RESIDUAL PLOTS FOR DETECTING
COVARIATE IMPORTANCE AND NONLINEARITY
IN REGRESSION MODELS WITH CENSORED DATA

Michael Parzen¹ and David Harrington²

Abstract

This paper illustrates some diagnostic techniques for regression models where the
dependent variable is subject to censoring. We present a graphical, residual based
diagnostic procedure for selecting explanatory variables and detecting the effects of
influential or outlying observations on a particular covariate. We also illustrate graphi-
cal diagnostics for determining if power transformations of covariates in the model
are required. To accomplish this, we use the added variable plot (AVP) and con-
structed variable plot (CVP). The AVP is a graphical analogue to the score statistic
for adding a new variable to a proposed model. Using the AVP technique, one may
also investigate the appropriate transformation on a covariate for the model. The
constructed variable plot (CVP) is based on the added variable plot and can help
the data analyst determine an appropriate transformation for a covariate as well as
identify cases with substantial influence on the choice of a transformation. Following
the development of the method, we apply the diagnostic techniques to two biomedical
data sets.

¹Emory University, Atlanta, GA 30322

²Harvard School of Public Health, Boston, MA 02115 and Dana Farber Cancer Insti-
tute, Boston, MA 02115
1 Introduction

Let $T$ be a failure time random variable, $x$ a $p \times 1$ vector of covariates and $\epsilon \sim F$, where $F$ is some known distribution. We are interested in diagnostics for linear failure time (LFT) models of the form

$$g(T) = x'\beta + \sigma \epsilon,$$  \hspace{1cm} (1)

where $g$ is a known monotonic function and $\sigma$ a scale parameter. In general, we only observe $T^* = \min(T, C)$, where $C$ is a censoring time random variable. If $g(t) = \log(t)$ this would be the accelerated failure time model. This approach of modeling failure time as a linear function of covariates is quite appealing to medical investigators due to its ease of interpretation. Unfortunately, LFT models are not very robust to model misspecification, making diagnostics an important issue. Although estimation and inference for linear failure time (LFT) models is simple and efficient, little has been done in the area of model diagnostics.

The detection of influential observations, that is, observations whose deletion result in substantial changes in parameter estimates or functions of the parameter estimates, is of great importance when models are fit to censored data. This problem has been considered for linear failure time models by (Hall, Rogers and Pregibon, 1982; Weissfeld and Schneider, 1990; Escobar, 1992). These authors have concentrated on methods such as deletion of observations and perturbation of data points to detect the influence of individual observations.

In this paper we discuss a graphical, residual based diagnostic procedure for selecting explanatory variables and detecting the effects of influential or outlying observations on a particular covariate. We also illustrate graphical diagnostics for determin-
ing if power transformations of covariates in the model are required. To accomplish this, we use the added variable and constructed variable plots. The added variable plot (AVP) was first developed for normal linear regression and dates back to Cox (1958, pg. 58). This procedure has been extended to generalized linear models by Wang (1985) (also Davison and Tsai, 1993). The AVP is a graphical analogue to the score statistic for adding a new variable to a proposed model. Using the AVP technique, one may also investigate the appropriate transformation on a covariate for the model. The constructed variable plot (CVP) is based on the added variable plot and can help the data analyst determine an appropriate transformation for a covariate as well as identify cases with substantial influence on the choice of a transformation. The CVP under normal linear models and generalized linear models has been considered by Cook and Weisberg (1982) and Wang (1987).

In the following sections we develop the added and constructed variable plots for censored regression models of the form (1). Following the development of the method, we apply the diagnostic techniques to two data sets. The first example involves a Mayo Clinic data set on primary biliary cirrhosis of the liver, analyzed using a Weibull regression model. The second example concerns an international non-Hodgkin’s lymphoma data set analyzed using the semi-parametric proportional odds regression model. In the proportional odds model, the link function \( g(t) \) is unspecified. Using techniques developed in an earlier paper (Parzen and Harrington, 1992), we show how the CVP can be used on this semi-parametric model.
2 Added Variable Plots for the LFT Model

Consider survival studies in which we have $n$ observations of the form $(t_i^0, \delta_i, x_i)$. Here $\delta_i$ is an indicator variable ($\delta_i=1$ if $t_i^0$ is a true failure and 0 otherwise), $x_i$ is a $p \times 1$ vector of covariates and $t_i^0$ is the failure time ($\delta_i = 1$) or the censoring time ($\delta = 0$). We use the variable $t_i$ to indicate the failure time of the $i$th individual which, because of censoring, may or may not be observed. Let $X = [x_1, \ldots, x_n]'$ be the $n \times p$ design matrix.

Let $g$ be a known monotonic function and suppose that $g(t_i)$ is linearly related to $x_i$. That is, there exists an unknown $p \times 1$ vector $\beta$ such that

$$g(t_i) = \beta'x_i + \sigma \epsilon_i, \quad (2)$$

where $\epsilon_i, i = 1, \ldots, n$, are independent and identically distributed random variables whose common distribution function $F$ is completely specified and $\sigma$ is a scale parameter.

We assume that

$$w_i = \frac{g(t_i) - \beta'x_i}{\sigma} \sim F.$$

Without loss of generality we assume that $\sigma$ is known. Let $z_i$ be a new covariate for the $i$th individual and $z$ the associated $n \times 1$ vector of observations. Our interest is in possibly adding the new covariate $z$ to an augmented model so that

$$w_i = \frac{g(t_i) - \beta'x_i - z_i \gamma}{\sigma} \sim F.$$

We start by examining the score test for the hypothesis that $\gamma = 0$. Under the assumption of an independent censoring mechanism, the likelihood may be written

$$L(\beta, \gamma) = \prod_{i=1}^{n} (\sigma^{-1} f(w_i))^{\delta_i} S(w_i)^{1-\delta_i},$$
where $S(w_i) = \int_{w_i}^{\infty} f(u) du$ and $w_i = (g(t_i^o) - \beta'x_i - z_i\gamma)/\sigma$.

Using notation from Kalbfleisch and Prentice (1980) the score statistic may be written

$$U_{\beta_j}(\beta, \gamma) = \frac{\partial \log L}{\partial \beta_j} = \sigma^{-1} \sum_{i=1}^{n} x_{ij}a_i \quad j = 1, \ldots, p,$$

and

$$U_{\gamma}(\beta, \gamma) = \frac{\partial \log L}{\partial \gamma} = \sigma^{-1} \sum_{i=1}^{n} z_i a_i,$$

where

$$a_i = -\delta_i \frac{\partial \log f(w_i)}{\partial w_i} + (1 - \delta_i) \lambda(w_i), \quad \text{with } \lambda(w_i) = \frac{f(w_i)}{S(w_i)}.$$

The observed information matrix $I(\beta, \gamma)$ has entries

$$-\frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k} = \sigma^{-2} \sum x_{ij}x_{ik}R_i,$$

$$-\frac{\partial^2 \log L}{\partial \beta_j \partial \gamma} = \sigma^{-2} \sum x_{ij}z_i R_i,$$

and

$$-\frac{\partial^2 \log L}{\partial \gamma^2} = \sigma^{-2} \sum z_i^2 R_i,$$

where

$$R_i = -\delta_i \frac{\partial^2 \log f(w_i)}{\partial w_i^2} + (1 - \delta_i) \left[ \lambda(w_i) \frac{\partial \log f(w_i)}{\partial w_i} + \lambda^2(w_i) \right].$$

Define $V = \text{diag}(R_i/\sigma^2)$ and let $\mathbf{s}$ be a vector of entries $a_i/\sigma$. Then (4) and $I(\beta, \gamma)$ can be written in simpler forms:

$$\frac{\partial \log L}{\partial \gamma} = \mathbf{s}'\mathbf{z}$$
and

\[ I(\beta, \gamma) = \begin{pmatrix} X'VX & X'Vz \\ z'VX & z'Vz \end{pmatrix}. \]

The maximum likelihood estimate (MLE) \( \hat{\beta} \) of \( \beta \) under \( H_0 : \gamma = 0 \) is the solution to

\[ U_{\beta_j}(\beta, 0) = 0, \quad j = 1, \ldots, p. \]

In general, the equations \( U_{\beta_j}(\beta, 0) = 0 \) are non-linear and must be solved by numerical iteration. If the Newton-Raphson method is used, then the \( m \)th approximation is given by

\[ \beta^{(m)} = \beta^{(m-1)} + \left( \frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k} \right)^{-1} U_{\beta_j}(\beta^{(m-1)}, 0). \]

Note that

\[ \left( \frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k} \right) = X'VX, \]

and

\[ U_{\beta} = X's. \]

We may re-write the updating equations as

\[ \beta^{(m)} = \beta^{(m-1)} + (X'VX)^{-1}X's, \]

or equivalently

\[ (X'VX)\beta^{(m)} = (X'VX)\beta^{(m-1)} + X's, \]

where both \( V \) and \( s \) are evaluated at \( (\beta^{(m-1)}, \gamma = 0) \).

Define \( y_* = X\beta^{(m-1)} + V^{-1}s \). Then the last equation may be written

\[ (X'VX)\beta^{(m)} = (X'V)y_* . \]
This equation has the same form as the normal equations for a linear model obtained by weighted least squares, except that (5) has to be solved iteratively because $y_\star$ and $V$ depend on $\beta$. This solution procedure is called iterative weighted least squares (IWLS). Note that the presence of censoring in the observations is incorporated into the estimation process through the use of $V$ and the “working vector” $y_\star$.

Let $\hat{\beta}$ denote the mle of $\beta$ under the hypothesis that $\gamma = 0$, obtained using IWLS. Henceforth let both $V$ and $s$ be evaluated at $(\hat{\beta}, \gamma = 0)$. For testing $H_0 : \gamma = 0$ the score statistic can be expressed as

$$S = \left[U_\gamma(\hat{\beta}, 0)\right]^2 I_{\gamma\gamma}(\hat{\beta}, 0),$$

where $I_{\gamma\gamma}$ is the last diagonal element of $I^{-1}(\beta, \gamma)$. Under $H_0 : \gamma = 0$ the asymptotic distribution of $S$ is a $\chi^2_1$ distribution.

After some algebraic manipulations the score statistic becomes

$$S = \frac{(s'z)^2}{z'V^{1/2}M V^{1/2}z},$$

where $M = I - V^{1/2}X(X'VX)^{-1}X'V^{1/2}$.

As we did in demonstrating the IWLS procedure, define $y_\star = X\hat{\beta} + V^{-1}s$. Then

$$V^{-1/2}s = MV^{-1/2}s \quad \text{and} \quad V^{-1/2}s = V^{1/2}(y_\star - X\hat{\beta})$$

which implies that

$$V^{-1/2}s = MV^{1/2}y_\star.$$

Using these identities, an equivalent expression of the score statistic is

$$S = \frac{(y'_\star V^{1/2}MV^{1/2}z)^2}{z'V^{1/2}M V^{1/2}z}.$$
Written in this form it is possible to recognize $S$ (see Seber, 1977, pg. 69 and Pregibon, 1982) as the difference in residual sum of squares between models

$$y_* = X\lambda + z\tau + \epsilon, \quad \epsilon \sim N(0, V^{-1}) \tag{8}$$

and

$$y_* = X\lambda + \epsilon, \quad \epsilon \sim N(0, V^{-1}). \tag{9}$$

That is, the score statistic is the additional sum of squares due to the hypothesis $H_0 : \tau = 0$ in the weighted linear regression of $y_*$ on $X$ and $z$.

A simple transformation of model (8) produces a model with constant variance:

$$V^{1/2}y_* = V^{1/2}X\lambda + V^{1/2}z\tau + V^{1/2}\epsilon. \tag{10}$$

It may be shown (Atkinson, 1985, pg. 68) that the least squares estimate of $\tau$ in model (10) is given by

$$\hat{\tau} = \frac{z'V^{1/2}MV^{1/2}y_*}{z'V^{1/2}MV^{1/2}z}.$$

We also obtain

$$\text{var}(\hat{\tau}) = \frac{1}{z'V^{1/2}MV^{1/2}z}.$$

Hence the test statistic for $H_0 : \tau = 0$ is given by

$$T = \frac{\hat{\tau}}{\sqrt{\text{var}(\hat{\tau})}} = \frac{z'V^{1/2}MV^{1/2}y_*}{\sqrt{z'V^{1/2}MV^{1/2}z}},$$

so that

$$S = T^2.$$
where $\mathbf{r}$ is the residual vector for the weighted regression of $\mathbf{y}_*$ on $\mathbf{X}$ and $\mathbf{z}^*$ corresponds to the residual vector for the weighted regression of $\mathbf{y}_*$ on $\mathbf{z}$.

Using this notation we find that

$$S = \frac{(\mathbf{z}^*)' \mathbf{r}^2}{(\mathbf{z}^*)' \mathbf{z}^*},$$

and

$$\hat{\tau} = \frac{(\mathbf{z}^*)' \mathbf{r}}{(\mathbf{z}^*)' \mathbf{z}^*}.$$

Thus $\hat{\tau}$ is simply the slope of the univariate regression of $\mathbf{r}$ on $\mathbf{z}^*$. Further, since $S = T^2$, we find that a statistically large value of $S$ occurs when there is evidence for correlation between $\mathbf{r}$ and $\mathbf{z}^*$.

A large value of $S$ does not necessarily indicate a systematic departure supported by all the data, but may be due to a few observations alone. Such observations can be detected by the added variable plot (AVP). The geometry of least squares implies that a large value of $S$ is due to substantial correlation between $\mathbf{y}_*$ and $\mathbf{z}$, after adjusting for the weighted regression on $\mathbf{X}$. If we examine the one-step maximum likelihood estimate of $\gamma$ in the full augmented model using IWLS and the initial value $(\mathbf{\beta}, \gamma) = (\hat{\mathbf{\beta}}, 0)$ we obtain

$$\hat{\gamma} = (I^{\gamma}(\hat{\mathbf{\beta}}, 0))^{-1} U_\gamma(\hat{\mathbf{\beta}}, 0) = \frac{s' \mathbf{z}}{z' V^{1/2} \mathbf{M} V^{1/2} \mathbf{z}} = \frac{(\mathbf{z}^*)' \mathbf{r}}{(\mathbf{z}^*)' \mathbf{z}^*} = \hat{\tau}.$$

Thus the one-step maximum likelihood estimate of $\gamma$ equals the value obtained from the regression through the origin of $\mathbf{r}$ on $\mathbf{z}^*$. Hence a plot of $\mathbf{r}$ against $\mathbf{z}^*$
provides a useful visual assessment of the evidence for regression and will indicate which observations are contributing to the relationship and which are deviating from it. We call this plot the *added variable plot* and for convenience, the vectors \( r \) and \( z^* \) are called *standardized residuals* and *z-residuals*, respectively. We note that if the value of \( \sigma \) is unknown, we may use a consistent estimate of \( \sigma \) under \( H_0 \) without disturbing any of the diagnostic results.

### 3 Constructed Variable Plots

One could search for an adequate transformation of a covariate by fitting many different models and then comparing them using a likelihood based criterion. Instead, we extend the procedure of Box and Tidwell (1962) (see also Cook and Weisberg (1982), section 2.4), which aids in the selection of transformations for explanatory variables in normal linear regression, to the accelerated failure time model.

We are interested in whether the covariate \( x_{ip} \) requires a transformation in the model. For simplicity we restrict the transformation required for \( x_{ip} \) to the family of power transformations proposed by (Box and Cox, 1964),

\[
x^{(\lambda)}_{ip} = \begin{cases} 
\frac{x_{ip}^{\lambda} - \lambda}{\lambda} & \lambda \neq 0 \\
\log x_{ip} & \lambda = 0
\end{cases}
\]

This family contains the usual log, square root, and inverse transformations as special cases and is scaled to be continuous at \( \lambda = 0 \).

We assume that the residual \( w_i \) can be written

\[
w_i = (y_i - x^{(\lambda)}_{ip} \beta_p - \sum_{j=1}^{p-1} x_{ij} \beta_j)/\sigma.
\]  

(11)

To check the nonlinearity of \( x_{ip} \), we consider the hypothesis \( H_0 : \lambda = 1 \). Expanding
\( x_{ip}^{(\lambda)} \) in a Taylor series around \( \lambda = 1 \) we find that

\[
x_{ip}^{(\lambda)} \approx x_{ip} + (\lambda - 1)(x_{ip} \log x_{ip} - x_{ip} + 1).
\]

Thus we obtain the approximated model to (11)

\[
w_i = (y_i - \mathbf{x}'_i \boldsymbol{\beta} - \beta_p (\lambda - 1)(x_{ip} \log x_{ip} - x_{ip} + 1))/\sigma. \tag{12}
\]

We now examine an augmented model as before with \( z_i = x_{ip} \log x_{ip} - x_{ip} + 1 \). The variable \( z_i \) is called a constructed variable and the resulting diagnostic plot of the standardized residuals versus the \( z \)-residuals is termed a constructed variable plot (CVP). The transformation parameter \( \lambda \) is related to the slope of the added variable plot for the constructed variable \( z \). A linear trend in such a plot may be taken as an indication that \( \lambda \neq 1 \). As before, the CVP can also be used to identify outlying cases that may be distorting the evidence for a transformation.

An approximate and easily constructed estimate of \( \hat{\lambda} \) is

\[
\hat{\lambda} = 1 + \frac{\hat{\gamma}}{\hat{\beta}_p},
\]

where \( \hat{\beta}_p \) is the MLE from model (11) when \( \lambda = 1 \) and \( \hat{\gamma} \) is the slope estimate of the regression line in the CVP. This approximate estimate of \( \lambda \) in combination with the CVP will usually suffice for diagnosing the need to transform and if so, what transformation is required. Further iteration of the constructed variable process using \( \hat{\lambda} \) as a starting value may be used to find better estimates.

### 4 Evaluation of Two Survival Data Sets

In this section we apply the constructed variable plot to two data sets in an effort to determine if power transformations of covariates are needed. The first example
Involves a Mayo Clinic data set on primary biliary cirrhosis of the liver, analyzed using a Weibull regression model. The second example concerns a non-Hodgkin’s lymphoma (NHL) data set analyzed using the proportional odds regression model. In the proportional odds model, the link function \( g(t) \) is unspecified. Using techniques developed in an earlier paper (Parzen and Harrington, 1992), we show how the CVP can be used on this semi-parametric model.

### 4.1 Liver Data Set

The primary outcome in the Mayo Clinic data set was time to death from primary biliary cirrhosis (PBC) of the liver, and published analyses have examined the association of that outcome with treatment and with a number of patient and disease characteristics available at diagnosis. A detailed description and analysis of the data is contained in Fleming and Harrington (1991). The Mayo Clinic has established a database of 424 patients having primary biliary cirrhosis (PBC). These 424 form the complete collection of all PBC patients, referred to Mayo between January 1974 and May 1984, who met standard eligibility criterion for a randomized, double-blinded, placebo-controlled, clinical trial of the drug D-penicillamine (DPCA). For each patient enrolled on the study, clinical, biochemical, serologic and histologic parameters were collected. In their book, Fleming and Harrington (1991) develop a natural history model for PBC based on the Cox proportional hazards model. Starting from 14 covariates they use stepwise regression to fit a final model of 5 covariates; age (years), albumin (gm/dl), presence of edema (0=no,.5=present but no diuretic therapy,1=present despite diuretic therapy), protime (prothrombin time in seconds), and serum bilirubin (mg/dl). We will examine a Weibull regression model (which is both a proportional hazards model and an accelerated failure time model) using the final
5 covariate model of Fleming and Harrington. The Weibull model is of the form of model (1) with \( g(t) = \log(t) \) and with an error term \( \epsilon \) that has an extreme value distribution.

The fit of the five covariate Weibull regression model is presented as Model 1 in Table 1. The covariate bilirubin is the dominating prognostic factor in the model; we are thus interested in testing whether the linear form of bilirubin is adequate. After fitting Model 1, we display the CVP for bilirubin in Figure 1, which corresponds to a p-value < .001 of the score statistic. Two outliers are evident in the CVP, cases 87 and 103. These cases however have \( z \)-residual values close to zero and hence have negligible influence on the regression. The outlier case 144 corresponds to the largest bilirubin value measured (28 mg/dl). Since 75\% of the bilirubin values are below 3.4 mg/dl and only 2 values are above 25 mg/dl (cases 156 and 144), this case may need to be examined further. Cases 81, 253 and 293 also appear as outliers and may need to be examined. These cases were noted by Fleming and Harrington as questionable.

The approximate estimate of the power transformation is \( \hat{\lambda} = -0.517 \). For ease of interpretation, we choose \( \lambda = -0.5 \) (the inverse square-root transformation). If we perform an additional CVP on the transformed bilirubin, we obtain a p-value of .04 with \( \hat{\lambda} = .12 \). Thus it would appear that the optimal transform for bilirubin is \( \log(1/\sqrt{\text{bilirubin}}) \) or equivalently, \( \log(\text{bilirubin}) \).

If the CVP’s for protime and albumin are computed using the transformed bilirubin, no transformations are required, as can be seen in Figure 2. If we do not transform bilirubin and compute the CVP’s, we obtain the plots in Figures 3a and 3b. The covariate albumin does not require a transformation, though case 231 is an extreme outlier. The CVP of protime using the untransformed bilirubin in the model does indicate a need for a transformation. The approximate transformation estimate of
λ produces the transform protime$^{-4}$. Note that case 107 (who has the largest pro-
time value) is driving the CVP and upon removal of this case a transformation is not
required. We are thus cautious to transform protime.

In their book, Fleming and Harrington used a log transform on bilirubin, protime
and albumin. Thus we end up with several models to compare:

1. MODEL 1: age, edema, protime, albumin, bilirubin

2. MODEL 2: age, edema, protime, albumin, log(bilirubin)

3. MODEL 3: age, edema, 1/protime$^4$, albumin, log(bilirubin)

4. MODEL 4: age, edema, log(protime), log(albumin), log(bilirubin)

The results of fitting the models may be found in Table 1. If we use the model
deviance as a metric for comparing competing models, Model 3 would be declared the
“best” model. We do, though, have trepidations about the transformation of protime
in Model 3. Figure 4 is a plot of the predicted survival times based on Models 2, 3
and 4. We see that for nearly every case the predicted times overlap, indicating that
the three models produce the same fit. Hence Model 2 gives just as good a fit as the
other models and is more parsimonious.

4.2 Proportional Odds Model and Lymphoma Data Set

As before, let $T$ be a random variable denoting time of failure and $\mathbf{x}$ an associated
vector of covariates. Following Bennett (1983), we define the proportional odds on
failure model as:

$$
\log \frac{F(t | \mathbf{x})}{1 - F(t | \mathbf{x})} = -\beta' \mathbf{x} + \log \Gamma(t),
$$

13
where $F(t \mid x) = P(T \leq t \mid x)$, and $\Gamma(t)$ is an unspecified increasing, continuous function on $[0, \infty)$ satisfying $\Gamma(0) = 0$ and $\Gamma(\infty) = \infty$. In an earlier paper (Parzen and Harrington, 1992) we provided a method for the adaptive estimation of both $\beta$ and $\Gamma$. Under the proportional hazards model, the hazard ratio for two individuals will converge to one, while maintaining a constant odds ratio.

The proportional odds model can be written as the linear transformation model

$$h(T) = \beta'x + \epsilon,$$

where $h(t) = \log \Gamma(t)$ and $\epsilon$ is a random error with standard logistic distribution.

In this section we will investigate the use of the CVP for the proportional odds model. The idea is to fit an initial proportional odds model with all covariates believed needed for explanation. From this fit we obtain $\hat{\Gamma}$. Using $\hat{\Gamma}$ as an approximation to the true $\Gamma$, we may apply the CVP to the model

$$\log \hat{\Gamma}(T) = \beta'x + \epsilon$$

to determine if transformation of any of the parameters is necessary. That is, assuming the shape of $\Gamma$ does not change substantially if covariates are transformed, we can assume we have a linear failure time type model and use the concepts presented in the previous section. We will illustrate this technique with a lymphoma data set.

### 4.2.1 Lymphoma Data Set

To establish a current universally recognized prognostic factor model for aggressive non-Hodgkin’s lymphoma (NHL), 16 cooperative groups or single institutions in Europe, Canada and the USA participated in an overview analysis of 3273 patients with
stage I-IV diffuse, mixed, large cell or immunoblastic lymphoma treated with intensive combination therapy between 1982 and 1985 (Shipp et al., 1993). In a step-wise Cox regression analysis using 1385 patients with complete data on all major clinical variables, five features were independently associated with survival; age (age: ≤ 60 vs. > 60), stage (stage:I/II vs. III/IV), number of extranodal sites (xtra:0,1 vs. > 1), performance status (ps:0,1 vs. ≥ 2), and serum LDH level (ldh).

Figure 5 displays plots of hazard function estimates stratified by the variables age, stage, xtra and ps. Note that the hazards are not proportional but do indeed seem to converge to one. Also, there is only 15% of the observation times past the 5 year mark so we tend to discount the wide fluctuations seen after that time point. Hence the proportional odds model would appear to be a logical choice for analysis.

The result of fitting a proportional odds model relating failure to the five covariates age, stage, xtra, ps and ldh is shown in Table 2. All covariates are significant in the model. The only continuous covariate in our model is ldh, and we are interested if ldh needs to be transformed.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-8.545</td>
<td>1.099</td>
</tr>
<tr>
<td>Stage</td>
<td>-5.933</td>
<td>1.278</td>
</tr>
<tr>
<td>LDH</td>
<td>-1.652</td>
<td>0.323</td>
</tr>
<tr>
<td>Performance Status</td>
<td>-9.133</td>
<td>1.434</td>
</tr>
<tr>
<td>No. of Extra Nodel Sites</td>
<td>-4.846</td>
<td>1.262</td>
</tr>
</tbody>
</table>

Using the estimated $\hat{\Gamma}$ from our initial fit, we use the CVP with ldh and obtain a p-value ≤ .01 for the hypothesis of no transform needed. The estimated transform is $\hat{\lambda} = -.16$ which for simplicity we take to be $\lambda = 0$ (indicating the log transform). In Figure 6 we see the CVP for ldh. There are two points that are clearly outliers,
cases 29 and 314. These two patients had the the largest ldh values and in fact were more than 75% larger than all the other ldh values. If we delete these two points we still need a log transform (p-value ≤ .01) so without any additional information we kept the individuals in the analysis. This raises an issue for future study of whether deleting a case from the search for a transformation also implies the case should be deleted from the entire analysis.

Since the estimated transform is \( \hat{\lambda} = -.16 \) to be thorough we compare models which have ldh transformed with \( \lambda = -.5 \) and with \( \lambda = 0 \):

1. MODEL 1: age,xtra,ps,stage,ldh
2. MODEL 2: age,xtra,ps,stage,log(ldh)
3. MODEL 3: age,xtra,ps,stage,1/sqrt(ldh)

Table 3 has the output from fitting the three models. Using the deviance as a goodness-of-fit measure, we see that Model 2 (the model with log(ldh)) gives a better fit. Recall though that we assumed that we knew the true \( \Gamma \) when using the diagnostics. Since the proportional odds model is semi-parametric, given transformed covariates, we might obtain a different value of \( \Gamma \). In (Parzen and Harrington, 1992) we found the estimates of \( \beta \) to be robust with regards to the estimate of \( \Gamma \). In Figure 7 we see that the converse might be true; the estimate of \( \Gamma \) is robust to the covariates in the model. Figure 7 has the computed \( \hat{\Gamma} \) under Model 1 and Model 2. We see that the shape of the functions are very similar. This will be an area of future research but it implies that that the CVP can be used for model diagnostics for the proportional odds model.
5 Conclusion

For the most part, model building is still a combination of *a priori* domain knowledge, interpreting and understanding diagnostics, and art. In this paper we have extended the use of the added variable and constructed variable plot to linear failure time models with censored data. The methods are easy to implement, and require little additional computation. A useful feature of the CVP is the ability to not just indicate that a transformation is needed, but also give an estimate of the necessary transformation. One area of future research is the synergy between deleting observations when searching for covariate transformations and then performing data analysis on the complete data set. A function written for the Splus statistical system is available from the authors to compute the AVP and CVP.
Table 1: Comparison of Four Regression Models for the PBC Liver Data Set.

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>MODEL 1</th>
<th>MODEL 2</th>
<th>MODEL 3</th>
<th>MODEL 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>3.006404</td>
<td>3.668508</td>
<td>1.027082</td>
<td>6.599862</td>
</tr>
<tr>
<td></td>
<td>(0.786438*)</td>
<td>(0.828919)</td>
<td>(0.653701)</td>
<td>(1.66007)</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.021137</td>
<td>-0.020558</td>
<td>-0.0204857</td>
<td>-0.020437</td>
</tr>
<tr>
<td></td>
<td>(0.006070)</td>
<td>(0.005471)</td>
<td>(0.005471)</td>
<td>(0.005472)</td>
</tr>
<tr>
<td>ALBUMIN</td>
<td>0.791639</td>
<td>0.5934192</td>
<td>0.5909394</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.149596)</td>
<td>(0.144176)</td>
<td>(0.145228)</td>
<td></td>
</tr>
<tr>
<td>PROTIME</td>
<td>-0.186051</td>
<td>-0.176960</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.049142)</td>
<td>(0.052643)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDEMA</td>
<td>-0.382593</td>
<td>-0.355688</td>
<td>-0.3408063</td>
<td>-0.344777</td>
</tr>
<tr>
<td></td>
<td>(0.156515)</td>
<td>(0.147543)</td>
<td>(0.149869)</td>
<td>(0.147518)</td>
</tr>
<tr>
<td>BILI</td>
<td>-0.080122</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.008979)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/(PRO^4)</td>
<td>9334.984</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2869.885)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log(ALB)</td>
<td></td>
<td>1.915876</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.439186)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>log(PRO)</td>
<td></td>
<td>-2.172475</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.645425)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>log(BILI)</td>
<td></td>
<td>-0.560616</td>
<td>-0.5443373</td>
<td>-0.555289</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.059212)</td>
<td>(0.0.06025)</td>
<td>(0.059087)</td>
</tr>
<tr>
<td>deviance</td>
<td>407.2493</td>
<td>389.4648</td>
<td>387.0245</td>
<td>387.7282</td>
</tr>
</tbody>
</table>

(*) Estimated Standard Error
Table 3: Comparison of Three Regression Models for the NHL Data Set.

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>MODEL 1</th>
<th>MODEL 2</th>
<th>MODEL 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>-0.08530269</td>
<td>-0.06789009</td>
<td>(-0.108082)</td>
</tr>
<tr>
<td></td>
<td>(0.00927121*)</td>
<td>(0.00945079)</td>
<td>(0.01033290)</td>
</tr>
<tr>
<td>STAGE</td>
<td>-0.05922869</td>
<td>-0.03296071</td>
<td>-0.07781356</td>
</tr>
<tr>
<td></td>
<td>(0.01101913)</td>
<td>(0.01174156)</td>
<td>(0.01159102)</td>
</tr>
<tr>
<td>LDH</td>
<td>-0.01654363</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00306800)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PS</td>
<td>-0.09115258</td>
<td>-0.07541994</td>
<td>-0.1113179</td>
</tr>
<tr>
<td></td>
<td>(0.01219877)</td>
<td>(0.01270744)</td>
<td>(0.01194947)</td>
</tr>
<tr>
<td>XTRA</td>
<td>-0.04841467</td>
<td>-0.03837594</td>
<td>-0.06398683</td>
</tr>
<tr>
<td></td>
<td>(0.01187186)</td>
<td>(0.01200894)</td>
<td>(0.01203929)</td>
</tr>
<tr>
<td>log(LDH)</td>
<td></td>
<td>-0.52152494</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.07845971)</td>
<td></td>
</tr>
<tr>
<td>1/(sqrt(LDH))</td>
<td></td>
<td>2.960065</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.51619558)</td>
<td></td>
</tr>
<tr>
<td>deviance</td>
<td>1583.807</td>
<td>1567.866</td>
<td>1582.687</td>
</tr>
</tbody>
</table>

(*) Estimated Standard Error
Figure 1: Constructed Variable Plot for covariate bilirubin. Score statistic p-value ≤ .001 and transformation estimate $\lambda = -0.517$. 
Figure 2: Constructed Variable Plot for protime and albumin using log(bilirubin). (a): Constructed Variable Plot for protime using log(bilirubin). Score statistic p-value = 0.236. (b): Constructed Variable Plot for albumin using log(bilirubin). Score statistic p-value = 0.244.
Figure 3: Constructed Variable Plot for protime and albumin using bilirubin. (a): Constructed Variable Plot for protime. Score statistic p-value = 0.029 and transformation estimate $\lambda = -4.38$. (b): Constructed Variable Plot for albumin. Score statistic p-value = 0.353.
Figure 4: Predicted survival times based on Models 2, 3 and 4. The plotted number corresponds to the model.
Figure 5: Stratified Non-Parametric Hazard Estimates for NHL Data Set.
Figure 6: Constructed Variable Plot for covariate ldh. Score statistic p-value ≤ .001 and transformation estimate λ = -0.16.
Figure 7: Baseline Odds of Failure for NHL Data Set.
References


