Гипотеза порождения данных

Грабовой Андрей Валериевич

Московский физико-технический институт

ВЦ РАН, Москва, 2018

Литература

- *S. G. Self and R. H. Mauritsen* Power/sample size calculations for generalized linear models // Biometrics, 1988. Vol. 44. P. 79–86.
- On power and sample size calculations for likelihood ratio tests in generalized linear models // Biometrics, 2000. Vol. 56. P. 1192–1196.
- On power and sample size calculations for Wald tests in generalized linear models // Journal of Statistical Planning and Inference, 2005. Vol. 128. P. 43–59.
- Кобзарь А. И. 2006. Прикладная математическая статистика. Москва. 816 с.

Обобщенно линейная модель

Пусть задана выборка размера m:

$$\mathfrak{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{m},$$

где $\mathbf{x} \in \mathbb{R}^n$, $y \in \mathbb{Y}$.
Пусть модель порождающая данные задается в следующем виде:

$$y = g(\mathbf{x}, \mathbf{w}, \beta),$$

где α некоторый параметр, который отвечает за шум в данных.

- Линейная регрессия $\mathbb{Y} = \mathbb{R}$:

$$g(\mathbf{x}, \mathbf{w}, \beta) = f(\mathbf{x}, \mathbf{w}) + \nu, \quad f(\mathbf{x}, \mathbf{w}) = \mathbf{x}^{\mathsf{T}}\mathbf{w}, \quad \nu \sim \mathcal{N}(0, \beta\mathbf{I}).$$

- Логистическая регрессия $\mathbb{Y} = \{0, 1\}$:

$$g(\mathbf{x}, \mathbf{w}, \beta) = \mathcal{B}e(f(\mathbf{x}, \mathbf{w})), \quad f(\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})}.$$

Линейная регрессия

Пусть задана выборка размера m:

$$\mathfrak{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{m},$$

где $\mathbf{x} \in \mathbb{R}^n$, $y \in \mathbb{R}$.
Пусть модель задается в следующем виде:

$$y = \mathbf{X}\mathbf{w} + \nu, \qquad \nu \sim \mathcal{N}(0, \beta\mathbf{I}).$$

- Базовый подход:

$$\hat{\mathbf{w}} = \left(\frac{1}{\beta}\mathbf{X}^\mathsf{T}\mathbf{X} + \frac{1}{\alpha}\mathbf{I}\right)^{-1} \frac{1}{\beta}\mathbf{X}^\mathsf{T}\mathbf{y}.$$

- Байесовский подход:

$$\mathbf{w} \sim \mathcal{N}\left(\hat{\mathbf{w}}, \mathbf{H}^{-1}\right), \quad \mathbf{H} = \frac{1}{\beta}\mathbf{X}^\mathsf{T}\mathbf{X} + \frac{1}{\alpha}\mathbf{I}.$$

Логистическая регрессия

Пусть задана выборка размера m:

$$\mathfrak{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{m},$$

где $\mathbf{x} \in \mathbb{R}^n$, $y \in \{0, 1\}$.
Пусть модель задается в следующем виде:

$$p(y = 1) = \frac{1}{1 + \exp(-\mathbf{x}^\mathsf{T}\mathbf{w})}.$$

- Базовый подход:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w} \in \mathbb{W}} L(\mathfrak{D}, \mathbf{w}),$$

  где $L(\mathfrak{D}, \mathbf{w})$ — правдоподобие выборки.
- Байесовский подход (аппроксимация Лапласа):

$$\mathbf{w} \sim \mathcal{N}\left(\hat{\mathbf{w}}, \mathbf{H}^{-1}\right), \quad \mathbf{H} = \frac{1}{\alpha}\mathbf{I} + \sum_{i=1}^{m} \lambda(\mathbf{x}_i, \hat{\mathbf{w}})(1 - \lambda(\mathbf{x}_i, \hat{\mathbf{w}}))\mathbf{x}_i\mathbf{x}_i^\mathsf{T}.$$

Определения размера выборки

Пусть задана выборка размера m:

$$\mathfrak{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{m},$$

где $\mathbf{x} = [\mathbf{u}, \mathbf{v}] \in \mathbb{R}^n$, $y \in \mathbb{Y}$.
Пусть модель порождающая данные задается в следующем виде:

$$y = g(\mathbf{x}, \mathbf{w}, \beta), \quad p(y|\mathbf{u}, \mathbf{v}, \mathbf{w}_u, \mathbf{w}_v) = \exp\left(\frac{y\theta - b(\theta)}{a(\psi)} + c(y, \psi)\right)$$

где α некоторый параметр, который отвечает за шум в данных.

- Линейная регрессия $\mathbb{Y} = \mathbb{R}$:

$$\theta = \mathbf{x}^{\mathsf{T}}\mathbf{w}, \quad a = 2\beta, \quad b(\theta) = \theta^2.$$

- Логистическая регрессия $\mathbb{Y} = \{0, 1\}$:

$$\theta = \mathbf{x}^{\mathsf{T}}\mathbf{w}, \quad a = 1, \quad b(\theta) = -\ln(1 - \sigma(\theta)).$$

Определения размера выборки

Пусть вектор параметров имеет следующее распределение:

$$\mathbf{w} \sim \mathcal{N}(\mathbf{m}, \mathbf{V}), \quad \mathbf{V} = \mathbf{I}^{-1}(\mathfrak{D}, \mathbf{m}),$$

где $\mathbf{I}(\mathfrak{D}, \mathbf{m})$ — информационная матрица Фишера.
Рассмотрим гипотезу:

$$H_0 : \mathbf{m}_u = \mathbf{m}_u^0, \quad \mathbf{m}_u \neq \mathbf{m}_u^0.$$

Рассмотрим следующие статистики:

- Lagrange test:
- Likelihood ratio test:
- Wald test:

Задав α и β можем получить оценки на достаточный объем выборки m^*.

Lagrange test

Рассмотрим гипотезу:

$$H_0 : \mathbf{m}_u = \mathbf{m}_u^0, \quad \mathbf{m}_u \neq \mathbf{m}_u^0.$$

- $S_{m,u}\left(\mathbf{w}_u, \mathbf{w}_v\right), S_{m,v}\left(\mathbf{w}_u, \mathbf{w}_v\right)$ — производные логарифма правдоподобия выборки $\mathfrak{D}_m$ по параметрам $\mathbf{w}_u$ и $\mathbf{w}_v$.
- $\mathbf{s}_m = S_{m,u}\left(\mathbf{m}_u^0, \hat{\mathbf{w}}_v^0\right)$, где $\hat{\mathbf{w}}_v^0$ из решения $S_{m,v}\left(\mathbf{m}_u^0, \mathbf{w}_v\right) = 0$.

Статистика Лагранжа:

$$LM = \mathbf{s}_m^{\mathsf{T}} \mathbf{Q}_m^{-1} \mathbf{s}_m$$

- $H_0 : LM \sim \chi^2\left(k\right)$
- $H_1 : LM \sim \chi^2\left(k, \gamma\right)$

- $\gamma = m\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\xi}$, $\boldsymbol{\xi}$ и $\boldsymbol{\Sigma}$ — средние оценки $\mathbf{s}_m$ по одному объекту.
- $\gamma : \chi^2_{k,1-\alpha} = \chi^2_{k,\beta}\left(\gamma\right)$

Likelihood ratio

Рассмотрим гипотезу:

$$H_0 : \mathbf{m}_u = \mathbf{m}_u^0, \quad \mathbf{m}_u \neq \mathbf{m}_u^0.$$

Статистика отношения логарифма правдоподобий:

$$LR = -2 \ln \frac{L\left(\mathfrak{D}_m, \hat{\mathbf{w}}_0\right)}{L\left(\mathfrak{D}_m, \hat{\mathbf{w}}\right)}$$

- $H_0 : LR \sim \chi^2\left(k\right)$
- $H_1 : LR \sim \chi^2\left(k, \gamma\right)$

- $\gamma = m\Delta^*$, где $\Delta^* = \mathsf{E}\left[2a^{-1}\left\{\left(\theta - \theta^*\right)\nabla b\left(\theta\right) - b\left(\theta\right) + b\left(\theta^*\right)\right\}\right]$
- $\gamma : \chi^2_{k,1-\alpha} = \chi^2_{k,\beta}\left(\gamma\right)$

Wald test

Рассмотрим гипотезу:

$$H_0 : \mathbf{m}_u = \mathbf{m}_u^0, \quad \mathbf{m}_u \neq \mathbf{m}_u^0.$$

Статистика Вальда:

$$W = \left(\hat{\mathbf{w}}_u - \mathbf{m}_u^0\right)^{\mathsf{T}} \hat{\mathbf{V}}_u^{-1} \left(\hat{\mathbf{w}}_u - \mathbf{m}_u^0\right)$$

- $H_0 : W \sim \chi^2(k)$
- $H_1 : W \sim \chi^2(k, \gamma)$

- $\gamma = m\delta$, где $\delta = \left(\mathbf{w}_u - \mathbf{m}_u^0\right)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{w}_u - \mathbf{m}_u^0\right)$
- $\gamma : \chi^2_{k,1-\alpha} = \chi^2_{k,\beta}(\gamma)$

Проверка на нормальность

Сравнение критериев проверки нормальности распределения случайных величин

Наименование критерия (раздел)	Характер альтернативного распределения					Ранг
	асимметричное		симметричное		≈ нормальное	
	$\alpha_4 < 3$	$\alpha_4 > 3$	$\alpha_4 < 3$	$\alpha_4 > 3$	$\alpha_4 \approx 3$	
Критерий Шапиро–Уилка (3.2.2.1)	1	1	3	2	2	1
Критерий K^2 (3.2.2.16)	7	8	10	6	4	2
Критерий Дарбина (3.1.2.7)	11	7	7	15	1	3
Критерий Д'Агостино (3.2.2.14)	12	9	4	5	12	4
Критерий α_4 (3.2.2.16)	14	5	2	4	18	5
Критерий Васичека (3.2.2.2)	2	14	8	10	10	6
Критерий Дэвида–Хартли–Пирсона (3.2.2.10)	21	2	1	9	1	7
Критерий χ^2 (3.1.1.1)	9	20	9	8	3	8
Критерий Андерсона–Дарлинга (3.1.2.4)	18	3	5	18	7	9
Критерий Филлибена (3.2.2.5)	3	12	18	1	9	10
Критерий Колмогорова–Смирнова (3.1.2.1)	16	10	6	16	5	11
Критерий Мартинеса–Иглевича (3.2.2.14)	10	16	13	3	15	12
Критерий Лина–Мудхолкара (3.2.2.13)	4	15	12	12	16	13
Критерий α_3 (3.2.2.16)	8	6	21	7	19	14
Критерий Шпигельхальтера (3.2.2.11)	19	13	11	11	8	15
Критерий Саркади (3.2.2.12)	5	18	15	14	13	16
Критерий Смирнова–Крамера–фон Мизеса (3.1.2.2)	17	11	20	17	6	17
Критерий Локка–Спурье (3.2.2.7)	13	4	19	21	17	18
Критерий Оя (3.2.2.8)	20	17	14	13	14	19
Критерий Хегази–Грина (3.2.2.3)	6	19	16	19	21	20
Критерий Муроты–Такеучи (3.2.2.17)	15	21	17	20	20	21