Механика Транспорт Комуникации

ISSN 1312-3823 (print)
ISSN 2367-6620 (online)
том 14, брой 3/2, 2016 г.
Научно списание **http://www.mtc-aj.com** **статия № 1370**

ДВУФАКТОРНИ ВЕРОЯТНОСТНИ МОДЕЛИ В ОБРАЗОВАНИЕТО

Райна Алашка
alraina@abv.bg

ВТУ „Тодор Каблешков", София 1574, ул. „Гео Милев"158, БЪЛГАРИЯ

Ключови думи: *Линеен вероятностен модел, Logit-модел, Probit-модел, Метод на най-малките квадрати, Вероятностни модели, Образование.*

Резюме: *Представени са линейният вероятностен модел (ЛВМ), Logit-моделът и Probit-моделът. Направен е двуфакторен анализ за получаване на вероятността за успешно полагане на изпит в зависимост от два фактора. Такива например са посещаемостта на учебните занятия и изпълнението на домашните задания. Представен е векторен вид на двуфакторния логит модел.*

1. ОСНОВНИ ХАРАКТЕРИСТИКИ НА ДВУФАКТОРНИТЕ ВЕРОЯТНОСТНИ МОДЕЛИ

При **линейния вероятностен модел** (ЛВМ) разглежданата зависима променлива Y е дихотомна и има само две значения – 0 и 1, които са индикатор за наличие и отсъствие на някакво явление.

$$(1) \qquad Y_i = \begin{cases} 1, & \text{при наличие на дадено явление,} \\ 0, & \text{при отсъствие на дадено явление.} \end{cases}$$

При двуфакторния ЛВМ за вероятностите да се сбъдне събитието при непретегления метод имаме:

$$(2) \qquad P_i = M(Y_i) = M(Y_i / X_{1,i}, X_{2,i}) = \hat{Y}_i = b_0 + b_1 X_{1,i} + b_2 X_{2,i}.$$

При **логистичния вероятностен (Logit) модел** зависимата променлива (следствието) Y^* е ненаблюдаема, обикновено наречена скрита (латентна) променлива. Това, което се наблюдава, е фиктивна променлива Y_i, чрез която се опитваме да опишем ненаблюдаемата променлива Y^*. За скритата (латентна) променлива Y^* фиктивната променлива Y_i има дихотомно проявление и се дефинира като:

$$(3) \qquad Y_i = \begin{cases} 1, & \text{ако } Y^* > 0; \\ 0, & \text{ако } Y^* \text{ е друг случай.} \end{cases}$$

Проявлението на латентната променлива ще разгледаме чрез оценка на така наречената **шансова пропорция,** която се дефинира като:

$$(4) \qquad \frac{P}{1-P} = \frac{1+e^{(b_0+b_1X_{1,i}+b_2X_{2,i}+\varepsilon)}}{1+e^{-(b_0+b_1X_{1,i}+b_2X_{2,i}+\varepsilon)}} = e^{(b_0+b_1X_{1,i}+b_2X_{2,i}+\varepsilon)}.$$

В горното равенство имаме частното на вероятността за наличие на дадено явление спрямо вероятността за неговото отсъствие, наречена пропорция, измерваща „шанса за реализация на дадено явление". Колкото е по-голям шансът, толкова е по-вероятно да се случи събитието (явлението).

Логаритмуваме с натурален логаритъм двете страни на равенство (4) и получаваме:

(5) $$L = \ln\frac{P}{1-P} = \ln e^{(b_0+b_1X_{1,i}+b_2X_{2,i}+\varepsilon)} = b_0 + b_1X_{1,i} + b_2X_{2,i} + \varepsilon\,.$$

В литературата L се нарича **logit** (логит), затова този модел се нарича logit-модел.

Оценяваме параметрите по метода на най-малките квадрати (МНМК) и намираме теоретичните стойности за вероятностите по формулата:

(6) $$\hat{P}_i = \frac{1}{1+e^{-\hat{L}_i}}$$, защото е в сила равенството $\hat{L}_i = \ln\frac{\hat{P}_i}{1-\hat{P}_i}$.

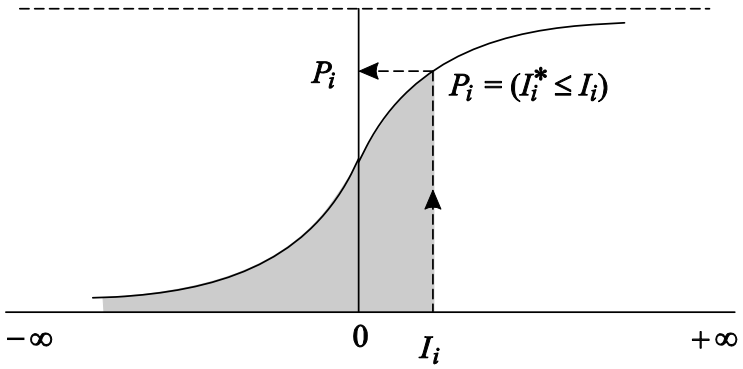
При използването на **пробит модела** се конструира ненаблюдаван индекс за полезност I_i (**нормит**). Ако ни е известна кумулативната вероятност P_i, намираме нормита, като използваме обратната функция на стандартното нормално разпределение:

(7) $$I_i = F^{-1}(P_i) = b_0 + b_1X_{1,i} + b_2X_{2,i}\,.$$

Оценяваме параметрите по МНМК и намираме теоретичните стойности за вероятностите по формулата:

(8) $$P_i = P(Y=1) = P(I_i^* \le I_i) = F(I_i) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{b_0+b_1X_{1,i}+b_2X_{2,i}} e^{-\frac{Z^2}{2}}\,dz\,,$$

където $F(z)$ е функцията на стандартното нормално разпределение, т.е. $I \in N(0;1)$.



Фиг.1

На Фиг. 1 е представена вероятността P_i събитието да се случи. Тази вероятност се измерва с площта под интегралната крива в граници $(-\infty; I_i]$

При разгледаните модели, директното прилагане на МНМК води до неточности, поради различния вид на отклоненията (остатъците) ε_i. За остатъците при МНМК трябва да е в сила **условието за хомоскедастицитет**, което означава да имаме еднакво разпределение на дисперсиите σ_i^2 на остатъците. Тогава е в сила: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_m^2$. За получаването на по-ефективни оценки се използват претеглените модели. При тях изходните променливи се претеглят с тегла $\sqrt{W_i}$ (колона 4 на таблица 1) [1], [2], [3].

Таблица 1

Модел	**Функция за вероятността**	**Дисперсия на остатъците**	**Тегла** $\sqrt{W_i}$
ЛВМ	$P(X_i) = b_0^* + b_1^* X_{1,i} + b_2^* X_{2,i}$	$\sigma_i^2 = P_i(1-P_i)$	$\frac{1}{\sqrt{P_i(1-P_i)}}$
Логит модел	$P(X_i) = \frac{1}{1+e^{-(b_0^* + b_1^* X_{1,i} + b_2^* X_{2,i})}}$	$\sigma_i^2 = \frac{1}{N_i P_i(1-P_i)}$	$\sqrt{N_i P_i(1-P_i)}$
Пробит модел	$P(X_i) = F(b_0^* + b_1^* X_{1,i} + b_2^* X_{2,i}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b_0^* + b_1^* X_{1,i} + b_2^* X_{2,i}} e^{-\frac{Z^2}{2}} dz$	$\sigma_i^2 = \frac{P_i(1-P_i)}{N_i (I_i)^2}$	$\sqrt{\frac{N_i (I_i)^2}{P_i(1-P_i)}}$

2. ПРИЛОЖЕНИЕ НА ВЕРОЯТНОСТНИТЕ МОДЕЛИ ЗА ИЗСЛЕДВАНИЯ И ИЗМЕРВАНИЯ В ОБРАЗОВАНИЕТО

Направено е изследване за резултатите от изпита по Висша математика 2 част на студенти от ВТУ „Тодор Каблешков”. Върху тези резултати са приложени вероятностните модели за пресмятане на вероятността за успешно полагане на изпита, в зависимост от количеството решени примери от зададените домашни работи и посещаемостта на учебните занятия (упражнения, лекции).

Цел на изследването е определяне на зависимостта на вероятността за успешно полагане на изпита от количеството решени примери от зададените домашни работи и посещаемостта на учебните занятия (упражнения, лекции).

Място на изследването е ВТУ „Тодор Каблешков”.

Обект на изследването са 50 студента – редовно и задочно обучение от I курс.

Предмет на изследването са резултатите от изпита по ВМ 2 част.

Данните са получени от справки от преподавателите за изпълнение на домашните задания и присъствени списъци, изпитните работи и изпитните протоколи.

Хипотеза за резултата от изследването: Вероятността за успех на изпита зависи от количеството на решените задачи от домашните задания и посещаемостта на учебните занятия (упражнения, лекции).

Методи на теоретичното изследване: Приложени са 3 модела. Те са линейният вероятностен модел (ЛВМ), Logit-моделът и Probit-моделът.

При разгледаните модели следствието е представено на слабата скала с две разновидности. За представянето на тези две разновидности се използват фиктивни означения $Y_i = \begin{cases} 1, & \text{студентът е издържал изпита;} \\ 0, & \text{студентът не е издържал изпита.} \end{cases}$

Факторните променливи са две. X_1 е факторната променлива – общ брой вярно решени примери от домашните работи. X_2 е факторната променлива – брой посещения на учебните занятия. Разглежданите стойности за X_1 са 4 (0, 100, 200, 300). Разглежданите стойности за X_2 са 3 (10, 20, 30). Тогава възможните различни двойки стойности са $3.4 = 12$. При наблюденията от изследването липсват 4 двойки: (0, 10), (100, 20), (100, 30) и (300, 10).

Обобщените данни от експеримента са дадени в таблица 2.

Таблица 2

Брой вярно решени примери	0	0	100	200	200	200	300	300
Брой посещения на учебните занятия	20	30	10	10	20	30	20	30
Брой изследвани студенти	4	6	7	3	5	15	6	4
Брой студенти издържали изпита	1	3	2	1	3	14	5	3

Резултати от изследването: В таблица 3 са поместени основните характеристики на двуфакторните модели получени при изследването.

Таблица 3

Модел	**Регресионно уравнение и очаквана стойност на вероятността**	**Критична права**
ЛВМ	$Y_i^* = -0,2168 + 0,0019.X_{1,i} + 0,0251.X_{2,i}$ $P_i = Y_i^*$	$P_i = 0,5$ $0,0019.X_{1,i} + 0,0251.X_{2,i} = 0,7168$
Логит модел	$\hat{L}_i^* = -2,7309 + 0,0069.X_{1,i} + 0,0951.X_{2,i}$ $P_i^* = \frac{1}{1+e^{-\hat{L}_i^*}}$	$P_i = 0,5 \Rightarrow L_i = 0$ $0,0069.X_{1,i} + 0,0951.X_{2,i} = 2,7309$
Пробит модел	$\hat{I}_i^* = -1,9893 + 0,0035.X_{1,i} + 0,0923.X_{2,i}$. $P_i^* = F(\hat{I}_i^*)$	$P_i = 0,5 \Rightarrow I_i = 0$ $0,0035.X_{1,i} + 0,0923.X_{2,i} = 1,9893$

Изводи и възможни приложения от изследването.

Първо, като разгледаме ЛВМ виждаме, че коефициентът b_0^* е отрицателен. Това показва, че разгледаните два фактора изчерпват основните фактори, от които зависи успеваемостта на студентите на изпита. Без известен брой посещения на учебните занятия или частично изпълнение на домашните работи (самоподготовка) е невъзможно да се положи успешно изпита. Ако студент, който не е посещавал нито едно учебно занятие и не е изпълнявал нито една домашна работа, си вземе изпита, то това се дължи на преписване, подсказване, предварителни компетентности или на случайни причини.

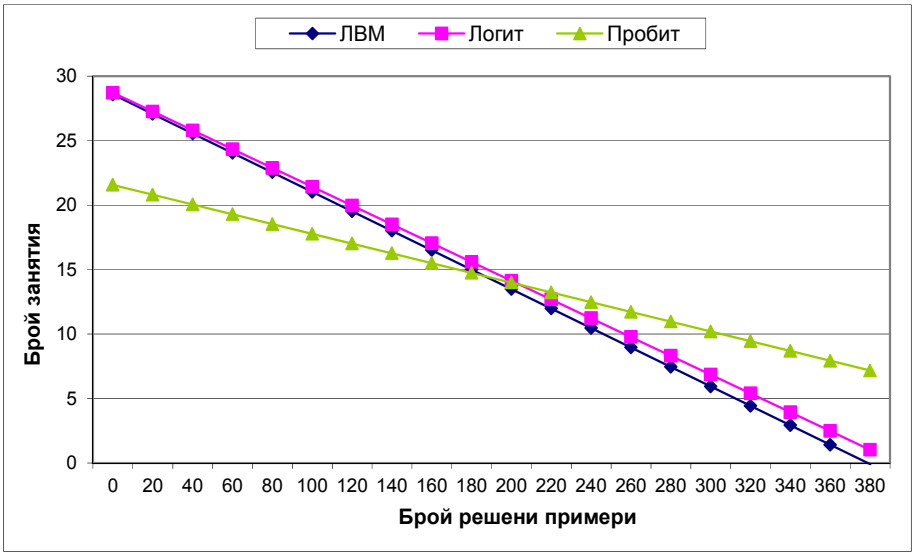
Второ, коефициентите пред факторните променливи в ЛВМ показват, че решаването на 10 примера повече повишава вероятността, изразена в проценти с 1,9 %. Посещението на всяко занятие води до повишаване на вероятността с 2,5%. Можем да направим извода, че посещението на едно занятие е еквивалентно на решаването на 14 примера.

Трето, при трите модела коефициентите пред факторните променливи са положителни. От това и от вида на функциите на вероятността се вижда, че те са растящи функции спрямо всяка от тях.

Четвърто, на Фиг. 2 са представени критичните прави в равнината X_1OX_2 за всеки от моделите. Ако една точка с координати (X_1^0, X_2^0) принадлежи на правата, то вероятността, според съответния модел е равна на 0,5, ако е под правата вероятността е по-малка от 0,5, а ако е над правата вероятността според съответния модел е над 0,5.

Ако обобщим трите модела и имаме точка с координати (X_1^0, X_2^0) от първи квадрант на равнината X_1OX_2, която лежи и под трите прави на моделите, то тогава вероятността студентът да си вземе изпита е по-малка от 0,5.

За точка с координати (X_1^1, X_2^1) от първи квадрант на равнината X_1OX_2, която лежи и над трите прави на моделите, вероятността студентът да си вземе изпита е по-голяма от 0,5.



Фиг.2

3. ВЕКТОРЕН ВИД НА ДВУФАКТОРНИЯ ЛОГИТ МОДЕЛ

Нека L е оценката за логита при двуфакторния модел от вида $L = b_0 + b_1X_1 + b_2X_2$.

Когато той е равен на нула ($L = 0$), вероятността за успех е равна на 0,5 ($P = 0,5$).

Пресмятаме градиента му (нормалния вектор $\vec{n}$ към правата, която се задава с уравнението $0 = b_0 + b_1X_1 + b_2X_2$ в равнината X_1OX_2):

(9) $$grad(L) = grad(b_0 + b_1.X_1 + b_2.X_2) = \frac{\partial L}{\partial X_1}.\vec{i} + \frac{\partial L}{\partial X_2}.\vec{j} = b_1\vec{i} + b_2\vec{j} = \vec{n}(b_1, b_2).$$

Тогава може да запишем логита във векторен вид:

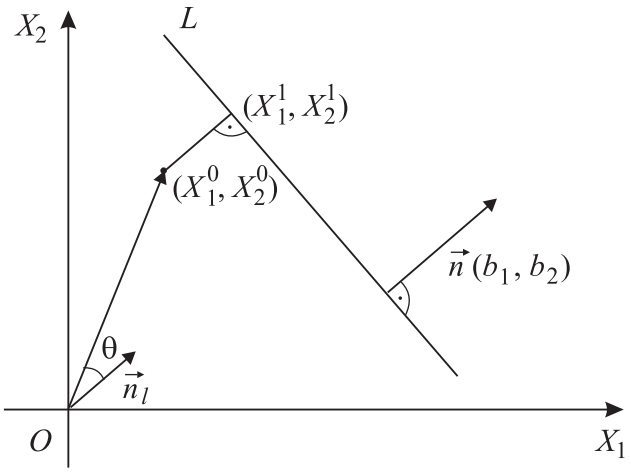
(10) $$L = b_0 + b_1X_1 + b_2X_2 = b_0 + \vec{n}.\overrightarrow{X}.$$

Нека $\overrightarrow{X^0} = (X_1^0, X_2^0)$ ($\overrightarrow{X^0} = X_1^0.\vec{i} + X_2^0.\vec{j}$) е векторът, характерен за студент с неговите стойности на факторните променливи X_1, X_2, за които вероятността да си вземе изпита е по-малка от 0,5. Тогава е в сила: $(grad(L).\overrightarrow{X^0}) = s < -b_0$.

Ако студентът иска да достигне критичната права или да я надмине (да повиши вероятността си за успех до и над 0,5) най-бързо, трябва да се движи по вектора на градиента. Трябва да увеличи показателите си и по двата фактора и да достигне поне до състояние $\overrightarrow{X^1}(X_1^1, X_2^1)$, такова че $(grad(L).\overrightarrow{X^1}) = -b_0$ (Фиг. 3).

За вектора $\overrightarrow{X^1}$ е изпълнено: $\overrightarrow{X^1} = \overrightarrow{X^0} + r.\underbrace{grad(L)}_{\vec{n}}$, където $r = \frac{(-b_0) - s}{|\vec{n}|^2} = \frac{(-b_0) - s}{b_1^2 + b_2^2}$.

Тъй като $s = (grad(L).\overrightarrow{X^0}) = (\vec{n}.\overrightarrow{X^0}) = |\vec{n}|.|\overrightarrow{X^0}|.\cos\theta$, способността зависи от два компонента - от дължината на вектора $\overrightarrow{X^0}(X_1^0, X_2^0)$ и от това, какъв е ъгълът θ между него и вектора $\vec{n} = grad(L)$.



Фиг.3

При различни вектори $\overrightarrow{X_i^0}\left(X_{1i}^0, X_{2i}^0\right)$ с еднаква дължина $\left|\overrightarrow{X^0}\right|$, за способността ще имаме най-голяма стойност при този вектор, който е колинеарен на вектора $\vec{n} = grad(L)$ (ъгълът, заключен между него и $\vec{n}$, е равен на нула).

Логита записваме във вида:

$$L = b_0 + \vec{n}.\overrightarrow{X} = \left|\vec{n}\right|\left(\frac{b_0}{\left|\vec{n}\right|} + \frac{\vec{n}}{\left|\vec{n}\right|}\cdot\overrightarrow{X}\right) = \left|\vec{n}\right|\left(\frac{b_0}{\left|\vec{n}\right|} + \overrightarrow{n_l}.\overrightarrow{X}\right) = \left|\vec{n}\right|\left(\overrightarrow{n_l}.\overrightarrow{X} - \frac{(-b_0)}{\left|\vec{n}\right|}\right),$$

където: трудността е $\frac{(-b_0)}{\left|\vec{n}\right|}$; способността е $\overrightarrow{n_l}.\overrightarrow{X}$; дискриминацията е $\left|\vec{n}\right| = \sqrt{b_1^2 + b_2^2}$.

ЛИТЕРАТУРА

[1] Вындев Д. Л. Записки по приложна статистика 2, СУ, София, 2003.
[2] Съйкова Ив. Д., Стойкова-Къналиева Адр. Ст., Съйкова Св. Ст.. Статистическо изследване на зависимости. Унив. изд. „Стопанство", София, 2002, 453 стр.
[3] Михалев Д., Алашка Р. Теория на вероятностите и статистика, София, 2012 г.

TWO-FACTOR PROBABILISTIC MODELS IN EDUCATION

Rayna Milkova Alashka
alraina@abv.bg

Todor Kableshkov Univercity of Transport,
1574 Sofia, 158 'Geo Milev' Street,
BULGARIA

Key words: *Linear probability model, Logit-model, Probit-model, The method of least squares, Item Response Theory, Education.*

Abstract*: Linear probability model, Logit-model and Probit-model are presented. A two-factor analysisis used to obtain the probability of passing the exam depending on two factors. Such as school attendance and performance of homework. A vector form of the two-factor logit model is presented.*