Анализ качественных переменных

Логлинейный анализ таблиц сопряженности

1. Понятие логлинейной модели

2. Логлинейный метод подбора модели

Понятие логлинейной модели

Логлинейная модель – множественная регрессионная модель, в которой категориальные переменные и их взаимодействия выступают в качестве предикторов, а роль зависимой переменной играет натуральный логарифм частот категорий. Использование логарифмической меры обусловливает линейность модели.

В этом уравнении частота – это частота текущей ячейки частотной таблицы, *λ*- воздействие со стороны одной или более независимых переменных, *μ*- общее среднее воздействие, *A*, *C*, *Y* – переменные агрессия, симпатия, условия:

$$\ln(\textit{частота}) = \mu + \lambda A + \lambda C + \lambda \mathrm{Y} + \lambda \mathrm{A} \cdot \mathrm{C} + \lambda \mathrm{A} \cdot \mathrm{Y} + \lambda \mathrm{C} \cdot \mathrm{Y} + \lambda \mathrm{A} \cdot \mathrm{C} \cdot \mathrm{Y}$$

Модель называется насыщенной, если она содержит все предикторы и их возможные взаимодействия.

Существуют более предпочтительные альтернативы в виде ненасыщенных моделей, которые отражают лишь статистически значимые главные эффекты и взаимодействия переменных.

Подменю Логлинейный анализ содержит три команды.

1. Общий — эта команда допускает вхождение в модель любых факторов и их взаимодействий и предполагает, что исследователь перед проведением анализа уже имеет гипотезы о составе модели.
2. Логит — применение этой команды позволяет рассматривать дихотомические переменные как зависимые, а одну (или более) категориальную переменную как независимую. При этом зависимая дихотомическая переменная используется не для прогнозирования частот категорий, а для разделения всех категорий на две группы. ^
3. Подбор модели — эта команда позволяет из всех возможных ненасыщенных моделей подобрать ту, которая в наибольшей степени соответствует исходным данным. Подбор осуществляется, как правило, автоматически. В результате выявляется совокупность значимых связей между категориальными переменными и вычисляются параметры μ и λ логлинейной модели.

Логлинейный метод подбора модели

Теоретически из насыщенной модели можно удалить любые элементы, получив произвольную ненасыщенную модель.
Далее можно проверить состоятельность этой модели и в случае несоответствия ее исходным данным перейти к анализу другой ненасыщенной модели.
Предпочтение отдается иерархическим логлинейным моделям, которые позволяют упорядочить процесс подбора окончательной состоятельной модели.
Основной особенностью иерархических моделей является то, что присутствие какого-либо взаимодействия переменных означает присутствие всех взаимодействий, имеющих более низкий порядок, и главных эффектов этих переменных. Например, если в модели присутствует взаимодействие агрессия х симпатия, то в ней присутствуют главные эффекты переменных агрессия и симпатия;
если в модели присутствует взаимодействие
агрессия х симпатия х условия,
то в ней также присутствуют взаимодействия агрессия х симпатия,
агрессия х условия и симпатия х условия, и т. д.

Существуют три вспомогательных метода, которые предназначены для подбора адекватной модели. Все три метода оказываются полезными и приводят к сходным результатам

Метод *исследования оценок параметров* предназначен для вычисления оценок параметров для насыщенной модели. SPSS вычисляет также стандартизованные оценки. Если значения последних невелики, то они не оказывают значимого влияния на модель и обычно исключаются.

Метод *вычисления частичного критерия хи-квадрат* в дополнение к оценкам параметров модели SPSS вычисляет критерий *хи-квадрат,* характеризующий степень соответствия модели исходным данным. При помощи этого критерия проверяется, являются ли все однофакторные эффекты, а также эффекты более высоких порядков статистически значимыми. При этом отсутствие общей значимости эффектов второго порядка вовсе не означает, что все эффекты первого порядка не являются значимыми. Аналогично, из отсутствия общей значимости эффектов любого порядка не следует отсутствие значимости отдельных взаимодействий этого порядка. Вследствие этих двух особенностей в SPSS предусмотрена возможность раздельной проверки главных эффектов и эффектов взаимодействий.

Суть метода *пошагового исключения* состоит в автоматической «подгонке» модели и сходна с методом исключения предикторов из уравнения регрессии: из насыщенной модели постепенно исключаются те элементы (переменные и их взаимодействия), которые не оказывают значимого воздействия. Данный метод построения модели относится к иерархическому логлинейному моделированию. Если обнаружено статистически значимое взаимодействие четырех переменных, не проверяется (на предмет исключения из модели) взаимодействие трех из этих переменных, иначе модель не являлась бы иерархической по определению. Окончательный результат «подгонки» модели наиболее приемлем, если все оставшиеся в ней элементы оказываются статистически достоверными.

Линейные регрессионные модели с фиктивными переменными

Фиктивные переменные в регрессионных моделях. Необходимость использования фиктивных переменных

В регрессионных моделях в качестве объясняющих переменных часто приходится использовать не только количественные (определяемые численно), но и качественные переменные. Например, спрос на некоторое благо может определяться ценой данного блага, ценой на заменители данного блага, ценой дополняющих благ, доходом потребителей и т.д. (эти показатели определяются количественно). Но спрос может также зависеть от вкусов потребителей, их ожиданий, национальных и религиозных особенностей и т.д. А эти показатели представить в численном виде нельзя. Возникает проблема отражения в модели влияния таких переменных на исследуемую величину.

Линейные регрессионные модели с фиктивными переменными

Обычно в моделях влияние качественного фактора выражается в виде фиктивной (искусственной) переменной, которая отражает два противоположных состояния качественного фактора.

Например, «фактор действует» — «фактор не действует», «курс валюты фиксированный» — «курс валюты плавающий», «сезон летний» — «сезон зимний» и т.д. В этом случае фиктивная переменная может выражаться в двоичной форме:

D =0 - фактор не действует,

D =1 - фактор действует.

Например, *D* = 0, если потребитель не имеет высшего образования, *D* = 1, если потребитель имеет высшее образование;

D = 0, если в обществе имеются инфляционные ожидания, *D* = 1, если инфляционных ожиданий нет.

Переменная *D* называется *фиктивной (искусственной, двоичной) переменной (индикатором).*

Линейные регрессионные модели с фиктивными переменными

Регрессионные модели, содержащие лишь качественные объясняющие переменные, называются **ANOVA-моделями** (моделями дисперсионного анализа).

Например, пусть У — начальная заработная плата.

D=0, если претендент не имеет высшего образования,

D=1, если претендент имеет высшего образование.

Тогда зависимость можно выразить моделью парной регрессии:

$Y=\beta + \gamma D + \varepsilon$

Очевидно,

$M(Y \mid D = 0) = \beta + \gamma\, 0 = \beta,$

$M(У \mid D = 1) = \beta + \gamma\, 1 = \beta + \gamma.$

При этом коэффициент β определяет среднюю начальную заработную плату при отсутствии высшего образования.

Нетрудно заметить, что ANOVA-модели представляют собой кусочно-постоянные функции. Однако такие модели в экономике крайне редки. Гораздо чаще встречаются модели, содержащие как качественные, так и количественные переменные.

Линейные регрессионные модели с фиктивными переменными

Модели ANCOVA – модели, в которых объясняющие переменные носят как количественный, так и качественный характер, называют *ANCOVA-моделями (моделями ковариационного анализа).*

ANCOVA модель при наличии у фиктивной переменной двух альтернатив

Вначале рассмотрим ANCOVA-модель с одной количественной и одной качественной переменной, имеющей два альтернативных состояния:

Пусть, например, У — заработная плата сотрудника фирмы, *X* — стаж сотрудника, *D* — пол сотрудника, т.е. D=0, если сотрудник - женщина, D=1, если сотрудник - мужчина.

Тогда ожидаемое значение заработной платы сотрудников при *x* годах трудового стажа будет:

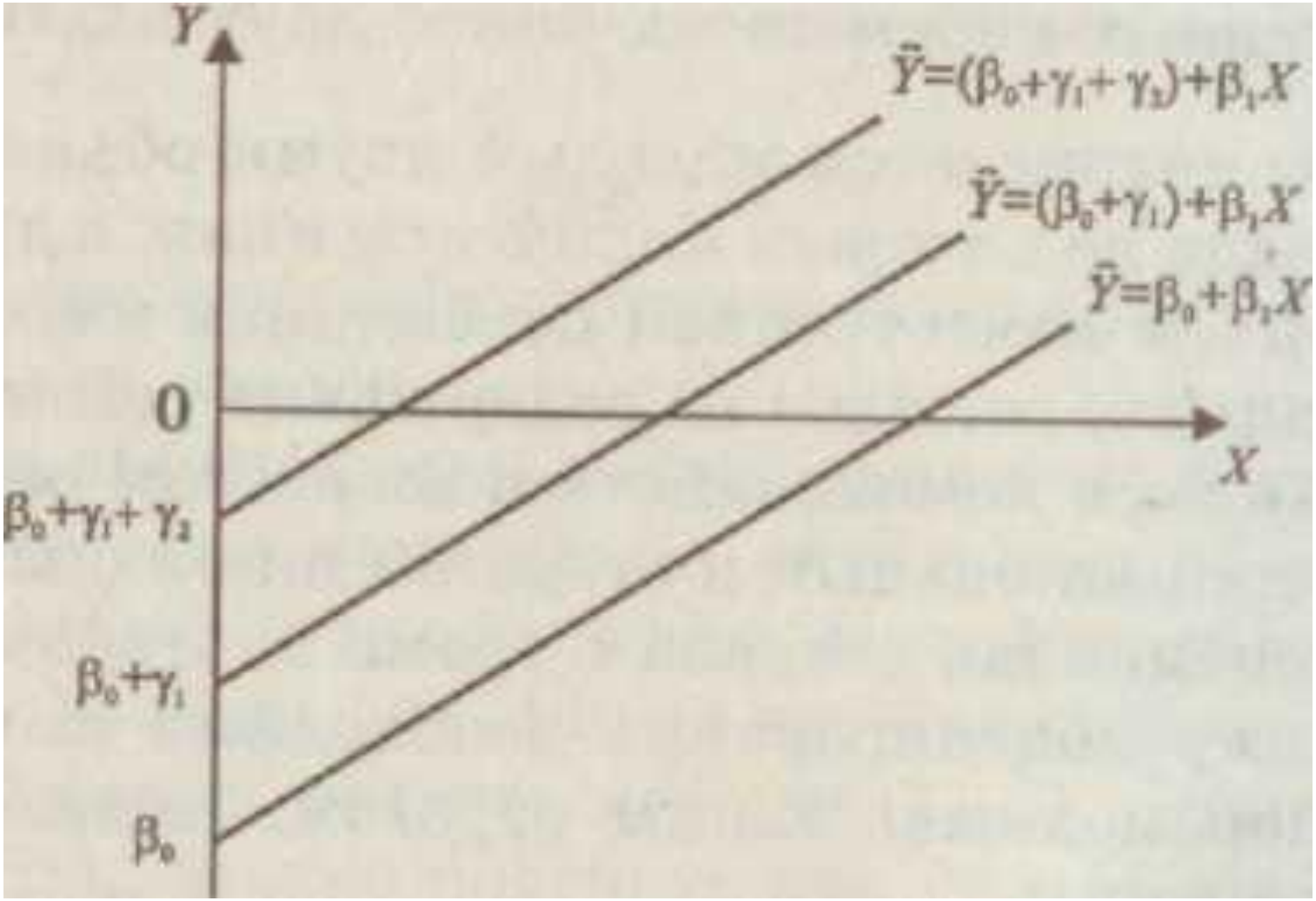
$M(Y \mid x, D = 0) = \beta 0 + \beta 1\, x$ для женщины,

$M(У \mid x, D = 1) = \beta 0 + \beta 1\, x + \gamma = (\beta 0 + \gamma) + \beta 1\, x$ для мужчины.

Если качественная переменная имеет *k* альтернативных значений, то при моделировании используются только (k-1) фиктивных переменных.

Коэффициент γ в модели – дифференциальный коэффициент свободного члена.

Линейные регрессионные модели с фиктивными переменными (качественная переменная имеет 3 альтернативы)



Использование фиктивных переменных в сезонном анализе

Легко видеть, что в модели рассматриваются такие ситуации, при которых квартальные различия отражаются лишь в различии свободных членов моделей. Если же различия затрагивают и изменения коэффициента пропорциональности, то этот факт может быть отражен в следующей модели (*):

$$Y_t = \beta_0 + \beta_1 X_t + \gamma_1 D_{1t} + \gamma_2 D_{2t} + \gamma_3 D_{3t} +$$
$$+ \gamma_4 D_{1t} X_t + \gamma_5 D_{2t} X_t + \gamma_6 D_{3t} X_t + \varepsilon_t$$

Стратегия выбора модели:

1. Рассмотреть модель (*);
2. Определить статистическую значимость коэффициентов.
3. Если дифференциальные угловые коэффициенты статистически незначимы, то перейти к модели без мультипликативных слагаемых. Если в этой модели дифференциальные свободные члены статистически незначимы, то сделать вывод, что квартальные (сезонные) изменения несущественны для рассматриваемой зависимости.

Логистическая регрессия (Logit-, Probit-, Tobit-анализ) **Модель LPM**

Фиктивная зависимая переменная

Фиктивные переменные могут быть использованы для объяснения поведения зависимой переменной. Например, если исследовать зависимость наличия автомобиля от дохода, пола субъекта и т.п., то зависимая переменная имеет как бы два возможных значения: 0, если машины нет, и 1, если машина есть.

Если для моделей данного типа использовать обыкновенный МНК, то оценки, получаемые с его помощью, не обладают свойствами наилучших линейных несмещенных оценок (BLUE). Поэтому для определения коэффициентов в этом случае используются другие методы.

Модель LPM

Рассмотрим модели, в которых зависимая переменная выражается в виде фиктивной (двоичной) переменной. Объясняющие переменные могут быть как количественными, так и качественными.

Например, анализируется наличие работы у субъекта в зависимости от возраста, образования, семейного положения, доходов остальных членов семьи и т.д. В этом случае зависимая переменная У имеет два возможных состояния:

У =0, субъект не имеет работы,

У =1, субъект имеет работу.

Или, например, при исследовании торгового баланса в качестве зависимой может быть использована следующая переменная:

У =0, если торговый баланс отрицательный,

У =1, если торговый баланс не отрицательный.

Модель LPM

$$Y = \beta_0 + \beta_1 X_1 + \ldots \beta_m X_m + \gamma_1 D_1 + \ldots + \gamma_k D_k + \varepsilon$$

Такие модели называются *линейными вероятностными моделями* (linear probability models) *(LPM-моделями).*

Модель LPM

Применимость МНК к моделям LPM имеет определенные ограничения:

1. *Случайные отклонения в данных моделях не являются нормальными случайными величинами, а скорее всего имеют биноминальное распределение.*

Невыполнимость предпосылки МНК о нормальном распределении случайных отклонений не столь существенна при определении оценок уравнения регрессии (они остаются несмещенными), но она достаточно важна при анализе проверок соответствующих гипотез. Однако с ростом объема выборки биноминальное распределение стремится к нормальному распределению.

2. *Случайные отклонения не обладают свойством постоянства дисперсии (гомоскедастичности).*

Следовательно, *дисперсия случайной ошибки в i-том наблюдении* зависит от вероятностей соответствующих значений У, которые в свою очередь зависят от выбранных значений Х. Это означает, что дисперсии отклонений могут быть различными для различных наблюдений.

Данная проблема гетероскедастичности также преодолима

3. *Использование формул может привести к ситуации, когда некоторые значения* У *будут либо меньше нуля, либо больше единицы.*

4. *Применение модели LPM весьма проблематично с содержательной точки зрения.*

Действительно, увеличение значения переменной *X* на одну единицу приводит к изменению значения *Y* на величину β1 вне зависимости от конкретного значения *X,* что противоречит теоретическим и практическим выкладкам (например, закону убывающей эффективности и т.п.).

Все вышеперечисленное позволяет сделать вывод о том, что непосредственное использование МНК в модели LPM приводит к серьезным погрешностям и необоснованным выводам. Поэтому в данном случае его использование не рекомендуется.

Логистическая регрессия

Logit модель

Для преодоления недостатков LPM-моделей необходимо использовать такие модели, в которых не будут, по крайней мере, нарушаться неравенства $0 \le P(Y = 1| x) \le 1$, и зависимость между $P(Y = 1 | x)$ и x не будет иметь линейный характер, а будет удовлетворять закону убывающей эффективности.

Для оценки параметров таких моделей применяются методы логистической регрессии, Logit-, Probit-, Tobit-анализа.

Например, логистическая регрессия используется, когда зависимая переменная — дихотомия, т. е. может принимать только два значения, например 0 и 1. При этом независимые переменные могут быть непрерывными или категориальными переменными.

Пусть зависимая переменная принимает значение 1 при появлении некоторого события А и 0, если событие А не появилось. При каждом наблюдаемом фиксированном наборе факторов вычисляется

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_i$$

Логистическая регрессия

Логистическая регрессия имеет много аналогий с обычной МНК-регрессией, хотя для оценки коэффициентов регрессии используется метод максимального правдоподобия, а не метод наименьших квадратов. В отличие от МНК-регрессии логистическая регрессия оценивает нелинейную связь между независимыми переменными и зависимой. При этом не возникает проблем гетероскедастичности, а требования менее строгие. Успех логистической регрессии может быть оценен по таблице числа правильных и неправильных классификаций дихотомической, зависимой переменной. Для проверки адекватности можно использовать критерии согласия, например критерий *хи-квадрат,* а проверку значимости коэффициентов можно проводить обычным способом.

Логистическая регрессия

Пример (файл helpLR.sav). Рассмотрим мнение партнера о том, полезна или нет оказанная ему помощь.

С логистической регрессией связаны такие математические понятия, как вероятность, шанс и натуральный логарифм шанса.

Вероятность — это ожидаемая относительная частота некоторого события.

Шанс представляет собой отношение вероятности того, что событие произойдет, к вероятности того, что событие не произойдет.

Шанс, в отличие от вероятности, не ограничен максимальным единичным значением. Единичное значение шанса соответствует ситуации, когда вероятности появления и не появления события равны.

Ключевым параметром логистической регрессии является *логит.* Логит равен натуральному логарифму шанса. Например, логит вероятности в 20 % равен -1,386...

Уравнение регрессионной модели, которую рассмотрели раньше, имеет следующий вид:

помощь = *B0+ B1 x* симпатия + *B2* x агрессия + B3 x польза

Согласно этому уравнению величина оказываемой помощи равна сумме константы (BO) и значений трех переменных (отражающих симпатию, агрессивность и пользу), умноженных на соответствующие коэффициенты регрессии. Несмотря на то что в логистическом анализе оценивается полезность или бесполезность помощи, а не ее величина, уравнение логистической регрессии похоже на уравнение множественной регрессии