The study of the multivariate logistic regression with increasing number of covariates

Arseniy Khaplanov

(Lomonosov Moscow State University)

Workshop "Stochastics, Geometry and Analysis"
Ulm, September 2–6, 2013

Introduction

There are random vector $(X^T, Y)^T \in \mathbb{R}^p \times \{0, 1\}$ and one constant $\alpha^0 \in \mathbb{R}^m$ defined in such a way that for some known function $f : \mathbb{R}^m \times \mathbb{R}^p \to [0, 1]$ the following equation holds:

$$\mathrm{P}(Y = 1|X = x) = f(\alpha^0, x)$$

and

$$\mathrm{P}(Y = 0|X = x) = 1 - f(\alpha^0, x).$$

For a sequence of i.i.d. random vectors $(X_q^T, Y_q)^T$, $q \in \mathbb{N}$ with the same distributions as for (X, Y) our aim is to construct any estimation $\widehat{\alpha}_n$ for parameter α^0.

Introduction

Logistic regression

One of the most commonly used functions is logistic function. Let set $m = p$. Define logistic function:

$$f(\alpha, x) = \frac{e^{-\alpha^T x}}{1 + e^{-\alpha^T x}}.$$

To construct estimator for this model the method of maximum likelihood is generally used with likelihood function

$$L_n(\alpha) = \prod_{q=1}^{n} \left[\mathrm{I}(Y_q = 1) f(\alpha, x) + \mathrm{I}(Y_q = 0)(1 - f(\alpha, x) \right].$$

So $\widehat{\alpha}_n = \operatorname{argmax}_{\alpha \in \mathbb{R}^p} L_n(\alpha)$.

Introduction

Multinomial logistic regression

Define some set $K = \{1, \dots, k\}$. Assume that $Y \in \{0\} \cup K$ and there are some nonrandom vectors $\alpha_1^0, \dots, \alpha_k^0 \in \mathbb{R}^p$ such that for every $j \in K$ and $x \in \mathbb{R}^p$

$$\mathrm{P}\,(Y = j\,|X = x) = \frac{\exp\{-(\alpha_j^0)^T x\}}{1 + \sum_{t=1}^k \exp\{-(\alpha_t^0)^T x\}},$$
$$\mathrm{P}\,(Y = 0\,|X = x) = \frac{1}{1 + \sum_{t=1}^k \exp\{-(\alpha_t^0)^T x\}},$$

Introduction

Multinomial logistic regression

Assume that $p = p_n, \; n \in \mathbb{N}$ and for each $n \in \mathbb{N}$ there exist i.i.d. random vectors $(X_q(n)^T, Y_q(n)^T)^T$ with parameters $\alpha_j^{0,n}$ of the logistic regression dependence.

Assumptions

A1 $p_n/n \to 0$ when $n \to \infty$;

A2 $\max_{i,q} |X_q^i| < \infty$ a.s. for all $n \in \mathbb{N}$. Define $S_n = \sum_{q=1}^n X_q X_q^T$. There exist two positive constants c_{min} and C_{max} such that the following equation holds for all $n \in \mathbb{N}$ a.s.

$$c_{min} n \leq \lambda_{min}(S_n) \leq \lambda_{max}(S_n) \leq C_{max} n.$$

Introduction

Theorem (Liang (2012))

Let assumptions (A1) and (A2) hold. Then there exists a sequence of a random variables $\widehat{\alpha}_n$ such that for $n \to \infty$

$$\mathrm{P}\left\{L_n(\widehat{\alpha}_n) = 1\right\} \to 1$$

and

$$u^T G_n^{1/2}\left(\widehat{\alpha}_n - \alpha_0\right) \overset{Law}{\to} Z,\ Z \sim N(0,1),$$

where u is a unit p_n-vector, and G_n - is a covariate matrix of $\nabla L_n(\alpha_n^0)$.

Main results

For some natural k there are random vectors $((X(n)^T, Y(n))^T$, where $X(n) \in \mathbb{R}^{p_n}$ and $Y(n) \in K \cup \{0\}$ for $K = \{1, 2, \ldots, k\}$. Assume that for every $n \in \mathbb{N}$, $j \in K$ and $x \in \mathbb{R}^{p_n}$ following equations hold:

$$\mathrm{P}\left(Y(n) = j \,|X(n) = x\right) = \frac{\exp\{-(\alpha_j^{0,n})^T x\}}{1 + \sum_{t=1}^{k} \exp\{-(\alpha_t^{0,n})^T x\}},$$

$$\mathrm{P}\left(Y(n) = 0 \,|X(n) = x\right) = \frac{1}{1 + \sum_{t=1}^{k} \exp\{-(\alpha_t^{0,n})^T x\}},$$

where $\alpha_j^{0,n} \mathbb{R}^{p_n}$, $j \in K$ are nonrandom vectors, which are parameters of the multinomial logistic regression model.

Main results

Our aim is to construct and investigate some estimate for the vector of parameters $\alpha^{0,n} := ((\alpha_1^{0,n})^T, \dots, (\alpha_k^{0,n})^T)^T$ if we have a sample $((X_1(n)^T, Y_1(n))^T, \dots, (X_n(n)^T, Y_N(n))^T)$. For $\alpha \in \mathbb{R}^{kp_n}$ determine double numerate of components:

$$\alpha^{(r,l)} := \alpha^{l+(r-1)p_n} = (\alpha_r)^l.$$

So $(\alpha^{0,n})^{r,\cdot}$ is equal to vector $\alpha_r^{0,n}$. Define some functions

$$\pi_j^n(\alpha, x) = \frac{\exp\left\{-\left(\alpha^{(j,\cdot)}\right)^T x\right\}}{1 + \sum_{t=1}^k \exp\left\{-\left(\alpha^{(t,\cdot)}\right)^T x\right\}},$$

$$\pi_0^n(\alpha, x) = \frac{1}{1 + \sum_{t=1}^k \exp\left\{-\left(\alpha^{(t,\cdot)}\right)^T x\right\}}.$$

Main results

Loss function:

$$\widetilde{L}_n(\alpha) = \prod_{q=1}^{n} \left(\pi^n_{Y_q(n)}(\alpha, X_q(n)) \right) = \prod_{q=1}^{n} \left\{ \prod_{j=0}^{k} \left[\pi^n_j(\alpha, X_q(n)) \right]^{\mathrm{I}\{Y_q(n)=j\}} \right\}.$$

Maximum likelihood estimate is $\widehat{\alpha}_n = \mathrm{argmax}_{\alpha \in \mathbb{R}^{kp}} \widetilde{L}_n(\alpha)$. But instead of finding maximum of this function we will look for a root for the gradient of it.

$$R_n(\alpha) = \left\{ R_n^{(1,1)}(\alpha), \ldots, R_n^{(1,p_n)}(\alpha), R_n^{(2,1)}(\alpha), \ldots, R_n^{(k,p_n)}(\alpha) \right\}^T,$$

where $R_n^{(r,l)}(\alpha) = \sum_{q=1}^{n} \left[\mathrm{I}\left\{ Y_q(n) = r \right\} - \pi^n_r(\alpha, X_q(n)) \right] X_q^l(n)$ and $\alpha \in \mathbb{R}^{kp_n}$.

Main results

It is easy to see that $R_n(\alpha) = \nabla\left[\ln \widetilde{L}_n(\alpha)\right]$. Define $\widehat{\alpha}_n$ as the root of equation $R_n(\alpha) = 0$ with the smallest norm. If there's no root of it then $\widehat{\alpha}_n := 0$. Determine new function:

$$b_{(m,s)}^{(r,l)}(n) = \mathbb{E}\, X^l(n) X^s(n) \cdot$$
$$\cdot \begin{cases} -\pi_r^n(\alpha^{0,n}, X(n))\pi_m^n(\alpha^{0,n}, X(n)), & \text{if } r \neq m; \\ \pi_r^n(\alpha^{0,n}, X(n))(1 - \pi_r^n(\alpha^{0,n}, X(n))), & \text{if } r = m. \end{cases}$$

Define matrix $B_n = \left(b_{(m,s)}^{(r,l)}(n)\right)$ with size $kp_n \times kp_n$ and $G_n = nB_n$.

Main results

Assumptions

- B1 *There exists constant* $C > 0$ *such that for every* $n \in \mathbb{N}$ $\|X(n)\| \leq C$ *holds a.s.*
- B2 *There exists constant* $c > 0$ *such that* $m_n^T B_n m_n \geq c\|m_n\|^2$ *holds for all* $n \in \mathbb{N}$ *and* $m_n \in \mathbb{R}^{kp_n}$.

$$U_n = \begin{pmatrix} \overbrace{\frac{1}{\sqrt{p_n}} \quad \dots \quad \frac{1}{\sqrt{p_n}}}^{p_n} & \overbrace{0 \quad \dots \quad 0}^{p_n} & \dots & \overbrace{0 \quad \dots \quad 0}^{p_n} \\ 0 \quad \dots \quad 0 & \frac{1}{\sqrt{p_n}} \quad \dots \quad \frac{1}{\sqrt{p_n}} & \dots & 0 \quad \dots \quad 0 \\ \vdots \qquad\quad \vdots & \vdots \qquad\quad \vdots & \ddots & \vdots \qquad\quad \vdots \\ 0 \quad \dots \quad 0 & 0 \quad \dots \quad 0 & \dots & \frac{1}{\sqrt{p_n}} \quad \dots \quad \frac{1}{\sqrt{p_n}} \end{pmatrix}.$$

Main results

Theorem

Let assumptions (B1) and (B2) hold. Than for every sequence $\delta_n > 0$ such that $\delta_n p_n/\sqrt{n} \to 0$ and $\sqrt{p_n}/\delta_n \to 0$ for $n \to \infty$ we have

$$\mathbb{P}\left(\|\alpha^{0,n} - \widehat{\alpha}_n\| \geq \delta_n/\sqrt{n}\right) \to 0.$$

Theorem

Let assumptions (B1) and (B1) hold and $p_n = o(n^{1/3})$ for $n \to \infty$. Than

$$U_n G_n^{1/2}\left(\widehat{\alpha}_n - \alpha^{0,n}\right) \overset{Law}{\to} Z, \quad Z \sim N(0, E_k), \quad n \to \infty,$$

where E_k is a unit matrix order k.

Main results

Corollary

Let assumptions of the last theorem hold, than for $n \to \infty$

$$\left\| \frac{Q_n(\widehat{\alpha}_n)}{n} - \frac{G_n}{n} \right\|_2 \xrightarrow{\mathrm{P}} 0$$

and

$$U_n Q_n(\widehat{\alpha}_n)^{1/2}(\widehat{\alpha}_n - \alpha^{0,n}) \xrightarrow{d} Z, \quad Z \sim N(0, E_k),$$

where $\|M\|_2$ *is an operator norm of matrix* M.

Lemma

Let Γ be a continuous injection from $\mathbb{R}^p$ to $\mathbb{R}^p$ with $\Gamma(x_0) = y_0$ for some fixed point $x_0 \in \mathbb{R}^{p_n}$. Assume that for some constants $\delta, R > 0$ the inequation $\inf_{||x-x_0||=\delta} \|\Gamma(x) - y_0\| \geq R$ holds. Than for every $y \in \{u \in \mathbb{R}^p : \|u - y_0\| \leq R\}$ there is $x = x(y)$ such that $\Gamma(x(y)) = y$ and $\|x(y) - x_0\| \leq \delta$.

Lemma

Let assumptions of the theorem 1 hold. Than for $n \to \infty$ there is some inequation with E_{kp_n} — unit matrix with order kp_n.

$$\sup_{\alpha \in N_n(\delta_n)} \| G_n^{-1/2} Q_n(\alpha) G_n^{-1/2} - E_{kp_n} \|_2 \xrightarrow{P} 0,$$

where $N_n(\delta_n) = \{\alpha \in \mathbb{R}^{kp_n} : ||G_n^{1/2}(\alpha - \alpha^{0,n})|| \leq \delta\}$.

Proof

The main idea in the proof of the first theorem

Determine $\Gamma_n(\alpha) = G_n^{-1/2}\left[R_n(\alpha) - R_n(\alpha^{0,n})\right]$. It is easy to prove that

$$P\left(\inf_{\alpha\in\partial N_n(\delta_n)}\left\|G_n^{-1/2}\{R_n(\alpha) - R_n(\alpha^{0,n})\}\right\| \geq \left\|G_n^{-1/2}R_n(\alpha^{0,n})\right\|\right) \to 1.$$

As matrix G_n is non trivial, so from the lemma 2 it follows that

$$P\left(\exists\alpha \in \partial N_n(\delta_n)\ :\ R_n(\alpha) = 0\right) \to 1,\ \ n \to \infty.$$

Proof

Proof of the second theorem

$$U_n G_n^{-1/2} R_n(\alpha^{0,n}) = U_n \left(\frac{G_n}{n}\right)^{-1/2} \frac{R_n(\alpha^{0,n})}{\sqrt{n}} \xrightarrow{d} Z,$$

where $Z \sim N(0, E_k)$, $n \to \infty$. From the multinomial center limit theorem we can say that

$$U_n G_n^{1/2}(\alpha^{0,n} - \widehat{\alpha}_n) = U_n G_n^{-1/2} R_n(\alpha^{0,n}) + U_n G_n^{-1/2} o_p(1) =$$
$$= Z + o_p(1).$$

- Anderson J.A. (1972). Separate sample logistic regression, Biometrika, 59, no. 1. pp. 19-35.
- Hossain S., Ejaz Ahmed S., Howlader H. (2012). Model selection and parameter estimation of a multinomial logistic regression model // J. Statist. Computation and Simulation. pp. 1–15.
- Khaplanov A. (2013). Asymptotic normality of the estimation of the multivariate logistic regression. Informatics and its app., 7, no. 2, pp. 69–74.
- Liang H. and Du P. (2012). Maximum likelihood estimation in logistic regression models with a diverging number of covariates // Electronic J. of Statist. 6, pp. 1838–1846.
- Rudin W. Principles of mathematical analysis // McGraw-hill book company. New York, Francisco, Toronto, London. 1964.

Thank you for your attention!