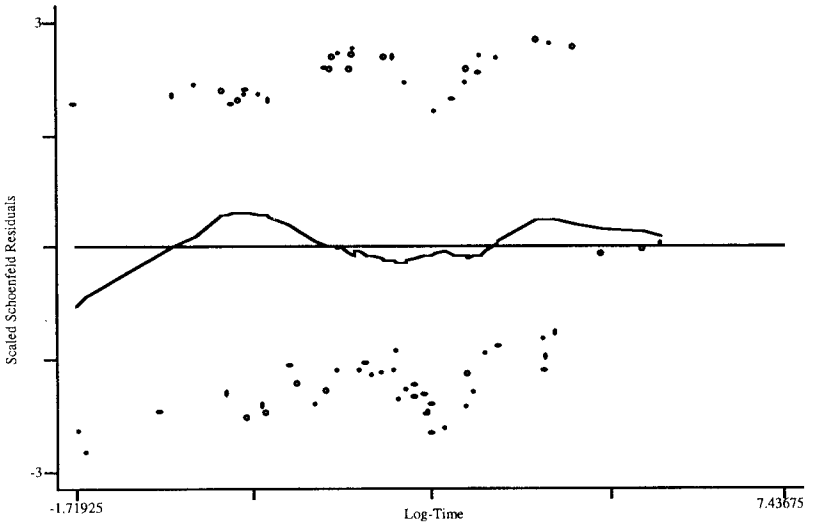
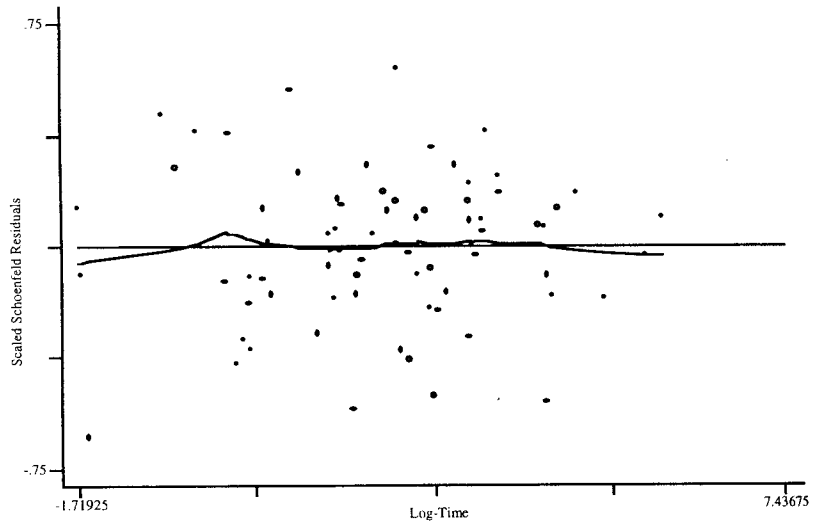


(a) Scaled Schoenfeld Residuals, Dichotomous Covariate



(b) Scaled Schoenfeld Residuals, Continuous Covariate

Figure 6.1 Graphs of the scaled Schoenfeld residuals and their lowess smooth obtained from main effects model in Table 6.1. Zero line is drawn for reference.

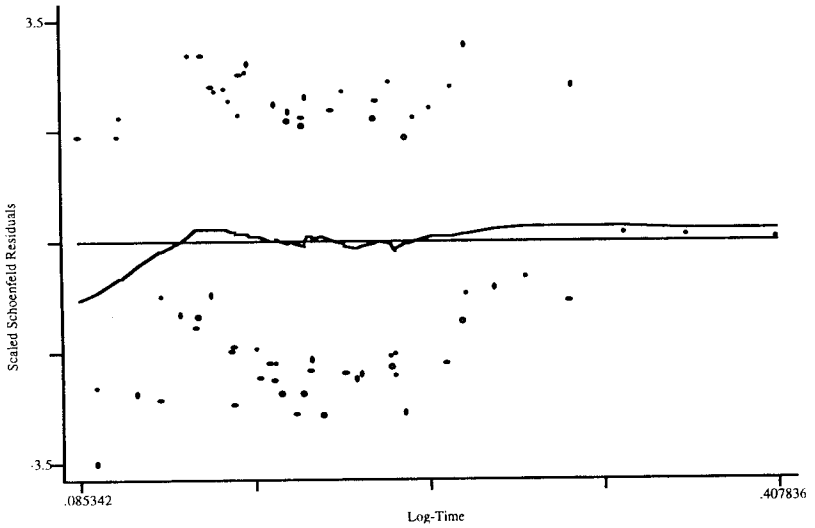
Table 6.2 Estimated Coefficients, Standard Errors, *z*-Scores, Two-Tailed *p*-Values and 95% Confidence Intervals for Models with a Nonproportional Hazard Function in the Continuous Covariate, *x* (*n* = 100 with 30% Censoring)

Variable	Coeff.	Std. Err.	*z*	*P*>|*z*|	95% CIE
d	0.561	0.256	2.19	0.028	0.059, 1.062
x	0.444	0.050	8.83	<0.001	0.343, 0.539
d	0.540	0.267	2.02	0.043	0.015, 1.063
x	0.539	0.070	7.68	<0.001	0.401, 0.676
d×ln(*t*)	0.498	5.149	0.10	0.923	–9.594, 10.590
x×ln(*t*)	2.337	1.037	2.25	0.024	0.304, 4.368

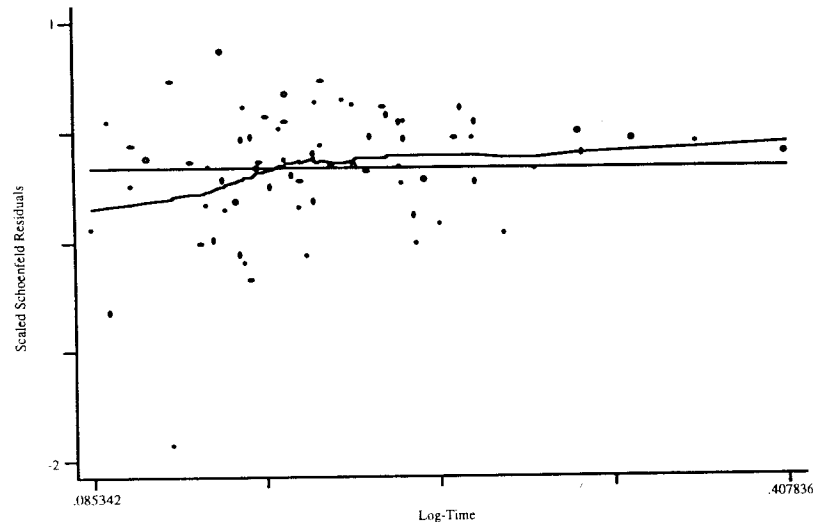
ficient for $d \times \ln(t)$. The Wald test for the coefficient for the $x \times \ln(t)$ term is not significant, suggesting that the model has a proportional hazard in the continuous covariate. The value of the partial likelihood ratio test for the addition of the two interaction variables is $G = 11.355$ and, with 2 degrees-of-freedom, the p-value is 0.003. The polygon connecting the smoothed scaled Schoenfeld residuals in Figure 6.3a shows a strong initial positive slope that levels off. The shape of this plot suggests that the dichotomous covariate may be an important determinant of survival initially, but not later in the follow-up period. The polygon of the smoothed scaled Schoenfeld residuals for the continuous covariate essentially has a zero slope, supporting the lack of significance of the interaction with time that was seen in Table 6.3.

These examples demonstrate the utility of the two-step procedure for assessing the proportional hazards assumption: (1) add the covariate by log-time interactions to the model and assess their significance using the partial likelihood ratio test, score test or Wald test and (2) plot the scaled and smoothed scaled Schoenfeld residuals obtained from the model without the interactions terms. The results of the two steps should support each other. Procedures for modeling in the presence of nonproportional hazards are discussed in Chapter 7, when extensions of the proportional hazards model are considered. We now turn to evaluating the model developed in Chapter 5 for the UIS, shown in Table 5.11. We leave evaluation of the model in Table 5.13 as an exercise.

The model shown in Table 5.11 for the UIS is relatively complex in that it contains 10 terms, two of which are interactions and two of which model nonlinear effects of a continuous covariate. As a first step in assessing the proportional hazards assumption, interactions of each main effect with log-time were added to the model, using only NDRUGFP1



(a) Scaled Schoenfeld Residuals, Dichotomous Covariate



(b) Scaled Schoenfeld Residuals, Continuous Covariate

Figure 6.2 Graphs of the scaled Schoenfeld residuals and their lowess smooth obtained from the main effects model in Table 6.2. Zero line is drawn for reference.

Table 6.3 Estimated Coefficients, Standard Errors, z-Scores, Two-tailed p-Values and 95% Confidence Intervals for Models with a Nonproportional Hazard Function in the Dichotomous Covariate, d (n = 100 with 30% Censoring)

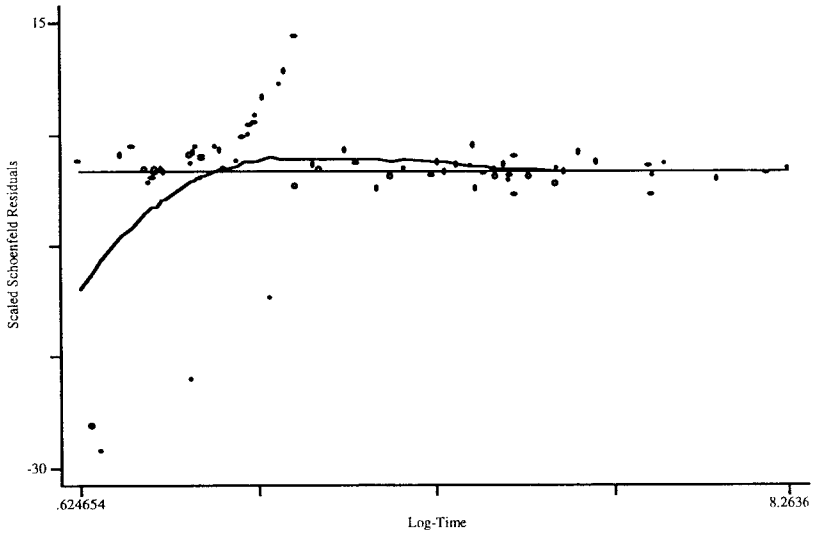
Variable	Coeff.	Std. Err.	z	$P>\|z\|$	95% CIE
d	4.238	0.608	6.98	<0.001	3.046, 5.430
x	0.171	0.032	5.40	<0.001	0.009, 0.133
d	8.977	1.884	4.77	<0.001	5.285, 12.669
x	0.185	0.034	5.54	<0.001	0.120, 0.251
$d \times \ln(t)$	2.709	0.837	3.24	0.001	1.069, 4.350
$x \times \ln(t)$	0.009	0.018	0.53	0.598	–0.025, 0.044

for number of drug treatments. Table 6.4 presents the estimated coefficients, standard errors, Wald statistics and p-values for the Wald statistics for the interactions with log-time. The value of the partial likelihood ratio test comparing the model in Table 5.11 to the 17 term model containing the seven interactions with log-time is $G = 5.538$ which, with 7 degrees-of-freedom, yields $p = 0.595$. These results suggests that the model may have proportional hazards in each of the seven covariates.

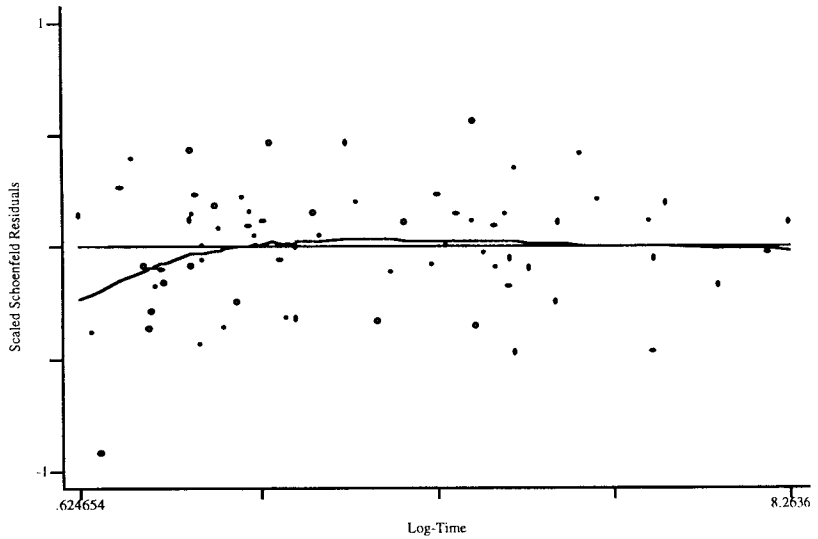
The next step is to examine a plot, similar to those in Figures 6.1–6.3, for each of the 10 terms in the model. The plots of the scaled Schoenfeld residuals and the lowess smooths shown in Figure 6.4 support the assumption of proportional hazards for each of the eight covariates shown. That is, each subplot in the figure has slope essentially equal to zero. The only possible exception is for the covariate

Table 6.4 Estimated Coefficients, Standard Errors, z-Scores and Two-Tailed p-Values for the Seven Interactions with Log-Time Added to the Model in Table 5.11 for the UIS (n = 575)

Variable	Coeff.	Std. Err.	z	$P>\|z\|$
AGE×ln(t)	0.002	0.009	0.20	0.838
BECKTOTA×ln(t)	–0.007	0.005	1.38	0.166
NDRUGFP1×ln(t)	–0.016	0.018	0.89	0.375
IVHX_3×ln(t)	–0.030	0.113	0.27	0.791
RACE×ln(t)	0.113	0.125	0.91	0.364
TREAT×ln(t)	0.128	0.100	1.28	0.201
SITE×ln(t)	–0.023	0.114	0.20	0.842



(a) Scaled Schoenfeld Residuals, Dichotomous Covariate



(b) Scaled Schoenfeld Residuals, Continuous Covariate

Figure 6.3 Graphs of the scaled Schoenfeld residuals and their lowess smooth obtained from the main effects model in Table 6.3. Zero line is drawn for reference.

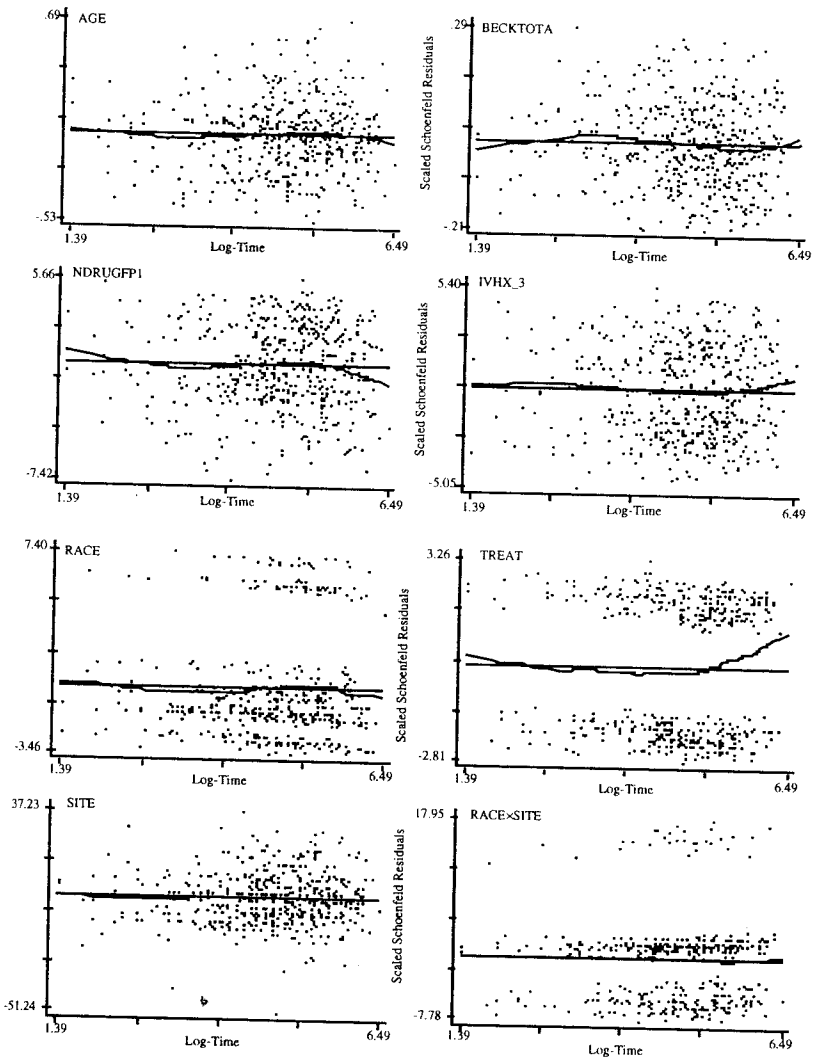


Figure 6.4 Graphs of the scaled Schoenfeld residuals and their lowess smooth obtained from the model in Table 5.11 for covariates (top left to bottom right): AGE, BECKTOTA, NDRUGFP1, IVHX_3, RACE, TREAT, SITE and RACE×SITE. The zero line is drawn for reference.

TREAT (third row, right plot). This plot could be interpreted to mean that the effect of the longer treatment (TREAT = 1) is most pronounced in the earlier and later periods of follow-up. However, we will not consider the possible departure from proportional hazards to be significant, since the Wald test of the treatment by log-time interaction is not significant, $p = 0.201$. We reexamine the effect of treatment in Chapter 7, when covariates that vary with time are discussed in detail. The plots for NDRUGFP2 and AGExSITE (the two terms in the model in Table 5.11 that are not shown in Figure 6.4), also support the proportional hazards assumption.

The two-step procedure for assessing proportional hazards yields results that support this assumption for the 10-term model for the UIS shown in Table 5.11. We now consider the evaluation of the subject-specific diagnostic statistics for leverage and influence.

6.4 IDENTIFICATION OF INFLUENTIAL AND POORLY FIT SUBJECTS

Another important aspect of model evaluation is a thorough examination of regression diagnostic statistics to identify which, if any, subjects: (1) have an unusual configuration of covariates, (2) exert an undue influence on the estimates of the parameters, and/or (3) have an undue influence on the fit of the model. Statistics similar to those used in linear and logistic regression are available to perform these tasks with a fitted proportional hazards model. There are some differences in the types of statistics used in linear and logistic regression and proportional hazards regression, but the essential ideas are the same in all three settings.

Leverage is a diagnostic statistic that measures how "unusual" the values of the covariates are for an individual. In some sense it is a residual in the covariates. In linear and logistic regression leverage [see Hosmer and Lemeshow (1989), Kleinbaum, Kupper, Muller and Nizam (1998), and Ryan (1997)] is calculated as the distance of the value of the covariates for a subject to the overall mean of the covariates. It is proportional to $(x-\bar{x})^2$. The leverage values in these settings have nice properties in that they are always positive and sum over the sample to the number of parameters in the model. While it is technically possible to break the leverage into values for each covariate, this is rarely done in linear and logistic regression. Leverage is not quite so easily defined nor does it have the same nice properties in proportional hazards regression. This is due to the fact that subjects may appear in multiple risk sets and thus may be present in multiple terms in the partial likelihood.

The score residuals defined in (6.16) and (6.17) form the nucleus of the proportional hazards diagnostics. The score residual for the ith subject on the kth covariate, see (6.14), is a weighted average of the distance of the value, x_{ik}, to the risk set means, $x_{w_i k}$, where the weights are the change in the martingale residual, $dM_i(t_j)$. The net effect is that, for continuous covariates, the score residuals have the linear regression leverage property that the further the value is from the mean the larger the score residual is, but "large" may be either positive or negative. Thus, the score residuals are sometimes referred to as the leverage or partial leverage residuals.

The graphs of the score residuals for the covariates AGE, BECKTOTA, NDRUGFP1 and the AGE × SITE interaction obtained from the fitted model in Table 5.11 are shown in Figure 6.5. These four terms were chosen because they are the continuous variables in the fitted model and are therefore most amenable to having their score residuals examined graphically. The graphs for the dichotomous covariates are less interesting in that all the values fall on two vertical bands at zero and one, the two covariate values.

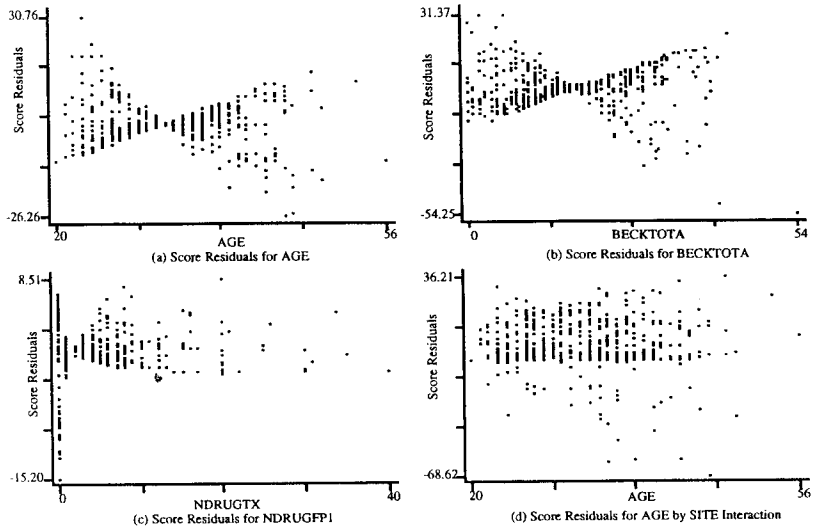


Figure 6.5 Graphs of the score residuals computed from the model in Table 5.11 for (a) AGE, (b) BECKTOTA, (c) NDRUGFP1 and (d) AGE × SITE Interaction.

The score residuals for AGE in Figure 6.5a display the fan shape expected, being smallest near the mean age of 32 and increasing in absolute value for ages increasingly older or younger than 32. The purpose of the plot is to see whether there are subjects whose ages yield unexpectedly large values. This would be seen in the graph as a point lying well away from the others in the plot. In Figure 6.5a there is one point in the top left and there are two in the bottom right that fall a bit away from the rest of the points. However, the distance between these points and the others is not striking. The two oldest subjects, ages 53 and 56, have score residuals that are well within the observed range of values. Thus, we conclude that there are no high leverage values for age.

The score residual values for BECKTOTA are plotted in Figure 6.5b. Recall that when we examined the scale of this covariate in Chapter 5 there was evidence in one of the plots, Figure 5.3, of some non-linearity, but it was attributed to a few high values. These same values appear in the bottom right corner of Figure 6.5b as high leverage points that fall well away from all the other points. For the moment we do nothing more than note this fact.

The covariate, NDRUGTX, entered the model non-linearly with two terms. The plot of the score residuals for the first term, NDRUGFP1, is shown in Figure 6.5c. The plot of the score residuals for the second term, NDRUGFP2, is nearly identical in appearance to Figure 6.5c and is not presented. The fan shape is not quite as apparent because the mean number of drug treatments is about 5 while the maximum number is 40. We chose to plot the score residuals versus the number of drug treatments, rather than the transformation, NDRUGFP1, in order to more easily identify values associated with large residuals. The plot of the residuals versus NDRUGFP1 is essentially the mirror image of this plot since the transformation is the inverse of the variable. The vertical line of values at the left of the plot corresponds to the residuals for subjects with zero previous treatments. The only possible high leverage point is the one on the bottom left for a subject with zero previous treatments, but this value is not too distant from the other values. Thus, we conclude that none of the score residuals for NDRUGFP1 are abnormally large.

The score residuals for the AGE $\times$ SITE interaction covariate are plotted versus AGE in Figure 6.5d. The plot does not have the fan shape seen in Figure 6.5a since, at any age, there is a mix of subjects from the two sites. There are a few points in the bottom of the plot that fall a bit away from the others. However, the plot tends to drift down

with no distinct break, so we conclude that none of the points have large residuals.

In summary, the plots in Figure 6.5 have shown that, except for the two subjects with the highest values for BECKTOTA, there are no strikingly large score residuals. Graphs and histograms, not shown, of the score residuals for the dichotomous covariates in the model did not yield any strikingly large values.

In linear and logistic regression, high leverage is not necessarily something to be concerned about. How high leverage contributes to a measure of the influence that a covariate value has on the estimate of a coefficient is of concern. The same is true in proportional hazards regression. To examine influence in the proportional hazards setting, we need statistics analogous to Cook's distance in linear regression. The purpose of Cook's distance is to obtain an easily computed statistic that approximates the change in the value of the estimated coefficients if a subject is deleted from the data. This is denoted as

$$\Delta\hat{\beta}_{ki} = \hat{\beta}_k - \hat{\beta}_{k(-i)}, \tag{6.22}$$

where $\hat{\beta}_k$ denotes the partial likelihood estimator of the coefficient computed using the entire sample of size n and $\hat{\beta}_{k(-i)}$ denotes the value of the estimator if the ith subject is removed. Cain and Lange (1984) show that an approximate estimator of (6.22) is the kth element of the vector of coefficient changes

$$\Delta\hat{\boldsymbol{\beta}}_i = \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}\right) = \hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\hat{\mathbf{L}}_i, \tag{6.23}$$

where $\hat{\mathbf{L}}_i$ is the vector of score residuals, (6.17), and $\hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ is the estimator of the covariance matrix of the estimated coefficients. These are commonly referred to as the *scaled score residuals* and their values may be obtained from some software packages, for example, SAS and S-PLUS.

Graphs of the scaled score residuals, (6.23), are presented in Figure 6.6 for the covariates whose score residuals were graphed in Figure 6.5. The plots in Figures 6.5 and 6.6 are quite similar in appearance, but the scaling has enhanced the fan shape. The points seen in the top left of Figure 6.5a and the bottom right of Figure 6.5b are more noticeable in Figures 6.6a and 6.6b, confirming that they may exert an undue influence on the estimates of the coefficients.

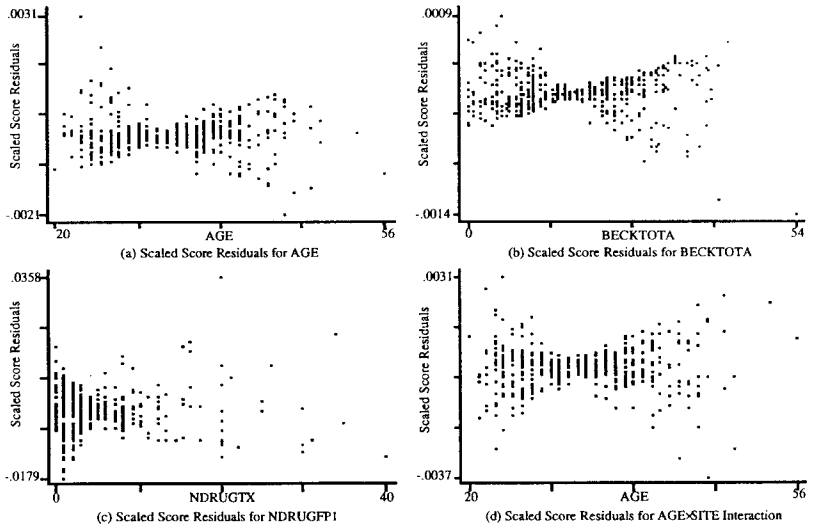


Figure 6.6 Graphs of the scaled score residuals computed from the model in Table 5.11 for (a) AGE, (b) BECKTOTA, (c) NDRUGFP1 and (d) AGE × SITE interaction.

One point in the top and middle of Figure 6.6c lies well away from the other scaled score residuals for subjects with the same number of drug treatments. This subject has some potential for influence on the coefficient for NDRUGFP1. We examine the effect this subject, as well as others, have on the model later in this section.

The plot in Figure 6.6d is much more distinctly fan shaped than its counterpart in Figure 6.5d, and none of the points seem to fall well away from the others. There are no distinct breaks in the points, they just slowly drift out, so nothing is noted in this plot for further examination. Plots of the scaled score residuals for the dichotomous covariates in the model revealed no points of potential high influence and thus are not shown.

Cook's distance in linear and logistic regression may be used to provide a single overall summary statistic of the influence a subject has on the estimators of all the coefficients. The overall measure of influence is

$$\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(-i)}\right)'\left[\widehat{\mathrm{Var}}\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(-i)}\right),$$

and using (6.23) it may be approximated using

$$\begin{aligned} ld_i &= \left(\Delta\hat{\boldsymbol{\beta}}_i\right)'\left[\widehat{\mathrm{Var}}\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\left(\Delta\hat{\boldsymbol{\beta}}_i\right) \\ &= \left(\hat{\mathbf{L}}_i\right)'\left[\widehat{\mathrm{Var}}\left(\hat{\boldsymbol{\beta}}\right)\right]\left[\widehat{\mathrm{Var}}\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\left[\widehat{\mathrm{Var}}\left(\hat{\boldsymbol{\beta}}\right)\right]\left(\hat{\mathbf{L}}_i\right) \end{aligned}$$

so

$$ld_i = \left(\hat{\mathbf{L}}_i\right)'\left[\widehat{\mathrm{Var}}\left(\hat{\boldsymbol{\beta}}\right)\right]\left(\hat{\mathbf{L}}_i\right). \tag{6.24}$$

The statistic in (6.24) has been shown by Pettitt and Bin Daud (1989) to be an approximation to the amount of change in the log partial likelihood when the ith subject is deleted. In this context the statistic is called the likelihood displacement statistic, hence the rationale for labeling it ld in (6.24). Thus

$$ld_i \cong 2\left[L_p\left(\hat{\boldsymbol{\beta}}\right)-L_p\left(\hat{\boldsymbol{\beta}}_{(-i)}\right)\right]. \tag{6.25}$$

Another form of the likelihood displacement statistic is obtained from a matrix form of (6.24). In particular let $\hat{\mathbf{L}}$ denote the n by p matrix whose ith row is $\hat{\mathbf{L}}_i'$, see (6.17), and let the n by n matrix of scaled score residuals be

$$\hat{\mathbf{L}}\left[\widehat{\mathrm{Var}}\left(\hat{\boldsymbol{\beta}}\right)\right]\hat{\mathbf{L}}'. \tag{6.26}$$

When the matrix in (6.26) is broken into its eigenvalues and associated eigenvectors, the n elements in the eigenvector associated with the largest eigenvalue are called the "l-max" statistics and are denoted lm_i for the ith subject. Since both ld_i and lm_i are overall summary statistics, we feel it makes the most sense to plot them versus another summary statistic. The one we like to use is the martingale residual. Other possible choices are the estimated survival probability or estimated cumulative hazard function. We feel that plots of these values against something like a study identification code or case number may be somewhat useful in locating large values in small data sets but provide little additional information about the subject. An additional enhancement that aids in the interpretation of the plot is to use a different symbol for the two values of the censoring variable.

Plots of the values of the likelihood displacement and l-max statistics versus the martingale residuals are shown in Figures 6.7a and 6.7b, respectively. Both plots have the same asymmetric "cup" shape with the bottom of the cup at zero. In linear and logistic regression, the influence diagnostic, Cook's distance, is a product of a residual measure and leverage. While the same concise representation does not hold in proportional hazards regression, it is approximately true in the sense that an influential subject will have a large residual and/or leverage. Thus, the largest values of both the likelihood displacement and l-max statistic form the sides of the cup and correspond to poorly fit subjects (ones with either large negative or positive martingale residuals).

Examining the plots in Figure 6.7a, we find that there is a group of four points lying well away from the others in the top left corner of the plot, with two other points slightly below this cluster. Five of the six subjects have censored survival times and all have martingale residuals less than about –2.0. Figure 6.7b is not quite as cup-shaped as Figure 6.7a, but the four points in the top left of each figure correspond to the same subjects. The principal difference in the two figures is that the two subjects on the right edge of Figure 6.7b correspond to a cluster with the next largest values of the l-max statistics, whereas in Figure 6.7a these same two subjects do not have large values of the likelihood displacement statistic. Thus the two statistics, while similar for the extreme values, do identify different subjects in the mid range. From this point of view it makes sense, in an applied setting, to look at both statistics to locate those subjects with large values on both or only one of the statistics.

In summary, use of the plots of the diagnostic statistics for change in individual coefficients identified four possible subjects whose effect on the model should be checked in more detail: one for AGE in Figure 6.6a, two for BECKTOTA in Figure 6.6b and one for NDRUGFP1 in Figure 6.6c. Four to six subjects were identified in Figures 6.7a and 6.7b as having extreme values for the summary change statistics, likelihood displacement and l-max. These subjects are possibly different from the ones previously identified in Figure 6.6. We emphasize "possibly different" because the summary measures take into account values of all the covariates. A subject extreme on only one covariate may be near enough to the middle for the others that an extreme value of the likelihood displacement or l-max statistic would not be generated.

The next step in the modeling process is to identify explicitly the subjects with the extreme values, refit the model deleting these subjects, and calculate the change in the individual coefficients. The final decision on the continued use of a subject's data to fit the model will depend on the observed percent change in the coefficients that results from deleting the subject's data and, more importantly, the clinical plausibility of that subject's data.

Deleting the subjects with extreme values in the change in coefficient diagnostic for AGE and NDRUGFP1 individually did not produce marked changes in the coefficients. However, deletion of the two subjects with the extreme values of the diagnostic for change in the BECKTOTA coefficient yielded a model in which the BECKTOTA coefficient was 33.5 percent larger than the coefficient in Table 5.11.

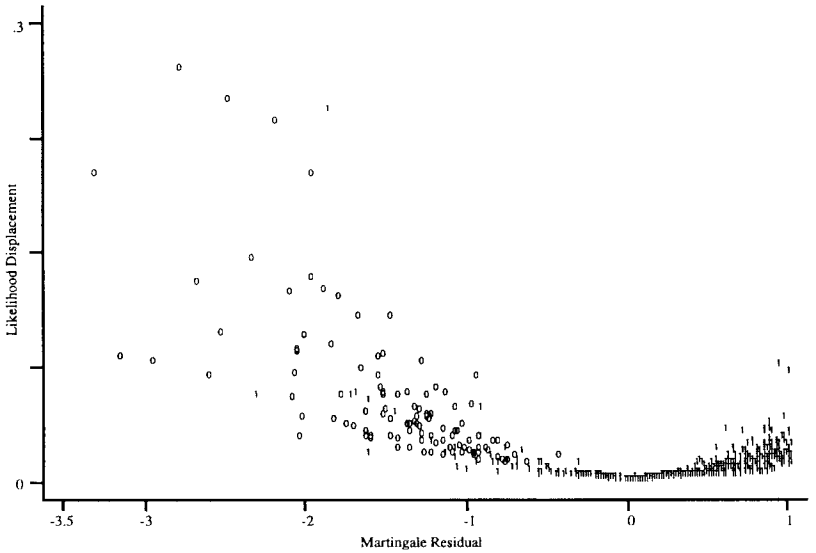
Through a process of deleting and refitting models, we determined that deletion of the four subjects with the most extreme values of the likelihood displacement or of the l-max statistics yielded models with important changes in several coefficients. The deletion of the two subjects with the next largest values of either statistic did not produce additional important changes in the coefficients.

Further examination of the data showed that the two subjects with the extreme values for change in the BECKTOTA coefficient were not among the four with the extreme values of either summary change measure. The model obtained by deleting six subjects (four [based on l-max] and two others [based on BECKTOTA]) had a coefficient for BECKTOTA that increased by 60.8 percent, a coefficient for the RACE $\times$SITE interaction that increased by 33.6 percent, a coefficient for the AGE $\times$ SITE interaction that increased by 20.3 percent, and a coefficient for SITE that increased by 14.9 percent. All other coefficients changed by less than 9 percent. In particular, the coefficient for TREAT increased by only 6.3 percent. At this point we reviewed the data for these six most influential subjects. We felt that only the value of 54 for BECKTOTA was a bit unusual, but not too extreme. In the end, therefore, we decided to keep all subjects in the data set.

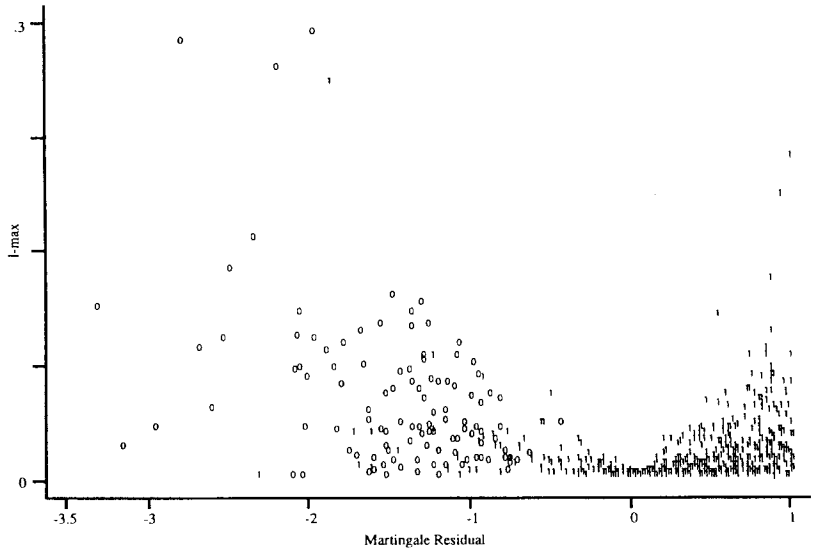
We leave as an exercise the fitting of the models with the specified subjects deleted and computation of the reported percent change in coefficients. We remind the reader that we calculate the percent change in a coefficient as

$$\Delta\hat{\beta}\% = 100\left(\hat{\beta}_{\text{reduced}} - \hat{\beta}_{\text{all}}\right)\Big/\hat{\beta}_{\text{all}},$$

where $\hat{\beta}_{\text{all}}$ stands for the estimate of the coefficient from the model with no subjects deleted and $\hat{\beta}_{\text{reduced}}$ stands for the estimate of the coefficient from the model with subjects deleted.



(a) Likelihood Displacement or Cook's Distance Statistic



(b) l-max Statistic

Figure 6.7 Graphs of the likelihood displacement or Cook's distance statistic and maximum eigenvalue or l-max statistic computed from the model in Table 5.11 versus the martingale residual (0 = censored, 1 = uncensored).

In summary, we feel that it is important to examine plots of the score residuals, scaled score residuals, the likelihood displacement statistic and the l-max statistic. The first two statistics are useful for identifying subjects with high leverage or who influence the value of a single coefficient. The latter two provide useful information for assessing influence on the vector of coefficients. Each statistic portrays an important aspect of the effect a particular subject has on the fitted model. One always hopes that major problems are not uncovered. However, if the model does display abnormal sensitivity to the subjects deleted, this is a clear indication of fundamental problems in the model and we recommend going back to "square-one" and redoing each step in the modeling process, perhaps with these subjects deleted.

The next step in the modeling process is to compute an overall goodness-of-fit test.

6.5 OVERALL GOODNESS-OF-FIT TESTS AND MEASURES

Until quite recently, all of the proposed tests for the overall goodness-of-fit of a proportional hazards model were difficult to compute in most software packages. For example, the test proposed by Schoenfeld (1980) compares the observed number of events to a proportional hazards regression model-based estimate of the expected number of events in each of G groups that are formed by partitioning the time axis and covariate space. Unfortunately, the covariance matrix required to form a test statistic comparing the observed to expected number of events is quite complex to compute. The test proposed by Lin, Wei and Ying (1993) is based on the maximum absolute value of partial sums of martingale residuals. This test requires complex and time consuming simulations to obtain a significance level. Other tests [e.g., O'Quigley and Pessione (1989) and Pettitt and Bin Daud (1990)] require that the time axis be partitioned and interactions between covariates and interval-specific, time-dependent covariates be added to the model. Overall goodness-of-fit is based on a significance test of the coefficients for the added variables.

Grønnesby and Borgan (1996) propose a test similar to the Hosmer-Lemeshow test [Hosmer and Lemeshow (1989)] used in logistic regression. They suggest partitioning the data into G groups based on the ranked values of the estimated risk score, $\mathbf{x}'\hat{\boldsymbol{\beta}}$. The test is based on the sum of the martingale residuals within each group, and it compares the