Advances in Applied Mathematics 应用数学进展, 2021, 10(7), 2353-2358
Published Online July 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.107246

Hans汉斯

自适应设计下基于COX回归模型的序贯压缩估计研究

鲁海波

新疆师范大学数学科学学院，新疆 乌鲁木齐
Email: andyluhaibo@foxmail.com

收稿日期：2021年6月12日；录用日期：2021年7月1日；发布日期：2021年7月13日

摘　要

在COX回归模型的应用中，我们经常会遇到包含太多变量的数据集，而这些变量中只有少数变量对模型有贡献。因此，在推断过程中估计“无效”变量会浪费大量的样本。在本文中，我们提出一种基于自适应压缩估计的序贯抽样策略来构造“有效”参数的固定长度的置信集，这样在忽略模型中的“无效”变量影响的同时，使用最少样本将模型中的“有效”变量快速地识别出来。最后，在自适应设计下对我们所提出的序贯抽样策略进行数值模拟并且数值模拟达到了预期的效果。

关键词

序贯抽样，样本量，压缩估计，停止法则

Sequential Shrinkage Estimate Based on COX Regression Model under Adaptive Design

Haibo Lu

School of Mathematics Science, Xinjiang Normal University, Urumqi Xinjiang
Email: andyluhaibo@foxmail.com

Received: Jun. 12th, 2021; accepted: Jul. 1st, 2021; published: Jul. 13th, 2021

Abstract

In the applications of COX regression models, we always encounter the data sets which contain too many variables that only a few of them contribute to the model. Therefore, it will waste much more samples to estimate the “non-effective” variables in the inference. In this paper, we use a sequential

文章引用: 鲁海波. 自适应设计下基于 COX 回归模型的序贯压缩估计研究[J]. 应用数学进展, 2021, 10(7): 2353-2358.
DOI: 10.12677/aam.2021.107246

procedure for constructing the fixed size confidence set for the "effective" parameters to the model based on an adaptive shrinkage estimate such that the "effective" coefficients can be efficiently identified with the minimum sample size. Adaptive design is considered for numerical simulation.

Keywords

Sequential Sampling, Sample Size, Shrinkage Estimate, Stopping Rule

1. 引言

COX 比例风险模型[1]是一种常用的处理具有协变量的生存数据分析方法。它被广泛应用于生物医学研究和工程等许多领域，用来评估协变量效应。在本文中，我们设失效时间为 T，条件风险函数为

$$h(t|x)=h_0(t)\mathrm{e}^{\beta' X} \tag{1}$$

其中，β (p 维向量)是回归系数，$h_0(t)$ (非负函数)称为基准风险函数。Cox (1984) [2]和 Fleming (1991) [3]对 COX 比例风险模型做了大量的研究。然而，在生物学、工程学和流行病学等研究中，数据集通常有大量的解释变量，但其中只有少数对模型有贡献。也就是说，在一个 p 维的回归系数中只有 p_0 ($p_0<p$ 且 p_0 未知)个分量是取非零值的。这些取非零值的变量被 Wang 和 Zhang (2013) [4]称为有效变量。目前有很多方法可以用来识别有效变量，如 LASSO [5]和 LARS [6]等。但另外需要关注的问题是，用多少样本才能既识别出有效变量，同时又能使参数估计达到预定的精度。这对于生物学和流行病学等需要考虑抽样成本的研究具有重要的意义。对于线性回归模型，Wang 和 Zhang (2013) [4]提出了一种序贯压缩估计方法来识别有效变量，从而达到参数估计的精度。数值模拟结果表明，与传统的序贯抽样方法相比，序贯压缩估计不仅可以从所有变量中识别出有效变量，而且可以节省大量样本。对于 COX 回归模型，还没有人提出相应的序贯方法。

本文针对 COX 回归模型提出了一种基于自适应压缩估计(*ASE*)来构造有效变量的固定窗宽的置信集的序贯抽样方法，使有效变量能以最小样本量快速识别。本文将在适应性设计(adaptive design)下研究所提出的自适应压缩估计(*ASE*)的大样本性质，同时在适应性设计下通过数值模拟得到了很好的模拟结果。

2. 序贯自适应压缩估计(*ASE*)

2.1. 最大偏似然估计(*MPLE*)

设样本容量为 n，T_i 和 C_i 分别是第 i $(i=1,2,\cdots,n)$ 个随机样本的失效时间和删失时间。假定 p 维向量 $X_i=(X_{i1},X_{i2},\cdots,X_{in})^{\mathrm{T}}$ 为第 i 个样本的协变量并且 T_i 和 C_i 关于 X_i 条件独立。令 $Y_i=\min\{T_i,C_i\}$ 是我们真正能观测到的个体失效时间，$\delta_i=I\{T_i\le C_i\}$ 是第 i 个样本被观测到的示性函数。假设 $t_1<t_2<\cdots<t_n$，R_j 是 t_j 时刻的风险集，即 $R_j=\{i:Y_i\ge t_j\}$。这样我们就可以定义模型(1)的偏似然函数为

$$\prod_{j=1}^{N}\frac{\exp\left(\beta^{\mathrm{T}}X_{(j)}\right)}{\sum_{i\in R_j}\exp\left(\beta^{\mathrm{T}}X_{(i)}\right)}$$

由此可得其对数似然函数为

$$L(\beta)=\sum_{j=1}^{N}\left\{\beta^{\mathrm{T}}X_{(j)}-\log\left[\sum_{i\in R_j}\exp\left(\beta^{\mathrm{T}}X_i\right)\right]\right\}$$

设回归参数 β 的最大偏似然估计(*MPLE*)为 $\tilde{\beta}$，则 $\tilde{\beta}$ 通可以通过解下面的得分方程获得：$U(\beta)=0$，其中

$$U(\beta)=\frac{\partial L(\beta)}{\partial\beta}=\sum_{j=1}^{N}\left\{X_{(j)}-\frac{\sum_{i\in R_j}X_i\exp\left(\beta^{\mathrm{T}}X_i\right)}{\sum_{i\in R_j}\exp\left(\beta^{\mathrm{T}}X_i\right)}\right\}$$

2.2. 自适应压缩估计(*ASE*)

设 $\kappa=\kappa(n)$，当 $n\to\infty$ 时，存在 $0<\delta<1/2$ 和 $\gamma>0$ 使得 $n^{\frac{1}{2}}\kappa\to 0$，$n^{\frac{1}{2}+\gamma\delta}\kappa\to\infty$。本文中我们需要以下假定：

(A1). 解释变量 x_i 满足条件 $\sup_i\|x_i\|<\infty$，并且残差项 $\varepsilon_i=\hat{\Lambda}(Y_i)\exp\left(\hat{\beta}^{\mathrm{T}}X_i\right)$ 具有有限二阶矩，即当 $\zeta>2$ 时，$E|\varepsilon_i|^{\zeta}<\infty$ 其中是 $\hat{\Lambda}$ 累计基准风险函数；

(A2). $\lim_{n\to\infty}I_n(\beta)/n=\Sigma$，其中 $I_n(\beta)$ 是回归参数 β 的信息矩阵，Σ 是一个正定矩阵。

定义 2.2.1 设 $\tilde{\beta}$ 为模型(1)的最大偏似然估计，则称 $\hat{\beta}_n=In(\varepsilon)\tilde{\beta}_n$ 为回归系数 β_0 的自适应压缩估计(ASE)，其中 $I_n(\varepsilon)=diag\{I_{n1}(\varepsilon),I_{n2}(\varepsilon),\cdots,I_{np}(\varepsilon)\}$ 是一个 $p\times p$ 维对角阵。同时可以证明 $\hat{\beta}_n=In(\varepsilon)\tilde{\beta}_n$ 满足相合性和渐进正态性。

2.3. 序贯抽样策略

由文献[7]，[8]的结论可以证明 $\sqrt{n}\left(\hat{\beta}_n-\beta_0\right)$, $n=1,2,\cdots$ 是依概率一致连续的，由此可得如下定理：

定理 2.3.1 设随机变量 *N(t)*取正整数值，当 $t\to\infty$ 有 $N(t)/t$ 依概率收敛于 1，且条件(A1)和(A2)成立，则当 $t\to\infty$ 时，

$$\sqrt{N(t)}\left(\hat{\beta}_{N(t)}-\beta_0\right)\to N\left(0,I_0\Sigma I_0^{-1}\right)$$

由定理 2.3.1 我们可以构造 β_0 的置信集和能够决定最小样本量的停止法则的序贯抽样策略。设 $\{(y_i,x_i):i=1,2,\cdots,k\}$ 是最先进入研究的 k 个样本，用 C_k 来表示。在任意给定小正数 ε 下，

$$\hat{p}_0(k)=\sum_{j=1}^{p}I_{kj}(\varepsilon)$$

是回归系数 p_0 基于条件 C_k 的估计量。令 $a_k^2\in R$ 对任意 $\alpha>0$，有 $P\left(\chi^2_{\hat{p}_0(k)}\le a_k^2\middle|C_k\right)=1-\alpha$ 成立。现在定义停时法则 N_d 为

$$N=N_d\equiv\inf\left\{k:k\ge n_0\text{ and }\frac{d^2}{a_k^2}\ge\nu_k\right\},\tag{2}$$

其中 ν_k 是 $kI_k(\varepsilon)(\Sigma)^{-1}I_k(\varepsilon)$ 的最大特征值，d 是置信集的预设精度。在本文的序贯估计策略中，一次只有一个新的观测进入研究直到满足(2)式的停止法则时就停止抽样，此时 β_0 的置信集为

$$R_N=\left\{Z\in R^p:\frac{S_N}{N}\le\frac{d^2}{\nu_N}\text{且当}I_{N_j}(\varepsilon)=0\text{时, }z_j=0,1\le j\le p\right\}\tag{3}$$

其中 $S_N=\left(Z_{N_1}-\hat{\beta}_{N_1}\right)^{\mathrm{T}}\tilde{\Sigma}_{11}\left(Z_{N_1}-\hat{\beta}_{N_1}\right)$。我们所提出的序贯抽样方法致力于找到有效变量的同时忽略无效变

量的影响，这是和传统序贯方法相比我们能够节省大量样本的关键，在下面的定理中我们给出停时 N_d 和置信集 R_N 的相关性质。

定理 2.3. 2 假定条件(A1)和(A2)都成立，设 N 是满足(2)式的停时，则：

(i). $\lim_{d\to 0}\frac{d^2N}{a^2\nu}=1$，a.s.；(ii). $\lim_{d\to 0}P\left(\beta_0\in R_N\right)=1-\alpha$；

(iii). $\lim_{d\to 0}\frac{d^2E(N)}{a^2\nu}=1$；　(iv). $\lim_{d\to 0}\hat{p}_0(N)=p_0$，a.s.且 $\lim_{d\to 0}E\left(\hat{p}_0(N)\right)=p_0$，

其中 ν 是矩阵 $I_0\Sigma^{-1}I_0$ 的最大特征值。

3. 数值模拟

我们在固定样本量下用所提方法对随机数据集合进行分析，以此来验证所提出的序贯压缩估计方法的性能。按照停止法则的定义，当抽样停止时，最终的置信集将满足预设精度和覆盖概率，因此我们可以比较分别基于 MPLE 和 ASE 的序贯抽样方法的平均停时。由于序贯压缩估计方法忽略无效变量的影响，故理论上平均所需停时应该显著小于不考虑变量选择的序贯方法。如果事先已知有效变量为 p_0 个同时无无效变量，那么只使用这 p_0 个有效变量的序贯方法无疑是效率最高的。所以，为便于比较，我们将所有(p_0 个)变量全部为有效变量的序贯估计方法作为基准线，在此情况下所获得的样本量应该是最小的。

在自适应设计下，随机模拟数据集中的 x_1 仍然由多元标准正态分布生成，$x_j\,(j>1)$ 由均值为 $\sum_{i=1}^{j-1}\left[x_j/(j-1)\right]$，方差协方差矩阵为单位阵的多元正态分布生成。不失一般性，选择模型(1)中的基准风险函数为 $h_0(t)=t^2$。回归系数真值取 $(-1.2,2.0,0,0,0,0,0,0,0,0)$，其中含有八个无效变量，回归系数置信集的预设精度 $d\in\{0.3,0.4,0.5,0.6\}$，取 $\alpha=0.05$，$\gamma=1$，$\delta=0.45$，$\theta=0.75$。另外当用 ASE 方法时我们用 BIC 方法来确定 ε，

$$BIC=-2\left(\sum_{j=1}^{N}\left(\beta^{\mathrm{T}}X_{(j)}-\log\left(\sum_{j\in R_j}\exp\left(\beta^{\mathrm{T}}X_i\right)\right)\right)\right)+\log(n)\times df/n$$

其中 df 表示 β 中非零分量的个数。

Table 1. Results of sequential sampling method based on ASE, MPLE with all variables and MPLE with only p_0 non-zero variables for COX regression model

表 1. COX 回归模型下分别应用 ASE，MPLE 和 MPLE_{p_0} 的序贯抽样方法的结果分析

$\beta=(-1.2,\ 2.0,\ 0.0,\ 0.0,\ 0.0,\ 0.0,\ 0.0,\ 0.0,\ 0.0,\ 0.0)$										
		MPLE_{p_0}			ASE			MPLE		
Design	d	N	κ^*	CP^+	N	κ	CP^+	N	κ	CP
Adaptive	0.6	92.68 (16.36)	1.019	0.95	106.795 (18.074)	1.040	0.935	272.71 (30.106)	1.006	0.95
	0.5	124.82 (17.189)	1.013	0.96	144.315 (23.01)	1.031	0.92	373.8 (33.836)	1.005	0.93
	0.4	182.2 (23.139)	1.01	0.95	204.6 (27.90)	1.019	0.935	558.85 (43.275)	1.003	0.93
	0.3	319.97 (31.557)	1.005	0.95	342.98 (39.208)	1.016	0.93	960.97 (55.86)	1.002	0.97

$\kappa^*=d^2N/(a^2\nu)$；CP^+ 是 95%置信集 R_N 的经验覆盖概率；**经验标准差在括号内。

表 1 描述了 COX 回归模型下的序贯抽样方法的数值模拟结果。在表 1 中我们列出了最终样本量 N (停时)，$\kappa^*=d^2N/(a^2\nu)$ 和 95%置信集的经验覆盖概率 R_N。所有三种情况(MPLE_{p_0}, ASE, MPLE)下的 κ 值都

非常接近 1，并且当 d 不断减小时经验覆盖概率 CP 越来越接近 95%，正如定理 2.3.2 描述的一样。然而，应用 *MPLE* 方法所得的样本量 N 比应用 ASE 方法和 MPLE_{p_0} 都大得多。而应用 *ASE* 的抽样策略所需的样本量和应用 MPLE_{p_0} 的抽样策略所需样本量差不多，这说明我们所提方法在变量选择的同时效率和回归参数中只有有效变量无无效变量的情况下的效率非常接近，而比不做变量选择情况下(即 *MPLE*)的抽样效率提高很多。

表 2 比较了在估计 COX 回归模型的回归系数时分别应用 *ASE* 和 *MPLE* 的抽样策略对识别回归系数中的有效变量和无效变量的效率。从结果可以看出应用 *ASE* 的抽样策略时不能被正确识别的零变量的平均个数几乎趋向于 0，而能被正确识别的非 0 变量的平均个数和模型中有效变量个数的真值非常接近(2 和 8)。结果表明基于 *ASE* 的序贯抽样策略下 $\hat{p}_0$ 是 p_0 的优良估计。而基 MPLE 的序贯抽样策略不能识别有效变量，因此无法获得 N_c^* 和 N_{ic}^* 的值。此外，所有参数的估计值和它们的真值都非常接近。

Table 2. Power of variable identification and estimation of nonzero components under sequential sampling method based on ASE and MPLE with COX regression model
表 2. COX 回归模型下分别应用 ASE 和 MPLE 的序贯抽样策略的变量识别和非零参数估计效率

		$\beta_1=-1.2,\ \beta_2=2.0$						
		ASE				*MPLE*		
Design	d	N_{ic}^*	N_c^*	β_1	β_2	N_{ic}^* N_c^*	β_1	β_2
Adaptive	0.6	0	7.86	−1.258 (0.16)	2.097 (0.183)	-	−1.228 (0.092)	2.074 (0.117)
	0.5	0	7.89	−1.251 (0.129)	2.104 (0.153)	-	−1.226 (0.077)	2.031 (0.095)
	0.4	0	7.97	−1.231 (0.111)	2.061 (0.15)	-	−1.213 (0.069)	2.021 (0.079)
	0.3	0	7.965	−1.216 (0.077)	2.043 (0.096)	-	−1.208 (0.044)	2.01 (0.068)

N_{ic}^*：β 中零分量(无效变量)被错误识别的平均个数；N_c^*：β 中非零分量(有效变量)被正确识别的平均个数。

4. 结论

在 COX 回归模型下基于自适应压缩估计(*ASE*)建立的序贯抽样方法不仅能够用最少的样本识别出回归参数中的有效变量，同时可以使回归参数的估计值达到预设的精度。数值模拟结果表明和传统的序贯抽样方法相比，我们提出的方法能够节省大量样本。然而，本文中所提方法涉及到的变量维数是固定的，后期我们将研究当变量维数随样本量变化时的序贯抽样方法的相关性质。

基金项目

1) 新疆师范大学博士科研启动基金项目：“基于广义线性模型的序贯分析研究”XJNUBS1539。

2) 新疆维吾尔自治区高校科研计划项目：“基于 COX 比例风险回归模型的序贯分析研究”(XJEDU2016I033)。

参考文献

[1] Cox, D.R. (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B*, **34**, 187-220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

[2] Cox, D.R. and Oakes, D. (1984) Analysis of Survival Data. Chapman and Hall, London.

[3] Fleming, T. and Harrington, D. (1991) Counting Processes and Survival Analysis. Wiley, New York.

[4] Wang, Z.F. and Chang Y.-C.I. (2013) Sequential Estimate for Linear Regression Models with Uncertain Number of Effective Variables. *Metrika*, **76**, 949-978. https://doi.org/10.1007/s00184-012-0426-4

[5] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, *Series B*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[6] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least Angle Regression. *Journal of Annals of Statistics*, **32**, 407-499. https://doi.org/10.1214/009053604000000067

[7] Anscombe, F.J. (1952) Large Sample Theory of Sequential Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, **48**, 600-607. https://doi.org/10.1017/S0305004100076386

[8] Woodroofe, M. (1982) Nonlinear Renewal Theory in Sequential Analysis. Society for Industrial and Applied Mathematics, Philadelphia. https://doi.org/10.1137/1.9781611970302