Бинарные модели

Определение 1.1. Переменная (или фактор) называется **дискретной**, если она принимает только целые конечные значения.

Бывают переменные:

Количественные (число детей в семье);

Качественные (да, нет и т.д.);

Порядковые, когда выбор ранжированный (упорядоченная альтернатива: низкий, средний, высокий).

Определение 1.2. Модели с дискретно-зависимой переменной называются **моделями множественного выбора** в случае, когда зависимые переменные принимают два значения, называют **моделями бинарного выбора.**

Определение модели бинарного выбора.

Если y – зависимая переменная, принимающая значения:
$y_i = 0$ и 1.
А $X=(x_1, x_2, ...,x_k)$ – независимые переменные;
$B = (b_1, b_2, ...,b_k)$ – вектор коэффициентов,
то линейная модель регрессии примет вид:
$y_i = b_1 \bullet x_1 + b_2 \bullet x_2 + ... + b_k \bullet x_k + \varepsilon_i$,
где $i = 1$ до n.
n – число наблюдений в каждой из переменных.
y_i принимает значения 0 и 1.
Следовательно, $M(\varepsilon_i) = 0$ – математическое ожидание.

Математическое ожидание

y_i:	0	1
P:	$p(y_i=0)$	$p(y_i=1)$

$M(y_i) = 1 \bullet p(y_i=1) + 0 \bullet p(y_i=0) = p(y_i=1) = X^T B^T$,
T- транспонированное, т.е. $p(y_i=1) = X^T B^T$ **(1.1)** или $p(y_i=0) = 1 - X^T B^T$
(1.1) **– модель линейной вероятности**.

Невозможность применения МНК

Рассмотрим однофакторную модель $y_i = a + b \bullet x_i + \varepsilon_i$, где y – бинарная.

Если к оценке данной модели применить МНК, то получим:

1) $y_{расчетное}$. Может быть $0 < y_{расчетное}. < 1$, что противоречит бинарности зависимой переменной.

2) Дисперсия остатков зависит от x_i.

$y_p = b \bullet x_i$; *тогда*

$\varepsilon_1 = b \bullet x_i$;

$\varepsilon_2 = 1 - b \bullet x_i$;

$D(\varepsilon_i) = b \bullet x_i \bullet (1 - b \bullet x_i)$ – т.е. дисперсия зависит от x, то при росте x дисперсия растет, т.е. имеет место гетероскедастичность остатков.

3. Т.к. ε_i может принимать два значения с вероятностью $p(y_i=1)$ и $1- p(y_i=1)$, следовательно, остатки не являются нормально распределенными величинами.

Т.о. нарушаются три предпосылки МНК. Следовательно, для моделирования значений модели (1.1) подбирают функции область значений, которых определяется [0;1], а выражение $b \bullet x_i$ играют роль аргумента этой функции.

$$\boldsymbol{P(y_i=1) = F(X_i B)} \text{ – } \textit{непрерывная и неубывающая.}$$

Выбор функции *F* определенный тип бинарной модели.

Функция стандартного нормального распределения

$$F(u) = \Phi(u) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz \qquad \textbf{(1.2)}$$

Нормальное стандартное распределение подразумевает, что мат. ожидание = 0, а среднеквадратичное отклонение σ=1.

Определение 1.3. Если бинарная модель имеет в качестве функции распределения функцию вида (1.2), то эта модель называется **Пробит – моделью**.

Функция стандартного логистического распределения

$$F(u) = \Lambda(u) = \frac{e^u}{1+e^u} \qquad \textbf{(1.3)}$$

Определение 1.4. Если бинарная модель имеет в качестве функции распределения функцию вида (1.3), то эта модель называется **Логит – моделью**

Функция экстремального (или Гомперца) распределения

$$F(u) = E(u) = e^{e^{-u}} \qquad \textbf{(1.4)}$$

Определение 1.4. Если бинарная модель имеет в качестве функции распределения функцию вида (1.4), то эта модель называется **экстрим – моделью или гомпит-моделью**

Селекция бинарных моделей

Спецификацию логит, пробит и гомпит модели проводят на основании теоретических предпосылок, а также исходя из минимума значений информационных критериев *Акайке, Шварца и Хана-Квина*.

$$AC = \ln(\sigma^2) + \frac{2k}{n}$$

$$SC = \ln(\sigma^2) + \frac{k \cdot \ln(n)}{n}$$

$$HQ = \ln(\sigma^2) + 2\frac{k \cdot \ln(\ln(n))}{n}$$

здесь n – общее число наблюдений ряда данных, k – число степеней свободы модели (равно числу факторов в модели +1)

σ^2 – остаточная или объясненная моделью дисперсия.

Маржинальные эффекты

Коэффициенты бинарной модели не могут интерпретироваться как предельный коэффициент влияния объясняющих переменных на зависимую.

Предельный коэффициент каждого объясняющего фактора x_j , j=1,..,k является непрерывным и зависит от значения остальных факторов и определяется:

$$\frac{\partial P(y=1)}{\partial x} = b \cdot F'(x^{T \cdot b}) = b \cdot f(x^{T \cdot b})$$

, где f - плотность вероятности

Для пробит-модели: $$\frac{\partial P(y=1)}{\partial x} = b \cdot \Phi'(x^{T \cdot b}) = b \cdot \varphi(x^{T \cdot b})$$, где $$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

Для логит-модели: $$\frac{\partial P(y=1)}{\partial x} = b \cdot \Lambda'(x^{T \cdot b}) = b \cdot \lambda(x^{T \cdot b})$$, где $$\lambda(u) = \frac{e^u}{(1+e^u)^2}$$

Для гомпит-модели: $$\frac{\partial P(y=1)}{\partial x} = b \cdot E(x^{T \cdot b}) = -b \cdot e^{e^{-x^{T \cdot b}}} \cdot e^{-x^{T \cdot b}}$$

Направление изменений эффекта зависит только от знака коэффициента регрессии.

Оценка моделей ММП

Для оценки параметров бинарных моделей применяют метод максимального ***правдоподобия с функцией правдоподобия:***

$$L=L(y_1,\dots,y_n)=\begin{cases} y_i=0;\ \ P(y_i=0)=1-P(y_i=1)=1-F(x^T b) \\ y_i=1;\ \ P(y_i=1)=F(x^T b) \end{cases}$$

y_i – рассмотрим как n случайных величин Y_i с одним возможным значением y_i. Эти случайные величины независимы. Их совместная вероятность = произведению их вероятности:

$$L=\prod_{y_i=0}(1-F(x^T\cdot b))\prod_{y_i=1}F(x^T\cdot b)=\prod_i(1-F(x^T\cdot b))^{y_i}F(x^T\cdot b)^{1-y_i}$$

Прологарифмируем выражение

Логарифмическая функция правдоподобия имеет вид:

$$l=\ln L=\sum_{y_i=0}y_i\ln F(x_i^T\cdot b)+\sum_{y_i=1}(1-y_i)\ln\left[1-F(x_i^T\cdot b)\right]$$

Для нахождения максимума необходимо найти частные производные по параметрам и приравнять их к «0». Решаем дифференциальное **уравнение правдоподобия**:

$$\frac{\partial l}{\partial b}=0 \text{ или } \sum_i\left(\frac{y_i f\left(x_i^T b\right)}{F\left(x_i^T b\right)}-\frac{(1-y_i)f\left(x_i^T b\right)}{1-F\left(x_i^T b\right)}\right)x_i=0$$

Проверка адекватности

Показатели качества подгонки:

1.1) Псевдо коэффициент детерминации , $R_{ps}^{\ 2} = 1 - \dfrac{n}{n + 2(l - \bar{l})}$

где n– количество наблюдений,

l – логарифмическая функция правдоподобия,

$\bar{l}$ – ограниченная логарифмическая функция правдоподобия, в которой все параметры кроме свободного члена равно нулю.

1.2) Коэффициент Макфаддена $R_{MF}^{\ 2} = 1 - \dfrac{l}{\bar{l}}$

Чем ближе показатели к 1, тем выше качество подгонки модели.

1.3) Гипотеза относительно значимости построенной модели бинарного выбора:

тест отношения правдоподобия Likelihood ratio test (LR), высчитывается в статистике, которые сравниваются с табличным значением $\chi^2(n)$, где n – число степеней свобод, равное числу ограничений в гипотезе. Для LR-теста LR- статистика в случае значимости построенной модели близка к 1.

Модели множественного выбора

Модели множественного выбора работают с зависимой переменной, которая имеет несколько альтернатив, то есть это дискретная переменная.

Модели множественного выбора:

1) с упорядоченными альтернативами;

2) с неупорядоченными альтернативами.

Зависимые переменные: 1) номинальные (качественные);

2) порядковые (то есть упорядоченные альтернативы).

Модели с неупорядоченными альтернативами имеют случайный уровень полезности.

Модели с неупорядоченными альтернативами

Модели с неупорядоченными альтернативами имеют случайный уровень полезности и выбираются альтернативы, приносящие наибольшую полезность. Пусть для i-ого индивида осуществляется выбор между J-альтернативами.

$$y_i = \begin{bmatrix} 1 \\ 2 \\ 3 \\ \dots \\ j \end{bmatrix}$$

Полезность выбора может быть представлена как линейная функция от независимых переменных z и j.

$U_{ij}=\beta^T \bullet z_{ij} \dots + \varepsilon_{ij}$,

где β^T – вектор параметров.

Если i-ый индивид делает выбор j-ой альтернативы, то в этом случае она будет ему максимально полезна.

Пусть y_i – случайная величина, которая описывает сделанный выбор.

То есть, модель описывает вероятность того, что выбор сделан в пользу j-ой альтернативы.

$P(y_i=j)= P(U_{ij} > U_{ik})$ для всех $k \neq j$, $k = 1,\dots,J$, где U_{ij} – наиболее полезная альтернатива, чем все остальные U_{ik}.

$F(U_{ij})$ – функция определения полезности: - логит, или - пробит.

Обычно в качестве объясняющих факторов выбирают характеристики специфические для альтернатив, которые могут изменяться в зависимости от вариантов выборов.

Модели множественного выбора с упорядоченными альтернативами

Определение 3.1.: Модели множественного выбора с упорядоченными альтернативами называются модели, для которых зависимая переменная является порядковой с ранжированными альтернативами (например оценки студента 2, 3, 4, 5).

Модель основана на введении латентной (ненаблюдаемой) переменной y^* порождающие 0, т.е. связанные с переменной y.

Выбор осуществляется между K-альтернативами.

Наша латентная переменная y* имеет вид:

$y^*=x_1b_1+x_2b_2+x_sb_s+b_0$,

где s – число независимых факторов b_j, $j=1,\ldots,s$ – коэффициенты регрессий.

Тогда латентная переменная y^* связана с y, следующим образом:

$$y=\begin{cases}0, y^*<0\\ 1, 0<y^*<\mu_1\\ 2, \mu_1<y^*<\mu_2\\ k, y^*>\mu_{k-1}\end{cases}$$

Пробит-модель

Вероятность выбора k-ой альтернативы, это вероятность того, что:

$\mu_{j-1} < y^* < \mu_j$, где $j=0,1,\ldots,k$.

Вероятность:

$$P(\mu_{j-1} < y^* < \mu_j) = |P(y^* < t) = F(t)| = F(\mu_j - y^*) - F(\mu_{j-1} - y^*) = |\mathbf{\Phi(t) - фун - я\ ЛЛплас}| =$$

$$= \Phi(\mu_j - y^*) - \Phi(\mu_{j-1} - y^*)$$

Тогда модель множественного выбора имеет вид:

Если $y^*=x^T b$, то $x^T=(1,x_1,x_2,\ldots,x_s)^T$, $b=(b_0,b_1,\ldots,b_s)$

$$(3.3)\begin{cases} P(y=0) = \Phi(-x^T b), \text{ где } b = (b_0, b_1, \ldots, b_s) \\ P(y=1) = \Phi(\mu_1 - x^T b) - \Phi(-x^T b) \\ P(y=2) = \Phi(\mu_2 - x^T b) - \Phi(\mu_1 - x^T b) \\ P(y=k) = 1 - \Phi(\mu_{k-1} - x^T b) \end{cases}$$

(3.3) – вероятностная модель множественного выбора с упорядоченными альтернативами, является пробит-моделью с нормальным стандартным распределением.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{x^2}{2}} dx$$

Логит-модель

Т.к. вероятность всегда положительная, P>0, то $0<\mu_1<\mu_2<\ldots<\mu_{k-1}$

$$
\begin{aligned}
&P(y=0)=\Lambda(-x^T b)\\
(3.4)\quad &P(y=1)=\Lambda(\mu_1-x^T b)-\Lambda(-x^T b)\\
&P(y=2)=\Lambda(\mu_2-x^T b)-\Lambda(\mu_1-x^T b)\\
&P(y=k)=1-\Lambda(\mu_{k-1}-x^T b)
\end{aligned}
$$

Где,

$$\Lambda=\frac{e^x}{1+e^x}$$

Оценки моделей (3.3) и (3.4) проводятся методом максимального правдоподобия.

Процедура проверки адекватности такая же как и для бинарных моделей.