PROPORTIONAL HAZARDS, TRANSFORMATION MODELS, PARTIAL LIKELIHOOD, THE ORDER BOOTSTRAP, AND ADAPTIVE INFERENCE. I.

BY

KJELL A. DOKSUM

TECHNICAL REPORT NO. 27

DECEMBER 1983

DEPARTMENT OF STATISTICS

UNIVERSITY OF CALIFORNIA, BERKELEY

RESEARCH PARTIALLY SUPPORTED BY

NATIONAL SCIENCE FOUNDATION

GRANT MCS-81-02349

PROPORTIONAL HAZARDS, TRANSFORMATION MODELS, PARTIAL LIKELIHOOD, THE ORDER BOOTSTRAP, AND ADAPTIVE INFERENCE. I.

Kjell A. Doksum*

Abstract

Connections between proportional hazard and transformation models are explored and exploited. It is shown that under certain assumptions, the transformation model is doubly adaptive in the sense that the linear model parameters can be estimated with full efficiency when both the transformation and error distribution are unknown.

This work was supported in part by the National Science Foundation Grant MCS-81-02349.

1. Introduction

(a) Transformation models

The independent random variables $Y_1,\ldots,Y_n$ are said to follow a linear transformation model if for some transformation h,

$$h(Y_i) = \underset{\sim\sim}{\beta} x_i + \sigma\varepsilon_i \ , \quad i = 1,\ldots,n \tag{1}$$

where $\underset{\sim}{x}_i' = (x_{i1},\ldots,x_{ip})$, $i = 1,\ldots,n$, are known constants, $\underset{\sim}{\beta} = (\beta_1,\ldots,\beta_p)$ is a vector of regression parameters and $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. with distribution F. Some common choices of h are $h(y) = y^\lambda$, $h(y) = \text{sign}(y)|y|^\lambda$,

$$h(y) = \frac{y^\lambda - 1}{\lambda} \ , \quad \lambda \neq 0$$
$$= \log y \ , \quad \lambda = 0 \ .$$

See for instance Tukey (1957), Box and Cox (1964), and Bickel and Doksum (1981).

(b) The proportional hazard model as a transformation model

Suppose Y_i is a survival time with distribution F_i and hazard rate $\lambda_i = f_i/[1-F_i]$, $i = 1,\ldots,n$. The Cox proportional hazard model is

$$\lambda_i(t) = \Delta_i\lambda(t), \quad \text{some } \lambda(t), \text{ where } \Delta_i = \exp(\underset{\sim\sim}{\beta} x_i) \ . \tag{2}$$

An equivalent form of this model is the Lehmann (1953) form

$$F_i(t) = 1 - [1 - F_0(t)]^{\Delta_i}, \quad \text{where } F_0(t) = 1 - \exp[-\int_0^t \lambda(x)dx] \ .$$

It follows that

$$\log\{-\log[1 - F_0(Y_i)]\} \overset{d}{=} -\underset{\sim\sim}{\beta} x_i + \varepsilon_i \tag{3}$$

where the $\{\varepsilon_i\}$ are i.i.d. with distribution $1 - \exp(-e^t)$ and $\overset{d}{=}$ means "distributed as." In other words, the proportional hazard model is a transformation model where after the transformation

$$h(Y_i, F_0) = \log\{-\log[1 - F_0(Y_i)]\}$$

with parameter F_0, the transformed data follows a parametric linear model. Estimation of $\underset{\sim}{\beta}$ is achieved by maximizing a partial (Cox (1972), (1975)) or marginal (Kalbfleisch and Prentice (1973)) likelihood which does not depend on F_0 (or h). Once an estimate is available for $\underset{\sim}{\beta}$, an estimate of F_0 can be obtained.

(c) Partial likelihoods

The partial likelihood idea can be applied to the transformation model (1) with h increasing but otherwise unknown, and F continuous. As in the proportional hazard model without censored data, the partial likelihood for $\underset{\sim}{\beta}$ is equivalent to the likelihood of the rank vector $(R_1,\ldots,R_n)$ where $R_i = \text{Rank}(Y_i) = \text{Rank}(h(Y_i))$. Thus we estimate $\underset{\sim}{\beta}$ by maximizing the likelihood

$$L(\underset{\sim}{\beta}) = P(R_1=r_1,\ldots,R_n=r_n) .$$

The case where F is standard normal (Box and Cox (1964)) and σ is known, but h is unknown will be considered first. Later, the cases F not normal but known, and F unknown will be considered, as well as the problem of estimating h and σ^2.

2. Partial likelihood for a nonparametric Box-Cox model

We consider the model (1) with F standard normal and with h an arbitrary function increasing on the range of the Y_i. It is clear

that since

$$\text{Rank}(h(Y_i)) = \text{Rank}(a + bh(Y_i)) \quad \text{for all } a, \text{ and } b > 0$$

we cannot use the partial likelihood to estimate the location of $h(Y_i)$ or the scale of $h(Y_i)$. When applying the partial likelihood (Cox (1972, 1975)) to the proportional hazard model (2), the same problem occurs. In this model, the intercept parameter is set equal to zero. Moreover $\text{Var}(h(Y_i,F_0)) = \text{Var}(\varepsilon_i)$ in (3) is necessarily equal to one. We will do the same, that is, in model (1) we set $\mu = E(h(Y_i)) = 0$ and $\text{Var}(h(Y_i)) = \sigma^2 = 1$. In other words, our model expresses a linear relationship between the covariates $x_1,\ldots,x_p$ and the dependent variable when the dependent variable is measured on a scale $h(Y)$ with mean zero and variance one. The intercept parameter is zero and $\sum_{i=1}^{n} x_{ij} = 0$, $j = 1,\ldots,p$. Moreover, we assume that $[\underset{\sim}{X}'\underset{\sim}{X}]^{-1}$ exists, where $\underset{\sim}{X} = (\underset{\sim}{x}_1',\ldots,\underset{\sim}{x}_n')'$ is the design matrix.

There is a connection between our model and the path analysis and structural equation models of genetics and economics, e.g. Wright (1968), Duncan (1975), and Johnson and Wichern (1982). To see this, reparametrize by dividing each x_{ij} by $(n^{-1}\sum_{i=1}^{n} x_{ij}^2)^{1/2}$. Now the model is

$$h(Y_i) = \sum_{j=1}^{p} t_{ij}\theta_j + \varepsilon_i$$

where $t_{ij} = x_{ij}/(n^{-1}\sum_{i=1}^{n} x_{ij}^2)^{1/2}$ and $\theta_j = \beta_j(n^{-1}\sum_{i=1}^{n} x_{ij}^2)^{1/2}$. In this equation, $h(Y_i)$; $t_{i1},\ldots,t_{ip}$ and ε_i all are measured on scales with mean zero and variance 1. Note that we can write

$$\theta_j = \beta_j \frac{(n^{-1}\sum_{i=1}^{n} x_{ij}^2)^{1/2}}{(\mathrm{Var}(h(Y_i))^{1/2}}$$

Thus θ_j is a "correlation" between $h(Y)$ and the j^{th} covariate when the j^{th} covariate is regarded as non-random. In the multivariate case where the covariates $x_1,\ldots,x_p$ are regarded as random, a connection between the θ's and (multiple) correlation still exists.

From Hoeffding (1950) and Terry (1952) we find that the partial or rank likelihood is

$$L(\underset{\sim}{\beta}) = E\left\{\exp\left[\frac{-1}{2}\sum_{i=1}^{n}[(Z^{(r_i)} - \underset{\sim}{\beta}\underset{\sim}{x}_i)^2 - Z_i^2]\right]\right\}$$

where $(Z_1,\ldots,Z_n)$ is a standard normal sample and $(Z^{(1)} < \cdots < Z^{(n)})$ is the order statistics vector for this sample. We turn to two approximations of $L(\underset{\sim}{\beta})$:

(a) <u>An order bootstrap</u>

Using the Law of Large Numbers, we can approximate $L(\underset{\sim}{\beta})$ by

$$\hat{L}_M(\underset{\sim}{\beta}) = \frac{1}{M}\sum_{j=1}^{M}\exp\left[\frac{-1}{2}\sum_{i=1}^{n}[(Z_j^{(r_i)} - \underset{\sim}{\beta}\underset{\sim}{x}_i)^2 - Z_{ji}^2]\right]$$

where $(Z_{j1},\ldots,Z_{jn})$, $j = 1,\ldots,M$, are independent standard normal samples and $Z_j^{(1)} < \cdots < Z_j^{(n)}$, $j = 1,\ldots,M$, are their corresponding order statistics.

This is a resampling scheme different from the bootstrap. The bootstrap method (Efron (1979)) is based on resampling from the empirical distribution which is determined by order statistics from the original data and does not use its order. The resampling scheme introduced above introduces new order statistics in each new sample but arranges them in the same order as the original data. This procedure

has to be implemented on the computer. It is recommended to find the values $\hat{\underset{\sim}{\beta}}_M$ that maximize $\hat{L}_M(\underset{\sim}{\beta})$ for $M = 50, 100, 150$, etc. and to stop when there is practically no change from one M to the next.

(b) A local approximation

Let $\mu_i = \underset{\sim}{\beta}\underset{\sim}{x}_i$ be the mean of Y_i. We consider the parameter set $B_n = \{\underset{\sim}{\beta}: \sum_{i=1}^{n} (\mu_i - \bar{\mu})^2 \leq K^2, \max_{1\leq i\leq n} |\mu_i - \bar{\mu}| \to 0\}$ where K^2 is a constant not dependent on n, while $\underset{\sim}{\beta}$, μ_i and p may all depend on n although this is surpressed in the notation.

Note that in our parametrization, $\bar{\mu} = 0$ so we can write

$$B_n = \{\underset{\sim}{\beta}: \sum_{i=1}^{n} \mu_i^2 \leq K^2, \max_{1\leq i\leq n} |\mu_i| \to 0\} .$$

The claim is that the parameter set B_n is of interest and is important; however, this is not meant to imply that the more usual parameter spaces such as R^p or a p-dimensional rectangle are not at least as important.

The parameter set B_n contains those parameter values where it is hard to distinguish between means μ_i and we need to do a good job of estimating these means. For instance, in the two sample cases with the usual appropriate choice of the x's, we have $\beta = \mu_1 - \mu_2$ and

$$B_n = [-(\frac{n}{n_1 n_2})^{1/2} K, (\frac{n}{n_1 n_2})^{1/2} K]$$

where n_1 and n_2 are the sample sizes for the first and second samples, respectively.

Similarly, for linear regression with $\mu_i = \beta x_i$, we have

$$B_n = [-K/(\sum x_i^2)^{1/2}, K/(\sum x_i^2)^{1/2}] .$$

Suppose that $\hat{\mu}_i$ is an estimate of μ_i, that $\hat{\sigma}(\hat{\mu}_i)$ is an estimate of the standard error of $\hat{\mu}_i$, $i = 1,\ldots,n$, and suppose that

$$\hat{\mu}_i \pm 1.96\hat{\sigma}(\hat{\mu}_i) \quad \text{and} \quad \hat{\mu}_j \pm 1.96\hat{\sigma}(\hat{\mu}_j)$$

are two interval estimates of μ_i and μ_j with approximate confidence coefficients 95%. If $\underset{\sim}{\beta} \in B_n$, then these two intervals will typically continue to overlap as n increases. Moreover, they shrink down to the same point.

On the other hand, if $\underset{\sim}{\beta} \notin B_n$ and μ_i and μ_j are fixed, then the intervals will shrink down to two separate points as $n \longrightarrow \infty$ for any reasonable estimates $\hat{\mu}_i$ and $\hat{\mu}_j$ not necessarily asymptotically optimal. Thus in this case, any two reasonable estimates will perfectly distinguish between μ_i and μ_j in the limit, while for $\underset{\sim}{\beta} \in B_n$, this is not the case and we really need to develop good estimates.

An objection to the above argument may be that when μ_i and μ_j are close, it is not that important to distinguish between them. This is why the fixed μ_i parameter space is at least as important as B_n.

In any case, we will need to check that the results obtained using B_n leads to good approximations for finite sample sizes. Here an approximate ballpark rule is that it does so for parameter values where the power of the level .05 likelihood ratio test of $H_0\colon \beta_1 = \cdots = \beta_p$ based on $h(Y_i)$ (assuming h known) has asymptotic power at most .95. Thus for the two sample problem, we have approximately

$$\beta\varepsilon[-(\frac{n}{n_1 n_2})^{1/2}3.6,(\frac{n}{n_1 n_2})^{1/2}3.6] \ .$$

For $n_1 = n_2 = 20$, this interval is $[-1.14,1.14]$.

Similarly, for linear regression where $E(h(X_i)) = \beta x_i$, properties derived for the local parameter space should give good approximations when

$$\beta \in [-3.6/(\sum x_i^2)^{1/2}, 3.6/(\sum x_i^2)^{1/2}] .$$

We will return to this question later.

For the local parameter space B_n, a standard approximation to the partial or rank likelihood $L(\underset{\sim}{\beta})$ is available, e.g. Hoeffding (1950), Terry (1952), Hajek (1962), and Hajek and Sidak (1967). If we maximize this approximation, we find that the local partial likelihood estimate of $\underset{\sim}{\beta}$ is

$$\hat{\underset{\sim}{\beta}} = \underset{\sim}{C}\underset{\sim}{A}$$

$$\text{with } \underset{\sim}{C} = (c_{ji}) = [\underset{\sim}{X}'\underset{\sim}{X}]^{-1}\underset{\sim}{X}' \text{ and } \underset{\sim}{A}' = (a(r_1),\ldots,a(r_n))$$

where $a(1),\ldots,a(n)$ are the normal scores defined by $a(k) = E(Z^{(k)})$, $Z^{(k)}$ being the k^{th} order statistic in a sample of size n from a $N(0,1)$ population. Again, dependence on n has been surpressed in the notation.

Theorem 1. *Let* $\hat{\beta}_j$ *denote the* j^{th} *component of* $\hat{\underset{\sim}{\beta}}$. *If* $\sum_{i=1}^{n} c_{ji}^2 / \max_{1\le i\le n} c_{ji}^2 \to \infty$, *then for* $\underset{\sim}{\beta} \in B_n$,

$$\frac{\hat{\beta}_j - \beta_j}{(\sum_{i=1}^{n} c_{ji}^2)^{1/2}} \xrightarrow{d} N(0,1) \quad \text{as} \quad n \to \infty .$$

Proof: According to Hajek (1962) and Hajek and Sidak (1967, p.216),

$$\frac{\sum_{i=1}^{n} (c_{ji} - \bar{c}_{j\cdot})a(R_i) - \mu_c}{\sigma_c} \xrightarrow{d} N(0,1)$$

where, in our parametrization

$$\bar{c}_{j\cdot} = \frac{1}{n}\sum_{i=1}^{n} c_{ij} = 0$$

and where

$$\mu_c = \sum_{i=1}^{n} c_{ji}(\mu_i - \bar{\mu}) \int_0^1 [\Phi^{-1}(u)]^2$$

$$= \sum_{i=1}^{n} c_{ji}\mu_i$$

$$= \beta_j$$

Furthermore, σ_c^2 is given by

$$\sigma_c^2 = \sum_{i=1}^{n} c_{ji}^2 \int_0^1 [\Phi^{-1}(u)]^2 du = \sum_{i=1}^{n} c_{ji}^2 \ . \qquad \square$$

It follows that an approximate 95% confidence interval for β_j is

$$\beta_j \overset{.95}{=} \hat{\beta}_j \pm 1.96 \left(\sum_{i=1}^{n} c_{ji}^2\right)^{1/2} .$$

This looks deceptively simple. The price we pay is that the confidence coefficient 95% is only valid when the dependent variable is measured on a scale $h(Y_i)$ with variance 1. On the other hand, the interval can be used to test H_0: $\beta_j = 0$: The test which rejects H_0 when 0 is not in the interval has approximate level $\alpha = .05$. Finally, note that estimation of h and scale and their effect on $\hat{\beta}_j$ will be treated later.

For the correlation type parameter θ_j, the approximate 95% confidence interval is

$$\theta_j \overset{.95}{=} \hat{\theta}_j \pm 1.96\left[\left(\frac{1}{n}\sum_{j=1}^{n} x_{ij}^2\right)\left(\sum_{i=1}^{n} c_{ji}^2\right)\right)^{1/2}$$

where $\hat{\theta}_j = \hat{\beta}_j(n^{-1}\sum_{i=1}^{n} x_{ij}^2)^{1/2}$. This interval is valid when $x_1,\ldots,x_p$ and $h(Y_i)$ are measured on scales with mean zero and variance 1. When $p = 1$, the above interval reduces to

$$\theta_j \overset{.95}{=} \hat{\theta}_j \pm 1.96/\sqrt{n}$$

which coincides with the approximate confidence interval for the correlation coefficient ρ between X and $h(Y)$ near $\rho = 0$.

Another consequence of the result is that $\hat{\beta}$ is asymptotically optimal for all h. In fact, if we consider the UMVUE (uniformly minimum variance unbiased estimate) $\tilde{\beta}_j$ based on $h(Y_i)$, $i = 1,\ldots,n$, which coincides with the MLE and LSE and is given by

$$\tilde{\beta}_j = \sum_{i=1}^{n} c_{ji} h(Y_i) ,$$

then $\tilde{\beta}_j$ is normally distributed with mean β_j and has exactly variance $\sum_{i=1}^{n} c_{ji}^2$, which is the asymptotic variance of the local partial likelihood estimate (LPLE) $\hat{\beta}_j$. It follows that $\hat{\beta}_j$ is adaptive in the sense of Bickel (1980), even though it can be computed without estimating h.

It can also be shown that

$$\frac{\hat{\beta}_j - \tilde{\beta}_j}{\sqrt{\sum_{i=1}^{n} c_{ji}^2}} \xrightarrow{P} 0$$

and that the above results can be extended to $\hat{\underset{\sim}{\beta}}$ and to linear parameters $\alpha = \sum a_j \beta_j$ as in Huber (1973, 1981).

Remarks

(a) It follows from Hajek and Sidak (1967) that the results of this section continue to hold if we replace the normal scores $a(k)$ by the approximate normal scores

$$a_A(k) = \Phi^{-1}(\frac{k}{n+1}), \quad k = 1,\ldots,n$$

which are much more readily available. Similarly, it follows from the results of Bell and Doksum (1963) that the results hold if we replace the $a(k)$ by standard normal order statistics $Z^{(k)}$. The advantage of using $Z^{(k)}$ is that the resulting estimates $\hat{\beta}_j$ are nearly normally distributed. It is recommended that several independent order statistics vectors $Z_i^{(1)} < \cdots < Z_i^{(n)}$, $i = 1,\ldots,M$, be used. For each vector, an estimate of β_j is formed and then $\hat{\beta}_{jM}$ is the average of these M values. Try $M = 10$, 20, etc. and stop when there is little change in $\hat{\beta}_{jM}$ as M changes to the next higher value. This is a local version of the resampling procedure in Section 2(a).

(b) The estimates introduced in Section 2(b) and Remark (a) above can be used as starting points for iterative procedures when maximizing $\hat{L}_M(\underset{\sim}{\beta})$ of Section 2(a).

(c) Suppose we compute $\hat{\underset{\sim}{\beta}}$ as in Section 2(b) and then compute $\hat{\underset{\sim}{\mu}} = \underset{\sim}{X}\hat{\underset{\sim}{\beta}}$. If $\| \hat{\underset{\sim}{\mu}} \| = (\sum_{i=1}^{n} \hat{\mu}_i^2)^{\frac{1}{2}} > 3.6$, then this is an indication that $\underset{\sim}{\beta}$ is not in the local parameter set and we need to use the estimate of Section 2(a).

(d) The approach of this section can be used to find conditions for the adaptability of Cox's partial likelihood estimates.

(e) It can also be used to find conditions for the adaptability of the Box-Cox estimates. That is, in the terminology of the transformation controversy, conditions under which, in the asymptotics, no allowance needs to be made for the estimation of the transformation parameter.

(f) Conditions under which $\hat{\underset{\sim}{\beta}}$ can be estimated efficiently when both h and F are unknown can be established.

(g) Estimation F, h and σ^2 will be considered in a forthcoming paper.

REFERENCES

Bell, C. B. and Doksum, K. A. (1965). Some new distribution-free statistics. Ann. Math. Statist., 36, 203-214.

Bickel, P. J. (1980). On adaptive estimation. Ann. Statist., 10, 647-671.

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. J. Amer. Statist. Assoc., 76, 293-311.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. J. Royal Statist. Soc., B, 26, 211-252.

Cox, D. R. (1972). Regression models and life tables. J. Roy. Statist. Soc., B, 34, 187-202.

Cox, D. R. (1975). Partial likelihood. Biometrika, 62, 269-276.

Duncan, O. D. (1975). Introduction to Structural Equation Models, Academic Press, New York.

Efron, B. (1979). Bootstrap methods: Another look at the jack-knife. Ann. Statist., 7, 1-26.

Hajek, J. (1962). Asymptotically most powerful rank order tests. Ann. Math. Statist., 33, 1124-1147.

Hajek, J. and Sidak, Z. (1967). Theory of Rank Tests, Academic Press, New York.

Hoeffding, W. (1950). "Optimum" nonparametric tests: Proc. 2nd Berkeley Symposium, 83-92.

Huber, P. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. Ann. Statist., 1, 799-821.

Huber, P. (1981). Robust Statistics, Wiley, New York.

Johnson, R. A. and Wichern, D. W. (1982). Applied Multivariate Statistical Analysis, Prentice Hall, New Jersey.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. Biometrika, 60, 267-278.

Lehmann, E. L. (1953). The power of rank tests. Ann. Math. Statist., 24, 23-43.

Terry, M. E. (1952). Some rank orders tests which are most powerful against specific parametric alternatives. Ann. Math. Statist., 23, 346-366.

Tukey, J. W. (1957). On the comparative anatomy of transformations. Ann. Math. Statist., 28, 602-632.

Wright, S. (1968). Evolution and the Genetics of Population, Vol. 1, University of Chicago Press.