Прикладной статистический анализ данных.
8. Обобщения линейной регрессии.

Рябенко Евгений
riabenko.e@gmail.com

I/2015

Постановка

Задача: оценить влияние одного или нескольких признаков на наступление какого-либо события и оценить его вероятность.

$1, \ldots, n$ — объекты;
$x_1, \ldots, x_k$ — предикторы;
y — отклик, $y_i \in \{0, 1\}$.

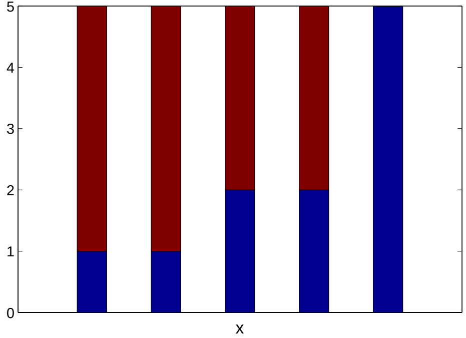
Хотим найти такую функцию π, что

$$P\left(y = 1 \left| x \right.\right) \approx \pi\left(x_1, \ldots, x_k\right)$$

.

Пример 1

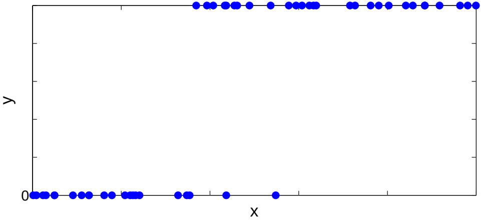
Повторяемый эксперимент с фиксированными уровнями фактора: разработка пестицидов, x_i — доза пестицида, y_i — смерть вредителя.



$$\hat{\pi}(x) = \frac{\sum_{i=1}^{n} y_i [x = x_i]}{\sum_{i=1}^{n} [x = x_i]}.$$

Пример 2

Неповторяемый эксперимент со случайными уровнями фактора: построение кривой спроса, x_i — цена товара, y_i — согласие купить товар.



Можно построить непараметрическую оценку при помощи ядерного сглаживания:

$$\hat{\pi}(x) = \frac{\sum\limits_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\sum\limits_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}.$$

Параметризация

Линейная регрессия:

$$\pi(x) = \beta_0 + \beta_1 x + \varepsilon.$$

- Оценка вероятности может выходить за $[0,1]$.
- В линейной регрессии $y = \mathbb{E}(y|x) + \varepsilon$, и МНК-оценка β хороша, когда $\varepsilon \sim N(0,\sigma)$. Здесь же, если $y = \pi(x) + \varepsilon$, то $\varepsilon = 1 - \pi(x)$ или $\varepsilon = \pi(x)$, и МНК-оценка будет плохой.

Нужно такое нелинейное преобразование

$$g(\pi(x)) = \beta_0 + \beta_1 x + \varepsilon,$$

чтобы:

- $\hat{\pi}(x) = g^{-1}\left(\hat{\beta}_0 + \hat{\beta}_1 x\right)$ принимала значения из $[0,1]$;
- изменения на краях диапазона значений x приводили к меньшим изменениям $\pi(x)$:
  x — годовой доход, y — покупка автомобиля,

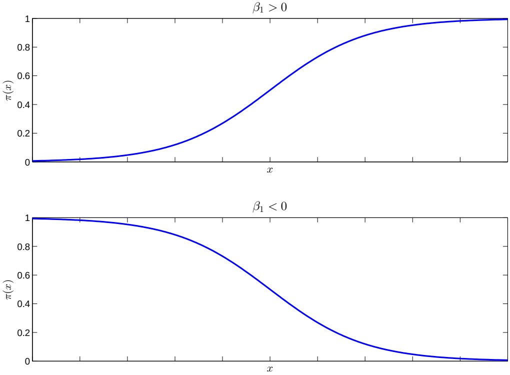
  $$\pi(10000000 + 200000) - \pi(10000000) < \pi(500000 + 200000) - \pi(500000).$$

Параметризация

Logit:

$$g(x) = g(\pi(x)) = \ln \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x + \varepsilon,$$

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}.$$



Относительный риск

Пусть $y \sim Ber(p)$, тогда **риск (odds)** события $y = 1$:

$$ODDS = \frac{p}{1-p}.$$

Если $y_1 \sim Ber(p_1)$, $y_2 \sim Ber(p_2)$, то **относительный риск (odds ratio)** события $y_1 = 1$ по сравнению с событием $y_2 = 1$:

$$OR = \frac{p_1/\left(1-p_1\right)}{p_2/\left(1-p_2\right)}.$$

Возраст / Серд. заболевания	≥ 55	≤ 55
есть	21	22
нет	6	51

$$OR = \frac{21/6}{22/51} \approx 8.1.$$

Роль коэффициентов логистической регрессии

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Пусть $x = [\text{возраст} \geq 55]$, $y = [\text{есть сердечные заболевания}]$. По $\hat{\beta}_1$ легко оценить относительный риск получения заболевания пожилыми людьми:

$$\widehat{OR} = e^{\hat{\beta}_1}.$$

Пусть $x = \text{возраст}$, $y = [\text{есть сердечные заболевания}]$. $e^{\hat{\beta}_1}$ имеет смысл мультипликативного прироста риска получения заболевания при увеличении возраста на 1 год.

Настройка параметров

Параметры оцениваются методом максимального правдоподобия:

$\pi(x)$ оценивает $P(y=1\,|x)$,
$1-\pi(x)$ оценивает $P(y=0\,|x)$ $\Rightarrow$

$$
\begin{aligned}
P(x_i, 1) &= \pi(x_i)\,,\\
P(x_i, 0) &= 1-\pi(x_i)\,,\\
l(\beta) &= \prod_{i=1}^{n} \pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i}\,,\\
L(\beta) &= \ln l(\beta) = \sum_{i=1}^{n}\left(y_i \ln\pi(x_i) + (1-y_i)\ln\left(1-\pi(x_i)\right)\right),\\
\hat{\beta} &= \operatorname*{argmax}_{\beta} L(\beta)\,.
\end{aligned}
$$

Свойства МП-оценки

$\hat{\beta}$:

- существует и единственна,
- находится методом Ньютона-Рафсона,
- состоятельна, асимптотически эффективна, асимптотически нормальна.

$\hat{\beta}$ может не существовать или не быть конечной, если:

- наблюдения $y = 0$ и $y = 1$ линейно разделимы в пространстве признаков X;
- матрица X вырождена.

Итерационный процесс может не сойтись, если число признаков k слишком велико относительно числа наблюдений n.

Дисперсия оценок

Пусть $I(\beta) \in \mathbb{R}^{(k+1)\times(k+1)}$ — матрица вторых производных $L(\beta)$:

$$\frac{\partial^2 L}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi(x_i)(1-\pi(x_i)),$$
$$\frac{\partial^2 L}{\partial \beta_j \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi(x_i)(1-\pi(x_i)).$$

Другая форма записи:

$$I(\beta) = X^T V X,$$
$$V = \operatorname{diag}(\pi(x_1)(1-\pi(x_1)), \dots, \pi(x_n)(1-\pi(x_n))).$$

Из теории оценок максимума правдоподобия: $\mathbb{D}\hat{\beta} = I^{-1}\left(\hat{\beta}\right)$.

Доверительные интервалы

Для отдельного коэффициента β_j:

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

Для $g(x_0)$ — логита нового объекта 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 вероятности $y = 1$ при $x = x_0$:

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

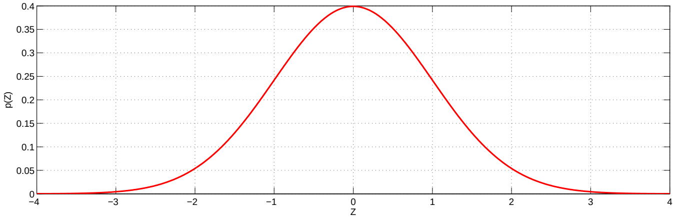
Критерий Вальда

нулевая гипотеза: $H_0\colon \beta_j = 0;$

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

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

$T \sim N(0,1)$ при H_0.



Критерий отношения правдоподобия

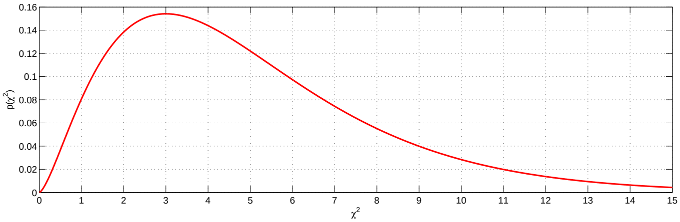
$$\underset{n\times(k+1)}{X} = \left(\underset{n\times(k+1-k_1)}{X_1}, \underset{n\times k_1}{X_2}\right); \quad \underset{(k+1)\times 1}{\beta^T} = \left(\underset{(k+1-k_1)\times 1}{\beta_1^T}, \underset{k_1\times 1}{\beta_2^T}\right)^T;$$

нулевая гипотеза: $H_0\colon \beta_2 = 0;$

альтернатива: $H_1\colon H_0$ неверна;

статистика: $G = 2\left(L_r - L_{ur}\right);$

$G \sim \chi^2_{k_1}$ при H_0.



Связь между критериями Вальда и отношения правдоподобия

При $k_1 = 1$ 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 и $\sum\limits_{i=1}^{n} [y_i = 1]$ разница между критериями невелика, но в случае, когда их показания расходятся, рекомендуется смотреть на результат критерия отношения правдоподобия.

Значимость категориальных предикторов

Значимость фиктивных переменных, кодирующих один категориальный предиктор, — тонкий вопрос.

- Необходимо включать или исключать категориальный предиктор целиком. Значимость соответствующих фиктивных переменных проверяется в совокупности с помощью критерия отношения правдоподобия.
- В случае, когда по отдельности какие-то фиктивные переменные не значимы, допустимо объединять уровни категориального предиктора, основываясь на интерпретации.
- Если какие-то уровни категориального предиктора лежат полностью в классе $y = 1$ или $y = 0$, их обязательно нужно объединить с другими уровнями, чтобы модель логистической регрессии могла быть построена.

Сравнение невложенных моделей

Невложенные модели можно сравнивать друг с другом по значению правдоподобия l, логарифма правдоподобия L или аномальности (deviance):

$$D = -2L.$$

Аномальность — аналог RSS в линейной регрессии; при добавлении признаков она не может убывать.

Для сравнения моделей с разным числом признаков можно использовать информационные критерии.
AIC — информационный критерий Акаике:

$$AIC = -2L + 2\left(k+1\right);$$

$AICc$ — он же с поправкой на случай небольшого размера выборки;

$$AICc = -2L + \frac{2k\left(k+1\right)}{n-k-1};$$

BIC (SIC) — байесовский (Шварца) информационный критерий:

$$BIC = -2L + \log n\left(k+1\right).$$

Мультиколлинеарность

Признаки мультиколлинеарности:

- правдоподобие модели высоко, но оценки многих коэффициентов близки к своим стандартным отклонениям;
- коэффициенты сильно меняются при включении и исключении других признаков.

Линейность логита

Проверка линейности логита по признакам — аналог визуального анализа остатков в обычной линейной регрессии.

Методы анализа линейности логита:

- сглаженные диаграммы рассеяния;
- фиктивные переменные по квартилям;
- дробные полиномы.

Сглаженные диаграммы рассеяния (smoothed scatterplots)

Рассмотрим оценку логита, полученную ядерным сглаживанием по x_j:

$$\bar{y}_{sm}\left(x_{ji}\right) = \frac{\sum\limits_{l=1}^{n} y_i K\left(\frac{x_{ji}-x_{li}}{h}\right)}{\sum\limits_{l=1}^{n} K\left(\frac{x_{ji}-x_{li}}{h}\right)},$$

$$\bar{l}_{sm}\left(x_{ji}\right) = \ln \frac{\bar{y}_{sm}\left(x_{ji}\right)}{1-\bar{y}_{sm}\left(x_{ji}\right)}.$$

График функции $\bar{l}_{sm}\left(x_j\right)$ должна быть похож на прямую.

Дробные полиномы (fractional polynomials)

Если логит нелинеен по признаку, можно попробовать добавлять в модель его осмысленные степени и проверять их значимость.
В автоматическом режиме это можно делать с помощью дробных полиномов.

1. Настраиваются модели с заменой x_j на допустимые степени признака x_j, например, из множества $S=\{-2,-1,-0.5,0,0.5,1,2,3\}$. Выбирается степень, максимизирующая правдоподобие.
2. Настраиваются модели с заменой x_j на двухкомпонентный полином x_j вида $\beta_{j_1}x_j^{p_1}+\beta_{j_2}x_j^{p_2}, \quad p_1,p_2\in S$ (если $p_1=p_2$, то берётся $\beta_{j_1}x_j^{p_1}+\beta_{j_2}x_j^{p_1}\ln x_j$). Выбираются степени, максимизирующая правдоподобие.
3. Если модель с полиномом второй степени значимо не лучше, чем линейная, используется линейная модель.
4. Если модель с полиномом второй степени значимо не лучше, чем с полиномом первой степени, используется модель с полиномом первой степени, иначе — с полиномом второй.

Содержательный отбор признаков

1. Если признаков достаточно много (например, больше 10), желательно сделать их предварительный отбор, основанный на значимости в однофакторной логистической регрессии. Для дальнейшего рассмотрения остаются признаки, достигаемый уровень значимости которых не превышает 0.25.
2. Строится многомерная модель, включающая все отобранные на шаге 1 признаки. Проверяется значимость каждого признака, удаляется небольшая группа незначимых признаков. Новая модель сравнивается со старой с помощью критерия отношения правдоподобия.
3. Чтобы убедиться, что удаление признаков не повлияло на оставшиеся, для каждого коэффициента $\hat{\beta}_j$ при оставшихся значимых признаках рассчитывается величина **delta-beta-hat-percent**:

$$\Delta\hat{\beta}\% = 100\frac{\hat{\beta}_j^{old} - \hat{\beta}_j^{new}}{\hat{\beta}_j^{new}}.$$

Если $\left|\Delta\hat{\beta}\%\right| > 20$, то какие-то из удалённых незначимых признаков были нужны, чтобы лучше определять коэффициенты значимых признаков; их стоит вернуть.

Содержательный отбор признаков

4. К признакам модели, полученной в результате циклического применения шагов 2 и 3, по одному добавляются удалённые признаки. Если какой-то из них становится значимым, он вносится обратно в модель.
5. Для непрерывных признаков полученной модели проверяется линейность логита. В случае обнаружения нелинейности признаки заменяются на соответствующие полиномы.
6. Исследуется возможность добавления в полученную модель взаимодействий факторов. Добавляются значимые интерпретируемые взаимодействия.
7. Проверяется адекватность финальной модели: близость y и $\hat{y}$; малость вклада наблюдений (x_i, y_i) на каждом объекте i в $\hat{y}$.

Порог классификации

Как по $\pi(x)$ оценить y?

$$y = [\pi(x) \geq p_0].$$

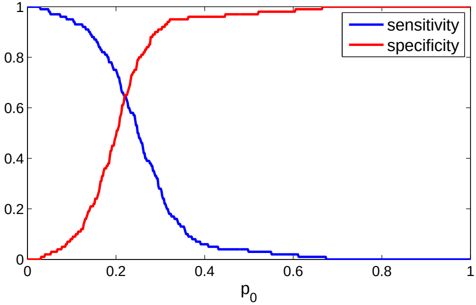
Чаще всего берут $p_0 = 0.5$, но можно выбирать по другим критериям, например, для достижения заданных показателей чувствительности или специфичности.

Порог классификации

Пример: эффективность терапии для наркозависимых, $p_0 = 0.5$:

$\hat{y}$ \ y	1	0
1	16	11
0	131	417

Чувствительность: $\frac{16}{16+131} \approx 10.9\%$.
Специфичность: $\frac{417}{11+417} \approx 97.4\%$.



Пример

Риск остеопоротических переломов у женщин:
https://yadi.sk/d/fkjAG2StfJm4J

Требования к решению задачи методом логистической регрессии

- визуализация данных, оценка наличия выбросов, анализ таблиц сопряжённости по категориальным признакам;
- содержательный отбор признаков: выбор наилучшей линейной модели, оценка линейности непрерывных признаков по логиту, анализ необходимости добавления взаимодействий, проверка адекватности финальной модели (анализ влиятельных наблюдений, классификация);
- выводы.

Постановка

$1, \dots, n$ — объекты;
$x_1, \dots, x_k$ — предикторы;
y — счётный отклик, $y_i \in \mathbb{N}$.

$$\mathbb{E}(y \,|x_1, \dots, x_k) = \mathbb{E}(y \,|x) = ?$$

Базовый метод — пуассоновская регрессия:

$$f(y \,|x) = \frac{e^{-\mu}\mu^y}{y!},$$
$$\mu = \mathbb{E}(y \,|x) = e^{x^T\beta},$$
$$\omega \equiv \mathbb{D}(y \,|x) = e^{x^T\beta}.$$

Настройка параметров

Оценка методом максимального правдоподобия:

$$L(\beta) = \ln l(\beta) = \sum_{i=1}^{n} \left(y_i x_i^T \beta - e^{x_i^T \beta} - ln(y_i!) \right),$$

$$\hat{\beta} = \operatorname*{argmax}_{\beta} L(\beta) \Leftrightarrow$$

$$\sum_{i=1}^{n} \left(y_i - e^{x_i^T \beta} \right) x_i = 0.$$

$\hat{\beta}$:

- существует и единственна,
- находится методом Ньютона-Рафсона,
- является состоятельной и асимптотически эффективной оценкой β,
- асимптотически нормальна.

Дисперсия оценок

Для оценки дисперсии $\hat{\beta}$ снова используется матрица вторых производных $L(\beta)$:

$$I(\beta) = \frac{\partial^2 L}{\partial\beta\partial\beta^T} = -\sum_{i=1}^{n} e^{x_i^T\beta} x_i x_i^T.$$

Из теории оценок максимума правдоподобия:

$$\begin{aligned}\mathbb{D}_{ML}\left(\hat{\beta}\right) &= I^{-1}\left(\hat{\beta}\right), \\ \hat{\beta} \overset{\text{a}}{\sim} N\left(\beta, \left(\sum_{i=1}^{n} e^{x_i^T\beta} x_i x_i^T\right)^{-1}\right) &\approx \\ \approx N\left(\beta, \left(\sum_{i=1}^{n} e^{x_i^T\hat{\beta}} x_i x_i^T\right)^{-1}\right)&.\end{aligned}$$

Доверительные интервалы

Для отдельного коэффициента β_j:

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

Для $\ln \mathbb{E}\left(y \,|\, x = x_0\right) = x_0^T\beta$:

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

Для $\mathbb{E}\left(y \,|\, x = x_0\right) = e^{x_0^T\beta}$:

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

Приближённый предсказательный интервал для $y\left(x_0\right)$ — отклика на новом объекте x_0:

$$e^{x_0^T\hat{\beta}} \pm 2\sqrt{e^{x_0^T\hat{\beta}}}.$$

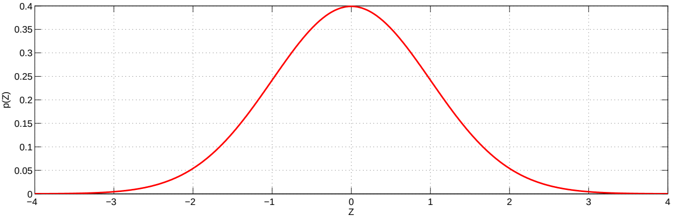
Критерий Вальда

нулевая гипотеза: $H_0\colon \beta_j = 0;$

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

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

$T \sim N(0,1)$ при H_0.



Критерий отношения правдоподобия

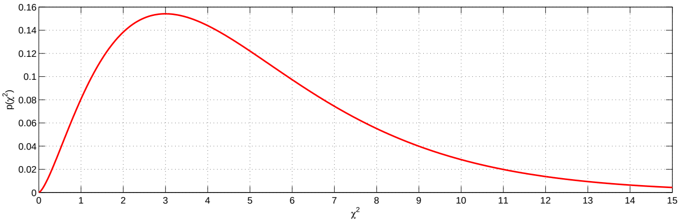
$$\underset{n\times(k+1)}{X} = \left(\underset{n\times(k+1-k_1)}{X_1}, \underset{n\times k_1}{X_2} \right); \quad \underset{(k+1)\times 1}{\beta^T} = \left(\underset{(k+1-k_1)\times 1}{\beta_1^T}, \underset{k_1\times 1}{\beta_2^T} \right)^T;$$

нулевая гипотеза: $H_0\colon \beta_2 = 0;$

альтернатива: $H_1\colon H_0$ неверна;

статистика: $G = 2\left(L_r - L_{ur}\right);$

$G \sim \chi^2_{k_1}$ при H_0.



Overdispersion/underdispersion

Пуассоновская модель предполагает, что $\omega = \mu$ (equidispersion).

- МП-оценки β остаются состоятельными, даже если распределение $y|x$ не является пуассоновским — достаточно того, что модель $\mathbb{E}\left(y\,|x\right)$ определена корректно.
- Оценки дисперсии $\hat{\beta}$ и соответствующие критерии требуют верного определения и $\mathbb{D}\left(y\,|x\right)$, поэтому они дают некорректные результаты, если матожидание и дисперсия не равны.
- Предположение о равенстве матожидания и дисперсии можно проверить; если оно не выполняется, можно изменить модель. Это позволит построить корректные критерии и более эффективные оценки β.

Overdispersion/underdispersion

Overdispersion — отрицательная биномиальная модель:

$$\omega(\alpha) = \mu + \alpha\mu^2,$$

$$f(y|\mu,\alpha) = \frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}.$$

Underdispersion — пороговая модель (hurdle model):

$$P(y=j) = \begin{cases} f_1(0), & j=0, \\ \frac{1-f_1(0)}{1-f_2(0)} f_2(j), & j>0. \end{cases}$$

Можно построить МП-оценки для α и β, а затем проверить гипотезу $\alpha = 0$ с помощью критерия отношения правдоподобия.

Устойчивая оценка дисперсии

Дисперсия оценки максимального квазиправдоподобия:

$$\mathbb{D}_{QML}\left(\hat{\beta}\right) = \left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} \left(\sum_{i=1}^{n} \omega_i x_i x_i^T\right) \left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1}.$$

Устойчивая состоятельная оценка дисперсии, подходящая для любого вида ω:

$$\mathbb{D}_{R}\left(\hat{\beta}\right) = \left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1} \left(\sum_{i=1}^{n} \left(y_i - \mu_i\right)^2 x_i x_i^T\right) \left(\sum_{i=1}^{n} \mu_i x_i x_i^T\right)^{-1}.$$

Меры качества модели

Относительные:

- аномальность:

$$D = -2L,$$

$$D_P = \sum_{i=1}^{n} \left(y_i \ln \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right),$$

$$D_{NB} = \sum_{i=1}^{n} \left(y_i \ln \frac{y_i}{\hat{\mu}_i} - \left(y_i + \alpha^{-1}\right) \ln \frac{y_1 + \alpha^{-1}}{\hat{\mu}_i + \alpha^{-1}} \right);$$

- AIC:

$$AIC = -2L + 2\,(k+1)\,.$$

Абсолютная:

- псевдо-R^2:

$$R^2_{DEV} = 1 - \frac{D}{D_0},$$

D_0 — аномальность модели с одной константой.

Пример

Число визитов к доктору:
https://yadi.sk/d/iaB-RbvRfNcC3

Требования к решению задачи методом пуассоновской регрессии

- визуализация данных, оценка наличия выбросов;
- отбор признаков: выбор наилучшей линейной модели, проверка равенства среднего и дисперсии, анализ необходимости добавления взаимодействий, проверка адекватности финальной модели (сравнение с устойчивой моделью, анализ влиятельных наблюдений);
- выводы.

Литература

Hosmer D.W., Lemeshow S., Sturdivant R.X. *Applied Logistic Regression*. — Hoboken: John Wiley & Sons, 2013.
Cameron C.A., Trivedi P.K. *Regression Analysis of Count Data*. — Cambridge University Press, 2013.