Прикладной статистический анализ данных.
8. Дополнения и обобщения регрессии.

Ольга Добролюбова
Юлиан Сердюк
cs.msu.psad@gmail.com

01.04.2022

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.

Оценка коэффициентов регрессии и их ковариационной матрицы методом AC реализована в функции lmac пакета regtools: `https://github.com/matloff/regtools` (устанавливается через install_github пакета devtools).

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

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

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

регрессионная модель:

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

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:

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

	$Pois(\lambda)$	$Bin(N,p)$	$N(\mu,\sigma^2)$
$a(\phi)$	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.

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

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

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

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

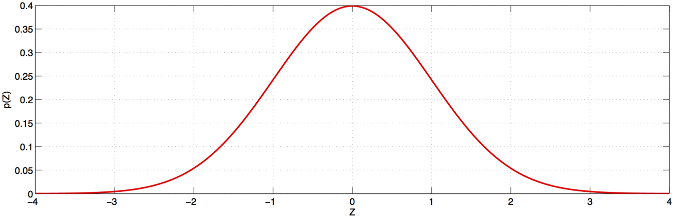
Критерий Вальда

нулевая гипотеза: $H_0: \beta_j = 0$

альтернатива: $H_1: \beta_j <\neq> 0$

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

нулевое распределение: $N(0,1)$



Критерий отношения правдоподобия

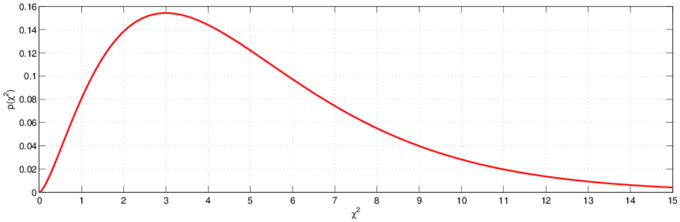
$$\underset{n\times(k+1)}{X} = \left(\underset{n\times(k+1-k_1)}{X_1}, \underset{n\times k_1}{X_2} \right); \quad \underset{(k+1)\times 1}{\beta^T} = \left(\underset{(k+1-k_1)\times 1}{\beta_1^T}, \underset{k_1\times 1}{\beta_2^T} \right)^T;$$

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.

Мультиколлинеарность. Фактор инфляции дисперсии

Рассмотрим линейную модель с константой и n независимыми переменными: $y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \varepsilon$. Тогда оцека дисперсии β_j:

$$\widehat{var}(\hat{\beta}_j) = \frac{s^2}{(n-1)\widehat{var}(x_j)} \cdot \frac{1}{1-R_j^2}$$

Фактор инфляции дисперсии (VIF, variance inflation factor) позволяет оценить увеличение дисперсии заданного коэффициента регрессии, происходящее из-за высокой корреляции данных.

$$VIF_j = \frac{1}{1-R_j^2}$$

Т.е. сравнимаем имеющуюся оценку с ситуацией "переменная имеет нулевую коррлеяцию с другими переменными" с помощью стандартной ошибки.

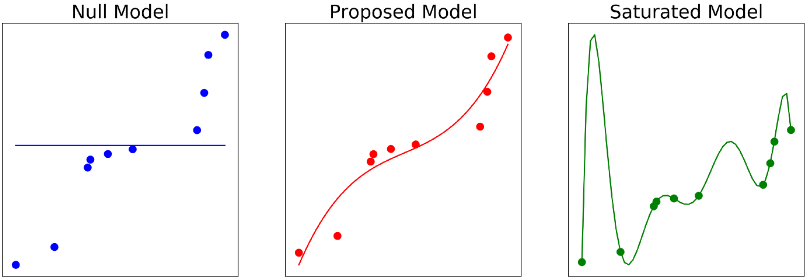
Эмпирическое правило: $VIF_j > 10$ - сильная мультиколлинеарность.

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.



Меры качества моделей

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 $n \geq 8$
- 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

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?

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

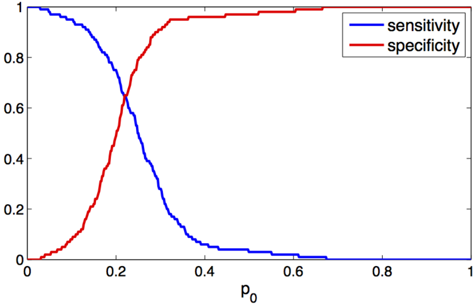
Чаще всего берут $p_0 = 0.5$, но можно выбирать по другим критериям, например, для достижения заданных показателей чувствительности или специфичности.

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: $\frac{417}{11+417} \approx 97.4\%$.



Выбросы

Остатки Пирсона:

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

Аналог расстояния Кука:

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

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.

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.