Анализ категориальных данных

Занятие 1. Модели бинарного выбора: спецификация

19 января 2021

Вопрос

Покажем, что такое линейная вероятностная модель (linear probability model)

Вопрос

Покажем, что такое линейная вероятностная модель (linear probability model)

Ответ

Это результат оценивания классической линейной регрессии применительно к случаю бинарного отклика:
$y_i = \beta_0 + \beta x_i + e_i$, где y_i принимает только два значения, где, к примеру, 1 – приняли рукопись к публикации, 0 – в противном случае.

Вопрос

Покажем, что такое линейная вероятностная модель (linear probability model)

Ответ

Это результат оценивания классической линейной регрессии применительно к случаю бинарного отклика:
$y_i = \beta_0 + \beta x_i + e_i$, где y_i принимает только два значения, где, к примеру, 1 – приняли рукопись к публикации, 0 – в противном случае.
В этом случае предсказанное значение отклика ($\hat{y}_i$) – это вероятность того, что Y принимает значение 1:
$E(y_i|x_i) = 1 \times P(y_i = 1|x_i) + 0 \times P(y_i = 0|x_i) = P(y_i = 1|x_i)$

Вопрос

В чем основные ограничения линейной вероятностной модели?

Вопрос

В чем основные ограничения линейной вероятностной модели?

Ответ

1. Предсказанные значения отклика выходят за допустимые границы, может быть меньше 0 или больше 1

Вопрос

В чем основные ограничения линейной вероятностной модели?

Ответ

1. Предсказанные значения отклика выходят за допустимые границы, может быть меньше 0 или больше 1
2. Содержательно не всегда правдоподобной является линейная взаимосвязь вероятности «успеха» и объясняющей переменной

Вопрос

Рассмотрим альтернативу. В чем суть подхода, основанного на латентной зависимой переменной?

Вопрос

Рассмотрим альтернативу. В чем суть подхода, основанного на латентной зависимой переменной?

Ответ

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

Условно ее можно интерпретировать как склонность к «успеху» (склонность к тому, что наблюдаемый $y_i = 1$)

На основе значений y_i^* определяются значения исходного y_i.
Если $y_i^* > 0$, то $y_i = 1$
Если $y_i^* \leq 0$, то $y_i = 0$

Вопрос

Запишем спецификацию модели с y_i^* в качестве отклика. Какие допущения делаем об ошибках?

Вопрос

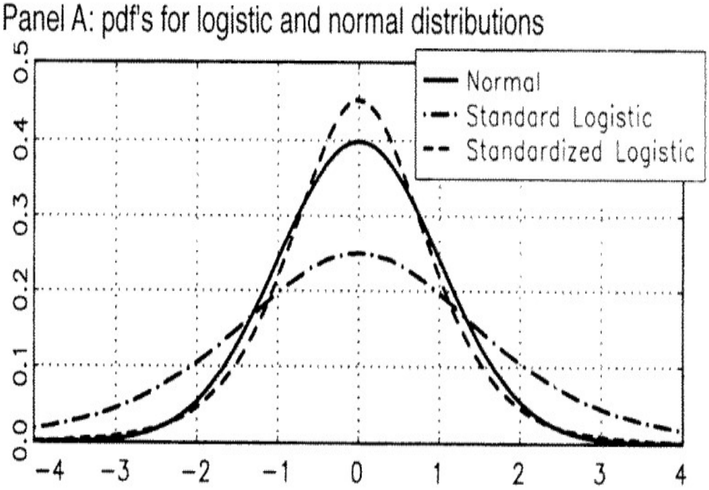
Запишем спецификацию модели с y_i^* в качестве отклика. Какие допущения делаем об ошибках?

Ответ

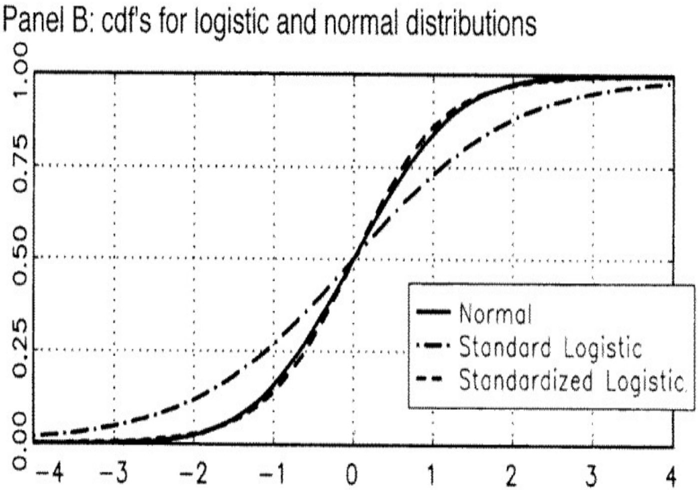
Важно, что латентная зависимая переменная линейным образом связана с объясняющими переменными:
$y_i^* = \beta_0 + \beta x_i + e_i$
Так как отклик ненаблюдаемый, нам нужны допущения о распределении ошибок:

1. $e \sim N(0, 1)$ (probit-model)
2. стандартное логистическое распределение $e \approx N(0, 3.29)$ (logit-model). $F(e) = \dfrac{exp(e)}{1 + exp(e)}$

Графики функций плотности



Графики функций распределения



Вопрос

Покажем, что $P(y_i = 1) = F(\beta_0 + \beta x_i)$, где F – функция распределения.

Вопрос

Покажем, что $P(y_i = 1) = F(\beta_0 + \beta x_i)$, где F – функция распределения.

Ответ

$P(y_i = 1) =$

Вопрос

Покажем, что $P(y_i = 1) = F(\beta_0 + \beta x_i)$, где F – функция распределения.

Ответ

$P(y_i = 1) = P(y_i^* > 0) =$

Вопрос

Покажем, что $P(y_i = 1) = F(\beta_0 + \beta x_i)$, где F – функция распределения.

Ответ

$P(y_i = 1) = P(y_i^* > 0) = P(\beta_0 + \beta x_i + e_i > 0) =$

Вопрос

Покажем, что $P(y_i = 1) = F(\beta_0 + \beta x_i)$, где F – функция распределения.

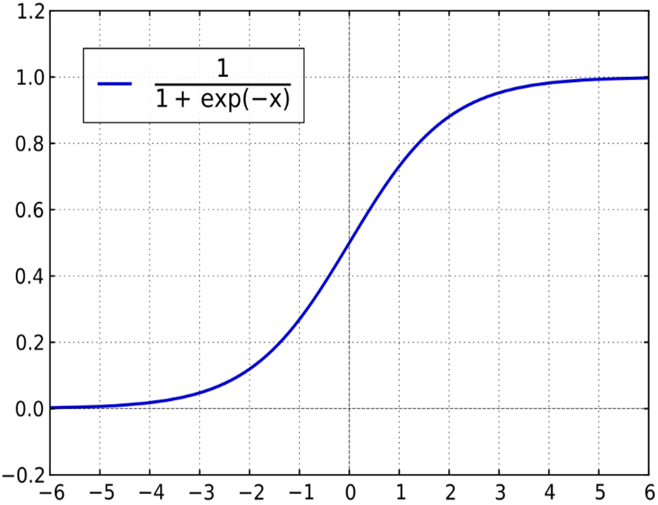
Ответ

$P(y_i = 1) = P(y_i^* > 0) = P(\beta_0 + \beta x_i + e_i > 0) = P(e_i \leq \beta_0 + \beta x_i)$, а функция распределения – это и есть вероятность того, что сл. величина не превышает указанное значение.

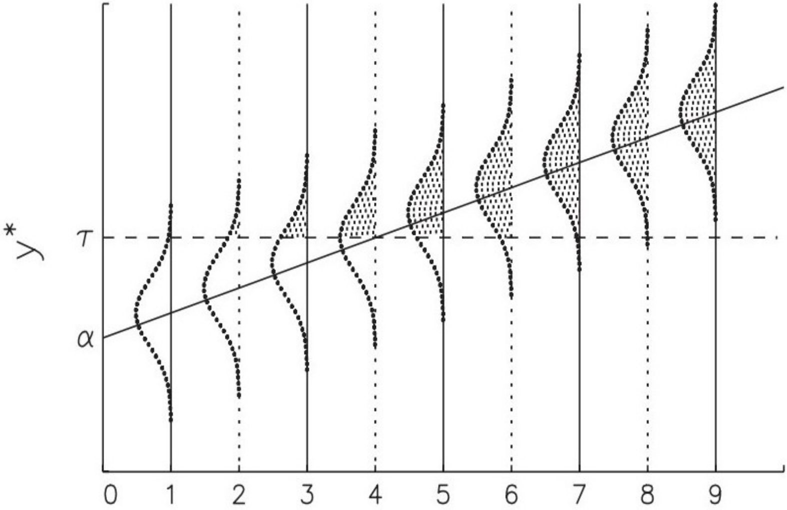
К примеру, для логит-модели:

$$P(y_i = 1) = F(\beta_0 + \beta x_i) = \frac{exp(\beta_0 + \beta x_i)}{1 + exp(\beta_0 + \beta x_i)}$$

Зависимость P(Y = 1) от X ...



…в результате той самой ползущей «улитки»



Можно обойтись и без латентного y_i^*:

Можно обойтись и без латентного y_i^*:

Ответ

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

Можно обойтись и без латентного y_i^*:

Ответ

1. Перейдем от $P(y_i = 1)$ к шансам $\dfrac{P(y_i = 1)}{1 - P(y_i = 1)}$
2. Запишем $P(y_i = 1)$ как функцию распределения:
$$\frac{\dfrac{exp(\beta_0 + \beta x_i)}{1 + exp(\beta_0 + \beta x_i)}}{1 - \dfrac{exp(\beta_0 + \beta x_i)}{1 + exp(\beta_0 + \beta x_i)}} = exp(\beta_0 + \beta x_i)$$

Можно обойтись и без латентного y_i^*:

Ответ

1. Перейдем от $P(y_i = 1)$ к шансам $\dfrac{P(y_i = 1)}{1 - P(y_i = 1)}$
2. Запишем $P(y_i = 1)$ как функцию распределения:
$$\frac{\dfrac{exp(\beta_0 + \beta x_i)}{1 + exp(\beta_0 + \beta x_i)}}{1 - \dfrac{exp(\beta_0 + \beta x_i)}{1 + exp(\beta_0 + \beta x_i)}} = exp(\beta_0 + \beta x_i)$$
3. $\ln\left(\dfrac{P(y_i = 1)}{1 - P(y_i = 1)}\right) = \beta_0 + \beta x_i$ (логит линейным образом связан с объясняющими переменными)