Многомерный логлинейный анализ разреженных таблиц сопряженности: особенности и ограничения в социологической практике

Ротмистров А.Н.

Щетинин А.Ю.

Аннотация

Логлинейный анализ в настоящее время становятся все более популярными, в том числе и в социальных науках. Это объясняется несколькими факторами, но прежде всего уникальной возможностью работать со счетными (count) данными и анализировать связи в многомерных таблицах сопряженности. Традиционный метод логлинейного анализа, который реализован на предпосылке о том, что данные подчиняются распределению Пуассона, в настоящее время достаточно сильно критикуется в связи с работой с данными с высокой дисперсией (overdispersed data). Тем не менее, из внимания упускается другая, намного более серьезная проблема, которая ограничивает применимость логлинейного анализа на разреженных данных: несуществующая оценка максимума правдоподобия. В данной статье мы попытаемся описать данную проблему, причины ее возникновения, опишем и систематизируем методы, которые могут снизить вероятность или исключить риск возникновения несуществующих оценок и предложим алгоритм работы с логлинейным анализом на социологических данных. Работа посвящена традиционному логлинейному анализу, однако описанная в ней проблема характерна и для ряда родственных методов, в том числе для отрицательной биномиальной модели (“57127 - Nonconvergence in log-linked Poisson and negative binomial models,” n.d.; J. M. C. Santos Silva & Tenreyro, 2010)*, которая рекомендуется как альтернатива логлинейному анализу на данных с высокой дисперсией* (Ver Hoef & Boveng, 2007).

Введение

Что такое логлинейный анализ и зачем он нужен

Логлинейный анализ - семейство методов, предназначенных для работы с многомерными таблицами сопряженности. На более простом уровне, логлинейный анализ может использоваться как более продвинутая альтернатива методу Хи-2, применимая на многомерных таблицах сопряженности (то есть способная работать более, чем с двумя

переменными). На более сложном - логлинейный анализ является полноценной обобщенной линейной моделью, которая с помощью итеративных механизмов максимизации правдоподобия может моделировать наблюдаемые частоты и выявлять признаки, наиболее влияющие на распределение частот в таблице сопряженности.

История развития метода

Логлинейный анализ был разработан в середине 20 века Лео Гудманом (Goodman, 1963). Изначально, задачей логлинейного анализа было расширение функционала метода Хи-2 (Fienberg & Rinaldo, 2007). Логлинейный анализ, как и метод Хи-2 работает с таблицами сопряженности и может выявлять наличие связи между переменными, но в отличие от метода Хи-2 не ограничивается только связью между двумя переменными, но может работать для таблиц сопряженности с более чем двумя измерениями, причем с учетом эффектов взаимодействия переменных (Upton, 1978). Содержательно, в классической версии логлинейного анализа строится так называемая насыщенная модель, которая объясняет 100% дисперсии в таблице сопряженности. Например, для трехмерной таблицы сопряженности насыщенная модель будет включать в себя 3 переменных (строки, столбцы, “глубина»), 3 двухмерных эффекта взаимодействия и один трехмерный эффект взаимодействия. В такой модели предсказанные переменные равны наблюдаемым. Константа в модели вычисляется как среднее по логарифму частот в каждой из ячеек таблицы.

Затем из модели можно исключать какой-либо из эффектов (например, если мы хотим проверить гипотезу о связи между двумя переменными, то необходимо исключить из модели эффект их взаимодействия). При исключении одного или нескольких эффектов предсказанные значения начинают отличаться от наблюдаемых и с помощью критерия Хи-2 можно провести оценку значимости эффекта. Кроме того, по коэффициентам в модели можно сделать предположения о силе и направленности связи между переменными.

Логлинейный анализ в настоящее время

В дальнейшем, логлинейный анализ был значительно доработан. Если предположить, что частоты в таблице сопряженности подчиняются распределению

Пуассона (что часто характерно для счетных данных), то при трансформации наблюдаемой переменной (в данном случае - наблюдаемых частот в таблице сопряженности) с помощью логарифмирования (log-transform) мы получаем линейную модель, которая может быть оптимизирована с помощью метода максимизации правдоподобия (Agresti, 2007; Haber, 1985).

В таком случае, функционал метода значительно расширяется: мы можем не только строить насыщенную модель и оценивать значимость эффектов, но и находить значения коэффициентов, при которых с наибольшей вероятностью наблюдались бы частоты в таблице. То есть логлинейный анализ становится обобщенной линейной моделью (Agresti, 2007), в которой переменная - частоты в таблице сопряженности моделируется на основании профилей независимых переменных. Стоит отметить, что в такой модели за константу в модели обычно берется логарифм частоты последнего профиля (есть исключения, например, функция loglin в R), что оказывает влияние на интерпретацию и является важным фактором при работе с моделью. Достаточно близкую модель можно получить, если использовать метод наименьших квадратов, а в качестве зависимой переменной использовать логарифм наблюдаемой частоты, но такой подход критикуется как не вполне валидный и менее точный (O'Hara & Kotze, 2010).

Ключевое преимущество метода заключается в том, что он позволяет работать с номинальными переменными (качественными характеристиками индивидов), в том числе с эффектами взаимодействия качественных переменных. На основании этих многомерных связей можно конструировать более содержательно наполненные предикторы, которые будут иметь более высокую предсказательную способность в других моделях. В некоторых случаях, полезной оказывается возможность моделировать частоты. Логлиненый анализ реализован практически во всех значимых современных пакетах: реализации существуют в SPSS, SAS, Stata, R.

В данной работе будет использоваться понятие разреженных данных. Хочется отметить, что мы не включаем в данное понятие «данные, насыщенные нулями» (zero inflated data). Такие данные часто встречаются в других отраслях (например, в лингвистике и естественных науках) и требуют особой методики работы с ними и специально разработанные модели. Условно проведем границу: если нули наблюдаются

хотя бы в двух[1] менее, чем в 50% ячеек таблицы сопряженности, то такие данные мы будем называть разреженными; если нули наблюдаются более, чем в 50% ячеек, то такие данные мы будем считать «данными, насыщенными нулями» и такие данные не будут предметом данной статьи.

Постановка проблемы

Несмотря на то, что у логлинейного анализа имеется большой список преимуществ, особенно для работы с социологическими данными, многими авторами отмечаются и недостатки данного метода. Большой пласт иностранной литературы посвящен проблеме высокой дисперсии (overdispersion) в логлинейных моделях(Dean, 1992; Ver Hoef & Boveng, 2007). По причине того, что в традиционной логлинейной модели критерий максимизации правдоподобия рассчитывается, основываясь на предпосылке о том, что наблюдаемые частоты подчиняются распределению Пуассона, модель проводит корректную оценку коэффициентов только в том случае, если наблюдаемые средние в ячейках равны дисперсии. Эта проблема хорошо известна как теоретикам, так и практикам в области статистики, предложены техники и рекомендации для преодоления данного ограничения. В частности, на таких данных рекомендуется использовать метод негативной биномиальной регрессии, который мы опишем далее(Ver Hoef & Boveng, 2007).

Нам в данной работе хотелось бы уделить внимание другой, намного менее известной проблеме, связанной с логлинейными моделями, а именно несуществующим оценкам критерия максимизации правдоподобия. В литературе отмечается, что несмотря на то, что данная проблема выявлена еще 40 лет назад, она практически неизвестна практикам и может приводить к ошибочным интерпретациям(Fienberg & Rinaldo, 2012).

Суть проблемы заключается в том, что при определенных паттернах наблюдаемых частот в ячейках итеративные механизмы максимизации правдоподобия не могут найти экстремум функции (по той причине, что его не существует) и потому проводят некорректную оценку коэффициентов(Fienberg & Rinaldo, 2012). Эта проблема родственна проблеме несуществующих оценок в логистической регрессии(Silvapulle, 1981), которая

[1] Это связано со спецификой описываемой проблемы. Она может возникнуть если нули наблюдаются хотя бы в двух ячейках в таблице сопряженности.

является более изученной, но тоже достаточно актуальной.

Проблема усугубляется тем, что существующие в современных статистических пакетах алгоритмы не имеют инструмента для выявления несуществующих оценок(Fienberg & Rinaldo, 2007). Фактически, программа строит модель и не выводит никаких ошибок, но содержательно модель оказывается неправильной. Вероятность возникновения проблемы возрастает с количеством нулевых ячеек в таблице сопряженности, именно поэтому наиболее релевантна эта проблема для разреженных данных. На стандартной выборке в 2000 человек, в многомерной таблице сопряженности, построенной всего на 6 переменных с 3 значениями, средняя частота в ячейке будет равняться примерно 2,7. Очевидно, что данные распределяются не гомогенно и в такой таблице сопряженности будет очень много нулевых ячеек. Усугубляет ситуацию существование структурных нулей: ситуации, когда данное сочетание наблюдаемых признаков невозможно. Таким образом, в практической работе логлинейный анализ практически всегда проводится на таблицах сопряженности с хотя бы одной наблюдаемой нулевой частотой.

В публикации 1974 года автор утверждает, что для решения поставленной проблемы достаточно того, чтобы в данных не было нулевых частот (Haberman, 1974). Следуя этому утверждению, большая часть современных алгоритмов автоматически добавляет маленькую константу к каждой ячейке в таблице сопряженности. Тем не менее, в ряде публикаций утверждается, что рекомендация не совсем верна и маленькие константы не решают проблему (Fienberg & Rinaldo, 2007; J. M. C. Santos Silva & Tenreyro, 2010).

В публикации 2010 года (J. M. C. Santos Silva & Tenreyro, 2010) авторы утверждают, что все ситуации несуществующей оценки сводятся к ситуациям коллинеарности предикторов, которые приводят к бесконечной оценке коэффициентов. Показано, что несуществующая оценка может наблюдаться даже в тех случаях, когда в таблице сопряженности нет ни одной нулевой частоты(J. M. C. Santos Silva & Tenreyro, 2010).

Вышеописанная проблема на настоящий момент не решается автоматическими алгоритмами и может приводить к неправильным результатам, а в итоге и к неправильным содержательным выводам, сделанным на основании анализа данных. Эта

проблема слабо известна практикующим специалистам (Fienberg & Rinaldo, 2007) и не предложены стандартные рекомендации по тому, как избежать вышеописанной проблемы. Именно разработка таких рекомендаций - задача данной статьи. Для начала, рассмотрим все известные стратегии, которые либо уже описаны в данных, либо применяются для работы с разреженными данными в других областях, но могут быть перенесены на логлинейный анализ.

Перемещение профилей

Авторами (Fienberg & Rinaldo, 2012) показано, что ситуация несуществующей оценки максимизации правдоподобия наблюдается всегда, когда в таблице сопряженности наблюдаются нули в первом и последнем профиле. Кроме того, отмечается, что с высокой вероятностью проблема будет возникать и в том случае, когда нулевые частоты наблюдаются в противолежащих ячейках. Исходя из данных зависимостей, можно разработать рекомендации по предварительной работе с данными. Так, если мы наблюдаем один из описанных выше паттернов, мы можем переместить ячейки в таблице сопряженности таким образом, чтобы избавиться от этих паттернов.

- Данная методика не искажает данные и, соответственно, результаты анализа и не требует никаких дополнительных теоретических предпосылок
- Методика не всегда может быть использована, если контрольный профиль в модели выбран содержательно по той причине, что в большей части реализаций он должен быть последним
- Выявлять противолежащие частоты затруднительно в таблицах сопряженности с более, чем 2 измерениями. Методика требует большой внимательности и высокой квалификации исследователя
- В некоторых случаях, с помощью перемещения ячеек невозможно избавиться от вышеописанных паттернов расположения нулей. Кроме того, методика не исключает, а только снижает вероятность возникновения проблемы

Обобщая, можно рекомендовать применять перемещение ячеек всегда, когда нулевым является последний профиль (нулевой последний профиль не только повышает риск возникновения проблемы, но и значительно искажает интерпретацию). Оптимально в

качестве последнего профиля выбирать профиль с частотой, близкой к средней геометрической. Кроме того, с помощью перемещения ячеек можно обнаружить проблему несуществующей оценки (если при перемещении профилей содержательно меняется интерпретация модели), так что данная методика рекомендуется к применению.

Слияние профилей / исключение части данных

Вторым способом борьбы с проблемой является слияние профилей или исключение части категорий из анализа. С помощью снижения размерности таблицы сопряженности и удаления пустых ячеек можно значительно снизить риск возникновения несуществующей оценки максимума правдоподобия.

- Главный недостаток данной методики заключается в том, что при ее использовании теряется часть информации.
- Данная методика требует от исследователя теоретических допущений, что обычно менее критично для ранговых шкал, но намного более сложно реализуемо для номинальных шкал, для которых в основном и используется логлинейный анализ. По этой же причине она менее применима для разведывательного анализа данных.
- Даже минимальные трансформации в данных могут значительно исказить коэффициенты в модели и изменить интерпретацию.
- Данная методика не всегда может исключить структурные нулевые частоты

Резюмируя, данная методика применима только в тех случаях, когда исследователь может принимать основанные на теории решения относительно данных, то есть, например, проводит анализ на основании обильного теоретического материала. Данная методика практически не может быть применена на этапе разведывательного анализа данных.

Достаточно близкой альтернативой будет также расчет нескольких логлинейных моделей отдельно. Такой подход позволит значительно снизить размерность таблицы сопряженности, но исключает значительное число эффектов взаимодействия между переменными (поиск которых в значительной степени и является целью логлинейного анализа)

Добавление константы

Практически все существующие пакеты автоматически добавляют константу во все ячейки модели. Ранее считалось, что этого шага достаточно, чтобы решить проблему, но недавние публикации (Fienberg & Rinaldo, 2007; J. M. C. Santos Silva & Tenreyro, 2010) показывают, что это не так. Более того, на данный момент не установлено насколько большие константы исключают риск возникновения проблемы, связано ли это каким-либо образом с тем, какова доля нулевых профилей в модели и как эти профили расположены. Кроме того, очевидно, что добавление константы может также вызывать смещение в коэффициентах, особенно если наблюдаемые частоты в ней малы. Например, возьмем значение константы, которое стандартно используется в пакете SPSS:

Таблица 1: модельная таблица сопряженности с 3 измерениями

1	3
2	8

4	3
2	1

Таблица 2: сравнение коэффициентов модели с и без добавления константы.

Переменная	Модель без константы		Модель с константной 0,5	
	Коэффициент	Значимость	Коэффициент	Значимость
константа	-0.18	0.83	0.25	0.73
x	0.96	0.33	0.74	0.38
y	1.34	0.17	1.05	0.20
z	1.61	0.09*	1.29	0.11
x*y	-0.05	0.95	0.06	0.94
x*z	-1.78	0.07*	-1.46	0.09*
y*z	-1.72	0.08*	-1.41	0.11

Как можно увидеть, на приведенной модельной таблице сопряженности с маленькими частотами добавление константы не изменило интерпретацию только 3 коэффициентов. В двух случаях коэффициенты перестали быть значимыми (с не меньшей вероятностью значимость может и повышаться, как это наблюдается в случае константы и эффекта взаимодействия переменных x и y, то есть не значимый коэффициент оказался бы

значимым). В двух случаях знак при коэффициенте изменился на обратный. При увеличении количества категорий и количества переменных влияние константы не снижается, а может даже увеличиваться, особенно если в модели много ячеек с маленькими частотами и несколько ячеек с большими.

Обобщая, мы не можем рекомендовать использовать константы по той причине, что они, с одной стороны, не исключают риск возникновения несуществующих оценок, а с другой стороны вносят значительные искажения в интерпретацию модели, особенно в случае с разреженными данными.

Методика Сильвы и Тенрейро

Авторы, анализируя проблемы алгоритма логлинейного анализа в пакете STATA, обращают внимание в том числе и на проблему несуществующей оценки максимума правдоподобия. Сильва и Тенрейро утверждают, что проблема с несуществующей оценкой для логлинейного анализа не ограничивается ситуацией с нулевыми частотами, но может возникнуть в тех случаях, когда существует полная коллинеарность одной или нескольких зависимых переменных с ненулевыми значениями зависимой переменной(J. Santos Silva & Tenreyro, 2011). Очевидно, что такая ситуация более вероятна если в таблице большое количество ячеек с нулевыми частотами, так что, в принципе, описанные проблемы близки(J. M. C. Santos Silva & Tenreyro, 2010).

Для решения вышеописанной проблемы авторы рекомендуют исключить все независимые переменные, которые являются коллинеарными с ненулевыми значениями зависимой переменной, а затем включать их одну за одной и контролировать изменения в модели. Они отмечают, что этот процесс не может быть автоматизирован по той причине, что при включении предикторов изменяются коэффициенты в модели и только исследователь может сделать содержательный вывод относительно того, закономерно ли такое изменение или наблюдается ситуация несуществующей оценки. Это же является и главным ограничением данного метода: у исследователя должны быть представления о характере взаимосвязей в модели, а выявление несуществующей оценки опирается не на объективные характеристики, а на интуицию аналитика.

Кратко об итеративных методах

Большая часть современных реализаций логлинейного анализа основаны на методе Ньютона-Рапсона (включая IWLS) (Brown & Fuchs, 1983). Помимо него, также возможна реализация алгоритма поиска максимума правдоподобия с помощью метода Iterative Proportional Fitting (Brown & Fuchs, 1983). Отмечается, что второе семейство методов, хоть и является менее быстрым (Brown & Fuchs, 1983), с большей вероятностью не сходится при несуществующей оценке правдоподобия и вообще менее подвержено различным вычислительным проблемам (Fienberg & Rinaldo, 2007). Таким образом, мы можем рекомендовать использовать реализации логлинейного анализа именно на этом алгоритме (например, метод loglin из статистического пакета R).

Рекомендации по работе с логлинейным анализом

Резюмируя сказанное ранее, можно говорить о том, что у проблемы несуществующей оценки максимума правдоподобия в логлинейных моделях в настоящий момент нет оптимального решения. Обозначенная проблема в значительной степени ограничивает применение логлинейного анализа для проведения разведочного анализа данных и использования в качестве инструмента описательной статистики. Для этих задач мы можем рекомендовать традиционный логлинейный анализ, предложенный Лео Гудманом. К сожалению, на данный момент нами не было найдено ни одной полностью функциональной версии данного метода в популярных статистических пакетах[2].

Синтезируя вышеописанные рекомендации, при работе с логлинейным анализом мы рекомендуем:

- Не использовать модели с очень большим количеством переменных по той причине, что при увеличении количества переменных очень быстро растет количество нулевых ячеек в многомерной таблице сопряженности. Для стандартной выборки в 2000 человек мы не рекомендуем использование более, чем 3-4 переменных
- Контролировать значения в профилях, особенно в первом и последнем профиле, а также наполненность категорий используемых переменных. Для алгоритмов, в котором константа рассчитывается по последнему профилю, последним профилем

[2] Единственное исключение - пакет loglin в R, в котором, однако, автоматически не рассчитывается значимость предикторов, что требует от исследователя дополнительных вычислений

рекомендуется задавать тот, частота которого наиболее близка к среднему геометрическому в модели. Это приближает модель к традиционной модели Гудмана, расширяет возможности интерпретации и снижает риск возникновения несуществующей оценки

- С осторожностью использовать добавление константы в ячейки, помня о том, что данная процедура не исключает риск возникновения несуществующей оценки и может вносить значительные искажения в результаты и содержательную интерпретацию
- Каждую модель пересчитывать несколько раз, исключая и включая переменные и эффекты взаимодействия. В особенности обращать внимание на те предикторы, которые коллинеарны ненулевым значениям зависимой переменной (частотам) а также те, которые рассчитываются по слабонаполненным ячейкам
- Контролировать модель на сходимость. По возможности, отдавать предпочтение реализациям логлинейного анализа, ищущим экстремум функции правдоподобия с помощью алгоритма IPF, в частности, функции loglin из статистического пакета R

Библиография

57127 - Nonconvergence in log-linked Poisson and negative binomial models. (n.d.). Retrieved March 13, 2017, from http://support.sas.com/kb/57/127.html

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed). Hoboken, NJ: Wiley-Interscience.

Brown, M. B., & Fuchs, C. (1983). On maximum likelihood estimation in sparse contingency tables. *Computational Statistics & Data Analysis*, *1*, 3–15.

Dean, C. B. (1992). Testing for Overdispersion in Poisson and Binomial Regression Models. *Journal of the American Statistical Association*, *87*(418), 451–457. https://doi.org/10.1080/01621459.1992.10475225

Fienberg, S. E., & Rinaldo, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *Journal of Statistical Planning and Inference*, *137*(11), 3430–3445.

Fienberg, S. E., & Rinaldo, A. (2012). Maximum likelihood estimation in log-linear models. *The Annals of Statistics*, *40*(2), 996–1023. https://doi.org/10.1214/12-AOS986

Goodman, L. A. (1963). On Methods for Comparing Contingency Tables. *Journal of the Royal Statistical Society. Series A (General)*, *126*(1), 94. https://doi.org/10.2307/2982447

Haber, M. (1985). Maximum likelihood methods for linear and log-linear models in categorical data. *Computational Statistics & Data Analysis*, *3*, 1–10.

Haberman, S. J. (1974). *The analysis of frequency data*. Chicago: University of Chicago Press.

O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, *1*(2), 118–122. https://doi.org/10.1111/j.2041-210X.2010.00021.x

Santos Silva, J. M. C., & Tenreyro, S. (2010). On the existence of the maximum likelihood

estimates in Poisson regression. *Economics Letters*, *107*(2), 310–312. https://doi.org/10.1016/j.econlet.2010.02.020

Santos Silva, J., & Tenreyro, S. (2011). poisson: Some convergence issues. Retrieved from http://repository.essex.ac.uk/3534/

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 310–313.

Upton, G. J. G. (1978). *The analysis of cross-tabulated data*. Chichester [Eng.] ; New York: Wiley.

Ver Hoef, J. M., & Boveng, P. L. (2007). QUASI-POISSON VS. NEGATIVE BINOMIAL REGRESSION: HOW SHOULD WE MODEL OVERDISPERSED COUNT DATA? *Ecology*, *88*(11), 2766–2772.