International Journal of Statistics and Probability; Vol. 2, No. 3; 2013
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Discriminating Among Several Semiparametric Models

M. M. E. Abd El-Monsef[1] & M. M. Seyam[1]

[1] Faculty of Science, Tanta University, Egypt

Correspondence: M. M. Seyam, Faculty of Science, Tanta University, Egypt. E-mail: m.seyam@science.tanta.edu.eg

Received: April 16, 2013 Accepted: June 19, 2013 Online Published: July 15, 2013

doi:10.5539/ijsp.v2n3p96 URL: http://dx.doi.org/10.5539/ijsp.v2n3p96

Abstract

To distinguish between two or more than two models one can use the T-optimality criterion. Another criterion using for discrimination between two or more than two models is KL-criterion, which depend on the Kullback-Leibler distance. KL-criterion can be used to discriminate between two non-normal models and a generalized of the KL-criterion was studied to discriminate more than two non-normal models. In this paper, more than two semiparametric models can be distinguished using generalized KL-criterion. An application was applied to illustrate the proposed technique by using three proportional hazard models via real data.

Keywords: optimal experimental design, semiparametric model, Kullback-Leibler distance, KL-optimality, T-oprimality

1. Introduction

Optimal designs are experimental designs that are generated, based on a optimality criterion and are generally optimal only for a specified statistical model. An optimality criterion showed how good a design is, based on some mathematical properties. One of these optimality criteria is T-optimality, which was proposed by Atkinson and Fedorov (1975a, 1975b). This criterion is used to distinguish between two or more than two models with normal errors. Ponce de Leon and Atkinson (1992) proposed a generalized T-optimality between two generalized linear models, which called generalized T-optimality criterion. Uciński and Bogacka (2005) introduced a generalization of this criterion for multi response models. A generalized T-optimality composed of maximizing the deviance from the model 2 when data are generated by model 1.

Recently, López-Fidalgo et al. (2005, 2007) extended the conventional T-optimality criterion, to handle any distribution for the random errors and introduced a new criterion depend on the Kullback-Leibler divergence, called KL-optimality criterion. A design which maximizes this criterion is called KL-optimal design. One of the most applicable distance for statistical distributions is Kullback-Liebler distance is proposed see, Burnham and Anderson (1998). The KL-criterion function includes the T-optimality criterion as a special case and is applicable to any parametric regression models. López-Fidalgo et al. (2007) applied KL-optimality criterion under non-normal distributions, as the lognormal and gamma distributions. When the discrimination between two binary response models then the KL-criterion and generalized T-criterion are identical, see López-Fidalgo et al. (2007). Otsu (2008) proposed the KL-optimal criterion by using López-Fidalgo et al. (2007) to a semiparametric setup to discriminate two regression models. Tomasi (2007) used a generalized KL-criterion to discriminate more than two non-normal models.

In this paper, more than two semiparametric models can be discriminated using generalized KL-criterion. In Section 2, Cox's proportional hazard model is introduced. In Section 3, a generalized KL-criterion for discriminating among several Cox models is considered. In Section 4, a real data is illustrated where three Cox-proportional hazards models are given. A conclusion is proposed in Section 5.

2. Cox's Proportional Hazards Model

The proportional hazard model was firstly introduced by Cox (1972), and this is the most common model in biostatistics. The advantages using this model are:

- The hazard ratio is an essay constant.
- The Cox model avoids making assumptions about the hazard.

Proportional hazards models are considered by Becker et al. (1989) who find D-optimal designs for models with one or two parameters and completely specified baseline hazard. They use geometric arguments and empirical values for the hazard rate to investigate how censoring affects the D-optimal designs for different shapes of the design region.

Survival analysis is a collection of statistical techniques used to examine and model the time it takes for events to occur. In survival analysis, when the event occur we use the term failure and survival time is the time taken for event failure to occur.

The Cox proportional hazards model is a semiparametric model which is given by:

$$h_i(t) = h_0(t)\exp\left(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}\right)$$

where

β_i's are the parameters;

t is the time;

$h_0(t)$ is the baseline hazard function;

x_i's are covariates.

If all of the x's are zero the exponential part of the previous equation equals 1, $h_i(t) = h_o(t)$, so $h_o(t)$ is called the baseline hazard function (when predictor variables all have a value of zero). Even though the baseline hazard function is unspecified, it is still possible to estimate the parameter estimates in the exponential part of the model. Cox (1972) showed how to derive a valid parameter estimate that does not require the estimate of the baseline hazard function.

The hazard function is the probability that an individual will experience an event (for example, death) within a small time interval, given that the individual has survived up to the beginning of that interval. It can therefore be interpreted as the risk of dying at time t. If the hazard function does not depend on time and its value is completely determined by the covariate and the unknown parameters, it means that the risk of failure is the same no matter how long the subject has been followed. The hazard function, denoted by $h(t)$, can be proposed as follows:

$$h(t) = \frac{number\ of\ individuals\ experiencing\ an\ event\ in\ interval\ beginning\ at\ t}{(number\ of\ individuals\ surviving\ at\ time\ t) \times (interval\ width)}$$

Assumptions of the Cox model are as follows:

- The ratio of the hazard function does not depend on time.
- Time is measured on a continuous scale.

There are three different tests to assess the significance of the coefficients: the partial likelihood ratio test, the score test, and Wald test.

3. Generalized KL-Criterion for Discriminating Among Several Cox Models

A statistical model is a collection of probability distribution functions or probability density functions. Let the statistical model can be written as $f_i(y, x, \theta_i)$, $i = 1, \ldots, k$ where y is the dependent variable, x is a vector of experimental conditions and $\theta_i \in \Omega_i \subset \mathbb{R}^{m_i}$ is the unknown parameter vector.

In order to discriminate these k rival models; an prolonged model which includes them is considered by Atkinson and Cox (1974). The k models are entrenched in a more general model, $f_{k+1}(y, x, \theta_{k+1})$. KL-criterion is used to discriminate between the i-th model and $f_{k+1}(y, x, \theta_{k+1})$. In this paper, the parameters of the extended model are supposed to be known.

The i-th KL–optimality criterion function is

$$I_{i,k+1}(\xi) = \min_{\theta_i \in \Omega_i} \int_{\chi} I\left[f_{k+1}(y, x, \theta_{k+1}),\ f_i(y, x, \theta_i)\right] \xi(dx), \tag{1}$$

where

$$I\left[f_{k+1}(y, x, \theta_{k+1}),\ f_i(y, x, \theta_i)\right] = \int f_{k+1}(y, x, \theta_{k+1})\ log\left[\frac{f_{k+1}(y, x, \theta_{k+1})}{f_i(y, x, \theta_i)}\right]\ dy$$

is the Kullback–Leibler distance between the true model $f_{k+1}(y, x, \theta_{k+1})$ and the alternative model $f_i(y, x, \theta_i)$.

If ξ is any design, the efficiency of ξ is the ratio of the criterion function (1) at ξ to its maximum value, i.e.

$$Eff_{i,\,k+1}(\xi) = \frac{I_{i,k+1}(\xi)}{I_{i,k+1}(\xi_i^*)},\ i = 1, \dots, k$$

where

$$\xi_i^* = \arg\max_{\xi} y_{i,k+1}(\xi)$$

is the KL-optimum design for discriminating model i from the general model.

Suppose that

$$I_\alpha(\xi) = \sum_{i=1}^{k} \alpha_i \cdot\ Eff_{i,k+1}(\xi) \tag{2}$$

be the generalized KL-criterion function which used to compare more than two models, and α is the $k \times 1$ vector of the coefficients α_i, which are such that $0 \le \alpha_i \le 1$ for $i = 1,\ \dots,\ k$ and $\int_{i=1}^{k} \alpha_i = 1$.

The following design

$$\Omega_i(\xi) = \left\{ \widehat{\theta_i} : \widehat{\theta_i}(\xi) = \arg\min_{\theta_i \in \Omega_i} \int_{\mathcal{X}} I\left[f_{k+1}(y, x, \theta_{k+1}),\ f_i(y, x, \theta_i)\right] \xi(dx) \right\},\ i = 1, \dots, k \tag{3}$$

is called a regular design, otherwise it is called singular design.

In this section, we will apply the KL-optimality criterion to discriminate more than two semiparametric models. One of the popular semiparametric models is *Cox proportional hazards model* given by:

$$\begin{aligned} h_i(t) &= h_0(t)\exp(X_1\beta_1 + \cdots + X_i\beta_i), \quad i = 1,\ \dots,\ k \\ &= h_0(t)\exp\left(\textstyle\sum_{i=1}^{k} X_i\beta_i\right) \end{aligned}$$

This model is based on two parts: $h_0(t)$ is called the baseline hazard function and depend on time only and the second part includes the covariates and does not conclude a time variable. So, the ratio of the hazards of two individuals does not depend on time, i.e. $h_0(t)$.

To find KL-optimum design for discriminating model i from the general model we first need to determine the i-th KL-criterion function given by (1).

Where

$$f_{k+1}(y, x, \beta_{k+1}) = \frac{\partial \eta^{k+1}}{\partial x_j} = h_0(t) \prod_{j=1}^{k+1} \beta_j \exp\left(\sum_{j=1}^{k+1} X_j\beta_j\right),\quad j = 1, 2, \dots, k+1$$

In our case, consider the following rival models:

$$\eta_1 = h_0(t)\exp(X_1\beta_1 + X_2\beta_2)$$

$$\eta_2 = h_0(t)\exp(X_2\beta_2 + X_3\beta_3)$$

$$\eta_3 = h_0(t)\exp(X_1\beta_1 + X_3\beta_3)$$

and the combined model

$$\eta_4 = h_0(t)\exp(X_1\beta_1 + X_2\beta_2 + X_3\beta_3)$$

In this paper, the parameters of the prolonged model are supposed to be identified. thus the optimal designs can be determined, so we let $\beta_1 = \beta_2 = \beta_3 = 1$.

$$f_1 = \frac{\partial^2 \eta_1}{\partial x_1 \partial x_2} = h_0(t)\exp(X_1 + X_2)$$

$$f_2 = \frac{\partial^2 \eta_2}{\partial x_2 \partial x_3} = h_0(t)\exp(X_2 + X_3)$$

$$f_3 = \frac{\partial^2 \eta_3}{\partial x_1 \partial x_3} = h_0\,(t) \exp(X_1 + X_3)$$

$$f_4 = \frac{\partial^3 \eta_4}{\partial x_1 \partial x_2 \partial x_3} = h_0\,(t) \exp(X_1 + X_2 + X_3)$$

The criterion function (2) becomes

$$I_\alpha\,(\xi) = \alpha_1 \frac{I_{1,4}(\xi)}{I_{1,4}(\xi_1^*)} + \alpha_2 \frac{I_{2,4}(\xi)}{I_{2,4}(\xi_2^*)} + (1 - \alpha_1 - \alpha_2) \frac{I_{3,4}(\xi)}{I_{3,4}(\xi_3^*)}$$

where the numerator is KL-optimality criterion function and given by:

$$I_{1,4}\,(\xi) = \int\int\int\int h_0\,(t) \exp(X_1 + X_2 + X_3)\; X_3\; dt\; dX_1 dX_2 dX_3$$

$$I_{2,4}\,(\xi) = \int\int\int\int h_0\,(t) \exp(X_1 + X_2 + X_3)\; X_1\; dt\; dX_1 dX_2 dX_3$$

$$I_{3,4}\,(\xi) = \int\int\int\int h_0\,(t) \exp(X_1 + X_2 + X_3)\; X_2\; dt\; dX_1 dX_2 dX_3$$

A design ξ_i^* which maximizing $I_{i,4}(\xi),\; i = 1, 2, 3$ is a KL-optimum design.

According to evaluate a KL-optimum design numerically the Kullback-Leibler used in the expression of the directional derivative.

Atkinson (1970) investigated a method for discriminating between models. It is desired to verify which of several alternative models adequately describe the data, the properties of a combined distribution containing the component models as special cases. Using this distribution, statistics are developed for testing for departures from one model in the direction of another and for testing the hypothesis that all models fit the data equally well.

4. An Application

In this section, a real data taken from Lee and Wang (2003) is applied in order to illustrate the proposed theoretical results. A sample of 200 cardiac patients was collected, and they were asked about some demographic variables then some clinical examinations were recorded. These patients were followed for ten years and the following variables were collected: age, SBP, LACR and LTG. The proportional hazards model used to identify which risk factors is the most important.

The event time of interest is CVD-free time, which is defined as the time in years. The covariates which used in this application are given by: systolic blood pressure (SBP), logarithm of ratio of urinary albumin and creatinine (LACR) and logarithm of triglycerides (LTG).

After computations KL-optimality criterion function becomes:

$$I_\alpha\,(\xi) = \alpha_1 \frac{I_{1,4}(\xi)}{0.42384} + \alpha_2 \frac{I_{2,4}(\xi)}{0.66369} + (1 - \alpha_1 - \alpha_2) \frac{I_{3,4}(\xi)}{0.39291}$$

with corresponding efficiencies 0.51367, 0.92001 and 0.65716, according to these efficiencies we get that the second model is more efficient than the first and third models.

5. Conclusion

In this paper, a generalization of the KL-optimality criterion was introduced to discriminate among several semiparametric models. The main core of the generalized KL-optimality criterion was applied to one of the most important semiparametric models, namely the Cox's proportional hazards model. A real data set was used to illustrate the new theoretical results. The generalized KL-optimality criterion enabled us to discriminate among four different Cox models and select the model with high efficiency.

References

Atkinson, A. C. (1970). A method for discriminating between models. *Journal of the Royal Statistical Society, Series B, 32*, 323-353.

Atkinson, A. C., & Cox, D. R. (1974). Planning experiments for discriminating between models. *Journal of the Royal Statistical Society, Series B, 36*, 321-348.

Atkinson, A. C., & Fedorov, V. V. (1975a). The design of experiments for discriminating between two rival models. *Biometrika, 62*(1), 57-70. http://dx.doi.org/10.1093/biomet/62.1.57

Atkinson, A., & Fedorov, V. V. (1975b). Optimal design: experiments for discriminating between several models. *Biometrika, 62*(2), 289-303. http://dx.doi.org/10.1093/biomet/62.2.289

Becker, N., Mcdonald, B., & Khoo, C. (1989). Optimal designs for fitting a proportional hazards regression model to data subject to censoring. *Australian and New Zealand Journal of Statistics, 31*, 449-468. http://dx.doi.org/10.1111/j.1467-842X.1989.tb00989.x

Burnham, K., & Anderson, D. (1998). *Model selection and inference: a practical information-theoretic approach*. New York: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4757-2917-7

Cox, D. (1972). Regression models and life-tables (with Discussion). *Journal of the Royal Statistical Society, Series B, 34*, 187-220.

Lee, E., & Wang, J. (2003). *Statistical methods for survival data analysis*. Hoboken, New Jersey: John Wiley & Sons, Inc. http://dx.doi.org/10.1002/0471458546

López-Fidalgo, J., Tommasi, C., & Trandafir, P. (2005). Optimal designs for discriminating between heteroscedastic models. In *Proceedings of the 5th St. Petersburg Workshop on Simulation*, 429-436. Saint Petersburg: NII Chemistry Saint Petersburg University Publishers.

López-Fidalgo, J., Tommasi, C., & Trandafir, P. (2007). An optimal experimental design criterion for discriminating between non-normal models. *Journal of the Royal Statistical Society, Series B, 69*, 231-242. http://dx.doi.org/10.1111/j.1467-9868.2007.00586.x

Otsu, T. (2008). Optimal experimental design criterion for discriminating semiparametric models. *Journal of Statistical Planning and Inference, 138*, 4141-4150. http://dx.doi.org/10.1016/j.jspi.2008.03.027

Ponce de Leon, A. , & Atkinson, A. (1992). The design of experiments to discriminate between two rival generalized linear models. In *Lecture Notes in Statistics -Advances in GLM and Statistical Modelling*, 159-164, New York: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4612-2952-0_25

Tommasi, C. (2007). Optimal designs for discriminating among several non-Normal models. In J. López-Fidalgo, J. M. Rodrìguez-Dìaz, & B. Torsney (Eds.), *mODa 8 - Advances in Model-Oriented Design and Analysis* (pp. 213-220). Physica-Verlag. http://dx.doi.org/10.1007/978-3-7908-1952-6_27

Uciński, D., & Bogacka B. (2005). T-optimum designs for discrimination between two multiresponse dynamic models. *Journal of the Royal Statistical Society, Series B, 67*, 3-18. http://dx.doi.org/10.1111/j.1467-9868.2005.00485.x

Copyrights