Прикладной статистический анализ данных
Дополнения и обобщения регрессии

Андрей Грабовой

Неслучайные пропуски

Иногда наличие пропуска в x_j информативно:

- отказ респондентов отвечать на вопрос
- Абрахам Вальд и повреждения самолётов
- признак не применим

В таких случаях необходимо:

1. создать новый бинарный признак

$$x_{j'} = \begin{cases} 1, & x_j = NA, \\ 0, & x_j \neq NA \end{cases}$$

2. заменить пропущенные значения в x_j на любую не встречающуюся в x_j константу c

Случайные пропуски

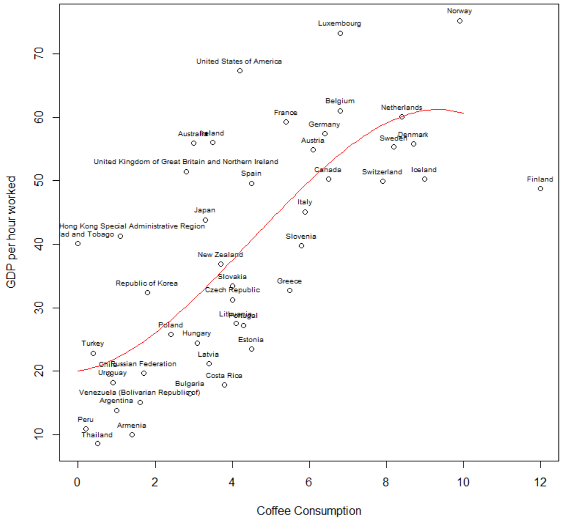
Способы борьбы с пропусками в X:

- удалить строки, содержащие пропуски (complete cases);
- заполнить пропуски (R packages: Amelia, mi, mice):
  - по ближайшему объекту
  - средними или медианами по столбцу
  - EM-алгоритмом (multiple imputation);
- считать X^TX и X^Ty только по полным парам (available cases):

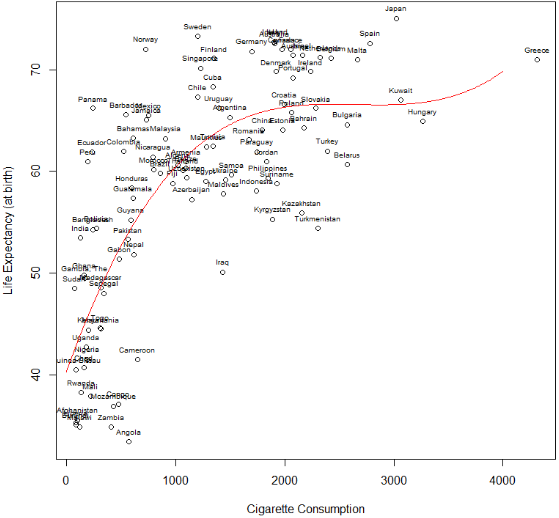
$$\left(X^TX\right)_{jl} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{il} \approx \frac{1}{n_{jl}}\sum_{i=1}^{n} x_{ij}x_{il}\left[x_{ij} \neq NA, x_{il} \neq NA\right],$$

n_{jl} — число полных пар.

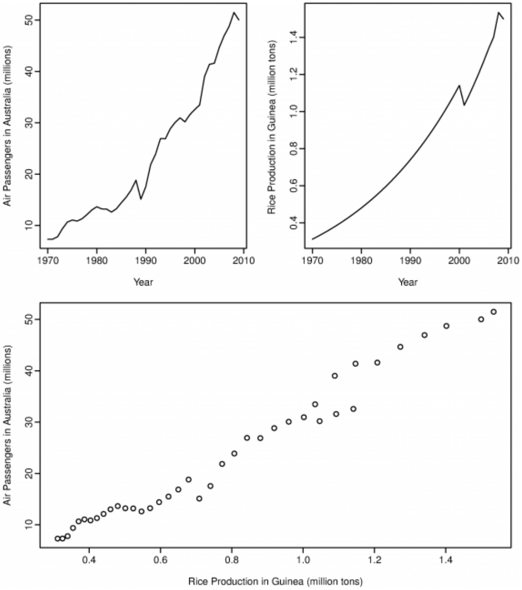
Интерпретация регрессионной модели



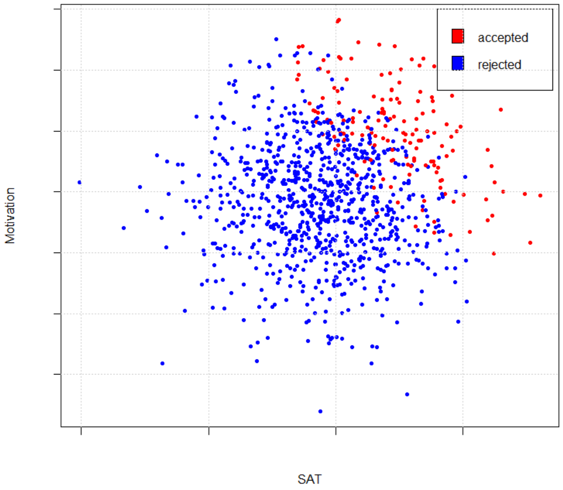
Интерпретация регрессионной модели



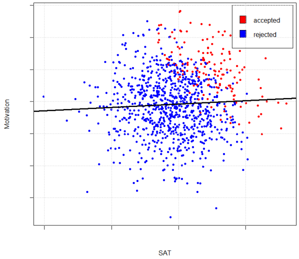
Интерпретация регрессионной модели



Интерпретация регрессионной модели



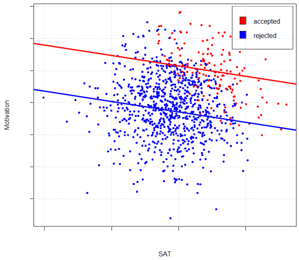
Интерпретация регрессионной модели




```
>summary(lm(motivation~sat, data=school))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.07415    0.03211  -2.309   0.0211 *
sat          0.05204    0.03189   1.632   0.1031
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.015 on 998 degrees of freedom
Multiple R-squared:  0.00266,Adjusted R-squared:  0.001661
F-statistic: 2.662 on 1 and 998 DF,  p-value: 0.1031
```

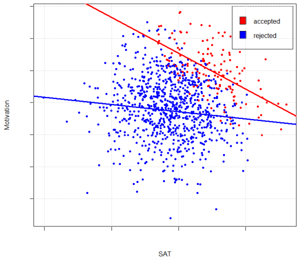
Интерпретация регрессионной модели




```
>summary(lm(motivation~sat+acceptance, data=school))
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.29246    0.03149  -9.289  < 2e-16 ***
sat            -0.16248    0.03121  -5.206 2.35e-07 ***
acceptanceTRUE  1.43688    0.08777  16.372  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9019 on 997 degrees of freedom
Multiple R-squared:  0.214,Adjusted R-squared:  0.2124
F-statistic: 135.7 on 2 and 997 DF,  p-value: < 2.2e-16
```

Интерпретация регрессионной модели




```
>summary(lm(motivation~sat*acceptance, data=school))
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.28296    0.03124  -9.059  < 2e-16 ***
sat                -0.11101    0.03284  -3.380 0.000752 ***
acceptanceTRUE      1.82261    0.12043  15.134  < 2e-16 ***
sat:acceptanceTRUE -0.44830    0.09692  -4.626 4.23e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8928 on 996 degrees of freedom
Multiple R-squared:  0.2305,Adjusted R-squared:  0.2282
F-statistic: 99.45 on 3 and 996 DF,  p-value: < 2.2e-16
```

Требования к решению задачи методом линейной регрессии

- визуализация данных, анализ распределения признаков (оценка необходимости трансформации), оценка наличия выбросов;
- оценка необходимости преобразования отклика и его поиск методом Бокса-Кокса;
- визуальный анализ остатков;
- проверка гипотез об остатках: нормальность, несмещённость, гомоскедастичность;
- отбор признаков с учётом множественной проверки гипотез и возможной гетероскедастичности;
- анализ необходимости добавления взаимодействий и квадратов признаков;
- расчёт расстояний Кука, возможное удаление выбросов, обновление модели;
- выводы.

Обобщённая линейная модель

$1, \dots, n$ — объекты;
$x_1, \dots, x_k$ — предикторы;
y — отклик;

$$X = \begin{pmatrix} x_{10} = 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} = 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}; \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix};$$

регрессионная модель:

$$\mathbb{E}\left(y \,|X\right) \equiv \mu = f\left(x_1, \dots, x_k\right);$$

линейная регрессионная модель:

$$\mu = X\beta;$$

обобщённая линейная регрессионная модель (GLM):

$$g\left(\mu\right) = X\beta, \quad \mu = g^{-1}\left(X\beta\right),$$

$g\left(x\right)$ — связующая функция — позволяет ограничить диапазон предсказываемых для μ значений.

Обобщённая линейная модель

В обычной линейной модели используется предположение о нормальности отклика:

$$y\,|X \sim N\left(X\beta, \sigma^2\right).$$

В обобщённой линейной модели распределение y берётся из экспоненциального семейства:

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

	$Pois(\lambda)$	$Bin(N, p)$	$N\left(\mu, \sigma^2\right)$
$a(\phi)$	1	1	σ^2
$b(\theta)$	e^θ	$n \ln\left(1 + e^\theta\right)$	$\theta^2/2$
$c(y, \phi)$	$\ln y!$	$\ln C_n^y$	$\frac{1}{2}\left(\frac{y^2}{\phi} + \ln(2\pi\phi)\right)$
$g(x)$	$\ln x$	$\ln \frac{x}{1-x}$	x
$g^{-1}(x)$	$e^x \in [0, \infty)$	$\frac{e^x}{1+e^x} \in [0, 1]$	$x \in \mathbb{R}$

Оценка параметров GLM

$\hat{\beta}$:

- оценивается методом максимального правдоподобия;
- существует и единственна,
- находится численно
  - методом Ньютона-Рафсона (Newton-Raphson method)
  - методом оценок Фишера (Fisher scoring method)
- состоятельна, асимптотически эффективна, асимптотически нормальна.

Итерационный процесс вычисления $\hat{\beta}$ может не сойтись, если k слишком велико относительно n.

$$\mathbb{D}\hat{\beta} = I^{-1}\left(\hat{\beta}\right),$$

$I(\beta) \in \mathbb{R}^{(k+1)\times(k+1)}$ — информационная матрица Фишера — матрица вторых производных логарифма правдоподобия $L(\beta)$.

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 $g\left(\mathbb{E}\left(y\,|x_0\right)\right)$ — 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:

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

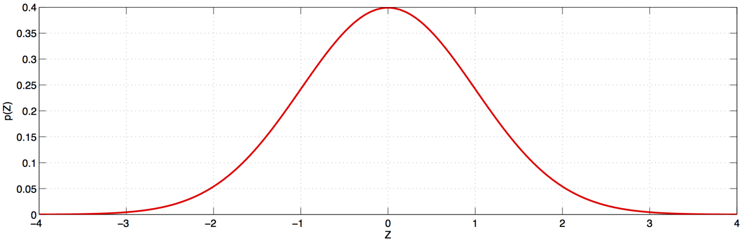
Критерий Вальда

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

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

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

нулевое распределение: $N(0,1)$



Критерий отношения правдоподобия

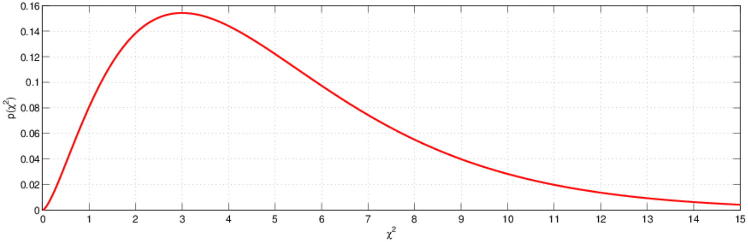
$$\underset{n\times(k+1)}{X} = \left(\underset{n\times(k+1-k_1)}{X_1}, \underset{n\times k_1}{X_2}\right); \quad \underset{(k+1)\times 1}{\beta^T} = \left(\underset{(k+1-k_1)\times 1}{\beta_1^T}, \underset{k_1\times 1}{\beta_2^T}\right)^T;$$

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

альтернатива: $H_1\colon H_0$ неверна

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

нулевое распределение: $\chi^2_{k_1}$



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 разница между критериями невелика, но в случае, когда их показания расходятся, рекомендуется смотреть на результат критерия отношения правдоподобия.

Меры качества моделей

Остаточная аномальность (residual deviance):

$$D_{res} = 2(L_{sat} - L_{fit})$$

Где L_{sat} – насыщенная (saturated) модель, имеющая число параметров равное числу объектов. Аномальность — аналог RSS в линейной регрессии; при добавлении признаков она не может убывать.

Для сравнения моделей с разным числом признаков можно использовать **информационные критерии**:

1. AIC — информационный критерий Акаике:

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

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

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

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

$$BIC = -2L + \ln n\,(k+1)\,.$$

Меры качества моделей

- $AIC < AICc$
- $BIC > AIC$ при $n \geqslant 8$
- выбор модели по BIC приводит к состоятельным оценкам — с ростом n вероятность выбора верного подмножества признаков стремится к 1
- минимизация AIC асимптотически даёт модель с наименьшей среднеквадратичной ошибкой предсказания
- модели со значением информационного критерия на расстоянии двух единиц от значения лучшей модели можно считать неотличимыми от лучшей

Бинарный отклик: постановка задачи

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

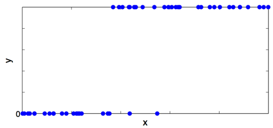
$1, \dots, n$ — объекты;
$x_1, \dots, x_k$ — предикторы;
y — отклик, $y_i \in \{0, 1\}$.

Хотим найти такой вектор β, что

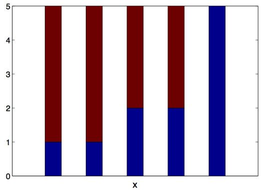
$$\mu = \mathbb{E}\,(y\,|X\,) = P\,(y = 1\,|X\,) \equiv \pi(x) \approx X\beta.$$

Примеры

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



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



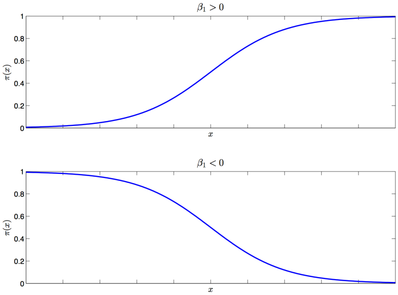
$\Longrightarrow$ логистическая регрессия может также использоваться для моделирования $y \in [0, 1]$.

Параметризация

Логит:

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

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



Параметризация

Логит:

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

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

- $\hat{\pi}\left(x\right) = g^{-1}\left(\beta_0 + \beta_1 x\right)$ принимает значения из $[0,1]$;
- изменения на краях диапазона значений x приводят к меньшим изменениям $\pi(x)$: x — годовой доход, y — покупка автомобиля,

$$\pi\left(10\,000\,000 + 200\,000\right) - \pi\left(10\,000\,000\right) < \pi\left(500\,000 + 200\,000\right) - \pi\left(500\,000\right).$$

Отношение шансов

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

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

Если $y_1 \sim Ber(p_1)$, $y_2 \sim Ber(p_2)$, то отношение шансов (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	$\geqslant 55$	< 55
есть	21	22
нет	6	51

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

Роль коэффициентов логистической регрессии

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \quad \Longleftrightarrow \quad \frac{p}{1-p} = e^{\hat{\beta}_0}\left(e^{\hat{\beta}_1}\right)^x$$

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

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

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

Настройка параметров

ММП:

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

Информационная матрица Фишера:

$$
I\left(\hat{\beta}\right) = X^T V X,
$$

$$
V = \operatorname{diag}\left(\hat{\pi}\left(x_i\right)\left(1 - \hat{\pi}\left(x_i\right)\right)\right).
$$

Проблемы настройки параметров

Если матрица X вырождена, некоторые коэффициенты модели не будут определены.

Если наблюдения $y = 0$ и $y = 1$ линейно разделимы в пространстве X, то:

- в теории коэффициенты бесконечно возрастают
- на практике коэффициенты и их дисперсии получаются большими, а почти все вероятности в обучающей выборке близки к 0 или 1.

Можно использовать регуляризацию Фирта. Функция меток исходной модели для коэффициента β_j:

$$\sum_{i=1}^{n} (y_i - \pi (x_i))\, x_{ij}.$$

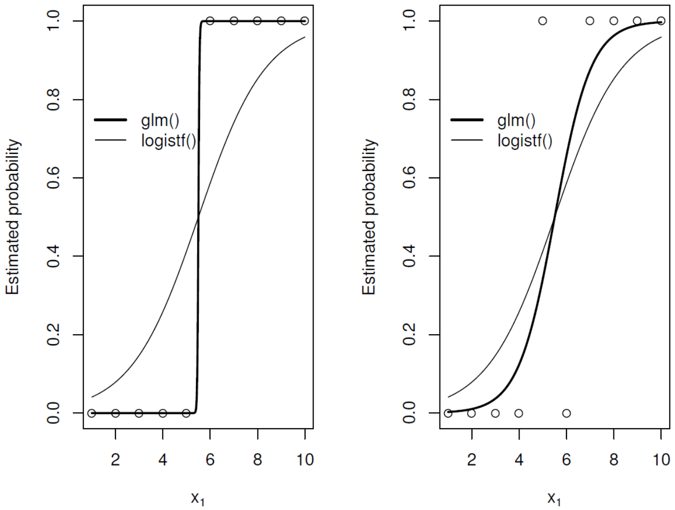
Регуляризованная версия:

$$\sum_{i=1}^{n} (y_i - \pi (x_i) + h_i (0.5 - \pi (x_i)))\, x_{ij},$$

h_i — диагональный элемент hat matrix:

$$H = V^{1/2} X \left(X^T V X\right)^{-1} X^T V^{1/2}.$$

Проблемы настройки параметров



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 $x = x_0$:

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

Линейность логита

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

Методы анализа линейности логита:

- сглаженные диаграммы рассеяния;
- дробные полиномы.

Сглаженные диаграммы рассеяния (smoothed scatterplots, LOESS)

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 $\bar{l}_{sm}\left(x_j\right)$ должен 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, максимизирующие 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. 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.

Содержательный отбор признаков

4. Для непрерывных признаков полученной модели проверяется линейность логита. В случае обнаружения нелинейности признаки заменяются на соответствующие полиномы.
5. Исследуется возможность добавления в полученную модель взаимодействий факторов. Добавляются значимые интерпретируемые взаимодействия.
6. 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?

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

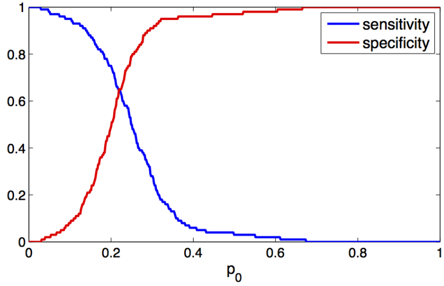
Чаще всего берут $p_0 = 0.5$, но можно выбирать по другим критериям, например, для достижения заданных показателей чувствительности или специфичности.

Порог классификации

Пример: тест на вирус, $p_0 = 0.5$:

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

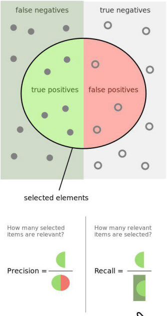
Чувствительность: $\frac{tp}{tp+fn} = \frac{16}{16+131} \approx 10.9\%$.
Специфичность: $\frac{tn}{fp+tn} = \frac{417}{11+417} \approx 97.4\%$.

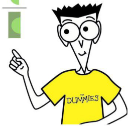


Порог классификации

$$\text{Полнота} = \text{Чувствительность: } \frac{tp}{tp+fn}.$$

$$\text{Точность: } \frac{tp}{fp+fp}.$$





Выбросы

Остатки Пирсона:

$$r_i = \frac{y_i - \hat{\pi}(x_i)}{\sqrt{\hat{\pi}(x_i)(1 - \hat{\pi}(x_i))}}.$$

Аналог расстояния Кука:

$$\Delta\hat{\beta}_i = \frac{r_i^2 h_i}{(1 - h_i)^2}.$$

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.

Натуральный отклик: постановка задачи

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

$$\mathbb{E}\left(y\,|x\,\right) =?$$

Базовый метод — пуассоновская регрессия:

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

Примеры

Стандартная пуассоновская модель:
x_{ij} — макроэкономические показатели, y_i — число банкротств банков,

$$\ln \mu = X\beta.$$

Может использоваться также для нормированных данных:
N_i — общее число банков, $\frac{1000 y_i}{N_i}$ — число банкротств на 1000 банков,

$$\ln \frac{1000\mu}{N} = X\beta, \;\; \ln \mu = \ln \frac{N}{1000} + X\beta.$$

Настройка параметров

ММП:

$$l\left(\beta\right) = \prod_{i=1}^{n} \frac{e^{-e^{x^T\beta}} \left(e^{x^T\beta}\right)^{y_i}}{y_i!},$$

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

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} L\left(\beta\right) \Leftrightarrow$$

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

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:

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

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:

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

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

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:

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

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

- 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.

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

- обработка пропусков — Gu;
- обобщённые линейные модели — Olsson;
- логистическая регрессия — Bilder, глава 2, Hosmer;
- регрессия на счётных данных — Bilder, глава 4, Cameron.

Bilder, C.R., Loughin, T.M. Analysis of Categorical Data with R, 2013.

Cameron C.A., Trivedi P.K. *Regression Analysis of Count Data*, 2013.

Gu X.M. *A Different Approach to the Problem of Missing Data*. In Joint Statistical Meetings, 2015, Seattle, WA.

Hosmer D.W., Lemeshow S., Sturdivant R.X. *Applied Logistic Regression*, 2013.

Olsson U. *Generalized Linear Models: An Applied Approach*, 2004.