Logit & probit модели

Чеботарь Полина
Мартьянова Елизавета

Содержание

Примечания: Z – линейная функция переменных, определяющих искомую вероятность, f(Z) – функция плотности распределения

Модели двоичного выбора

Примеры

Часто интересны факторы, определяющие подобные ситуации:

- Почему одни люди поступают в вузы, а другие – нет?
- Почему одни люди меняют место жительства, а другие – нет?
- И т.п. (ответ можно закодировать как «нет» = 0, «да» = 1)

Типы

- Линейная модель
- **Логит-модель**
- **Пробит-модель**
- Тобит-модель

Метод оценки

- Метод максимального правдоподобия
- МНК (только для линейной модели)

Функция вероятности события

$p_i = p(Y_i = 1) = \beta_1 + \beta_2 X_i$ • Линейная модель

$p_i = F(Z_i) = \frac{1}{1 + e^{-Z_i}}$ • **Логит-модель**

$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}$ • **Пробит-модель**

Примечания: Z – линейная функция переменных, определяющих искомую вероятность, f(Z) – функция плотности распределения, Y – зависимая переменная, принимающая значения 1 и 0

Содержание

Примечания: Z – линейная функция переменных, определяющих искомую вероятность, f(Z) – функция плотности распределения

Логит-модель. Области применения

Историческая справка:

В 1950-х зарождалась в работах разных авторов, в нынешнем виде сформулирована в середине 1960х (D.R. Cox *Some procedures associated with the logistic qualitative response curve*).

Используется:

- Медицина (определение вероятности успешного лечения и т.п.)
- Социология
- Маркетинговые исследования (предсказание склонности к покупке)
- Задачи классификации (скоринг в банках, маркетинг и пр.)

Примечания: Z – линейная функция переменных, определяющих искомую вероятность, f(Z) – функция плотности распределения

Логит-модель. Математический смысл

• **Вероятность события определяется функцией**:

$$p_i = F(Z_i) = \frac{1}{1+e^{-Z_i}}.$$, где Z:

$Z_i = \beta_1 + \beta_2 X_i$ - Линейная комбинация независимых факторов

• **Предельное воздействие вел-ны Z на вероятность есть производная функции вероятности**: $f(Z) = \frac{dp}{dZ} = \frac{e^{-Z}}{(1+e^{-Z})^2}.$

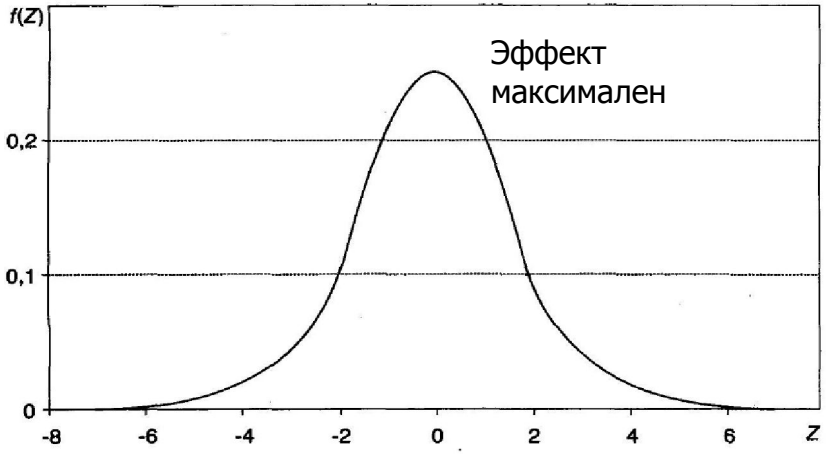
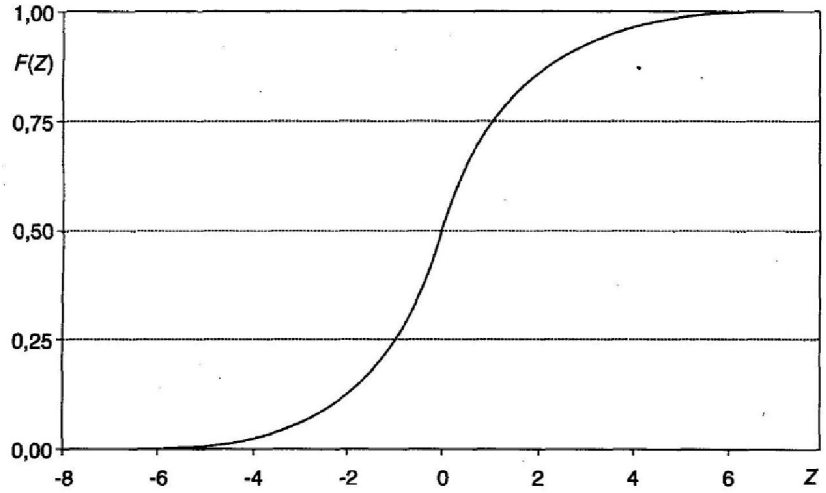


Рис. 10.4. Предельное воздействие *Z* на вероятность



Исправление недостатка линейной модели, в которой вероятность могла получаться больше 1 (что логически неверно):

•Z ☐ бесконечность, e^{-Z} ☐ 0, вероятность ограничена сверху 1

•Z ☐ - бесконечность, e^{-Z}☐ бесконечность, вероятность ограничена снизу 0

Примечания: Z – линейная функция переменных, определяющих искомую вероятность, f(Z) – функция плотности распределения

Логит-модель. Этапы оценки.

1) Определение зависимой переменной и факторов
2) Построение переменной Z, как линейной комбинации независимых переменных
3) Построение уравнения для искомой вероятности события и нахождение производных (для оценки кумулятивного и предельного воздействия факторов)
4) Проведение вычислений с помощью программы (используется метод максимального правдоподобия)
5) Интерпретация результатов
6) Качество оценивания

Примечания: Z – линейная функция переменных, определяющих искомую вероятность, f(Z) – функция плотности распределения

Пример. Окончание средней школы (1)

1)

Переменная	Описание
GRAD	•Зависимая переменная •1- если индивид окончил школу, 0 – в противном случае
ASVABC	•Независимая переменная •Совокупный результат тестирования познавательных способностей
SM	•Независимая переменная •Число лет обучения матери респондента
SF	•Независимая переменная •Число лет обучения отца респондента
MALE	•Независимая переменная, фиктивная переменная •Пол, 1=мужской, 0=женский

2) $$Z = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + \beta_5 MALE.$$

Примечания: Z – линейная функция переменных, определяющих искомую вероятность, f(Z) – функция плотности распределения

Пример. Окончание средней школы (2)

3) $$p_i = F(Z_i) = \frac{1}{1 + e^{-Z_i}}.$$ (Подставляется полученное выражение для Z)

4)

```
. logit GRAD ASVABC SM SF MALE
Logit estimates Number of obs = 540
LR chi2(4) = 43.75
Prob > chi2 = 0.0000
Log likelihood = -96.804844 Pseudo R2 = 0.1843
```

GRAD	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ASVABC	.1329127	.0245718	5.41	0.000	.0847528	.1810726
SM	-.023178	.0868122	-0.27	0.789	-.1933267	.1469708
SF	.0122663	.0718876	0.17	0.865	-.1286307	.1531634
MALE	.1279654	.3989345	0.32	0.748	-.6539318	.9098627
_cons	-3.252373	1.065524	-3.05	0.002	-5.340761	-1.163985

Таблица оцененных коэффициентов. Далее для оценки кумулятивного и предельного эффектов необходимо произвести дальнейшие расчеты, подставив полученные коэффициенты в формулы.

Пример. Окончание средней школы (3)

Переменная	Среднее	*b*	Среднее × *b*	*f*(*Z*)	*bf*(*Z*)
ASVABC	51,36	0,1329	6,8257	0,0281	0,0037
SM	11,58	–0,0231	–0,2687	0,0281	–0,0007
SF	11,84	0,0123	0,1456	0,0281	0,0003
MALE	0,50	0,1280	0,0640	0,0281	0,0036
Постоянный член	1,000	–3,2524	–3,2524		
Итого			3,5143		

$$\frac{dp}{dASVABC} = \frac{dp}{dZ} \times \frac{dZ}{dASVABC} = f(Z)\beta_2$$

Пример нахождения выражения предельного эффекта для одной из переменных

Столбец предельных эффектов

Пример. Окончание средней школы (4)

5)

- Увеличение ASVABC на один балл увеличивает вероятность успешного окончания школы на 0,4 процентных пункта.
- Аналогично, влияет принадлежность к мужскому полу.
- Образование родителей влияет незначительно
- Кроме того, на 10% уровне значимости значим только коэффициент при переменной ASVABC

Пример. Окончание средней школы (4)

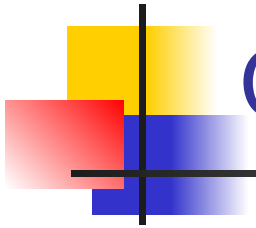
6) **Для метода максимального правдоподобия нет коэффициента, аналогичного R-square, поэтому используются следующие способы**:

-Число правильно предсказанных исходов, если в наблюдении i, считать предсказанием 1 при p(i)>0,5, 0 – в противном случае

-Сумма квадратов отклонений $\sum_{i=1}^{n}(Y_i - p_i)^2$

-Коэффициент корреляции между исходными и предсказанными значениями

Кроме того, значимость отдельных коэффициентов по-прежнему можно оценить с помощью t-статистики (или z-статистики для больших выборок).



Содержание

Примечания: Z – линейная функция переменных, определяющих искомую вероятность, f(Z) – функция плотности распределения

Пробит-модель. Обзор

- 1935 год – Chester Bliss «THE CALCULATION OF THE DOSAGE-MORTALITY CURVE», Annals of Applied Biology
- 1)1934 год - Chester Bliss «The method of probits», Science
  2)1947 - David John Finney «Probit Analysis», Cambridge University Press

- Сферы использования
  - Медицина
  - Социология
  - Маркетинг
  - Любые статистические исследования

Пробит-модель. Математическая составляющая 1(2)

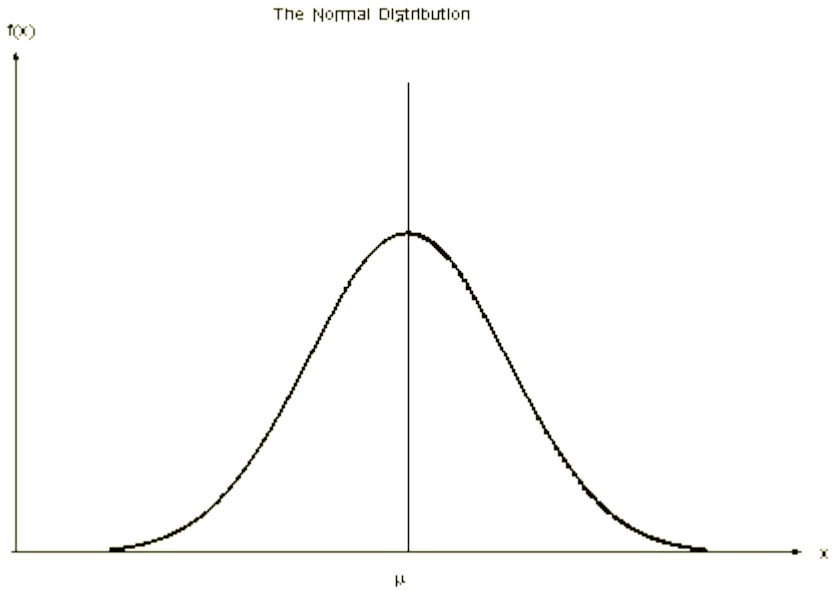
Пробит-модель – альтернативная модель двоичного выбора

Для пробит-анализа используется стандартное нормальное распределение для моделирования зависимости F(Z)

$$p_i = F(Z_i).$$

- функция вероятности зависит от переменной Z, которая в свою очередь зависит от выбранных факторов

$$Z = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k.$$



Пробит-модель. Математическая составляющая 2(2)

Для оценки параметров, как и в логит-модели, используется метод максимального правдоподобия

Предельный эффект переменной Xi - равен производной функции вероятности по этой переменной

$$\frac{\partial p}{\partial X_i} = \frac{dp}{dZ}\frac{\partial Z}{\partial X_i} = f(Z)\beta_i.$$

Так как f(Z) – производная функции (функция плотности) стандартного нормального распределения F(Z), то она выглядит следующим образом

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}.$$

Примечания: Z – линейная функция переменных, определяющих искомую вероятность, f(Z) – функция плотности распределения

Пробит-модель

Расчет общей статистики предельного эффекта:

1.Рассчитать значение Z для средних значений объясняющих переменных

$$Z = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k.$$

1.Рассчитывается f(Z) по формуле

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}.$$

1.Рассчитывается предельный эффект Xi равный f(z)bi

Примечания: Z – линейная функция переменных, определяющих искомую вероятность, f(Z) – функция плотности распределения

Пробит-модель. Применение 1(3)

Переменная	Описание
GRAD	•Зависимая переменная •1- если индивид окончил школу, 0 – в противном случае
ASVABC	•Независимая переменная •Совокупный результат тестирования познавательных способностей
SM	•Независимая переменная •Число лет обучения матери респондента
SF	•Независимая переменная •Число лет обучения отца респондента
MALE	•Независимая переменная, фиктивная переменная •Пол, 1=мужской, 0=женский

$$Z = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + \beta_5 MALE.$$

Примечания: Z – линейная функция переменных, определяющих искомую вероятность, f(Z) – функция плотности распределения

Пробит-модель. Применение 2(3)

. probit GRAD ASVABC SM SF MALE
Probit estimates Number of obs = 540
LR chi2(4) = 44.11
Prob > chi2 = 0.0000
Log likelihood = -96.624926 Pseudo R2 = 0.1858

GRAD	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ASVABC	.0648442	.0120378	5.39	0.000	.0412505	.0884379
SM	-.0081163	.0440399	-0.18	0.854	-.094433	.0782004
SF	.0056041	.0359557	0.16	0.876	-.0648677	.0760759
MALE	.0630588	.1988279	0.32	0.751	-.3266368	.4527544
_cons	-1.450787	.5470608	-2.65	0.008	-2.523006	-.3785673

Примечания: Z – линейная функция переменных, определяющих искомую вероятность, f(Z) – функция плотности распределения

Пробит-модель. Применение 3(3)

Пробит оценивание – зависимая переменная GRAD

Переменная	Среднее	*b*	Среднее × *b*	*f*(*Z*)	*bf*(*Z*)
ASVABC	51,36	0,0648	3,3281	0,0680	0,0044
SM	11,58	–0,0081	–0,0938	0,0680	–0,0006
SF	11,84	0,0056	0,0663	0,0680	0,0004
MALE	0,50	0,0631	0,0316	0,0680	0,0043
Постоянный член	1,00	–1,4508	–1,4508		
Итого			1,8814		

Примечания: Z – линейная функция переменных, определяющих искомую вероятность, f(Z) – функция плотности распределения

Сравнение результатов оценки logit и probit

logit

Переменная	Среднее	*b*	Среднее × *b*	*f*(*Z*)	*bf*(*Z*)
ASVABC	51,36	0,1329	6,8257	0,0281	0,0037
SM	11,58	–0,0231	–0,2687	0,0281	–0,0007
SF	11,84	0,0123	0,1456	0,0281	0,0003
MALE	0,50	0,1280	0,0640	0,0281	0,0036
Постоянный член	1,000	–3,2524	–3,2524		
Итого			3,5143		

probit

Переменная	Среднее	*b*	Среднее × *b*	*f*(*Z*)	*bf*(*Z*)
ASVABC	51,36	0,0648	3,3281	0,0680	0,0044
SM	11,58	–0,0081	–0,0938	0,0680	–0,0006
SF	11,84	0,0056	0,0663	0,0680	0,0004
MALE	0,50	0,0631	0,0316	0,0680	0,0043
Постоянный член	1,00	–1,4508	–1,4508		
Итого			1,8814		

Незначительные изменения

Логит и пробит анализ. Преимущества и недостатки

Плюсы

- Исправление недостатка линейной модели, в которой вероятность могла получаться больше 1 (что логически неверно): вероятность от 0 до 1
- При решении задач классификации объекты можно разделять на несколько групп:
  - Например, в скоринге не только -(0 - плохой, 1 - хороший), но и несколько групп (1, 2, 3, 4 группы риска).

Минусы

- Систематическое завышение оценки коэффициентов регрессии при размере выборки – менее 500
- При построении модели нужно минимально **10** исходов на каждую независимую переменную (рекомендованное значение **30-50**):
  - Например, интересующий исход – смерть пациента. Если 50 пациентов из 100 умирают – максимальное число независимых переменных в модели = 50/10=5

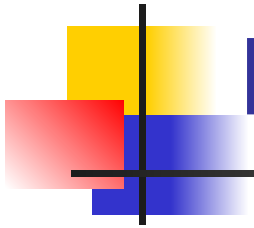
Реальные исследования 1(2)

- 2010 – «**Predicting Foreign Bank Exits? Logit and Probit Regression Approach**», Aneta Hryckiewicz *(Goethe University, Frankfurt),* Oskar Kowalewski (*Warsaw School of Economics)*
- Данные:
  - 81 закрытый филиал в 37 странах
  - период 1999-2006
  - Анализ данных для филиала и домашнего региона, для года закрытия и предшествующего ему года

Переменная	расшифровка
Assets	Log total assets
Agrowth	Annual change of total assets
Equity	Equity to total assets ratio
Loans	Net loans to total assets ratio
Liquidity	Liquid assets to customer and short term funding ratio
LQuality	Loan loss provision to net interest revenue ratio
ROAA	Return on average assets
Costs	Cost to income ratio

Реальные исследования. Результаты 2(2)

- Основная причина закрытия зарубежных отделений – не низкие финансовые показатели филиала, а внутренние проблемы материнского банка: выявлена прямая взаимосвязь между падением показателей материнского банка и ростом вероятности закрытия зарубежного подразделения.
- При этом в год закрытия показатели материнского банка показывали значительный рост
- Результаты логит и пробит анализа отличаются незначительно



Конец

Спасибо за внимание!

Источники

- Nemes S, Jonasson JM, Genell A, Steineck G. 2009 Bias in odds ratios by logistic regression modelling and sample size. BMC Medical Research Methodology
- Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996). "A simulation study of the number of events per variable in logistic regression analysis". *J Clin Epidemiol* 49 (12): 1373–9.
- Agresti A (2007). "Building and applying logistic regression models". *An Introduction to Categorical Data Analysis*. Hoboken, New Jersey: Wiley. p. 138
- Lennox, Clive
  S., Identifying Failing Companies: A Re-evaluation of the Logit, Probit and MDA Appro aches (February 1998)
- Hryckiewicz, Aneta and Kowalewski, Oskar, Predicting Foreign Bank Exits? A Logit and Probit Regression Approach (January 15, 2010)

Примечания: Z – линейная функция переменных, определяющих искомую вероятность, f(Z) – функция плотности распределения