

Линейная регрессия. Выбор модели.

ПМИ ФКН ВШЭ, 23 ноября 2019 г.

Денис Деркач, Алексей Артёмов

ФКН ВШЭ

Денис Деркач

Оглавление

Мотивация

Обобщённые линейные модели

Оптимизация в GLM

Оценка качества GLM

Мотивация

Линейная регрессия: основные модели

Мы можем записать основную (общую) 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 и не зависят друг от друга.

Применимость общих линейных моделей

Фактически:

$y_i \sim N(\mu_i; \sigma)$ - случайная часть,

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

Построение обобщённой модели

Что будет, если мы не можем утверждать:

- › что случайные ошибки распределены нормально?
- › что случайные ошибки не коррелируют?

Пример: регрессия с бинарными откликами

- › Местная клиника отправила своим клиентам предложение сделать прививку от гриппа для защиты от ожидаемой эпидемии.
- › Через некоторое время у случайных 50 пациентов спросили сделали ли они прививку, при этом записываем информацию о возрасте и общем отношении к своему здоровью.
- › Необходимо построить модель, которая предсказывает привит ли человек в зависимости от его возраста и отношения к здоровью.

Таким образом, мы говорим о модели, где $Y_i = \{0; 1\}$.

Регрессия с бинарными откликами: проблемы

Для каждого ответа у нас получается вероятностное распределение Бернулли с вероятностью успеха π_i :

$\mathbb{P}(Y_i = y_i) = \pi_i^{y_i} (1 - \pi_i)^{(1 - y_i)}$. То есть:

- › среднее $\mu_i = \mathbb{E}(Y_i) = \pi_i$;
- › дисперсия $\text{Var}(Y_i) = \sigma_i^2 = \pi_i(1 - \pi_i)$

Таким образом дисперсия коррелирует со средним, обычные уравнения регрессии не применимы!

Аналогичные выражения мы получим для n испытаний.

Регрессия с бинарными откликами: проблемы

Ещё одна проблема. Очевидная форма регрессии,

$$y_i = x_i\beta,$$

в левой части должна быть ограничена $(0; 1)$, правая часть же может расти неограниченно.

Идея: преобразование Logit

- › Введём преобразование:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \in (-\infty; \infty);$$

- › обратная функция:

$$\text{logit}^{-1}(x) = \frac{e^x}{1+e^x} \in (-1; 1);$$

- › производная:

$$\frac{d}{d\pi}\text{logit}(\pi) = \frac{1}{\pi(1-\pi)} \sim \frac{1}{\text{Var}(\pi)}.$$

Logit, Probit и log-log

Вообще, существуют три популярных преобразования бинарной регрессии:

- › Logit: $\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$;
- › Probit: $\text{probit}(\pi) = \Phi^{-1}(\pi_i)$, Φ - функция распределения (cdf) нормальной величины;
- › Log-log: $\eta_i = \log(-\log(1 - \pi_i))$.

Каждая имеет свою интерпретацию и применяется в специальных типах задач.

Logit, Probit и log-log

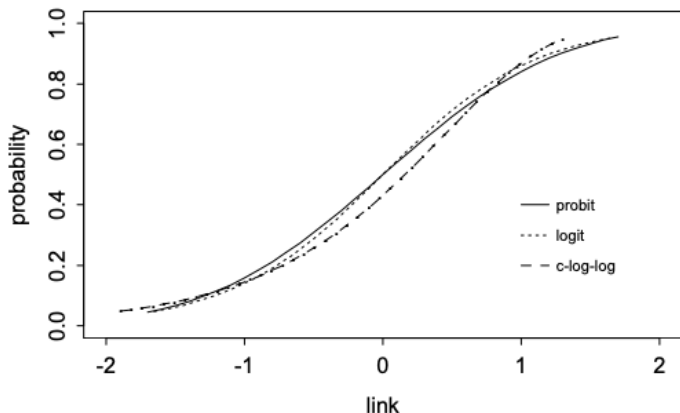


FIGURE 3.7: The Standardized Probit, Logit and C-Log-Log Links

Картинка отсюда.

Денис Деркач

Итоговая задача

Мы хотим максимизировать:

$$\mathcal{L}(\pi_i; Data) = \prod_{i=1}^n \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i}$$

При этом, используя logit получаем:

$$\pi_i = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}$$

Подставляя в первое уравнение, получим:

$$\mathcal{L}((\beta_0; \beta_1); Data) = \prod_{i=1}^n \left(\frac{(e^{\beta_0 + \beta_1 X_i})^{Y_i}}{1 + e^{\beta_0 + \beta_1 X_i}} \right)$$

К сожалению, прямое нахождение экстремумов у такой функции правдоподобия затруднено, потому приходится использовать итеративные процедуры.

Обобщённые линейные модели

Экспоненциальное семейство распределений

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:

- › η_1, \dots, η_k и B от θ ,
- › T_1, T_2, \dots, T_k и h от $x \in \mathbb{R}^q$,

такие, что плотность вероятности этого семейства записывается:

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

Экспоненциальное семейство распределений

Линеаризуя, получим для скалярных x и θ :

$$p_{\theta, \phi}(x) = \exp \left(\frac{x\theta - b(\theta)}{a(\phi)} + c(y; \phi) \right),$$

где ϕ зависит от конкретного распределения.

Пример экспоненциального семейства

$$p_{\theta, \phi}(x) = \exp \left(\frac{x\theta - b(\theta)}{a(\phi)} + c(y; \phi) \right).$$

Нормальное распределение можно переписать в нужной форме:

$$p_{\theta, \phi}(x) = \exp \left[\frac{x\theta - \theta^2/2}{\phi} - \frac{1}{2} \left(\frac{x^2}{\phi} + \ln(2\pi\phi) \right) \right].$$

То есть,

$$a(\phi) = \phi; b(\theta) = \theta^2/2; c(y; \phi) = -\frac{1}{2} \left(\frac{x^2}{\phi} + \ln(2\pi\phi) \right).$$

Другие экспоненциальные семейства

Table 15.9 Functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ for Constructing the Exponential Families

Family	$a(\phi)$	$b(\theta)$	$c(y, \phi)$
Gaussian	ϕ	$\theta^2/2$	$-\frac{1}{2} \left[y^2/\phi + \log_e(2\pi\phi) \right]$
Binomial	$1/n$	$\log_e(1+e^\theta)$	$\log_e\binom{n}{ny}$
Poisson	1	e^θ	$-\log_e y!$
Gamma	ϕ	$-\log_e(-\theta)$	$\phi^{-2} \log_e(y/\phi) - \log_e y - \log_e \Gamma(\phi^{-1})$
Inverse-Gaussian	ϕ	$-\sqrt{-2\theta}$	$-\frac{1}{2} \left[\log_e(\pi\phi y^3) + 1/(\phi y) \right]$

NOTE: In this table, n is the number of binomial observations, and $\Gamma(\cdot)$ is the gamma function.

Из John Fox, Applied Regression Analysis and Generalized Linear Models.

Свойства экспоненциальных семейств

$$p_{\theta, \phi}(x) = \exp \left(\frac{x\theta - b(\theta)}{a(\phi)} + c(y; \phi) \right).$$

- › θ часто называется каноническим параметром, причём $g_c(\theta)$ — функция математического ожидания и не зависит от ϕ ;
- › ϕ называется дисперсионным параметром.

Заметим:

- › $\mathbb{E}(X) = b'(\theta) = \mu$.
- › $\text{Var}(X) = b''(\theta)a(\phi) = a(\phi)V(\mu)$, $V(\mu)$ — дисперсионная функция.

Обобщённые линейные модели:

компоненты

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

- › стохастическая (random) компонента:

$$y_i \sim f(y; \theta)$$

- › фиксированный линейный предиктор (систематическая компонента):

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

- › функция связи

$$g(\mathbb{E}(y_i)) = \eta_i.$$

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

Функция связи (link function)

Функция связи $g(\mu)$ используется разная в зависимости от распределения $f(y; \theta)$. Обычно предполагают, что $g(\mu)$ монотонна и дифференцируема в области разрешённых μ .

Количество распределений из экспоненциального семейства довольно мало, для каждого из них есть канонические функции связи.

Определение

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

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

Резюме

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

В обобщённой 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.
- › $\eta_i = \sum x_{ij}\beta_j$ - линейный предиктор.
- › $g(\mu_i) = \mu_i$ - 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.

Оптимизация в GLM

Метод наименьших квадратов

В общем случае GLM, у нас нет идентичных и одинаково распределённых остатков. Потому метод наименьших квадратов здесь также применять нельзя (вернее, его нужно модифицировать).

Из-за этого используют метод максимального правдоподобия.

Метод максимального правдоподобия

Вернёмся к примеру с бинарной регрессией:

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

$$\mathcal{L}(\beta; \mathbf{y}) = \prod_{i=1}^n \exp \left\{ y_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) + \log(1 - \pi_i) \right\}$$

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

Наблюдения

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

Оценка параметров GLM

$\hat{\beta}$:

- › оценивается методом максимального правдоподобия;
- › существует и единственна,
- › находится численно,
- › состоятельна, асимптотически эффективна, асимптотически нормальна.

Способы оптимизации

- › метод Ньютона-Рафсона;
- › метод IRLS (итерационный взвешенный метод наименьших квадратов);
- › метод скоринга Фишера;
- › стохастические методы;
- › оценка максимального квази-правдоподобия.

В случае использования канонической функции связи первые три эквивалентны, иначе сказывается разница в том, какой гессиан используется в методе (см. обсуждение здесь).

IRLS в GLM

Алгоритм:

1. инициализировать значения $\hat{\mu}_i$ и $\hat{\beta}_j$ (для бинарной регрессии необходимо брать значения около 0 или 1 в зависимости от Y_i);
2. оценить значение $Z_i = g(\hat{\mu}_i) + g'(\hat{\mu})(Y_i - \hat{\mu}_i)$;
3. использовать веса $W^{-1} = g^2(\hat{\mu})V(\hat{\mu}_i)$ и применить обычную регрессию Z по X , получить $\hat{\beta}_j$;
4. получить новую оценку $\hat{\mu} = g^{-1}(\sum X_{ij}\beta_j)$;
5. повторить шаги 2-4 до сходимости.

Доверительные интервалы $\hat{\beta}$

Ассимптотически, мы можем использовать:

$$\hat{\beta} \sim \mathcal{N}(\beta, i^{-1}),$$

где $i(\beta) = \phi^{-1} X^T W X$, то есть мы можем использовать последний шаг IRLS для оценки матрицы ковариаций:

$$\widehat{\text{cov}}(\hat{\beta}) = \phi(X^T \hat{W} X)^{-1}.$$

Для некоторых семейств ϕ необходимо оценивать отдельно.

Оценка ϕ

К сожалению, оценка ϕ с помощью ОМП затруднена. Потому мы используем метод моментов, который даёт нам оценку:

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

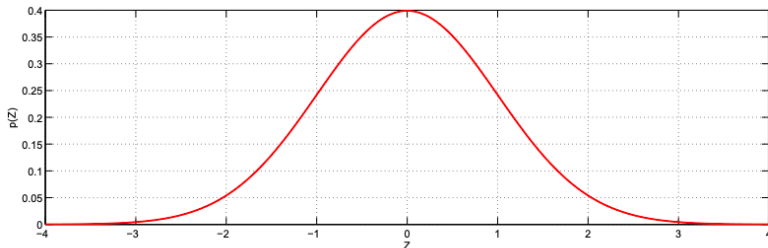
где n - количество точек, p - количество параметров.

Тест Вальда для коэффициентов GLM

нулевая гипотеза: $H_0: \beta_j = 0;$

альтернатива: $H_1: \beta_j \neq 0;$

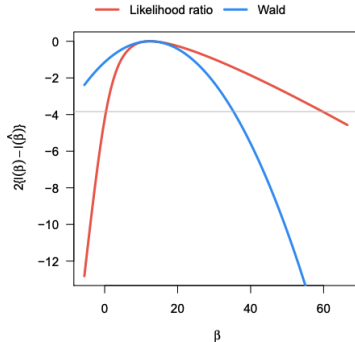
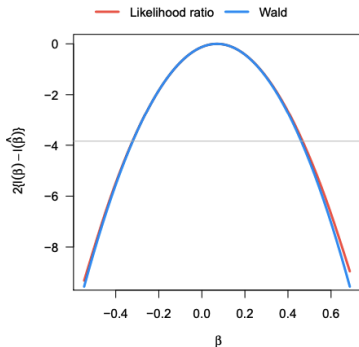
статистика: $T = \frac{\hat{\beta}_j}{\sqrt{(I^{-1}(\hat{\beta}))_{jj}}};$
 $T \sim N(0, 1)$ при H_0 .



Тест Вальда и отношение правдоподобий

- › Критерии Вальда и отношения правдоподобия не эквивалентны.
- › При больших n разница между критериями невелика, но в случае, когда их показания расходятся.
- › Вспомним, что мы в реальности получаем ОМП.

Сравнение Вальдовских интервалов и интервалов из функции правдоподобия



Слева направо увеличение количества примеров в выборке. В правой картинке теста Вальда не исключает 0, а тест отношения правдоподобий исключает.

Оценка качества GLM

Disclaimer: общий тест χ^2

- › Часто используется в статпакетах.
- › Это тест отношения правдоподобия полученной модели по сравнению с моделью только с постоянным коэффициентом β_0 .
- › Как и все прочие выводы на основании модельных статистик, тест отношения правдоподобия обоснован в предположении, что модель существует.
- › Единственное, что можно сказать при значимом общем χ^2 тесте, что, если модель верна, некоторые из ее коэффициентов ненулевые (насколько это полезно, решайте сами).
- › Простое правило: не используйте тест χ^2 для оценки качества.
- › Или хотя бы не используйте χ^2 в сравнениях.

Анализ остатков

Для проверки качества регрессий используют также анализ остатков.

› остатки Пирсона (для логрессии):

$$r_i = \frac{y_i - \pi_i}{\sqrt{\hat{\pi}_i(1 - \hat{p}_i)}}$$

Коэффициент детерминации

Очень хочется использовать коэффициент детерминации, который мы использовали для простых линейных моделей. В GLM существуют два подхода:

- › через квадрат коэффициента корреляций между Y_i и \hat{Y}_i .
- › напрямую, по определению.

К сожалению, коэффициент детерминации плохо определяет качество для ненормальных данных (например, R^2 для $Y_i = 0$ и $\hat{Y}_i = 0.9$ или $\hat{Y}_i = 0.99$ почти не отличаются, а $\mathbb{P}(Y_i = 0)$ отличаются в 10 раз).

ДевIANса (Аномальность)

Это приводит к идее осуществлять оценку качества через правдоподобие.

Определение

ДевIANсом обобщённой линейной модели является:

$$D(\mu, Y) = -2 \log \mathcal{L}(\mu, Y) + -2 \log \mathcal{L}(Y, Y)$$

где μ - оценка Y .

ДевIANса строится для оптимизации методом скоринга Фишера.

Анализ девианса: 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:

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

ДевIANса (deviance)

ДевIANса - мера различия правдоподобий двух моделей:

- › Остаточная девIANса

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

- › 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

Эффективная остаточная девианса

По аналогии с R^2 можно эффективно учесть количество степеней свободы:

$$D = 1 - \frac{D/(n-p)}{D_{null}/(n-1)} \sim \chi^2_{n-p}$$

Очень приближённо, мы можем сравнить две девиансы с помощью теста Вальда:

$$W = \frac{D(R) - D(F)}{\hat{\phi}} \sim \chi^2_{df_R - df_F}$$

В принципе, в анализе мы можем использовать X^2 Пирсона.

ROC

В специализированном случае бинарной регрессии, мы можем использовать совпадение ROC кривых, построенных по предсказаниям логрессии. В таком случае нам необходимо проводить тесты на равенство AUC (см. семинар).

Резюме

- › Обобщённые линейные модели добавляют возможность рассматривать данные, в которых зависимые переменные являются ненормальными.
- › Важным частным случаем является обобщённая модель для бинарных откликов с канонической функцией связи logit - логистическая регрессия.
- › Критерии качества определяются похожим способом с обычной линейной регрессией.
- › Лучшим критерием качества всё ещё является визуальный анализ.