Анализ категориальных данных

Мультиномиальные логит-модели

27 марта 2021

Мультиномиальная модель как набор моделей с бинарным откликом:

Мультиномиальная модель как набор моделей с бинарным откликом:

Случай трех категорий: A, B, C, где C – база

$$\ln \frac{P(y_i = A)}{P(y_i = C)} = \beta_{0,A|C} + \beta_{1,A|C}\, x_i$$

$$\ln \frac{P(y_i = B)}{P(y_i = C)} = \beta_{0,B|C} + \beta_{1,B|C}\, x_i$$

Мультиномиальная модель как набор моделей с бинарным откликом:

Случай трех категорий: A, B, C, где C – база

$$\ln \frac{P(y_i = A)}{P(y_i = C)} = \beta_{0,A|C} + \beta_{1,A|C}\, x_i$$

$$\ln \frac{P(y_i = B)}{P(y_i = C)} = \beta_{0,B|C} + \beta_{1,B|C}\, x_i$$

Выполним преобразование:

$$exp(\ln \frac{P(y_i = A)}{P(y_i = C)}) = exp(\beta_{0,A|C} + \beta_{1,A|C}\, x_i)$$

Мультиномиальная модель как набор моделей с бинарным откликом:

Случай трех категорий: A, B, C, где C – база

$$\ln \frac{P(y_i = A)}{P(y_i = C)} = \beta_{0,A|C} + \beta_{1,A|C}\, x_i$$

$$\ln \frac{P(y_i = B)}{P(y_i = C)} = \beta_{0,B|C} + \beta_{1,B|C}\, x_i$$

Выполним преобразование:

$$exp(\ln \frac{P(y_i = A)}{P(y_i = C)}) = exp(\beta_{0,A|C} + \beta_{1,A|C}\, x_i)$$

$$\frac{P(y_i = A)}{P(y_i = C)} = exp(\beta_{0,A|C} + \beta_{1,A|C}\, x_i)$$

Мультиномиальная модель как набор моделей с бинарным откликом:

Случай трех категорий: A, B, C, где C – база

$$\ln \frac{P(y_i = A)}{P(y_i = C)} = \beta_{0,A|C} + \beta_{1,A|C}\, x_i$$

$$\ln \frac{P(y_i = B)}{P(y_i = C)} = \beta_{0,B|C} + \beta_{1,B|C}\, x_i$$

Выполним преобразование:

$$exp(\ln \frac{P(y_i = A)}{P(y_i = C)}) = exp(\beta_{0,A|C} + \beta_{1,A|C}\, x_i)$$

$$\frac{P(y_i = A)}{P(y_i = C)} = exp(\beta_{0,A|C} + \beta_{1,A|C}\, x_i)$$

$$P(y_i = A) = exp(\beta_{0,A|C} + \beta_{1,A|C}\, x_i) P(y_i = C)$$

Мультиномиальная модель как набор моделей с бинарным откликом (продолжение):

Мультиномиальная модель как набор моделей с бинарным откликом (продолжение):

Случай трех категорий: A, B, C, где C – база

Выразим вероятность попасть в базовую категорию:

$$P(y_i = C) = 1 - \sum_{j=A}^{B} P(y_i = j) = 1 - \sum_{j=A}^{B} P(y_i = C) exp(\beta_{0j} + \beta_{1j} x_i)$$

$$= 1 - P(y_i = C) \sum_{j=A}^{B} exp(\beta_{0j} + \beta_{1j} x_i)$$

Мультиномиальная модель как набор моделей с бинарным откликом (продолжение):

Случай трех категорий: A, B, C, где C – база

Выразим вероятность попасть в базовую категорию:

$$P(y_i = C) = 1 - \sum_{j=A}^{B} P(y_i = j) = 1 - \sum_{j=A}^{B} P(y_i = C) exp(\beta_{0j} + \beta_{1j} x_i)$$

$$= 1 - P(y_i = C) \sum_{j=A}^{B} exp(\beta_{0j} + \beta_{1j} x_i)$$

$$P(y_i = C) = \frac{1}{1 + \sum_{j=A}^{B} exp(\beta_{0j} + \beta_{1j} x_i)},$$ следовательно, для A:

Мультиномиальная модель как набор моделей с бинарным откликом (продолжение):

Случай трех категорий: A, B, C, где C – база

Выразим вероятность попасть в базовую категорию:

$$P(y_i = C) = 1 - \sum_{j=A}^{B} P(y_i = j) = 1 - \sum_{j=A}^{B} P(y_i = C) exp(\beta_{0j} + \beta_{1j} x_i)$$

$$= 1 - P(y_i = C) \sum_{j=A}^{B} exp(\beta_{0j} + \beta_{1j} x_i)$$

$$P(y_i = C) = \frac{1}{1 + \sum_{j=A}^{B} exp(\beta_{0j} + \beta_{1j} x_i)},$$ следовательно, для A:

$$P(y_i = A) = \frac{exp(\beta_{0,A|C} + \beta_{1,A|C}\, x_i)}{1 + \sum_{j=A}^{B} exp(\beta_{0j} + \beta_{1j} x_i)}$$

Очень хочется перейти к «шансам»:

$$\frac{P(y_i = A)}{P(y_i = C)} = \frac{\dfrac{exp(\beta_{0,A|C} + \beta_{1,A|C}\, x_i)}{1 + \sum_{j=A}^{B} exp(\beta_{0j} + \beta_{1j} x_i)}}{\dfrac{1}{1 + \sum_{j=A}^{B} exp(\beta_{0j} + \beta_{1j} x_i)}} = exp(\beta_{0,A|C} + \beta_{1,A|C}\, x_i)$$

Очень хочется перейти к «шансам»:

$$\frac{P(y_i = A)}{P(y_i = C)} = \frac{\dfrac{exp(\beta_{0,A|C} + \beta_{1,A|C}\, x_i)}{1 + \sum_{j=A}^{B} exp(\beta_{0j} + \beta_{1j} x_i)}}{\dfrac{1}{1 + \sum_{j=A}^{B} exp(\beta_{0j} + \beta_{1j} x_i)}} = exp(\beta_{0,A|C} + \beta_{1,A|C}\, x_i)$$

А шансы ли это?

У S. Long Вы прочитаете привычную интерпретацию в терминах отношения шансов, однако, строго говоря, это не шансы, а риски (risk ratio). Вместо отношения шансов в мультиномиальной модели считаются отношения вероятностей.

Допущение о независимости от посторонних альтернатив (IIA)

Допущение о независимости от посторонних альтернатив (IIA)

Риски попасть в категорию не зависят от альтернатив.

Пример

A - черные ботинки; $P(A) = \dfrac{1}{2}$

B - черные туфли; $P(B) = \dfrac{1}{2}$

Допущение о независимости от посторонних альтернатив (IIA)

Риски попасть в категорию не зависят от альтернатив.

Пример

A - черные ботинки; $P(A) = \dfrac{1}{2}$

B - черные туфли; $P(B) = \dfrac{1}{2}$

Вдруг вмешались C - красные туфли. Тогда если $P(A) = \dfrac{1}{3}$; если $P(B) = \dfrac{1}{3}$; если $P(C) = \dfrac{1}{3}$, то соблюдается IIA.

Допущение о независимости от посторонних альтернатив (IIA)

Риски попасть в категорию не зависят от альтернатив.

Пример

A - черные ботинки; $P(A) = \frac{1}{2}$

B - черные туфли; $P(B) = \frac{1}{2}$

Вдруг вмешались C - красные туфли. Тогда если $P(A) = \frac{1}{3}$; если $P(B) = \frac{1}{3}$; если $P(C) = \frac{1}{3}$, то соблюдается IIA.

Вопрос

В каком случае IIA будет нарушаться?

Тестирование IIA

Hausman-McFadden test

1. H0: IIA
2. Переоценить модель на сокращенной выборке (без категории J)
3. $S = (\hat{b_r} - \hat{b_f})^T (Cov(\hat{b_r}) - Cov(\hat{b_f}))^{-1} (\hat{b_r} - \hat{b_f}) \sim \chi^2_k$, где k – количество параметров в restricted model; $\hat{b_r}$ – оценки в restricted model; $\hat{b_f}$ – оценки в полной модели