0.6983	-0.6605	2.9343	2.9522	2.9828
	800	-0.7390	-0.7244	-0.7016	3.1415	3.1538	3.1711
Deviance	20	-0.8808	-0.8101	-0.6780	2.2151	2.3076	2.5062
	40	-0.9570	-0.8985	-0.7963	2.5126	2.5816	2.7246
	80	-0.9996	-0.9571	-0.8833	2.7671	2.8155	2.9067
	200	-1.0606	-1.0316	-0.9789	3.0709	3.0984	3.1457
	400	-1.0911	-1.0703	-1.0291	3.2848	3.3023	3.3323
	800	-1.1144	-1.0987	-10739	3.4877	3.4998	3.5167

estimates of the analysis. In generalized linear regression, influential observations in the data set are usually detected using the delete-case method. The method computes the changes in each regression parameter by dropping an individual one at a time and refitting the model again. Here, the same procedure is introduced in the Cox PHM, where the changes in parameter coefficient obtained when the *j*th subject is eliminated are observed. Any large changes in the values of parameter coefficient suggest that the observations are influential. This method is the standard method that is commonly used to identify influential observations in survival model. For further references see also Therneau *et al.* (1990), Cain and Lange (1984), Reid and Crépeau (1985), Storer and Crowley (1985) Collet (2003) and Wang *et al.* (2006).

Let $\hat{\beta}$ be the true regression coefficients without dropping any observations from the data set, and $\hat{\beta}_{(j)}$ be the regression coefficients produced by eliminating *j*th observation. Therefore the equation of standard deletion-case method is defined as $\Delta\hat{\boldsymbol{\beta}}_{(j)} = \left|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(j)}\right|$. In the process of dropping the observation one at a time, the changes on the inferences made by the model are monitored. The graphical display can be viewed when plotting the $\Delta\hat{\boldsymbol{\beta}}_{(j)}$ values versus rank order of variable values to screen the influence observations as illustrated in Figure 5.9 based on *pf* factors of prostate cancer data.

From Figures 5.9, every extreme point is investigated and is considered to be a candidate of influence observation. For example, point in a red circle is far from the other points. In this case some of the influence observations may be identified as outliers. However, this approach is very time-consuming but is still used as the standard approach in detecting influential observations in survival analysis.

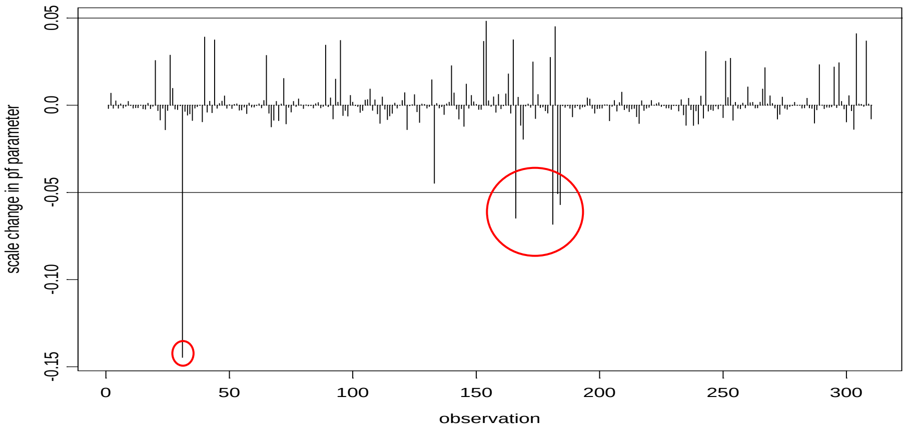


Figure 5.9
The *dfbetas* plot on *pf* factors in prostate cancer data

5.4 Real data analysis - Prostate Cancer

We consider the prostate cancer data (Andrews and Herzberg, 1985) as an illustration. Only 310 patients were considered in this example. The data consist of patients of age below 75 years old with prognostic factors summarized as follows:

- treatment (rx; $\leq$ 0.2mg diethylstilbestrol vs. > 0.2mg)
- weight index: weight (kg) – height (cm) + 200 (wt; < 100 vs. $\geq$ 100)
- performance rating (pf; normal vs. limitation of activity),
- serum haemoglobin (hg; < 12g/ml vs. $\geq$12g/ml)
- size of primary lesion (sz; < 30cm^2 vs. $\geq$ 30cm^2)
- gleason stage/grade category: combined index of tumour stage and histological grade (sg; $\leq$10 vs. > 10)
- history of cardiovascular disease (HX; no vs. yes).

Table 5.5 gives the descriptive values of censored and uncensored observation in the data. In general, censored observations have longer survival times but smaller deviation from the survival mean. Patients with survival time equal to *0* indicate they die less than 1 month after entering the study. We have modeled the data using the Cox PHM. Here, we perform the outliers and influential observation detection procedure discussed in this chapter.

Table 5.5
Descriptive analysis of prostate cancer data

Survival Times	**Overall**	**Censored**	**Uncensored**
Min	0	51	0
Median	39.5	64	24
Mean	38.8	62.7	26.8
1st Qu.	18	57	11
3rd Qu.	59.7	68	39
Max	76	76	74
Std. Dev.	23.2	7.5	18.6

5.4.1 Outliers Detection

For detecting outliers, we use the cut points ±1.96 for normal deviate r_{Ni} residuals and deviance r_{Di} residuals, and ±3.66 on log-odds r_{Li} residuals with respect to level of significant $\alpha = 0.025$.

Figure 5.10 gives the r_{Di} versus prognostic index plot for prostate cancer data. If cut points used are ±1.96, there are 22 patients that have r_{Di} values exceeding the cut points. By analyzing the profile of the 22 patients as shown in Table 5.6, it is found that the said individuals do not follow the definition of outliers given by Therneau *et al.* (1990). For example, for patients 79, 127, and 366 with short survival time, they are expected to die early with unfavorable prognostic factors which are >100 of weight index and $\geq$12g/ml of serum haemoglobin.

Table 5.6
Profile of individuals defined as outlier identified by deviance residual

patient (*i*)	**MF**	**status**	**prognostic factors**							$S(t_i)$	r_{D_i}
			wt	**pf**	**HX**	**hg**	**sz**	**sg**	**rx**		
151	0	1	0	1	1	0	0	1	1	0.9549	2.0596
173	0	1	0	0	0	0	0	1	0	0.9614	2.1324
202	0	1	1	0	0	0	1	1	1	0.9543	2.0538
231	0	1	0	0	1	1	0	1	0	0.9534	2.0444
382	0	1	1	0	1	1	1	1	1	0.9449	1.9630
437	0	1	1	0	1	0	0	0	1	0.9845	2.5186
492	0	1	0	1	1	1	0	1	1	0.9643	2.1686
116	1	1	1	0	1	1	0	1	0	0.9471	1.9834
200	1	1	0	0	1	1	0	0	1	0.9751	2.3259
243	1	1	0	1	0	0	0	0	0	0.9686	2.2255
362	1	1	1	0	1	1	0	0	0	0.9737	2.3020
451	4	1	1	0	0	1	0	0	1	0.9793	2.4033
414	5	1	0	0	0	1	0	0	1	0.9650	2.1768
79	6	1	1	0	1	1	0	0	1	0.9479	1.9910
127	6	1	1	0	0	1	0	0	0	0.9519	2.0298
366	8	1	1	0	0	1	0	0	1	0.9554	2.0656
408	56	0	1	0	1	0	0	1	0	0.1368	-1.9946
41	60	0	0	0	1	1	0	1	0	0.0995	-2.1481
392	64	0	0	0	1	1	0	1	0	0.0851	-2.2200
427	70	0	1	0	1	1	0	1	0	0.1248	-2.0400
50	72	0	0	1	1	0	0	1	0	0.0181	-2.8331
293	76	0	0	0	1	0	0	1	1	0.0592	-2.3774

Figure 5.11 gives the r_{L_i} versus prognostic index plot for prostate cancer data. It shows that five points exceed the suggested cut points (±3.66). We observe that patients 50, 200, 293, 437, and 451 are identified as outliers listed in Table 5.7 with r_{L_i} values given in column 12. Patients 50 and 293 have long censored survival times. Both are known to have longer survival time although the prognostic factors indicate

that they should have died earlier. However, the other three patients unexpectedly died too early with favorable prognostic.

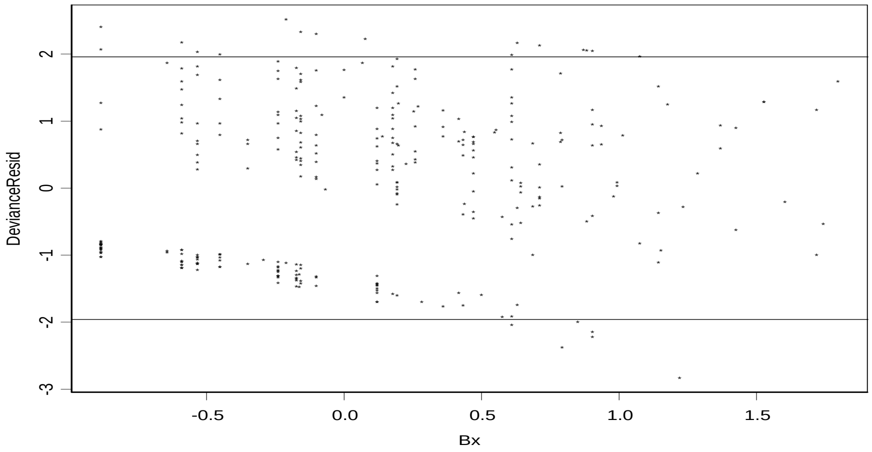


Figure 5.10
r_{Di} vs. prognostic index plot for prostate cancer data

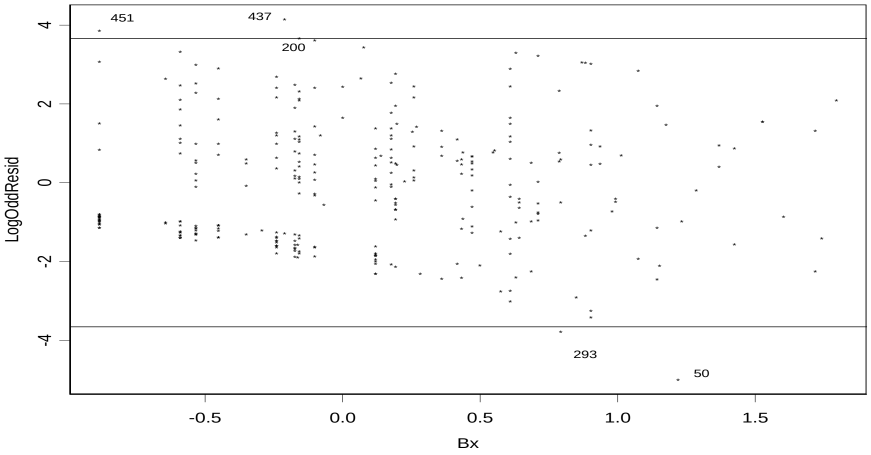


Figure 5.11
r_{L_i} vs. prognostic index plot for prostate cancer data

Figure 5.12 gives the r_{N_i} versus prognostic index for prostate cancer data. It is found that, five points are identified as outliers. We observe that the patients are similar to that detected by the r_{N_i} as listed in Table 5.7. The corresponding values of r_{N_i} are given in column 13.

Table 5.7
Profile of outliers detected by normal deviate and log-odd residuals

patient (*i*)	MF	status	prognostic factors							$S(t_i)$	r_{L_i}	r_{N_i}
			wt	pf	HX	hg	sz	sg	rx			
437	0	1	1	0	1	0	0	0	1	0.9845	4.1481	2.1558
200	1	1	0	0	1	1	0	0	1	0.9751	3.6669	1.9614
451	4	1	1	0	0	1	0	0	1	0.9793	3.8566	2.0395
50	72	0	0	1	1	0	0	1	0	0.0181	-5.0041	-2.4578
293	76	0	0	0	1	0	0	1	1	0.0592	-3.7958	-1.9908

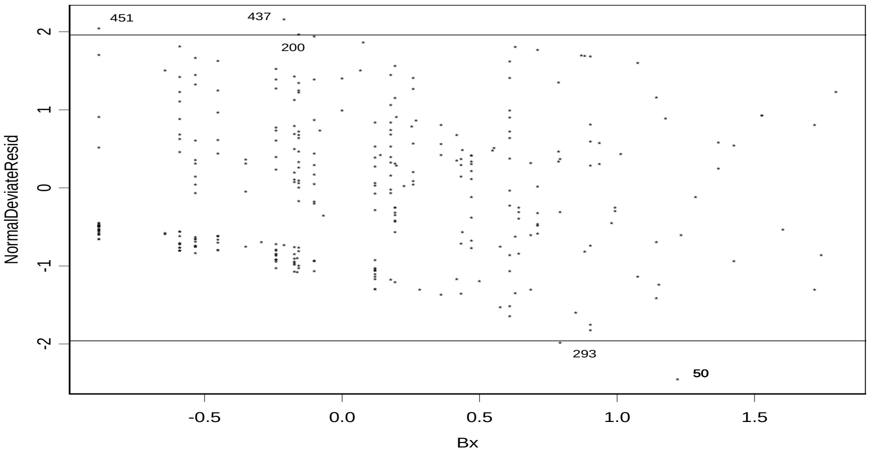


Figure 5.12
r_{N_i} vs. prognostic index plot for prostate cancer data

5.4.2 Influential Observations Detection

The delete-case method can be used to identify the influential observations. Figure 5.13 give the influence plot for each prognostic factor. We observe that there exist points that give large $\left(\Delta_i \boldsymbol{\beta}_{(j)}\right)$ for all factors except *wt*. We note that patients 5, 41, 50, 148, 150, 208, 243, 266, 292, 293, 294, 296, 392, 408 and 477 are classified as influential observations. These 11 patients out of 15 patients (as given in Table 5.8) have long survival times (MF) and known to be alive until the end of study. We observe that patient 243 is unexpectedly die early with early stage of cancer. Patients

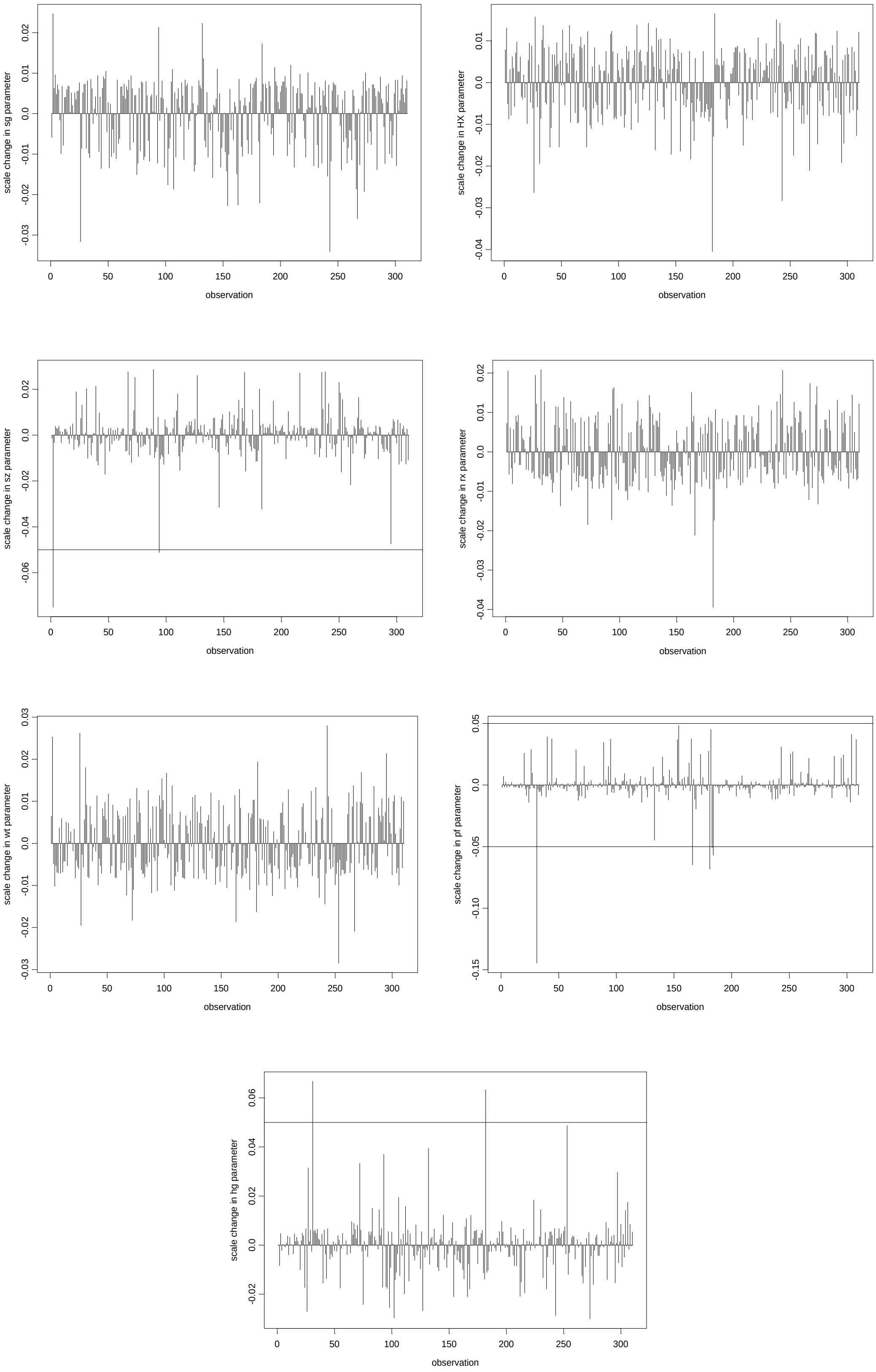


Figure 5.13
Delta-betas for each factor of prostate cancer data

50, 293 and 477 who are identified as outliers in Section 5.4.1 are also detected as influential observation.

Table 5.9 gives the parameter estimate, hazard and global test of PHA for the Cox PHM with and without omitting the influential patients. In general, the parameter estimate, relative hazard and global test show some changes. The changes are not too large except the relative hazard for *sg*, *HX*, *sz*, *pf* and *hg*. The changes are more than 2%. After inspection once again on the data set, there is no typo error. Therefore, further study should be carried out to determine other factors that contribute to the survival of these influential patients. The final model is:

$$h_i\left(t_k\right) \;=\; \exp\,\Big(0.845x_{sg} + 0.683x_{HX} + 0.963x_{sz} - 0.401x_{rx} - 0.446x_{wt} + 0.376x_{pf} - 0.374x_{hg}\Big)\; h_0\left(t_k\right)$$

Table 5.8
Observations considered as influential observations using delete-case method

prognostic factor	sg	HX	Sz	rx	Wt	Hg	pf
patients numbers	41, 392	293	5, 150, 477	293	-	50, 243, 266, 292, 293, 294, 296	50, 148, 208, 293, 408

5.5 Summary

From the analysis, we find several interesting results. Firstly, the normal deviate r_{N_i} and log-odds r_{L_i} residuals are able to detect "meaningful" outliers compared to deviance r_{D_i} residuals. Generally, patients selected by the normal deviate and log-odds residual are a subset of patients selected by the deviance residual. Secondly, we find that the standard normal distribution is not appropriate to be the reference distribution of the deviance residuals. Thirdly, the log-odds and normal deviate residuals do follow their respective sampling distribution for up to 20% of censoring only. Fourthly, the normal deviate and log-odds residual are able to identify outliers corresponding to the patients who die "too early" or "too soon". Finally, the influential observation in this study is detected using the delete-case method. The method is able to identify influential observation when applied on the prostate cancer data.

Table 5.9
Cox PHM result

Individuals omitted	**Variables**	***β*-parameters estimate**	**Relative hazard**	**Global test**
None	sg	0.711	2.035	0.754
	HX	0.433	1.541	
	sz	0.816	2.261	
	rx	-0.350	0.705	
	wt	-0.294	0.746	
	pf	0.077	1.080	
	hg	-0.240	0.786	
Influential observations (listed in Table 5.8)	sg	0.845	2.329	0.640
	HX	0.683	1.979	
	sz	0.963	2.620	
	rx	-0.401	0.669	
	wt	-0.446	0.640	
	pf	0.376	1.456	
	hg	-0.374	0.688	