Entropy **2015**, *17*, 3679-3691; doi:10.3390/e17063679

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Density Regression Based on Proportional Hazards Family

Wei Dang [1] and Keming Yu [2,3,*]

[1] Business School, Shihezi University, Xinjiang, 831300, China

[2] School of Management, Hefei University of Technology, Hefei, 230009, China

[3] Department of Mathematics, College of Engineering, Design and Physical Sciences, Brunel University London, Uxbridge, UB8 3PH, UK

* Author to whom correspondence should be addressed; E-Mail: keming.yu@brunel.ac.uk; Tel.: +44-1895-266128.

Academic Editors: Carlos Alberto de Bragança Pereira and Adriano Polpo

Received: 16 March 2015 / Accepted: 29 May 2015 / Published: 4 June 2015

Abstract: This paper develops a class of density regression models based on proportional hazards family, namely, Gamma transformation proportional hazard (Gt-PH) model . Exact inference for the regression parameters and hazard ratio is derived. These estimators enjoy some good properties such as unbiased estimation, which may not be shared by other inference methods such as maximum likelihood estimate (MLE). Generalised confidence interval and hypothesis testing for regression parameters are also provided. The method itself is easy to implement in practice. The regression method is also extended to Lasso-based variable selection.

Keywords: best linear unbiased estimators (BLUE); density regression; exact inference; gamma random variable; proportional hazards distribution family; regression analysis

1. Introduction

Many regression models can be derived by allowing probability distribution parameter(s) to depend on covariates, such as generalised linear model (GLM) [1] which typically assumes the mean of a distribution depends on covariates via a link function. Examples include scale parameter regression [2–6] and variance function regression [7]. In particular, scale parameter is often well-defined when variance function may be infinity under heavy tail circumstance. Scale function provides a more nature dispersion measure than variance for the non-Gaussian case [8]. Moreover, scale function is a robust measure [2,9] of loss amounts paid by an insurance company and demand of a product as depicted by its sales via the popular 'The SEVERITY Procedure' in software *SAS* (www.sas.com).

In this paper we propose a class of density regression models based on proportional hazards family, which is introduced via the scale parameter of the proportional hazards distribution. The distribution family is defined as follows: assume that the response variables Y has the probability distribution belong to proportional hazards family $F(y;\lambda,\theta)$ or proportional reverse hazards family $F_r(y;\lambda,\theta)$ below:

$$\begin{aligned} F(y;\lambda,\theta) &= 1-[1-G(y;\lambda)]^{\theta}, \\ F_r(y;\lambda,\theta) &= [G(y;\lambda)]^{\theta}, \end{aligned} \tag{1}$$

where θ is usually a scale parameter and $G(\cdot;\lambda)$ is a distribution function possibly dependent only on λ. This family of distributions $\{F(y;\lambda,\theta),\theta>0\}$ is discussed in [10] (Section 7.E. ff.). In general, we will call θ the proportional parameter and λ the G-parameter. $G(y;\lambda)$ could even belong to a one-parameter exponential family: its density function is given by $a(x)\exp(\lambda^T T(x)-A(\lambda))$. Therefore, the distribution family (1) is indeed a big family of distributions. Examples of family (1) include Weibull distribution, Gompertz distribution, Lomax distribution, Exponential distribution, Burr type XII distribution, Kumaraswamy distribution and so on. For example, when $G(y;\lambda)=1-\exp\{-y^{\lambda}\}$ in family (1), we have the two-parameter Weibull distribution.

The proportional hazard family (1) has been extensively used to model failure time and carry out survival analysis. For example, based on the proportional hazards family $F(y;\lambda,\theta)=1-[1-G(y;\lambda)]^{\theta}$, the hazard rate of F equals to $\theta\frac{g(y;\lambda)}{\overline{G}(y;\lambda)}$, which is proportional to the hazard rate $\frac{g(y;\lambda)}{\overline{G}(y;\lambda)}$ of G. Where g is the density function of G, and $\overline{G}(y;\lambda)=1-G(y;\lambda)$ is the survival function or reliability function of $G(y;\lambda)$.

Like GLM, the effect of covariates $\boldsymbol{x}$ on Y could be set up by modelling parameter θ as a function of $\boldsymbol{x}$. In this paper we model $\log(\theta)$ as a linear regression function of covariates $\boldsymbol{x}$, that is,

$$\log(\theta(\boldsymbol{x}))=\boldsymbol{x}^T\boldsymbol{\beta}, \tag{2}$$

where $\boldsymbol{\beta}$ standards for regression coefficients and $\boldsymbol{x}^T$ is the transpose of $\boldsymbol{x}$. Clearly, model (2) includes the methods in the SAS SEVERITY procedure which can model the effect of exogenous or regressor variables on a probability distribution, as long as it has a scale parameter.

Next we develop a method to inference $\boldsymbol{\beta}$ in Equation (2) via a combination of Gamma random variable based transformation and ordinary least squares (OLS) estimate. This method is particularly suitable for small sample without using bootstrap or Bayesian inference. This method is totally different from existing methods used to fit parametric regression models such as maximum likelihood estimate (MLE). Section 2 details the exact inference with known parameter λ or $G(y;\lambda)$, including derivation of an unbiased minimum variance estimate for $\boldsymbol{\beta}$ and hazard ratio of the family (1).Generalised confidence interval estimation and hypothesis testing of regression models are also provided. Section 3 extends the method to unknown parameter λ. Section 4 further introduces adaptive-Lasso based method for the proposed Gamma transformation proportional hazard (Gt-PH) regression when variable selection is required for checking the effect of a big number of covariates $\boldsymbol{x}$. The numerical performance of the proposed regression model estimation, particularly, the algorithm and the comparison with MLE and asymptotic confidence interval, as well as a real data analysis, are illustrated in Section 5. Finally, a brief conclusion is presented in Section 6.

2. Inference Method for the Proportional Hazard Model

From now on we focus on proportional hazards family $F(y;\lambda,\theta) = 1 - [1 - G(y;\lambda)]^{\theta}$ but the method can be applied to proportional reverse hazards family without much change, see the brief discussion in Section 6.

First assume that λ or $G(y;\lambda)$ is known for the proportional hazards family (1). Given independent observations $\{\boldsymbol{x}_i, Y_i\}_{i=1}^{n}$ of $(\boldsymbol{x}, Y)$, the MLE of regression coefficient $\boldsymbol{\beta}$ in Equation (2) is usually derived by a likelihood function $L(\boldsymbol{\beta})$ such as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\,[\overline{G}(Y_i;\lambda)]^{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})-1} g(Y_i;\lambda).$$

The MLEs of $\boldsymbol{\beta}$ have no explicit form and only asymptotic unbiased under some regular conditions.

In this section we aim at a simply explicit estimation of $\boldsymbol{\beta}$ and derivation of its best linear unbiased estimators (BLUE). Here "best" means the lowest variance of the estimate among linear unbiased estimates.

Note that $Y \sim F(y;\lambda,\theta)$, so $F(Y;\lambda,\theta) \sim U[0,1)$, and $-\log[1 - F(Y;\lambda,\theta)] \sim Exp(1)$. That is, $-\theta\,\log(\overline{G}(Y;\lambda)) \sim Exp(1)$.

Given $\{\boldsymbol{x}_i, Y_i\}_{i=1}^{n}$, let $\theta_i \equiv \theta(\boldsymbol{x}_i)$, $S_i = -\log(\overline{G}(Y_i;\lambda))$ then $2\,\theta_i\,S_i \sim \chi^2(2)$. As the distribution $\chi^2(2)$ is a special case of a Gamma distribution from a Gamma random variable, saying, Γ, and note that a Gamma distribution is the maximum entropy probability distribution of Γ for which $E(\log(\Gamma)) = \psi(\text{shape} - \text{parameter}) - \log(1/(\text{scale} - \text{parameter}))$ is fixed, where $\psi(t) = d\log(\Gamma(t))/dt$ is the digamma function. Therefore we have

$$\begin{aligned} E[\log(S_i) + \log(\theta_i)] &= \psi(1), \\ Var[\log(S_i) + \log(\theta_i)] &= \psi'(1), \end{aligned} \qquad (3)$$

where $\psi'(t) = d^2\log(\Gamma(t))/dt^2$, $\psi(1) = -\gamma$ with the Euler-Mascheroni constant $\gamma \sim 0.5772$ and $\psi'(1) = \frac{\pi^2}{6}$.

Let

$$U_i = -\log(S_i) - \gamma, \qquad (4)$$

then Equation (3) can be re-written as a standard GLM:

$$E(U_i) = \boldsymbol{x}_i^T\boldsymbol{\beta},\ Var(U_i) = \psi'(1).$$

Or, in terms of vector and matrix, we have that the regression parameter vector $\boldsymbol{\beta}$ satisfies GLM

$$E(\boldsymbol{U}) = \boldsymbol{X}\boldsymbol{\beta},\ Var(\boldsymbol{U}) = \psi'(1)\mathbf{1}, \qquad (5)$$

with known model variance $\psi'(1)$. Where $\boldsymbol{X}$ is the matrix consisting of observations on covariates and has 1 as the 1st column, $\boldsymbol{U}$ is the vector consisting of observations of U_i. We name this method as Gt-PH model.

Therefore, according to the Gauss-Markov theorem, an exploit form of the BLUE of regression parameter vector $\boldsymbol{\beta}$ and the variance of the estimators of Gt-PH Model can be obtained as follows:

Theorem 1. *Under the distribution family (1) and Gt-PH Model (2), the BLUE of $\boldsymbol{\beta}$ is given by*

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{U},$$

and the variance of this estimator is given by

$$Var(\widehat{\boldsymbol{\beta}}) = \psi'(1)\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

Specially, under a simple linear model of (2): $\log(\theta) = \beta_0 + \beta_1 x$, *we have* $\hat{\beta}_0 = \bar{U} - \bar{x}\hat{\beta}_1 = -\frac{1}{n}\sum_i \log(S_i) - \gamma - \bar{x}\hat{\beta}_1$ *and* $\hat{\beta}_1 = \frac{\sum_{j=1}^n (x_{1j}-\bar{x})U_j}{\sum_{j=1}^n (x_j-\bar{x})^2} = -\frac{\sum_j (x_j-\bar{x})\log S_j}{\sum_j (x_j-\bar{x})^2}$. $var(\hat{\beta}_0) = \psi'(1)\,\frac{\bar{x}^2}{\sum_{j=1}^n (x_j-\bar{x})^2}$, $var(\hat{\beta}_1) = \frac{\psi'(1)}{\sum_{j=1}^n (x_j-\bar{x})^2}$, $cov(\hat{\beta}_0,\ \hat{\beta}_1) = -\psi'(1)\,\frac{\bar{x}}{\sum_{j=1}^n (x_j-\bar{x})^2}$.

When Theorem 1 provides the BLUE of regression parameter $\boldsymbol{\beta}$, we should discuss the estimation of $\theta(\boldsymbol{x}_0) = \exp(\boldsymbol{x}_0^T\boldsymbol{\beta})$ at covariates $\boldsymbol{x} = \boldsymbol{x}_0$ and the hazard ratio $HR = \exp((\boldsymbol{x}_a - \boldsymbol{x}_b)^T\boldsymbol{\beta})$ at covariates $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$, which are often interests of practical issues. Theorem 2 below gives the unbiased estimators of $\theta(\boldsymbol{x}_0)$ and HR.

Theorem 2. *Under the assumptions of Theorem 1, $\theta(\boldsymbol{x}_0)$ and hazard ratio HR are estimated by $\widehat{\theta}(\boldsymbol{x}_0) = \exp(\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}})$ and $\widehat{HR} = \exp((\boldsymbol{x}_a - \boldsymbol{x}_b)^T\widehat{\boldsymbol{\beta}})$, respectively. And their unbiased estimators are provided as follows.*

For $i = 1,\cdots,n$, let the indicator vector $\mathbf{e}_i = (0,\cdots,0,1,0,\cdots,0)^T$ with ith component as non-zero value 1 only, and let constant $c_i = -\boldsymbol{x}_0^T[\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\mathbf{e_i}$, $d_i = -(\boldsymbol{x}_a - \boldsymbol{x}_b)^T[\boldsymbol{X}^T\boldsymbol{X}]^{-1}\boldsymbol{X}^T\mathbf{e_i}$, then

(1) if all $1 + c_i > 0$, the unbiased estimator of $\theta(\boldsymbol{x}_0)$ is given by

$$\widetilde{\theta}(\boldsymbol{x}_0) = \exp(-\gamma\sum_{i=1}^n c_i)\frac{\hat{\theta}(\boldsymbol{x}_0)}{\prod_{i=1}^n \Gamma(1+c_i)},$$

with variance (if $1 + 2c_i > 0$):

$$Var(\tilde{\theta}(\boldsymbol{x}_0)) = \prod_{j=1}^n\left(\frac{\Gamma(1+2c_j)}{\Gamma^2(1+c_j)} - 1\right)\theta(\boldsymbol{x}_0)^2.$$

(2) if all $1 + d_i > 0$, the unbiased estimator of HR is given by

$$\widetilde{HR} = \exp(-\gamma\sum_{i=1}^n d_i)\frac{\widehat{HR}}{\prod_{i=1}^n \Gamma(1+d_i)},$$

with variance (if all $1 + 2d_i > 0$):

$$Var(\widetilde{HR}) = \prod_{j=1}^n\left(\frac{\Gamma(1+2d_j)}{\Gamma^2(1+d_j)} - 1\right)HR^2.$$

Proof. When λ is known, we try to derive the expectation of $\theta(\boldsymbol{x}_0)$ at covariate vector $\boldsymbol{x}_0 = (1, x_{01}, \cdots, x_{0n})^T$.

From

$$\log \widehat{\theta}(\boldsymbol{x}_0) - \log \theta(\boldsymbol{x}_0) = \boldsymbol{x}_0^T \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$$
$$= \boldsymbol{x}_0^T [\boldsymbol{X}^T \boldsymbol{X}]^{-1} \boldsymbol{X}^T [\boldsymbol{U} - E(\boldsymbol{U})],$$

and note that

$$\boldsymbol{U} = -(\log S_1, \cdots, \log S_n)^T - \gamma,$$

and

$$E(\boldsymbol{U}) = (\log \theta_1, \cdots, \log \theta_n)^T.$$

where each $S_i \sim Exp(\theta_i)$.

Let constant $c_i = -\boldsymbol{x}_0^T [\boldsymbol{X}^T \boldsymbol{X}]^{-1} \boldsymbol{X}^T \mathbf{e_i}$, then

$$\log \widehat{\theta}(\boldsymbol{x}_0) - \log \theta(\boldsymbol{x}_0) = \gamma \sum_{i=1}^{n} c_i + \log\left(\prod_{j=1}^{n} (\theta_j S_j)^{c_j}\right).$$

Note that each $\theta_j S_j \sim Exp(1)$ and all $\theta_j S_j$ are independent, so

$$\frac{\widehat{\theta}(\boldsymbol{x}_0)}{\theta(\boldsymbol{x}_0)} = \exp(\gamma \sum_{i=1}^{n} c_i) \prod_{j=1}^{n} (\theta_j S_j)^{c_j}.$$

$$E\left(\frac{\widehat{\theta}(\boldsymbol{x}_0)}{\theta(\boldsymbol{x}_0)}\right) = \exp(\gamma \sum_{i=1}^{n} c_i) \prod_{j=1}^{n} \Gamma(1 + c_j),$$

subject to all $1 + c_j > 0$.

$$Var\left(\frac{\widehat{\theta}(\boldsymbol{x}_0)}{\theta(\boldsymbol{x}_0)}\right) = \exp(2\gamma \sum_{j=1}^{n}) \prod_{j=1}^{n} \left(\Gamma(1 + 2c_j) - \Gamma^2(1 + c_j)\right).$$

subject to all $1 + 2c_j > 0$. Therefore, the unbias estimator of $\theta(\boldsymbol{x}_0)$ in Theorem 2 can be obtained from a simple modification of the estimator $\widehat{\theta}(\boldsymbol{x}_0)$. Along the exact same line, the unbias estimator of HR in Theorem 2 can be obtained from a simple modification of the estimator $\widehat{HR} = \exp((\boldsymbol{x}_a - \boldsymbol{x}_b)^T \widehat{\boldsymbol{\beta}})$. □

Clearly, these estimators provide exact estimation while exact statistical inference is preferable for many reasons, particularly when the sample size is small or not big enough.

2.1. Confidence Intervals

Confidence intervals (or prediction intervals) for fitting a regression function $\boldsymbol{x}_0^T \boldsymbol{\beta}$ of $\boldsymbol{x}^T \boldsymbol{\beta}$ at $\boldsymbol{x} = \boldsymbol{x}_0$ are useful and often required in practice. Recall an asymptotic normal based confidence interval is given by point estimate ± (critical value) (standard error of the estimate). The asymptotic normality based confidence interval can also be derived for $\boldsymbol{x}_0^T \boldsymbol{\beta}$ and $\theta(\boldsymbol{x}_0) = (\exp(\boldsymbol{x}_0^T \boldsymbol{\beta}))$. In fact, as the BLUE of a regression function $\boldsymbol{x}_0^T \boldsymbol{\beta}$ could be estimated by $\boldsymbol{x}_0^T \widehat{\boldsymbol{\beta}}$ with known variance $\frac{\pi^2}{6} \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0$, given a significant level $0 < \alpha < 1$, the approximate $(1 - \alpha)\%$ confidence interval of the regression function $\boldsymbol{x}_0^T \boldsymbol{\beta}$ is given by

$$\left(\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}} - Z_{\alpha/2}\,\pi\,\sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0/6},\ \ \boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}} - Z_{1-\alpha/2}\,\pi\,\sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0/6}\right).$$

Similarly, the approximate $(1-\alpha)\%$ confidence interval of $\theta(\boldsymbol{x}_0)$ is given by

$$\left(\widehat{\theta}(\boldsymbol{x}_0)\,e^{-Z_{\alpha/2}\,\pi\,\sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0/6}},\ \ \widehat{\theta}(\boldsymbol{x}_0)\,e^{-Z_{1-\alpha/2}\,\pi\,\sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0/6}}\right),$$

where Z_α is the α quantiles of standard normal distribution. Because $\widehat{\theta}(\boldsymbol{x}_0) \equiv \exp(\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}})$ is approximately log-normal while $\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}}$ is approximately normal.

However, under the proposed approach, Generalised confidence intervals for the regression function $\boldsymbol{x}_0^T\boldsymbol{\beta}$ and $\theta(\boldsymbol{x}_0)$ are available. In fact, under the assumptions and conclusions of Theorems 1 and 2, note that $\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}} - \boldsymbol{x}_0^T\boldsymbol{\beta} - \gamma\sum_{i=1}^n c_i = \sum_{j=1}^n c_j\log(\xi_j)$ with $\xi_j = \theta_j S_j$ independently following Exp(1). For a given dataset $(n, \boldsymbol{x}_0, \widehat{\boldsymbol{\beta}})$, consider pivotal quantity for $\boldsymbol{\beta}$: $\eta = \boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}} - \boldsymbol{x}_0^T\boldsymbol{\beta} - \gamma\sum_{i=1}^n c_i$, which is a log-linear function of independent Exp(1) variables. If η_α denotes the upper α percentile of η, then the values $\eta_{1-\alpha}$ and η_α can be obtained using Monte Carlo simulations, that is, repeatedly generating the values of η m-times via sampling from log-exponential vector $\log\xi$ for the fixed values of $(n, \boldsymbol{x}_0, \widehat{\boldsymbol{\beta}})$. Let $\eta_{1-\alpha}$ and η_α+ are the $1-\alpha$ generalized lower and upper confidence limits for $\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}} - \boldsymbol{x}_0^T\boldsymbol{\beta} - \gamma\sum_{i=1}^n c_i$, respectively. Therefore, the $(1-\alpha)\%$ confidence interval for the regression function $\boldsymbol{x}_0^T\boldsymbol{\beta}$ and $\theta(\boldsymbol{x}_0)$ are given respectively by

$$\left(\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}} - \gamma\sum_{i=1}^n c_i - \eta_{\alpha/2},\ \ \boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}} - \gamma\sum_{i=1}^n c_i - \eta_{1-\alpha/2}\right),$$

and

$$\left(\widehat{\theta}(\boldsymbol{x}_0)\exp(-\gamma\sum_{i=1}^n c_i - \eta_{\alpha/2}),\ \ \widehat{\theta}(\boldsymbol{x}_0)\exp(-\gamma\sum_{i=1}^n c_i - \eta_{1-\alpha/2})\right).$$

Section 5 will illustrate the performance of asymptotic confidence intervals and exact ones numerically.

2.2. Regression Parameter Testing

Theorems 1 and 2 give the estimation of regression parameter and hazard ratio. In this section we discuss hypothesis testing for regression parameter.

In practice we could test a simple regression parameter or a subset of vector $\boldsymbol{\beta}$. Without loss of generality, consider a single regression parameter test: to test if the kth regression coefficient β_k $(k = 0, 1, \cdots, p)$ equal to a known value β_{k0}.

$$H_0 : \beta_k = \beta_{k0}\ \ vs\ \ H_0 : \beta_k \neq \beta_{k0}.$$

Note that

$$\beta_k = (0, \cdots, 0, 1, 0, \cdots, 0)^T\boldsymbol{\beta},$$

and let $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$ except kth component $\beta_k = \beta_{k0}$.

Let $\boldsymbol{x}_k = (0, \cdots, 0, 1, 0, \cdots, 0)^T$, then the test is equivalent to

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0\ \ vs\ \ H_0 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0.$$

Consider the test statistic $\overrightarrow{T}$ of the from:

$$\overrightarrow{T} = \boldsymbol{x}_k^T\widehat{\boldsymbol{\beta}} - \boldsymbol{x}_k^T\boldsymbol{\beta}_0,$$

For $i = 1, \cdots, n$, let $h_i = -\boldsymbol{x}_k^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\mathbf{e_i}$. Note that, when H_0 is true,

$$\overrightarrow{T} = \gamma\sum_{i=1}^{n} h_i + \sum_{i=1}^{n} h_i \log \xi_i,$$

where ξ_i are independent Exp(1) variables.

For a given dataset $(n, \boldsymbol{X}, \boldsymbol{U})$, consider test statistic: $\eta = \overrightarrow{T} - \gamma\sum_{i=1}^{n} h_i$. If η_α denotes the upper α percentile of η, then given a significant level $0 < \alpha < 1$, reject H_0 when $\eta > \eta_{\alpha/2}$ or $\eta < \eta_{1-\alpha/2}$.

The values η_α can be obtained using Monte Carlo simulations, that is, repeatedly generating the values of η m-times via sampling from log-exponential vector $\log(\xi)$ for the fixed values of $(n, \boldsymbol{X}, \boldsymbol{U})$.

3. Estimation of G-Parameter λ

When λ or G is unknown, whatever the method to obtain λ or the survival function $\overline{G}$ of G, as long as $\overline{G}$ then data $\mathbf{U}$ are available, Theorems 1 and can be applied to estimate regression parameter $\boldsymbol{\beta}$ and hazard ratio HR respectively. Below provides a method to estimate both $\boldsymbol{\beta}$ and λ simultaneously.

Let $V(\lambda, \boldsymbol{\beta}) \equiv 2\sum_{i=1}^{n} \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\left(-\log(\overline{G}(Y_i;\lambda))\right)$, then conditional on $\boldsymbol{x}$,

(1) $V(\lambda, \boldsymbol{\beta}) \sim \chi^2(2n)$ and

(2) $V(\lambda, \boldsymbol{\beta})$ is a monotone function of λ.

Because, conditional on $\boldsymbol{x}$,

(1) $2\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\left(-\log(\overline{G}(Y_i;\lambda))\right) = 2\theta_i S_i \sim \chi^2(2)$ from Section 2 and

(2) $\frac{\partial V(\lambda,\boldsymbol{\beta})}{\partial\lambda} = 2\sum_{i=1}^{n} \theta_i \frac{g'(\lambda)}{\overline{G}(Y_i;\lambda)}$ and assume $g'(\lambda) = \frac{\partial g(Y;\lambda)}{\partial\lambda} > 0$.

Therefore, we may combine Theorem 1 and

$$V(\lambda, \boldsymbol{\beta}) = 2(n-1) \tag{6}$$

to inference both $\boldsymbol{\beta}$ and λ simultaneously. The numerical studies in Section 5 provide the details of an algorithm for obtaining $\hat{\lambda}$ and $\widehat{\boldsymbol{\beta}}$.

At the same time, the conditional interval estimate of λ can be given by

$$\left(V_\lambda^{-1}(\chi^2_{\alpha/2}(2n)),\ \ V_\lambda^{-1}(\chi^2_{1-\alpha/2}(2n))\right).$$

4. Regression Variable Selection Via Adaptive Lasso

If variable selection is required for the proposed Gt-PH model, we outline that any modern Lasso-type of estimate for ordinary linear regression can be implemented here straightaway. We use the adaptive lasso [11] to outline variable selection in the regression model.

The lasso estimates for a regression model $E(\boldsymbol{y}) = \boldsymbol{x}^T\boldsymbol{\beta}$ are defined as

$$\hat{\boldsymbol{\beta}}(lasso) = \text{argmin}_{\boldsymbol{\beta}}||\boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{x}_j\beta_j||^2 + \delta\sum_{j=1}^{p}|\beta_j|, \tag{7}$$

where δ is a nonnegative regularization parameter. The second term in (7) is the so-called L_1 penalty, which is crucial for the success of the lasso.

An ideal lasso procedure should be able to identify the true model with probability one, and provide consistent and efficient estimators for the relevant regression coefficients. We use a convex adaptive lasso penalty to illustrate its suitability of our regression model, but many penalties can be applied with regression model (2). This adaptive penalty adapts each coefficient with a weight to reflect the importance of the corresponding covariate, which is equivalent to using different tuning parameters for different coefficients. The coefficients of unimportant covariates are assigned larger weights so that they can be shrunk to zero more easily, leading to the oracle property [11].

When $\boldsymbol{\beta}$ in regression function (2) satisfies regression model (5), an adaptive lasso for estimating $\boldsymbol{\beta}$ could be derived via

$$\hat{\boldsymbol{\beta}}(adapt_{lasso}) = \text{argmin}_{\boldsymbol{\beta}}||\boldsymbol{U} - \sum_{j=0}^{p} \boldsymbol{x}_j\beta_j||^2 + \delta_n\sum_{j=0}^{p}w_j|\beta_j|, \tag{8}$$

where $\mathbf{w} = (w_0, .., w_p)^T$ is a known weights vector. If the weights are data-dependent such as $w_j = 1/|\hat{\beta}_j^{(initial)}|$ with an initial estimator $\hat{\beta}_j^{(initial)}$, then the weighted lasso can have the oracle properties. The reciprocal of any consistent estimator of $\boldsymbol{\beta}$ can be used as the adapting weights; here we may suggest the maximum likelihood estimator of $\boldsymbol{\beta}$.

That is, let $\mathcal{A} = \{j : \beta_j \neq 0\}$ and assume that $|\mathcal{A}| = q < p$, then the true regression model depends only on a subset of $\boldsymbol{x}$. According to Theorem 1 that the $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ has zero bias and variance $\psi'(1)(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n})^{-1}$, so that $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_p(1)$ when $n \to \infty$. According to the Theorem 2 in [11] we have

Theorem 3. *suppose that $\delta_n = o(\sqrt{n})$ and $\delta_n n^{(\nu-1)/2} \to \infty$, then*

(i) $\widehat{\boldsymbol{\beta}}(adapt_{lasso})$ can identify the right subset model $\mathcal{A}$.

(ii) $\widehat{\boldsymbol{\beta}}(adapt_{lasso})$ has the optimal estimation rate,

$$\sqrt{n}\Big(\widehat{\boldsymbol{\beta}}(adapt_{lasso}) - \boldsymbol{\beta}(adapt_{lasso})\Big) \to N(0,\ \Sigma),$$

where $\Sigma = \psi'(1)(\frac{\boldsymbol{X}_{\mathcal{A}}^T\boldsymbol{X}_{\mathcal{A}}}{n})^{-1}$ with $\boldsymbol{X}_{\mathcal{A}}$ derived from a sub-matrix of $\boldsymbol{X}$ which corresponds to the true subset model. Clearly, $\boldsymbol{X}_{\mathcal{A}}^T\boldsymbol{X}_{\mathcal{A}}$ is a $q \times q$ matrix.

Finally, existing algorithms and software for adaptive lasso such as the R-package Package "parcor" [12] can be implemented for the proposed Gt-PH model straightaway.

5. Numerical Analysis

In this section we first carry out some numerical analysis to illustrate the performance of the proposed Gt-PH regression method and make some comparison with MLE-based inference, which we focus on finite sample performance of estimators of regression parameter $\boldsymbol{\beta}$ and G-parameter λ. Then we apply the proposed Gt-PH model to a real data analysis which models the effect of age, gender and body mass index (BMI) on the length of stay (LOS) of heart attach patients.

Our algorithm for estimating λ and $\boldsymbol{\beta}$ in the proposed Gt-PH model is in Algorithm 1:

Algorithm 1: Algorithm for estimating λ and $\boldsymbol{\beta}$ in the proposed Gt-PH model.

```
  (I) Given data $(\boldsymbol{x}, \boldsymbol{Y})$, use $\boldsymbol{Y}$ only to fit $F(y; \lambda, \theta)$ by a method such as MLE with R-function *fitdistr* to obtain an initial estimate of $\lambda$;
 (II) known $\lambda$, obtain observed vector $\boldsymbol{U}$ from $U_i = -\log(\overline{G}(Y_i; \lambda) - \gamma$;
(III) obtain the estimators of $\boldsymbol{\beta}$ via a linear regression model (5) with data $(\boldsymbol{x}, \boldsymbol{U})$ with R-function *lm*;
 (IV) plug the estimators of $\boldsymbol{\beta}$ into (III) and estimate $\lambda$ via Equation (6): $V(\lambda, \boldsymbol{\beta}) = 2(n-1)$;
  (V) repeat steps (II), III) and (IV) until convergence.
```

In contrast, given data $(\boldsymbol{x}, \boldsymbol{Y})$ the MLE-based algorithm to fit regression model (2) is based on the log-likelihood function of $(\boldsymbol{\beta}, \lambda)$, which is given by

$$l(\boldsymbol{\beta}, \lambda) \propto \sum_{i=1}^{n} \sum_{j=1}^{q} \beta_j x_{ij} + \sum_{i=1}^{n} \log g(Y_i; \lambda) + \sum_{i=1}^{n} \sum_{j=1}^{q} (\beta_j x_{ij}) \log(\bar{G}(Y_i; \lambda)).$$

Then MLEs of $\boldsymbol{\beta}$ and λ can be derived via the partial derivatives of $l(\boldsymbol{\beta}, \lambda)$.

For example, consider a simple linear regression model via parameter θ: $\theta = \exp(\beta_0 + \beta_1 x)$ depends on x, for a Weibull distribution with $Y \sim F(y; \lambda, \theta) = 1 - \exp(-\theta y^{\lambda})$, then $G(y; \lambda) = 1 - \exp(-y^{\lambda})$ and the (conditional) likelihood function is given by

$$L((\boldsymbol{x}, \boldsymbol{y}); \beta_0, \beta_1, \lambda) = \lambda^n \prod_{i=1}^{n} \theta_i Y_i^{\lambda - 1} \exp(-\theta_i Y_i^{\lambda}),$$

and then the log-likelihood function is given by

$$\begin{aligned} l(\beta_0, \beta_1, \lambda) &= n \log(\lambda) + \sum_i \log(\theta_i) + (\lambda - 1) \sum_i \log(Y_i) - \sum_i \theta_i Y_i^{\lambda}. \\ &= n \log(\lambda) + n\beta_0 + \beta_1 \sum_i x_i + (\lambda - 1) \sum_i \log Y_i + \exp(-\beta_0) \sum_i \exp(-\beta_1 x_i) Y_i^{\lambda}. \end{aligned}$$

Then the MLEs for $(\lambda, \beta_0, \beta_1)$ satisfy

$$\begin{aligned} 0 &= \lambda^2 \exp(-\beta_0)\frac{1}{n}\sum_i \exp(-\beta_1 x_i)Y_i^{\lambda-1} + \lambda\frac{1}{n}\sum_i \log(Y_i) + 1 \\ \beta_0 &= \log(\frac{1}{n}\sum_i \exp(-\beta_1 x_i)Y_i^{\lambda}) \\ \exp(\beta_0)\sum_i x_i &= \sum_i x_i Y_i^{\lambda}\exp(-\beta_1 x_i). \end{aligned} \tag{9}$$

Now let the true values $(\beta_0, \beta_1, \lambda) = (-1,\ -1,\ 0.5)$. We assess the performance of the proposed Gt-PH algorithm and MLE via the experiment with three different sample sizes: $n = 20, 50, 100$ for data $(x_i,\ y_i)$ $(i = 1, \cdots, n)$. Table 1 summaries the biases and mean square error (MSE)s of each parameter estimator from both Gt-PH model and MLE method under 2000 times of replications.

Table 1. The biases and mean square error (MSE)s of the estimators of the parameters $(\lambda, \beta_0, \beta_1)$.

n	**Parameter**	**Method**	**Bias**	**MSE**
20	λ	Gt-PH	–0.0081	0.0905
		MLE	0.0522	0.1083
	β_0	Gt-PH	0.0036	0.0904
		MLE	0.0152	0.0989
	β_1	Gt-PH	–0.0014	0.0989
		MLE	0.01632	0.1231
50	λ	Gt-PH	–0.0032	0.0712
		MLE	0.0374	0.0923
	β_0	Gt-PH	0.0020	0.0336
		MLE	–0.0145	0.0892
	β_1	Gt-PH	–0.0054	0.0359
		MLE	0.0095	0.0892
100	λ	Gt-PH	0.0018	0.0523
		MLE	0.0075	0.0801
	β_0	Gt-PH	0.0001	0.0173
		MLE	–0.0023	0.0154
	β_1	Gt-PH	0.0001	0.0174
		MLE	0.0015	0.0452

Clearly, the Gt-PH model is premising for sample size less than 100, but it's does not always outperform over MLE for sample size equal 100 or more than 100.

We have also checked the performance of new method with other values of $\lambda = 1, 1.5$. It seem that the G-parameter has little impact on regression estimation. That is, given the regression model (2), selection of $\lambda = 1$ or $\lambda > 1$ or $\lambda < 1$ has very little impact on the estimate of $\boldsymbol{\beta}$.

In terms of confidence interval for regression function and hazard function, under the liner regression $\log(\theta(x)) = \beta_0 + \beta_1 x$, we have $\boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x} = \frac{\bar{X}^2 - 2x\bar{X} + x^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$. Figure 1 at the bottom of this paper plots the 95% confidence intervals for both fitted regression line and hazard function. Clearly, generalised confidence intervals have good coverage properties and much shorter interval lengths than approximate intervals.

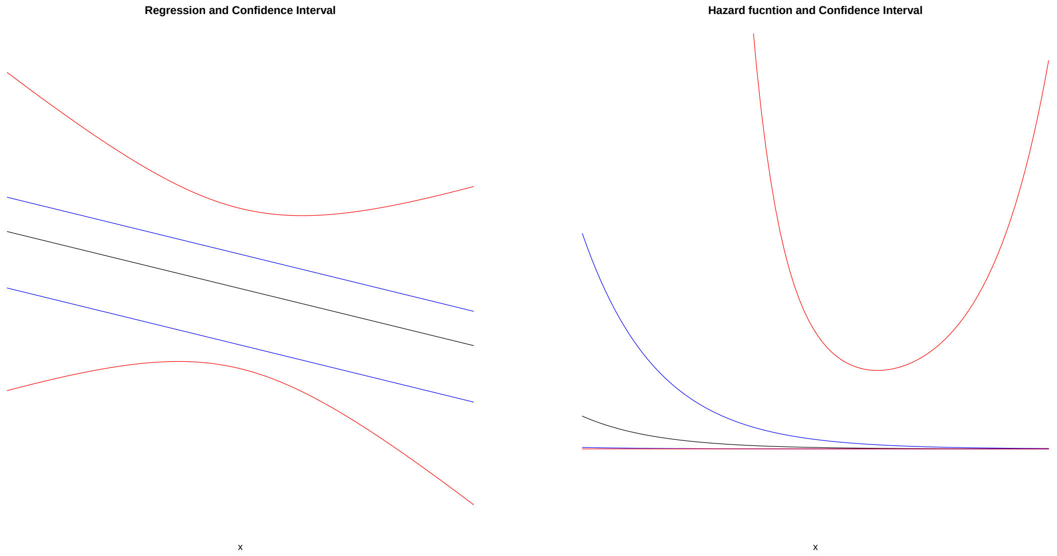

Figure 1. (**Left**): Fitted regression line (black line) for $\log\theta(x) = -1 - x$ and its 95% generalised confidence intervals by the proposed method (blue lines) and normal approximate method (red lines). (**Right**): Fitted hazard function $\theta(\boldsymbol{x})$ (black curve) and its 95% generalised confidence intervals by the proposed method (blue curves) and normal approximate method (red curves).

LOS of Worcester Heart Attack Study

Based on the Worcester Heart Attack Study [13] we aim to investigate how the age (years), gender (female $= 1$ and male $= 0$) and BMI affect LOS. As the distribution of LOS is typically skewed, we use a Weibull distribution to modeling LOS and then fit a regression of LOS over gender, age, interaction of gender and age as well as BMI via the proposed Gt-PH regression, based on the data WHAS100 Data [13] whose size is 100.

The regression model (2) for this special case and aim is introduced as

$$\log(\theta) = \beta_0 + \beta_1 Gender + \beta_2 Age + \beta_3 Age \times Gender + \beta_4 BMI.$$

We are then able to check and compare the effect of the factors on the distribution of LOS via the estimators of these regression parameters.

According to our algorithm, we first fit a Weibull distribution $F(y;\lambda,\theta) = 1 - \exp(-\theta y^{\lambda})$ to obtain an initial value of $\lambda_0 = 1.421$, then replace the θ by the regression above and start the Gt-PH algorithm. After around 10 times of iteration we observe the convergence and obtain the fitted regression model as

$$\log(\hat{\theta}) = -2.745 + 25.825 \times Gender - 0.104 \times Age - 0.461 \times Age \times Gender - 0.224 \times BMI.$$

and $\lambda = 1.432$. Then we could have many different ways for interpretation of the analysis. For example, we could get proper interpretation of the analysis via the median of LOS: note that the logarithm of the median of LOS under the Weibull assumption and fitted Gt-PH model is given by

$$\begin{aligned} & \log(median) \\ = \; & (-\log(\theta) + \log(2))/\lambda \\ = \; & 2.610 - 18.034 \times Gender + 0.0727 \times Age + 0.322 \times Gender \times Age + 0.156 \times BMI. \end{aligned}$$

Clearly, in terms of logarithm of the median of LOS, female patients stay about 18 days shorter than male patients in hospital, and it increases 0.0727 days when patient age increases 1 year. Finally, the interaction of gener and age as well BMI have positive effect on LOS.

6. Discussion

Regression analysis is one of the most important methods in statistics, which is widely used in almost all science and social science research. The proposed Gt-PH (proportional hazard family-based regression) models and their inference methods in this paper are suitable for not only small data analysis but also big data based variable selection. The method and algorithm are easy to implement and have good interpretation in practice.

The method can also be applied to the proportional reverse hazard family $F_r(y;\lambda,\theta)$ defined in (1) straightway. In fact, if a random variable Y belongs to the family, then $-\log(F_r(Y;\lambda,\theta)) \sim Exp(1)$, so $\theta(-\log(G(\lambda))) \sim Exp(1)$. Let the random variable $S = -\log(G(Y;\lambda))$, then Equation (3) holds.

However, the new method is only suitable for the proportional hazard family and reverse hazard family. Extension to more general family of distribution will be discussed in another paper.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant No. 71490725, 71071087 and 11261048).

Author Contributions

Under the idea and detailed guidance of the corresponding author, the first author carried out all numerical calculations and a first draft. Both authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Nelder, J.A.; Wedderburn, R.W.M. Generalized linear models. *J. R. Stat. Soc. A* **1972**, *135*, 370–384.
2. Carroll, R.J.; Ruppert, D. *Transformationand Weighting in Regression*; Chapman & Hall: London, UK, 1988.
3. Smyth, G.K. Generalized linear models with varying dispersion. *J. R. Stat. Soc. B* **1989**, *51*, 47–60.
4. Smyth, G.K. An efficient algorithm for REML in heteroscedastic regression. *J. Comput. Graph. Stat.* **2002**, *11*, 836–847.
5. Engel, J.; Huele, A.F. A generalized linear modeling approach to robust design. *Technometrics* **1996**, *38*, 365–373.
6. Lee, Y.; Nelder, J.A. Double hierarchical generalized linear models. *Appl. Stat.* **2006**, *55*, 139–185.
7. Western, B.;Bloome, D. Variance function regression for studying inequality. *Sociol. Methodol.* **2009**, *39*, 293–325.
8. Bickel, P.J.; Lehmann, E. Descriptive statistics for nonparametric models. III: dispersion. *Ann. Stat.* **1976**, *4*, 1139–1158.
9. Newey, W.K.; Powell, J.L. Asymmetric Least Squares Estimation and Testing. *Econometrica* **1987**, *55*, 819–847.
10. Marshall, A.W.; Olkin, I. *Life Distributions; Structure of Nonparametric, Semiparametric, and Parametric Families*; Springer: New York, NY, USA, 2007.
11. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429.
12. Package "parcor". Available online: www.cran.r-project.org/web/packages/parcor/parcor.pdf (accessed on 2 June 2015).
13. Hosmer, D.; Lemeshow, S.; May, S. *Applied Survival Analysis: Regression Modeling of Time to Event Data*, 2nd ed.; Wiley: New York, NY, USA, 2008.