MWSUG 2018, Paper HS-038

Assessing Model Adequacy in Proportional Hazards Regression

Michael G. Wilson
Indianapolis IN, USA

ABSTRACT

Proportional Hazards regression has become an exceedingly popular procedure for conducting analysis on right-censored, time-to-event data. A powerful, numerically stable and easily generalizable model can result from careful development of the candidate model, assessment of model adequacy, and final validation. Model adequacy focuses on overall fitness, validity of the linearity assumption, inclusion (or exclusion) of a correct (or an incorrect) covariate, and identification of outlier and highly-influential observations. Due to the presence of censored data and the use of the partial maximum likelihood function, diagnostics to assess these elements in proportional hazards regression compared to most modeling exercises can be slightly more complicated. In this paper, graphical and analytical methods using a rich supply of distinctive residuals to address these model adequacy challenges are compared.

1. Introduction

1.1 The Model

Proportional Hazards (PH) Regression using a partial maximum likelihood function to estimate the covariate parameters in the presence of censored time to failure data (Cox, 1972) has become widely used for conducting survival analysis. The PHREG procedure in SAS®/STAT (SAS Institute, 2016) has appeared as the prevailing procedure with which to conduct such analyses. The specification for the model is:

$$\lambda(t) = \lambda_0(t)\, exp\left\{\sum_{i=1}^{p} \beta_i x_i\right\} \quad (1.1)$$

Where, $\lambda_0(t)$ is the baseline hazard function and is a non-negative arbitrary hazard function when all covariates are zero.

The summation in braces, which is sometimes called the risk score in proportional hazards, when expanded is given by:

$$\left\{\sum_{i=1}^{p} \beta_i x_i\right\} = \left[\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right] \quad (1.2)$$

1.2 Classes of Model Assumptions

Model adequacy in the PH regression model has two classes of assumptions, that when satisfied ordinarily allow one to rely on the resultant statistical inferences and predictions. The first assumption is that time is independent of the covariates in the hazard function. In other words, the ratio of the hazard function for two individuals with different regression covariates, does not vary with time. This is more commonly known as the PH assumption.

The second assumption is that the relationship between log cumulative hazard and a covariate is linear. Some research methodologists refer to this as the linearity assumption.

Several approaches to detecting, testing and modeling non-proportional hazards are available in the literature. There are several reputable sources providing guidance on identifying and modeling non-proportional hazards (Wilson, 2010). Those approaches have been extensively evaluated and shown to perform satisfactorily (Michael Schemper, Wakounig, & Heinze, 2009).

1.3 Verification

Fewer resources are available that focus on verifying the second assumption of model adequacy regarding the relationship between the log cumulative hazard and the covariate. The presence of missing, or incorrect covariates, incorrect functional forms and highly influential observations are known to produce a violation of this second assumption. The application of a statistical method to data in which the model assumptions are violated can result in wrong conclusions. Fortunately, diagnostics are available in

the form of residuals and methods to assess these potentially detrimental precursors.

Verifying that assumptions are satisfied for PH models is slightly more complicated than it is for general linear regression for at least three reasons. Firstly, PH regression directly models the hazard function and not simply dependent observations. Secondly, estimates of the modeled hazard function are difficult to display, so substitutions are often used. Thirdly, failure-time data are usually distributed by the exponential, the Weibull, or the log-normal, which might be less familiar to the analyst than normally-distributed data.

1.4 Data Patterns and Methods

For illustrative purposes, two synthetic, doubly-censored, time-to-event datasets were generated. These datasets are patterned after a retrospective chart review of the effect of early vs. late tracheostomy on survivorship for 88 consecutive patients undergoing thoracic surgery at a particular research institution (Ladowski, Ladowski, & Wilson, 2013).

Tracheostomy is commonly conducted procedure in critically ill patients. It has many potential advantages but the procedure is not without modest risks. However, the effect of the timing of the procedure has on survivability is not well documented (Griffiths, Barber, Morgan, & Young, 2005). The National Association of Medical Directors of Respiratory Care recommended that translaryngeal (endotracheal) intubation be used only for patients requiring less than 10 days of artificial ventilation. They further recommended tracheostomy should be placed in patients who still require artificial ventilation 21 days after admission. These recommendations are based only on expert opinion, descriptive review (Kane, Rodriguez, & Luchette, 1997) and a systematic review (Maziak, Meade, & Todd, 1998), which did not include a formal meta-analysis of the data.

For both the confirmatory (n=500) and pilot (n=120) datasets, nine (p=9) covariates were generated including, (1) an indicator variable for early vs. late tracheostomy (0, if early, or <= 10 days; 1, if late, or > 10 days), (2) serum creatinine (in mg/dl), (3) continuous age (in years), (4) body-mass index (kg/m^2), (5) glycosylated hemoglobin (percent), (6) fasting levels of low-density lipoprotein (mg/dL), (7) systolic blood pressure (mmHg), (8) pre-operative Forced Expiratory Volume in one second (FEV1; in L), and (9) number of previous surgeries. The continuous covariates were generated with balance within the categorical covariate. Although in the original dataset, statistically significant interactions were observed between creatinine and age, these datasets were simulated without it or any other interaction.

The failure times were generated from the proportional hazards case of the exponential hazard by selecting random failure time from the Weibull hazard, $h(t) = \lambda\gamma(\lambda t)^{\gamma-1}$, where gamma ($\gamma$) is 1. All failure times were non-negative and their distribution right skewed. The two censoring mechanisms were (1) singly, fixed (Type I) at ten years and (2) random with a small, but non-zero λ hazard set to generate a one hundredth percent dropout.

A proportional hazards model was fit to these datasets. The models were examined for adequacy using several diagnostics offered in the PHREG procedure. Admittedly, these datasets are for illustrative purposes and are without imperfections analysts can possibly find experimentally and empirically. Namely these data structures have six agreeable features. First, the multiplicative structure (Equation 1.1) of the model (Fleming and Harrington, 1991) and not additive (Aalen, 1989) is appropriate. Secondly, the effects from missing data have been contained (Horstman 2013). Thirdly, competing risks have been regulated (Gooley, Leisenring, Crowley, & B Storer, 1999) and (Dagis, 2010). Fourthly, informative censoring has been reduced (Allison, 1995). Fifthly, separation or the problem of monotone likelihood has been Firth's corrected (Tsiatis, 1981) and there are no failure to converge. Finally, any non-proportionality has been managed (Grambsch & Therneau, 1994). These diagnostics might not perform as expected in the presence of these structural issues.

The application of generalized, Martingale, deviance, and score residuals are explored to assess general lack of fit, incorrect or missing covariates, incorrect functional form, and impact of extreme observations on the parameter estimation.

2. General Lack of Fit

2.1 Estimation of the Cumulative Hazard

In proportional hazards regression, a likelihood function is maximized to obtain parameter estimates and estimates of the cumulative hazard function or adjusted survival function (Equation 2.1). This semi-parametric method of estimation for proportional hazards model, properly called the method of partial maximum likelihood (PL), is remarkable on its own and is one of the most significant ideas of modern statistical theory. It is so significant in applied statistics that many authors have asserted that its importance eclipsed the PH model itself. It is slightly different than the method of maximum likelihood estimation (Fisher, 1925) in that the number of terms it contains is equal to the number of untied

events (D) and none for censored observations. It is semi-parametric since there is no need to specify the baseline hazard function, $\lambda_0(t)$. As indicated earlier, the baseline hazard function is a non-negative arbitrary hazard function when all covariates are zero.

$$PL = \prod_{i=1}^{D} L_i \qquad (2.1)$$

The L_i terms are a ratio of the hazard function for the i^{th} individual who experiences the event at time t_i in the numerator and the sum of the hazard function for all individuals who have not yet experienced the event (including some individuals who will be censored later). These individuals comprise what is called the risk set at time t_i.

Similar to many semi- and non-parametric methods, the PL depend on the ranks of the event times. So, if the actual event times are monotonic transformations like adding a constant, multiplying by a constant, or taking the logarithm, the estimated coefficients are unchanged. Also, the estimates from PL are not fully efficient, so the standard errors are slightly larger when compared to using the entire likelihood function (Bradley Efron, 1977).

The benefit is that the estimates are robust regardless of the actual shape of the baseline hazard function. The beta estimates are, however, consistent and asymptotically normal.

For each of the D terms, the L_i are the hazards for the individual subject that has the event in the interval divided by the sum of the hazards for all subjects at risk for the event in the interval. The denominators for the L_i are called W_i and are also used in the estimation of the empirical baseline cumulative hazard function for discrete failure times given in Equation 2.2.

$$\widehat{H_0}(t_i) = \widehat{\Lambda_0}(t_i) = \sum_{t_i<t} \frac{d_i}{W_i} \ , for\ i = 1, 2, \dots, D. \qquad (2.2)$$

The d_i are the number of failures in the interval (t_{i-1}, t_i). This is a step function that jumps at observed failure times. When the covariates from the PL are zero, equation 2.2 reduces to the Nelson-Aalen estimators, which have been available in SAS/STAT 9.4 in PROC LIFETEST. Adjusted survival estimates are the Napierian base, *e*, raised to the arithmetic inverse of these values.

2.2 Generalized Residuals

Generalized Residuals sometimes referred to as Cox-Snell residuals, can be used to assess the overall fit of a model based on a proportional hazards regression. If the PH model (Equation 1.1) is correct, the Cox-Snell residual is defined as the negative log of the survival estimate for a given subject (Equation 2.3). The inverse of this residual is precisely provided in PHREG using the OUTPUT statement using the *keyword=name* convention where *name* is the logarithm of survival.

$$r_j = \widehat{H_0}(t_j) \exp\left\{ \sum_{k=1}^{p} Z_{jk}\ b_k \right\}, \quad for\ j = 1, 2, \dots, n. \qquad (2.3)$$

Plots of these residuals can provide an impression of the overall fit. The plot of these residuals is similar to the empirical cumulative density plots from linear models, which include a reference line for the normal distribution. When the cumulative hazard rate, given by (Equation 2.4),

$$\widehat{\Lambda}(t) = \int_0^t \lambda(x)dx \qquad (2.4)$$

is plotted against a sample from a unit exponential distribution, it will follow a 45-degree line on Cartesian coordinates. If the PH model is correct, then the Generalized Residuals will appear to be a censored sample from a unit exponential distribution and fall roughly along the 45-degree line as shown in Figure 2.1 for a moderate sized study and Figure 2.2 for a smaller sized study. Values above the 45-degree line are those where the model over-predicts failure and conversely values below the reference line are those where the model under-predicts failure.

Generalized Residuals can be used to examine if separate levels of subgroups based on an included covariate share the same baseline hazard. Two plots are created based on generalized residuals from two PH regression analyses using the BY statement in PHREG, each of the levels separately and stratified using the STRATA statement (See Table 2.1). In the first plot, overlay the residuals from the two separate models as in Figure 2.3. In the second plot, overlay the residuals from the two strata (Figure 2.4).

Table 2.1 SAS Commands to fit the Stratified Proportional Hazard Model and Plot the Generalized Residuals

```
proc phreg data = _cencov01;
     strata trt;
     model survtime*event(0) = cc1 cc2 cc3
           / ties = &ties;
     output out = _genres08 LOGSURV = h
           / method = ch;
             run;
data _genres09;
     set _genres08;
     csresid = -h;
     cons = 1;
     run;
proc sort data = _genres09;
     by trt;
     run;
proc phreg data = _genres09;
     by trt;
     model  csresid*event(0) = cons;
     output out = _genres10 logsurv = ls
           /method = ch;
           run;
data _genres11;
     set _genres10;
     if trt = 0 then haz0 = -ls;
     if trt = 1 then haz1 = -ls;
     run;
```

The difference in the height of these Cox-Snell plots for separate, or stratified, treatment groups is the difference between groups in the empirical cumulative hazard function, $\{\widehat{\Lambda}(t)\}$. Therefore, it can give you a hint into the group differences.

Admittedly, there are at least three limitations of the generalized residual plots. Firstly, the interpretation is less intuitive again because the shape of the exponential distribution is less familiar to the analyst. Secondly, the rationale for the expected values of x-prime beta following the unit exponential distribution isn't immediately obvious. Finally, the reference line is the expected of the expected. It is the expected from the model with the expectation the model fits well.

2.4 Generalized R-Squares

Two Generalized Neigelkirke R-Squares have been proposed by some authors as a measure of overall fit (Heinzl, 2000). The SAS code used to calculate them is provided in Table 2.3. R-square values are in Table 2.2 for three models.

Table 2.3 SAS Commands required to Calculate the Generalized R-Square

```
data gt02;
     set gt01;
     length str $64.;
     if lowcase(test) =: 'likelihood';
     genrsq01 = 1 - exp(-1*(ChiSq/&nobs));
     rsqunadj = 1 - ( exp
              (-1*(ChiSq/2))**(2/&nobs) );
     put 'The Generalized (Cox-Snell)
          R-Square value is '  genrsq01;
     str = "The Generalized (Cox-Snell)
            R-Square value is";
     value = genrsq01;
     run;
data fs02 fs03;
     set fs01;
     length str $64.;
     if lowcase(criterion) =: '-2';
     w01 = exp(-1*WithoutCovariates/2);
     w02 = exp(-1*WithCovariates/2);
     w03 = exp(-1*(WithoutCovariates-
           WithCovariates)/2);
     w04 = 2/&nobs;
     r2unadj = 1 - w03**w04;
     r2max   = 1 - w01**w04;
     genrsq02 = r2unadj/r2max;
     put 'The unadjusted Generalized
          R-Square value is ' r2unadj ;
     put 'The Generalized (Kent-Oquigley)
          R-Square value is ' genrsq02 ;
     output fs02;
     str = 'The unadjusted Generalized
            R-Square value is ';
     value = r2unadj;
     output fs03;
     str = 'The Generalized (Kent-Oquigley)
     R-Square value is ';
     value = genrsq02;
     output fs03;
     run;
```

Table 2.2 R-square values for the Complete Model, the Model including an incorrect covariate and the Model with a missing covariate.

	Complete Model	Incorrect Covariate	Missing Covariate
Kent-O'Quigley	0.55592	0.55592	0.72319
Cox-Snell	0.55481	0.55481	0.71977

3. Incorrect or Missing Covariates

3.1 Model Selection Procedures

In selecting covariates for any multiple regression model, the analyst needs to protect against two different type of errors. Firstly, including an incorrect covariate is a false positive, Type I error and increases the variability and reduces the precision of the model. Secondly, excluding a true predictor is a false negative, Type II error and increases the bias of the model. So, covariates unrelated to the outcome may reduce power but should not introduce bias. Conversely covariates spuriously related will. In this selection process, keeping several concepts in mind will help including, the modeling aim, power, selection size, subject matter expertise, and minimizing interactions.

Prior to initiating any assessment of model adequacy, it is useful to clarify the purpose or aim of the modeling. There are at least three purposes. Firstly, a single covariate is under investigation for its association with survival, but several other predictors exist for which there is an interest to adjust as in a randomized clinical trial. Secondly, a collection of factors of known relevance are under investigation for their ability to predict survival for example when the interest is in developing a prognostic index. Thirdly, where a collection of factors is under investigation for their potential association with survival, possibly with additional known factors as when the interest is in reducing the number of covariates. Although this list is not exhaustive, these purposes drive the choice of suitable model adequacy criterion (Bradburn, Clark, Love, & Altman, 2003). The first purpose has been chosen for the illustrative purposes in this paper.

In addition, it is important to keep in mind that the power and the assessment of model adequacy are related to the number of events rather than the number of participants. Simulation work has suggested that at least 10 events need to be observed for each covariate considered, and anything less will lead to problems, for example, the regression coefficients become biased (Peduzzi P, Concato J, Feinstein AR, 1995) and (Kocak & Onar-Thomas, 2012).

In the case when the number of events is limited, additional covariates could be reduced to a single variable using principal components or another scaling technique. This single variable may not be interpretable, but using a single score could be better than deleting all covariates from consideration. In addition, it could also reduce potential problems with collinearity, as will be seen in the next section.

Subject matter knowledge should guide the selection of candidate predictors. Early deletion of those with little chance of being predictive or of being measured reliably will result in models with less over-fitting and greater generalizability (Henderson & Velleman, 1981). Too often, 'semi-automated' methods, such as stepwise selection which will be discussed more in a moment, are used. However, models based purely on statistical significance may not be meaningful or useful.

Likewise, careful inclusion of interactions in a statistical model is essential so that, if present, interactions represent a true phenomenon rather than general lack of fit of the model. Lists of types of plausible interactions have been made available by some thoughtful authors (M. Schemper, 1988).

As common as it is, stepwise selection less preferable as other methods available. Using simulation results, it has previously been shown to generate a misleading model with known incorrectly included covariates (Derksen & Keselman, 1992). On the other hand, the method of Best Subsets using Mallows' C(p) has been recommended (Hosmer & Lemeshow, 1999).

3.2 Method of Best Subsets

The method works as follows. Using the candidate terms, all possible subsets are fitted and then ranked within the number of fixed predictor variables (p) by the value of the Score Test chi-square statistic. The Score Test is based on the first derivative of the log likelihood, is sometimes called the Rao Test and can be used to test the global null hypothesis that all betas equal zero (Bera & Bilias, 2001). Each statistic has an asymptotic chi-square distribution with p degrees of freedom (Cook & DeMets, 2007). The value of p is also the number of betas in the model.

Criticisms of the Score Test are based on the idea that it is difficult to compare models of different sizes because the score test tends to increase with the number of predictors variables in the model. However, the Score Test can be used to approximate the value of Mallows' Cp. This statistic is a measure of model bias where large values of Mallows' Cp indicate an important variable was omitted from the model. For the full model, Cp = p (Mallows, 1973). Mallows' C(p) for reduced models can be approximated using the formula below:

$$Approximation\ to\ Mallows' C(p) = Score(q) + (p - q) \quad (3.1)$$

where, p = then number of parameters in the full model,
q = the difference between p and the number of parameters in the subset model.

Table 3.1 SAS Commands required to Calculate the Mallows' C(p)

```
ods output bestsubsets = bss01;
proc phreg data = _cencov01;
     model survtime*event(0) = &cov
           / ties = &ties selection = score
             best = 3  ;
             run;
data bss02;
     set bss01;
     call symput
('ChisqFullModel',scorechisq);
     call symput
('ParmsFullModel',numberinmodel);
     run;
%put ChisqFullModel = &ChisqFullModel;
%put ParmsFullModel = &ParmsFullModel;

data bss03;
     set bss01;
     format scoreq MallowsCp 8.4;
     scoreq = &ChisqFullModel-scorechisq;
     q = &ParmsFullModel-numberinmodel;
     MallowsCp = scoreq +
                (&ParmsFullModel - (2*q));
                 run;
```

The Method of Best Subsets does not necessarily maintain model hierarchy. Hierarchically well-formed (HWF) models are models that contain all main effects that were involved in interaction terms. Choose the first hierarchically well formulated model with a Mallow's C(p) lower than the number of variables in the model. Figure 3.1 shows an example of a plot of Mallows' C(p) for a dataset with 18 covariates.

Figure 3.2 and 3.4 show that when the model is missing a covariate the generalized residuals wonder away from the reference line. These observations are shuddering under the weight of the larger influence they must shoulder when a covariate is missing. The Figures 3.5 – 3.7 show the plots of Mallows' C(p) for the complete model, the model including an incorrect covariate and the model with a missing covariate. The R-square values were shown in Table 2.2.

The most stringent test of a model is an external validation, which is the application of the 'frozen' model to a new population. Validation is important because over-fitting is such a common problem, especially with small datasets. In the absence of external validation, using an internal validation (or sometimes called a hold-out) dataset, bootstrapping or cross-validation will help prevent including spuriously related covariates (Harrell, Lee, & Mark, 1996).

Shrinkage coefficient can be used to evaluate possible over-fitting (Van Houwelingen & Le Cessie, 1990). A concordance statistic (Hanley & McNeil, 1982) and Somers' D (Somers, 1962) serve as general discrimination indices. Bias can estimated for Somers D by bootstrapping 200 replicates (B Efron & Tibshirani, 1993). Acceleration can be estimated by a jackknife procedure (DiCiccio & Efron, 1996). The bias-corrected, accelerated confidence interval was constructed (Bradley Efron, 1987) as a means to gauge internal validity (Harrell et al., 1996).

3.3 Goodness of Fit

Models can be assessed for Overall Goodness-of-Fit. One test proposed by Gronnesby and Borgan, which partitions the data into G groups based on the ranked values of the estimated linear predictors (Gronnesby & Borgan, 1996). The test compares the observed number of events in each group to the model-based estimate of the expected number of events. Because the Gronnesby and Borgan test is asymptotically equivalent to the likelihood ratio test (May & Hosmer, 1998), it can be simplified to using partial likelihood ratio test.

$$-2\ln\left[\frac{L(\hat{\beta},0)}{L(\hat{\beta},\hat{\gamma})}\right] \sim X^2, \text{ with F-R } df. \qquad (3.2)$$

The problems with this test can identify include, having outliers in the data, omitting important terms in the model, such as interactions, needing to transform some of the predictor variables and having a non-linear relationship between the log hazard and the continuous predictor variables.

Interestingly, when using the tie-down Brownian process to assess the PH assumptions and the model is missing a covariate, the Score Process Plots will look like they violate the PH assumption. In those cases, you end up chasing a phantom problem and might damage the predictive power of your model. On the other hand, the ASSESS option are not sensitive to the inclusion of an incorrect covariate.

4. Incorrect Functional Form

The partial likelihood will yield parameter estimates for the covariates in the proportional hazards model that fit the hazard as a linear coefficient. However, this method assumes that the predictors operate linearly. If the relationship between an included covariate and the model fit is something other than linear then the interpretation of the hazard ratio would be incorrect. Therefore, the assessment of linearity, or sometimes called function form, is important. There are at least three methods that can be used to assess linearity, including the Method of Categorizing the Covariate, assessing the Martingale residuals and plots of the cumulative Martingale process.

4.1 Method of Categorizing the Covariate

In the Method of Categorizing the Covariate, categorize the covariate into k (4 or 5) quantiles. Construct k – 1, zero-one, indicator variables. Add them to the model. Plot the k – 1 parameter estimates against the k – 1 means of the categories. Add to the plot a point for the reference category (Mean of the reference category and beta = 0). Look for a relationship that is linear, quadratic or threshold.

Figure 4.1 shows a linear relationship for a continuous covariate that has been categorized into 4 quantiles. Although the slope for this linear relationship was generated as negative 1, this plot shows a slope of positive 1. This sign reversal is not surprising since the data were generated for the log-survival format and these estimates from PHREG are in the log-hazard format.

4.2 Assessing the Martingale Residuals

The second method is to assess the relationship between the Martingale residuals from the model without the covariate and the covariate. The Martingale is defined below in equation 4.1.

$$\widehat{M}_j = \delta_j - r_j \tag{4.1}$$

Remember the δ_j are 1 if died, 0 otherwise and the r_j closely approximate to the unit exponential, which has values from 0 to about 3. So, we can expect the values of the Martingale Residual to range between -3 and +1. As can be seen from this definition, their interpretation is the difference between the observed and expected. An important property of Martingale residuals is that they sum to zero, so their mean is also zero. In addition, the covariance between any two residuals is also zero (See equations 4.2 and 4.3 below).

$$\sum_{j=1}^{n} \widehat{M_j} = 0 \tag{4.2}$$

$$\mathrm{Cov}\left(\widehat{M}_i, \widehat{M}_j\right) = 0, \quad \text{for all } i \neq j \tag{4.3}$$

The Martingale residuals have been suggested as possible diagnostics for the correct functional form, PH assumption, leverage on the beta estimates and for lack of model fit (Therneau, Grambsch, & Fleming, 1990).

Plot these Martingale residuals against the value of the covariate for each subject. Fit a loess regression to the plot and look for relationship that is linear, quadratic or threshold. Figure 4.2 shows Martingale residuals against the value of the covariate for each subject. Since these values are observed minus expected, those values above the loess line are events, specifically failures, not predicted by the model. An alternative interpretation is that large positive values indicate that the observed death came before the model predicted it and large negative values indicate that the observed death came after the model predicted it. Looking at the figure, notice again how the Martingale residuals have a maximum value of +1. Also, the LOESS line with a smoothing parameter of 0.6 has been overlaid on this scatterplot to show the linear relationship with a continuous covariate.

When the functional form of the covariate is quadratic then neither the categorized quantile estimates of beta (Figure 4.3) or the loess line of the martingale residuals (Figure 4.4) are no longer linear. Neither diagnostic displays linearity for logarithmic function forms (Figure 4.5 and 4.6). Likewise, when the functional form of the covariate is z * log(z) then neither the categorized quantile estimates of beta (Figure 4.7) or the loess line of the martingale residuals (Figure 4.8) are no longer linear. Interestingly, this has a false negative impact on the graphic check for the PH assumption, which fails when in fact PH is not violated (Figures 4.9 and 4.10).

If the path of the observed loess line is above the abscissa, the covariate needs to be pulled back; this can be done by taking the logarithm. If the path is below then the covariate needs to be expanded; this can be done by squaring it.

4.3 Cumulative Martingale Process Plots

Finally, some authors have recommended the use of the Cumulative Martingale Process Plots, which is implemented in PHREG using the ASSESS option (Lin, Wei, & Ying, 1993). It is a tied-down Brownian Bridge of the cumulative sum of the Martingale process versus the covariate. The covariate must be in the model that generated the residuals. If the observed path crosses the simulated paths, it suggests there is a functional form violation. This method is useful for showing gross (crude) non-linearity. However, the Cumulative Martingales are not very sensitive for fine-tuning function form and would need to be used in conjunction with other checks to suggest a functional transformation. For example, the cumulative sum of the Martingale processes for the quadratic, logarithmic or z*log(z) covariates studied in this paper do not suggest a functional transformation.

5. Extreme Observations

As in linear models, extreme observations in PH regression are to be carefully assessed. Unlike linear models where the dependent variable can be evaluated independent of the model, all residuals in proportional hazards regression are some function or transformation of observed minus expected values. But that might be sufficient since the interest is only in the influence of the observation on the model anyway. So, there are two types of extreme observations in proportional hazards regression. The first type is where the individual records a relatively extended life and has a high-risk score as estimated by the model. The second type is where the individual records a relatively short life and has a low-risk score.

5.1 Framework for Assessing Extreme Values

The thorough analyst knows the database well, which includes having carefully identified extreme observations and in particular, and understands what those observations mean for the model. A three-step evaluation process is recommended, which is similar to the process used in the linear models (Thompson, Brunelle, & Wilson, 2002). First, determine if the observation is notable, then secondly, examine it for accuracy. If accurate then thirdly determine if it influences the model.

First, determine if the observation is notable. A notable observation is one that is distinguished from the others and by any definition, nearly every dataset contains notable observations but the decisive critical level must be selected judiciously. Large levels over-exclude and smaller levels over-include. Many understandably prefer the comfort of the 0.05 alpha level to identify those observations greater than 1.96 standard errors from the mean. It has been argued that not all data are normally distributed and that a 5% level is too exclusive and that 98th percentiles have performed well (Wilson, 2000). If not notable then it can be safely included.

Secondly, in practice by far the most common explanation for notable observations is that they contain recording, data entry, or coding errors. The reason it is recommended that all observations be systematically checked for correctness is that there is a danger in selectively targeting some observations for error checking. If found to be inaccurate then it should be corrected or deleted.

Thirdly, if an observation influences the model, attempt to understand what it insights it provides for strengthening the model. For example, is there another covariate that ought to be included or is there some non-linearity that needs that is inadequately modeled? If not influential then the observation can be considered for down-weighting or dampening (Hogg, 1979).

In this section, three datasets will be used to illustrate the detection of extreme values. The first dataset has no extremes values (None). The second dataset has four known extreme values two that strengthen each covariate for both categorical groups (Strengthen). The third dataset also has four known extreme values with two that weaken each covariate for both categorical groups (Weaken).

Authors have recommended for the identification of highly influential observations (Step 3) several, at least six, influence diagnostics: Martingale, Deviance, Score, DFBETAs, Leverage Displacement, and LMAX. Perhaps some perform better than others. Nevertheless, all are provided in the PHREG procedure.

5.2 Martingale Residuals and Extreme Values

In the previous section, it was shown that the Martingale residual gave a measure of the difference between the observed and the fitted value as expected from the model. This measure has been recommended as a candidate for the identification of highly influential observations (Therneau et al., 1990). A plot of the Martingale residuals from the model with no extreme values versus risk score is provided in Figure 5.1. The risk score was previously defined in Equation 1.2.

$$Risk\ Score = \sum_{i=1}^{p} \beta_i x_i = \left[\beta_1 x_1 + \beta_2 x_2 + \ \ldots \ + \beta_p x_p\right] \quad (5.1)$$

5.3 Deviance Residuals

As can be seen in Figure 5.1, Martingale residuals are highly skewed. Their maximum value is +1 and their minimum possible value is negative infinity. On the other hand, the Deviance residual, D_j, is defined by a transformation of the Martingale has a more normally-shaped distribution than the Martingale.

$$Deviance\ Residuals = D_j = Sign\left[\widehat{M_j}\right]\left\{-2\left[\widehat{M_j} + \delta_j \log\left(\delta_j - \widehat{M_j}\right)\right]\right\}^{\frac{1}{2}} \quad (5.2)$$

Figure 5.2, shows the relationship between Deviance and Martingale residuals. From equation 5.2, it can be seen that the Deviance Residual has a value of zero when the Martingale is zero. The logarithm tends to inflate the value

of the residual when the Martingale is close to one and shrink large negative values. In the presence of light to moderate censoring and no influential observations, the plots of the Deviance residual against the risk score will appear as normally-distributed scatter as can be seen in Figure 5.3. When there is heavy censoring, a large collection of points near zero distort the normally-distributed scatter. Nevertheless, possibly influential observations will have deviance residuals with large absolute values.

Even though it is a transformation, the Deviance residual, like the Martingale, is a measure of observed minus expected hazard. Therefore, large positive values indicate that the observed death actually came before the model predicted it. Likewise, large negative values indicate that the observed death came after the model predicted it.

Figure 5.4 and 5.5 show the Martingale and Deviance residuals for the dataset with four known extreme values. Figure 5.4 does not suggest any potential extreme observations with the possible exception of the individual with the risk score of 2.8 and who had a Martingale residual -3.6. Examination of the Deviance plot shows that this individual had a Deviance residual of -2.2, which is within the 98th percentile and the acceptable range for Deviance residuals. Because this observation has a negative Deviance residual that means the observed death came after the model predicted it.

The observations with risk scores of -1.2, -1.0, 0.1, and 0.9 cannot be identified in the plot of the Martingale residuals. On the other hand, in the Deviance residual plots they are obvious. These residuals were positive so the observed death actually came before the model predicted it.

Therneau et al. (1990) conducted Monte Carlo studies which show that both types of residuals detect extreme observations from subjects that lived longer than expected by the model. On the other hand, those individuals who die sooner than expected by the model are detected only by the deviance residual.

5.4 Method of Deleted Observations

If the sample size is small enough, the preferred method of checking the influence of individual observations is, for lack of a better term what will be called in this paper, the Method of Deleted Observations. In this method, using PHREG, estimate the *p* parameters using all of the n observations, as usual. Call those estimates beta-hat(k, n+1), where k = 1, 2, . . . , *p* and save them. Then temporarily delete the first observation in the dataset and re-estimate the *p* parameters, calling those estimates beta-hat(k, 1). Estimate beta-hat(k, 2) by deleting the second observation from the full dataset and beta-hat(k, j) by deleting the jth observation from the full dataset. Repeat for all n observations. The total number of beta-hat estimates generated will be (n+1). The influence of an observation, say j, has on the model parameter, k, is defined as shown in Equation 5.3:

$$Diff\ (k,j) = \left[\left(\hat{\beta}_{(k,n+1)}\right) - \left(\hat{\beta}_{(k,j)}\right)\right] \tag{5.3}$$

Where k = 1, 2, 3, . . ., p and j = 1, 2, 3, . . ., n. A plot of Diff(k, j) against j, the observation number, for each parameter k, can gage the influence of the *j*th observation on the *k*th covariate. If Diff(k, j) is close to zero, the jth observation has little influence, conversely large values suggest a large influence.

5.5 Score Residuals

The Method of Deleted Observations is not computational feasible with larger datasets. Fortunately an approximation of Diff(k, j) can be derived based on the score residual, S(k, j). The score residuals are a decomposition of the first partial derivative of the log likelihood and are defined by:

$$S_{jk}(t) = \int_0^t \left\{Z_{jk}(u) - \bar{Z}_k(u)\right\} d\widehat{M_j(u)}. \tag{5.4}$$

To assess the influence of an observation, first the score residual is evaluated at infinity and when all covariates are fixed at time 0, which reduces to:

$$S_{jk}(\infty) = \delta_j \left[Z_{jk} - \bar{Z}_k(T_j)\right] - \sum_{t_n \le T_j} \left[Z_{jk} - \bar{Z}_k(t_n)\right] * \exp\left(\underline{\boldsymbol{b}}\,\underline{\boldsymbol{Z}}_j\right) * \left[\widehat{H_0}(t_n) - \widehat{H_0}(t_{n-1})\right]. \tag{5.5}$$

Where k = 1, 2, 3, . . ., p and j = 1, 2, 3, . . ., n. The first term of the approximation shown in Equation 5.4 is the difference between the covariate Z(j,k) at the failure time and the expected value of the covariate at this time. This is recognizable as Schoenfeld's partial residual (Schoenfeld, 1982). Schoenfeld's partial residual is useful in the graphical assessment of non-proportional hazards as has been previously discussed and illustrated (Wilson, 2010), but not further elaborated here in the discussion of extreme observations.

Secondly, the final quantity is the product of Equation 5.5 and the inverse observed Fisher information. As a whole, these covariate-wise residuals gage the influence of the *j*th

observation on the *k*th covariate. Examples for the three covariates in the model without extreme values are provided in Figures 5.6, 5.7, and 5.8. Likewise, Figures 5.9, 5.10 and 5.11 illustrate the effect of the extreme values on each covariate for the Strengthen dataset. Finally, Figures 5.12, 5.13 and 5.14 illustrate the effect of the extreme values on each covariate for the Weaken dataset.

5.6 DFBetas

When it is discovered that a few observations seem to have an influence on the model, the next step is to estimate the size of that influence. DFBETAs are approximations of the difference in the parameter estimates Diff(k, j) = [beta-hat(k, n+1) – beta-hat(k, j)] when the *j*th observation is omitted. These variables are a weighted transformation of the score residual variables and have been shown to be good approximations.

The effect of the extreme values that strengthen the effects of the continuous covariates as measured by DFBETAs can be seen in Figure 5.15 and 5.16. Alternatively, the effect of the extreme values that weaker the effects of the continuous covariates as measured by DFBETAs can be seen in Figure 5.17 and 5.18.

5.7 Gharibvand Plots

In a previous tutorial of using SAS for survival analysis, Gharibvand suggested a Deviance residual bubble plot by risk with the diameter of the bubbles being proportional to the LMAX statistic (Gharibvand, 2008). Examples of the Gharibvand Plots are provided in Figures 5.19 and 5.20 for the strengthening and weakening datasets, respectively.

5.8 Combination Plots for Extreme Values

Here a modified Gharibvand plot is suggested by combining three residuals into a single plot. This plot has the percent change in the DFBETAs versus observation number using the Leverage Displacement statistic in place of the LMAX and labeling observations with extreme Deviance values. This plot converts the DFBETAs to a percent change scale which measure the overall effect an observation has on a given covariate and is more intuitive for some clients. The Leverage Displacement statistic is also a little easier to understand and has almost the same magnitude of the LMAX statistic. Finally, only extreme one-percent of Deviance values are labeled.

If the observations in the dataset have equal influence the plot will appear to be random scatter about the abscissa. Observations with the greatest percent changes in DFBETAs, with the largest diameter bubble and which are labeled are considered the most influential on the model parameter estimates.

Examples of the Combined Residual Plots for extreme values that strengthen the effects of the two continuous covariates dataset can be seen in Figure 5.21 and 5.22 and alternatively, the Combined Residual Plots for the extreme values that weaken the effects of the two continuous covariates dataset can be seen in Figure 5.23 and 5.24.

Interestingly, slightly larger percent change in the DFBETAs can be seen when a model is missing a covariate, since these observations have to shoulder a larger influence when a covariate is missing. So, if adding a covariate to a model causes the residuals decrease uniformly, an important missing covariate will probably have been found.

6. Clustered and Repeated Events

Until this point in the discussion, only PH regression diagnostics for independent events with a single occurrence have been considered. However, the analyst is too frequently confronted with datasets containing events that are neither.

6.1 Clustered Events

Consider a study of cell-based therapy for subjects with critical limb ischemia (CLI) for promoting amputation-free survival (Murphy et al., 2011). CLI pathogenesis can be systemic, as in the case of diabetes, and these subjects will as a result often have disease in their contralateral limb. Therefore, the assumption that the index and contralateral limb are independence might be suspect. Another example of a dataset from a research study of diabetic retinopathy can examine the time to macular edema in each of the subjects' eyes (Geraldes, Hiraoka-yamamoto, Matsumoto, & Clermont, 2012). The eyes of a single subject are not independent of each other and are therefore clustered.

6.2 Repeated Events

Secondly, the same adverse event, such as headache, can be reported repeatedly by the same subject over the course of a psychopharmacological clinical trial (Goldstein & Wilson, 1993). The field of clinical oncology has several examples of circumstances of repeated events. superficial bladder tumors have been known to reoccur (Wie, Lin, & Weissfeld, 1989). Repeated events from the same subject are likely to be correlated (Li & Lagakos, 1997). In the case of events with a positive intra-subject correlation, a subject with shorter time to first event is likely to have a shorter time to the next event. Without adjustment for these correlations the standard errors of the betas are incorrect. In general, these standard errors for the cluster-level covariates, like event,

would be under-estimated and the standard errors for the subject-level covariates would be over-estimated. Performing the analysis of repeated events without adjustment for the correlation of repeated can be misleading (Chaichana et al., 2012).

In addition to the analysis of time-dependent covariates, the exceptionally useful programming steps available in the PHREG procedure available in SAS/STAT, simple cases of the analysis of clustered or repeated events can also be implemented. When these programming steps are invoked, the ASSESS, BASELINE, OUTPUT statements are no longer available. Although understandable, no residuals are subsequently available for assessing model adequacy.

6.3 Intra-cluster Correlation Adjustment

In clustered events, failure times have an intra-cluster correlation. Adjustments for those correlations can be achieved by the analysis of a marginal proportional hazard model (Lee, Wei, & Amato, 1992) using a robust sandwich covariance matrix estimate or alternatively, use a shared frailty model where cluster effects are incorporated into the model as independently, identically-distributed, normal random variables (Lin, 1994) using the RANDOM statement, which has been available since SAS/STAT 9.3.

Analysis with PHREG for data with repeated time-to-event can be input using Counting Style Process of Input (Therneau & Grambsch, 2000). This input style allows for multiple records per subject. Ake and Carpenter describe an excellent data creation macro as well as an example of the PHREG syntax (Ake & Carpenter, n.d.). But again understandably, no residuals are available for assessing model adequacy. Martingale residuals and score residuals can be constructed by accumulating within subject and taking the average within covariate.

6.4 Analytic Approaches

In cases of these complex models, multiple analyses are recommended. Consider fitting the Intensity Model (Andersen & Gill, 1982) and the Proportional Means Model (Lin, Wei, Yang, & Ying, 2000). In these models, different estimates of the variance are used. In the Intensity Model, the COVM option is specified to use the model-based covariance estimate. In the Proportional Means Model, the COVB(AGGREGATE) option is used to estimate the robust sandwich covariance.

Two conditional models for the analysis of repeated events has proposed (Prentice, Williams, & Peterson, 1981). First, in a total time model, the time-to-event dataset is recoded to examine time to the (k+1) occurrence. A subject that experiences two occurrences provides the time to the second event. However, in the analysis for the third event, this subject is censored. Secondly, the time-to-event data can be re-coded in the gap time model.

Finally, it has been proposed that recurrent events be considered a special case of multivariate failure times and use a marginal approach (Wie et al., 1989). Authors have shown that the joint distribution of the vector of parameter estimates can be approximated by a multivariate normal distribution. This WLW method fits a proportional hazards model to each of the component times simultaneously and assisted by the STRATA ensuring identical baseline hazard function. The standard errors of the regression parameters are estimated using the robust sandwich covariance again with the COVS(AGGREGATE) option.

These models make slightly different assumptions so careful interpretation is recommended. Although SAS Documentation for provides excellent examples of implementing these 5 approaches, the issue of model adequacy in those examples is not considered.

Few researchers of statistical methodology provide guidance on the assessment of model adequacy for PH regression when events are clustered or repeated, although a tutorial on the frailty model, with some attention to analytical, non-graphical, assessment of model adequacy is available (Govindarajulu, Lin, Lunetta, & D'Agostino, 2011).

7. Summary

Validity of statistical inferences and predictions in PH regression can depend on how appropriate two fundamental assumptions are for the data. The first assumption is time independence of the covariates in the hazard function, that is, the PH assumption. The second assumption is that the relationship between log cumulative hazard and a covariate is linear.

It is possible or at least suspected that violations of the second assumption might be responsible for what appears to be violations of the first. Several examples were shown where data with known problems were fit to a model generating violations of model adequacy. Those models were also assessed for the PH assumption and found to have violated it also.

Methods for assessing model adequacy for proportional hazard regression were described. Several PH regression diagnostics were reviewed including the generalized, Martingale, deviance, and score. The application of these diagnostics to assess overall fit, covariate selection, functional form, and the leverage exerted by each subject in

parameter estimation. Examples were provided that illustrated how these inadequacies can result in misleading or invalid models. Some remedial measures for the analyst to implement were suggested.

8. References

Ake, C. F., & Carpenter, A. L. (n.d.). Extending the Use of PROC PHREG in Survival Analysis.

Allison, P. D. (1995). *Survival Analysis Using SAS: A Practical Guide*. Cary, N.C.: SAS Institute, Inc.

Andersen, P. K., & Gill, R. D. (1982). Cox's Regression Model counting Process: A Large Sample Study. *Annals of Statistics*, *10*, 1100–1120.

Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival analysis Part III: multivariate data analysis -- choosing a model and assessing its adequacy and fit. *British Journal of Cancer*, *89*(4), 605–11. http://doi.org/10.1038/sj.bjc.6601120

Chaichana, K. L., Zadnik, P., Weingart, J. D., Olivi, A., Gallia, G. L., Blakeley, J., … Quiñones-Hinojosa, A. (2012). Multiple resections for patients with glioblastoma: prolonging survival. *Journal of Neurosurgery*, 1–9. http://doi.org/10.3171/2012.9.JNS1277

Cox, D. (1972). Regression Models and Lifetables (with discussion). *Journal of the Royal Statisticial Society, Series B*, *34*, 187–220.

Dagis, A. (City of H. (2010). A Discussion of Generating Cumulative Incidence Curve in Cases of Competing Risk in Survival Analysis. *Procedings of the Western SAS Users Group*.

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*, 265–282.

DiCiccio, T. J., & Efron, B. (1996). Better bootstrap confidence intervals. *Statistical Science*, *11*(3), 189–228. http://doi.org/10.1002/sim.4134

Efron, B. (1977). The Efficiency of Cox's Likelihood function for Censored Data. *Journal of American Statistical Association*, *76*, 312–319.

Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, *82*(397), 171–185. http://doi.org/10.2307/2289144

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. (D. R. Cox, D. V Hinkley, N. Reid, D. B. Rubin, & B. W. Silverman, Eds.), *Refrigeration And Air Conditioning* (Vol. 57). Chapman & Hall. http://doi.org/10.1111/1467-9639.00050

Fisher, R. A. (1925). Theory of Statistical Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*. http://doi.org/10.1017/S0305004100009580

Geraldes, P., Hiraoka-yamamoto, J., Matsumoto, M., & Clermont, A. (2012). Activation of PKCδ and SHP1 by hyperglycemia causes vascular cell apoptosis and diabetic retinopathy, *15*(11), 1298–1306. http://doi.org/10.1038/nm.2052.Activation

Gharibvand, L. (2008). A Step-by-Step Guide to Survival Analysis. University of California, Riverside: WUSS. Retrieved from https://www.lexjansen.com/wuss/2008/tut/tut08.pdf

Goldstein, D. J., & Wilson, M. G. (1993). Adverse event frequencies generate hypotheses of efficacy and safety. *Clinical Pharmacology and Therapeutics*, *54*(3), 245–251. http://doi.org/10.1038/clpt.1993.144

Gooley, T. A., Leisenring, W., Crowley, J., & B Storer, B. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine*, *18*(6), 695–706. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10204198

Govindarajulu, U. S., Lin, H., Lunetta, K. L., & D'Agostino, R. B. (2011). Frailty models: Applications to biomedical and genetic studies. *Statistics in Medicine*, *30*(22), 2754–64. http://doi.org/10.1002/sim.4277

Grambsch, P. M., & Therneau, T. M. (1994). Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika*, *81*(3), 515–526.

Griffiths, J., Barber, V. S., Morgan, L., & Young, J. D. (2005). Systematic review and meta-analysis of studies of the timing of tracheostomy in adult patients undergoing artificial ventilation. *British Medical Journal*, *330*(7502), 1–5. http://doi.org/10.1136/bmj.38467.485671.E0

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. http://doi.org/10.1016/j.msea.2008.11.058

Harrell, F. E. J., Lee, K. L., & Mark, D. B. (1996). Tutorial In Biostatistics Multivariable Prognostic Models : Issues In Developing Models , Evaluating Assumptions And Adequacy , And Measuring And Reducing Errors. *Statistics in Medicine*, *15*(4), 361–387. Retrieved from http://www.unt.edu/rss/class/Jon/MiscDocs/Harrell_1996.pdf

Heinzl, H. (2000). Using SAS to calculate the Kent and O'Quigley measure of dependence for Cox proportional hazards regression model. *Computer Methods and Programs in Biomedicine*, *63*(1), 71–76.

Henderson, H., & Velleman, P. (1981). Building multiple regression models interactively. *Biometrics*, *37*, 391–411.

Hogg, R. V. (1979). Statistical Robustness: One View of Its Use in Applications Today. *The American Statistician*, *33*(3), 108–115.

Hosmer, D. W., & Lemeshow, S. (1999). *Applied Survival*

Analysis. New York: John Wiley & Sons.

Kane, T. D., Rodriguez, J. L., & Luchette, F. A. (1997). Early versus late tracheostomy in the trauma patient. *Respir Care Clin N Am*, *3*, 1–20.

Kocak, M., & Onar-Thomas, A. (2012). A Simulation-Based Evaluaton of the Asymptotic Power Formulas for cox Models in Small Sample Cases. *American Statistician*, *66*(3), 173–179.

Ladowski, J. S., Ladowski, J. M., & Wilson, M. G. (2013). Early versus Late Tracheostomy After Major Cardiovascular Surgery (in preparation). *Journal of Cardiothoracic Surgery*.

Lee, E. W., Wei, L. J., & Amato, D. A. (1992). Cox-Type Regression Anlaysis for Large Numbers of Small Groups of correlated Failure Time Observations. In J. P. Klein & P. K. Goel (Eds.), *Survival Analysis: Statem of the Art* (pp. 237–247). Dordrecht, Netherlands: Kluwer Academic Publishers.

Li, Q. H., & Lagakos, S. W. (1997). Use of the Wei-Lin-Weissfeld method for the analysis of a recurring and a terminating event. *Statistics in Medicine*, *16*(8), 925–40. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9160489

Lin, D. Y. (1994). Cox Regression Analysis of Multivariate Failure Time Data: The Marginal Approach. *Statistics in Medicine*, *13*, 2233–2247.

Lin, D. Y., Wei, L. J., Yang, I., & 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.

Lin, D. Y., Wei, L. J., & Ying, Z. (1993). Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals. *Biometrika*, *80*(3), 557–572. http://doi.org/10.2307/2337177

Maziak, D. E., Meade, M. O., & Todd, T. R. (1998). The timing of tracheotomy: a systematic review. *Chest*, *114*, 605–609.

Murphy, M. P., Lawson, J. H., Rapp, B. M., Dalsing, M. C., Klein, J., Wilson, M. G., … March, K. L. (2011). Autologous bone marrow mononuclear cell therapy is safe and promotes amputation-free survival in patients with critical limb ischemia. *Journal of Vascular Surgery*, *53*(6), 1565–74. http://doi.org/10.1016/j.jvs.2011.01.074

Peduzzi P, Concato J, Feinstein AR, H. T. (1995). Importance of events per independent variable in proportional hazards regression analysis II: accuracy and precision of regression estimates. *Journal of Clinical Epidemiology*, *48*, 1503–1510.

Prentice, R. L., Williams, B. J., & Peterson, A. V. (1981). On the Regression Analysis of Multivariate Failure Time Data. *Biometrika*, *68*, 378–379.

SAS Institute, I. (2016). *SAS/STAT(r) 14.2 User's Guide*. Cary NC.

Schemper, M. (1988). Non-parametric analysis of treatment-covariate interaction in the presence of censoring. *Statistics in Medicine*, *7*(12), 1257–1266. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3221690&tool=pmcentrez&rendertype=abstract

Schemper, M., Wakounig, S., & Heinze, G. (2009). The estimation of average hazard ratios by weighted Cox regression. *Statistics in Medicine*, *28*(19), 2473–2489. http://doi.org/10.1002/sim.3623

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, *69*, 239–241.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, *27*(6), 799–811.

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.

Therneau, T. M., Grambsch, P. M., & Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, *77*(1), 147–160. http://doi.org/10.1093/biomet/77.1.147

Thompson, W. L., Brunelle, R. L., & Wilson, M. G. (2002). Performance and Interpretation of Laboratory Test. In A. Cato, L. Sutton, & A. Cato III (Eds.), *Clinical Drug Trials and Tributions* (Second, pp. 65–78). Marcel Dekker, Inc.

Tsiatis, A. (1981). A Large Sample Study of the Estimates for the Integrated Hazard Function in Cox's Regression Model for Survival Data. *Annals of Statistics*, *9*, 93–108.

Van Houwelingen, J. C., & Le Cessie, S. (1990). Predictive value of statistical models. *Statistics in Medicine*, *9*(11), 1303–1325. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2277880

Wie, L. J., Lin, D. Y., & Weissfeld, L. (1989). Regression analysis of multivariage incomplete failuret time data by modeling marginal distributions. *Journal of the American Statistical Association*, *84*, 1065–1073.

Wilson, M. G. (2000). Lilly Reference Ranges. In S. S-C (Ed.), *Encyclopedia of Biopharmaceutical Statistics*. New York: Marcel Dekker, Inc.

Wilson, M. G. (2010). Assessing and Modeling Time to Event Data with Non-Proportional Hazards. *Procedings of the Mid-West SAS Users Group*, Paper 125-2010.

9.0 Notices

TRADEMARK INFORMATION

ACKNOWLEDGMENTS

The author thanks the 2018 MWSUG Co-chairs, Cindy and Dave, and the Health Sciences Section Co-chair, Jessica for accepting this contribution for presentation. In addition, he is appreciative of the mentor for an earlier version of this manuscript, Bob Hamer, who made suggestions for improvements and discovered errors. Any remaining errors are nevertheless the author's full responsibility.

CONTACT INFORMATION

Your comments and questions are encouraged. Contact me at:

Michael G. Wilson
11630 Diamond Pointe Ct.
Indianapolis, IN 46236
317.620.0077
michael.g.wilson@biostat.us

10.1 Index of Figures

Section Page
10.2 Figures for Section 2 (4 Figures) 15
10.3 Figures for Section 3 (8 Figures) 17
10.4 Figures for Section 4 (10 Figures) 21
10.5 Figures for Section 5 (24 Figures) 26

10.2 Figures for Section 2 (4 Figures)

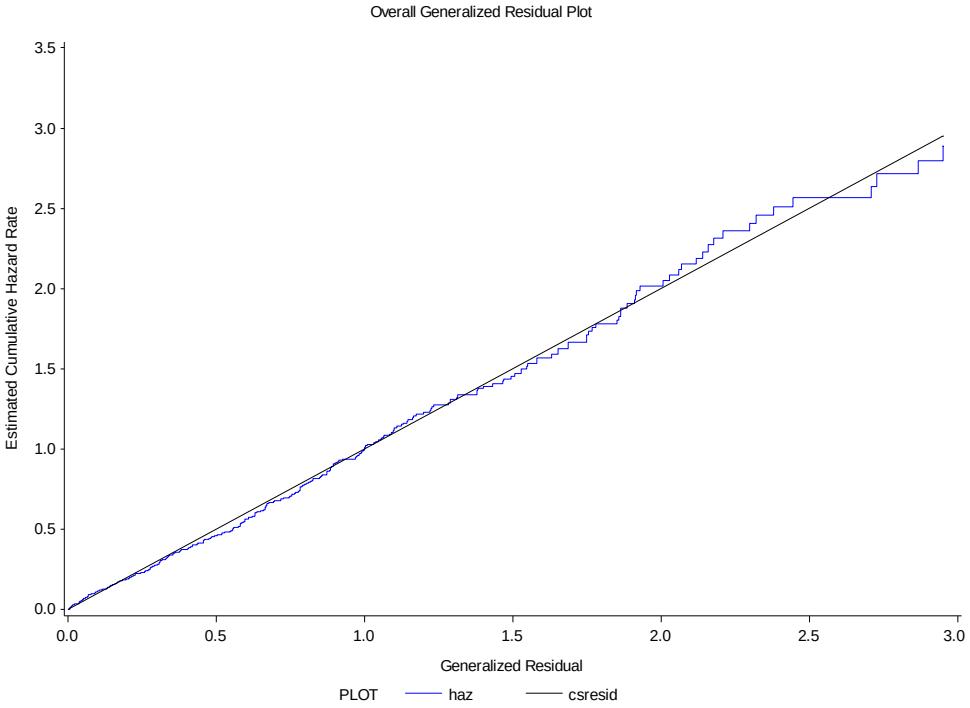

Figure 2.1: Overall Generalized Residual Plot for a confirmatory study (n=500), without incorrect or missing covariates, misspecification of the functional form, or extreme values. (Dataset = 04 {Confirmatory, No Inadequacies}; Graph Type = Cox-Snell Generalized {CS}; Covariates = 01 Categorical Treatment; One Level; [SimExp04_CS01])

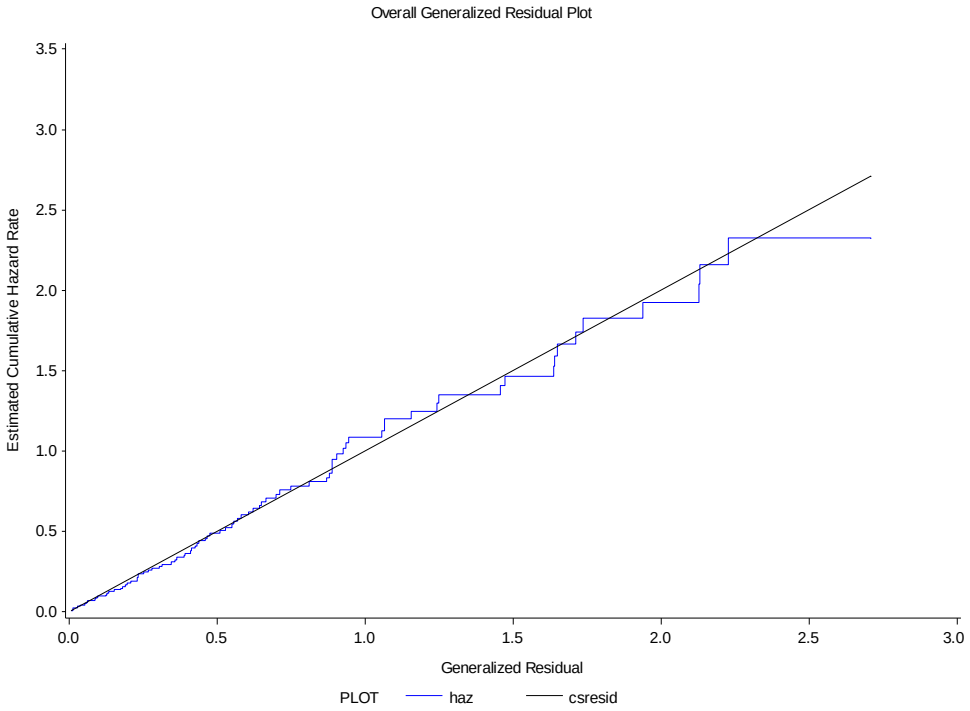

Figure 2.2: Overall Generalized Residual Plot for a small-to-moderate sized study (n=80), without incorrect or missing covariates, misspecification of the functional form, or extreme values. (Dataset = 51 {Confirmatory, No Inadequacies}; Graph Type = Cox-Snell Generalized {CS}; Covariates = 01 Categorical Treatment; One Level; [SimExp04_CS01])

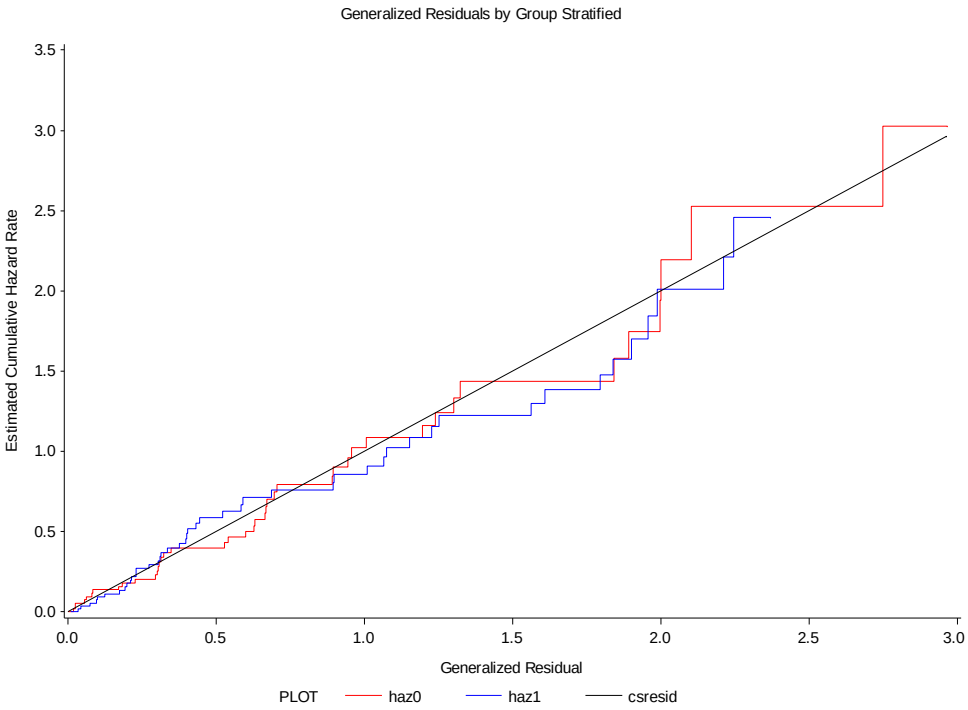

Figure 2.3 Overall Generalized Residual Plot for a confirmatory study (n=500), without incorrect or missing covariates, misspecification of the functional form, or extreme values. (Dataset = 42 {Confirmatory, No Inadequacies}; Graph Type = Cox-Snell Generalized {CS}; Covariates = 01 Categorical Treatment; Two Levels [SimExp42_CS02])

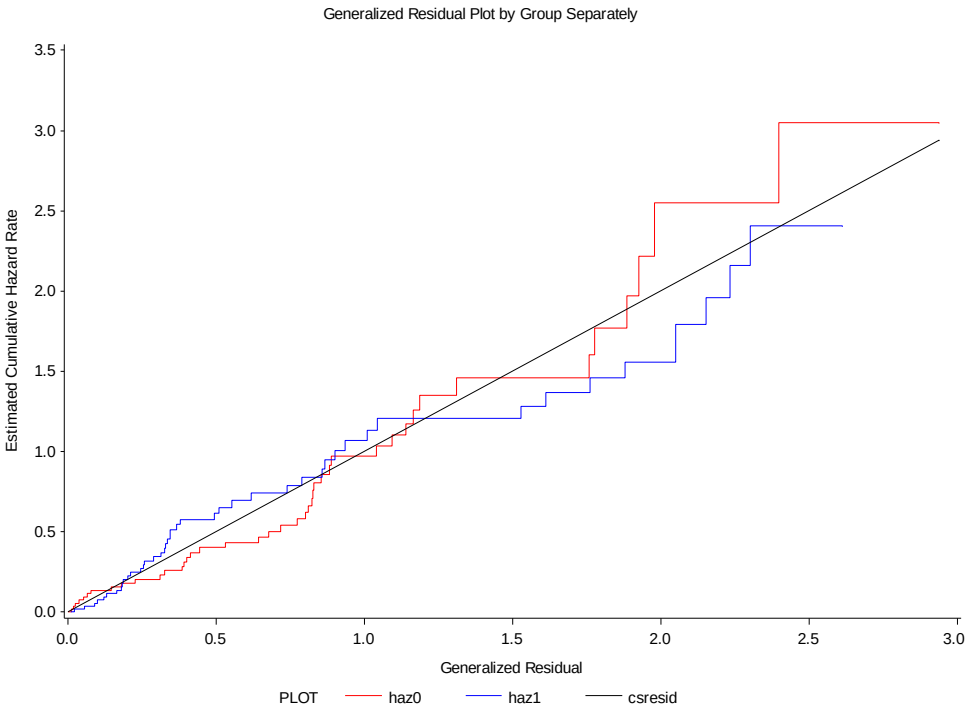

Figure 2.4: Overall Generalized Residual Plot for a small-to-moderate sized study (n=80), without incorrect or missing covariates, misspecification of the functional form, or extreme values. (Dataset = 42 {Confirmatory, No Inadequacies}; Graph Type = Cox-Snell Generalized {CS}; Covariates = 01 Categorical Treatment; Two Levels; [SimExp42_CS02])

10.3 Figures for Section 3 (8 Figures)

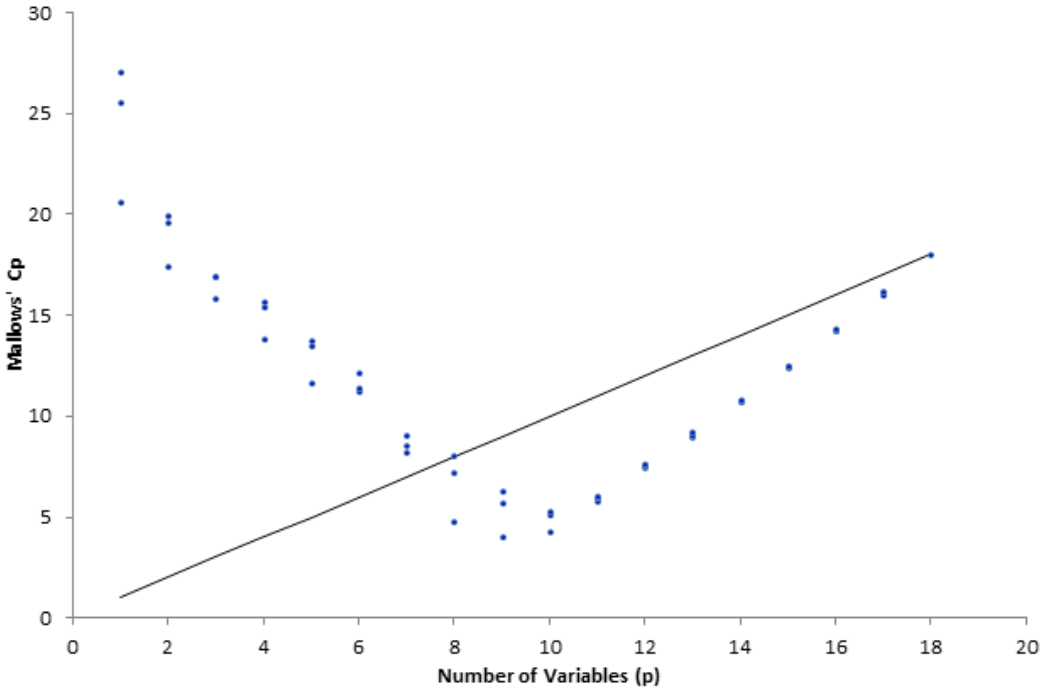

Figure 3.1: Mallows' C(p) is a measure of model bias large values indicate that an important variable was omitted from the model. Value below the reference line are a measure of bias [MallowsCp.xlsx].

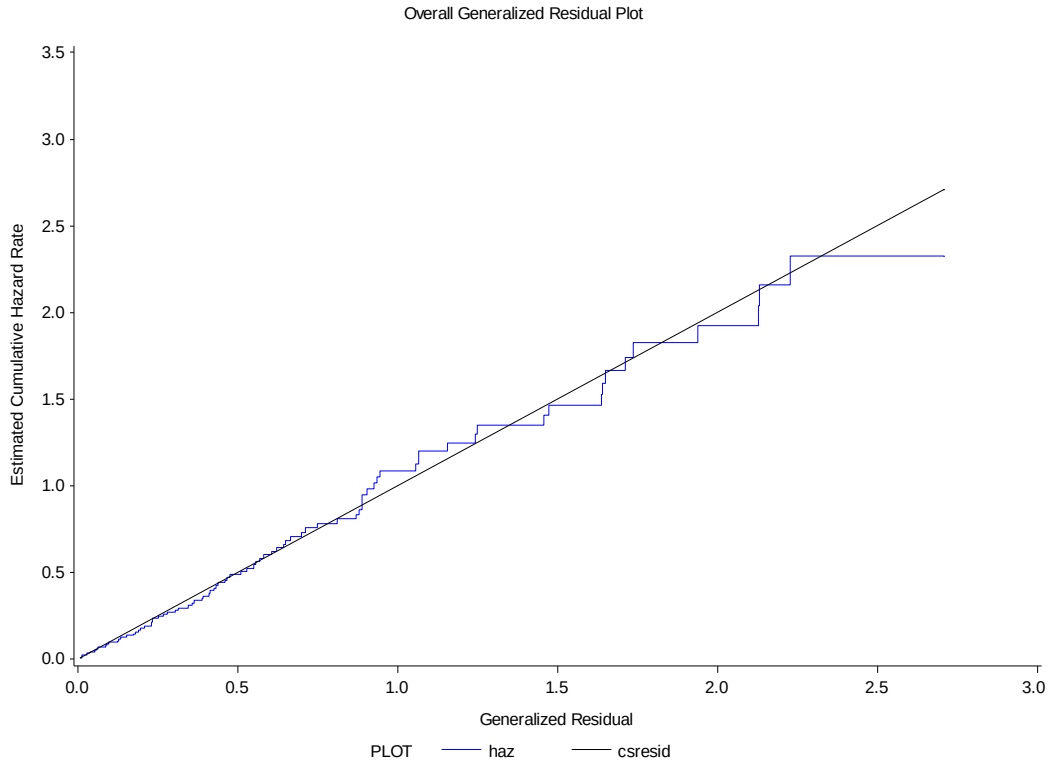

Figure 3.2: Cumulative Hazard for a Well-fit model (Dataset 30 [SimExp30_CS01]).

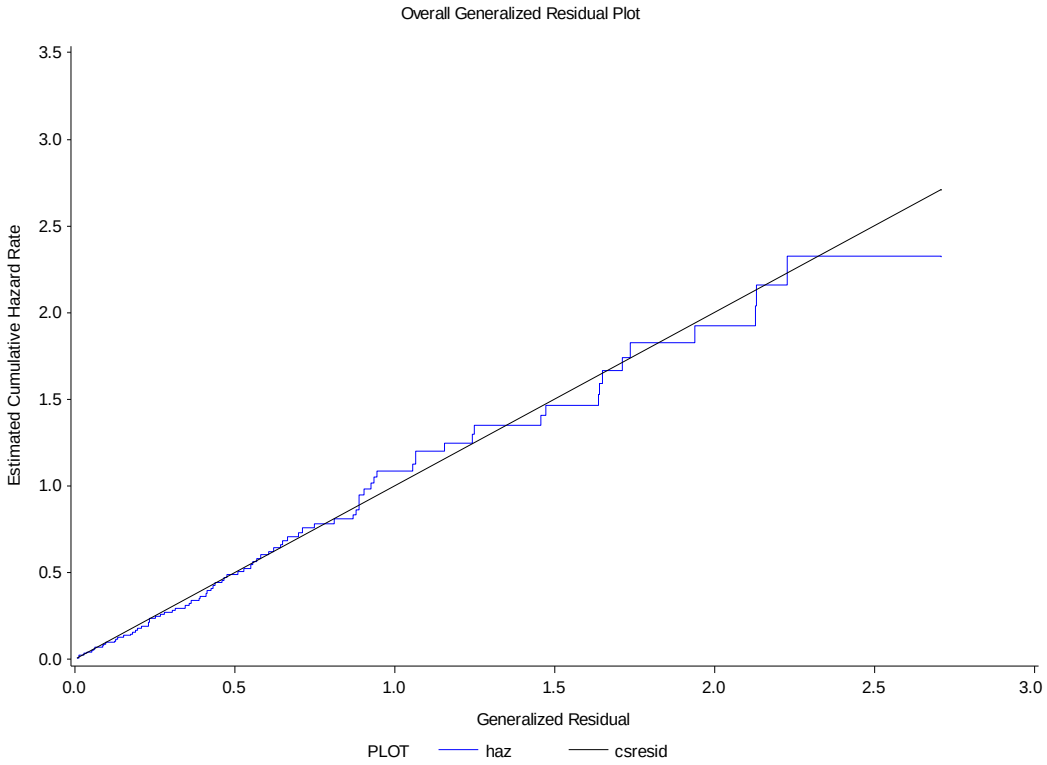

Figure 3.3: Cumulative Hazard for a model with an incorrect covariate included (Dataset 51 [SimExp51_CS01]).

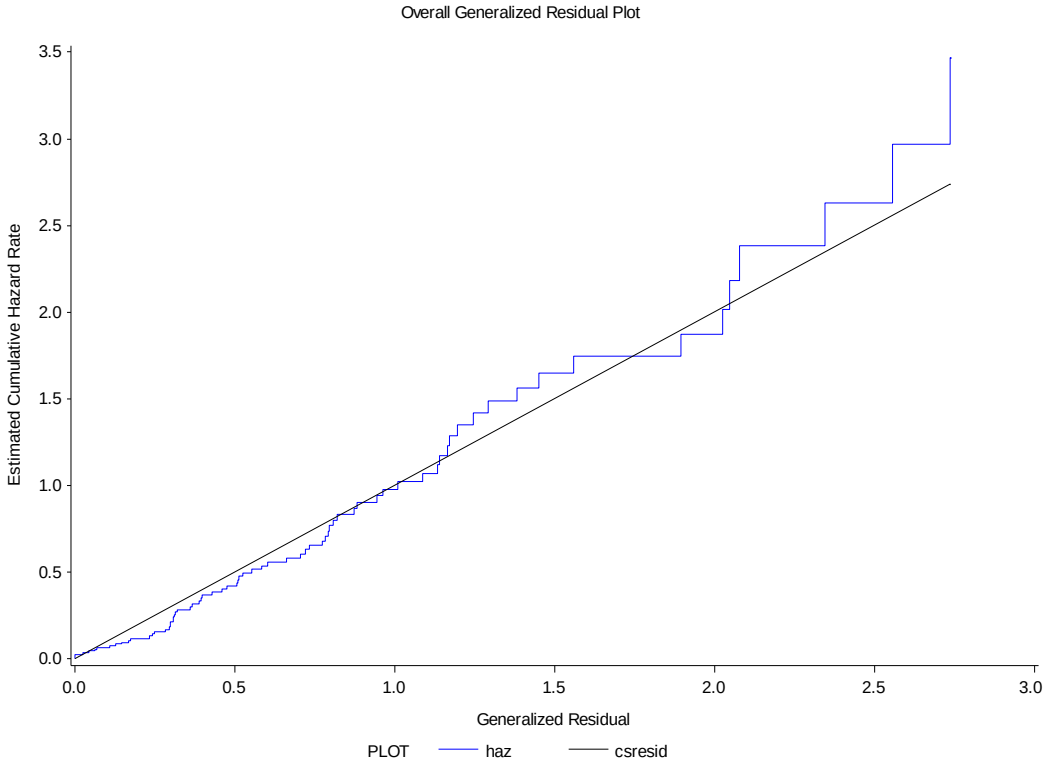

Figure 3.4: Cumulative Hazard for a model with a missing covariate (Dataset 31 [SimExp31_CS01]).

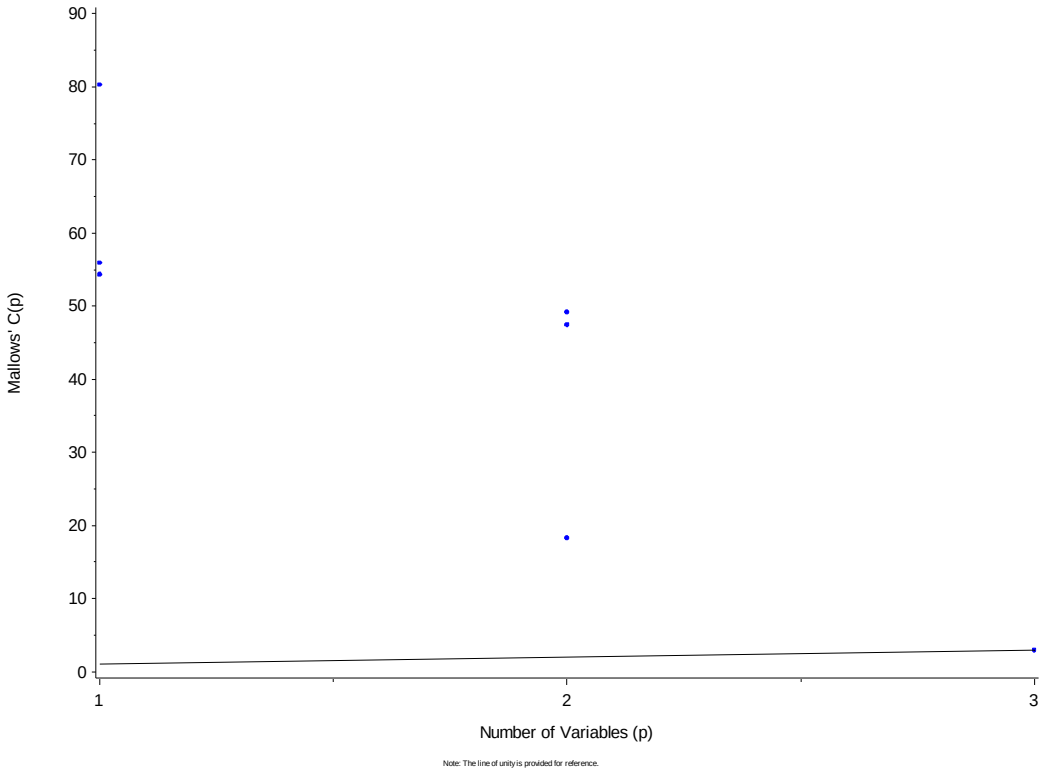

Figure 3.5: Mallows' C(p) plot for a well-fit model and shows no models that omitted an important variable since there were no values below the reference line are a measure of bias (Dataset 30).

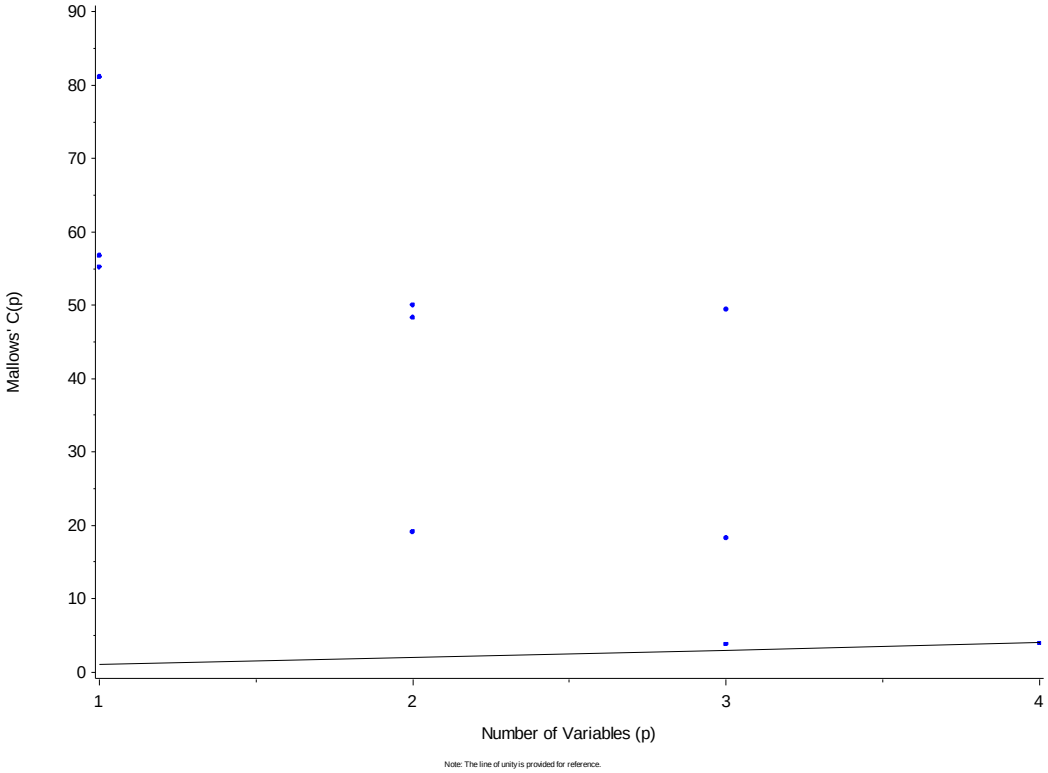

Figure 3.6: Mallows' C(p) plot for a for a model with an incorrect covariate included. It shows no important variable has been omitted (Dataset 51).

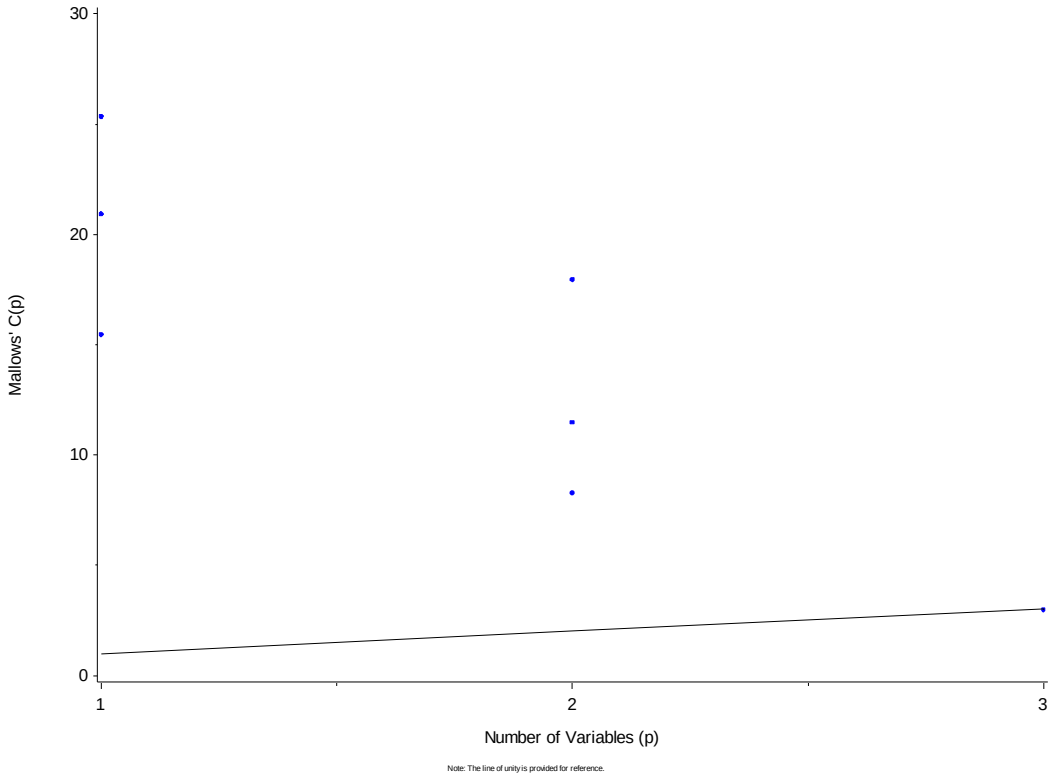

Figure 3.7: Mallows' C(p) plot for a model with a missing covariate (Dataset 31).

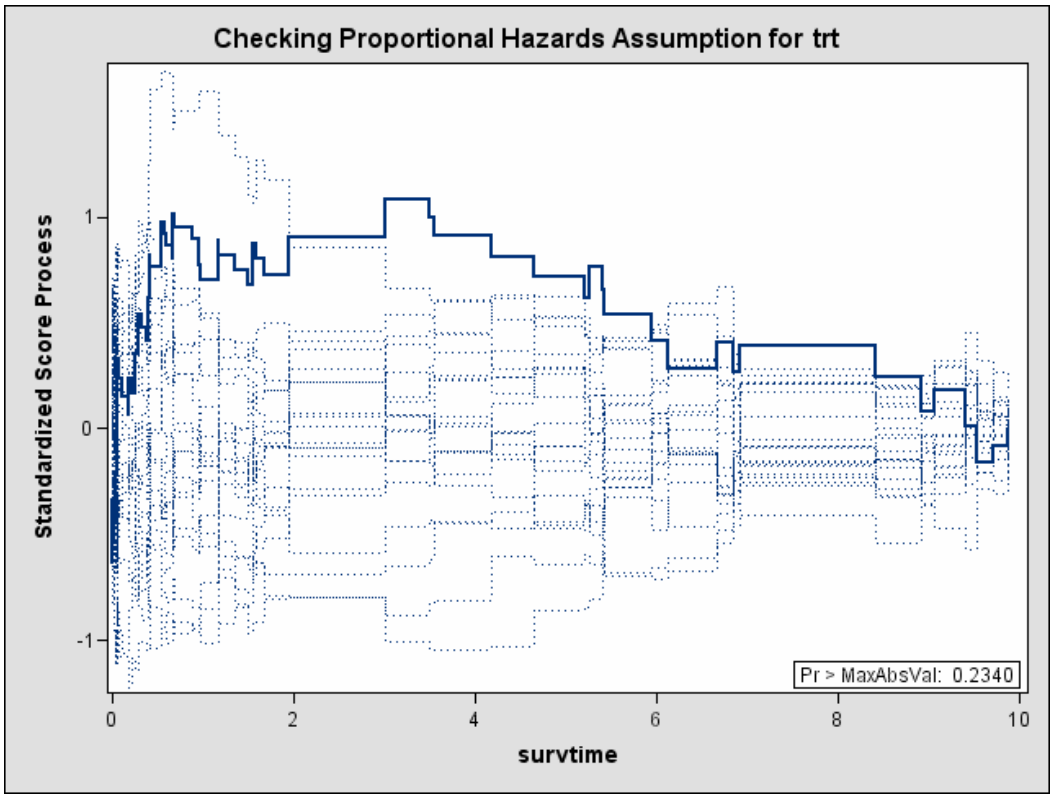

Figure 3.8: Standardized Score Process for a model with data whose hazards are proportional but a missing an important covariate (Dataset 31). The conclusion of this standardized score process graph alone misinforms the analyst.

10.4 Figures for Section 4 (10 Figures)

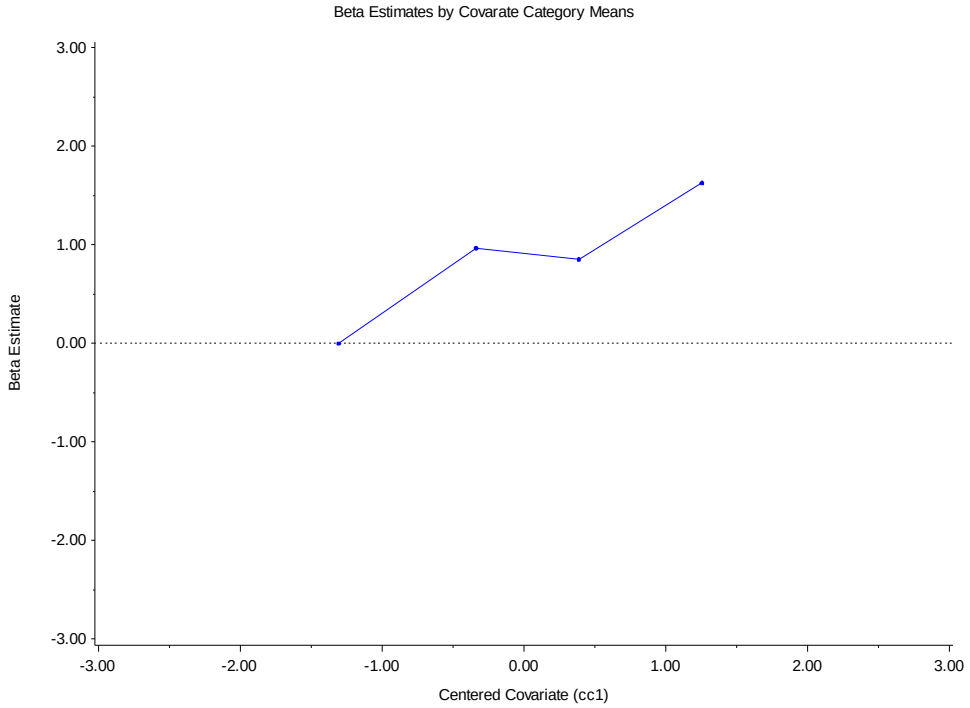

Figure 4.1: Categorized Quantile Estimates of Beta for a linear covariate. (Dataset = 41 {Confirmatory, No Inadequacies}; Graph Type = Quantile Categories {QC}; Covariates = 01 Categorical Treatment; Two Levels; [SimExp41_QC01.emf])

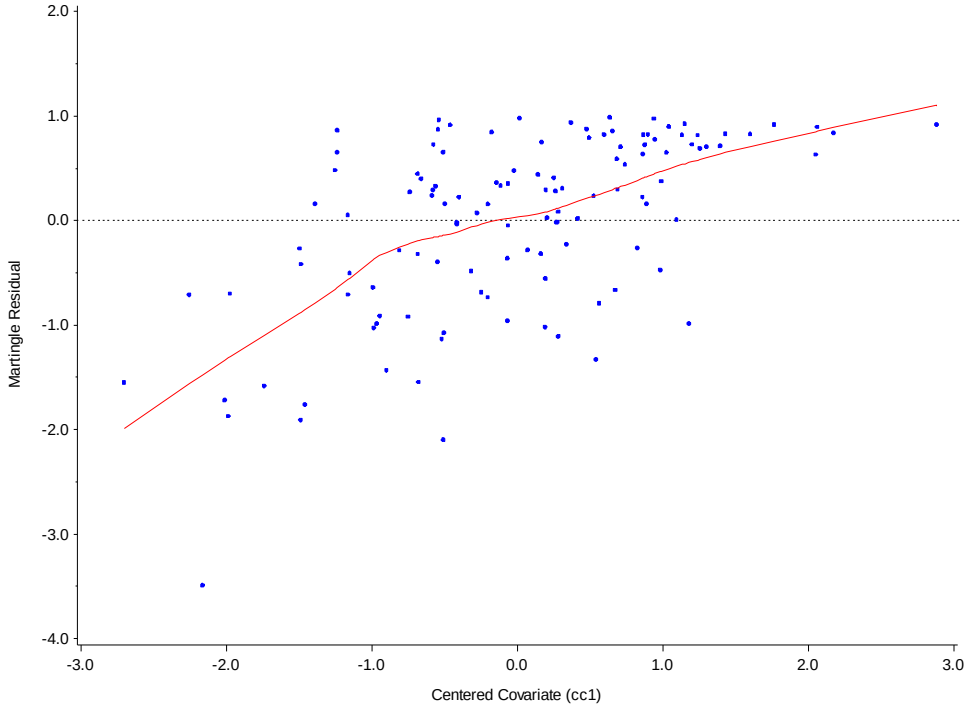

Figure 4.2: Cumulative Hazard for a Well-fit model (Dataset 30).

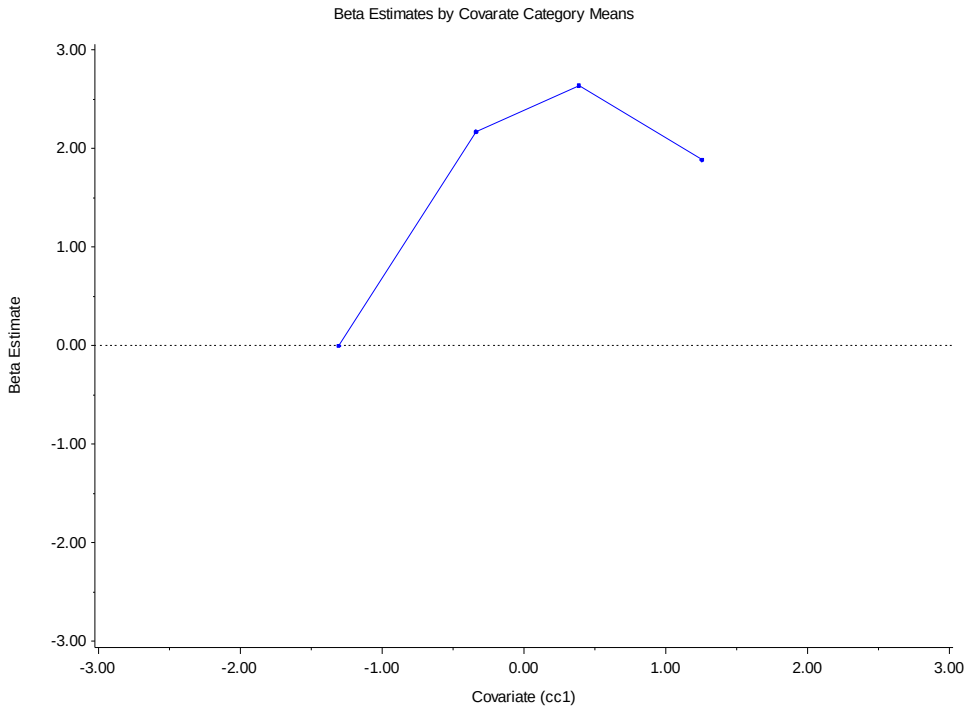

Figure 4.3: Categorized Quantile Estimates of Beta for a quadratic covariate. Dataset 42; SimExp42.

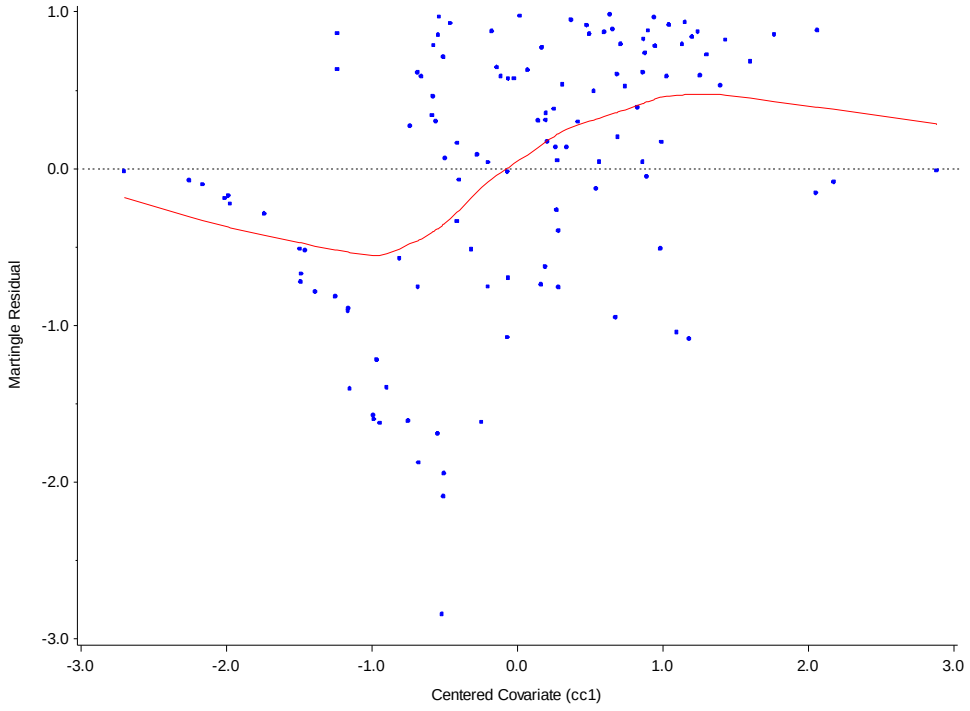

Figure 4.4: Martingale residuals and loess regression line for a model containing a quadratic covariate. SimExp42

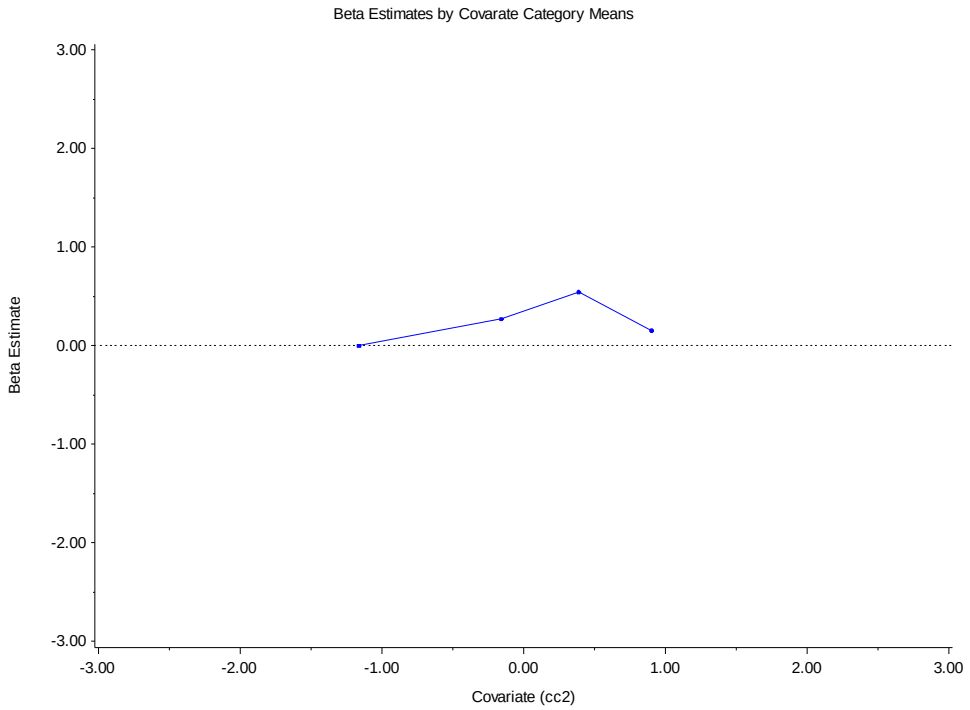

Figure 4.5: Categorized Quantile Estimates of Beta for a logarithmic covariate. SimExp44_QC02

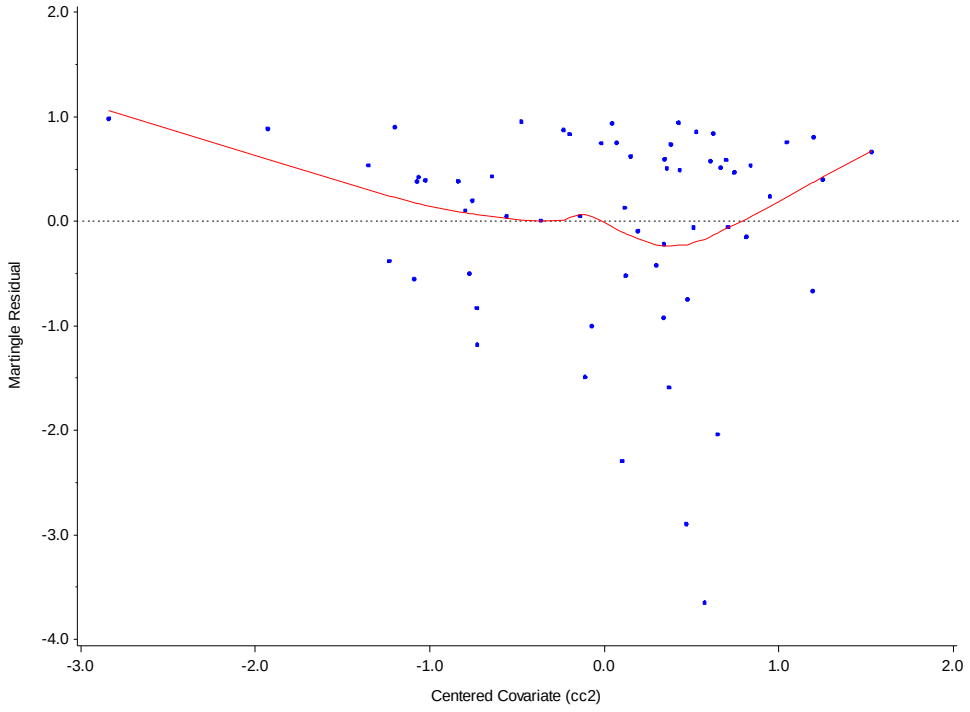

Figure 4.6: Martingale residuals and loess regression line for a model containing a logarithmic covariate (Dataset 44).

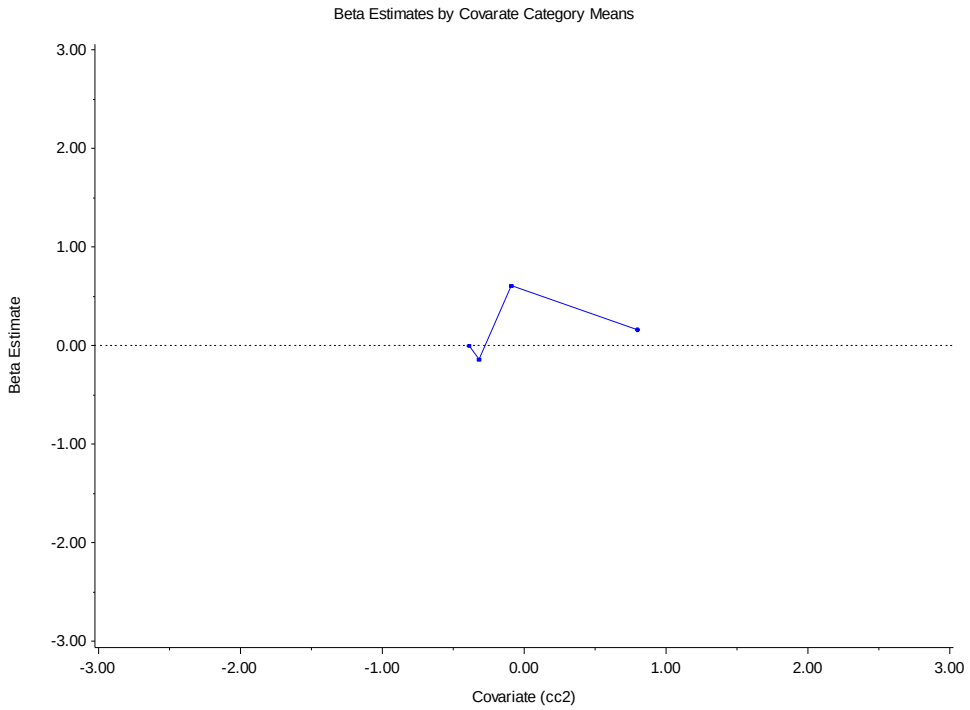

Figure 4.7: Categorized Quantile Estimates of Beta for a z*log(z) covariate. SimExp43_QC02

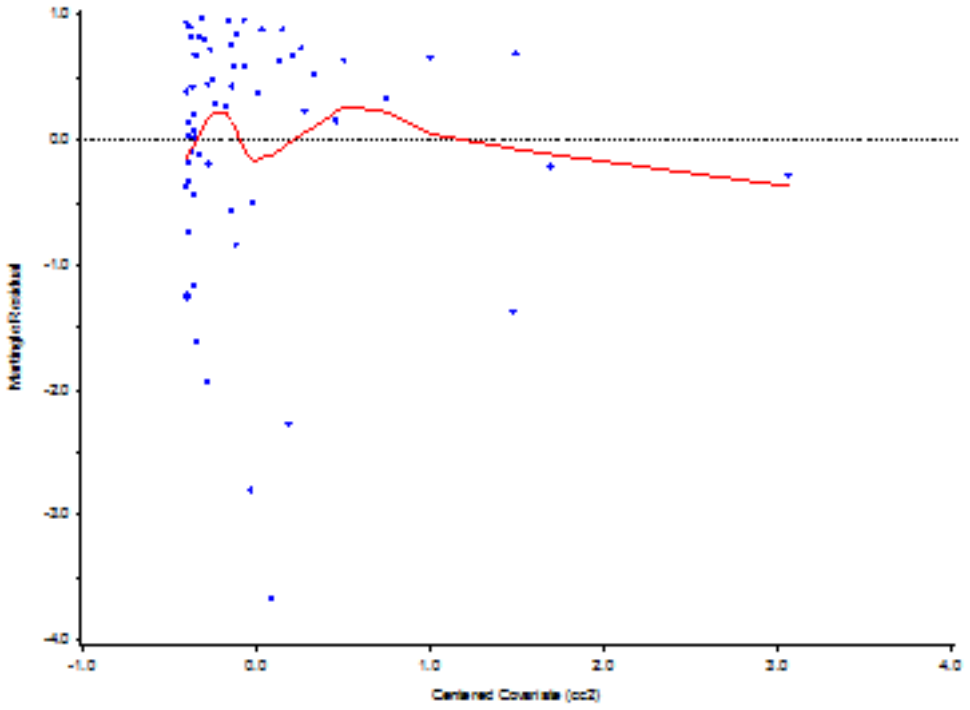

Figure 4.8: Martingale residuals and loess regression line for a model containing a z*log(z) covariate. SimExp43_MGcc2

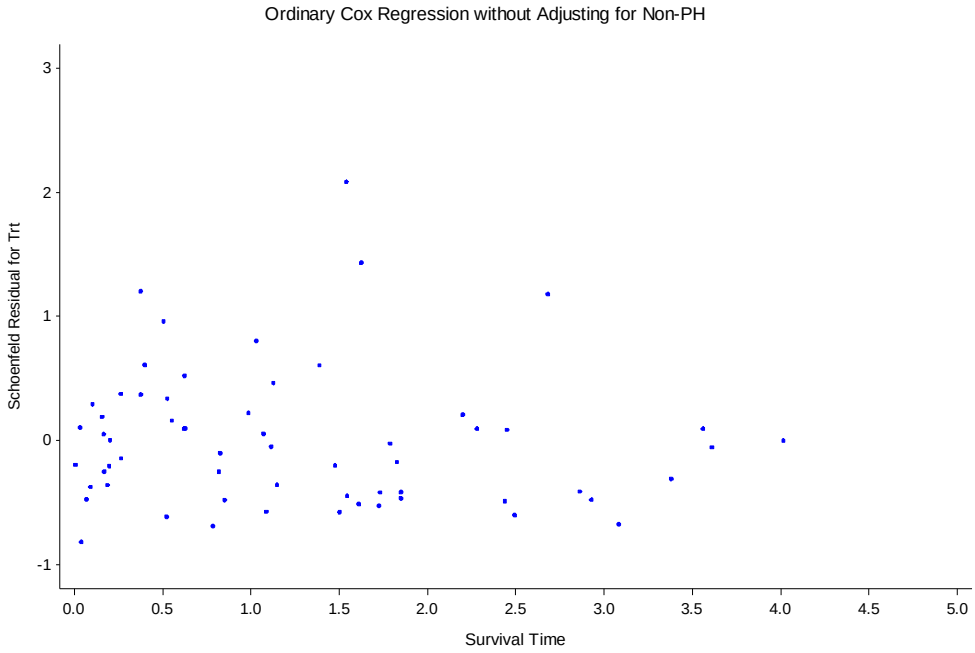

Figure 4.9: Graphical Check for Non-proportional Hazards using Schoenfeld's Residuals for a model containing a z*log(z) covariate using the log of the negative log of survival (See Wilson 2010 for more details). SimExp43_GCPH05A

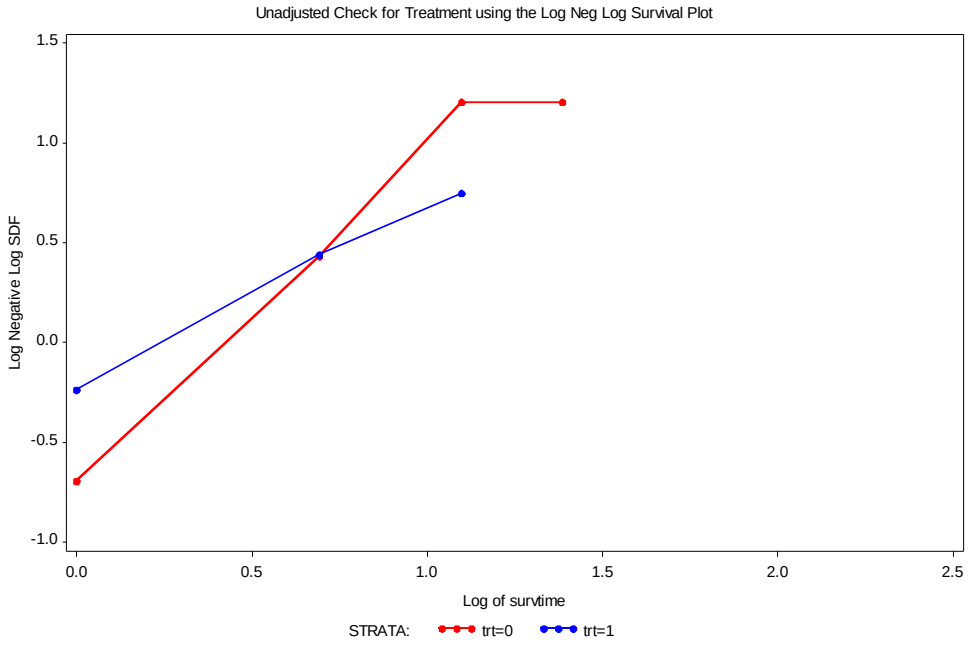

Figure 4.10: Another Graphical Check for Non-proportional Hazards for a model containing a z*log(z) covariate using the log of the negative log of survival.) SimExp43_GCPH01 Also see the std score process graph; p = 0.4760.

10.5 Figures for Section 5 (24 Figures)

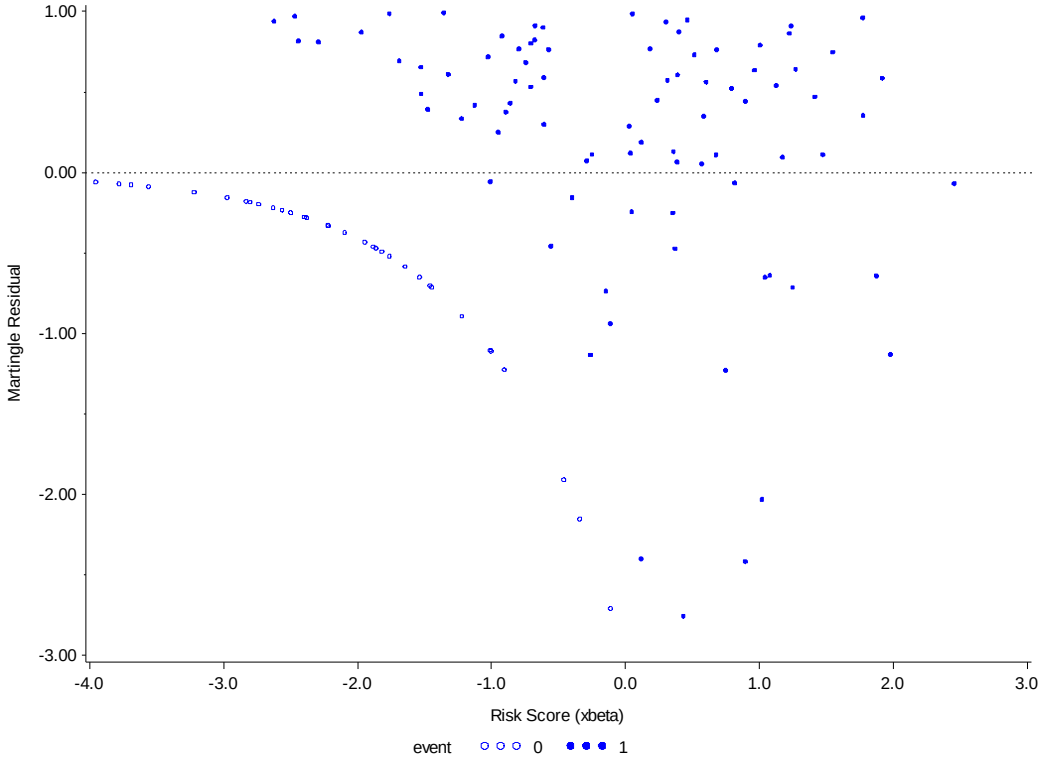

Figure 5.1: Martingale Residuals Plotted against Risk Score for data with no extreme values. Notice the negatively skewed distribution and the logarithmic curve bound by censored observations. Dataset 51.

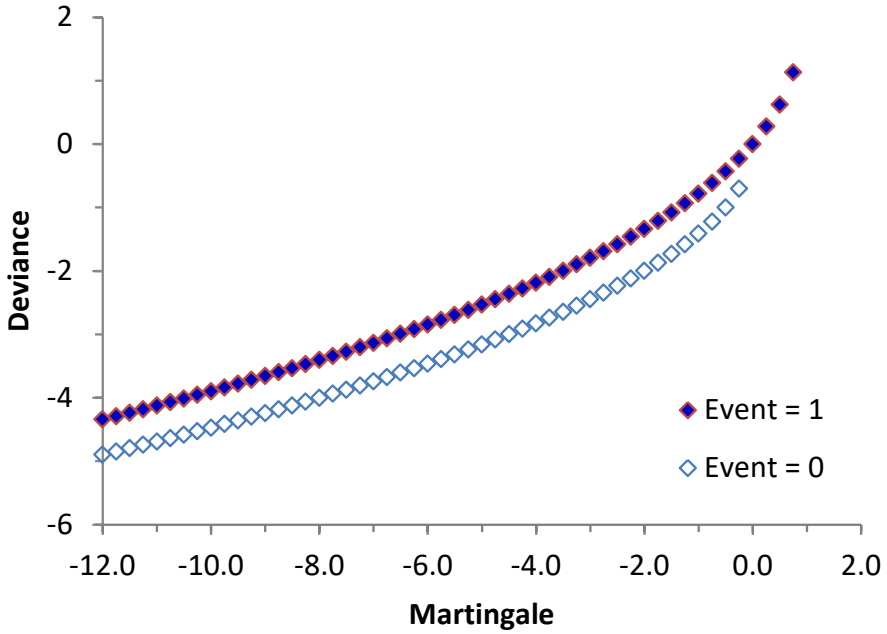

Figure 5.2: The Relationship between the Deviance and Martingale Residuals. The Deviance Residuals are a logarithmic transformation of the Martingales.

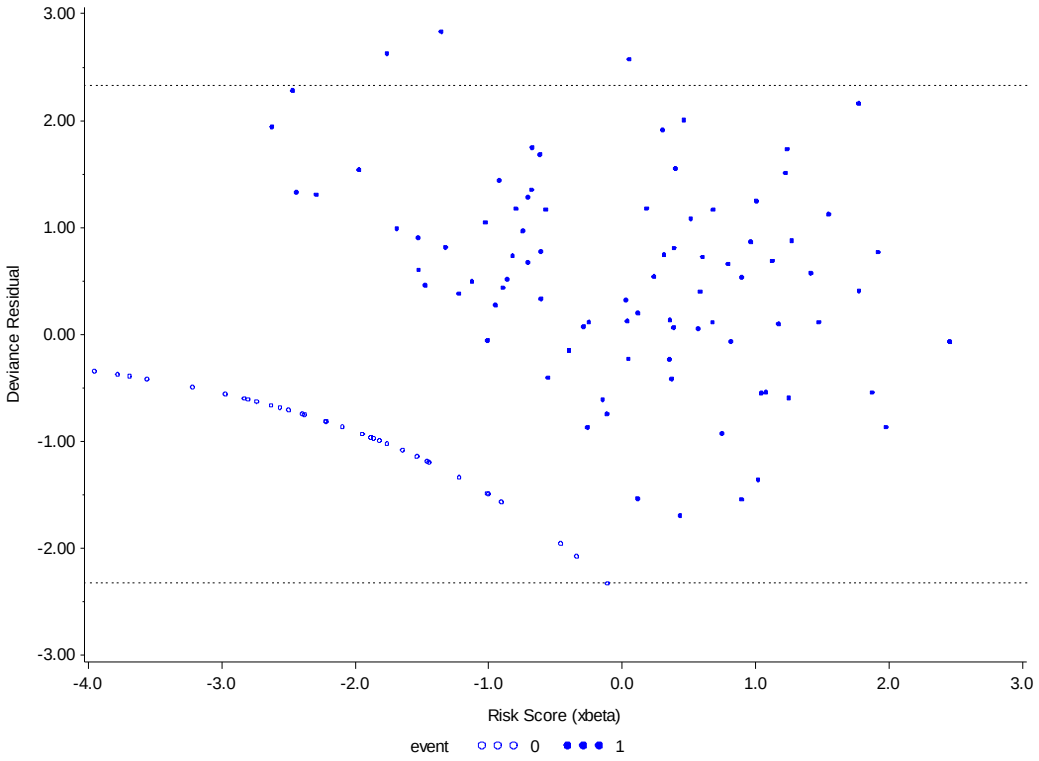

Figure 5.3: Deviance Residuals Plotted against Risk Score for data with no extreme values. Dataset 51

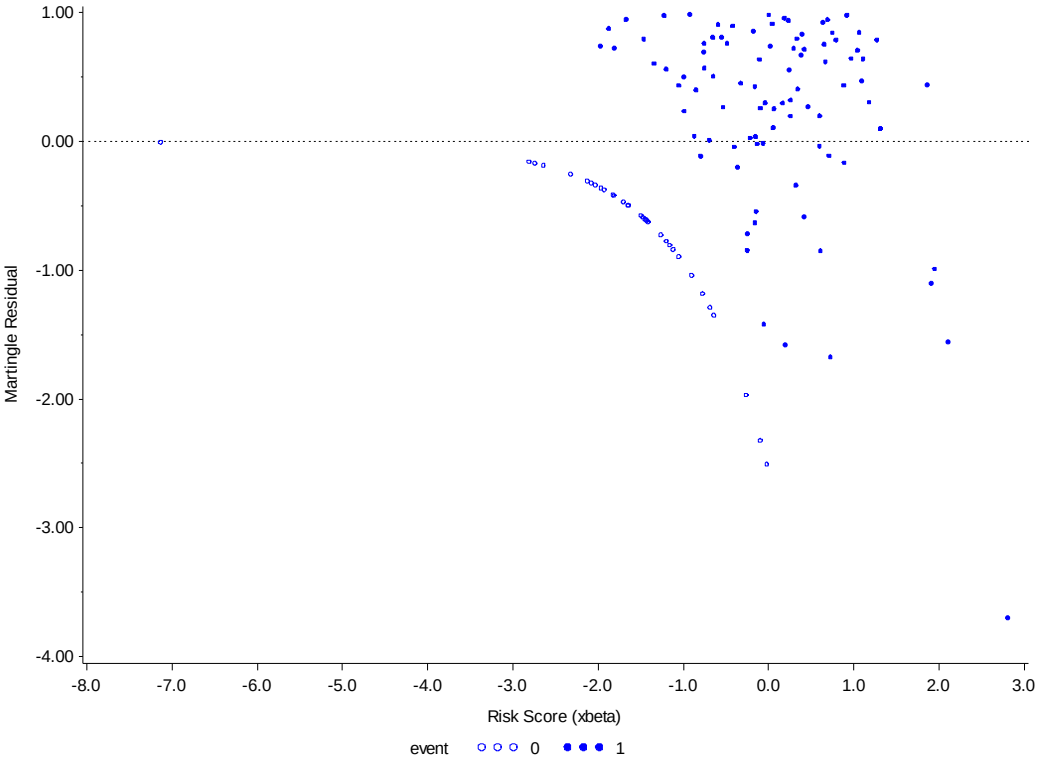

Figure 5.4: Martingale Residuals Plotted against Risk Score for a Model with four known extreme values. Dataset 52

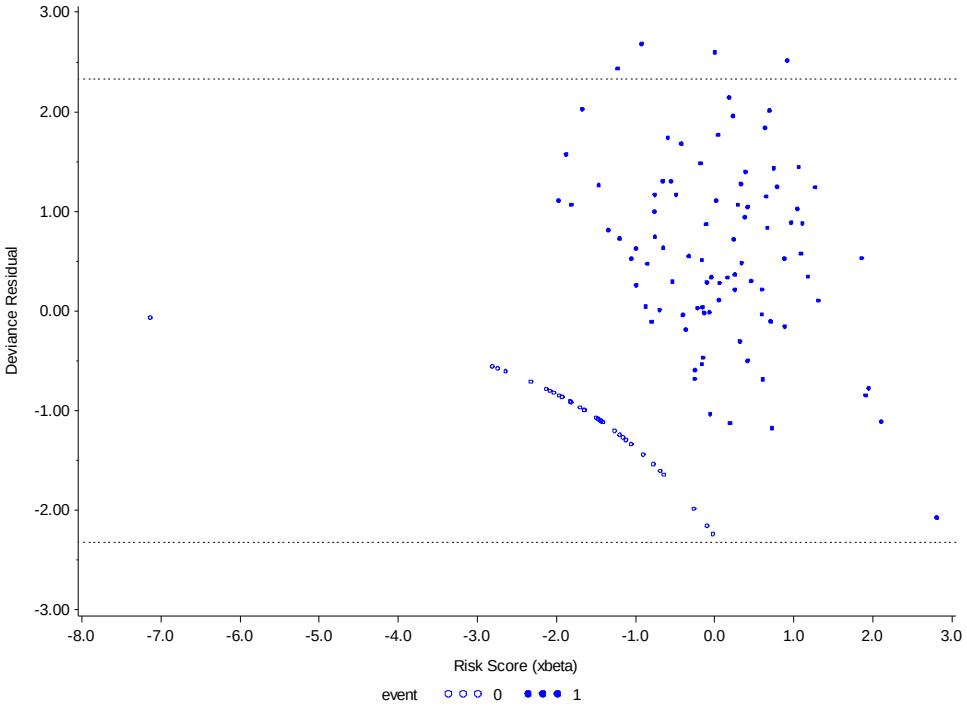

Figure 5.5: Deviance Residuals Plotted against Risk Score for a Model with four known extreme values. Dataset 52

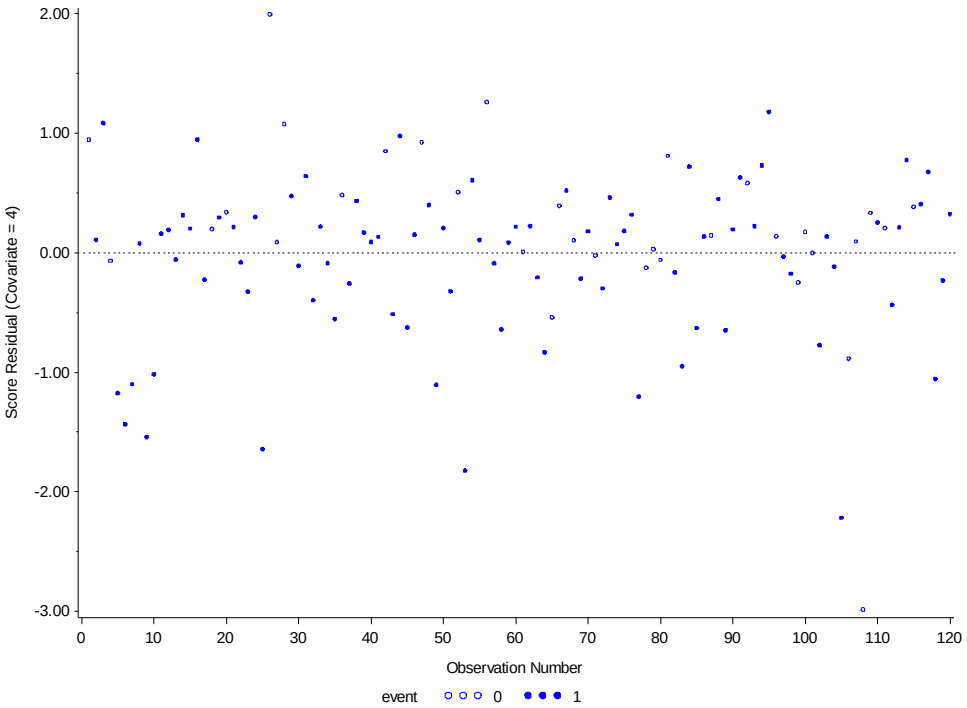

Figure 5.6: Score Residuals for the categorical covariate by Observation number for a Model without extreme values. Dataset 51

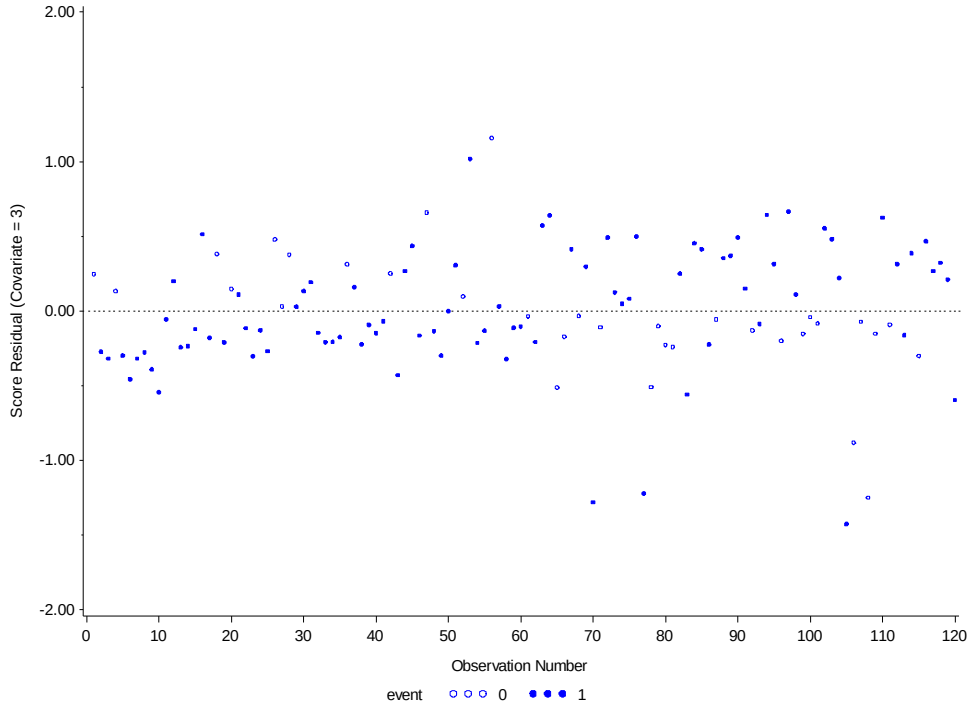

Figure 5.7: Score Residuals for the first continuous covariate by Observation number for data without extreme values. Dataset 51

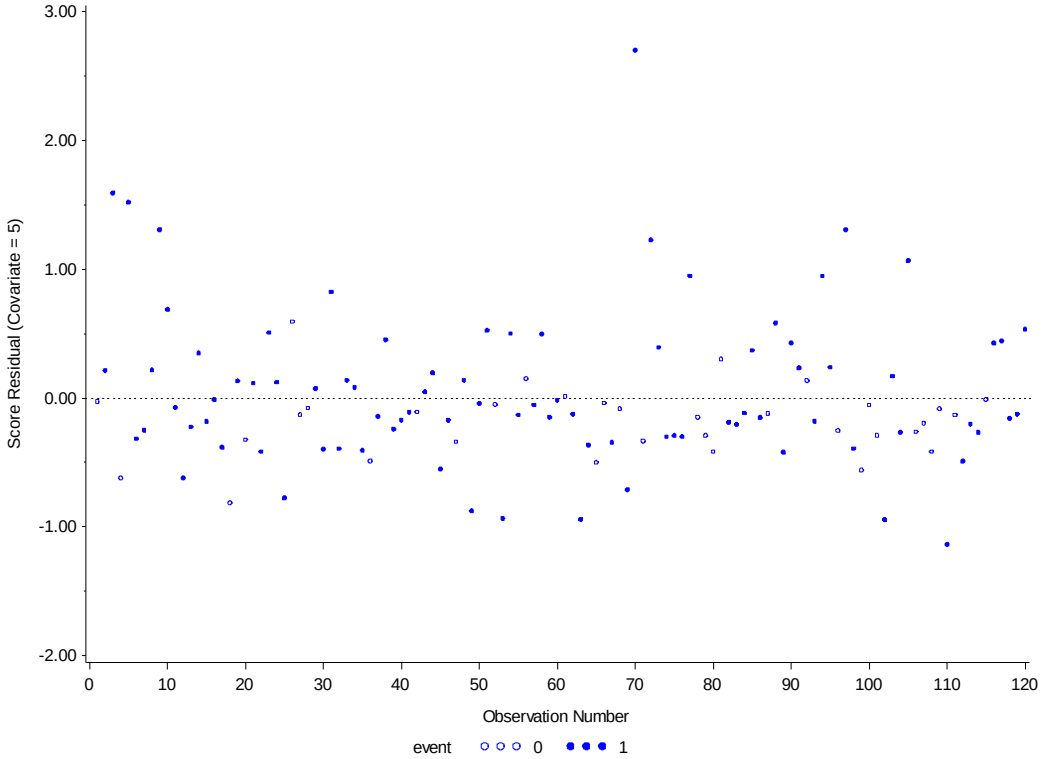

Figure 5.8: Score Residuals for the second continuous covariate by Observation number for data without extreme values. Dataset 51

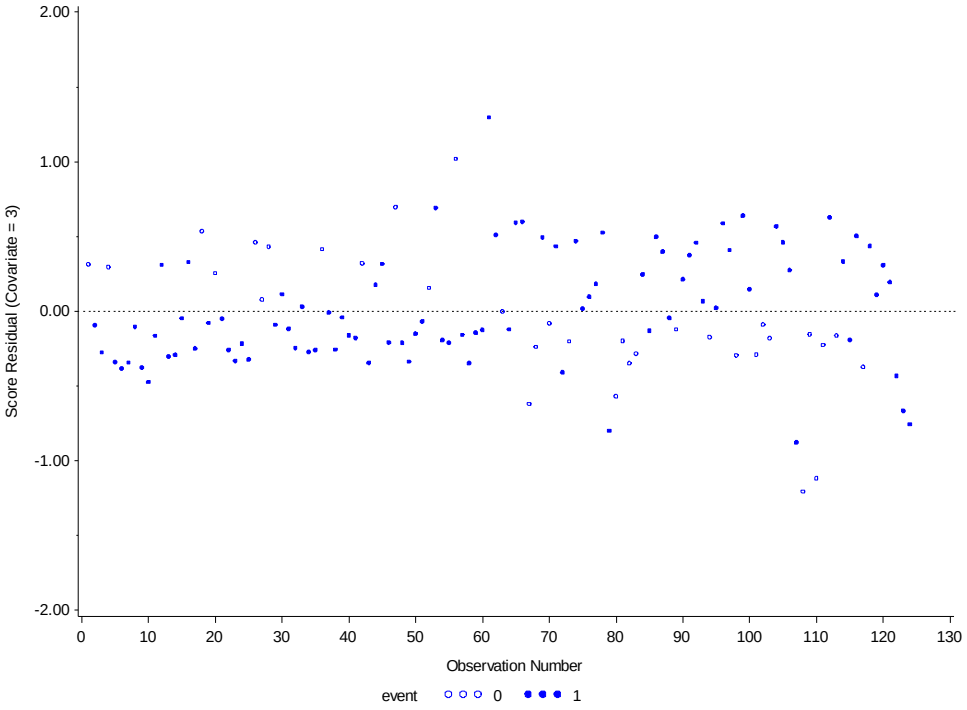

Figure 5.9: Score Residuals for the first continuous covariate by Observation number for data with known extreme values. Dataset 52

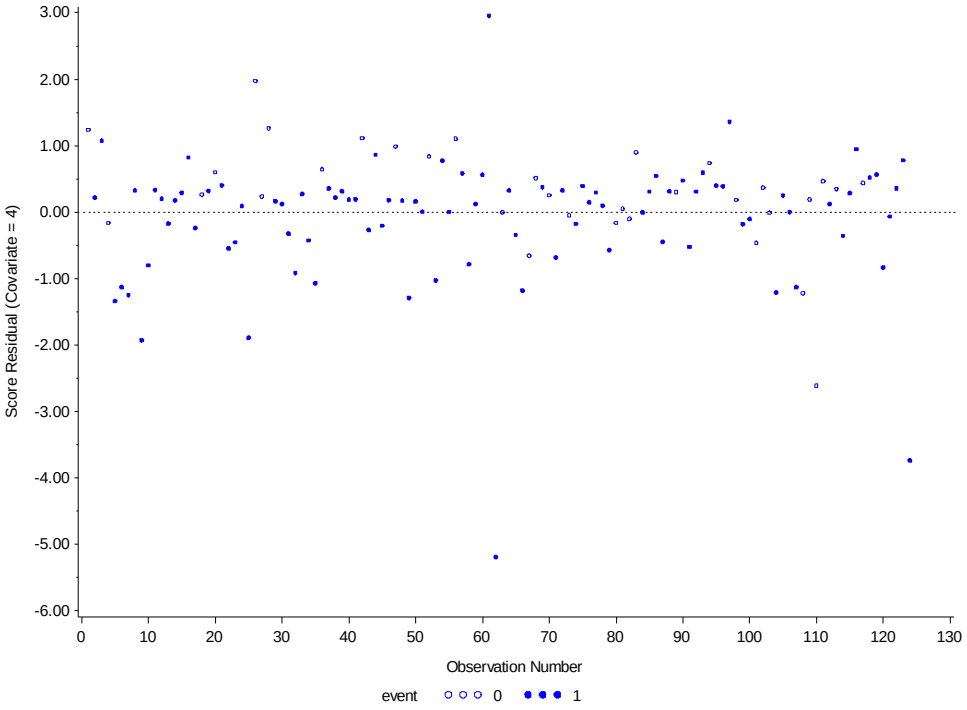

Figure 5.10: Score Residuals for the second continuous covariate by Observation number for data with known extreme values. Dataset 52

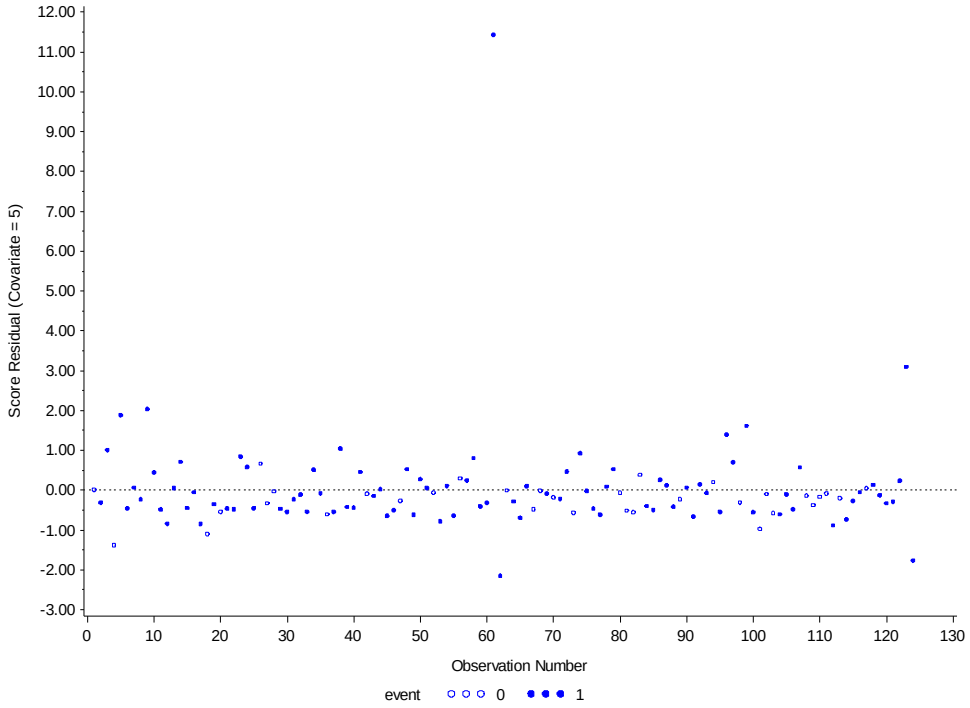

Figure 5.11: Score Residuals for the second continuous covariate by Observation number for data with known extreme values. Dataset 52

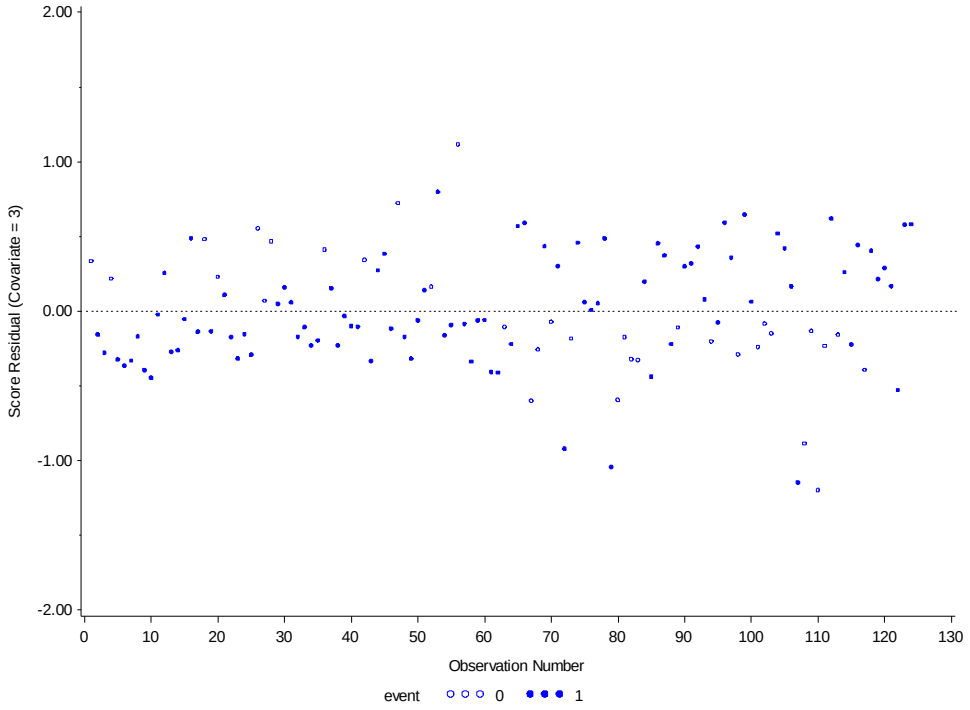

Figure 5.12: Score Residuals for the categorical covariate by Observation number for data with known extreme values. Dataset 53

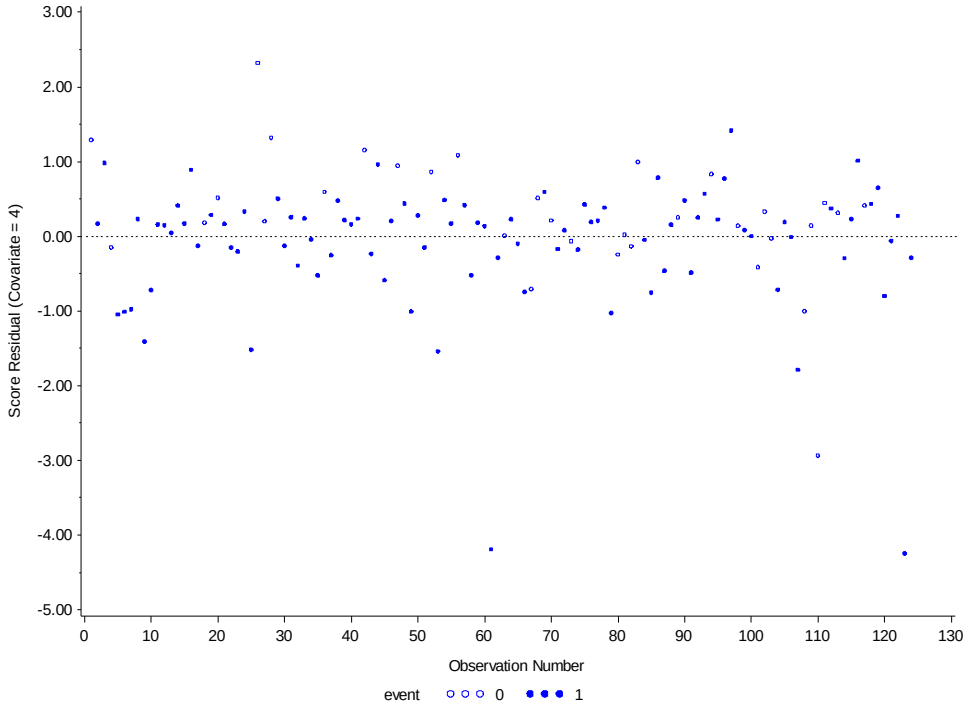

Figure 5.13: Score Residuals for the categorical covariate by Observation number for data with known extreme values. Dataset 53

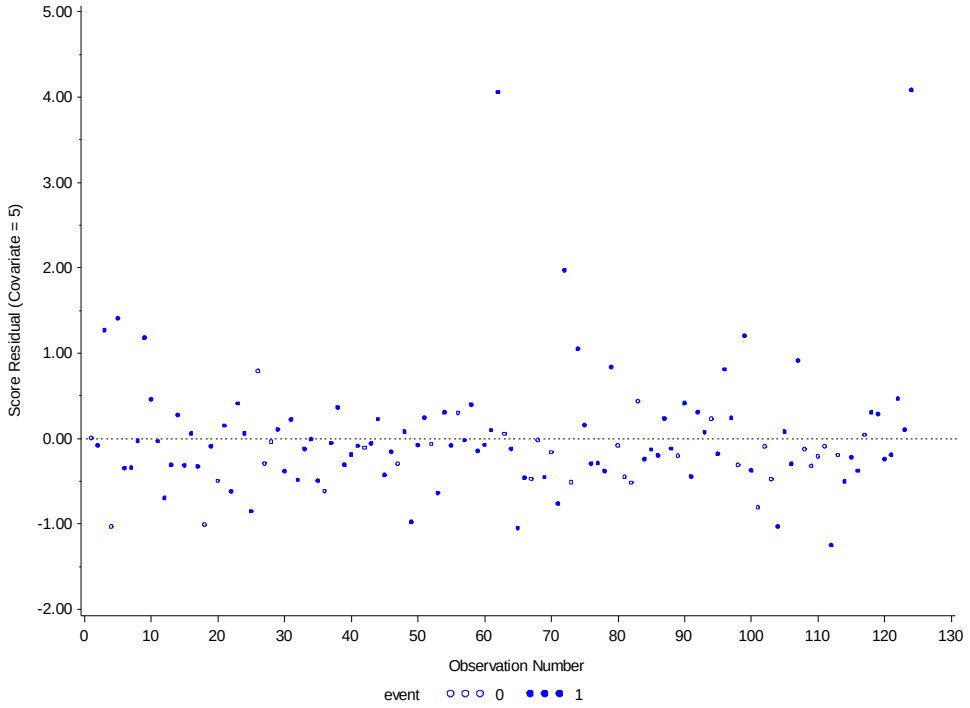

Figure 5.14: Score Residuals for the categorical covariate by Observation number for data with known extreme values. Dataset 53

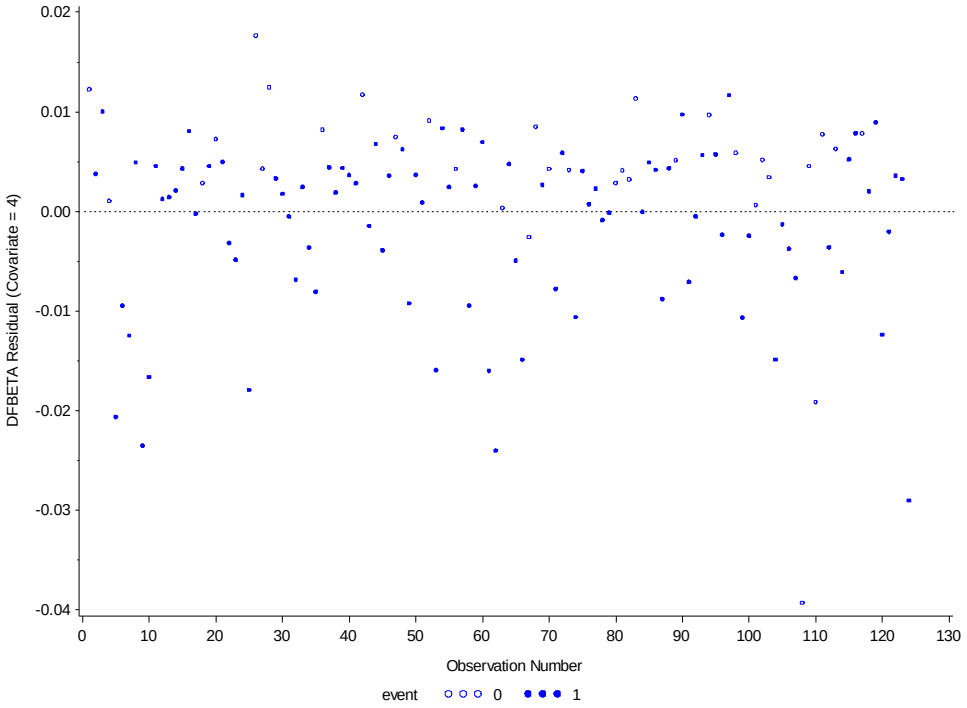

Figure 5.15: DFBETA Residuals for the first continuous covariate by Observation number for data with known strengthening extreme values. Dataset 52

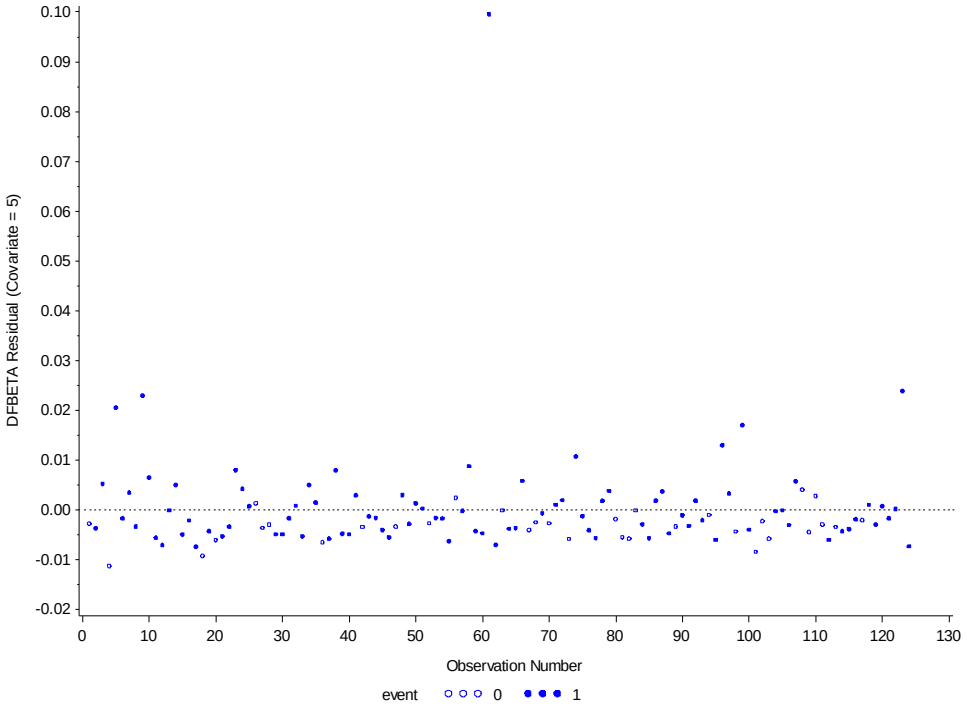

Figure 5.16: DFBETA Residuals for the second continuous covariate by Observation number for data with known strengthening extreme values. Dataset 52

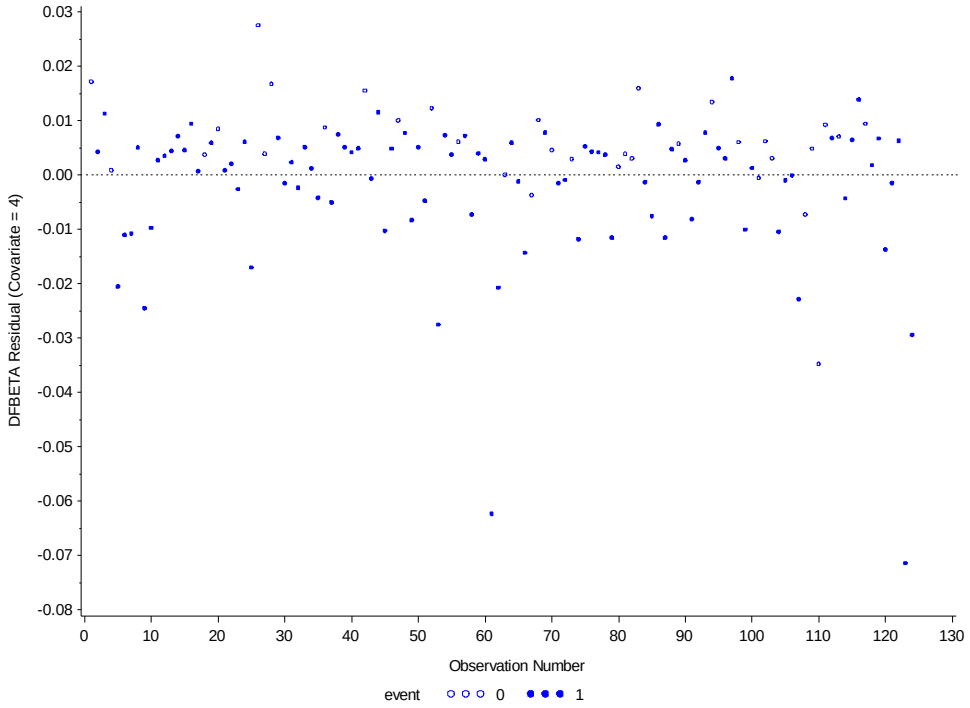

Figure 5.17: DFBETA Residuals for the first continuous covariate by Observation number for data with known weaken extreme values. Dataset 53

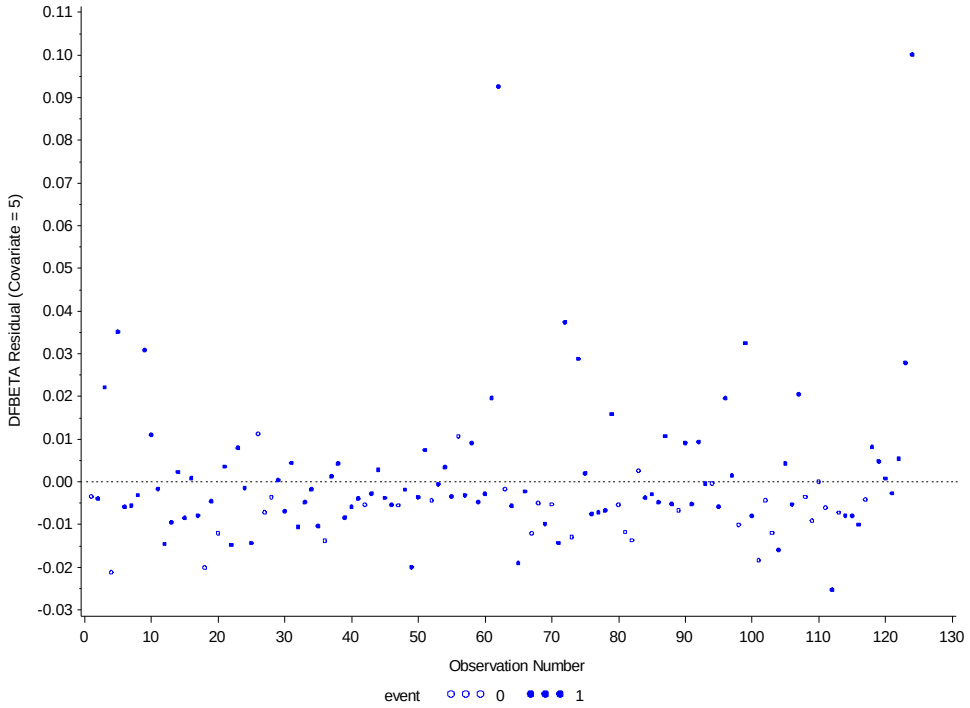

Figure 5.18: DFBETA Residuals for the second continuous covariate by Observation number for data with known weaken extreme values. Dataset 53

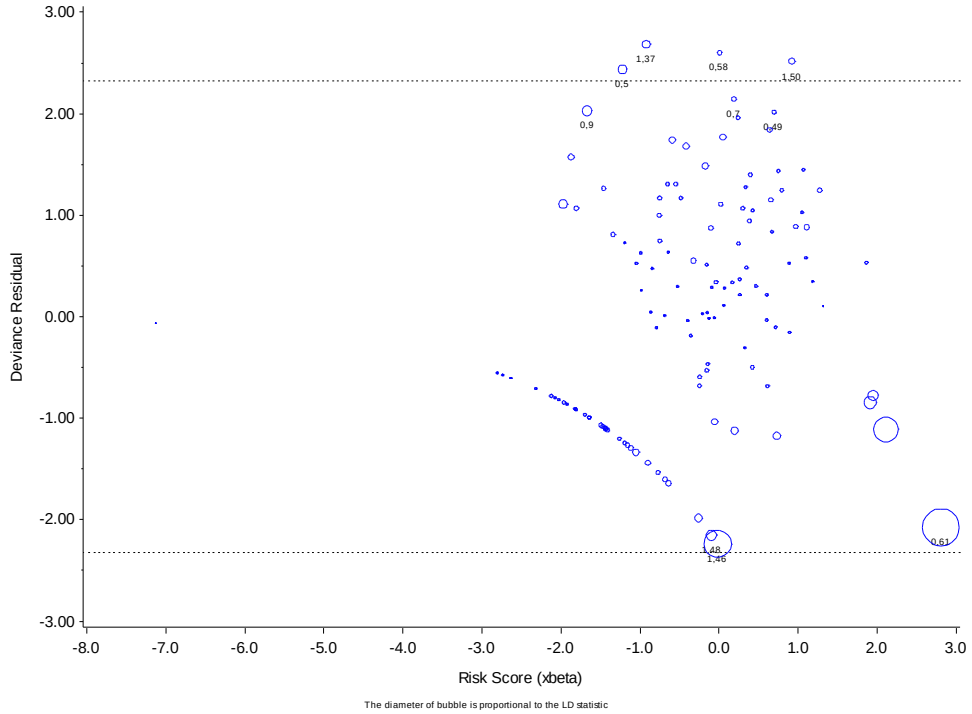

Figure 5.19: Gharibvand Plots for the strengthening Dataset 52. SimExp52_DV04

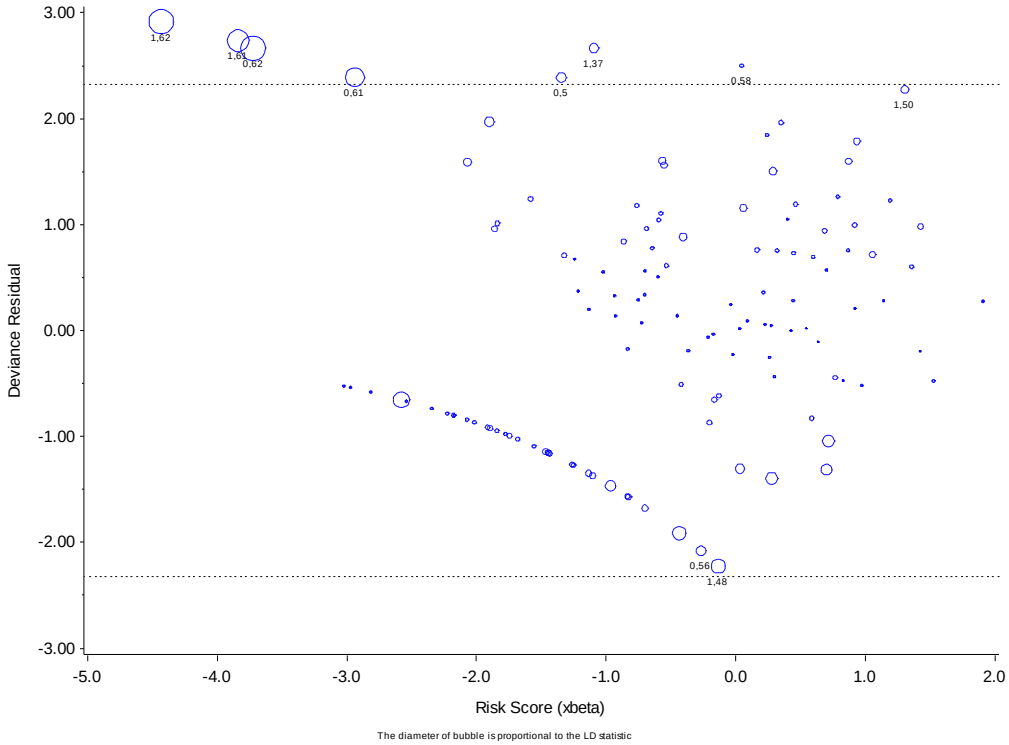

Figure 5.20: Gharibvand Plots for the weakening Dataset 53. SimExp53_DV04

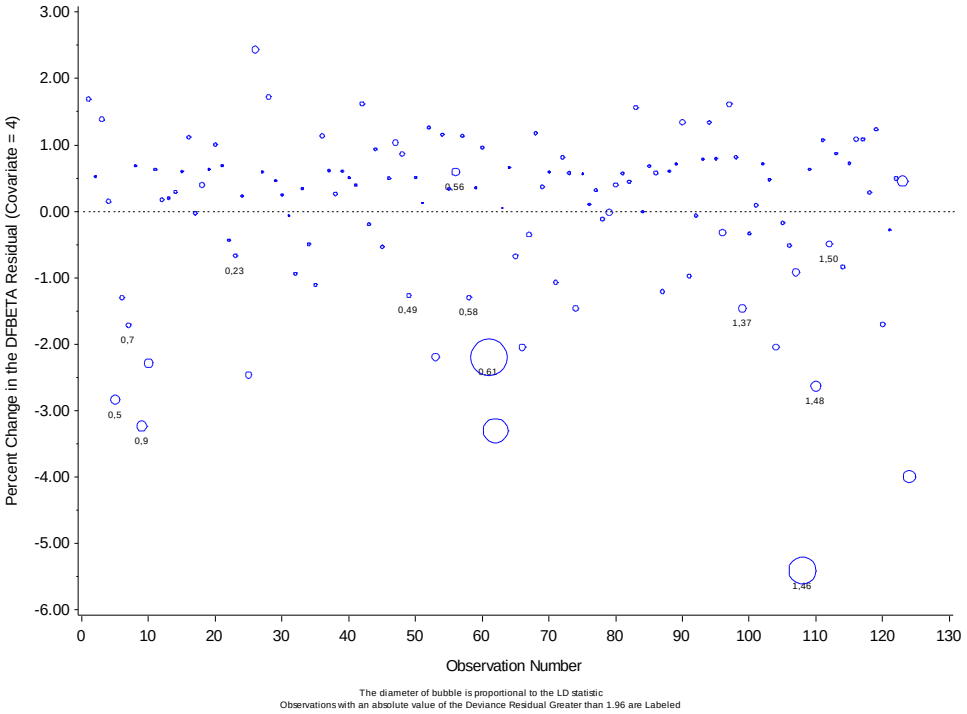

Figure 5.21: Combined Residual Plots for the strengthening dataset for the first continuous covariate by Observation number Dataset 52. SimExp52_DFB04

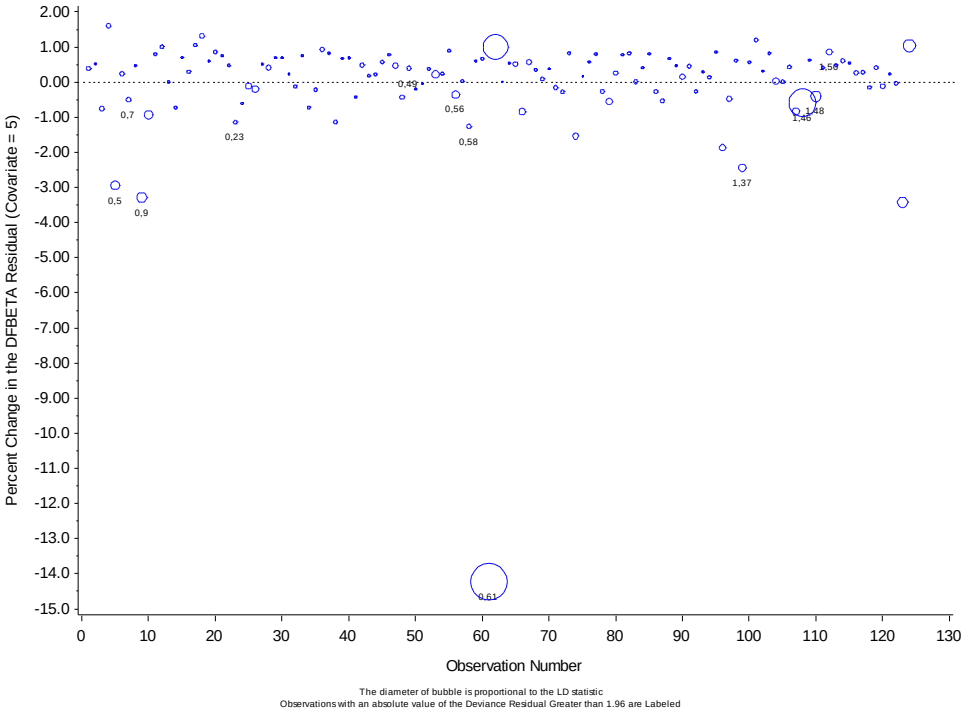

Figure 5.22: Combined Residual Plots for the strengthening dataset for the second continuous covariate by Observation number. Dataset 52. SimExp52_DFB05

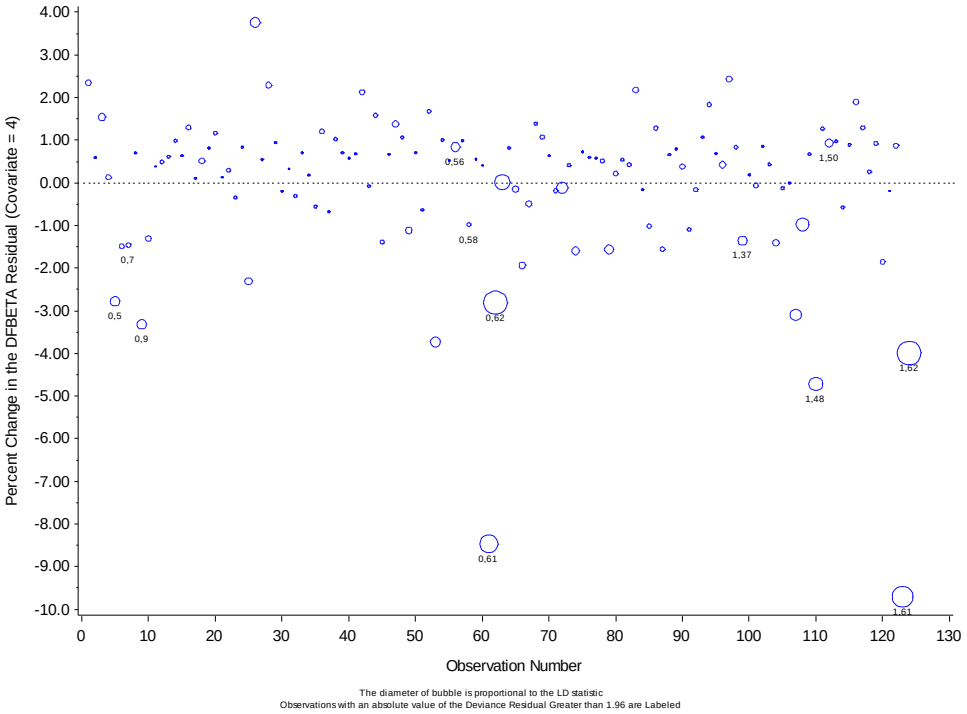

Figure 5.23: Combined Residual Plots for the weakening dataset for the first continuous covariate by Observation number. Dataset 53. SimExp53_DFB04

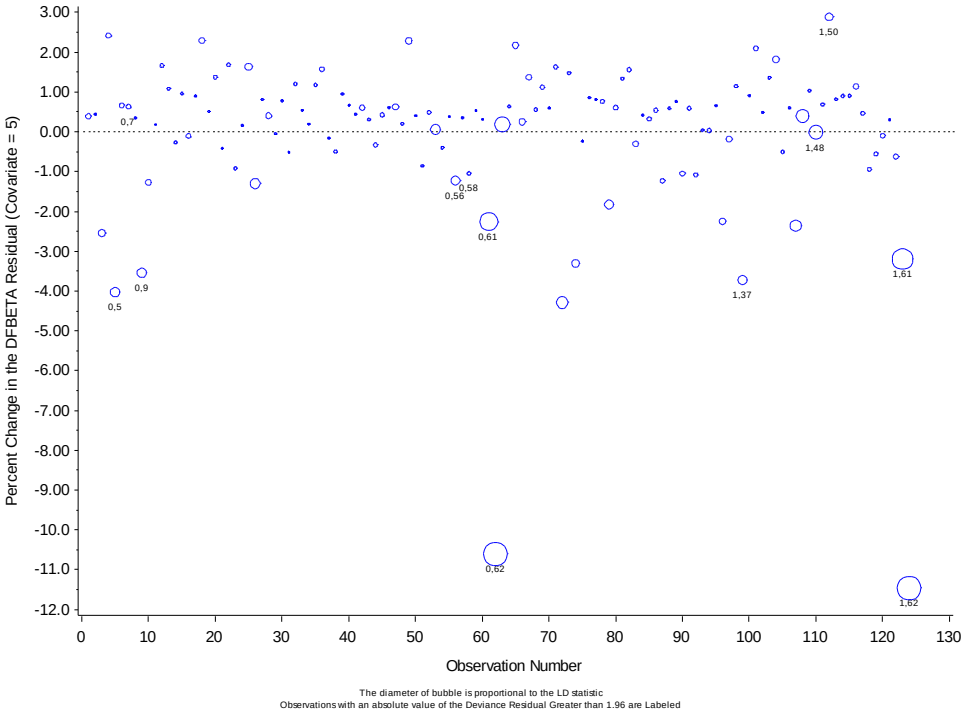

Figure 5.24: Combined Residual Plots for the weakening dataset for the second continuous covariate by Observation number. Dataset 53. SimExp53_DFB04