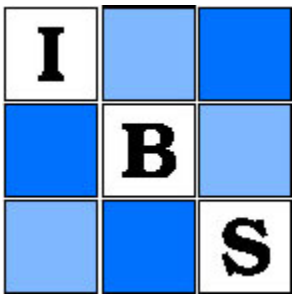


BIOMETRICS 58, 1–12
March 2002

Model-Checking Techniques Based on Cumulative Residuals

D. Y. Lin,[1,*] L. J. Wei,[2,] and Z. Ying[3,***]**

[1]Department of Biostatistics, University of North Carolina,
CB 7420 McGavran-Greenberg, Chapel Hill, North Carolina 27599-7420, U.S.A.
[2]Department of Biostatistics, Harvard University,
677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.
[3]Department of Statistics, 618 Mathematics, Columbia University,
New York, New York 10027, U.S.A.
email: lin@bios.unc.edu
**email:* wei@sdac.harvard.edu
***email:* zying@stat.columbia.edu

SUMMARY. Residuals have long been used for graphical and numerical examinations of the adequacy of regression models. Conventional residual analysis based on the plots of raw residuals or their smoothed curves is highly subjective, whereas most numerical goodness-of-fit tests provide little information about the nature of model misspecification. In this paper, we develop objective and informative model-checking techniques by taking the cumulative sums of residuals over certain coordinates (e.g., covariates or fitted values) or by considering some related aggregates of residuals, such as moving sums and moving averages. For a variety of statistical models and data structures, including generalized linear models with independent or dependent observations, the distributions of these stochastic processes under the assumed model can be approximated by the distributions of certain zero-mean Gaussian processes whose realizations can be easily generated by computer simulation. Each observed process can then be compared, both graphically and numerically, with a number of realizations from the Gaussian process. Such comparisons enable one to assess objectively whether a trend seen in a residual plot reflects model misspecification or natural variation. The proposed techniques are particularly useful in checking the functional form of a covariate and the link function. Illustrations with several medical studies are provided.

KEY WORDS: Generalized linear models; Goodness of fit; Link function; Longitudinal data; Marginal models; Model misspecification; Regression diagnostics; Residual plots; Transformation.

1. Introduction

Regression plays a central role in the statistical analysis of experimental and observational data. With the advancement of computers and software, even sophisticated models are commonly used in applications. Although model misspecification can seriously affect the validity and efficiency of regression analysis, model checking has not become a routine practice, partly because of the lack of suitable tools, especially for complex data structures and nonlinear models.

Residuals, defined as the differences between the observed and fitted values of the response, are highly informative about the aptness of a regression model. If the model is correct, the residuals are centered at zero and the plot of the residuals against any coordinate, such as a covariate or the fitted value, should exhibit no systematic tendency. The appearance of a systematic trend may indicate wrong functional form of the covariate or lack of linearity. However, determination of whether a trend seen in a residual plot reflects model misspecification or natural variation can be quite challenging.

As an illustration, we consider the surgical unit example described in Neter et al. (1996, Section 8.2). The data contain the survival time and several covariates for 54 patients undergoing a particular type of liver operation. After an elaborate model selection process, Neter et al. (1996, Section 9.6) arrived at the following final model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon, \tag{1.1}$$

where Y is the logarithm (with base 10) of the survival time, and X_1, X_2, and X_3 are, respectively, *blood-clotting score*, *prognostic index*, and *enzyme function score*. The estimation results for Neter's model are given in Table 1a.

Figure 1a plots, by the circles, the residuals under (1.1) against X_1. It is difficult to ascertain whether or not a systematic pattern exists in this residual plot. This difficulty reflects a remark made by Cook and Weisberg (1999, p. 337) that "Anomalies can be found in all residual plots if we look hard enough." Similar statements appear in Atkinson (1985, pp. 34–35), McCullagh and Nelder (1989, p. 392), and other texts.

The interpretation of residual plots may be facilitated by

Table 1

Linear regression for the surgical unit data

Covariate	Estimate	SE	Estimate/SE
(a) Neter's Model			
Blood-clotting score	0.0692	0.0041	16.98
Prognostic index	0.0093	0.00038	24.30
Enzyme function score	0.0095	0.00031	31.08
(b) Revised Model			
Log(*blood-clotting score*)	0.396	0.021	18.58
Prognostic index	0.0095	0.00035	26.99
Enzyme function score	0.0096	0.00028	33.76

a suitable smoothing algorithm. Figure 1a also shows, by the gray curve, the *lowess* smooth (Cleveland, 1979) of the raw residuals, which suggests curvature in the plot. "Such smoothed curves must be treated with some caution, however, since the algorithm is quite capable of producing convincing-looking curves from entirely random configurations" (McCullagh and Nelder, 1989, pp. 394–395).

The subjective nature of the aforementioned residual analysis is due to the fact that the variabilities of individual residuals are unknown. By contrast, it is possible to determine the distributions of certain types of aggregates of residuals. In particular, goodness-of-fit tests can be constructed by discretizing the covariate vector and generating nonoverlapping

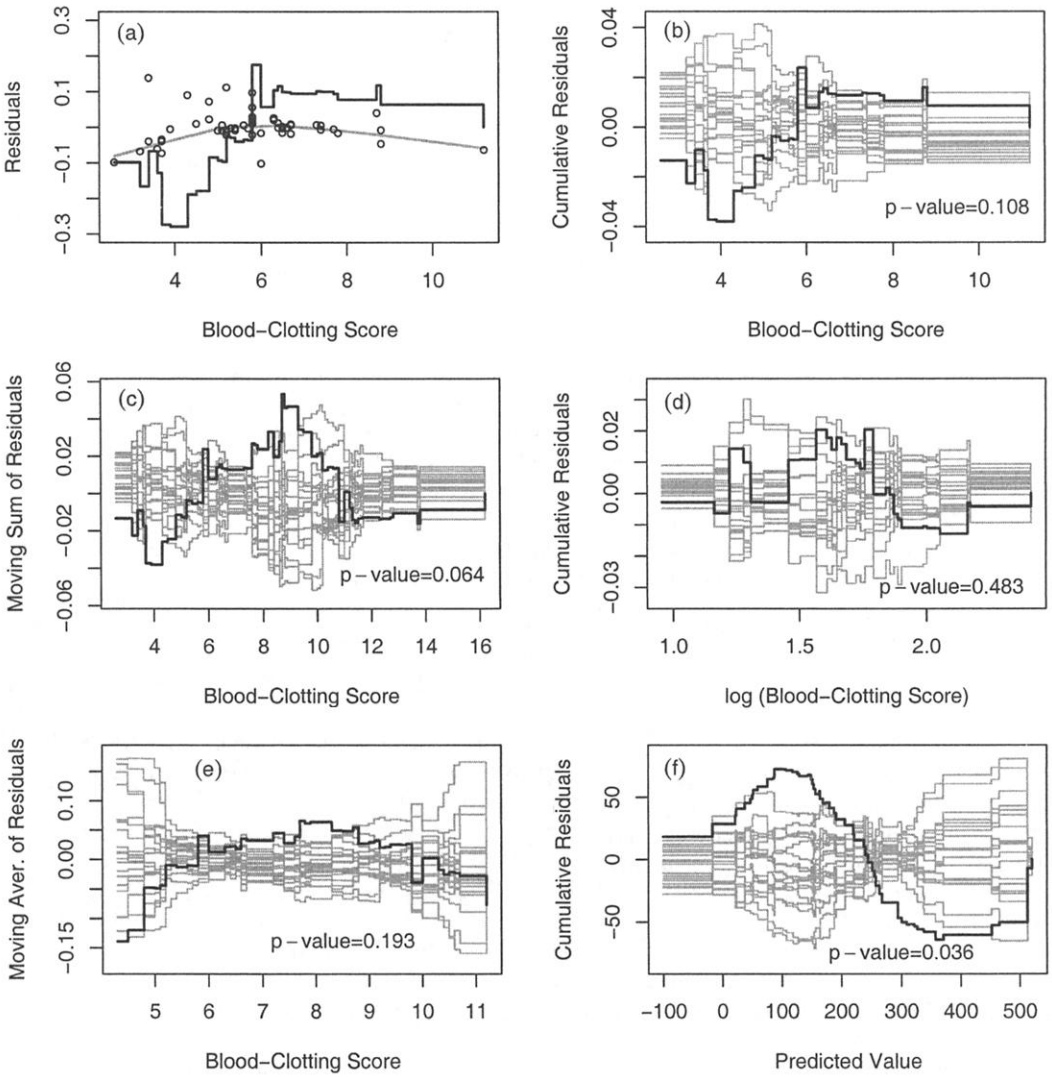


Figure 1. Residual plots for the surgical unit example: (**a**) raw residuals vs. *blood-clotting score* in Neter's model, shown by circles, and the corresponding *lowess* smooth and cumulative sum, shown by the gray and black curves; (**b**) cumulative sum of residuals vs. *blood-clotting score* in Neter's model; (**c**) moving sum of residuals with $b = 5$ vs. *blood-clotting score* in Neter's model; (**d**) cumulative sum of residuals vs. log(*blood-clotting score*) in the revised model; (**e**) moving average of residuals with $b = 4$ vs. *blood-clotting score* in Neter's model; and (**f**) cumulative sum of residuals vs. the predicted value when the survival time is untransformed. In (**b**)–(**f**), the observed pattern is shown by the black curve, and 20 simulated realizations are shown by the gray curves; the P-value pertains to the supremum test with 10,000 realizations.

groups of residuals (e.g., Schoenfeld, 1980; Tsiatis, 1980). Unforunately, the partition of the covariate space is arbitrary and different partitions may result in conflicting conclusions.

In this paper, we present objective model-checking techniques based on the cumulative sums of residuals over certain coordinates. For example, the black curve in Figure 1a is the observed cumulative sum of the residuals over X_1 under model (1.1): for any value x on the horizontal axis, the corresponding value on the vertical axis is the sum of the residuals associated with the covariate values less than or equal to x. Like raw residuals and their smooths, the cumulative sums are centered at 0 if the assumed model is correct. The main motivation for considering cumulative sums is because their natural variations can be ascertained. Specifically, under the null hypothesis of correct model specification, the distribution of the cumulative sum, when regarded as a stochastic process, can be approximated by that of a zero-mean Gaussian process whose realizations can be generated by computer simulation. To assess whether the observed residual pattern reflects anything beyond random fluctuation, we may compare the observed cumulative sum to a number of realizations from the Gaussian process.

For illustration, we plot in Figure 1b the observed cumulative sum of the residuals along with realizations from the zero-mean Gaussian process. (From now on, the cumulative sum is standardized by the square root of the total sample size.) The curves generated from the null distribution tend to be closer to and intersect the horizontal axis more often than the observed curve. The maximum absolute value of the observed cumulative sum is 0.038. Out of 10,000 realizations from the null distribution, only 10.8% have a maximum greater than 0.038. Thus, the P-value for the Kolmogorov-type supremum test is 0.108. These graphical and numerical results suggest that model (1.1), especially the functional form for *blood-clotting score*, may be inappropriate.

In the next section, we explain how to create plots such as Figure 1b and how to identify the nature of model misspecification from such plots. In fact, we develop a class of graphical and numerical methods for checking the mean structure (including the functional forms of covariates and the link function) of any generalized linear model (McCullagh and Nelder, 1989). In Section 3, we provide parallel development for the marginal regression models with dependent observations (e.g., repeated measures in longitudinal studies). Some concluding remarks are given in Section 4.

2. Generalized Linear Models

2.1 *Models and Estimators*

Let Y be the response, and $\mathbf{X} = (1, X_1, \ldots, X_p)'$ be a $(p+1)\times 1$ vector of covariates. We assume that the distribution of Y belongs to the exponential family with density function

$$f_Y(y;\theta,\phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y,\phi)\}, \tag{2.1}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are some specific functions, θ is the parameter of interest, and ϕ is a nuisance parameter. The parameter θ depends on $\mathbf{X}$ through the mean $\mu(\theta)$ of Y in the following way:

$$g\{\mu(\theta)\} = \boldsymbol{\beta}'\mathbf{X}, \tag{2.2}$$

where g is a known link function, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is a $(p+1)\times 1$ vector of unknown regression parameters. In this paper, $\mathbf{V}'$ denotes the transpose of a vector or matrix $\mathbf{V}$.

Suppose that the data consist of n independent replicates of $(Y, \mathbf{X})$. The likelihood score function for $\boldsymbol{\beta}$ takes the form

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} h(\boldsymbol{\beta}'\mathbf{X}_i)\mathbf{X}_i(Y_i - \nu(\boldsymbol{\beta}'\mathbf{X}_i)), \tag{2.3}$$

where $\nu(r) = g^{-1}(r)$, and $h(r) = \partial\{g\circ\mu(r)\}^{-1}/\partial r$. Denote the solution to $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$ by $\widehat{\boldsymbol{\beta}}$. Also, write $\mathcal{I}(\boldsymbol{\beta}) = -\partial\mathbf{U}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}'$. If the model specified by (2.1) and (2.2) is valid, then $n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically zero-mean normal with covariance matrix $\mathbf{A}^{-1}(\boldsymbol{\beta})$, where $\mathbf{A}(\boldsymbol{\beta}) = \lim_{n\to\infty}\{n^{-1}\mathcal{I}(\boldsymbol{\beta})\}$.

A rigorous proof of the aforementioned asymptotic results can be found in Fahrmeir and Kaufmann (1985). By extending their arguments, one can show that, if (2.2) holds but (2.1) fails, then $n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ continues to be asymptotically zero-mean normal, but with covariance matrix $\mathbf{A}^{-1}(\boldsymbol{\beta})\mathbf{B}(\boldsymbol{\beta}) \times \mathbf{A}^{-1}(\boldsymbol{\beta})$, where $\mathbf{B}(\boldsymbol{\beta}) = \lim_{n\to\infty} n^{-1}\,\Sigma_{i=1}^{n}\,\mathbf{U}_i(\boldsymbol{\beta})\mathbf{U}_i'(\boldsymbol{\beta})$, and $\mathbf{U}_i(\boldsymbol{\beta})$ is the ith term on the right-hand side of (2.3). Thus, by using the robust sandwich covariance matrix estimator $\mathcal{I}^{-1}(\widehat{\boldsymbol{\beta}})\,\Sigma_{i=1}^{n}\,\mathbf{U}_i(\widehat{\boldsymbol{\beta}})\mathbf{U}_i'(\widehat{\boldsymbol{\beta}})\mathcal{I}^{-1}(\widehat{\boldsymbol{\beta}})$ for $\widehat{\boldsymbol{\beta}}$, one can make valid inference about $\boldsymbol{\beta}$ in (2.2) even if (2.1) is incorrect (White, 1982). The mean function specified by (2.2) has two major aspects: the functional form for each component of $\mathbf{X}$, and the link function g. We wish to examine these two aspects without imposing (2.1).

2.2 *Model-Checking Techniques*

The residuals are defined by $e_i = Y_i - \nu(\widehat{\boldsymbol{\beta}}'\mathbf{X}_i)$ $(i = 1, \ldots, n)$. To check the functional form for the jth $(j = 1, \ldots, p)$ component of the covariate vector $\mathbf{X}$, we consider the cumulative sum of the e_i over the X_{ji}, i.e.,

$$W_j(x) = n^{-1/2}\sum_{i=1}^{n} I(X_{ji} \le x)e_i, \tag{2.4}$$

where X_{ji} is the jth component of $\mathbf{X}_i$, $x \in \mathcal{R}$, and $I(\cdot)$ is the indicator function. Note that $W_j(\cdot)$ is a step function with possible jumps at the distinct values of the X_{ji}. The black curve in Figure 1b is an example of (2.4). We regard $W_j(x)$ as a stochastic process indexed by x. Su and Wei (1991) and Stute (1997) studied the following process:

$$W(\mathbf{x}) = n^{-1/2}\sum_{i=1}^{n} I(\mathbf{X}_i \le \mathbf{x})e_i, \tag{2.5}$$

where $\mathbf{x} = (x_1, \ldots, x_p)' \in \mathcal{R}^p$, and $I(\mathbf{X}_i \le \mathbf{x}) = I(X_{1i} \le x_1, \ldots, X_{pi} \le x_p)$. The process given in (2.4) is a special case of (2.5) with $x_k = \infty$ for all $k \ne j$.

Define

$$\widehat{W}(\mathbf{x}) = n^{-1/2}\sum_{i=1}^{n}\{I(\mathbf{X}_i \le \mathbf{x}) + \boldsymbol{\eta}'(\mathbf{x};\widehat{\boldsymbol{\beta}})\mathcal{I}^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{X}_i h(\widehat{\boldsymbol{\beta}}'\mathbf{X}_i)\} \times e_i Z_i, \tag{2.6}$$

where

$$\boldsymbol{\eta}(\mathbf{x};\boldsymbol{\beta}) = -\sum_{i=1}^{n} I(\mathbf{X}_i \le \mathbf{x})\partial\nu(\boldsymbol{\beta}'\mathbf{X}_i)/\partial\boldsymbol{\beta},$$

and $(Z_1, \ldots, Z_n)$ are independent standard normal variables that are independent of $(Y_i, \mathbf{X}_i)$ $(i = 1, \ldots, n)$. Under the null hypothesis H_0 that model (2.2) holds, the conditional distribution of $\widehat{W}(\mathbf{x})$ given the data $(Y_i, \mathbf{X}_i)$ $(i = 1, \ldots, n)$ is the same in the limit as the unconditional distribution of $W(\mathbf{x})$ (Su and Wei, 1991). To approximate the null distribution of $W(\mathbf{x})$, we simulate a number of realizations from $\widehat{W}(\mathbf{x})$ by repeatedly generating the normal samples $(Z_1, \ldots, Z_n)$ while fixing the data $(Y_i, \mathbf{X}_i)$ $(i = 1, \ldots, n)$ at their observed values.

We are primarily interested in the graphical examinations of specific model assumptions, such as the functional form of each covariate. As mentioned earlier, $W_j(x)$ is a special case of $W(\mathbf{x})$. Thus, the null distribution of $W_j(x)$ can be approximated through simulating the corresponding zero-mean Gaussian process $\widehat{W}_j(x)$, say. To assess how unusual the observed process $w_j(\cdot)$ is under H_0, one may plot $w_j(\cdot)$ along with a few realizations from the $\widehat{W}_j(\cdot)$ process (see Figure 1b).

To further enhance the objectivity of this graphical technique, one may complement the cumulative residual plot with some numerical values measuring the extremity of $w_j(\cdot)$. Because $W_j(\cdot)$ fluctuates randomly around 0 under H_0, a natural numerical measure is the Kolmogorov-type supremum statistic $S_j \equiv \sup_x |W_j(x)|$. An unusually large observed value s_j would suggest faulty functional form of X_j. The P-value, $\Pr(S_j \geq s_j)$, can be approximated by $\Pr(\widehat{S}_j \geq s_j)$, where $\widehat{S}_j = \sup_x |\widehat{W}_j(x)|$. We estimate $\Pr(\widehat{S}_j \geq s_j)$ by generating a large number, say 1000 or 10,000, of the realizations from $\widehat{W}_j(\cdot)$. For Figure 1b, $s_j = 0.038$, and the P-value is 0.108.

Because it accumulates all the residuals associated with covariate values less than x, $W_j(x)$ tends to be dominated by the residuals associated with small covariate values, and, for relatively large x, $W_j(x)$ does not resemble the conventional plot of raw residuals. These phenomena motivate us to study the following modification of (2.4):

$$W_j(x; b) = n^{-1/2} \sum_{i=1}^{n} I(x - b < X_{ji} \leq x) e_i, \tag{2.7}$$

where b is a positive constant. By definition, $W_j(x; b)$ can take nonzero values for x between $\min_i X_{ji}$ and $\max_i X_{ji} + b$. For $x \leq \min_i X_{ji} + b$, (2.7) is identical to (2.4); for $\min_i X_{ji} + b \leq x \leq \max_i X_{ji}$, (2.7) represents a sum of residuals with blocks of size b; for $x \geq \max_i X_{ji}$, (2.7) pertains to the cumulative sum of residuals from $\max_i X_{ji}$ to $\max_i X_{ji} - b$. We refer to (2.4) and (2.7) as cumulative and moving sums, respectively. Clearly, (2.4) is a special case of (2.7) with $b = \infty$. The null distribution of $W_j(x; b)$ can be approximated by the conditional distribution of

$$\widehat{W}_j(x; b) = n^{-1/2} \sum_{i=1}^{n} \{I(x - b < X_{ji} \leq x) + \boldsymbol{\eta}_j'(x; b, \widehat{\boldsymbol{\beta}}) \mathcal{I}^{-1}(\widehat{\boldsymbol{\beta}}) \mathbf{X}_i h(\widehat{\boldsymbol{\beta}}' \mathbf{X}_i)\} e_i Z_i \tag{2.8}$$

given $(Y_i, \mathbf{X}_i)$ $(i = 1, \ldots, n)$, where

$$\boldsymbol{\eta}_j(x; b, \boldsymbol{\beta}) = -\sum_{i=1}^{n} I(x - b < X_{ji} \leq x) \partial \nu(\boldsymbol{\beta}' \mathbf{X}_i) / \partial \boldsymbol{\beta}. \tag{2.9}$$

Figure 1c shows the moving sum with $b = 5$, which is the range of the lower half of the covariate values. The fact that the raw residuals tend to be positive for the middle values of *blood-clotting score* and negative at the two ends is more transparent in Figure 1c than in Figure 1b. The observed value of the supremum statistic $S_1(b) \equiv \sup_x |W_1(x; b)|$ with $b = 5$ is 0.053, which corresponds to a positive rather than negative value of w_1. The associated P-value is 0.064, providing slightly stronger evidence against the selected functional form for *blood-clotting score* than $S_1 \equiv S_1(\infty)$.

Simulation studies indicated that the supremum tests based on $S_j(b)$ have proper sizes even when n is as small as 50 and when b is as small as the range of the lowest 10% of the covariate values. In addition, the choice of b that is roughly the range of the lower half of the covariate values results in a test that is slightly more powerful than S_j (i.e., $b = \infty$) in detecting misspecification of the power transformation for X_j (e.g., misspecifying $\log X_j$ as X_j or omitting X_j^2).

A question naturally arises as to whether or not one can guess the right functional form for the covariate by examining residual plots such as those of Figure 1b and 1c. To answer this question, we display in Figure 2 some prototype mean functions for the moving sums of residuals with various choices of b when the functional form of the covariate is misspecified. This figure was constructed under linear regression models, but similar patterns were observed for logistic and other models. The resemblance of an observed residual pattern to a specific curve in Figure 2 would provide a useful hint on how to correct the misspecification.

The observed residual patterns in Figure 1b and 1c resemble, respectively, the black and gray curves of Figure 2a, suggesting the log-transformation for *blood-clotting score*. The results shown in Figure 1d confirm that the log-transformation is indeed more appropriate than the original scale. In contrast, the log-transformation was not found to improve the fit appreciably by the conventional residual analysis (Neter et al., 1996, p. 389).

The moving sum defined in (2.7) is based on blocks of the same size only for $\min_i X_{ji} + b \leq x \leq \max_i X_{ji}$, and even within that range the moving sum may not resemble the familiar raw residuals if the covariate values are not evenly distributed. To mimic the raw residuals and their smooths, we consider the moving average

$$\overline{W}_j(x; b) = \frac{n^{1/2} \sum_{i=1}^{n} I(x - b < X_{ji} \leq x) e_i}{\sum_{i=1}^{n} I(x - b < X_{ji} \leq x)},$$

for

$$\min_i X_{ji} + b \leq x \leq \max_i X_{ji}.$$

The null distribution of $\overline{W}_j(x; b)$ can be approximated by $\widehat{W}_j(x; b)/n^{-1} \Sigma_{i=1}^{n} I(x - b < X_{ji} \leq x)$. A moving-average analog of Figure 1b and 1c is given in Figure 1e. The black curve in Figure 1e has a similar shape to the gray curve of Figure 1a. This is not surprising because *lowess* is essentially a weighted moving average. Because Figure 2 was constructed

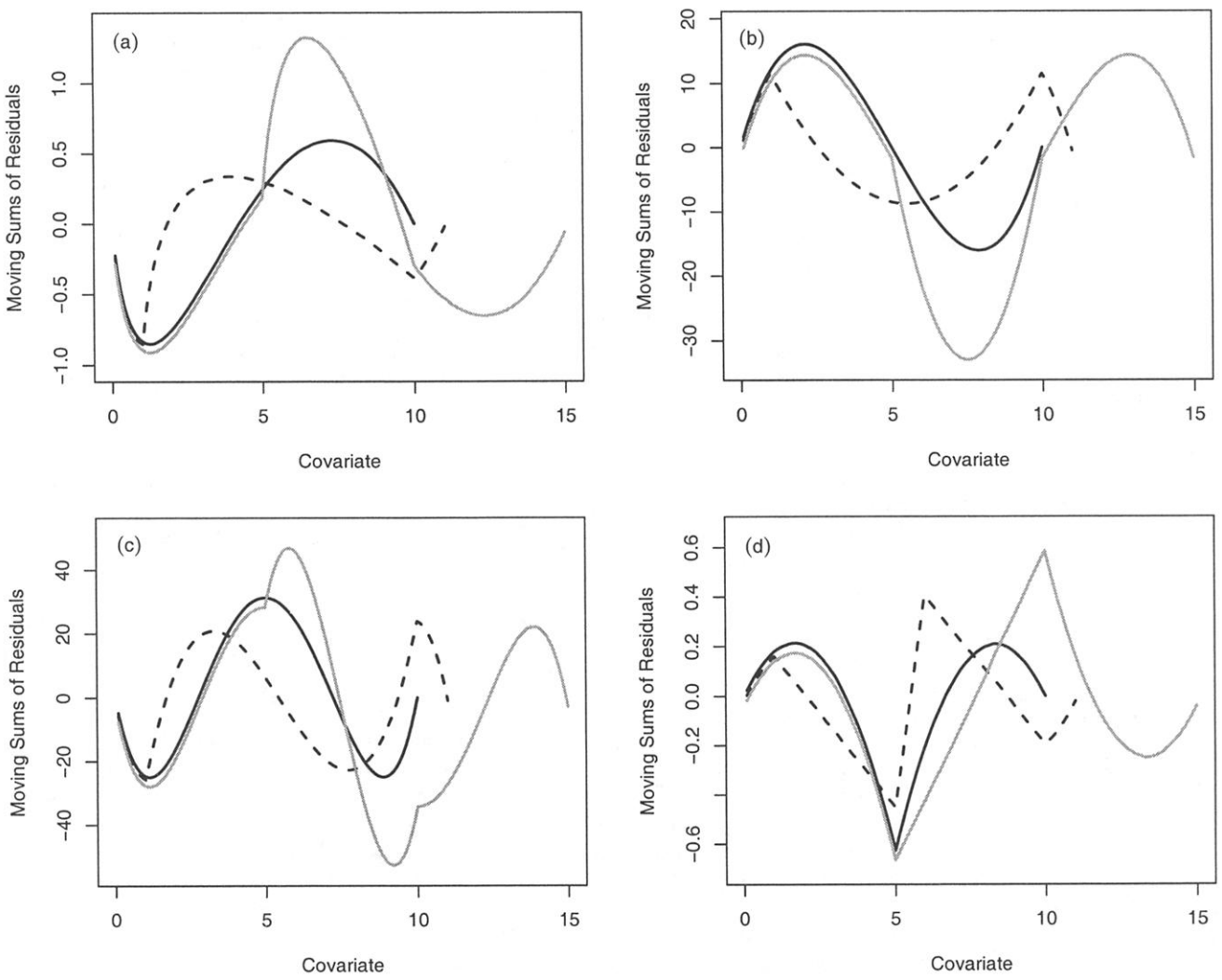


Figure 2. The mean functions of the moving sums $W_1(\cdot; b)$ with different values of b when the functional form of X_1 is misspecified: (**a**) model $E(Y) = \beta_0 + \beta_1 X_1$ is fitted to data generated by $E(Y) = \beta_0 + \gamma \log X_1$; (**b**) model $E(Y) = \beta_0 + \beta_1 X_1$ is fitted to data generated by $E(Y) = \beta_0 + \beta_1 X_1 + \gamma X_1^2$; (**c**) model $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$ is fitted to data generated by $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \gamma X_1^3$; and (**d**) model $E(Y) = \beta_0 + \beta_1 X_1$ is fitted to data generated by $E(Y) = \beta_0 + \gamma I(X_1 > 5)$. The data are generated by simulating normal responses under $\gamma = 1$ and $X = 1, 1.1, 1.2, \ldots, 9.9, 10$ with equal probability. The mean functions do not depend on other regression parameters. The curves look upside down under $\gamma = -1$. The dashed, gray and (solid) black curves correspond to $b = 1$, 5, and ∞. The dashed and gray curves are shifted downward to avoid overlaps of the curves.

with a uniformly distributed covariate, the patterns for the moving sums shown in Figure 2, when restricted to the range of $\min_i X_{ji} + b \le x \le \max_i X_{ji}$, would be what to expect of the moving averages. In general, the cumulative and moving sums are preferable to moving averages because the latter, although easier to interpret, are more variable.

To assess the linearity of model (1.1) and more generally the link function $g(\cdot)$ of model (2.2), we consider the moving sum of residuals over the fitted values

$$W_g(x; b) = n^{-1/2} \sum_{i=1}^{n} I(x - b < \widehat{\boldsymbol{\beta}}' \mathbf{X}_i \le x) e_i. \qquad (2.10)$$

The null distribution of $W_g(x; b)$ can be approximated by the conditional distribution of $\widehat{W}_g(x; b)$, which is obtained from $\widehat{W}_j(x; b)$ by replacing $I(x - b < X_{ji} \le x)$ in (2.8) and (2.9) with $I(x - b < \widehat{\boldsymbol{\beta}}' \mathbf{X}_i \le x)$. This approximation is a special case of a result established in the Appendix. As in the case of W_j, one may plot the observed process $w_g(\cdot; b)$ along with a few realizations of $\widehat{W}_g(\cdot; b)$, and supplement the graphical display with an estimated P-value for the supremum test $S_g(b) \equiv \sup_x |W_g(x; b)|$. Although we refer to S_g as the link function test, anomalies in W_g may reflect misspecification of the link function, the functional form of the response variable or the linear predictor.

Su and Wei (1991) confined their attention to the omnibus test $\sup_{\mathbf{x}} |W(\mathbf{x})|$. They showed that this test is consistent

against the general alternative that there does not exist a vector $\boldsymbol{\beta}$ such that (2.2) hold for all the $\mathbf{x}$ in the sample space generated by $\mathbf{X}$. By extending their arguments and those of Lin, Wei, and Ying (1993, Appendix 3), we can show that $S_g(b)$ is consistent against the general alternative that the link function is not the one specified in (2.2). Furthermore, $S_j(b)$ is consistent against any alternative under which $W_j(x;b)$ is not centered at 0 for all x, that is, $\lim_{n\to\infty} n^{-1/2}W_j(x;b)$ is nonzero for some x. In general, $W_j(x;b)$ will not be centered at 0 for all x if the functional form for X_j is misspecified. This is particularly true for linear regression when there is no additional model misspecification and X_j is independent of all other covariates.

Irregularities seen in the residual plot against X_j may be an artifact caused by a faulty functional form for another covariate closely correlated with X_j. Furthermore, a faulty functional form for a covariate may manifest itself in the residual plot against the fitted values, suggesting wrong choice of link function. Thus, all the proposed methods are checking the fit of the entire model specified by (2.2). Nevertheless, in general, W_j is most informative about the functional form for X_j, and W_g about the link function.

To identify a satisfactory model, one may adopt the following strategy: (1) construct a good initial model based on the scientific interest and substantive knowledge, and examine all components of this model; (2) fix the component that is the most problematic to arrive at a new model; (3) examine all components of the new model; and (4) iterate between steps 2 and 3 until no significant model misspecification can be detected.

In the surgical unit example, the sample correlation coefficients are 0.52, 0.28, and 0.87 between X_1 and X_2, X_1 and X_3, and X_2 and X_3. Because survival time is always positive, log-transformation is often employed. If one uses the logarithm (with base 10) of the survival time as the response variable, then the initial model is Neter's model given in (1.1). For illustration, suppose that we pick (1.1) as the initial model, but with Y being the untransformed survival time. Figure 1f displays the cumulative residuals against the fitted values of this model, with the P-value of the supremum test S_g being 0.036; the P-values of the supremum tests S_j $(j = 1, 2, 3)$ are 0.17, 0.18, and 0.11. Thus, the linearity assumption is the most problematic. The observed residual pattern in Figure 1f is opposite to that of Figure 1b, suggesting log-transformation for either Y or $E(Y)$. The use of the logarithm (with base 10) for Y of course yields Neter's model. For this model, the functional form of *blood-clotting score* turns out to be the most problematic, with the P-value of the supremum test S_1 being 0.108; the P-values of the supremum tests S_2 and S_3 for *prognostic index* and *enzyme function score* are 0.37 and 0.45, and that of the S_g test is 0.21. As discovered earlier, log(*blood-clotting score*) is more appropriate than *blood-clotting score*. This transformation leads to the revised model shown in Table 1b. The residual analysis indicates that the revised model is reasonable, with the P-values for S_j $(j = 1, 2, 3)$ and S_g being 0.48, 0.20, 0.63, and 0.65.

One may extend the omnibus test of Su and Wei (1991) by using $\sup_{\mathbf{x}} |W(\mathbf{x};\mathbf{b})|$, where

$$W(\mathbf{x};\mathbf{b}) = n^{-1/2}\sum_{i=1}^{n} I(\mathbf{x}-\mathbf{b} < \mathbf{X}_i \le \mathbf{x})e_i, \quad (2.11)$$

and $\mathbf{b}$ is a $p \times 1$ vector of constants. Clearly, (2.11) contains (2.7) as a special case. The null distribution of $W(\mathbf{x};\mathbf{b})$ can be approximated by the conditional distribution of $\widehat{W}(\mathbf{x};\mathbf{b})$, which is an obvious modification of (2.6) and (2.8). It is computationally demanding to implement the supremum tests $\sup_{\mathbf{x}} |W(\mathbf{x})|$ and $\sup_{\mathbf{x}} |W(\mathbf{x};\mathbf{b})|$ when $\mathbf{X}$ is of high dimension and n is large. To alleviate this burden, we recommend that the supremum be evaluated at a selected set of points, say the deciles of each covariate, rather than all possible covariate values.

The aforementioned graphical and numerical methods are targeted at the mean function given in (2.2), and do not require assumption (2.1) at all. Furthermore, all the results continue to hold even if $h(\boldsymbol{\beta}'\mathbf{X}_i)$ is replaced by any other weight function involving $\boldsymbol{\beta}$ and $\mathbf{X}_i$.

2.3 *Child Health and Development Study*

To further illustrate the proposed model-checking techniques, we consider the Child Health and Development Study described in Selvin (1995, pp. 58–59). The data consist of 680 live-born, white, male infants in the San Francisco, California area. A major scientific question is how the mother's age and smoking exposure affect the risk of a low-weight baby. Table 2a shows the results from fitting the following logistic model:

$$\text{logit}\,\pi = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \quad (2.12)$$

where X_1 is the mother's age, X_2 indicates, by the values 1 versus 0, whether or not the mother is a smoker, and π is the probability that the baby has low birth weight. Based on this model, one would conclude that maternal age is not a significant factor.

Figure 3a shows the moving sum of the residuals against X_1 with $b = 10$, which is the interquartile range of X_1; the P-value of 0.034 suggests wrong choice of the functional form for X_1. The P-value for the S_g test is 0.11. The observed residual pattern in Figure 3a resembles the gray curve in Figure 2b, suggesting inclusion of the square term. As evident from Table 2b, X_1^2 is indeed required. Once age^2 is added to model (2.12), the residual plot against *age* reveals no systematic trend, with the P-value for the supremum test $S_1(10)$ being 0.49 (see Figure 3b). The logistic link function is reasonable, with the P-value for the S_g test being 0.25. The results in Table 2b reveal that the risk of low birth weight is the lowest among women in their mid-20s and higher among the younger and older women.

Table 2
Logistic regression for the Child Health and Development Study data

Covariate	Estimate	SE	Estimate/SE	P-value
(a) Linear Age Effect				
Age	0.028	0.029	0.97	0.333
Smoking	1.080	0.345	3.13	0.002
(b) Quadratic Age Effect				
Age	−0.443	0.223	−1.98	0.047
Age^2	0.0083	0.0039	2.13	0.033
Smoking	1.079	0.347	3.11	0.002

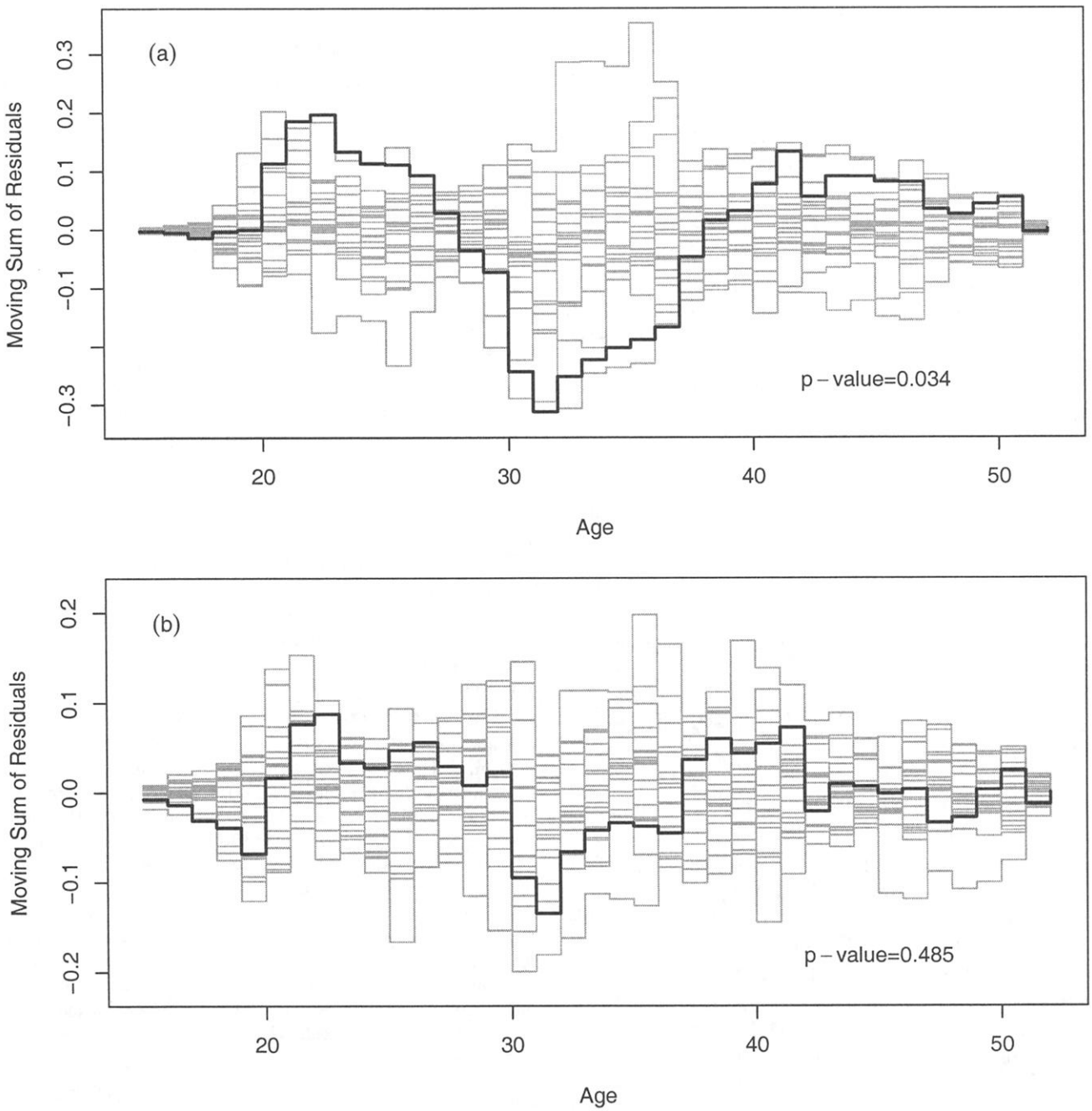


Figure 3. Plots of the moving sums of residuals with $b = 10$ against *age* for the Child Health and Development Study: (**a**) linear age effect model; and (**b**) quadratic age effect model. The observed pattern is shown by the black curve, and 20 simulated realizations are shown by the gray curves; the P-value pertains to the supremum test with 10,000 realizations.

Obviously, there is no need to use the proposed test to detect the omission of a quadratic term or any other terms. In fact, the Wald and score statistics are (asymptotically) efficient in testing against such alternatives. However, the latter statistics can only be used to test nested alternatives, that is, extra parameters in embedded parametric models. In contrast, the proposed supremum statistics can be used to assess which functional form is more appropriate and whether the chosen functional form is satisfactory. Numerical studies indicated that the efficiencies of the proposed supremum tests are quite high relative to the Wald and score statistics in testing nested alternatives.

3. Marginal Models for Dependent Observations

3.1 *Marginal Models and Generalized Estimating Equations*

Dependent observations occur when a response variable is measured repeatedly over time on a subject, when several response variables are measured on each subject, or when the

subjects are sampled in clusters. For simplicity of description, we focus on repeated measures from longitudinal studies, although the results apply to other types of dependent data.

For $i = 1, \ldots, n$, and $k = 1, \ldots, K_i$, let Y_{ik} be the response of the ith subject on the kth occasion, and let $\mathbf{X}_{ik}$ be the corresponding $(p+1) \times 1$ vector of covariates. We assume that the marginal mean of the response, $E(Y_{ik}) \equiv \mu_{ik}$, depends on the covariate vector $\mathbf{X}_{ik}$ by

$$g(\mu_{ik}) = \boldsymbol{\beta}' \mathbf{X}_{ik}, \tag{3.1}$$

where $g(\cdot)$ is a known link function, and $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of unknown regression parameters. The marginal variance depends on the marginal mean according to $\mathrm{Var}(Y_{ik}) = \xi(\mu_{ik})\phi$, where $\xi(\cdot)$ is a known function, and ϕ is a scale parameter that may need to be estimated.

Let $\mathbf{R}_i(\boldsymbol{\alpha})$ be a "working" correlation matrix for $(Y_{i1}, \ldots, Y_{iK_i})$ that perhaps depends on a vector of unknown parameters $\boldsymbol{\alpha}$. Define $\mathbf{A}_i = \mathrm{diag}\{\xi(\mu_{i1}), \ldots, \xi(\mu_{iK_i})\}$. We estimate $\boldsymbol{\beta}$ by solving the generalized estimating equation

$$\sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} \boldsymbol{\epsilon}_i = \mathbf{0}, \tag{3.2}$$

where $\boldsymbol{\epsilon}_i = (Y_{i1} - \mu_{i1}, \ldots, Y_{iK_i} - \mu_{iK_i})'$, $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \times \mathbf{A}_i^{1/2}$, and $\mathbf{D}_i = \{\partial \mu_{ik} / \partial \beta_j;\ k = 1, \ldots, K_i;\ j = 1, \ldots, p+1\}$. The resulting estimator $\widehat{\boldsymbol{\beta}}$ is consistent and asymptotically normal with covariance matrix

$$\left(\sum_{i=1}^{n} \widehat{\mathbf{D}}_i' \widehat{\mathbf{V}}_i^{-1} \widehat{\mathbf{D}}_i \right)^{-1} \sum_{i=1}^{n} \widehat{\mathbf{D}}_i' \widehat{\mathbf{V}}_i^{-1} \mathbf{e}_i \mathbf{e}_i' \widehat{\mathbf{V}}_i^{-1} \widehat{\mathbf{D}}_i \times \left(\sum_{i=1}^{n} \widehat{\mathbf{D}}_i' \widehat{\mathbf{V}}_i^{-1} \widehat{\mathbf{D}}_i \right)^{-1}$$

even if ξ and $\mathbf{R}_i$ $(i = 1, \ldots, n)$ are misspecified, where $\mathbf{e}_i = (e_{i1}, \ldots, e_{iK_i})' = (Y_{i1} - \widehat{\mu}_{i1}, \ldots, Y_{iK_i} - \widehat{\mu}_{iK_i})'$, $\widehat{\mu}_{ik} = g^{-1}(\widehat{\boldsymbol{\beta}}' \mathbf{X}_{ik})$, and $\widehat{\mathbf{D}}_i$ and $\widehat{\mathbf{V}}_i$ are obtained by replacing the unknown parameters in $\mathbf{D}_i$ and $\mathbf{V}_i$ with their sample estimators (Liang and Zeger, 1986).

3.2 *Model-Checking Techniques*

Because of the similarity between (3.1) and (2.2), it is natural to extend the methods of Section 2.2 to the current setting. Under model (3.1), the residuals e_{ik} $(i = 1, \ldots, n;\ k = 1, \ldots, K_i)$ are centered at 0. To check the functional forms of covariates and to construct omnibus tests, we consider the following extension of (2.11):

$$W(\mathbf{x}; \mathbf{b}) = n^{-1/2} \sum_{i=1}^{n} \sum_{k=1}^{K_i} I(\mathbf{x} - \mathbf{b} < \mathbf{X}_{ik} \le \mathbf{x}) e_{ik}; \tag{3.3}$$

to check the link function, we consider the analogue of (2.10)

$$W_g(x; b) = n^{-1/2} \sum_{i=1}^{n} \sum_{k=1}^{K_i} I(x - b < \widehat{\boldsymbol{\beta}}' \mathbf{X}_{ik} \le x) e_{ik}. \tag{3.4}$$

A new technical challenge in dealing with (3.3) and (3.4) is that the repeated measures within the same subject are potentially correlated. Nevertheless, we prove in the Appendix that the null distribution of $W(\mathbf{x}; \mathbf{b})$ can be approximated by the conditional distribution of

$$\widehat{W}(\mathbf{x}; \mathbf{b}) = n^{-1/2} \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K_i} I(\mathbf{x} - \mathbf{b} < \mathbf{X}_{ik} \le \mathbf{x}) e_{ik} + \boldsymbol{\eta}'(\mathbf{x}; \mathbf{b}, \widehat{\boldsymbol{\beta}}) \left(\sum_{l=1}^{n} \widehat{\mathbf{D}}_l' \widehat{\mathbf{V}}_l^{-1} \widehat{\mathbf{D}}_l \right)^{-1} \times \widehat{\mathbf{D}}_i' \widehat{\mathbf{V}}_i^{-1} \mathbf{e}_i \right\} Z_i \tag{3.5}$$

given the data $(Y_{ik}, \mathbf{X}_{ik})$ $(i = 1, \ldots, n; k = 1, \ldots, K_i)$, where

$$\boldsymbol{\eta}(\mathbf{x}; \mathbf{b}, \boldsymbol{\beta}) = -\sum_{i=1}^{n} \sum_{k=1}^{K_i} I(\mathbf{x} - \mathbf{b} < \mathbf{X}_{ik} \le \mathbf{x}) \partial \mu_{ij} / \partial \boldsymbol{\beta}. \tag{3.6}$$

In addition, the null distribution of $W_g(x; b)$ can be approximated by the conditional distribution of $\widehat{W}_g(x; b)$, which is obtained from $\widehat{W}(\mathbf{x}; \mathbf{b})$ by replacing $I(\mathbf{x} - \mathbf{b} < \mathbf{X}_{ik} \le \mathbf{x})$ in (3.5) and (3.6) with $I(x - b < \widehat{\boldsymbol{\beta}}' \mathbf{X}_{ik} \le x)$.

Given the above asymptotic approximations, graphical and numerical inspections of model (3.1) can be carried out in the same manner as in Section 2.2. These methods are valid even if ξ and $\mathbf{R}_i$ $(i = 1, \ldots, n)$ are misspecified, and may also be used even if $\mathbf{D}_i' \mathbf{V}_i^{-1}$ $(i = 1, \ldots, n)$ are replaced by other $(p+1) \times K_i$ weight matrices. Furthermore, the supremum tests have the same consistency properties as in the case of independent observations.

3.3 *AIDS Clinical Trial*

We now apply the proposed methods to the CD4 data taken from an AIDS clinical trial reported by Fischl et al. (1990). The study randomly assigned 360 HIV patients to azidothymidine (zidovudine; AZT) and 351 to placebo. The CD4 counts were measured repeatedly over the course of the study. We confine our attention to the 4328 measurements taken during the first 40 weeks of the study.

A major challenge in modeling longitudinal data is to specify the time trend of the response. Biological rationale and descriptive data analysis suggest that the CD4 counts among the placebo patients tend to decline monotonically over the entire study period, whereas those of the AZT patients tend to rise for the first few weeks and then decline over time. Thus, we first consider the following model:

$$E(Y_{ik}) = \beta_0 + \beta_1 T_{ik} + \beta_2 T_{ik}^2 + \beta_3 R_i T_{ik} + \beta_4 R_i T_{ik}^2, \tag{3.7}$$

where T_{ik} is the time (in weeks) when the kth measurement of the CD4 count is made on the ith patient, Y_{ik} is the CD4 measurement at T_{ik} for the ith patient, and R_i is the indicator of AZT for the ith patient. There is no need to include R_i itself

---

$\longrightarrow$

Figure 4. Plots of the cumulative sums of residuals against *time* for the CD4 data: (**a**)–(**e**) pertain to the quadratic model, and (**f**) pertains to the cubic model. For (**a**)–(**d**), two simulated realizations are shown in gray and dotted curves along with the black observed curve. For (**e**) and (**f**), the observed pattern is shown by the black curve, and 20 simulated realizations are shown by the gray curves; the P-value pertains to the supremum test with 10,000 realizations.

Table 3

Marginal linear regression for the CD4 data[a]

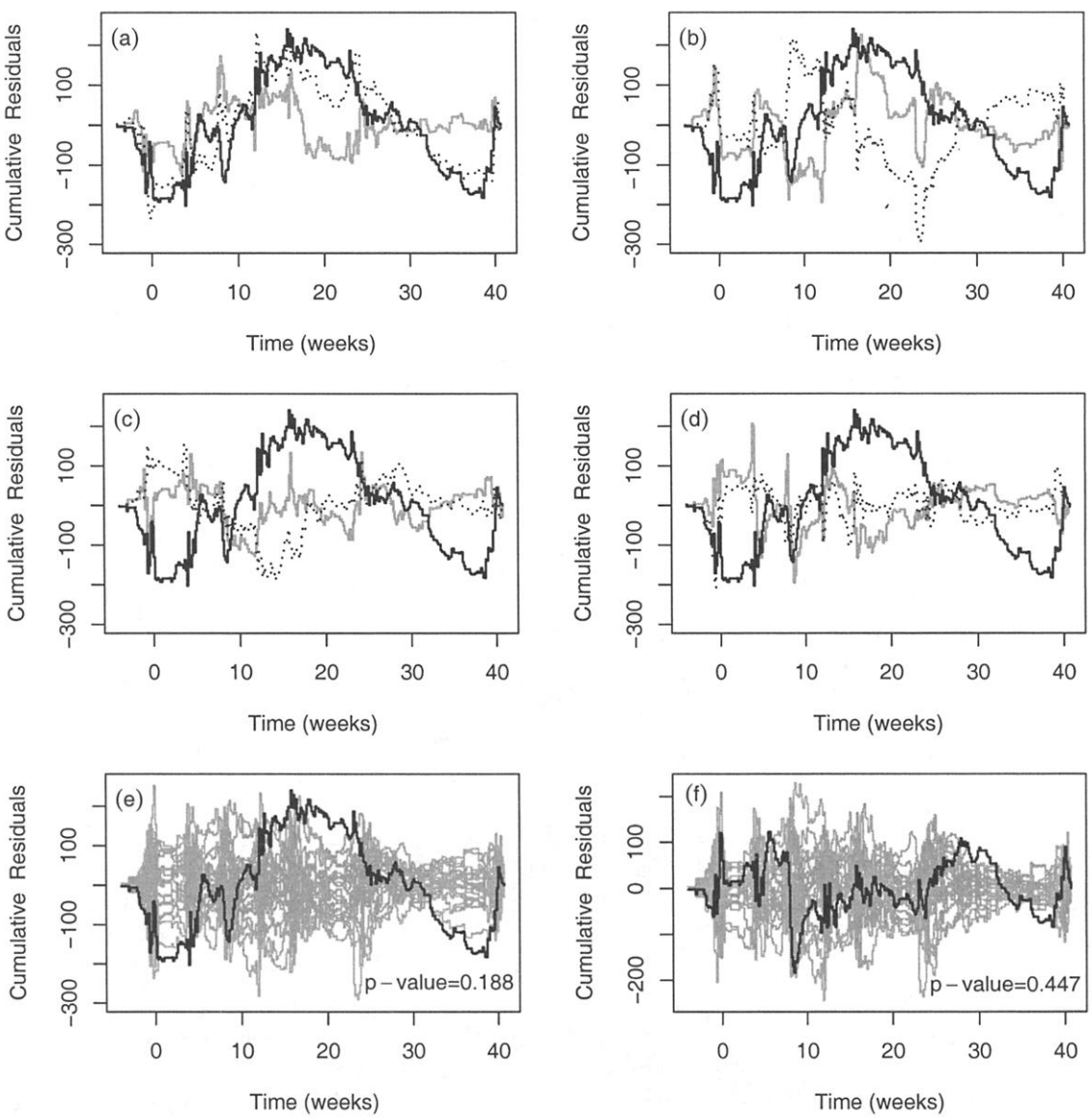
Covariate	Estimate	SE	Estimate/SE	*P*-value
(a) Quadratic Time Trend				
Time	−1.957	0.854	−2.29	0.022
$Time^2$	0.037	0.022	1.64	0.101
Treatment × *time*	5.369	1.407	3.82	0.0001
Treatment × $time^2$	−0.146	0.035	−4.15	<0.0001
(b) Cubic Time Trend				
Time	0.077	1.726	0.045	0.965
$Time^2$	−0.111	0.118	−0.95	0.344
$Time^3$	0.0026	0.002	1.27	0.206
Treatment × *time*	7.672	2.737	2.80	0.005
Treatment × $time^2$	−0.350	0.178	−1.97	0.049
Treatment × $time^3$	0.0038	0.003	1.26	0.208

[a] The independence working assumption is used in all analyses.

in the model because the treatment difference is zero at the baseline because of randomization. The estimation results for (3.7) are summarized in Table 3a. The estimates are consistent with the anticipated time trends for the placebo and AZT groups.

Figure 4a–4e displays the cumulative residuals against time under model (3.7); the simulated realizations are shown both in pairs and in aggregate of 20. The *P*-value of the associated test is 0.19. The *P*-value for the S_g test is 0.22. These results indicate that model (3.7), although reasonable, is not entirely satisfactory. The observed residual pattern under this model resembles the black curve of Figure 2c, suggesting cubic time trends.

Table 3b shows the estimation results when T_{ik}^3 and $R_i T_{ik}^3$ are added to the right-hand side of equation (3.7). The *P*-values for these two terms are both about 0.21, which are similar to the aforementioned *P*-values of the supremum



goodness-of-fit tests for model (3.7). The residual analysis reveals no systematic pattern under this new model, with the P-values for testing the time trend and link function being 0.45 and 0.70, respectively (see Figure 4f).

4. Discussion

This paper expands upon the existing work on model-checking techniques based on cumulative residual processes. New graphical and numerical methods for examining individual components of a generalized linear model (as opposed to omnibus tests) are introduced. Model-checking techniques for dependent observations are developed, and the issue of determining the nature of model misspecification from cumulative residual plots is addressed. The use of general moving sums and averages of residuals to enhance the versatility of the model-checking techniques also is explored.

The results of the residual analysis may depend on the block size b. Different values of b are potentially sensitive to different types of model misspecification: large values of b are desirable in detecting global misspecification, whereas small values are more sensitive to local departures. One may try different values of b in exploratory data analysis. For formal goodness-of-fit tests, b ought to be prespecified. This is especially important in small samples because then the results may vary substantially with b. In theory, b can be any specific constant or a data-dependent quantity that becomes deterministic as $n \to \infty$.

The aggregate plots in Figure 1 and other figures are reminiscent of the simulation envelopes for assessing normality (Atkinson, 1985, p. 35). We may display the empirical quantiles instead of the actual simulated curves. There is also an interesting connection between the residual analysis discussed here and the sampling inspection used in statistical quality control. The plots of raw residuals, cumulative sum, and moving sum and average are analogous to Shewhart, cusum, and moving-average control charts (Montgomery, 1997, Chapters 4 and 7). The distributions of the control charts are simpler than those of the residual plots because there is essentially no "plug-in" estimator.

The basic building blocks for the new methods are the ordinary residuals in the form of the observed minus predicted values of the response. Although the proposed approach can be applied to other forms of residuals (e.g., standardized residuals and partial residuals), the ordinary residuals are most informative about misspecification of the mean function of the model, which is the focus of this paper. One should examine outliers and influential observations by using standardized residuals, deviance residuals, and other measures.

Lin et al. (1993) showed how to use the cumulative sums of martingale residuals (Barlow and Prentice, 1988; Therneau, Grambsch, and Fleming, 1990) to check the proportional hazards model of Cox (1972) with possibly censored survival data, whereas Spiekerman and Lin (1996) and Lin et al. (2000) provided parallel developments for marginal Cox models with correlated survival data and for proportional means and rates models with recurrent events data. These authors confined their attention to the exponential regression function and time-independent covariates. We are currently exploring extensions to nonexponential regression functions and time-dependent covariates. We also are investigating the use of cumulative residual processes in identifying the nature of model misspecification for these semiparametric models.

The implementation of the proposed methods requires special software. FORTRAN programs for several models, including the ones used in the examples, are available from the authors. Commercial software packages that implement the proposed graphical and numerical techniques currently are being developed by the Cytel and StatSci corporations. The software allows dynamic colored graphics.

Acknowledgements

This research was supported by the National Institutes of Health and the National Science Foundation. The authors are grateful to the reviewers for their helpful comments.

Résumé

Depuis longtemps, on utilise les résidus pour apprécier graphiquement et numériquement l'adéquation de modèles de régression. L'analyse conventionnelle des résidus, fondée sur les graphes des résidus bruts ou des courbes lissées est éminemment subjective, tandis que la plupart des tests numériques d'ajustement informent peu sur la nature de l'inadéquation. Nous développons des techniques d'examen d'un modèle, objectives et informatives, qui reposent soit sur les sommes des résidus cumulées selon certaines coordonnées (comme une covariable ou le prédicteur linéaire), soit sur des regroupements partiels (sommes ou moyennes mobiles) selon ces coordonnées. Pour une grande variété de modèles statistiques et de structures des observations, dont les modèles linéaires généralisés sur observations indépendantes ou corrélées, les distributions de ces processus stochastiques sous le modèle postulé peuvent être approchées par des processus gaussiens centrés, faciles à simuler. On peut donc comparer graphiquement et numériquement chaque processus observé à une série de réalisations du processus gaussien. Ces comparaisons permettent d'évaluer objectivement dans quelle mesure une tendance observée sur un graphique serait attribuable à une mauvaise spécification du modèle plutôt qu'aux fluctuations d'échantillonnage. Ces techniques sont particulièrement utiles pour l'examen de la fonction de lien ou de la forme fonctionnelle d'une covariable. Plusieurs exemples médicaux illustrent cette approche.

References

Atkinson, A. C. (1985). *Plots, Transformations and Regression.* Oxford: Oxford University Press.

Barlow, W. E. and Prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika* **75,** 65–74.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74,** 829–836.

Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics.* New York: Wiley.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics* **13,** 342–368.

Fischl, M. A., Richman, D. D., Hansen, N., et al. (1990). The safety and efficacy of zidovudine (AZT) in the treatment of subjects with mildly symptomatic human immunodeficiency virus type 1 (HIV) infection. *Annals of Internal Medicine* **112,** 727–737.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Lin, D. Y., Wei, L. J., and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80,** 557–572.

Lin, D.Y., Wei, L.J., Yang, I., and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Series B* **62,** 711–730.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. New York: Wiley.

Montgomery, D. C. (1997). *Introduction to Statistical Quality Control*, 3rd edition. New York: Wiley.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Statistical Models*, 4th edition. Chicago: IRWIN.

Pollard, D. (1990). *Empirical Processes: Theory and Applications.* Hayward, California: IMS.

Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika* **67,** 145–153.

Selvin, S. (1995). *Practical Biostatistical Methods.* Belmont, California: Duxbury Press.

Spiekerman, C. F. and Lin, D. Y. (1996). Checking the marginal Cox model for correlated failure time data. *Biometrika* **83,** 143–156.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics* **25,** 613–641.

Su, J. Q. and Wei, L. J. (1991). A lack-of-fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association* **86,** 420–426.

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika* **77,** 147–160.

Tsiatis, A. A. (1980). A note on a goodness-of-fit test for the logistic regression without replication. *Biometrika* **67,** 250–251.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50,** 1–26.

Received March 2001. *Revised October* 2001.
Accepted October 2001.

Appendix

Weak Convergence of $W(\mathbf{x};\mathbf{b})$, $W_g(x;b)$, $\widehat{W}(\mathbf{x};\mathbf{b})$, *and* $\widehat{W}_g(x;b)$

We shall establish the weak convergence for the processes $W(\mathbf{x};\mathbf{b})$, $W_g(x;b)$, $\widehat{W}(\mathbf{x};\mathbf{b})$, and $\widehat{W}_g(x;b)$ defined in Section 3. (Note that Section 2 is a special case of Section 3.) We assume that $(Y_{i1},\ldots,Y_{iK_i};\ \mathbf{X}_{i1},\ldots,\mathbf{X}_{iK_i};\ K_i)$ $(i=1,\ldots,n)$ are independent and identically distributed, and K_i, $\|\mathbf{X}_{ik}\|$ and the second moments of Y_{ik} $(i=1,\ldots,n;\ k=1,\ldots,K_i)$ are bounded by a constant not dependent on i or k.

We first establish the weak convergence of $W(\mathbf{x};\mathbf{b})$ as a two-parameter process in $\mathbf{x}$ and $\mathbf{b}$. By the consistency of $\widehat{\boldsymbol{\beta}}$ and the Taylor series expansion, uniformly in $\mathbf{x}$ and $\mathbf{b}$,

$$W(\mathbf{x};\mathbf{b}) = n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K_i} I(\mathbf{x}-\mathbf{b}<\mathbf{X}_{ik}\le\mathbf{x})\epsilon_{ik} + n^{-1}\boldsymbol{\eta}'(\mathbf{x};\mathbf{b},\boldsymbol{\beta})\left(n^{-1}\sum_{l=1}^{n}\mathbf{D}_l'\mathbf{V}_l^{-1}\mathbf{D}_l\right)^{-1} \times n^{-1/2}\sum_{i=1}^{n}\mathbf{D}_i'\mathbf{V}_i^{-1}\boldsymbol{\epsilon}_i + o_p(1). \tag{A.1}$$

By the law of large numbers, $n^{-1}\Sigma_{l=1}^{n}\mathbf{D}_l'\mathbf{V}_l^{-1}\mathbf{D}_l$ converges to a constant matrix. Furthermore, the uniform law of large numbers (Pollard, 1990, p. 39) implies that $n^{-1}\boldsymbol{\eta}(\mathbf{x};\mathbf{b},\boldsymbol{\beta})$ converges, uniformly in $\mathbf{x}$ and $\mathbf{b}$, to a nonrandom function. Thus, for fixed $\mathbf{x}$ and $\mathbf{b}$, the right-hand side of (A.1) is essentially a normalized sum of n independent and identically distributed zero-mean random variables. It then follows from the multivariate central limit theorem that $W(\mathbf{x};\mathbf{b})$ converges in finite-dimensional distributions to a zero-mean Gaussian process. The tightness for the second term on the right-hand side of (A.1) follows from the asymptotic normality of $n^{-1/2}\Sigma_{i=1}^{n}\mathbf{D}_i'\times\mathbf{V}_i^{-1}\boldsymbol{\epsilon}_i$, together with the uniform convergence of $n^{-1}\boldsymbol{\eta}(\mathbf{x};\mathbf{b},\boldsymbol{\beta})$. The first term consists of sums of monotone step functions, which are clearly manageable (Pollard, 1990, p. 38). Thus, the functional central limit theorem (Pollard, 1990, p. 53) entails that this term is tight. Hence, $W(\mathbf{x};\mathbf{b})$ converges weakly to a zero-mean Gaussian process with covariance function between $(\mathbf{x},\mathbf{b})$ and $(\mathbf{x}^\dagger,\mathbf{b}^\dagger)$ being

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\left\{\sum_{k=1}^{K_i} I(\mathbf{x}-\mathbf{b}<\mathbf{X}_{ik}\le\mathbf{x})\epsilon_{ik} + \boldsymbol{\eta}'(\mathbf{x};\mathbf{b},\boldsymbol{\beta})\left(\sum_{l=1}^{n}\mathbf{D}_l'\mathbf{V}_l^{-1}\mathbf{D}_l\right)^{-1}\mathbf{D}_i'\mathbf{V}_i^{-1}\boldsymbol{\epsilon}_i\right\} \times\left\{\sum_{k=1}^{K_i} I(\mathbf{x}^\dagger-\mathbf{b}^\dagger<\mathbf{X}_{ik}\le\mathbf{x}^\dagger)\epsilon_{ik} + \boldsymbol{\eta}'(\mathbf{x}^\dagger;\mathbf{b}^\dagger,\boldsymbol{\beta})\left(\sum_{l=1}^{n}\mathbf{D}_l'\mathbf{V}_l^{-1}\mathbf{D}_l\right)^{-1}\times\mathbf{D}_i'\mathbf{V}_i^{-1}\boldsymbol{\epsilon}_i\right\}. \tag{A.2}$$

We next show that, conditional on the data $(Y_{ik},\mathbf{X}_{ik})$ $(i=1,\ldots,n;k=1,\ldots,K_i)$, the process $\widehat{W}(\mathbf{x};\mathbf{b})$ converges weakly to the same limiting Gaussian process as that of $W(\mathbf{x};\mathbf{b})$. Conditional on the data, the only random elements in $\widehat{W}(\mathbf{x};\mathbf{b})$ are $(Z_1,\ldots,Z_n)$. Thus, it follows from the multivariate central limit theorem and a straightforward covariance calculation that, conditional on the data, $\widehat{W}(\mathbf{x};\mathbf{b})$ converges in finite-dimensional distributions to a zero-mean Gaussian process whose covariance function converges to (A.2). The tightness of $\widehat{W}(\mathbf{x};\mathbf{b})$ follows from the above arguments for showing the tightness of $W(\mathbf{x};\mathbf{b})$.

To establish the weak convergence of $W_g(x;b)$ and $\widehat{W}_g(x;b)$, we define the three-parameter process

$$W_g(x;b,\boldsymbol{\gamma}) = n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K_i} I(x-b<\boldsymbol{\gamma}'\mathbf{X}_{ik}\le x)e_{ik}.$$

Obviously, $W_g(x;b) = W_g(x;b,\widehat{\boldsymbol{\beta}})$. We wish to show that

$$W_g(x;b,\widehat{\boldsymbol{\beta}}) = W_g(x;b,\boldsymbol{\beta}) + o_p(1). \tag{A.3}$$

To this end, we assume that $\Sigma_{k=1}^{K_i} I(\boldsymbol{\gamma}'\mathbf{X}_{ik} \le x)\epsilon_{ik}$ is L_2-continuous in a small neighborhood of $\boldsymbol{\beta}$ in the sense that

$$E\left\{\sum_{k=1}^{K_i} I(\boldsymbol{\gamma}_1'\mathbf{X}_{ik}\le x_1)\epsilon_{ik} - \sum_{k=1}^{K_i} I(\boldsymbol{\gamma}_2'\mathbf{X}_{ik}\le x_2)\epsilon_{ik}\right\}^2 \to 0$$

as $\boldsymbol{\gamma}_1 \to \boldsymbol{\gamma}_2$ and $x_1 \to x_2$, for all $\boldsymbol{\gamma}_2$ sufficiently close to $\boldsymbol{\beta}$. This assumption corresponds to condition (v) in the functional central limit theorem of Pollard (1990, p. 53), and ensures that $W_g(x;b,\boldsymbol{\gamma})$ is stochastically equicontinuous in $(x;b,\boldsymbol{\gamma})$ when $(x;b,\boldsymbol{\gamma})$ is equipped with the usual Euclidean metric. Thus, we can apply the above arguments for establishing the weak convergence of $W(\mathbf{x};\mathbf{b})$ to show that $W_g(x;b,\boldsymbol{\gamma})$ converges weakly and then use the stochastic equicontinuity to establish (A.3). The rest of the proof is similar to the preceding proof for $W(\mathbf{x};\mathbf{b})$ and $\widehat{W}(\mathbf{x};\mathbf{b})$.