Логарифмически-линейные модели

А.Р.Бессуднов
(bessudnov@gmail.com)

НИУ ВШЭ, факультет социологии

18 мая 2012

Логарифмически-линейные модели

- Применяются, когда все переменные модели являются категориальными (данные могут быть представлены в виде таблицы сопряженности)
- Наиболее полезны, когда нас интересует не одна, а несколько зависимых переменных
- Статистически эквивалентны биномиальной/мультиномиальной логистической регрессии, а также регрессии Пуассона (в зависимости от спецификации)
- В социологии наиболее часто применяются в исследованиях социальной мобильности

Таблица сопряженности: пол и вера в загробную жизнь

	да	нет	всего
мужчины	375	134	509
женщины	435	147	582
всего	810	281	1091

Источник: A.Agresti, An Introduction to Categorical Data Analysis

Структура данных (1)

id	пол	вера
1	m	y
2	f	y
3	f	n
4	m	n
..	..	..
1091	m	n

Структура данных (2)

пол	вера	n
m	y	375
m	n	134
f	y	435
f	n	147

Те же данные как доли

	да	нет	всего
мужчины	0.3437	0.1228	0.4665
женщины	0.3987	0.1347	0.5335
всего	0.7424	0.2576	1

Только маргинальные доли

	да	нет	всего
мужчины			0.4665
женщины			0.5335
всего	0.7424	0.2576	1

При условии независимости пола и веры в загробную жизнь (1)

	да	нет	всего
мужчины	$p_{i+}p_{+j}$	$p_{i+}p_{+j}$	0.4665
женщины	$p_{i+}p_{+j}$	$p_{i+}p_{+j}$	0.5335
всего	0.7424	0.2576	1

При условии независимости пола и веры в загробную жизнь (2)

	да	нет	всего
мужчины	0.4665*0.7424	0.4665*0.2576	0.4665
женщины	0.5335*0.7424	0.5335*0.2576	0.5335
всего	0.7424	0.2576	1

При условии независимости пола и веры в загробную жизнь (3)

	да	нет	всего
мужчины	0.3463	0.1202	0.4665
женщины	0.3961	0.1374	0.5335
всего	0.7424	0.2576	1

Для получения ожидаемых частот при условии независимости необходимо умножить каждую ячейку таблицы на 1091 (общее число наблюдений)

	да	нет	всего
мужчины	377.9	131.1	509
женщины	432.1	149.9	582
всего	810	281	1091

Связаны ли пол и вера в загробную жизнь?

- Критерий хи-квадрат (χ^2) позволяет протестировать гипотезу о независимости переменных в таблице сопряженности
- Он основан на сравнении наблюдаемых частот в ячейках таблицы и частот, ожидаемых при условии независимости. Если разница между наблюдаемыми и ожидаемыми частотами большая, то вероятность того, что она является случайной, мала
- $\chi^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$
- n_{ij} – наблюдаемая частота в ячейке ij, μ_{ij} – ожидаемая частота в ячейке ij при условии независимости

Наблюдаемые и ожидаемые частоты

пол	вера	n_{ij}	μ_{ij}	$\frac{(n_{ij}-\mu_{ij})^2}{\mu_{ij}}$
m	y	375	377.9	0.02
m	n	134	131.1	0.06
f	y	435	432.1	0.02
f	n	147	149.9	0.06

$$\chi^2 = 0.02 + 0.06 + 0.02 + 0.06 = 0.16$$

Проверка гипотезы о независимости

- Зная количество степеней свободы в модели, мы можем проверить гипотезу о независимости переменных
- В двумерной таблице сопряженности количество степеней свободы $df = (I-1)(J-1)$, т.е. в нашем случае 1
- Критическое значение χ^2 при одной степени свободы на 95% уровне значимости равно 3.8
- $0.16 < 3.8$ (p=0.69), таким образом мы не можем отвергнуть гипотезу о том, что наблюдаемые частоты статистически значимо отличаются от ожидаемых при условии независимости (т.е. о том, что две переменные не связаны между собой)

Логлинейная модель независимости

- В модели независимости вероятность попасть в ячейку таблицы сопряженности ij определяется маргинальным распределением по строкам и столбцам: $\pi_{ij} = \pi_{i+}\pi_{+j}$
- В логлинейных моделях используются частоты, а не вероятности: $\mu_{ij} = n\pi_{i+}\pi_{+j}$
- Модель удобнее оценивать как сумму, а не произведение эффектов, соответственно: $\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$
- λ – эффект общей суммы, λ_i^X – эффект строк, λ_j^Y – эффект столбцов
- Эта формула представляет модель независимости для двумерной таблицы сопряженности. Эта модель предполагает, что две переменные не связаны между собой

Наш пример

параметр	коэф.	ст.ош.
λ	4.88	0.07
λ^F	0.13	0.06
λ^{AL}	1.06	0.07

- В этом примере женщины закодированы как 1 (мужчины 0), вера в загробную жизнь как 1 (ее отсутствие 0)
- Таким образом, в модели независимости предсказанный логарифм частоты в ячейке (1;1)= 4.88+0.13+1.06=6.07
- $e^{6.07}$=432.7
- $\log \mu_{1;0} = 4.88 + 0.13 = 5.01.\ e^{5.01} = 149.9$
- $\log \mu_{0;1} = 4.88 + 1.06 = 5.94.\ e^{5.94} = 379.9$
- $\log \mu_{0;0} = 4.88.\ e^{5.01} = 131.6$

Насыщенная (saturated) модель

- Насыщенная модель в двумерной таблице сопряженности предполагает, что между двумя переменными наличествует связь
- Эта модель полностью описывает наблюдаемые данные
- $\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$

Параметры насыщенной модели

параметр	коэф.	ст.ош.
λ	4.898	0.09
λ^F	0.0925	0.12
λ^{AL}	1.029	0.1
λ^{F*AL}	0.0558	0.14

- $\log \mu_{1;1} = 4.898 + 0.0925 + 1.029 + 0.0558 = 6.0753.\ e^{6.0753} = 435$
- $\log \mu_{1;0} = 4.898 + 0.0925 = 4.9905.\ e^{4.9905} = 147$
- $\log \mu_{0;1} = 4.898 + 1.029 = 5.927.\ e^{5.927} = 375$
- $\log \mu_{0;0} = 4.898.\ e^{4.898} = 134$

Какая модель лучше: модель независимости или насыщенная?

- Иными словами, связаны ли между собой две переменные?
- Формально мы можем сравнить модели, используя показатели качества модели: хи-квадрат Пирсона (χ^2) и отклонение (deviance, G^2)

модель	χ^2	G^2	df	p-value
F, A	0.1620	0.1621	1	0.687
FA	0	0	0	

Анализ многомерных таблиц сопряженности

- В двумерных таблицах сопряженности логлинейный анализ фактически мало отличается от анализа с помощью критерия хи-квадрат (а также логистической и мультиномиальной логистической регрессии)
- В многомерных таблицах сопряженности логлинейный анализ дает более интересные результаты

Пример: анализ связи между употреблением алкоголя, сигарет и марихуаны

- Данные: опрос учеников выпускного класса американской средней школы около Дейтона, Огайо

алкоголь (A)	сигареты (C)	марихуана (M) - да	марихуана (M) - нет
да	да	911	538
да	нет	44	456
нет	да	3	43
нет	нет	2	279

Источник: A.Agresti, An Introduction to Categorical Data Analysis

Возможные модели

- (A,C,M) – модель независимости
- (M, AC) – употребление алкоголя и сигарет связаны между собой, употребление марихуаны ни с чем не связано
- (C, AM); (A, CM)
- (AC, AM) – употребление алкоголя связано с употреблением сигарет и марихуаны; если контролировать употребление алкоголя, связь между употреблением сигарет и марихуаны отсутствует
- (AC, CM); (AM, CM)
- (AC, AM, MC) – существует парная связь между всеми тремя переменными. Отношения шансов между любыми двумя переменными одинаковы на всех уровнях третьей переменной. Модель гомогенности. Иными словами, эта модель предполагает отсутствие взаимодействия между переменными
- (ACM) – насыщенная модель, полностью описывает данные

Предсказанные значения

- Для каждой модели мы можем рассчитать предсказанные значения частот в ячейках таблицы сопряженности

A	C	M	(A,C,M)	(M, AC)	(AM, CM)	(AC, AM, MC)	(ACM)
да	да	да	540	611.2	909.24	910.4	911
да	да	нет	740.2	837.8	438.84	538.6	538
да	нет	да	282.1	210.9	45.76	44.6	44
да	нет	нет	386.7	289.1	555.16	455.4	456
нет	да	да	90.6	19.4	4.76	3.6	3
нет	да	нет	124.2	26.6	142.16	42.4	43
нет	нет	да	47.3	118.5	0.24	1.4	2
нет	нет	нет	64.9	162.5	179.84	279.6	279

Сравнение моделей

модель	G^2	χ^2	df	p-value (по G^2)
(A,C,M)	1286	1411.4	4	< 0.001
(M, AC)	843.8	704.9	3	< 0.001
(AM, CM)	187.8	177.6	2	< 0.001
(AC, AM, MC)	0.4	0.4	1	0.54
(ACM)	0	0	0	

- Модель (AC, AM, MC) статистически не значимо отличается от насыщенной

Параметры модели (AC, AM, MC)

параметр	коэф.	ст.ош.
λ	5.63	0.06
λ_a	0.49	0.08
λ_c	-1.89	0.16
λ_m	-5.31	0.48
λ_{ac}	2.05	0.17
λ_{am}	2.99	0.46
λ_{mc}	2.85	0.16

- Во всех парах переменных наличествует значимая положительная связь. Т.е. если ученик курит, то повышается вероятность того, что он употребляет алкоголь; если употребляет алкоголь, то повышается вероятность употребления марихуаны; если курит, то повышается вероятность употребления марихуаны. Однако сила связи между курением и употреблением алкоголя не зависит от того, употреблял ли ученик марихуану (т.е. для употреблявших и не употреблявших марихуану сила связи между курением и употреблением алкоголя примерно одинакова)

Заключение

- Логлинейные модели входят в класс обобщенных линейных моделей, generalized linear models (т.е. по сути являются одним из вариантов регрессии, регрессией Пуассона)
- В современной исследовательской практике в социологии используются редко, т.к. в большинстве случаев могут быть заменены (мультиномиальной) логистической регрессией, позволяющей также включать количественные (интервальные) предикторы
- Логлинейные модели традиционно используются в исследованиях социальной мобильности (переменные: социальный класс респондента, социальный класс отца, страна, период)
- В случае, если переменные являются порядковыми, используются более сложные логмультипликативные модели