Анализ категориальных данных

Модели с порядковым откликом

27 февраля 2021

Спецификация модели: подход, основанный на ЛАТЕНТНОЙ зависимой переменной

Спецификация модели: подход, основанный на ЛАТЕНТНОЙ зависимой переменной

Мы допускаем, что существует некоторая ненаблюдаемая переменная y_i^*, принимающая любые значения $(-\infty; +\infty)$

$y_i = j$ (определенная категория наблюдаемой переменной), если $c_{j-1} \leq y_i^* < c_j$, где с – cutpoint (пороговое значение)

На основе значений y_i^* определяются значения исходного y_i.
Для крайних категорий:
Если $-\infty \leq y_i^* < c_1$, то $y_i = 1$
Если $c_{J-1} \leq y_i^* < \infty$, то $y_i = J$

Спецификация модели: подход, основанный на ЛАТЕНТНОЙ зависимой переменной

Латентная зависимая переменная линейным образом связана с предикторами:
$y_i^* = \beta_0 + \beta x_i + e_i$
Так как отклик ненаблюдаемый, нам нужны допущения о распределении ошибок:

1. $\epsilon \sim N(0, 1)$ (probit-model)
2. стандартное логистическое распределение $\epsilon \approx N(0, 3.29)$ (logit-model)

Покажем, как рассчитывается вероятность того, что наблюдаемая зависимая переменная принимает конкретное значение.

Покажем, как рассчитывается вероятность того, что наблюдаемая зависимая переменная принимает конкретное значение.

$P(y_i = j|x) =$

Покажем, как рассчитывается вероятность того, что наблюдаемая зависимая переменная принимает конкретное значение.

$$P(y_i = j|x) = P(c_{j-1} \leq y_i^* < c_j|x)$$

Покажем, как рассчитывается вероятность того, что наблюдаемая зависимая переменная принимает конкретное значение.

$$P(y_i = j|x) = P(c_{j-1} \leq y_i^* < c_j|x) = P(c_{j-1} \leq \beta_0 + \beta x_i + e_i < c_j|x)$$

Покажем, как рассчитывается вероятность того, что наблюдаемая зависимая переменная принимает конкретное значение.

$P(y_i = j|x) = P(c_{j-1} \leq y_i^* < c_j|x) = P(c_{j-1} \leq \beta_0 + \beta x_i + e_i < c_j|x) = P(c_{j-1} - \beta_0 - \beta x_i \leq e_i < c_j - \beta_0 - \beta x_i|x) = F(c_j - \beta_0 - \beta x_i) - F(c_{j-1} - \beta_0 - \beta x_i)$, где F – функция распределения.

Покажем, как рассчитывается вероятность того, что наблюдаемая зависимая переменная принимает конкретное значение.

$P(y_i = j|x) = P(c_{j-1} \leq y_i^* < c_j|x) = P(c_{j-1} \leq \beta_0 + \beta x_i + e_i < c_j|x) = P(c_{j-1} - \beta_0 - \beta x_i \leq e_i < c_j - \beta_0 - \beta x_i|x) = F(c_j - \beta_0 - \beta x_i) - F(c_{j-1} - \beta_0 - \beta x_i)$, где F – функция распределения.

Для крайних категорий:

$P(y_i = 1) = F(c_1 - \beta_0 - \beta x_i) - F(-\infty - \beta_0 - \beta x_i)$

Покажем, как рассчитывается вероятность того, что наблюдаемая зависимая переменная принимает конкретное значение.

$P(y_i = j|x) = P(c_{j-1} \leq y_i^* < c_j|x) = P(c_{j-1} \leq \beta_0 + \beta x_i + e_i < c_j|x) = P(c_{j-1} - \beta_0 - \beta x_i \leq e_i < c_j - \beta_0 - \beta x_i|x) = F(c_j - \beta_0 - \beta x_i) - F(c_{j-1} - \beta_0 - \beta x_i)$, где F – функция распределения.

Для крайних категорий:

$P(y_i = 1) = F(c_1 - \beta_0 - \beta x_i) - F(-\infty - \beta_0 - \beta x_i) = F(c_1 - \beta_0 - \beta x_i)$

Покажем, как рассчитывается вероятность того, что наблюдаемая зависимая переменная принимает конкретное значение.

$P(y_i = j|x) = P(c_{j-1} \leq y_i^* < c_j|x) = P(c_{j-1} \leq \beta_0 + \beta x_i + e_i < c_j|x) = P(c_{j-1} - \beta_0 - \beta x_i \leq e_i < c_j - \beta_0 - \beta x_i|x) = F(c_j - \beta_0 - \beta x_i) - F(c_{j-1} - \beta_0 - \beta x_i)$, где F – функция распределения.

Для крайних категорий:

$P(y_i = 1) = F(c_1 - \beta_0 - \beta x_i) - F(-\infty - \beta_0 - \beta x_i) = F(c_1 - \beta_0 - \beta x_i)$

$P(y_i = J) = F(\infty - \beta_0 - \beta x_i) - F(c_{J-1} - \beta_0 - \beta x_i)$

Покажем, как рассчитывается вероятность того, что наблюдаемая зависимая переменная принимает конкретное значение.

$P(y_i = j|x) = P(c_{j-1} \leq y_i^* < c_j|x) = P(c_{j-1} \leq \beta_0 + \beta x_i + e_i < c_j|x) = P(c_{j-1} - \beta_0 - \beta x_i \leq e_i < c_j - \beta_0 - \beta x_i|x) = F(c_j - \beta_0 - \beta x_i) - F(c_{j-1} - \beta_0 - \beta x_i)$, где F – функция распределения.

Для крайних категорий:

$P(y_i = 1) = F(c_1 - \beta_0 - \beta x_i) - F(-\infty - \beta_0 - \beta x_i) = F(c_1 - \beta_0 - \beta x_i)$

$P(y_i = J) = F(\infty - \beta_0 - \beta x_i) - F(c_{J-1} - \beta_0 - \beta x_i) = 1 - F(c_{J-1} - \beta_0 - \beta x_i)$

Исходя из предыдущего определения $P(y_i = j|x)$, покажем, чему равна вероятность того, что наблюдаемая зависимая переменная принимает значение, НЕ превышающее указанную категорию (функция распределения от j):

Исходя из предыдущего определения $P(y_i = j|x)$, покажем, чему равна вероятность того, что наблюдаемая зависимая переменная принимает значение, НЕ превышающее указанную категорию (функция распределения от j):

Пусть j = 3, тогда
$F(y_i = 3|x) = P(y_i = 1|x) + P(y_i = 2|x) + P(y_i = 3) =$

Исходя из предыдущего определения $P(y_i = j|x)$, покажем, чему равна вероятность того, что наблюдаемая зависимая переменная принимает значение, НЕ превышающее указанную категорию (функция распределения от j):

Пусть j = 3, тогда
$F(y_i = 3|x) = P(y_i = 1|x) + P(y_i = 2|x) + P(y_i = 3) =$
$F(c_1 - \beta_0 - \beta x_i) +$

Исходя из предыдущего определения $P(y_i = j|x)$, покажем, чему равна вероятность того, что наблюдаемая зависимая переменная принимает значение, НЕ превышающее указанную категорию (функция распределения от j):

Пусть j = 3, тогда
$F(y_i = 3|x) = P(y_i = 1|x) + P(y_i = 2|x) + P(y_i = 3) =$
$F(c_1 - \beta_0 - \beta x_i) + F(c_2 - \beta_0 - \beta x_i) - F(c_1 - \beta_0 - \beta x_i) +$

Исходя из предыдущего определения $P(y_i = j|x)$, покажем, чему равна вероятность того, что наблюдаемая зависимая переменная принимает значение, НЕ превышающее указанную категорию (функция распределения от j):

Пусть j = 3, тогда
$F(y_i = 3|x) = P(y_i = 1|x) + P(y_i = 2|x) + P(y_i = 3) = F(c_1 - \beta_0 - \beta x_i) + F(c_2 - \beta_0 - \beta x_i) - F(c_1 - \beta_0 - \beta x_i) + F(c_3 - \beta_0 - \beta x_i) - F(c_2 - \beta_0 - \beta x_i) =$

Исходя из предыдущего определения $P(y_i = j|x)$, покажем, чему равна вероятность того, что наблюдаемая зависимая переменная принимает значение, НЕ превышающее указанную категорию (функция распределения от j):

Пусть j = 3, тогда
$F(y_i = 3|x) = P(y_i = 1|x) + P(y_i = 2|x) + P(y_i = 3) = F(c_1 - \beta_0 - \beta x_i) + F(c_2 - \beta_0 - \beta x_i) - F(c_1 - \beta_0 - \beta x_i) + F(c_3 - \beta_0 - \beta x_i) - F(c_2 - \beta_0 - \beta x_i) = F(c_3 - \beta_0 - \beta x_i)$, где F – функция распределения.

Исходя из предыдущего определения $P(y_i = j|x)$, покажем, чему равна вероятность того, что наблюдаемая зависимая переменная принимает значение, НЕ превышающее указанную категорию (функция распределения от j):

Пусть j = 3, тогда
$F(y_i = 3|x) = P(y_i = 1|x) + P(y_i = 2|x) + P(y_i = 3) = F(c_1 - \beta_0 - \beta x_i) + F(c_2 - \beta_0 - \beta x_i) - F(c_1 - \beta_0 - \beta x_i) + F(c_3 - \beta_0 - \beta x_i) - F(c_2 - \beta_0 - \beta x_i) = F(c_3 - \beta_0 - \beta x_i)$, где F – функция распределения.
Для крайних категорий:
$F(y_i = 1|x) = F(c_1 - \beta_0 - \beta x_i)$

Исходя из предыдущего определения $P(y_i = j|x)$, покажем, чему равна вероятность того, что наблюдаемая зависимая переменная принимает значение, НЕ превышающее указанную категорию (функция распределения от j):

Пусть j = 3, тогда
$F(y_i = 3|x) = P(y_i = 1|x) + P(y_i = 2|x) + P(y_i = 3) = F(c_1 - \beta_0 - \beta x_i) + F(c_2 - \beta_0 - \beta x_i) - F(c_1 - \beta_0 - \beta x_i) + F(c_3 - \beta_0 - \beta x_i) - F(c_2 - \beta_0 - \beta x_i) = F(c_3 - \beta_0 - \beta x_i)$, где F – функция распределения.
Для крайних категорий:
$F(y_i = 1|x) = F(c_1 - \beta_0 - \beta x_i)$
$F(y_i = J|x) = 1$

Второй подход к спецификации модели: через ШАНСЫ, БЕЗ латентного y_i^*:

Второй подход к спецификации модели: через ШАНСЫ, БЕЗ латентного y_i^*:

1. Перейдем от $P(y_i = j)$ к шансам $\dfrac{P(y_i \leq j)}{P(y_i > j)} = \dfrac{P(y_i \leq j)}{1 - P(y_i \leq j)}$

Второй подход к спецификации модели: через ШАНСЫ, БЕЗ латентного y_i^*:

1. Перейдем от $P(y_i = j)$ к шансам $\dfrac{P(y_i \leq j)}{P(y_i > j)} = \dfrac{P(y_i \leq j)}{1 - P(y_i \leq j)}$
2. В допущении о стандартном логистическом распределении шансы можно представить:

$$\frac{\dfrac{exp(c_j - \beta_0 - \beta x_i)}{1 + exp(c_j - \beta_0 - \beta x_i)}}{1 - \dfrac{exp(c_j - \beta_0 - \beta x_i)}{1 + exp(c_j - \beta_0 - \beta x_i)}} = exp(c_j - \beta_0 - \beta x_i)$$

Второй подход к спецификации модели: через ШАНСЫ, БЕЗ латентного y_i^*:

1. Перейдем от $P(y_i = j)$ к шансам $\dfrac{P(y_i \le j)}{P(y_i > j)} = \dfrac{P(y_i \le j)}{1 - P(y_i \le j)}$
2. В допущении о стандартном логистическом распределении шансы можно представить:
$$\frac{\dfrac{exp(c_j - \beta_0 - \beta x_i)}{1 + exp(c_j - \beta_0 - \beta x_i)}}{1 - \dfrac{exp(c_j - \beta_0 - \beta x_i)}{1 + exp(c_j - \beta_0 - \beta x_i)}} = exp(c_j - \beta_0 - \beta x_i)$$
3. $\ln(\dfrac{P(y_i \le j)}{P(y_i > j)} = c_j - \beta_0 - \beta x_i$

Классическая модель с порядковым откликом (пропорциональных шансов) оценивается в допущении о параллельности регрессий

Классическая модель с порядковым откликом (пропорциональных шансов) оценивается в допущении о параллельности регрессий

Эффект предиктора одинаковый для любой кумулятивной логит-модели: к примеру, при сравнении 1-ой категории и всех остальных, при сравнении 1,2 и всех остальных и т.д.

$P(y \leq j|x) = F(c_j - \beta_0 - \beta x_i)$, то есть, меняется только константа, а эффект переменных остается постоянным.

Условие параллельности регрессий нужно тестировать:

Условие параллельности регрессий нужно тестировать:

1. предварительно можно оценить набор логистических моделей с бинарным откликом: новая зависимая переменная $= 1$, если $y > j$, 0 – в противном случае (или наоборот). Далее сравнить оценки коэффициентов в $J - 1$ моделях

Условие параллельности регрессий нужно тестировать:

1. предварительно можно оценить набор логистических моделей с бинарным откликом: новая зависимая переменная $= 1$, если $y > j$, 0 – в противном случае (или наоборот). Далее сравнить оценки коэффициентов в $J - 1$ моделях
2. тест Бранта (принцип как в тесте Вальда): можно протестировать гипотезу о параллельности как для отдельных предикторов, так и в целом для всей модели

Что делать, если условие о параллельности регрессий нарушается?

Что делать, если условие о параллельности регрессий нарушается?

1. оценить gologit (generalized ordered logit) без ограничений (наименее экономная модель): равносильно оцениванию $J-1$ моделей с бинарным откликом (см. предыдущий слайд)

Что делать, если условие о параллельности регрессий нарушается?

1. оценить gologit (generalized ordered logit) без ограничений (наименее экономная модель): равносильно оцениванию $J-1$ моделей с бинарным откликом (см. предыдущий слайд)
2. оценить модель с частично пропорциональными шансами (partial proportional odds): ослабить допущение о параллельности только для некоторых переменных

Что делать, если условие о параллельности регрессий нарушается?

1. оценить gologit (generalized ordered logit) без ограничений (наименее экономная модель): равносильно оцениванию $J-1$ моделей с бинарным откликом (см. предыдущий слайд)
2. оценить модель с частично пропорциональными шансами (partial proportional odds): ослабить допущение о параллельности только для некоторых переменных
3. посредством теста отношения правдоподобия (likelihood-ratio test) сравнить альтернативные спецификации и выбрать оптимальную