UNIVERSITATEA DIN BUCUREȘTI

FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ

MODELUL LINIAR GENERALIZAT

Mocanu Alexandru Gabriel

București, iunie 2025

Cuprins

1 Preliminarii **3**

1.1 Estimatorul de verosimilitate maximă. Introducere 3

1.2 Consistenţa estimatorului de verosimilitate maximă 5

1.3 Normalitatea asimptotică a EVM 9

2 Modelul liniar generalizat **13**

2.1 Introducere 13

2.2 Familia exponenţială 13

2.3 Modelul liniar generalizat 15

2.4 Studiu de caz: regresie logistică 15

Capitolul 1

Preliminarii

1.1 Estimatorul de verosimilitate maximă. Introducere

Fie $(\mathcal{X},\ (\mathbb{P}_\theta)_{\theta\in\Theta})$ un model statistic asociat unui eşantion de volum n, X_1,..., X_n şi fie $\theta^* \in \Theta$ parametrul real. Vom presupune că modelul nostru este identifiabil, i.e. aplicaţia $\theta \to \mathbb{P}_\theta$ este injectivă. Scopul nostru este să găsim un estimator $\widehat{\theta}(X_1, ..., X_n)$ astfel încât $\mathbb{P}_{\hat{\theta}}$ este "apropiată" de $\mathbb{P}_\theta$. Acest demers motivează următoarea definiţie.

Definiţia 1.1.1 (Variaţia totală)

Fie $\mathbb{P}_\theta$ şi $\mathbb{P}_{\theta'}$ două măsuri de probabilitate definite pe un spaţiu măsurabil $(\Omega, \mathcal{F})$. Variaţia totală dintre cele două măsuri este definită ca $\mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'})$=$\sup_{A\in\mathcal{F}} |\mathbb{P}_\theta(A) - \mathbb{P}_{\theta'}(A)|$.

Propoziţia 1.1.2 *Dacă $\mathbb{P}_\theta$ şi $\mathbb{P}_{\theta'}$ au densitate în raport cu măsura de numărare, atunci*

$$TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2}\sum_{x\in\Omega} \left|p_\theta(x) - p_{\theta'}(x)\right|.$$

Dacă $\mathbb{P}_\theta$ şi $\mathbb{P}_{\theta'}$ au densitate în raport cu măsura Lebsesgue, atunci

$$TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2}\int_\Omega |f_\theta(x) - f_{\theta'}(x)| dx$$

Demonstraţie. Vom da demonstraţia pentru cazul continuu, cazul discret fiind complet analog înlocuind integralele cu sume.

Fie $A^* = \{x \in \Omega \mid f_\theta(x) > f_{\theta'}\}$. Avem că

$$\frac{1}{2}\int_\Omega |f_\theta(x) - f_{\theta'}(x)| dx = \frac{1}{2}\int_{A*} (f_\theta(x) - f_{\theta'}(x)) dx + \frac{1}{2}\int_{(A*)^c} (f_{\theta'}(x) - f_\theta)) dx. \tag{1.1}$$

Ştim că $\int_\Omega f_\theta(x)dx = 1$ şi $\int_\Omega f_{\theta'}(x)dx = 1 \Rightarrow \int_\Omega (f_\theta(x) - f_{\theta'}(x))dx = 0$

Aşadar $\int_{(A*)^c}(f_{\theta'}(x) - f_\theta)dx = \int_{A*}(f_\theta(x) - f_{\theta'}(x))dx$. Înlocuind în (1.1) ajungem la

$$\frac{1}{2}\int_\Omega |f_\theta(x) - f_{\theta'}(x)|dx = \int_{A*}(f_\theta(x) - f_{\theta'}(x))dx = |\mathbb{P}_\theta(A^*) - \mathbb{P}_{\theta'}(A^*)| \leq \mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}).$$

O proprietate elementară a integralelor este că $\forall g$ şi $\forall A$, $\int_A g(x)dx \leq \int_{\{x|g(x)\geq 0\}} g(x)dx$, de unde obţinem $\mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2}\int_\Omega |f_\theta(x) - f_{\theta'}(x)|dx$.

■

Observaţia 1.1.3 *Deşi avem formule explicite pentru variaţia totală, în practică este foarte laborios să o minimizăm. O strategie ar fi să găsim un estimator pentru variaţia totală, iar apoi să îl minimizăm în funcţie de θ.*

Definiţia 1.1.4 (Divergenţa Kullback-Leibler) *Fie $\mathbb{P}_\theta$ şi $\mathbb{P}_{\theta'}$ două distribuţii pe un spaţiu $(\Omega, \mathcal{F})$. Divergenţa Kullback-Leibler este*

$$KL(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \begin{cases} \sum_{x\in\Omega} p_\theta(x) log\left(\frac{p_\theta(x)}{p_{\theta'}(x)}\right) & \textit{dacă } \Omega \textit{ e discret,} \\ \int_\Omega f_\theta(x) log\left(\frac{f_\theta(x)}{f_{\theta'}(x)}\right) dx & \textit{dacă } \Omega \textit{ e continuu.} \end{cases}$$

Revenind la modelul nostru statistic, $\mathrm{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) = \mathbb{E}_{\theta^*}\left[\log\left(\frac{p_{\theta*}(X)}{p_\theta(X)}\right)\right] = \mathbb{E}_{\theta^*}[\log(p_{\theta^*}(X))] - \mathbb{E}_{\theta^*}[\log(p_\theta(X))]$.

Fie acum $\mathrm{KL}_n(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) = \mathbb{E}_{\theta^*}[\log(p_{\theta^*}(X))] - \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i)$. Din legea numerelor mari ştim că $\mathrm{KL}_n(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) \xrightarrow{\text{a.s}} \mathrm{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$. Putem deduce principiul verosimilităţii maxime din

$$\underset{\theta\in\Theta}{\text{argmin}}\ \mathrm{KL}_n(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) = \underset{\theta\in\Theta}{\text{argmin}}\ -\frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i) = \underset{\theta\in\Theta}{\text{argmax}}\ \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i).$$

Definiţia 1.1.5 (Funcţia de verosimilitate) *Fie $X_1, ..., X_n$ un eşantion de volum n şi fie $(\mathcal{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ un model statistic asociat. Vom presupune că $\mathbb{P}_\theta$ are densitate f_θ (în raport cu măsura Lebesgue sau măsura de numărare. Funţia de verosimilitate a modelului, $\mathcal{L} : \mathbb{R}^n \times \Theta \to \mathbb{R}$ este definită prin*

$$\mathcal{L}(x_1, ..., x_n, \theta) = \prod_{i=1}^n f_\theta(x_i).$$

Definiţia 1.1.6 (Estimatorul de verosimilitate maximă) *Fie $X_1, ..., X_n$ un eşantion de volum n şi fie $(\mathcal{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ un model statistic asociat. Fie $\theta^* \in \Theta$ parametrul real. Estimatorul de verosimilitate maximă a lui θ^* este definit ca*

$$\underset{\theta\in\Theta}{argmax}\ \mathcal{L}(X_1, ..., X_n, \theta).$$

Introducerea euristică de mai sus lasă loc pentru două întrebări. Este de ajuns ca $\mathrm{KL}_n(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) \xrightarrow{\text{a.s}} \mathrm{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$ pentru ca $\underset{\theta\in\Theta}{\operatorname{argmin}} \mathrm{KL}_n(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) \xrightarrow{\text{a.s}} \underset{\theta\in\Theta}{\operatorname{argmin}} \mathrm{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$? (în general convergenţa punctuală nu implică convergenţa uniformă). Minimizând divergeţa Kullback Leibler, minimizăm variaţia totală? Un răspuns pentru a doua întrebare îl dă inegalitatea lui Pinsker, ce provine din teoria informaţiei.

Teorema 1.1.7 (Inegalitatea lui Pinsker) *Fie $\mathbb{P}$ şi $\mathbb{Q}$ două repartiţii pe un spaţiu măsurabil $(\Omega, \mathcal{F})$. Atunci $\sqrt{\frac{1}{2}KL(\mathbb{P},\mathbb{Q})} \geq TV((\mathbb{P},\mathbb{Q})$.*

1.2 Consistenţa estimatorului de verosimilitate maximă

Teorema 1.2.1 *Să presupunem că $(\mathcal{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ este un model statistic identifiabil şi spaţiul parametrilor Θ este finit. Fie $X_1, ..., X_n \sim \mathbb{P}_{\theta^*}$ iid. În aceste condiţii, estimatorul de verosimilitate maximă $\hat{\theta}_n = \underset{\theta\in\Theta}{argmax} \frac{1}{n}\sum_{i=1}^n logp_\theta(X_i)$ este consistent, i.e $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta^*$.*

Demonstraţie. Vom fixa $\theta \neq \theta^*$. Din legea tare a numerelor mari ştim că

$$Z_n = \frac{1}{n}\sum_{i=1}^n \mathrm{log}p_{\theta^*}(X_i) - \frac{1}{n}\sum_{i=1}^n \mathrm{log}p_\theta(X_i) \xrightarrow{a.s} \mathrm{KL}(\mathbb{P}_{\theta*}, \mathbb{P}_\theta) > 0, \forall\theta \neq \theta^*.$$

Fie acum $A_\theta = \{\omega \mid \lim_{n\to\infty} Z_n(\omega) = \mathrm{KL}(\mathbb{P}_{\theta*}, \mathbb{P}_\theta)\}$. $\mathbb{P}_{\theta^*}(A_\theta) = 1$. Fie $\omega \in A_\theta$ şi fie $\epsilon = \frac{1}{2}\mathrm{KL}(\mathbb{P}_{\theta*}, \mathbb{P}_\theta)$. Din convergenţa punctuală, există $N \in \mathbb{N}$ astfel încât

$$|Z_n(\omega) - \mathrm{KL}(\mathbb{P}_{\theta*}, \mathbb{P}_\theta)| < \frac{1}{2}\mathrm{KL}(\mathbb{P}_{\theta*}, \mathbb{P}_\theta), \forall n \in \mathbb{N}, i.e.$$

$$0 < \frac{1}{2}\mathrm{KL}(\mathbb{P}_{\theta*}, \mathbb{P}_\theta) < Z_n(\omega), \forall n > N.$$

Pentru fiecare $\omega \in A_\theta$, definim $N_\theta(\omega) = \inf\{N \in \mathbb{N} | \forall n > N, Z_n(\omega) > 0\}$. Astfel, N_θ este o v.a. finită a.s. $\forall\theta \neq \theta^*$. Deoarece mulţimea parametrilor Θ este finită, v.a. N definită prin $N(\omega) = \underset{\theta\neq\theta^*}{\max} N_\theta(\omega)$ este finită a.s. Definim $N = \inf\{m | \mathbb{P}(\{\omega | N(\omega) > m\}) = 0\}$.

Există deci un $N \in \mathbb{N}$ astfel încât $\forall n > N, \frac{1}{n}\sum_{i=1}^n \mathrm{log}p_{\theta^*}(X_i) > \frac{1}{n}\sum_{i=1}^n \mathrm{log}p_\theta(X_i)$ a..s $\forall\theta \neq \theta^*$. Rezultă deci că $\underset{\theta\in\Theta}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^n \mathrm{log}p_\theta(X_i) = \theta^*, \forall n > N$. ■

Pentru a demonstra consistenţa în cazul general, avem nevoie de un instrument în plus şi anume o lege uniformă a numerelor mari de tipul Glivenko Cantelli.

Definiţia 1.2.2 *Fie $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{H}$ o mulţime de funcţii $f : \mathbb{R} \to \mathbb{R}$ şi $X_1, ..., X_n$ un eşantion de volum n. Spunem că $\mathcal{H}$ satisface o lege uniformă a numerelor mari dacă $\underset{f\in\mathcal{H}}{sup} |\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]| \xrightarrow{\mathbb{P}} 0$.*

Definiţia 1.2.3 *Fie* (Θ, d) *un spaţiu metric. Pentru* $\epsilon > 0$, *spunem că* $\{\theta^i\}_{i=1}^N$ *este o* ϵ*-acoperire a lui* Θ *dacă* $\forall \theta \in \Theta$, *există un* i *astfel încât* $d(\theta, \theta^i) \leq \epsilon$.

Definiţia 1.2.4 *Numărul de* ϵ*-acoperire a lui* Θ *este definit ca*

$$N(\Theta, d, \epsilon) = inf\{N \in \mathbb{N}^* | \exists \text{ o } \epsilon\text{-acoperire } \{\theta^i\}_{i=1}^N \text{ a lui } \Theta\}.$$

Definiţia 1.2.5 *Fie* $\mathcal{H} \subseteq \{f : \mathcal{X} \to \mathbb{R}\}$, $\mathcal{X} \subseteq \mathbb{R}$ *o mulţime de funcţii şi fie* λ *o măsură pe* $\mathcal{X}$. *O mulţime* $\{[l_i, u_i]\}_{i=1}^N \subset \mathbb{R}^{\mathcal{X}}$ *este un* ϵ*-bracket a lui* $\mathcal{H}$ *în* $L_P(\lambda)$ *dacă*

$$\forall f \in \mathcal{H}, \exists i, a.î.\ l_i \leq f \leq u_i \text{ şi } ||u_i - l_i||_{L_p(\lambda)} \leq \epsilon.$$

Definiţia 1.2.6 *Numărul de bracketing al lui* $\mathcal{H}$ *este*

$$N_{[]}(\mathcal{H}, L_p(\lambda), \epsilon) := inf\{N \in \mathbb{N} | \exists \text{ un } \epsilon\text{-bracket } \{[l_i, u_i]\}_{i=1}^n \text{ al lui } \mathcal{H} \text{ în } L_p(\lambda)\}.$$

Teorema 1.2.7 (Legea uniformă a numerelor mari) *Fie* $X_1, ..., X_n \sim \mathbb{P} \circ X_1^{-1}$ *i.i.d pe un spaţiu de probabilitate* $(\Omega, \mathcal{F}, \mathbb{P})$. *Fie* $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}, \mathcal{X} \subseteq \mathbb{R}$, *o mulţime de funcţii ce satisface* $N_{[]}(\mathcal{H}, L_1(\mathbb{P} \circ X_1^{-1}), \epsilon) < \infty, \forall \epsilon > 0$. *În aceste condiţii*

$$\sup_{f \in \mathcal{H}} |\frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]| \xrightarrow{\mathbb{P}} 0.$$

Demonstraţie. Fie $\epsilon > 0$ şi fie $\{[l_i, u_i]\}_{i=1}^N$ un ϵ-bracket al lui $\mathcal{H}$. Oricare ar fi $f \in \mathcal{H}$, există un j astfel încât $l_j \leq f \leq u_j$, de unde rezultă

$$\frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \leq \frac{1}{n} \sum_{i=1}^n u_j(X_i) - \mathbb{E}[l_j(X)] =$$

$$= \frac{1}{n} \sum_{i=1}^n u_j(X_i) - \mathbb{E}[u_j(X)] + \mathbb{E}[u_j(X)] - \mathbb{E}[l_j(X)] =$$

$$= \frac{1}{n} \sum_{i=1}^n u_j(X_i) - \mathbb{E}[u_j(X)] + ||u_j(X) - l_j(X)||_{L_1(\mathbb{P} \circ X_1^{-1})} \leq$$

$$\leq \frac{1}{n} \sum_{i=1}^n u_j(X_i) - \mathbb{E}[u_j(X)] + \epsilon$$

Aşadar rezultă că

$$\sup_{f \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right) \leq \max_j \left(\frac{1}{n} \sum_{i=1}^n u_j(X_i) - \mathbb{E}[u_j(X)] \right) + \epsilon.$$

În mod similar obţinem că

$$\mathbb{E}[f(X)] - \frac{1}{n}\sum_{i=1}^{n} f(X_i) \leq \mathbb{E}[l_j(X)] - \frac{1}{n}\sum_{i=1}^{n} l_j(X_i) + \epsilon$$

şi deci

$$\sup_{f\in\mathcal{H}}(\mathbb{E}[f(X)] - \frac{1}{n}\sum_{i=1}^{n} f(X_i)) \leq \max_j(\mathbb{E}[l_j(X)] - \frac{1}{n}\sum_{i=1}^{n} l_j(X_i)) + \epsilon.$$

Combinând cele două relaţii,

$$\sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]\right| \leq \max_j\left(\frac{1}{n}\sum_{i=1}^{n} u_j(X_i) - \mathbb{E}[u_j(X)]\right)\vee(\mathbb{E}[l_j(X)]-\frac{1}{n}\sum_{i=1}^{n} l_j(X_i))+\epsilon.$$

Aadar

$$\lim_{n\to\infty}\sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]\right| \leq \epsilon.$$

Definiţia 1.2.8 *Fie $l_\theta(x) = logp_\theta(x)$. Matricea informaţională a lui Fisher este definită ca $I(\theta) = \mathbb{E}_\theta[\nabla l_\theta(X)\nabla l_\theta(X)^T]$, unde $\nabla l_\theta(x) := \left[\frac{\partial}{\partial\theta_j} logp_\theta(x)\right]_{j=1}^{d} \in \mathbb{R}^d$.*

Propoziţia 1.2.9 *Vom presupune că $l_\theta(x)$ satisface următoarele condiţii de regularitate:*

- *modelul este identifiabil*
- *$suppp_\theta(x) = suppp_{\theta^*}(x), \forall\theta \neq \theta^*$*
- *$\forall\theta \in \Theta, \forall x$, derivatele parţiale de ordinul 2 ale lui $p_\theta(x)$ există şi sunt finite*
- *integrala $\int p_\theta(x)dx$ poate fi diferenţiată de două ori sub semnul integral ca funcţie de θ.*

În aceste condiţii $I(\theta) = \mathbb{E}_\theta[\nabla l_\theta(X)\nabla l_\theta(X)^T] = -\mathbb{E}_\theta[\nabla^2 l_\theta(X)]$, unde

$$\nabla^2 l_\theta(x) = \left[\frac{\partial^2}{\partial\theta_i\theta_j} log(p_\theta(x))\right]_{i,j=1}^{d} \in \mathbb{R}^{d\times d}.$$

Demonstraţie.

$$\mathbb{E}_\theta[\nabla^2 l_\theta(X)]_{(i,j)} = \mathbb{E}_\theta\left[\frac{\partial^2 \mathrm{log} p_\theta(X)}{\partial_i\partial\theta_j}\right] = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_j}\left(\frac{\partial \mathrm{log} p_\theta(X)}{\partial\theta_i}\right)\right] = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_j}\left(\frac{\frac{\partial p_\theta(X)}{\partial\theta_i}}{p_\theta(X)}\right)\right] =$$

$$= \mathbb{E}_\theta\left[\frac{\frac{\partial^2 p_\theta(X)}{\partial\theta_i\theta_j}p_\theta(X) - \frac{\partial p_\theta(X)}{\partial\theta_i}\frac{\partial p_\theta(X)}{\partial\theta_j}}{p_\theta(X)^2}\right] = \mathbb{E}_\theta\left[\frac{\frac{\partial^2 p_\theta(X)}{\partial\theta_i\theta_j}}{p_\theta(X)}\right] - \mathbb{E}\left[\frac{\frac{\partial p_\theta(X)}{\partial\theta_i}}{p_\theta(X)}\frac{\frac{\partial p_\theta(X)}{\partial\theta_j}}{p_\theta(X)}\right] =$$

$$= \int \frac{\frac{\partial^2 p_\theta(x)}{\partial\theta_i\theta_j}}{p_\theta(x)} p_\theta(x)dx - I(\theta)_{(i,j)} = \frac{\partial^2}{\partial\theta_i\theta_j}\int p_\theta(x) - I(\theta)_{(i,j)} = -I(\theta)_{(i,j)}.$$

■

Propoziţia 1.2.10 *Fie $l_\theta(x) = logp_\theta(x)$. În condiţiile propoziţiei de mai sus,* $\mathbb{E}_\theta[\nabla l_\theta(X)] = 0$

Demonstraţie. $\mathbb{E}_\theta[\nabla l_\theta(X)]_i = \int \frac{\partial p_\theta(x)}{\partial\theta j} p_\theta(x)dx = \int \frac{\frac{\partial}{\partial\theta j}\log p_\theta(x)}{p_\theta(x)} p_\theta(x)dx = \frac{\partial}{\partial\theta j}\int p_\theta(x)dx = 0$ ■

Ne aflăm acum în poziţia de a demonstra consistenţa EVM în cazul general.

Teorema 1.2.11 *Fie $(\mathcal{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ un model statistic asociat unui eşantion de volum n, $X_1,..., X_n \sim \mathbb{P}_{\theta^*}$ iid. Vom face următoarele presupuneri despre modelul nostru:*

- *modelul este identifiabil*
- $supp p_\theta(x) = supp p_{\theta^*}(x), \forall\theta \neq \theta^*$
- $\Theta \subset \mathbb{R}^d$ *este o muţime compactă.*
- *funcţiile din familia $\mathcal{H} = \{l_\theta(x) = -logp_\theta(x) | \theta \in \Theta\ \}$ sunt $L(x)$ Lipschitz în θ i.e. $\forall x$ şi $\forall\theta_1, \theta_2$ $|l_{\theta_1}(x) - l_{\theta_1}(x)| \leq L(x)||\theta_1 - \theta_2||$.*
- $\mathbb{E}_{\theta^*}[L(X)] < \infty$.
- $\sup\limits_{\theta: d(\theta,\theta^*)\geq\epsilon} \mathbb{E}[logp_\theta(X)] < \mathbb{E}[logp_{\theta^*}(X)]$.

În aceste condiţii, estimatorul de verosimilitate maximă $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta^*$.

Demonstraţie. Mulţimea parametrilor Θ este compactă, deci $\forall\epsilon > 0$ $N(\Theta, ||\cdot||, \epsilon/2) < \infty$. Fie $\{\theta^i\}_{i=1}^N$ o $\epsilon/2$-acoperire a lui Θ. Oricare ar fi $\theta \in \Theta$, există $1 \leq i \leq N$ astfel încât $||\theta - \theta_i|| \leq \epsilon/2$. Din condiţia Lipschitz avem

$$|l_\theta(x) - l_{\theta_i}(x)| \leq L(x)||\theta - \theta_i||, \forall x \in \mathbb{R}.$$

Definim $u_i(x) = l_{\theta_i}(x) + \frac{\epsilon}{2}L(x)$ şi $h_i(x) = l_{\theta_i}(x) - \frac{\epsilon}{2}L(x)$. Rezultă că

$$h_i(x) \leq l_\theta(x) \leq u_i(x),$$

$\forall x \in \mathbb{R}$ şi $\mathbb{E}[|u_i(X) - h_i(X)|] \leq \epsilon\mathbb{E}[L(X)]$. Am obţinut că $N_{[]}(\mathcal{H}, L_1(\mathbb{P}_{\theta^*}), \epsilon\mathbb{E}[L(X)]) \leq N(\Theta, ||\cdot||, \epsilon/2) < \infty$.

Putem concluziona deci că $\mathcal{H}$ satisface legea uniformă a numerelor mari, i.e.

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\log p_\theta(X_i)-\mathbb{E}_{\theta^*}[\log p_\theta(X)]\right|\xrightarrow{\mathbb{P}}0.$$

Avem că

$$\mathbb{E}[\log p_{\theta^*}(X)]-\mathbb{E}[\log p_{\hat{\theta}_n}(X)]=(\mathbb{E}[\log p_{\theta^*}(X)]-\frac{1}{n}\sum_{i=1}^{n}\log p_{\theta^*}(X_i))+$$

$$+\left(\frac{1}{n}\sum_{i=1}^{n}\log p_{\theta^*}(X_i)-\frac{1}{n}\sum_{i=1}^{n}\log p_{\hat{\theta}_n}(X_i)\right)+$$

$$+\left(\frac{1}{n}\sum_{i=1}^{n}\log p_{\hat{\theta}_n}(X_i)-\mathbb{E}[\log p_{\hat{\theta}_n}(X)]\right).$$

Primul termen converge la 0 în probabilitate din legea slabă a numerelor mari, al doilea este negativ din modul cum este definit $\hat{\theta}_n$ iar pentru al treilea vom folosi legea uniformă a numerelor mari. Aşadar

$$\mathbb{E}[\log p_{\theta^*}(X)]-\mathbb{E}[\log p_{\hat{\theta}_n}(X)]\leq(\mathbb{E}[\log p_{\theta^*}(X)]-\frac{1}{n}\sum_{i=1}^{n}\log p_{\theta^*}(X_i))+0+$$

$$+\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\log p_\theta(X_i)-\mathbb{E}[\log p_\theta(X)]\right|\xrightarrow{\mathbb{P}}0.$$

Din ipoteză ştim că $\sup_{\theta:d(\theta,\theta^*)\geq\epsilon}\mathbb{E}[\log p_\theta(X)]<\mathbb{E}[\log p_{\theta^*}(X)]$. Deci $\forall\epsilon,\exists\delta$ astfel încât $d(\theta,\theta^*)\geq\epsilon\Rightarrow\mathbb{E}[\log p_\theta(X)]<\mathbb{E}[\log p_{\theta^*}(X)]-\delta$. Avem incluziunea de evenimente $\{d(\hat{\theta}_n,\theta^*)\geq\epsilon\}\subseteq\{\mathbb{E}[\log p_{\hat{\theta}_n}(X)]<\mathbb{E}[\log p_{\theta^*}(X)]-\delta\}$. Rezultă că

$$\mathbb{P}_{\theta^*}(\{d(\hat{\theta}_n,\theta^*)\geq\epsilon\})\leq\mathbb{P}_{\theta^*}(\{\mathbb{E}[\log p_{\hat{\theta}_n}(X)]<\mathbb{E}[\log p_{\theta^*}(X)]-\delta\})\to0.$$

Am demonstrat astfel că $\hat{\theta}_n\xrightarrow{\mathbb{P}}\theta^*$.

1.3 Normalitatea asimptotică a EVM

Lema 1.3.1 *Fie $f:\mathbb{R}^d\to\mathbb{R}^k$, diferenţiabilă. Presupunem că Df este L-Lipschitz, i.e.*

$$\|Df(x)-Df(y)\|_{op}\leq L\|x-y\|_2.$$

În aceste condiţii

$$f(y)=f(x)+Df(x)(y-x)+R(y-x),$$

unde R este o matrice de rest (care depinde de x, y) și satisface:

$$\|R\|_{op} \le \frac{L}{2}\|y-x\| \quad \text{și} \quad \|R(y-x)\| \le \frac{L}{2}\|y-x\|^2.$$

Demonstraţie.

Definim $\phi_i(t) = f_i((1-t)x+ty)$, $\phi_i : [0,1] \to \mathbb{R}$. Observăm că $\phi_i(0) = f_i(x)$, $\phi_i(1) = f_i(y)$ și

$$\phi_i'(t) = \left(\nabla f_i((1-t)x+ty)\right)^T (y-x).$$

Atunci:

$$Df((1-t)x+ty)(y-x) = \begin{bmatrix} \nabla f_1^T((1-t)x+ty) \\ \nabla f_2^T((1-t)x+ty) \\ \vdots \\ \nabla f_k^T((1-t)x+ty) \end{bmatrix} (y-x) = \begin{bmatrix} \phi_1'(t) \\ \phi_2'(t) \\ \vdots \\ \phi_k'(t) \end{bmatrix}.$$

Deoarece $\phi_i(1) - \phi_i(0) = \int_0^1 \phi_i'(t)dt$, rezultă:

$$f(y) - f(x) = \int_0^1 Df((1-t)x+ty)(y-x)dt =$$

$$= \int_0^1 \left(Df((1-t)x+ty) - Df(x)\right)(y-x)dt + Df(x)(y-x).$$

Pentru a mărgini termenul de rest, avem:

$$\left\|\int_0^1 \left(Df((1-t)x+ty) - Df(x)\right)(y-x)dt\right\| \le \int_0^1 \left\|\left(Df((1-t)x+ty) - Df(x)\right)(y-x)\right\| dt \le$$

$$\le \int_0^1 \|Df((1-t)x+ty) - Df(x)\|_{op} \cdot \|y-x\|dt,$$

$$\le \int_0^1 L\|(1-t)x+ty-x\|\cdot\|y-x\|dt = \int_0^1 Lt\|y-x\|\cdot\|y-x\|dt = L\|y-x\|^2 \int_0^1 tdt = \frac{L}{2}\|y-x\|^2.$$

Teorema 1.3.2 (Normalitate asimptotică) *Fie $(\mathcal{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ un model statistic asociat unui eşantion de volum n, $X_1,\ldots, X_n \sim \mathbb{P}_{\theta^*}$ iid. Vom face următoarele presupuneri despre modelul nostru:*

- $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta^*$, *împreună cu tot ce presupune*

- *matricea Hessiană este $M(x)$-Lipschitz, i.e.*

$$||\nabla^2 l_{\theta_1} - \nabla^2 l_{\theta_2}||_{op} \leq M(x)||\theta_1 - \theta_2||_2,\ \mathbb{E}_{\theta^*}[M(X)] < \infty.$$

- *$I(\theta^*)$ este inversabilă*
- *$\forall\theta \in \Theta, \forall x$, derivatele parţiale de ordinul 2 ale lui $p_\theta(x)$ există şi sunt finite*
- *integrala $\int p_\theta(x)dx$ poate fi diferenţiată de două ori sub semnul integral ca funcţie de θ.*

În aceste condiţii $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} N(0, I(\theta^*)^{-1})$.

Demonstraţie. Fie estimatorul de verosimilitate maximă $\hat{\theta} = \underset{\theta\in\Theta}{\text{argmax}} \sum_{i=1}^n \text{log} p_\theta(X_i)$ şi fie funcţia $\nabla_\theta(x) : \mathbb{R}^d \to \mathbb{R}^d$. Dezvoltăm în serie Taylor în jurul lui θ^* gradientul fiecărui termen din sumă şi obţinem

$$\nabla l_{\hat{\theta}_n}(X_i) = \nabla l_{\theta^*}(X_i) + \nabla^2 l_{\theta^*}(X_i)(\hat{\theta}_n - \theta^*) + \hat{r}(X_i)(\hat{\theta}_n - \theta^*), i = 1, ..., n,$$

unde din lema 1.3 $||\hat{r}(x)||_{op} \leq M(x)||\hat{\theta}_n - \theta^*||$. Adunând cele n relaţii şi folosind liniaritatea gradientului avem

$$0 = \nabla\left(\frac{1}{n}\sum_{i=1}^n l_{\hat{\theta}_n}(X_i)\right) = \frac{1}{n}\sum_{i=1}^n \nabla l_{\hat{\theta}_n}(X_i) = \frac{1}{n}\sum_{i=1}^n \nabla l_{\theta^*}(X_i) + \frac{1}{n}\sum_{i=1}^n \nabla^2 l_{\theta^*}(X_i)(\hat{\theta}_n - \theta^*) +$$

$$+\frac{1}{n}\sum_{i=1}^n \hat{r}(X_i)(\hat{\theta}_n - \theta^*).$$

Aplicăm lema 1.3 şi ajungem la

$$||\sum_{i=1}^n \hat{r}(X_i)||_{op} \leq \frac{1}{n}\sum_{i=1}^n M(X_i)||\hat{\theta}_n - \theta^*|| = o_P(1).$$

Înlocuind în expresia de mai sus

$$0 = \frac{1}{n}\sum_{i=1}^n \nabla l_{\theta^*}(X_i) + \frac{1}{n}\sum_{i=1}^n \nabla^2 l_{\theta^*}(X_i)(\hat{\theta}_n - \theta^*) + o_P(1)(\hat{\theta}_n - \theta^*) =$$

$$= \frac{1}{n}\sum_{i=1}^n \nabla l_{\theta^*}(X_i) + (\mathbb{E}[\nabla^2 l_{\theta^*}] + (\frac{1}{n}\sum_{i=1}^n \nabla^2 l_{\theta^*}(X_i) - \mathbb{E}[\nabla^2 l_{\theta^*}]) + o_P(1))(\hat{\theta}_n - \theta^*).$$

Din legea numerelor mari $\frac{1}{n}\sum_{i=1}^n \nabla^2 l_{\theta^*}(X_i) - \mathbb{E}[\nabla^2 l_{\theta^*}] = o_P(1)$ şi înmulţind cu $\sqrt{n}$ ambele părţi ale egalităţii obinem

$$\sqrt{n}(\mathbb{E}[\nabla^2 l_{\theta^*}] + o_P(1)) = -\sqrt{n}\frac{1}{n}\sum_{i=1}^n \nabla l_{\theta^*}(X_i).$$

Ştim din teorema limită centrală că

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\nabla l_{\theta^*}(X_i) \xrightarrow{d} N(0, I(\theta^*)),$$

deci putem concluziona din teorema lui Slutsky că

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} N(0, (\mathbb{E}[\nabla^2 l_{\theta^*}])^{-1} I(\theta^*)^{-1} \mathbb{E}[\nabla^2 l_{\theta^*}])^{-1}) = N(0, I(\theta^*)^{-1}).$$

■

Capitolul 2

Modelul liniar generalizat

2.1 Introducere

Într-un model liniar clasic presupunem că dependenţa dintre variabila de răspuns şi variabila predictor este de forma $Y_i = X_i^T\beta + \epsilon_i$ sau în formă matriceală $\mathbb{Y} = \mathbb{X}\beta + \epsilon$. Presupunerile pe care le facem asupra modelului sunt următoarele:

- $Y_i|X_i \sim N(X_i^T\beta, \sigma^2) = N(\mu(X_i), \sigma^2)$,
- $\mathbb{E}[Y_i|X_i] = \mu(X_i) = X_i^T\beta$.

Într-un model liniar generalizat, vom generaliza aceste componente în felul următor:

- $Y_i|X_i \sim$ o distribuţie din familia exponenţială
- $g(\mu(X_i)) = X_i^T\beta$ sau, mai intuitiv $\mu(X_i) = g^{-1}(X_i^T\beta.)$ (funcţia de legătură), unde $\mu(X_i) = \mathbb{E}[Y_i|X_i]$.

Exemplul 2.1.1 *Avem un set de date format din măsurători făcute pe 81 de copii post operaţie pentru a corecta cifoza. Variabilele predictor sunt: vârsta, numărul de vertebre pe care s-a operat şi numărul vertebrei cele mai de sus pe care s-a operat. Variabila răspuns ne spune dacă au existat sau nu degenerări post operaţie.*

Variabila de răspuns este binară deci suntem obligaţi să modelăm Y_i ca $Y_i|X_i \sim$ Bernoulli(p). Avem că $\mu(X_i) \in (0,1)$ deci nu putem spune că $\mu(X_i) = X_i^T\beta$. Avem nevoie de o funcţie g inversabilă astfel încât $g(X_i^T\beta) \in (0,1)$.

2.2 Familia exponenţială

Definiţia 2.2.1 *O familie de repartiţii $\{\mathbb{P}_\theta, \theta \in \Theta \subseteq \mathbb{R}^k\}$ face parte din familia exponenţială cu k parametrii dacă densitatea p_θ în raport cu o măsură μ pe $\mathcal{X} \subset \mathbb{R}^q$ satisface:*

- $S_p = \{x|p_\theta(x) > 0\}$ *nu depinde de θ.*

- *există funcţii reale* $g_1, ..., g_k, B : \Theta \to \mathbb{R}$ *şi* $T_1, ..., T_k, h : \mathcal{X} \to \mathbb{R}, h(x) \geq 0$ *astfel încât* f_θ *să poată fi scrisă sub forma*

$$p_\theta(x) = h(x) \exp\left(\sum_{i=1}^{k} g_i(\theta) T_i(x) - B(\theta)\right), \quad \forall x \in \mathcal{X} \subset \mathbb{R}^q.$$

Definiţia 2.2.2 *Spunem că familia de repartiții* $\{\mathbb{P}_\theta,\ \theta \in \Theta \subset \mathbb{R}\}$ *face parte din Familia Exponențială cu un parametru dacă densitatea* $p_\theta(x)$ *în raport cu o măsură* μ *verifică:*

- $S_p = \{x \mid p_\theta(x) > 0\}$ *nu depinde de* θ
- *Există funcțiile* $g,\ B : \Theta \to \mathbb{R}$ *și* $T,\ h : \mathcal{X} \to \mathbb{R},\ h(x) \geq 0,$ *astfel încât densitatea* $p_\theta(x)$ *să poată fi scrisă sub forma*

$$p_\theta(x) = h(x) e^{g(\theta)T(x) - B(\theta)}$$

Spațiul parametrilor

$$\Theta_0 = \left\{\theta : \int h(x)\ \exp\left(g(\theta)\, T(x)\right) d\mu(x) < \infty\right\}$$

este cel mai mare spațiu pentru care densitatea din Familia Exponențială cu un parametru există.

O reparametrizare importantă este $\eta = g(\theta)$ și Familia Exponențială devine

$$p_\eta(x) = h(x)\ \exp\left\{\eta\, T(x) - A(\eta)\right\}$$

unde $A(\eta) = \log \int h(x)\ \exp\left(\eta\, T(x)\right) d\mu(x)$. Această reprezentare se numește *reprezentarea canonică* iar η este *parametrul canonic.*

Mulțimea

$$\Theta_\eta = \left\{\eta : \int h(x)\ \exp\left(\eta\, T(x)\right) d\mu(x) < \infty\right\}$$

se numește *spațiul natural al parametrilor.*

Propoziţia 2.2.3
- $\mathbb{E}[T(X)] = A'(\eta)$,
- $Var(T(X)) = A''(\eta)$.

Demonstraţie. Fie $l_\eta(x) = \log p_\eta(X) = \log(h(X)) + \eta T(X) - A(\eta)$. Derivând avem că $\frac{\partial l_\eta}{\partial \eta} = T(X) - A'(\eta)$. Din 1.2.10 reiese că $0 = \mathbb{E}\left[\frac{\partial l_\eta}{\partial \eta}\right] = \mathbb{E}[T(X)] - A'(\eta)$, de unde rezultatul dorit.

Pe de altă parte $\frac{\partial^2 l_\eta}{\partial l_\eta^2} + \left(\frac{\partial l_\eta}{\partial \eta}\right)^2 = -A''(\eta) + (T(X) - A'(\eta))^2 = -A''(\eta) + (T(X) - \mathbb{E}[T(X)])^2$. Aplicând 1.2.9 avem că $Var(T(X)) = A''(\eta)$.

Vom nota de acum înainte parametrul η din densitatea familiei exponenţiale canonice cu θ.

2.3 Modelul liniar generalizat

Fie $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}, i = 1, ..., n$ perechi independente astfel încât

- $Y_i|X_i$ are densitate $p_{\theta_i}(y|x) = h(y)\, \exp\{\theta_i\, T(y) - A(\theta_i)\}\,, T(y) = y$,
- $\mathbb{E}[Y_i|X_i] = \mu_i = g^{-1}(X_i^T\beta)$, unde g este o funcţie de legătură strict monotonă şi diferenţiabilă.

Definiţia 2.3.1 *Vom numi funcţie de legătură canonică o funcţie g strict monotonă şi diferenţiabilă ce satisface $g(\mu_i) = \theta_i$. Dar $\mu_i = \mathbb{E}[T(Y)] = \mathbb{E}[Y] = A'(\theta_i)$. Rezultă că $g(x) = (A')^{-1}(x)$.*

Parametrul de interes în modelul nostru este β. În consecinţă, trebuie să găsim un mod de a îl lega de θ_i, pentru a îl putea estima din funcţia de verosimilitate. Avem următoarea relaţie între cele două:

$$\theta_i = (A')^{-1}(\mu_i) = (A')^{-1}(g^{-1}(X_i^T\beta)) := j(X_i^T\beta).$$

Dacă g este funcţia de legătură canonică, j este identitatea. Funcţia de log-verosimilitate pentru eşantionul $Y_1, ..., Y_n$ este

$$\mathcal{L}(\theta; Y|X) = \sum_{i=1}^{n} \log(h(Y_i)) + Y_i\theta_i - A(\theta_i) = \sum_{i=1}^{n} \log(h(Y_i)) + Y_i j(X_i^T\beta) - A(j(X_i^T\beta)).$$

Ignorând termenii care nu depind de θ şi folosind funcţia de legătură canonică, obţinem

$$\mathcal{L}(\theta; Y|X) = \sum_{i=1}^{n} Y_i X_i^T\beta - A(X_i^T\beta).$$

2.4 Studiu de caz: regresie logistică

Să presupunem că $Y_i|X_i \sim Bernoulli(p_i), i = 1, ..., n$ independente şi X_i este un vector predictor de dimensiune $p \times 1$.

În acest caz $p_{\theta_i}(y|x) = p_i^y(1-p_i)^{1-y} = \exp\,\{y\log p_i + (1-y)\log(1-p_i) = \exp\,\{\log\left(\frac{p_i}{1-p_i}\right)y - \log\left(\frac{1}{1-p_i}\right)\}$. Rezultă că

$$\theta_i = \log\left(\frac{p_i}{1-p_i}\right), p_i = \frac{1}{1+e^{-\theta_i}}, A(\theta_i) = \log\left(\frac{1+e^{-\theta_i}}{e^{-\theta_i}}\right).$$

Mai ştim că $\mathbb{E}[Y_i|X_i] = p_i$, de unde obţinem că funcţia de legătură canonică este $g(x) = \log\left(\frac{x}{1-x}\right)$ (cuantila distribuţiei logistice).

Funcţia de log-verosimilitate este dată de

$$\mathcal{L}(\theta; Y|X) = \sum_{i=1}^{n} Y_i\theta_i - \log\left(\frac{1+e^{-\theta_i}}{e^{-\theta_i}}\right) = \sum_{i=1}^{n} Y_i X_i^T\beta - \log\left(\frac{1+e^{-X_i^T\beta}}{e^{-X_i^T\beta}}\right) =$$

$$\sum_{i=1}^{n} Y_i X_i^T\beta - \log(1+e^{X_i^T\beta}).$$

Putem face schimbarea de variabilă $Z_i = 2Y_i - 1 \in \{-1, 1\}$ şi obţinem

$$\mathcal{L}(\theta; Y|X) = -\sum_{i=1}^{n} \log(1+e^{-Z_i X_i^T\beta}) = \sum_{i=1}^{n} \log\left(\frac{1}{1+e^{-Z_i X_i^T\beta}}\right).$$

Observaţia 2.4.1 *Considerăm funcția*

$$\ell_\theta(x,y) = \log(1+\exp(-y\, x^T\theta)).$$

Observăm că funcția

$$\phi(t) = \log(1+e^{-t})$$

este 1-Lipschitz, deoarece $\left|\phi'(t)\right| = \frac{e^{-t}}{1+e^{-t}} \leq 1$. *Prin urmare, pentru orice* $\theta_1, \theta_2 \in \Theta$ *și orice* (x, y),

$$\left|\ell_{\theta_1}(x,y) - \ell_{\theta_2}(x,y)\right| = \left|\phi(y\, x^T\theta_1) - \phi(y\, x^T\theta_2)\right| \leq \left|y\, x^T(\theta_1-\theta_2)\right| \leq \|x\|\,\|\theta_1-\theta_2\|,$$

unde ultima inegalitate rezultă din inegalitatea Cauchy–Schwarz.

Aplicând teorema 1.2.11 pentru clasa de funcții Lipschitz $(\ell_\theta)_{\theta\in\Theta}$ *cu* $L(x) = \|x\|$, *dacă mulțimea de parametri* Θ *este compactă și* X *are momentul de ordin I finit, rezultă că*

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\ell_\theta(X_i,Y_i) - \mathbb{E}[\ell_\theta(X,Y)]\right| \xrightarrow{\mathbb{P}} 0.$$

Verificând şi celelalte condiţii de regularitate putem deduce că estimatorul de verosimilitate maximă din regresia logistică $\widehat{\beta}_n \xrightarrow{\mathbb{P}} \beta^*$.

Observaţia 2.4.2 *Fie*

$$\ell_\theta(x,y) = \log(1+e^{-y\, x^T\theta}), \qquad u(\theta) = y\, x^T\theta, \qquad \phi(t) = \log(1+e^{-t}).$$

Atunci, aplicând succesiv regula lanţului obţinem

$$\nabla_\theta^2 \ell_\theta(x,y) = \phi''(u(\theta))\; x\, x^T,$$

unde

$$\phi''(t) = \frac{e^{-t}}{(1+e^{-t})^2}, \quad \phi'''(t) = \frac{e^{-t}(1-e^{-t})}{(1+e^{-t})^3},$$

și $\sup_t |\phi'''(t)| = \frac{1}{4}$. *Am obţinut că derivata a doua a lui* ϕ *este Lipschitz, şi deci*

$$\|\nabla^2\ell(\theta_1) - \nabla^2\ell(\theta_2)\|_{op} = \left|\phi''(u(\theta_1)) - \phi''(u(\theta_2))\right| \|x\, x^T\|_{op} \leq \frac{1}{4}\,|u(\theta_1) - u(\theta_2)|\,\|x\|^2.$$

Dar

$$|u(\theta_1) - u(\theta_2)| = |y\, x^T(\theta_1 - \theta_2)| \leq \|x\|\,\|\theta_1 - \theta_2\|.$$

Prin urmare

$$\|\nabla^2\ell(\theta_1) - \nabla^2\ell(\theta_2)\|_{op} \leq \frac{1}{4}\,\|x\|^3\,\|\theta_1 - \theta_2\|,$$

adică Hessiana este Lipschitz în θ *cu*

$$M(x) = \frac{1}{4}\,\|x\|^3.$$

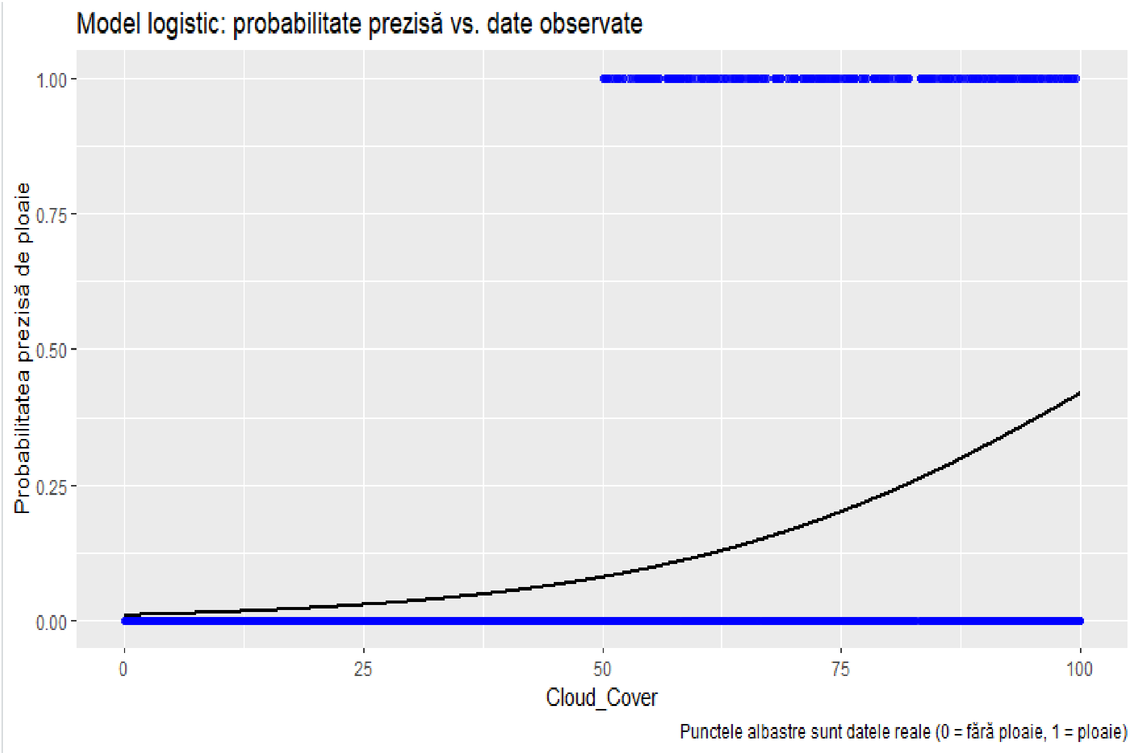
Putem aplica teorema 1.3.2 pentru a concluziona că $\sqrt{n}(\hat{\beta}_n - \beta^*) \xrightarrow{d} N(0, I(\beta^*)^{-1})$.

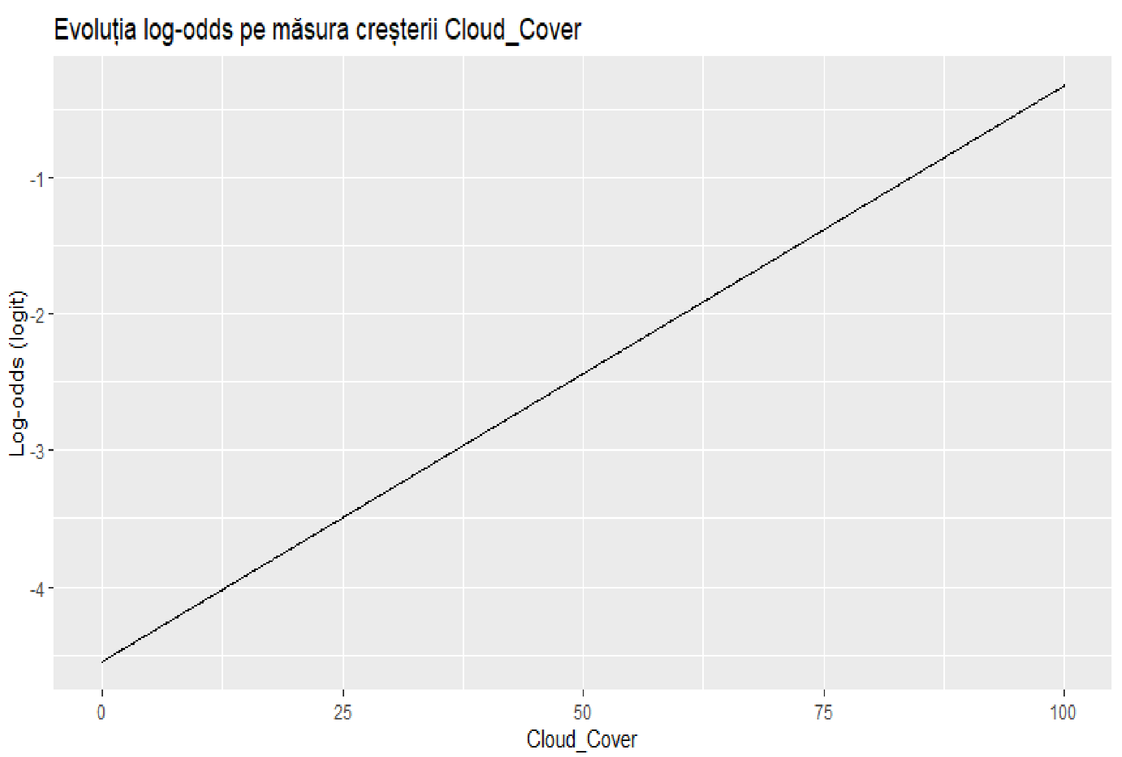
Exemplul 2.4.3 *Setul de date weatherforecastdata.csv conţine observaţii despre vreme pentru un număr de zile. Vrem să vedem care este relaţia între cât de înorat este afară şi dacă plouă sau nu. Este natural să modelăm răspunsul binar cu o distribuţie Bernoulli şi să apliăm regresia logistică.*

```
getwd()

library(dplyr)
library(ggplot2)

setwd("C:\\Users\\alexandru\\Documents\\seturi_de_date")

forecast = read.csv(
  "C:\\Users\\alexandru\\Documents\\seturi_de_date\\weather_forecast_
     data.csv"
)

# Creeaza o coloana noua, RainTodayBinary, care este 1 daca în
   coloana Rain apare textul "rain" si 0 în rest
forecast$RainTodayBinary <- ifelse(forecast$Rain == "rain", 1, 0)

head(forecast)
```

```r

# Scatterplot intre Cloud_Cover si variabila binara:
# - Cloud_Cover pe axa x
# - RainTodayBinary (0 sau 1) pe axa y
forecast %>%
  ggplot() +
  geom_point(aes(x = Cloud_Cover, y = RainTodayBinary))

# Fit-uim un model de regresie logistica (GLM cu link logit), unde
    explicativa este Cloud_Cover, iar raspunsul este RainTodayBinary:
    RainTodayBinary ~ Cloud_Cover
forecast_glm <- glm(
  RainTodayBinary ~ Cloud_Cover,
  data   = forecast,
  family = binomial(link = "logit")
)

# Construim un nou data-frame pentru a evalua modelul in valori ale
    Cloud_Cover de la 0 la 100, cu pas de 0.1
glm_predictions = data.frame(
  "Cloud_Cover" = seq(0, 100, by = 0.1)
)

# Prezicem log-odds
glm_predictions$LogOdds = predict(
  forecast_glm,
  newdata = glm_predictions,
  type = "link"
)

# Prezicem probabilitatea  p = E[RainTodayBinary | Cloud_Cover]
glm_predictions$PredictedProb = predict(
  forecast_glm,
  newdata = glm_predictions,
  type = "response"
)

# Grafic al log-odds in functie de Cloud_Cover
glm_predictions %>%
  ggplot() +
  geom_line(aes(x = Cloud_Cover, y = LogOdds)) +
```

```
  labs(
    y = "Log-odds (logit)",
    title = "Evolutia log-odds pe masura cresterii Cloud_Cover"
  )

# Grafic al probabilitatii prezise versus observatiile reale:
glm_predictions %>%
  ggplot() +
  geom_line(aes(x = Cloud_Cover, y = PredictedProb), size = 1) +
  geom_point(
    data = forecast,
    aes(x = Cloud_Cover, y = RainTodayBinary),
    color = "blue",
    alpha = 0.6
  ) +
  labs(
    y = "Probabilitatea prezisa de ploaie",
    title = "Model logistic: probabilitate prezisa vs. date observate",
    caption = "Punctele albastre sunt datele reale (0 = fara ploaie, 1 = ploaie)"
  )

# Afiseaza rezumatul modelului glm:


summary(forecast_glm)
```


Model logistic: probabilitate prezisă vs. date observate
Probabilitatea prezisă de ploaie
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Cloud_Cover
Punctele albastre sunt datele reale (0 = fără ploaie, 1 = ploaie)


Evoluția log-odds pe măsura creșterii Cloud_Cover
Log-odds (logit)
-1
-2
-3
-4
0
25
50
75
100
Cloud_Cover

Bibliografie

[1] Aad Van der Vaart, Asymptotic Statistics, Cambridge University Press, Cambridge, 2012

[2] Aad Van der Vaart, Weak Convergence and Empirical Processes, Springer, 1996

[3] Peter McCullagh, John A. Nelder, Generalized Linear Models, Chapman and Hall, 1989