Paper 255-28
Let the Data Speak: New Regression Diagnostics Based on Cumulative Residuals

Gordon Johnston and Ying So
SAS Institute Inc.
Cary, North Carolina, USA

Abstract

Residuals have long been used as the basis for graphical and numerical examination of the adequacy of regression models. Conventional residual analysis based on plotting raw residuals or their smoothed versions is highly subjective, whereas most goodness-of-fit tests provide little information about the nature of model inadequacy.

In this paper, new model-checking techniques of Lin et al. (1993, 2002) based on cumulative sums and other aggregates of residuals are described. These techniques provide objective and informative checks on the adequacy of the fitted model for a variety of statistical models and data structures. Specific aspects of the adequacy of the fitted model can be assessed, depending on the particular model. In generalized linear models and marginal models for dependent responses (GEEs), special attention is given to checking the functional form of a covariate in the linear predictor and the form of the link function. In proportional hazards models, the focus is on checking the functional form of a covariate and the validity of the proportional hazards assumption. These methods are available for release 9.1 of SAS/STAT® software in the GENMOD and PHREG procedures, and use ODS graphics for graphical output. Use of these graphical and numerical methods is illustrated with several examples.

Introduction

Regression models seek to express a population measure (for instance, the mean) as a linear or nonlinear function of some regression parameters, usually related to the physical or experimental setup. Regression plays an important role in the analysis of experimental and observational data, and many SAS/STAT procedures are devoted to fitting regression models. Modern computing capabilities enable the routine fitting of both linear and nonlinear regression models with complex data structures, such as discrete responses, dependent responses, and censoring. While model misspecification can affect the validity and efficiency of regression models, model checking has not become routine practice, in part due to lack of suitable tools, especially for the more complicated models.

Residuals, defined as the difference between the measured responses and the fitted values, are highly informative about the fit of a model. If the model is correct, the residuals are centered at zero, and a plot of the residuals against a coordinate such as one of the covariates or the fitted values will display no systematic patterns. The appearance of a systematic trend may indicate the absence of an important covariate or an incorrect functional form. However, determining whether a pattern observed in a residual plot is due to a model misspecification or due to natural variation can be difficult. Techniques introduced by Lin et al. (1993, 2002) based on cumulative sums of residuals or other aggregates of residuals, such as moving sums or LOWESS smoothed residuals, with respect to certain coordinates, provide objective and informative criteria with which to check the adequacy of the model.

As an illustration, consider the surgical unit example described in Neter et al. (1996) and further analyzed in Lin et al. (2002). The data contains the survival time and covariates for 54 patients undergoing a particular type of liver surgery. Neter et al. (1996) arrived at the following linear model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

where Y is the logarithm of survival time, and X_1, X_2,

and X_3 are, respectively, *blood-clotting score*, *prognostic index*, and *enzyme function score*, and ϵ is a $N(0, \sigma^2)$ error term. The parameter estimates table from a GENMOD fit for the above model is shown in Table 1. All the parameters appear to be highly significant.

Table 1. Original Model for Surgical Unit Example

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	0.4836	0.0426	0.4001	0.5672	128.71	<.0001
x1	1	0.0692	0.0041	0.0612	0.0772	288.17	<.0001
x2	1	0.0093	0.0004	0.0085	0.0100	590.45	<.0001
x3	1	0.0095	0.0003	0.0089	0.0101	966.07	<.0001
Scale	0	0.0469	0.0000	0.0469	0.0469		

NOTE: The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

Raw residuals, along with LOWESS smoothed raw residuals from the model above, are shown in Figure 1, plotted versus X_1. The raw residuals, indicated by the points in Figure 1, seem to show no particular pattern. The LOWESS smoothed raw residuals, plotted as the solid line in Figure 1, show some curvature, but possibly not enough to indicate a convincing trend. A problem in the interpretation of such plots is the subjective nature of the interpretation.

Lin et al. (1993, 2002) present a more objective way of checking model fit based on cumulative sums of residuals over certain coordinates, for specific types of regression models. The following sections present examples of model-checking and selection for generalized linear models, marginal models for dependent responses (GEEs), and proportional hazards models for censored survival data. The GENMOD and PHREG procedures were used to produce the graphical and numerical results for these examples. The graphical results were created with new features in the procedures using ODS graphics.

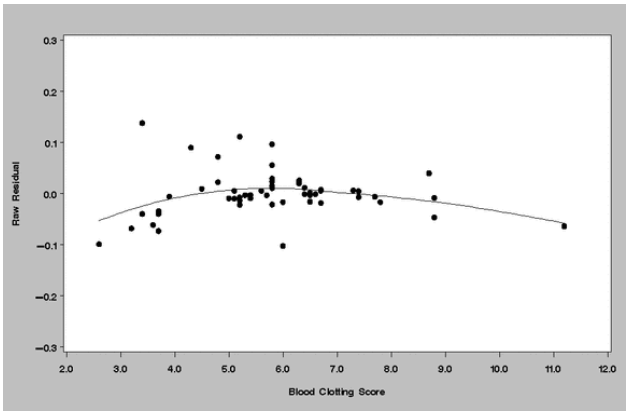

Figure 1. Raw Residuals for Surgical Unit Example

Generalized Linear Models

The solid line in Figure 2 shows the cumulative sum of residuals from the model for the surgical unit data, with respect X_1. For any value x on the horizontal axis, the solid line represents the cumulative sum of the residuals for all values of X_1 less than or equal to x. Like the raw residuals or smoothed raw residuals, cumulative residuals will be centered at zero if the model fit is correct. The motivation for considering cumulative sums of residuals is that the asymptotic distribution can be determined. Under the null hypothesis of a correct model fit, they can be approximated as a zero mean Gaussian process with a covariance structure determined by the particular type of regression model. Realizations of the Gaussian process can be simulated by computer and compared with the observed process to assess whether the observed residual process represents anything beyond random variation.

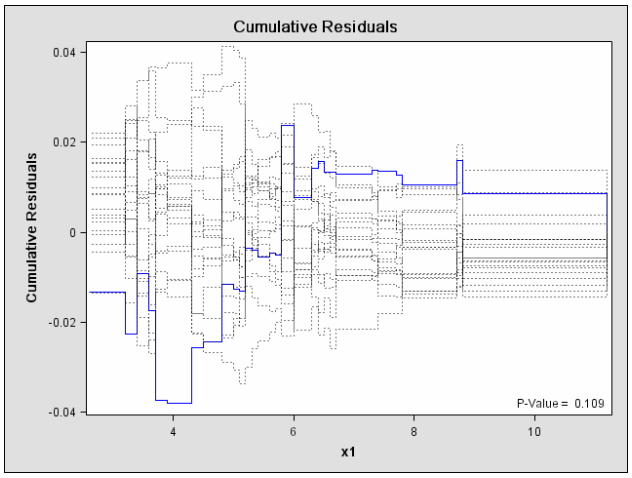

Figure 2. Cumulative Residuals for Surgical Unit Example

The light dashed lines in Figure 2 are the first 20 realizations of 10,000 simulated paths of the cumulative residual process under the null hypothesis of a correct model fit. Most of the paths tend to be closer to and intersect the horizontal axis more than the observed residuals. The maximum absolute value of the observed cumulative residuals is 0.038. Of the 10,000 realizations under the null hypothesis, about 10.9% were more extreme in maximum absolute value. Thus, the *p*-value for a Kolmogorov-type supremum test is 0.109. These results suggest that there may be a better fitting model for the surgical unit data.

Lin et al. (2002) find that a model with $\log(X_1)$ instead of X_1 provides a better fit than the original model. Figure 3 shows the cumulative residuals and the first 20 realizations of the null hypothesis process. The observed residual process appears to be more typical of the null hypothesis simulations, and the *p*-value, based on 10,000 simulated sample paths, is 0.483, indicating a more appropriate model.

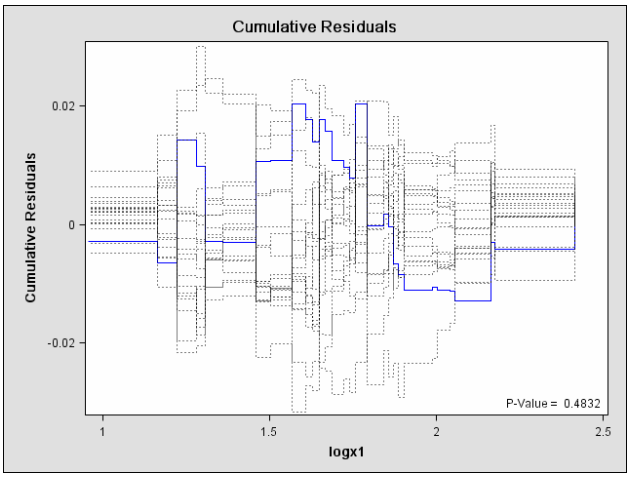

Figure 3. Cumulative Residuals for Revised Model

The parameter estimates for the revised model are shown in Table 2.

Table 2. Revised Model for Surgical Unit Example

Parameter	DF	Estimate	Standard Error	Pr > ChiSq
Intercept	1	0.1844	0.0504	0.0003
logx1	1	0.3961	0.0213	<.0001
x2	1	0.0095	0.0004	<.0001
x3	1	0.0096	0.0003	<.0001
Scale	0	0.0434	0.0000	

NOTE: The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

You can perform model checking using other aggregates of residuals, such as moving sums of residuals and LOWESS smoothed residuals, which are weighted averages of residuals. These are both available in GENMOD, and the analysis is performed in the same way as in the surgical unit example.

Although the surgical unit example was for a linear model, you can perform similar model-checking for any generalized linear model. The emphasis is on checking the form of the model for the mean, either by cumulative residuals summed over a covariate value, or by cumulative residuals summed over the value of the linear predictor. If you compute cumulative residuals over the linear predictor, then the resulting supremum statistic and graphical analysis are more appropriate for checking the form of the link function. Cumulative residuals over a covariate value are more appropriate for checking the functional form of the covariate.

Marginal Models for Dependent Responses (GEEs)

You can also use the methods described here to check the form of the mean in a marginal model for dependent responses fit by solving generalized estimating equations (GEEs). Lin et al. (2002) apply cumulative residual analysis to the CD4 data taken from an AIDS clinical trial. The study randomly assigned 360 HIV patients to an AZT treatment and 351 to placebo. CD4 counts were measured repeatedly over the course of the study. Lin et al. (2002) considered two models to describe the time trend of the response. The first was a quadratic model

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

and the second was a cubic model

$$\begin{aligned} E(Y_{ik}) &= \beta_0 + \beta_1 T_{ik} + \beta_2 T_{ik}^2 + \beta_3 T_{ik}^3 \\ &\quad + \beta_4 R_i T_{ik} + \beta_5 R_i T_{ik}^2 + \beta_6 R_i T_{ik}^3 \end{aligned}$$

where R_i is an indicator variable for treatment for patient i, and T_{ik} is the time (in weeks) of the kth measurement of CD4 count for patient i.

The models described above were fit using GEE methods with a normal distribution and an independence working correlation assumption. To check the form of the two proposed models for the mean, cumulative residuals were computed and plotted for each of the two models. Figure 4 shows the cumulative residuals for the quadratic model, and the first 20 of 10,000 simulations of the cumulative residuals under the null hypothesis of a correct model for the mean.

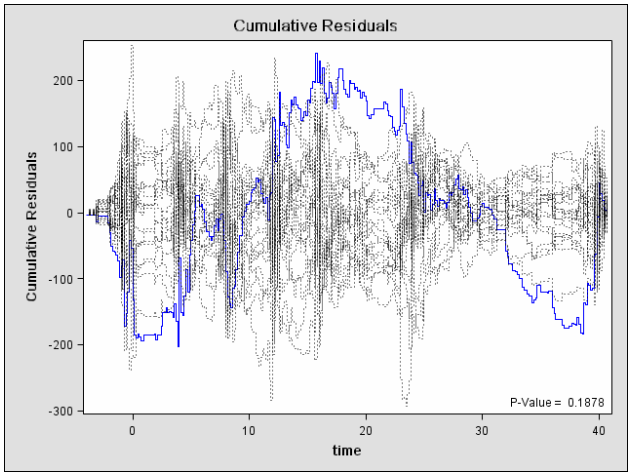

Figure 4. Cumulative Residuals for the CD4 Data, Quadratic Model

Figure 5 shows cumulative residuals and the first 20 realizations for the cubic model.

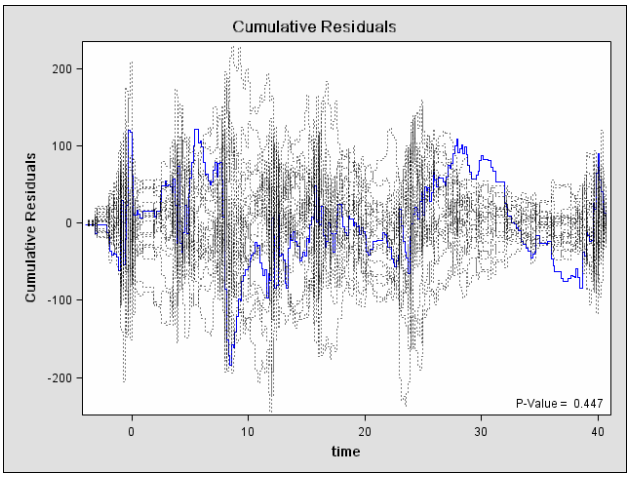

Figure 5. Cumulative Residuals for the CD4 Data, Cubic Model

The *p*-values, based on 10,000 simulations, are 0.188 for the quadratic model, and 0.447 for the cubic model. Based on these, and the shape of the cumulative residual curves, Lin et al. (2002) conclude that the quadratic model, while reasonable, is not entirely satisfactory. The cubic model appears to provide a better fit for the time trend.

Proportional Hazards Models

Checking the adequacy of the Cox model includes checking the functional form of a covariate, the link function, and the validity of the proportional hazards assumption. In PROC PHREG, the checking of the link function is omitted for two reasons: 1) the log link is desirable since it ensures that the hazard function is positive, and 2) the log link function is adequate when the correct functional forms of the covariates are used.

As an illustration of the techniques of Lin et al. (1993), their Mayo liver disease example is reproduced using PROC PHREG. The Cox model for primary biliary cirrhosis (PBC) data consists of five covariates: log(Bilirubin), log(Protime), log(Albumin), Age, and Edema. Parameter estimates are shown in Table 3.

Table 3. Parameter Estimates for PBC Data

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Variable Label
logbili	1	0.87068	0.08263	111.0333	<.0001	2.389	log(Bilirubin)
logalbumin	1	-2.53278	0.64819	15.2681	<.0001	0.079	log(Albumin)
logprotime	1	2.37969	0.76659	9.6364	0.0019	10.802	log(Protime)
age	1	0.03940	0.00765	26.5302	<.0001	1.040	Age
edema	1	0.85919	0.27114	10.0416	0.0015	2.361	Edema

To check the functional form of a covariate, the partial sums of the martingale residuals are computed and plotted against the values of the covariate. When the model fit is correct, the partial-sum process converges to a zero-mean Gaussian process whose distribution can be approximated by Monte Carlo simulations. First consider using the covariate Bilirubin instead of log(Bilirubin). Figure 6 displays the cumulative martingale residuals (in solid line) for Bilirubin and the first 20 of the 10,000 simulated curves (in light dotted lines).

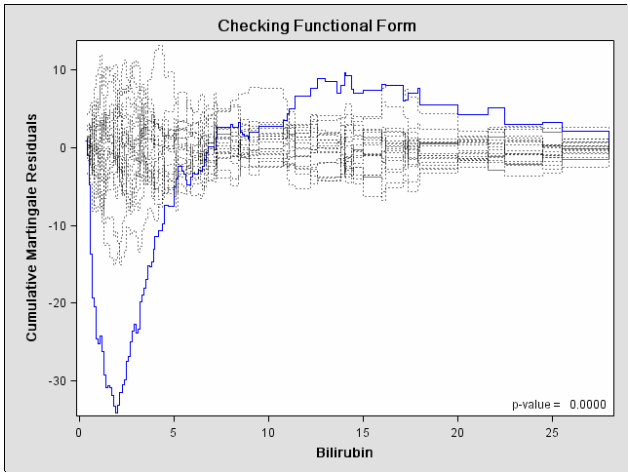

Figure 6. Cumulative martingale residuals for surgical unit data

The observed cumulative martingale residual process is obviously atypical of the first 20 simulated null processes; moreover, none of the 10,000 simulated paths has an absolute maximum exceeding that of the observed process. The fitted model overestimates the hazards for the low end of the Bilirubin values and underestimate the hazards for high Bilirubin values. The pattern suggests a logarithmic transform. When the log transform is applied to Bilirubin, the observed process appears to be more typical of the simulated processes (Figure 7). The *p*-value, based on 10,000 simulated samples, is 0.0572, indicating a much improved model.

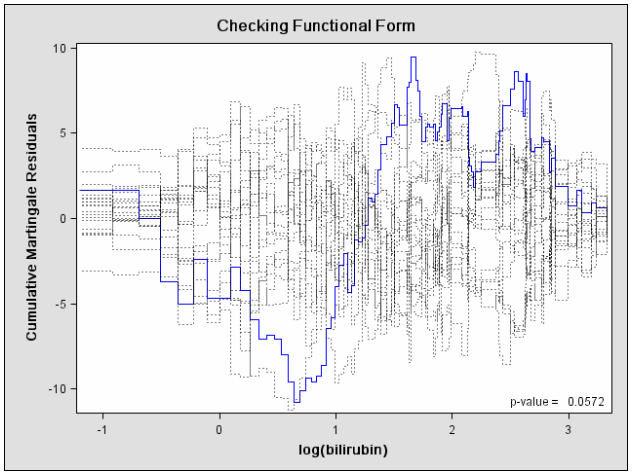

Figure 7. Cumulative Martingale Residuals for Revised Model

To check the proportional hazards assumption, the score process (which is a transformed partial sum process of the martingle residuals) is compared to the simulated processes under the null hypothesis that the proportional hazards assumption holds. Figure 8 shows the observed standardized score process for log(Protime) and the first 20 of 10,000 simulated null processes, revealing violation of the proportional hazards assumption. As Lin et al. (1993) suggests, the violation may be corrected using time-dependent covariates or stratification.

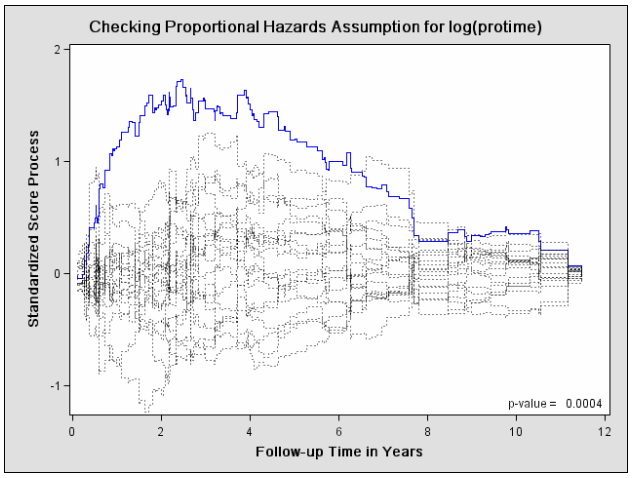

Figure 8. Cumulative Martingale Residuals for Revised Model

Conclusions

The model-checking techniques described here provide more objective methods for assessing model adequacy than have been previously available. Flexible numerical and graphical implementations of the techniques have been included in the SAS/STAT GENMOD and PHREG procedures for SAS Release 9.1.

References

Lin, D. Y., Wei, L. J., and Ying, Z. (2002), "Model-Checking Techniques Based on Cumulative Residuals," *Biometrics*, 58, 1-12.

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.

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

Contact Information

Gordon Johnston, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513.
Email gordon.johnston@sas.com

Ying So, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513.
Email ying.so@sas.com