REGRESSION MODELS FOR LONGITUDINAL DATA

MARTINA ORSÁKOVÁ

Abstract. V článku referuji vlastnosti parametrického odhadu v regresním modelu $Z_i^k = m(\theta_0, \mathbb{X}_i(T_i^k)) + \varepsilon_i^k$, kde T_i^k představuje čas k-tého pozorování vysvětlované proměnné Z u i-tého subjektu. Obecně mnohorozměrný proces kovariát $\mathbb{X}_i(s) = (X_i^{(1)}(s), ..., X_i^{(d)}(s))$ závisí na událostech do času s, tj. na $\{(Z_i^k, T_i^k) | k : T_i^k < s\}$. Odhad neznámých parametrů θ_0 je založen na metodě nejmenších vážených čtverců.

Резюмэ. В работе реферирую качества параметрического оценивания в регрессивной модели $Z_i^k = m(\theta_0, \mathbb{X}_i(T_i^k)) + \varepsilon_i^k$, где T_i^k есть время k-наблудения зависимой переменнии Z для i-субъекта. Вообще многомерной процесс независимых переменний $\mathbb{X}_i(s) = (X_i^{(1)}(s), ..., X_i^{(d)}(s))$ зависит от события до времени s, т.е. от $\{(Z_i^k, T_i^k) | k : T_i^k < s\}$. Оценивание неизвестных параметров θ_0 основано на методе наименьшич взвшеных квадратов.

1. Introduction

Consider the triples $(T_i^k, Z_i^k, \mathbb{X}_i(T_i^k))$, where T_i^k is the time of the k-th observations of i-th subject, Z_i^k is dependent variable and $\mathbb{X}_i(T_i^k)$ is a d-dimensional process of covariates: $\mathbb{X}_i(s) = (X_i^{(1)}(s), X_i^{(2)}(s), ..., X_i^{(d)}(s))$. We assume that we have n i.i.d. replicates from a generic model. The T_i's can be exposed to different censoring systems. The covariates $\mathbb{X}_i(s)$ can depend on the events prior to time s, i.e. on the $\{(Z_i^k, T_i^k) | k : T_i^k < s\}$. The process $\mathbb{X}_i(s)$ may include quantitative as well as qualitative variables, and could include indicators of sex, treatment group, geographical regions or age, and various other continuous measurements. In this paper we investigate properties of a particular estimator for the regression parameters in the model $Z_i^k = m(\theta_0, \mathbb{X}_i(T_i^k)) + \varepsilon_i^k$.

2. Parametric model

Assume that $m(.)$ is a regression function - known and (sufficiently) smooth function of unknown regression parameters θ_0 :

$$Z_i^k = m(\theta_0, \mathbb{X}_i(T_i^k)) + \varepsilon_i^k, \tag{1}$$

where $i = 1, ..n$, $k = 1, ..N_t^i$, ε_i^k is a noise. The process N_t^i counts number of observations for i-th subject in $[0, t]$, i.e. $N_t^i = \sum_k I(T_i^k \le t)$.

2000 *Mathematics Subject Classification.* Primary 62J02.
Key words and phrases. Longitudinal data, regression model.
Tato práce vznikla za podpory grantu MSM 113200008.

2.1. **Assumptions.** Define $R_t^i = \sum_{k=1}^{N_t^i} Z_i^k$ and a filtration

$$\mathcal{F}_t = \sigma(R_s^i, N_s^i : s \le t,\ i = 1, ..n) \vee \mathcal{A},$$

the history of the process R_t^i and N_t^i for $i = 1, ..n$. $\mathcal{A}$ is assumed to be a σ-algebra independent of $\sigma(R_s^i, N_s^i : s \le t,\ i = 1, ..n)$, and represents knowledge prior to time 0. It is assumed that no two of the counting processes N_t^i jump at the same time (almost surely). Let $\mathbb{X}_i(s)$ is predictable process with respect to the filtration $\mathcal{F}_s$ and is "cadlag" - the right continuous with left limits process. The process N_t^i has a random intensity $\lambda_t^i \ge 0$, which is "cadlag". One particular form for the intensity is Aalen's multiplicative model: $\lambda_t^i = \alpha(t) Y_t^i$, where $\alpha(.)$ is deterministic function and Y_t^i is $\mathcal{F}_t$-predictable process. In practice, censoring indicators play the role of Y_t^i, i.e. the variable is 1 if the i-th subject is at risk and 0 otherwise.

The conditional distribution of the noise terms ε_i^k is denoted

$$F_s(z) = P(\varepsilon_i^{N_u^i+1} \le z | \mathcal{F}_u, T_i^{N_u^i+1} = s), \tag{2}$$

and it is assumed that ε_i^k have conditional mean and variance given by

$$E(\varepsilon_i^{N_u^i+1} | \mathcal{F}_u, T_i^{N_u^i+1} = s) = 0, \tag{3}$$

$$E((\varepsilon_i^{N_u^i+1})^2 | \mathcal{F}_u, T_i^{N_u^i+1} = s) = \sigma^2(\mathbb{X}_i(s)), \tag{4}$$

where $\sigma^2(.)$ is deterministic, continuous and bounded function.

2.2. **Estimation.** Let $\mathcal{B}$ denote the Borel σ-field on $\mathbb{R}$. For $A \in \mathcal{B}$, define the counting process

$$N_s^i(A) = \sum_k I(Z_i^k \in A) I(T_i^k \le s), \tag{5}$$

that counts the number of observations of "size-A" in the time interval $[0, t]$ for the i-th subject, and the associated marked point process $P^i(ds \times dz) : P^i([0, s] \times A) = N_s^i(A)$. Note the notation

$$\int_0^t \int H_i(s, z) P^i(ds \times dz) = \sum_k H_i(T_i^k, Z_i^k) I(T_i^k \le t).$$

Our estimator $\hat{\theta}$ of θ_0 is based on the conditional weighted least squares. The objective function is

$$L(\theta, t) = \frac{1}{2n} \sum_{i=1}^{n} \sum_{k=1}^{N_t^i} (Z_i^k - m(\theta, \mathbb{X}_i(T_i^k))^2 W(\mathbb{X}_i(T_i^k)), \tag{6}$$

where $W(.)$ is some weight function independent of θ. Denote

$$H_i(\theta, s, z) = \frac{1}{2}(z - m(\theta, \mathbb{X}_i(s))^2 W(\mathbb{X}_i(s))$$

and (6) can be rewritten as

$$L(\theta, t) = \frac{1}{n} \sum_{i=1}^{n} \int_0^t \int H_i(\theta, s, z) P^i(ds \times dz). \tag{7}$$

Denote $U(\theta,t) = ((\frac{d}{d\theta_k})L(\theta,t))_{k=1,..d}$ and $I(\theta,t) = ((\frac{d^2}{d\theta_j d\theta_k})L(\theta,t))_{j,k=1,..d}$. The estimator $\hat{\theta}$ is defined as a solution to the equations:

$$U(\theta,t) = 0. \tag{8}$$

Such a $\hat{\theta}$ does not necessarily minimize $L(\theta,t)$, it may not be a consistent estimator and may not exist. However, if we assume that there exists a unique maximum for $L(\theta,t)$, the estimator is consistent.

Before giving the asymptotic results about the estimator of θ_0, some further conditions must be stated. We discus the specific model where the processes $R_s^1,\ ...R_s^n$ are assumed to be independent and identically distributed. These assumptions could be weakened, for more general results see [6].

ASSUMPTION A

Assume that $m(\theta,x)$ is three times continuously differentiable in a neighborhood $B(\theta_0,r)$ around θ_0.

ASSUMPTION B

Let further $W(.)$, $E(\int_0^t \lambda_s^i ds)$ and the derivatives of $m(.,.)$ in a neighborhood $B(\theta_0,r)$ are bounded.

ASSUMPTION C

The processes

$$q_s^i := \frac{d}{d\theta_k} m(\theta_0, \mathbb{X}_i(s)) \frac{d}{d\theta_j} m(\theta_0, \mathbb{X}_i(s)) W(\mathbb{X}_i(s)) \lambda_s^i$$

and

$$Q_s^i := \frac{d}{d\theta_k} m(\theta_0, \mathbb{X}_i(s)) \frac{d}{d\theta_j} m(\theta_0, \mathbb{X}_i(s)) \sigma^2(\mathbb{X}_i(s)) W^2(\mathbb{X}_i(s)) \lambda_s^i$$

satisfy

$$E\left(\sup_{s\in[0,t]} |q_s^i|\right) < \infty, \quad E\left(\sup_{s\in[0,t]} |Q_s^i|\right) < \infty$$

for all $i = 1,...n$, and $j,k = 1,...d$.

Assume further that there exist non-negative definitive symmetric matrices Σ_I, Σ_U such that $\Sigma_I = \left(E(\int_0^t q_s^i ds)\right)_{j,k=1,...d}$ and $\Sigma_U = \left(E(\int_0^t Q_s^i ds)\right)_{j,k=1,...d}$.

The following theorems give the asymptotic results about the estimator $\hat{\theta}$. All proofs can be found in [6].

Theorem 1:

Under the assumptions A-C there exists a consistent solution to $U(\theta,t) = 0$, that provides a local minimum of $L(\theta,t)$ with probability tending to one. □

Theorem 2:

Under the assumptions A-C and with $\hat{\theta}$ a consistent solution of $U(\theta,t) = 0$, one has

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, \Sigma_I^{-1} \Sigma_U \Sigma_I^{-1}).$$

Further $-I(\hat{\theta},t)$ and $\hat{\Sigma}_U$ provide consistent estimates of Σ_I and Σ_U respectively, where $\hat{\Sigma}_U := \left(\frac{1}{n}\sum_{i=1}^{n}\int_0^t\int(\frac{d}{d\theta_j}H_i(\hat{\theta},s,z))(\frac{d}{d\theta_k}H_i(\hat{\theta},s,z))P^i(ds\times dz)\right)_{j,k=1,\ldots d}$.

□

The following result is of interest for the purpose of testing a simple hypothesis $H:\theta=\theta_0$ against the composite alternative $G:\theta\neq\theta_0$. Let $W_d(\Sigma)$ denote the Wishart distribution corresponding to a d-dimensional normal distribution $N(0,\Sigma)$.

Theorem 3:

Under the hypothesis H and under the assumptions A-C and if $\hat{\theta}$ is a consistent solution of $U(\theta,t)=0$ one has

$$n(L(\hat{\theta},t)-L(\theta_0,t))\xrightarrow{D}W_d\left(\frac{1}{2}\Sigma_I^{-1/2}\Sigma_U\Sigma_I^{-1/2}\right).$$

□

When the conditional variance $\sigma^2(.)$ of the error term ε_i^j is a known function, a natural choice of the weight function is $W(.)=\frac{1}{\sigma^2(.)}$.

Theorem 4:

The asymptotic variance of $\sqrt{n}(\hat{\theta}-\theta_0)$ is minimized by choosing the weight function $W(.)$ as the inverse of the conditional variance, $\frac{1}{\sigma^2(.)}$; that is $\Sigma^{opt}\leq\Sigma_I^{-1}\Sigma_U\Sigma_I^{-1}$, where Σ^{opt} is the asymptotic variance of $\sqrt{n}(\hat{\theta}-\theta_0)$ for the choice $W(.)=\frac{1}{\sigma_2(.)}$.

□

The conditional variance, $\sigma^2(.)$ is usually unknown and in some cases one can replace $\sigma^2(.)$ by a uniformly consistent estimator, $\hat{\sigma}^2(.)$, and still obtain the theorems of this section. In the case of Aalen's multiplicative intensity model, Sheike in [6] proposed a non-parametric estimator of $\sigma^2(.)$, which is uniformly consistent under some regularity conditions. An estimator is $\hat{\sigma}^2(y)=\frac{\hat{V}(y)}{\hat{o}(y)}$, where

$$\begin{aligned}\hat{V}(y) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_t^i}({Z_i^j}^2-(\hat{m}(y))^2)\frac{1}{\beta}K(y-\mathbb{X}_i(T_i^j),b)\frac{J^i(T_i^j)}{Y^i_{T_i^j}},\\ \hat{o}(y) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_t^i}\frac{1}{\beta}K(y-\mathbb{X}_i(T_i^j),b)\frac{J^i(T_i^j)}{Y^i_{T_i^j}}.\end{aligned}$$

$K(.)$ is a d-dimensional kernel, $b=(b_1,...,b_d)$ is a bandwidth, $\beta=b_1...b_d$ and $J^i(s)=I(Y_s^i>0)$. The kernel estimation of the regression function (see [7]) is defined as

$$\hat{m}(y)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{N_t^i}Z_i^jK(y-\mathbb{X}_i(T_i^j),b)}{\sum_{i=1}^{n}\sum_{j=1}^{N_t^i}K(y-\mathbb{X}_i(T_i^j),b)}.$$

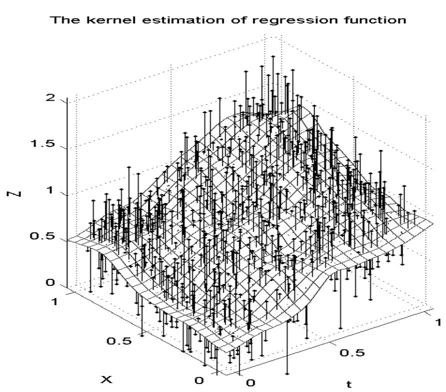
Sheike in [6] proved that under some "technical assumptions" and provided $\hat{\sigma}^2$ is a uniformly consistent estimator of the σ^2, theorems 1-3 remain true for $W(.)=\frac{1}{\hat{\sigma}^2}$.

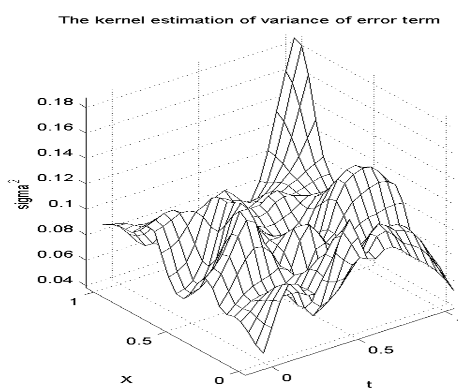
3. Simulation study

We studied the properties of our estimator in the following limited simulation study. We simulated a two-dimensional longitudinal data. Suppose the covariate process $\mathbb{X}_i(T_i^j) = (X_i(T_i^j), T_i^j)$, where X_i is an additional covariate for the i-th individual and T_i^j is the j-th random observation time of i-th individual.

The covariate processes X_i were generated from the random variables with density function $\frac{5}{6} + \frac{x}{3}$ for $x \in [0, 1]$. For each individual the observation times were generated from the Poisson process over the unit time period of $[0, 1]$ with parameter λ, where $\lambda = 10$, if the previous response value $Z_i^{j-1} < 0.8$, and $\lambda = 20$ otherwise. Notice that the random observation time depends on the value of the previous measurement. The response values were generated as $Z_i^j = \theta_1 + \theta_2 \cdot T_i^j + \theta_3 \cdot X_i^j + \varepsilon_i^j$, where $\theta_1 = 0.2$, $\theta_2 = 0.8$ and $\theta_3 = 0.4$ and ε_i^j has a normal distribution with mean zero and standard deviation 0.3. This linear regression model satisfies needed assumptions. We generate 80 individuals in the sample. On the average there are about 13 observations for each individual. Together we have got 1068 observations.

The following figure provides a plot of a non-parametric estimate of the regression function. The estimate of the conditional variance of the noise term $\sigma^2(.)$ is plotted in the second figure. We used Gaussian kernel with bandwidth $b_1 = b_2 = 0.002$.





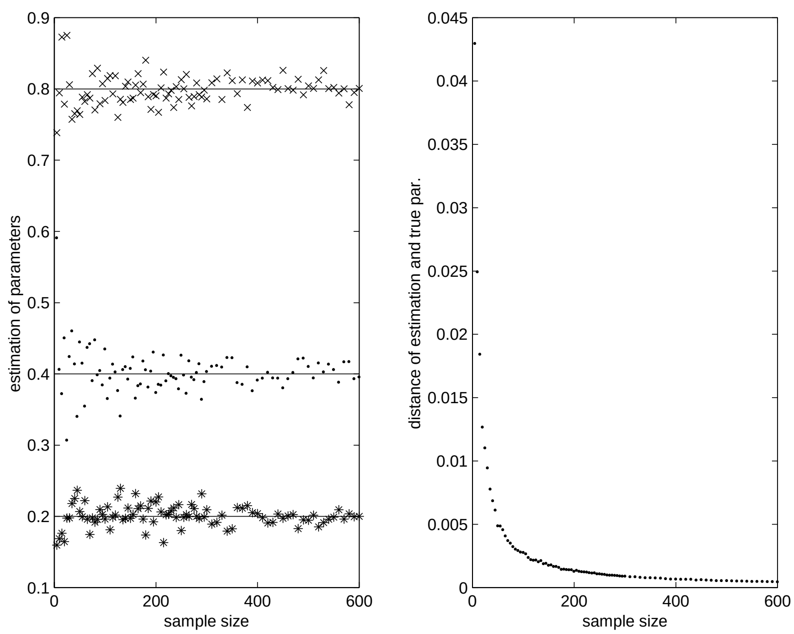
The estimation of θ_0 is done for two different choices of the weight function. The first weight function examined is $W(.) \equiv 1$. For this choice the estimate of θ_0 and the estimate of their standard errors and correlations are given in the table:

	Estimates	Standard errors	Estimated correlations		
i	$\hat{\theta}_i$	$\sqrt{var\ \theta_i}$	1	2	3
1	0.1912	0.0026	1	-0.6838	0.0285
2	0.8158	0.0034	-	1	-0.6401
3	0.3799	0.0034	-	-	1

The second choice of the weight function is $W(.) = \frac{1}{\hat{\sigma}^2}(.)$, where $\hat{\sigma}^2(.)$ is the Sheike's estimate of $\sigma^2(.)$. The weighted estimate of θ_0, standard errors and correlations are given in this table:

	Estimates	Standard errors	Estimated correlations		
i	$\hat{\theta}_i$	$\sqrt{var\ \theta_i}$	1	2	3
1	0.2019	0.0021	1	-0.6585	0.0123
2	0.8125	0.0029	-	1	-0.6397
3	0.3684	0.0028	-	-	1

The last pictures show the estimated parameters and the distance of the estimated parameter and its true value for different sample size. The estimate was computed for the choice of weights $W(.) \equiv 1$.



4. Some remarks

The Sheike's methods mesh well with the modeling in terms of the conditional distribution of the current observation given the past. When there is a strong time dependence, the conditional least square methods seem preferable to the generalized estimation equations. It was mentioned that the optimal choice of the weight function is the inverse of the conditional variance. Some caution has to be taken in choosing the smoothing parameters b. The appropriate choice of the bandwidth ensures that the regression function estimates are not too unstable and at the same time do not introduce too much bias. A small difficulty arises from our non-parametric estimator: the bias is more severe at the edges of the data and therefore one must limit the results to nonedge areas.

References

[1] Flemming T.R., Harrington, D.P. (1991): Counting Processes and Survival Analysis. Wiley - Sons.
[2] Jacobsen M. (1982): Statistical Analysis of Counting Processes. Springer-Verlag.
[3] Lehmann E. (1983): Theory of Point Estimation. Wiley.
[4] Martinussen, T., Scheike, T.H. (1998): A Non-Parametric Dynamic Additive Regression Model for Longitudinal Data. Research Report. Copenhagen.
[5] Martinussen, T., Scheike, T.H. (1998): A Semi-Parametric Additive Regression Model for Longitudinal Data. Research Report. Copenhagen.
[6] Scheike, T.H. (1994): Parametric Regression for Longitudinal Data with CP Measurement Times. Scandinavian Journal of Statistics 21.
[7] Scheike, T.H., Zhang M. (1998): Cumulative Regression Tests for Longitudinal Data. Annals of Statistics.

UK MFF, KPMS, Sokolovsk 83, 186 75 Praha
E-mail: orsakova@karlin.mff.cuni.cz