Additive hazard Regression mit zensierten Datensätzen

Paul Schulz

19.01.2015

Inhaltsverzeichnis

Ausgangsproblematik

- Wir gehen von zensierten Datensätzen aus.

Ausgangsproblematik

- Wir gehen von zensierten Datensätzen aus.
- Unter zensierten Datensätzen verstehen wir solche, in denen nicht nur durch Ausfall Individuen aus der Studie ausscheiden, sondern auch durch andere Gründen.

Ausgangsproblematik

- Wir gehen von zensierten Datensätzen aus.
- Unter zensierten Datensätzen verstehen wir solche, in denen nicht nur durch Ausfall Individuen aus der Studie ausscheiden, sondern auch durch andere Gründen.
- In diesen Modellen wollen wir die Überlebenswahrscheinlichkeit und die Risikofaktoren schätzen.

Die additive hazard-Funktion

Sei T die Ausfallszeit, C die Abbruchszeit, $\Delta = I(T \leq C)$ der Fehlerindikator, $X = T \wedge C$ die Ausscheidezeit und $Z(\cdot)$ ein Vektor von vorhersagbaren Kovarianzprozessen. Wir beobachten (X_i, Δ_i, Z_i) und nehmen an das die bedingte hazard-Funktion gegeben ist durch:

$$\lambda(t|Z) = \lambda_0(t) + \beta_0^T Z(t) \tag{1}$$

Die additive hazard-Funktion

Sei T die Ausfallszeit, C die Abbruchszeit, $\Delta = I(T \leq C)$ der Fehlerindikator, $X = T \wedge C$ die Ausscheidezeit und $Z(\cdot)$ ein Vektor von vorhersagbaren Kovarianzprozessen. Wir beobachten (X_i, Δ_i, Z_i) und nehmen an das die bedingte hazard-Funktion gegeben ist durch:

$$\lambda(t|Z) = \lambda_0(t) + \beta_0^T Z(t) \tag{1}$$

Die bedingte Hazardfunktion, gibt an mit welcher Wahrscheinlichkeit ein Individuum ausfällt gegeben Z. λ_0 ist der Basehazard, welcher beschreibt wie wahrscheinlich ein Ausfall generell ist und $\beta_0^T Z$ modelliert den Einfluss der Risikofaktoren auf die Ausfallswahrscheinlichkeit.

Darstellung des Zählprozesses

Sei $N_i(t) = I(X_i \leq t, \Delta_i = 1)$ der 'Ausfall beobachtet' -Zählprozess und $Y_i(t) = I(X_i \geq)$ der Risikoindikator. Dann lässt sich $N_i(t)$ nach der Doob-Zerlegung umschreiben zu :

$$N_i(t) = M_i(t) + \int_0^t Y_i(s) \left\{ \lambda_0(s) + \beta_0^T Z_i(s) \right\} ds \qquad (2)$$

,wobei $M_i(t)$ein Martingal und $\int_0^t Y_i(s) \left(\lambda_0(s) + \beta_0^T Z_i(s) \right) ds$ vorhersagbar ist.

Die Schätzfunktion

Wir möchten β_0 mit Hilfe folgender Schätzfunktion annähern:

$$U(\beta) = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} Y_i(t) \left\{ Z_i(t) - \bar{Z}_i(t) \right\} \left\{ dN_i(t) - Y_i(t) \beta_0^T Z_i(t) \, dt \right\}$$

,wobei $\beta \in \mathbb{R}$, $\bar{Z}(t) = \sum_{i=1}^{n} Y_i(t) Z_i(t) / \sum_{i=1}^{n} Y_i(t)$ und τ die maximale Beobachtungszeit ist.

Darstellung der Schätzfunktion

Dies können wir umschreiben zu $U(\beta) = b - V\beta$ mit

$$b = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_i(t) - \bar{Z}(t) \right\} dN_i$$

'Mittlere Abweichung der Parameter von $-\bar{Z}$ zum Ausfallzeitpunkt',

Darstellung der Schätzfunktion

Dies können wir umschreiben zu $U(\beta) = b - V\beta$ mit

$$b = \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \left\{ Z_i(t) - \bar{Z}(t) \right\} dN_i$$

'Mittlere Abweichung der Parameter von $-\bar{Z}$ zum Ausfallzeitpunkt',

$$V = \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau Y_i(t) \left\{ Z_i(t) - \bar{Z}(t) \right\}^{\otimes 2} dt$$

'Empirische Covarianzmatrix aller Aktiven gemittelt über den Beobachtungszeitraum'.

Darstellung der Schätzfunktion

Wir erhalten durch integrieren von $-U(\beta)$ nach β eine Fehlerfunktion

$$L(\beta) = \frac{1}{2}\beta^T V \beta - b^T \beta \tag{3}$$

$L(\beta)$ lässt sich als empirischer Vorhersagefehler interpretieren.

Der Regularisierte Schätzer

Regularisierter Schätzer

$$\hat{\beta} = arg\ min_{\beta \in (R)^p} \left\{ Q(\beta) \equiv L(\beta) + \sum_{i=1}^{p} p_\lambda (|\beta_j|) \right\} \tag{4}$$

Mögliche penalty Funktionen:

- Lasso, L_1-penalty, $p(\Theta) = \Theta, \Theta > 0$.
- SICA, Familie von penalty-Funktionen welche glatt zwischen L_0- und L_1-penalties liegt,

$$\rho(\Theta) = \frac{(a+1)^\Theta}{a+\Theta}, \Theta \leq 0, a > 0 \text{ fest.} \tag{5}$$

Massart, 2000, Theorem 9

Lemma 1

Sei $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit Werten in $\mathbb{R}^N$. Für gewisse reellen Zahlen $a_{i,t}$ und $b_{i,t}$, so dass $a_{i,t} \leq X_{i,t} \leq b_{i,t}$, für alle $i \leq n$ und alle $t \leq N$. Sei

$$Z = sup_{1 \leq t \leq N} \sum_{i=1}^{n} X_{i,t}$$

und definiere $L^2 = sup_{1 \leq t \leq N} \sum_{n}^{i=1} (b_{i,t} - a_{i,t})^2$. Dann gilt für alle positiven x:

$$P\left[Z \geq E\left[Z\right] + x\right] \leq \exp\left(-\frac{x^2}{2L^2}\right).$$

Konzentration von empirischen Matrizen

Lemma 2

Unter Bedingung 1-3, $\mu > 0$ und $\varphi \vee \mu^{-1} = O\left(\sqrt{n}/s\right)$, gibt es Konstanten $D, K > 0$, so dass

$$P\left(\left\|V_{AA}^{-1}\right\|_\infty \geq 2\left\|D_{AA}^{-1}\right\|_\infty \mid \Omega_L\right) \leq s^2 D exp\left(-K\frac{n}{L^4}\left(\frac{1}{\varphi^2 s^2} \wedge 1\right)\right), \tag{6}$$

Konzentration von empirischen Matrizen

Lemma 2

$$P\left(\left\|V_{A^cA}V_{AA}^{-1}\right\|_\infty \geq \underbrace{\left((1-\frac{\alpha}{2})\frac{\rho'(0+)}{\rho'_\lambda(d)}\wedge(2cn^\gamma)\right)}_{\geq\left\|D_{A^cA}D_{AA}^{-1}\right\|_\infty} \mid \Omega_L\right)$$
$$\leq (p-s)sD\exp\left(-K\frac{n}{L^4}\left(\frac{\left(\rho'_\lambda(d)^{-1}\wedge n^\gamma\right)^2}{\varphi^2 s^2}\wedge 1\right)\right) \tag{7}$$
$$+s^2\exp\left(-K\frac{n}{L^4}\left(\frac{1}{\varphi^2 s^2}\wedge 1\right)\right)$$

'V weicht nicht stark von D ab'

Konzentration von empirischen Matrizen

Lemma 2

$$P\left(\Lambda_{min}\left(V\right) \leq \lambda\kappa_0 \middle|\, \Omega_L\right) \leq s^2 \exp\left(-K\frac{n}{L^4}\left(\frac{\mu^2}{s^2} \wedge 1\right)\right) \tag{8}$$

'Der Effekt von V wird nicht durch den Effekt von ρ überlagert'

Charakterisierung des regularisierten Schätzers

Lemma 3

Unter Bedingung 1 ist $\hat{\beta} \in \mathbb{R}^{\rho}$ ein strikter lokaler Minimierer von 4, falls gilt:

$$U_{\hat{A}}\left(\hat{\beta}\right) - \lambda \rho'_{\lambda}\left(\left|\hat{\beta}_{\hat{A}}\right|\right) \circ sgn\left(\hat{\beta}_{\hat{A}}\right) = 0, \tag{9}$$

$$\left\| U_{\hat{A}^C}\left(\hat{\beta}\right)\right\|_{\infty} < \lambda \rho\left(0+\right), \tag{10}$$

$$\Lambda_{min}\left(V_{\hat{A}\hat{A}}\right) > \lambda_{\kappa}\left(\rho_{\lambda}; \hat{\beta}_{\hat{A}}\right) \tag{11}$$

'Bedingung vom Typ Ableitung verschwindet'.

Schwache Oracle Eigenschaft

Satz

Zusätzlich zu Bedingung 1-3 gelte:

$$\frac{\left(\rho_\lambda'(d)^{-1} \wedge n^\gamma\right)^2}{\varphi^2 s^2 (\log p)^{r_1}} \to \infty, \quad \frac{(\varphi^{-1} \wedge \mu)}{s^2 (\log s)^{r_1}} \to \infty, \tag{12}$$

$$\frac{n\lambda^2}{(\log p)^{r_1}} \to \infty, \quad \frac{n^{1-2\gamma}\lambda^2}{(\log s)^{r_1}} \to \infty, \tag{13}$$

$$d \geq c_1 \varphi \lambda \rho'(0+), \tag{14}$$

mit $\mu > 0, r_1 = (r+4)/r$, und $c_1 = 2 + 1/(4c)$. Dann gilt für Konstanten $D, K > 0$, mit Wahrscheinlichkeit mindestens

Schwache Oracle Eigenschaft

Satz

$$1 - D\exp\left(-Kn^{(1/r_1)}\left(\frac{(\varphi^{-1} \wedge \mu)}{s^2} \wedge 1\right)^{(1/r_1)}\right) \tag{15}$$

$$-D\exp\left(-Kn^{(1/r_1)}\left(\frac{(\lambda^2)}{n^2\gamma} \wedge 1\right)^{(1/r_1)}\right) \to 1 \tag{16}$$

existiert ein regulärer Schätzer $\hat{\beta}$, welcher folgende Bedingungen erfüllt:

- (Sparsamkeit) $\hat{\beta}_{A^C} = 0$
- (Abweichung in der L_∞-Norm) $\left\|\hat{\beta} - \beta_{0A}\right\|_\infty \leq c_1\varphi\lambda\rho'(0+)$

Berechnen von $\hat{\beta}$

- Benutze, wie beim Lasso-Schätzer, die Methode des Koordinatenweisen Abstiegs.

Berechnen von $\hat{\beta}$

- Benutze, wie beim Lasso-Schätzer, die Methode des Koordinatenweisen Abstiegs.
- Beachte die penalty-Funktion ist nicht konvex. Nutze die Konvexität von L aus.