Обобщённые линейные модели

Центр биоэлектрических интерфейсов, 25 марта 2019 г.

Денис Деркач, Влад Белавин

Оглавление

Мотивация

Линейная регрессия: общие модели

Мы можем записать общую (основную) линейную модель (General Linear Model):

$$Y = X\beta + \varepsilon$$

- Y - вектор наблюдаемых зависимых переменных (откликов);
- X - матрица независимых переменных (дизайн эксперимента);
- β - матрица, включающая параметры, представляющие интерес для исследования;
- ε - матрица случайных ошибок.

NB: случайные ошибки распределены нормально.

Применимость общих линейных моделей

Перефразируя: мы имеем ввиду, что переменная-отклик подчиняется нормальному распределению.

$$y_i \sim N(\mu; \sigma) \text{ - случайная часть,}$$

$$\mathbb{E}(y_i) = \mu_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_{p-1} x_{p-1,i} \text{ - фиксированная часть.}$$

Построение обобщённой модели

Что будет, если мы не можем утверждать, что случайные ошибки распределены нормально?
Необходимо построить обобщённую модель (Generalized Linear Model, GLM, GLZ).
NB: общие модели иногда также сокращают GLM.

Обобщённые линейные модели

Обобщённые линейные модели: компоненты

Для $f(y;\theta)$ принадлежащему экспоненциальному семейству распределений:

$y_i \sim f(y;\theta)$ - случайная (random) компонента

$g\left(\mathbb{E}(y_i)\right) = \eta_i$, где g - функция связи (link function),

$$\eta_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_{p-1} x_{p-1,i}$$

- фиксированная часть (systematic, fixed).

Таким образом, вместо двух компонент общей линейной модели, у обобщённой линейной модели есть три компоненты.

Экспоненциальное семейство распределений: напоминание

Определение

Семейство распределений $\{P_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$ называется (k-параметрическим) экспоненциальным семейством на $\mathbb{R}^q$, если существуют такие вещественнозначные функции:

- $\eta_1, \ldots, \eta_k$ и B от θ,
- $T_1, T_2, \ldots, T_k$ и h от $x \in \mathbb{R}^q$,

такие, что плотность вероятности этого семейства записывается:

$$p_\theta(x) = \exp\left[\sum_{i=1}^{k}(\eta_i(\theta)T_i(x) - B(\theta))h(x)\right]$$

Примеры экспоненциальных семейств

Для непрерывных величин:

- Нормальное распределение
- Гамма распределение

Для дискретных величин:

- Биномиальное распределение
- Распределение Пуассона
- Отрицательное биномиальное распределение

Каноническое экспоненциальное семейство

Перепишем определение экспоненциального семейства:

$$f(y;\theta) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y;\phi)\right)$$

Заметим:

- $\mathbb{E}(Y) = \mu = b'(\theta)$.
- $\mathbb{V}ar(Y) = b''(\theta)\phi$.

Экспоненциальные семейства

	Normal	Poisson	Bernoulli
Notation	$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{P}(\mu)$	$\mathcal{B}(p)$
Range of y	$(-\infty, \infty)$	$[0, -\infty)$	$\{0, 1\}$
ϕ	σ^2	1	1
$b(\theta)$	$\frac{\theta^2}{2}$	e^θ	$\log(1 + e^\theta)$
$c(y, \phi)$	$-\frac{1}{2}(\frac{y^2}{\phi} + \log(2\pi\phi))$	$-\log y!$	1

Функция связи (link function)

Функция связи $g(\mu)$ используется разная в зависимости от распределения $f(y;\theta)$. Обычно предполагают, что $g(\mu)$ монотонна и дифференциируема в области разрешённых μ.
Количество распределений из экспоненциального семейства довольно мало, для каждого из них есть канонические функции связи.

Определение

Канонической функцией связи для экспонеициального семейства, g, называют функцию, которая связывает среднее μ и канонический параметр θ.

Заметим, так как $\mu = b'(\theta)$, то каноническая $g(\mu) = (b')^{-1}(\mu)$.

Канонические функции связи

$y_i \sim N(\mu_i, \sigma)$

$E(y_i) = \mu_i$

Функция связи идентичность

$\mu_i = \eta_i$

$\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_{p-1} x_{p-1i}$

Обратная функция:

$g^{-1}(\eta_i) = \eta_i$

$y_i \sim \text{Poisson}(\mu_i)$

$E(y_i) = \mu_i$

Функция связи логарифм

$\ln(\mu_i) = \eta_i$

$\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_{p-1} x_{p-1i}$

Обратная функция:

$g^{-1}(\eta_i) = e^{\eta_i}$

$y_i \sim \text{Binomial}(n_i, \pi_i)$

$E(y_i) = \pi_i$

Функция связи логит

$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$

$\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_{p-1} x_{p-1i}$

Обратная функция:

$g^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$

NB: Такие функции связи не являются единственно возможными.

Резюме

В общей линейной модели, соотношение между $\mathbb{E}(y_i)$ и параметрами линейно.
В обобощённой линейной модели, соотношение между функцией от $\mathbb{E}(y_i)$ и параметрами линейно.

GLM для нормального распределения

Выпишем три компоненты GLM:

- $y_i \sim \mathcal{N}(\mu; \sigma)$ - случайная компонента.
- $\eta_i = \sum x_{ij}\beta_j$ - linear predictor.
- $g(\mu_i) = \mu_i$ - функция связи.

Тогда:

$$\mu_i = g(\mu_i) = \eta_i = \sum x_{ij}\beta_j.$$

Таким образом, общая линейная модель — частный случай обобщённой линейной модели.

GLM для других распределений

Пример

Правительство хочет изучить зависимость бинарную возврата кредита от следующих переменных: оборот (непрерывный), сектора экономики (12 факторов) и целевого рынка (6 факторов).

Тогда

$$\mathbb{P}\left[Y=y\right]=\pi^{y}(1-\pi)^{(1-y)}=\exp\left(y\log\frac{\pi}{1-\pi}+\log(1-\pi)\right)$$

Заметим, что $\eta(\pi)=\log\frac{\pi}{1-\pi}$, часто функции связи выбирают близкими к $b(\theta)$.

В нашем случае $g(\mu_i)=g(\pi_i)=\log\frac{\pi_i}{1-\pi_i}$. Такая функция связи иногда называется логистической.

Оптимизация в GLM

Метод наименьших квадратов

В общем случае GLM, у нас нет идентичных и одинаково распределённых остатков. Потому метод наименьших квадратов здесь применять нельзя (вернее, его нужно модифицировать). Из-за этого используют метод максимального правдоподобия.

Метод максимального правдоподобия

Вернёмся к примеру с кредитами:

$$\mathbb{E}[Y_i] = \pi_i = g^{-1}(\eta_i) = \frac{1}{1 + e^{-\eta_i}}$$

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}; \mathbf{y}) &= \prod_{i=1}^{n} \exp\left\{ y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \right\} \\
\log \mathcal{L}(\boldsymbol{\beta}; \mathbf{y}) &= \sum_{i=1}^{n} y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} \log(1-\pi_i) \\
&= \sum_{i=1}^{n} y_i \sum_{j=1}^{p} x_{ij}\beta_j - \sum_{i=1}^{n} \log\left(\exp\left\{ \sum_{j=1}^{p} x_{ij}\beta_j \right\} + 1 \right)
\end{aligned}$$

Таким образом, мы можем выписать правдоподобие в явном виде и попытаться его оптимизировать.

Связь с β

Для произвольной функции связи g:

$$\theta_i = (b')^{-1}(\mu) =$$
$$= (b')^{-1}(g^{-1}(X_i^T \beta)) \equiv h(X^T \beta),$$

где $h = (b')^{-1} \circ g^{-1} = (g \circ b')^{-1}$.
NB: Для канонической функции связи $h = 1$.

Общий случай

В общем случае:

$$\log \mathcal{L}_n = \sum_i \frac{Y_i\theta_i - b(\theta_i)}{\phi} =$$

$$= \sum_i \frac{Y_i h(X_i^T\beta) - b(h(X_i^T\beta))}{\phi},$$

Для канонической функции связи:

$$\log \mathcal{L}_n = \sum_i \frac{Y_i X_i^T\beta - b(X_i^T\beta)}{\phi}$$

Наблюдения

- Лог-правдоподобие строго вогнуто для канонической функции связи (в случае, если $\phi(x) > 0$).
- Как следствие, ОМП сходится к единственному глобальному максимуму.
- В случае неканонической функции связи, это не так.

Методы оптимизации

- метод Ньютона-Рафсона;
- метод Фишера;
- метод итеративно перевешиваемых наименьших квадратов (Iteratively Re-weighted Least Squares).

Оценка параметров GLM

$\hat{\beta}$:

- оценивается методом максимального правдоподобия;
- существует и единственна,
- находится численно (например, методом Ньютона-Рафсона),
- состоятельна, асимптотически эффективна, асимптотически нормальна.

Доверительные интервалы $\hat{\beta}$

Для отдельного коэффициента β_j:

$$\hat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\left(I^{-1}\left(\hat{\beta}\right)\right)_{jj}}.$$

Для $g\left(\mathbb{E}\left(y\,|x_0\right)\right)$ — преобразованного матожидания отклика на новом объекте x_0:

$$x_0^T\hat{\beta} \pm z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}.$$

Для матожидания отклика на новом объекте x_0:

$$\left[g^{-1}\left(x_0^T\hat{\beta} - z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}\right), g^{-1}\left(x_0^T\hat{\beta} + z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}\right)\right].$$

I – информационная матрица Фишера.

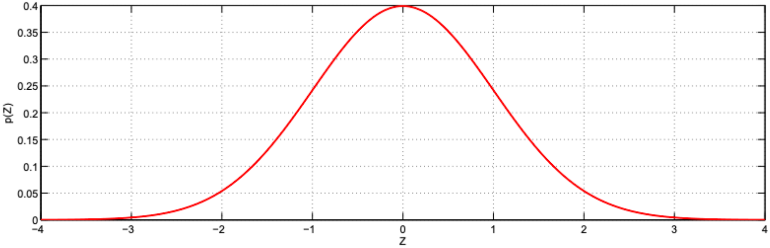
Тест Вальда для коэффициентов GLM

нулевая гипотеза: $H_0: \beta_j = 0;$

альтернатива: $H_1: \beta_j <\neq> 0;$

статистика: $T = \dfrac{\hat{\beta}_j}{\sqrt{\left(I^{-1}(\hat{\beta})\right)_{jj}}};$

$T \sim N(0,1)$ при H_0.



Тест Вальда и отношение правдопободобий

При $k1 = 1$ критерии Вальда и отношения правдоподобия не эквивалентны, в отличие от случая линейной регрессии, когда в этом случае достигаемые уровни значимости критериев Стьюдента и Фишера совпадают.
При больших n разница между критериями невелика, но в случае, когда их показания расходятся, рекомендуется смотреть на результат критерия отношения правдоподобия.

Оценка качества GLM

Анализ аномальности: шкала

Для получения шкалы рассматривают два граничных случая:

- Насыщенная модель — каждое уникальное наблюдение (сочетание значений предикторов) описывается одним из n параметров.
- Предложенная модель — модель, подобранная в данном анализе.
- Нулевая модель — все наблюдения описываются одним параметром (средним).

Степени свободы:

$df_{\text{saturated}} = 0$; $df_{\text{model}} = n - p_{\text{model}}$; $df_{\text{null}} = n - 1$;

Аномальность (deviance)

Аномальность - мера различия правдоподобий двух моделей:

- Остаточная аномальность

$$D_{\text{residual}} = 2(\log \mathcal{L}_{\text{saturated}} - \log \mathcal{L}_{\text{model}})$$

- Нулевая аномальность

$$D_{\text{null}} = 2 \log \mathcal{L}_{\text{saturated}} - \log \mathcal{L}_{\text{null}}$$

Сравнение нулевой и остаточной аномальности позволяет судить о статистической значимости модели в целом (при помощи теста отношения правдоподобий).

Анализ аномальности

- Для тестирования значимости модели целиком:

$$LRT = 2\log\left(\frac{\log\mathcal{L}_{\text{model}}}{\log\mathcal{L}_{\text{null}}}\right) = D_{\text{null}} - D_{\text{residual}}$$

$$df = df_{\text{null}} - df_{\text{model}} = p_{\text{model}} - 1$$

- Для тестирования значимости предикторов:

$$LRT = 2\log\left(\frac{\log\mathcal{L}_{\text{model}}}{\log\mathcal{L}_{\text{reduced}}}\right) = D_{\text{full}} - D_{\text{residual}}$$

$$df = df_{\text{full}} - df_{\text{model}} = p_{\text{model}} - p_{\text{reduced}}$$

В дальнейшем происходит сравнение с χ^2 с df степенями свободы.
NB: применение похоже на R^2