A Robust Parameterization for Unbounded Covariates Within the Cox Proportional Hazards Model

Richard J. Jackson* and Trevor F. Cox

Cancer Research UK Liverpool Cancer Trials Unit, University of Liverpool, Liverpool, UK

Abstract: The Cox proportional hazards model is widely used in the analysis of medical data either for survival or time to a particular event. Factors and continuous covariates can be easily incorporated into the model and hazard ratios calculated. The model can however be distorted when extreme value observations occur within a continuous covariate and the hazard ratio can become extremely large. To overcome this, transformations of the covariate are often made, which can be simple, e.g. log, or more sophisticated such as the fitting of a fractional polynomial. This paper takes a different approach and makes a transformation based on the logistic function that has the property that the hazard ratio is bounded. The models are introduced and discussed. Model diagnostics based on Schoenfeld residuals and the influence function are established and then data from a pancreatic cancer trial are used to illustrate the model.

Keywords: ESPAC 3 trial, hazard ratio, influence function, logistic function, Schoenfeld residuals.

INTRODUCTION

Cox proportional hazards modeling is in widespread use in medical and other contexts [1]. Here robust hazard models are proposed that extend the Cox model. Survival data will be the main focus but the robust models proposed can be used in any time-to-event situation. Survival data from a pancreatic cancer randomized controlled trial will be used for illustration.

For the Cox model, the hazard function, λ, is modeled as

$$\lambda = \lambda_0 \exp\left(\beta^T \mathrm{x}\right)$$

where x is a vector of covariates and β a vector of coefficients. For a factor, x_i, the associated coefficient, β_i, leads to the hazard ratio, $\gamma_i = \exp(\beta_i)$, whilst for a continuous variable β_i gives the increase in hazard ratio per unit increase in the value of the continuous variable. One problem with continuous variables is that modeling the hazard in such a linear way can mean large or extreme values of x are associated with very high hazards when this may be unrealistic in practice. The presence of only a single extreme value observation can be enough to violate any model assumptions of proportionality [2] and biased estimates produced. Also, the interpretation of the hazard ratio is awkward or inappropriate. To overcome these problems, sometimes a simple transformation can be made, for instance, log(x), or x^{-1} and this can be sufficient for obtaining a good model fit. A more sophisticated model is achieved by using a fractional polynomial approach [3], which uses a mixture of transformations, but both simple transformations and fractional polynomials can still be influenced by extreme observations. In this paper a transformation based on the logistic function is proposed that can improve the model fit and guard against extreme observations having undue influence on the overall model.

Some previous methods to account for extreme value observations have concentrated on amendments to the likelihood formulation. A good overview is given by Farcomeni and Ventura [4] with two approaches in particular given specific attention: an approach based on a weighted likelihood formulation for the Cox model, notably proposed by Bednarski [5] and Minder and Bednarski [6] and secondly, an approach using `trimmed' likelihoods given by Viviani and Farcomeni [2]. For weighted Cox regression, a likelihood is proposed in the form:

$$\mathrm{l}(\beta) = \log\left(\mathrm{L}(\beta)\right) = \sum_{i=1}^{N} \mathrm{A}\left(\mathrm{t_i}, \mathrm{z_i}\right)\left[\mathrm{z_i} - \frac{\sum_{\mathrm{R(j)}} \mathrm{A}\left(\mathrm{t_i}, \mathrm{z_i}\right) \mathrm{z_j} \exp\left(\beta^{\mathrm{T}} \mathrm{z_j}\right)}{\sum_{\mathrm{R(j)}} \mathrm{A}\left(\mathrm{t_i}, \mathrm{z_i}\right) \exp\left(\beta^{\mathrm{T}} \mathrm{z_j}\right)}\right]$$

Here $\mathrm{A}(\mathrm{t_i}, \mathrm{z_i})$ is a smooth non-negative function which takes a value zero for either large values of t or β^Tz and R (j) the usual risk set at time t_i. This method down-weights or completely ignores patients who either have large covariate values or who live longer than may be expected. The second approach uses a trimmed likelihood by excluding observations that give the smallest contribution to the likelihood. Whilst either procedure may produce more robust hazard ratios they do not cure the problem of non-proportionality. More troublesome may be that the model is explicitly treating some data as less valuable than others and a possible

*Address correspondence to this author at the Cancer Research UK Liverpool Cancer Trials Unit, University of Liverpool, Block C Waterhouse Building, 1-3 Brownlow Street, Liverpool, L69 3GL, UK; Tel: +44 (0)151 7948834;
E-mail: richj23@liv.ac.uk

E-ISSN: 1929-6029/14

criticism is that the methods can be seen as trying to amend the data to fit a model as opposed to producing a model to fit the data.

A ROBUST PARAMERIZATION

Let the hazard function be modelled as

$$\lambda = \lambda_0 f\left(\theta, x\right)$$

where θ=(α, δ, ω, β) are parameters to be estimated. The family of transformations proposed here has the form

$$f\left(\theta, x\right) = \frac{\delta + \alpha \exp\left(\beta x\right)}{\omega + \exp\left(\beta x\right)} \qquad \text{(model 1)}$$

This is an adaption of the logistic function and has asymptotes α and δ/ω. Restrictions are needed on the parameters: α≥0, δ≥0, ω≥0, in order for f(θ,x) to be non-negative. The first derivative of f(θ,x) is $(\omega\alpha-\delta)\beta\exp(\beta x)/\{\omega+\exp(\beta x)\}^2$ and in order for a positive β to have positive slope for f, and correspondingly, a negative β to have a negative slope for f, then δ<ωα. Also, f(θ,x) is monotonically increasing in x which is usually a useful property in practice. A particular fractional polynomial might not possess this property. Model 1 has the property that the hazard function is symmetric regarding the baseline hazard, i.e. f(θ,x) and 1/f(θ,x) have the same functional form for the two reciprocal models $\lambda=\lambda_0 f(\theta,x)$, and $\lambda_0=\lambda/f(\theta,x)$.

A desirable property for f(θ,x) is that when β = 0, implying that the covariate has no effect on survival, then f(θ,x) should have the value unity. This implies ω=δ+α-1 and leads to model 2.

$$f\left(\theta, x\right) = \frac{\delta + \alpha \exp\left(\beta x\right)}{\delta + \alpha - 1 + \exp\left(\beta x\right)} \qquad \text{(model 2)}$$

The asymptotes for model 2 are α and δ/(δ+α-1) and it still retains baseline hazard symmetry. For α> 1, positive β will give a positive slope for f and negative β a negative slope. The value of x which has no effect on the baseline hazard is x=0. If this should be a different value then the variable x should be adjusted accordingly with a linear transformation. Note if a parameter is entered into the model, replacing x by x-ζ with estimation of ζ, then it can be shown that, by rearranging parameters, model 2 reverts back to model 1.

The slopes of the logistic function at +x and –x are identical. For model 2 to have this property, then δ=2-α and hence

$$f\left(\theta, x\right) = \frac{2 - \alpha + \alpha \exp\left(\beta x\right)}{1 + \exp\left(\beta x\right)} \qquad \text{(model 3)}$$

where the asymptotes are α and 2-α. This model loses its baseline hazard symmetry.

For model 2 to have reciprocal asymptotes, α and 1/α, then δ=1 giving

$$f\left(\theta, x\right) = \frac{1 + \alpha \exp\left(\beta x\right)}{\alpha + \exp\left(\beta x\right)} \qquad \text{(model 4)}$$

This model retains baseline hazard symmetry.

Lastly the standard Cox model is obtained by letting α become infinite in model 4, or by letting α = 0 which will negate the β coefficient. If α = 1 then f (θ, x) = 1 with x having no effect on the hazard function.

In this paper, concentration is on model 4 although the other models could be used and fitted to data in a similar manner to that for model 4.

Fitting The Model

Suppose the explanatory variables consist of p-1 binary variables, $x_1,\ldots,x_{p-1}$, representing various factors and one continuous covariate, x_p, to be fitted in the proportional hazards models. The hazard for the *ith* observation is

$$\lambda_i = \lambda_0 \exp\left(\beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}\right)\left\{1 + \alpha \exp\left(\beta_p x_p\right)\right\} / \left\{\alpha + \exp\left(\beta_p x_p\right)\right\}$$

and then the partial likelihood is given by

$$\prod_{i=1}^{N} \left\{ \frac{\lambda_i}{\sum_{j \in R(i)} \lambda_i} \right\}^{\delta i}$$

where *R(i)* is the risk set at the*i*th survival time and δ_i is the censoring value (1 – the event occurred, 0 – the time is censored). Here, the partial likelihood was maximized using the “optim” package within the statistical package R.

A Simulation Study

A small simulation study was carried out to investigate model fitting and accuracy. Survival times were simulated from an exponential distribution with baseline hazard set at 0.5 and with 5% of observations

Table 1: Results of Fitting the Standard and the Robust Model to Data Simulated from the Standard Model (α=0)

N	Param.	Standard model					New model				
		Est. (s.e.)	Bias	Acc.	Cov.	ACIL	Est. (s.e.)	Bias	Acc.	Cov.	ACIL
100	β_{trt}	0.128 (0.231)	0.022	0.054	0.94	0.848	0.129 (0.227)	0.021	0.052	0.93	0.844
	β_{cov}	0.05 (0.005)	0	0	0.99	0.022	0.05 (0.005)	0	0	0.92	0.02
250	β_{trt}	0.127 (0.134)	0.023	0.018	0.69	0.518	0.126 (0.133)	0.024	0.018	0.68	0.516
	β_{cov}	0.05 (0.003)	0	0	0.68	0.013	0.049 (0.004)	0.001	0	0.63	0.012
500	β_{trt}	0.151 (0.093)	-0.001	0.009	0.97	0.361	0.144 (0.097)	0.006	0.009	0.94	0.36
	β_{cov}	0.05 (0.002)	0	0	0.96	0.009	0.049 (0.002)	0.001	0	0.92	0.009
1000	β_{trt}	0.152 (0.063)	-0.002	0.004	0.95	0.253	0.148 (0.067)	0.002	0.005	0.93	0.252
	β_{cov}	0.05 (0.002)	0	0	0.93	0.006	0.049 (0.002)	0.001	0	0.78	0.006

randomly censored. Explanatory variables were a two-level factor representing treatment within a two-arm clinical trial alongside a continuous covariate. Parameters for the robust model are denoted by (β_{trt}, β_{cov}, α_{cov}). Firstly, a check was made on whether the new model could be fitted adequately to data arising from the standard Cox model by fixing the value of α_{cov} to be zero when simulating the data. The parameters (β_{trt}, β_{cov}) were given the values (0.15, 0.05). Patients were split equally between the two arms. The covariate was simulated as log (x) ~ N (3.5, 1.5). Sample sizes of 100, 250, 500 and 1000 were used, each time simulating 1000 datasets. Each fitted model was assessed in terms of bias, accuracy, coverage and average confidence interval length (ACIL) [7]. Table **1** shows the results of fitting the standard Cox model and the new model. It can be seen that there is very good agreement between the estimates for two models. The estimated values of α_{cov} (not shown) for the robust model were large enough to essentially make the model equivalent to the standard Cox model.

Next, data were simulated from the robust model formulation with the parameters (β_{trt}, β_{cov}, α_{cov}) and given the values (0.15, 0.05, 5). Table **2** shows the results of the simulations where the new model, the standard Cox model, the standard Cox model with log-transformed covariate values and a fractional polynomial model are fitted. The new model fits the data well and recovers the true parameter values accurately. The bias in the treatment coefficient is very small. There is some small reduction in the coverage for β_{cov} in the robust models and upon further inspection, this can be attributed to some skewness in the distribution of β_{cov}. The standard model underestimates the treatment coefficient even for a sample size of 1000. Note, β_{cov} cannot be compared across the two models. The log-transformed model achieves similar bias to the robust model but the fractional polynomial model offers little improvement over the standard model.

MODEL DIAGNOSTICS

Two model diagnostics are explored for the new model, (i) residuals based on standard Schoenfeld residuals [8] and (ii) an analytical form of an influence function following the method of Reid and Crapeau [9].

Residuals

Schoenfeld residuals for a particular covariate are calculated using the partial derivative of the partial log-likelihood function with respect to the covariate's associated parameter and evaluating this at the maximum likelihood estimate. For robust model 4, there are two parameters α and β and so two Shoenfeld type residuals will be calculated. Differentiating the log-likelihood with respect to α gives

$$\sum_{i=1}^{N}\delta_i\left[\frac{\exp(\beta x_i)-1}{\alpha\{\alpha-1+\exp(\beta x_i)\}}-\frac{\sum_{j\in R}\frac{(\exp(\beta x_j)-1)(\exp(\beta x_j))}{\{\alpha-1+\exp(\beta x_j)\}^2}}{\sum_{j\in R}\frac{\alpha\exp(\beta x_j)}{\alpha-1+\exp(\beta x_j)}}\right]$$

and with respect to β,

$$\sum_{i=1}^{N}\delta_i\left[\frac{(\alpha-1)x_i}{\alpha-1+\exp(\beta x_i)}-\frac{\sum_{j\in R}\frac{\alpha(\alpha-1)x\exp(\beta x_j)}{\{\alpha-1+\exp(\beta x_j)\}^2}}{\sum_{j\in R}\frac{\alpha\exp(\beta x_j)}{\alpha-1+\exp(\beta x_j)}}\right].$$

Table 2: Simulation Results to Assess Fitting of the Robust Model

N	Param.	Standard model					New model				
		Est. (s.e.)	Bias	Acc.	Cov.	ACIL	Est. (s.e.)	Bias	Acc.	Cov.	ACIL
100	β_{trt}	0.127 (0.203)	0.023	0.042	0.954	0.811	0.154 (0.204)	−0.004	0.042	0.9522	0.820
β_{cov}		0.003 (0.002)					0.056 (0.028)	−0.006	0.001	0.916	0.094
α_{cov}							6.35 (3.682)	−1.35	15.381	0.968	13.206
250	β_{trt}	0.132 (0.130)	0.018	0.017	0.95	0.515	0.152 (0.135)	−0.002	0.018	0.938	0.507
β_{cov}		0.002 (0.001)					0.052 (0.016)	−0.002	0	0.932	0.055
α_{cov}							5.313 (1.120)	−0.313	1.353	0.968	4.33
500	β_{trt}	0.135 (0.092)	0.015	0.009	0.946	0.355	0.154 (0.091)	−0.004	0.008	0.938	0.356
β_{cov}		0.002 (0.001)					0.051 (0.010)	−0.001	0	0.932	0.037
α_{cov}							5.197 (0.725)	−0.197	0.565	0.966	2.911
1000	β_{trt}	0.131 (0.061)	0.019	0.004	0.954	0.25	0.150 (0.061)	0	0.004	0.964	0.25
β_{cov}		0.001 (0.001)					0.050 (0.007))	0	0	0.914	0.026
α_{cov}							5.067 (0.493)	−0.067	0.247	0.962	1.984
N	Param.	log transformed models					Fractional polynomial model				
		Est. (s.e.)	Bias	Acc.	Cov.	ACIL	Est. (s.e.)	Bias	Acc.	Cov.	ACIL
100	β_{trt}	0.156 (0.220)	-0.006	0.048	0.952	0.813	0.122 (0.249)	0.028	0.063	0.914	0.777
β_{cov}		0.357 (0.083)									
α_{cov}											
250	β_{trt}	0.146 (0.129)	0.004	0.017	0.95	0.505	0.136 (0.169)	0.014	0.029	0.932	0.503
β_{cov}		0.344 (0.049)									
α_{cov}											
500	β_{trt}	0.147 (0.092)	0.003	0.008	0.948	0.355	0.14 (0.131)	0.01	0.017	0.928	0.353
β_{cov}		0.344 (0.035)									
α_{cov}											
1000	β_{trt}	0.145 (0.062)	0.005	0.004	0.95	0.25	0.137 (0.079)	0.013	0.006	0.95	0.249
β_{cov}		0.341 (0.025)									
α_{cov}											

The individual terms to the right in the overall sums give the Shoenfeld type residuals, a pair for each observed survival time. The residuals are not linked to x directly, but to the terms to the left within the overall sums.

Influence Function

The influence function measures the rate of change in a statistical functional when there is a small amount of contamination from another distribution and is defined as

$$I(x) = \lim_{\in \to 0}\left[\frac{T\{(1-\varepsilon)F+\varepsilon\delta_x - T(F)\}}{\in}\right],$$

where *T* is the statistical functional giving the parameter of interest, *F* is the underlying distribution of the data and δ_X is the contamination introduced into the distribution. Replacing *F* by F_n, the empirical distribution function, $T(F_n)$, will be the estimate of *T*(*F*) and the corresponding empirical influence function will measure the dependence of the estimate on particular data values.

Read and Crepeau establish the influence function for the proportional hazards model. They show this to be

$$\mathrm{I} = \mathrm{A}^{-1}\left(\hat{\beta}\right)\delta_{\mathrm{i}}\left\{\mathrm{z}_{\mathrm{i}} - \sum\nolimits_{\mathrm{Ri}} \mathrm{z}_{j} \exp\left(\beta^{T}\mathrm{z}_{j}\right) / \sum\nolimits_{\mathrm{Ri}} \exp\left(\beta^{T}\mathrm{z}_{j}\right)\right\}$$
$$+\mathrm{A}-1\left(\hat{\beta}\right)C_{i}\left(\hat{\beta}\right)$$

where

$$\mathrm{A}\left(\hat{\beta}\right) = \mathrm{n}^{-1}\sum\nolimits_{\mathrm{i}=1}^{N}\delta_{\mathrm{i}}\left[\sum\nolimits_{Ri}\mathrm{z}_{j}\mathrm{z}_{j}{}^{\mathrm{T}}\exp\left(\beta^{T}\mathrm{z}_{j}\right) / \sum\nolimits_{Ri}\exp\left(\beta^{T}\mathrm{z}_{j}\right) - \right.$$
$$\left.\left\{\sum\nolimits_{Ri} zj \exp\left(\beta^{T}\mathrm{z}_{j}\right) / \sum\nolimits_{Ri}\exp\left(\beta^{T}\mathrm{z}_{j}\right)\right\}\left\{\sum\nolimits_{Ri} zj \exp\left(\beta^{T}\mathrm{z}_{j}\right) / \sum\nolimits_{Ri}\exp\left(\beta^{T}\mathrm{z}_{j}\right)\right\}^{\mathrm{T}}\right]$$

and

$$Ci\left(\beta\right) = \exp\left(\beta^{T}z_{i}\right)\left(\begin{array}{l} \sum\nolimits_{tj\le ti}\delta_{j} \\ \left[\sum\nolimits_{Rj}\mathrm{z}_{k}\exp\left(\beta^{T}\mathrm{z}_{k}\right) / \left\{\sum\nolimits_{Rj}\exp\left(\beta^{T}\mathrm{z}_{k}^{T}\right)\right\}^{2}\right] \\ -\mathrm{z}_{i}\sum\nolimits_{tj\le ti}\delta_{j}\left\{1 / \sum\nolimits_{Rj}\exp\left(\beta^{T}\mathrm{z}_{k}\right)\right\} \end{array}\right)$$

The algebra involved to arrive at this result is heavy and not particularly informative. A similar result was found for the robust models 1, 2, 3 and 4 where the algebra was even more involved and lengthy and so is not repeated here. Details are available from the authors and also will appear in a PhD thesis written by Jackson.

APPLICATION TO DATA FROM THE ESPAC 3 TRIAL

Robust model 4 was applied to data from the ESPAC-3 trial set up to investigate the effect of adjuvant chemotherapy on patients with resectable pancreatic cancer. Of particular interest are the group of patients who had pancreatic ductal adenocarcinomas (PDAC) and for whom a value of post operative CA19.9 was recorded (n=759). It is reasonably assumed that information for this covariate is missing completely at random and no bias is introduced by considering a complete case analysis. Previously published analyses [10] are followed, forcing the terms `Resection Margin' (Negative vs. Positive) and `Treatment Arm' (5FU vs. Gemcitabine) into the model as stratification factors. Also identified as important are `Lymph Nodes' (Negative vs. Positive), `Tumour Differentiation' (Poor vs. Moderate vs. Well) and `Smoking Status' (Never vs. Past vs. Present vs. Missing).

Figure **1** gives a histogram of CA19.9 values which is seen to have a very skewed distribution and prone to extreme value observations. The median (inter quartile range) is 24 (10, 63) but there are a number of observations greater than 1,000; only values up to 2,000 are displayed, the largest recorded being 37,000.

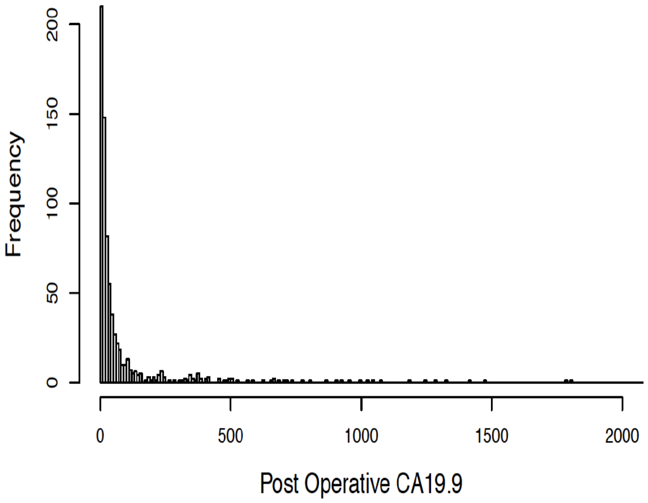

Figure 1: Histogram of post operative CA19.9 values.

Four models were fitted to the data, (i) standard Cox proportional hazards with raw CA19.9 data (Reference model), (ii) standard Cox proportional hazards with log transformed CA19.9 (Log model), (iii) a fractional polynomial model for CA19.9 (Frac. polyn. model) and (iv) robust model 4(Robust model). Table **3** shows the log-likelihood, Akaiki's information criterion (AIC), the model coefficients and their estimated standard errors. For the reference model, small estimates of β and for the estimated standard error are obtained. This is a consequence of the large extreme values observed. Taking as an example, the median value for CA19.9 as 24, a hazard ratio of 1.02 is obtained showing very modest increases in the baseline hazard. For extreme values of 2,000, 5,000 and 37,000, hazard ratios of 1.17, 1.49 and 18.9 are obtained. A clinician, however, may find it difficult to believe that a patient with CA19.9 value of 37,000 has an instantaneous risk of death of almost 20 times that of a patient with a zero value. The log-transformed model gives an improved model fit as shown by an AIC of 6882 compared to 6912 for the reference model. The hazard ratios for the reference values of 24, 2,000, 5,000 and 37,000 for CA19.9 are 1.95, 4.93, 5.90 and 9.11 respectively. Here, extreme hazard ratios are avoided to a small extent. Patients with a median value of CA19.9 are almost twice more likely to die at any given time point as those with a zero value.

Table 3: Results of Fitting Four Models to the ESPAC 3 Pancreatic Adenocarcinoma Survival Data

		Model							
		(i) Reference		(ii) Log		(iii) Frac. polyn.		(iv) Robust model 2	
		Log-lik.	AIC	Log-lik.	AIC	Log-lik.	AIC	Log-lik.	AIC
		-3447	6912	-3432	6882	-3412	6847	-3420	6861
Factor	Level	coef.	s.e.	coef.	s.e.	coef.	s.e.	coef.	s.e.
Resec. Margin	Neg.								
	Pos.	0.21	0.09	0.19	0.09	0.20	0.09	0.18	0.09
Treatment	5FU								
	Gem.	-0.12	0.08	-0.10	0.08	-0.11	0.08	-0.09	0.08
Lymph N.	Neg.								
	Pos.	0.55	0.10	0.48	0.10	0.46	0.10	0.46	0.10
Tumour Diff.	Poor								
	Mod.	-0.29	0.10	-0.29	0.10	-0.27	0.10	-0.30	0.10
	Well	-0.64	0.15	-0.62	0.15	-0.69	0.15	-0.63	0.15
Smoke	Never								
	Past	0.09	0.10	0.08	0.10	0.07	0.10	0.08	0.10
	Present	0.24	0.12	0.26	0.12	0.27	0.12	0.27	0.12
	Missing	0.22	0.18	0.22	0.18	0.21	0.18	0.17	0.18
CA19.9	α					β_1=0.02	3.87e-3	3.77	0.70
	β	7.95e-5	1.54e-5	0.21	0.03	β_2=0.32	0.03	0.01	0.00

The fractional polynomial fitted was $\beta_1 \times$ 100/(CA19.9+1) + $\beta_2 \times$ log{(CA19.9+1)/100}.

The fractional polynomial model had the lowest AIC with a value of 6847. The fractional polynomial that was produced was

$$\beta_1 \times \frac{100}{\mathrm{CA19.9}+1} + \beta_2 \times \log\left\{\frac{(\mathrm{CA19.9}+1)}{100}\right\}$$

This pair of transformations chosen by the fractional polynomial software might not be interpretable by clinicians and one problem with fractional polynomial regression is that a new data set generated under the same conditions can easily give rise to different transformations. As an illustration, the ESPAC 3 data were randomly split into two equal sized subsets of the data and fractional polynomial models fitted separately to both. The functional form of the two fractional polynomials differed. They were

$$\beta_1 \times \{(\mathrm{CA19.9}+1)/1000\}^{-0.5} + \beta_2 \times \log\{(\mathrm{CA19.9}+1)/1000\}$$

and

$$\beta_1 \times \{(\mathrm{CA19.9}+1)/100\}^{-2} + \beta_2 \times \log\{(\mathrm{CA19.9}+1)/100\}$$

Returning to the fractional polynomial model fitted to the whole dataset, for the reference points of 24, 2,000, 5,000 and 37,000,hazard ratios of 1.73, 3.64, 4.55 and 7.79 are obtained.

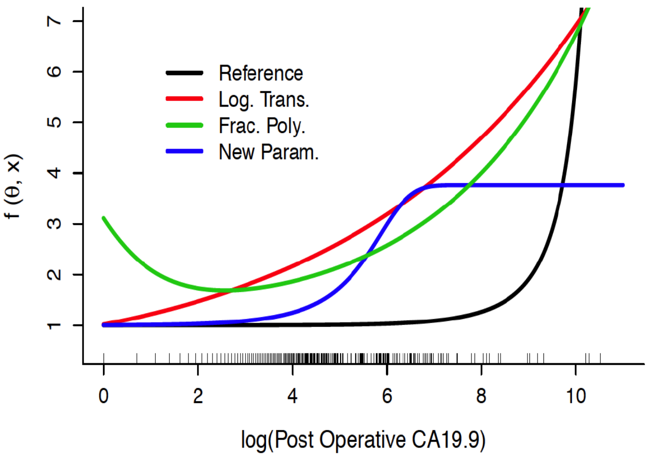

Figure 2: Hazard ratios plotted against log (CA19.9) for the four fitted models.

Model 4 has an AIC value of 6861 which is less than that for the log transformed model but more than that for the fractional polynomial model. The upper asymptote is 3.77 which corresponds to a maximum hazard ratio of also 3.77. A hazard ratio of 1.19 is obtained for the median CA19.9 value of 24. The other reference values of 2,000, 5,000 and 37,000 all have a hazard ratio of 3.77 obtained from the upper asymptote. From a clinical perspective, this is the most attractive model with modest small increases in the CA19.9 resulting in modest increases in the hazard ratio and larger values curtailed to ensure that unrealistically large hazard ratios are not obtained. This is highlighted in Figure **2** where the hazard ratio is plotted against log (CA19.9) for all four models. The hazard ratio for the log transformed model follows an exponential curve while for the standard Cox model the hazard ratio follows a curve exp (exp(x)) because the x-axis is on the log-scale. For the fractional polynomial model the hazard ratio first decreases and then increases which is unrealistic in practice and could make clinical interpretations troublesome. Furthermore, there is no value of CA19.9 that has zero effect on the baseline hazard function within the observed range of data and this may affect confounding in other covariates as the baseline hazard function is amended to account for this. This can be seen somewhat in the analysis of the ESPAC-3 dataset with some amendments in the point estimates, especially for the Tumour Differentiation covariates. Model 2 has the desired shape of curve for the hazard function and is bounded whereas all other models can have the hazard ratio increase indefinitely.

Model diagnostics

Model diagnostics were calculated for the four models, firstly residuals and then influence measures. Figure **3** shows Shoenfeld residuals for the parameters associated with CA19.9 for the four models fitted. The scales on the graphs are not comparable. The residuals from the extreme values can be seen in the plot for the reference model and also the two plots for the fractional polynomial model. The variance of residuals for the log model decreases as survival time increases. The two plots for robust model **4** show how the asymptote controls the residuals and also shows that the variability of the residuals with time is much less than for the other models.

Figure **4** shows the influence measures obtained for the four models, plotted against log (CA19.9) and where crosses mark observed events and circles censored events. For the reference model, there is no obvious relationship of the influence measures with

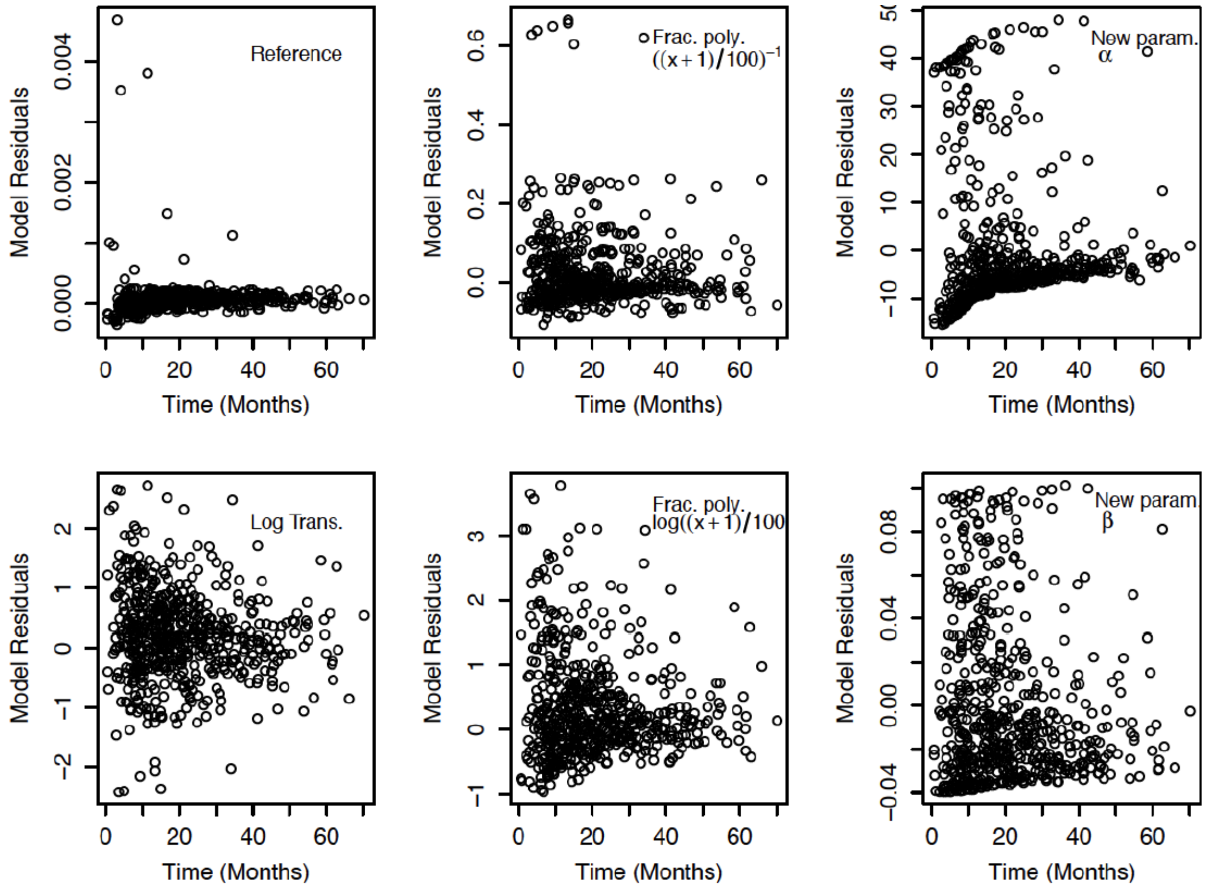

Figure 3: Schoenfeld residuals for the four models.

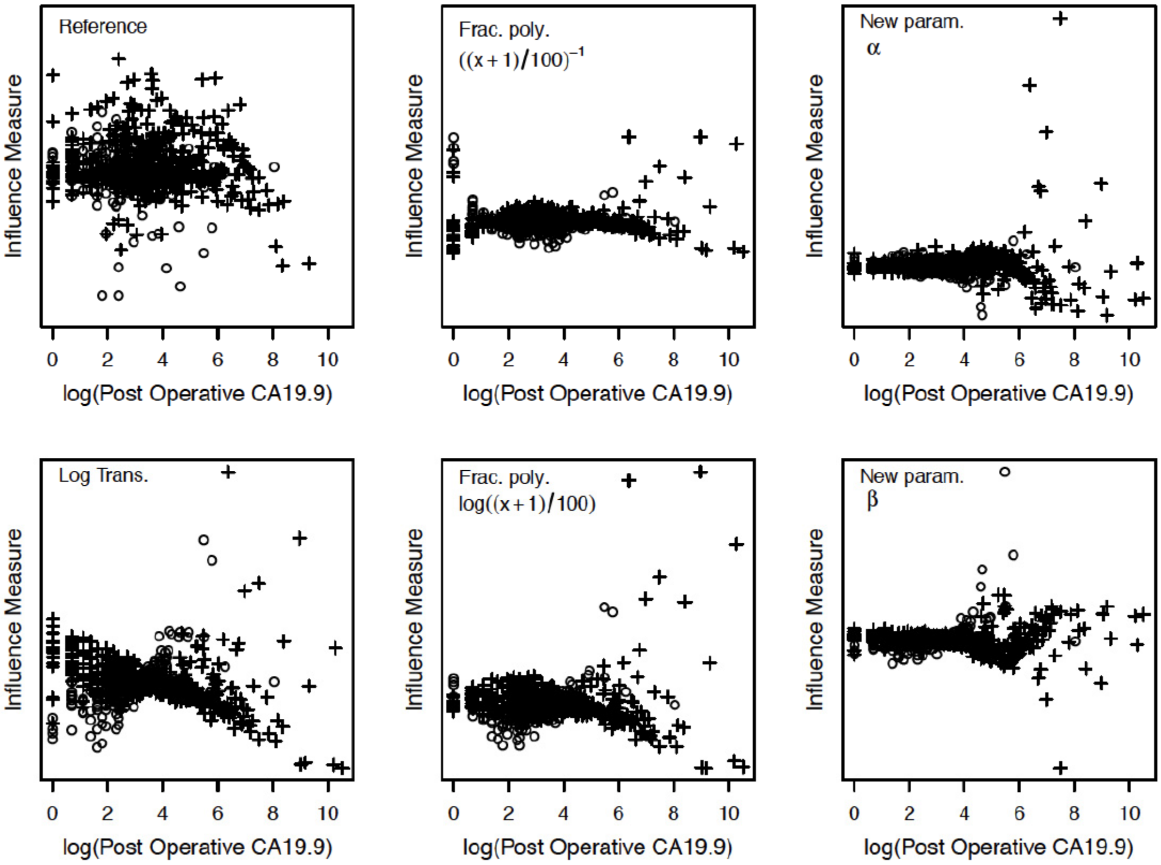

Figure 4: Influence measures for the four models.

CA19.9, whereas for the log-transformed model there is a central point of CA19.9 around the value 4. Either side of this point, there is a general divergence with both small and large values of CA19.9 having relatively large effects upon parameter estimation. For the fractional polynomial model, the influence measures associated with the $\{(x+1)/100\}^{-1}$ term show that very small values of CA19.9 can have a disproportionally large influence upon parameter estimation. There are also some large positive influence measures associated with log (CA19.9) values greater than 6. For the term given by log $\{(x+1)/100\}$, there is a similar relationship to that seen for the log transformed model although here the divergence from some central point is less pronounced. Large values of CA19.9 are again associated with typically large, positive influence measures. For robust model 4, there is neither the divergence away from some central point, nor any large influence measures associated with small values of CA19.9. However, the parameter α is associated with some large positive influence measures. This is to be expected, as this is the parameter associated with setting the upper asymptote. The estimate of the parameter is driven by the amount of `large' data that are observed and any single data value can have a relatively large effect on the estimate. Upon first inspection of the plot for β, apart from two large influence measures, there is fairly flat relationship with log(CA19.9), even at large values. Upon closer inspection, there is some change in the relationship between log (CA19.9) values of 4 and 6, and again between 6 and 8. This change, although small, can be seen to correspond to the points in the function that are chiefly concerned with the growth of the functional relationship and immediately afterwards as the asymptote is approached.

DISCUSSION

A method has been given to robustly model a continuous covariate in the Cox proportional hazards situation that automatically guards against extreme values and sets asymptotes for the minimum and maximum hazard ratios. This can be very useful in the clinical context. The model was successfully demonstrated on survival data following resection for pancreatic adenocarcinoma where CA19.9 is used as a biomarker. The distribution of CA19.9 is highly skewed and so was a good candidate for the robust parameterization.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. JP Neoptolemos for permission to use the ESPAC 3 trial data.

REFERENCES

[1] Cox DR. Regression Models and Life-Tables. JR Stat Soc Series B 1972; 34: 187-220.

[2] Viviani S, Farcomeni A. Robust estimation for the Cox regression model based on trimming. Biometrical Journal 2011; 53: 956-973. http://dx.doi.org/10.1002/bimj.201100008

[3] Royston P, Lambert PC. Flexible parametric survival analysis using Stata: Beyond the Cox model. Texas: Stata Press 2011.

[4] Farcomeni A, Ventura L. An overview of robust methods in medical research. Stat Methods Med Res 2010; 21: 111-133. http://dx.doi.org/10.1177/0962280210385865

[5] Bednarski T. Robust estimation in Cox's Regression models. Scand J Stat 1993; 20: 213-225.

[6] Minder CE, Bednarski T. A robust method for proportional hazards regression. Statist Med 1996; 15: 1033-1047. http://dx.doi.org/10.1002/(SICI)1097-0258(19960530)15:10<1033::AID-SIM215>3.0.CO;2-Y

[7] Burton A, Altman DG, Royston P, *et al.* The design of simulation studies in medical statistics. Statist Med 2006; 25: 4279-4292. http://dx.doi.org/10.1002/sim.2673

[8] Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika1982; 69, 239-241. http://dx.doi.org/10.1093/biomet/69.1.239

[9] Reid N, Crepeau H. Influence function for proportional hazards regression. Biometrika1985; 72: 1-9. http://dx.doi.org/10.1093/biomet/72.1.1

[10] Neoptolemos JP, Stocken DD, Bassi C, *et al.* Adjuvant Chemotherapy With Fluorouracil Plus Folinic Acid vs Gemcitabine Following Pancreatic Cancer Resection: A Randomized Controlled Trial. JAMA 2010; 304: 1073-1081. http://dx.doi.org/10.1001/jama.2010.1275

Received on 14-07-2014 Accepted on 03-09-2014 Published on 06-11-2014

http://dx.doi.org/10.6000/1929-6029.2014.03.04.1