Логистическая Регрессия

Evgeny Sedashov, PhD
esedashov@hse.ru

5/02/2025

Введение

- В социальных науках часто возникают задачи моделирования бинарных зависимых переменных или бинарной классификации.
- Явка на выборы.
- Состояние войны/мира между двумя странами.
- Выдача/невыдача кредита клиенту банка.
- Классификация автоматизированных аккаунтов в социальных сетях.

Логистическая Регрессия

- В задачах моделирования использование OLS модели является проблематичным – матожидание ошибки для таких данных не равно 0, а распределение ошибки не является нормальным.
- В задачах классификации бинарные категории логически требуют соответствующих методов анализа.
- Мы сконцентрируемся на моделировании, т.к. задача классификации будет детально разбираться в следующем году.

Бинарные Зависимые Переменные

- Бинарные зависимые переменные удобно рассматривать как результаты Бернулли-испытаний.
- Мы уже рассмотрели, как работает метод максимального правдоподобия в контексте Бернулли-последовательностей для самого простого кейса – подброс одинаковой монеты с неизвестной вероятностью орла p.
- Логистическая регрессия не слишком отличается от данного простого сценария.

Функция Правдоподобия I

- Предположим, что мы подбрасываем не одну одинаковую монету, а несколько разных, с разными вероятностями p.

- Иллюстрация процесса – автомат выдаёт Вам монетки K раз, при этом Вы не знаете заранее вероятность орла p_i для каждой конкретной монеты; i – порядковый индекс.

- Функция правдоподобия в этом случае:

$$\mathcal{L} = \prod_{i=1}^{K} p_i^{v_i} (1 - p_i)^{1 - v_i}$$

- $v_i \in \{0; 1\}$, $0 \leq p_i \leq 1$

Функция Правдоподобия II

- Мы не знаем вероятность орла p_i, но есть визуальные свойства монетки, которые могут связаны с этой вероятностью.
- Например, у монетки может быть больше царапен на одной из сторон, что может быть свидетельством о большой вероятности выпадения данной стороны; также одна из сторон может быть более выцветшей.
- Обозначим данные характеристики как матрицу $\mathbf{X}$ и строчку данной матрицы, соответствующую конкретной монетке, как $\mathbf{x}_i$
- Мы можем связать данные характеристики с вероятностью p_i с помощью линк-функции f:

$$p_i = f(\mathbf{b}, \mathbf{x}_i)$$

где $\mathbf{b}$ – вектор параметров, определяющих эффекты предикторов на вероятность p_i.

Функция Правдоподобия III

- Перепишем функцию правдоподобия:

$$\prod_{i=1}^{K} f(\mathbf{b}, \mathbf{x}_i)^{v_i}(1 - f(\mathbf{b}, \mathbf{x}_i))^{1-v_i}$$

- f может быть любой функцией, удовлетворяющей определённым требованиям.
- Самое важное требование: $0 \leq f(\mathbf{b}, \mathbf{x}_i) \leq 1 \; \forall \; \mathbf{b}, \; \mathbf{x}_i$.
- Второе требование: $\mathbf{x}_i \neq \mathbf{x}_j \implies f(\mathbf{b}, \mathbf{x}_i) \neq f(\mathbf{b}, \mathbf{x}_i)$.

Типичные Формы f

- Функция f обычно параметеризируется логистической (логистическая модель) и нормальной (пробит модель) функциями распределения.

- Если используется логистическая функция, то функция f определяется как:

$$f(\mathbf{b}, \mathbf{x}_i) = \frac{1}{1 + exp(-(\mathbf{x}_i \mathbf{b}))}$$

- Если используется нормальная функция, то имеем: **probit** regression:

$$f(\mathbf{b}, \mathbf{x}_i) = \mathbf{\Phi}(\mathbf{b}, \mathbf{x}_i)$$

Стохастическая Функция Полезности

- В экономике, модели бинарного выбора часто выводятся с применением случайной функции полезности.

- Допустим индивид i может выбрать из двух опций, 0 или 1.

- Например, индивид принимает решение, голосовать или нет.

- Полезность голосования может быть выражена как:

$$U_i = \mathbf{x}_i \mathbf{b} + \epsilon_i$$

  где ϵ_i – ошибка (причина, по которой функция называется стохастической).

Выбор Между Альтернативами

- Логично ввести следующее правило принятия решения: if $U_i > 0$, индивид приходит на выборы, в противном случае индивид не приходит на выборы:

$$U_i > 0 \implies Turnout$$

$$U_i \leq 0 \implies \sim Turnout$$

- Подставляя в функцию полезности, имеем:

$$\mathbf{x}_i\mathbf{b} + \epsilon_i > 0 \implies Turnout$$

$$\mathbf{x}_i\mathbf{b} + \epsilon_i \leq 0 \implies \sim Turnout$$

Функция Распределения Ошибки I

- Подставляя, получаем $\epsilon_i > -\mathbf{x}_i\mathbf{b}$.
- Применим оператор вероятности:

$$\mathbb{P}[Turnout] = \mathbb{P}[\epsilon_i > -\mathbf{x}_i\mathbf{b}] = 1 - \mathbb{P}[\epsilon_i \leq -\mathbf{x}_i\mathbf{b}]$$

  где

$$\mathbb{P}[\epsilon_i \leq -\mathbf{x}_i\mathbf{b}]$$

  – функция распределения ошибки.

- Оставшаяся часть рассуждений – параметеризировать функцию распределения.

Функция Распределения Ошибки II

- Можно предположить, что ошибки следуют логистической функции распределения:

$$\mathbb{P}[\mathit{Turnout}] = 1 - \frac{1}{1 + \mathit{exp}(\mathbf{x}_i\mathbf{b})} = \frac{\mathit{exp}(\mathbf{x}_i\mathbf{b})}{1 + \mathit{exp}(\mathbf{x}_i\mathbf{b})} =$$

$$\frac{1}{1 + \mathit{exp}(-\mathbf{x}_i\mathbf{b})} = \mathrm{Logit}(\mathbf{x}_i\mathbf{b})$$

- Ошибки также могут следовать нормальной функции распределения:

$$\mathbb{P}[\mathit{Turnout}] = 1 - \boldsymbol{\Phi}(-\mathbf{x}_i\mathbf{b}) = \boldsymbol{\Phi}(\mathbf{x}_i\mathbf{b})$$

потому что нормальное распределение симметрично относительно среднего.

Метод Максимального Правдоподобия I

- Финальный шаг – выписать функцию правдоподобия:

$$\mathcal{L}(\mathbf{b}, \mathbf{X}) = \prod_{i=1}^{K} \mathbb{P}[Turnout_i]^{v_i}(1 - \mathbb{P}[Turnout_i])^{1-v_i}$$

где v_i – бинарный индикатор явки для индивида i.

- В оптимизационных алгоритмах используется натуральный логарифм от данной функции.

Метод Максимального Правдоподобия II

- Цель – найти вектор $\mathbf{b}$, максимизирующий функцию правдоподобия.
- Итеративные градиентные алгоритмы (градиентный спуск, метод Ньютона).
- Самый простой алгоритм – градиентный спуск с фиксированным правилом апдейта.

Тестирование Гипотез

- Гипотезы тестируются похожим на МНК способом:

$$Z = \frac{\hat{\beta} - \mu_0}{se(\hat{\beta})}$$

где *se* – стандартная ошибка регрессионного коэффициента, а μ_0 – значение для тестирования нулевой гипотезы.

- Базовая идея следующая:

$$\hat{\mathbf{b}} \xrightarrow{d} \mathcal{N}(\mathbf{b}, \mathbf{H}^{-1}_{\log(\mathcal{L})})$$

- Подробные математические доказательства можно найти в Newey, W.K, and McFadden, D. 1994. "Chapter 36 Large sample estimation and hypothesis testing." *Handbook of Econometrics* 4: 2111-2245.

Коэффициенты Регрессии

- Значение коэффициентов регрессии в таких моделях отличается от МНК-регрессии.
- Знаки коэффициентов сохраняют интуитивное значение.
- Величина коэффициентов – более сложный вопрос.
- Есть ряд способов решения данной проблемы.

Логарифм Отношения Шансов I

- Предположим, мы хотим интерпретировать коэффициенты в знакомых нам терминах “единица увеличение x – β единиц изменения в y”.

- Рассмотрим логистическую модель:

$$p_i = \frac{1}{1 + exp(-\mathbf{x}_i\mathbf{b})}$$

- Можно переформатировать данное выражение, получив логарифм отношения шансов (**log-odds**):

$$\mathbf{x}_i\mathbf{b} = \ln(\frac{p_i}{1 - p_i})$$

где $\frac{p_i}{1-p_i}$ – отношение шансов.

Логарифм Отношения Шансов II

- Логарифм отношения шансов имеет простую функциональную форму.
- Увеличение в log-odds означает увеличение вероятности p_i.

Логарифм Отношения Шансов III

- Вычислить $\mathbf{X}\hat{\mathbf{b}}$.
- Извлечь вариационно-ковариационную матрицу $\mathbf{V}$.
- Вычислить
$$\mathbf{X}\mathbf{V}\mathbf{X}^T$$
.
- Извлечь главную диагональ, взять квадратный корень и умножить на 1.96 – обозначим как $\mathbf{ci}$.
- $[\mathbf{X}\hat{\mathbf{b}} + \mathbf{ci}; \mathbf{X}\hat{\mathbf{b}} - \mathbf{ci}]$.

Предсказанные Вероятности

- Более интуитивные и универсальный подход – вычисление самих вероятностей исхода 1.
- Мы определили, что для каждого наблюдения i мы моделируем p_i как $f(\mathbf{x}_i, \mathbf{b})$.
- Чтобы вычислить предсказанные вероятности, нужно просто подставить $\mathbf{X}\hat{\mathbf{b}}$ в f.

Предсказанные Вероятности – Доверительные Интервалы

- Первый подход идентичен подходу лог-оддс, просто рассчитанные доверительные интервалы вставляются в линк-функцию.
- Второй подход (King et. al. (2000)) опирается на идеи о сходимости регрессионных оценок.
- $$\hat{\mathbf{b}} \xrightarrow{d} \mathcal{N}(\mathbf{b}, \mathbf{H}^{-1}_{\log(\mathcal{L})})$$
- King et.al. (2000) CLARIFY использует данную идею для вычисления доверительных интервалов.

Предсказанные Вероятности – Доверительные Интервалы

- **Шаг 1:** используя генератор случайных чисел, сделайте выборку из S наблюдений, следующих многомерному нормальному распределению с параметрами $\hat{\mathbf{b}}$ и $\mathbf{V}$; назовём эту матрицу $\mathbf{D}$ (размерность – S x n, где n – количество независимых переменных).
- **Шаг 2:** для каждого вектора $\mathbf{x}_i$, вычислите $\mathbf{D}\mathbf{x}_i^T$; назовём получившийся вектор $\mathbf{\hat{y}_i}$.
- **Шаг 3:** для каждого вектора $\mathbf{\hat{y}_i}$ вычислите среднюю (или медиану), 2.5 и 97.5 перцентили.
- **Шаг 4:** постройте график с независимой переменной на горизонтальной шкале и предсказанными вероятностями на вертикальной шкале.

Другие Проблемы

- Графическая интерпретация модели обычно предполагает, что все переменные, кроме интересующей нас, устанавливаются на некоторые постоянные значения (средние или медианы).

- Такой подход позволяет делать красивые иллюстрации, но также может вести к неточностям, потому что, в отличие от МНК-модели, эффект конкретной переменной $\mathbf{x}$ в нелинейных моделях может зависеть от всех переменных.

- Есть несколько стандартных способов решения данной проблемы.

Постановка Проблемы

- В некоторыз случаях, данные выглядят как временные промежутки (последовательности непрерывающихся единиц и нулей, time spells).

- Например, пара стран может находиться в состоянии войны несколько лет, в результате чего во всех таких годах индикатор войны $= 1$.

Решение I

- Для подобных данных мы должны учитывать данную временную динамику.
- Carter and Signorino (2010) предложили сконструировать переменную, которая считает количество непрерывающихся лет войны, а затем использует кубический полином данной переменной в регрессионной модели.

Решение II

- В более общем смысле мы имеем

$$\mathbb{P}[War_{it}] = f(\mathbf{x}_{it}\mathbf{b} + g(Duration_{it}))\gamma$$

где $Duration_{it}$ – продолжительность войны в годах для диады i в момент времени t.

- Подход Carter and Signorino (2010):

$$g(Duration_{it}) = Duration_{it} + Duration_{it}^2 + Duration_{it}^3$$

- Beck, Katz and Tucker (1998) предлагали использовать сплайн-функцию в качестве g.