**Глава 14**

**Модели логистической регрессии**

В модели линейной регрессии  существует два типа переменных – объясняющие переменные  и исследуемая переменная . Эти переменные могут быть измерены по непрерывной шкале, а также как индикаторная переменная. Когда объясняющие переменные являются качественными, тогда их значения expressed как индикаторные переменные, и затем используются модели фиктивных переменных.

При исследовании переменная в качественной переменной, то ее значения могут быть выражены с помощью индикатора переменная принимает только два возможных значения 0 и 1. В таком случае используется логистическая регрессия. Например,  can обозначает такие значения, как успех или неудача, да или нет, нравится или не нравится, которые могут быть обозначены двумя значениями 0 и 1.

Рассмотрим модель



где .

Переменная исследования принимает два значения как  или 1. Предположим, что  следует распределению Бернулли с a parameter  so its probability distribution is



Предполагая 



Из модели  мы имеем



Таким образом, функция  - это просто вероятность того, что 

Обратите внимание , что  поэтому

* когда 
* 

Напомним, что ранее  was assumed to follow a normal distribution when  не было индикаторной переменной.

Когда  является индикаторной переменной, то  принимает только два значения, поэтому нельзя предполагать, что она соответствует нормальному распределению.

В обычной регрессионной модели ошибки гомоскедастичны, т.е.  и так . Когда  является индикаторной переменной, тогда



Таким образом depends on  and is a function mean of . В Более того, поскольку  и  is the вероятностью, поэтому  и, следовательно, существует ограничение on   что накладываетbig constraint on the choice of the linear функции отклика. Невозможно соответствовать модели in, в которой прогнозируемые значения лежат за пределами интервала 0 и 1.

Когда  является дихотомической переменной, то эмпирические данные свидетельствуют о том, что функция  на всей реальной прямой, которую можно отобразить на  имеет сигмовидную форму. Это нелинейная форма , подобная

|  |  |
| --- | --- |
|  |  |

Естественным выбором для  была бы кумулятивная функция распределения случайной величины. В частности, логистическое распределение, чья кумулятивная распределения является упрощенной logistic function yields a good link and is given by



**Линейный предиктор и функции связи:**

Систематический компонент в *E*(*y*) является линейным предиктором и обозначается как



Функция связи in обобщенной линейной модели связывает линейный предиктор  со средним откликом .

Таким образом



В обычных линейных моделях, основанных на нормально распределенной исследуемой переменной, the link  is usedссылкаand is called an **identity link**,,. Функция связи отображает диапазон  на всю реальную , обеспечивает хорошее эмпирическое приближение и несет значимые интерпретации в реальных приложениях.

В случае логистической регрессии функция связи определяется как



Это преобразование называется **логитным** преобразованием вероятности  и  называется **коэффициентами**. Ссылка  также называется **логарифмическими коэффициентами**. Функция этой ссылки получается следующим образом:



или 

или 

или .

**Примечание:** Аналогично функции logit, также другие функции, которые имеют форму, аналогичную форме логистической функции. Эти функции также могут быть преобразованы с помощью . Существуют две такие популярные функции – пробитное преобразование и дополнительное преобразование log-log. Пробит-преобразование основано на преобразовании  с использованием кумулятивной функции распределения нормального распределения и на этом основана **пробит-регрессионная модель.**

The Является .

**Оценка максимального правдоподобия параметров:**

Рассмотрим общую форму модели логистической регрессии



где  - независимая случайная величина Бернулли с a параметром  с



Функция плотности вероятности  является



Функция правдоподобия является





С Тех пор , как



итак

.

Предположим что на каждом уровне переменных доступны повторные наблюдения. Пусть  - числа 1, наблюдаемых for  observation and  be the количество испытаний при каждом наблюдении. Затем

.

Максимального правдоподобия оценку  среди  получается путем численного максимизации.

Если  тогда асимптотически



После получения  линейный предиктор оценивается с помощью



Установленное значение равно



**Интерпретация параметров:**

Чтобы понять интерпретацию связанного  в модели логистической регрессии, сначала рассмотрим простой случай с только одной переменной в качестве



После подгонки модели,  получаются оценки  соответственно. Тогда подогнанный линейный предиктор в  равен



который является логарифмическим коэффициентом при  подобранном значении при  , равно



с какой логарифмическойвероятностью 

Таким образом

****

Это называется **нечетным отношением,** которое представляет собой предполагаемое увеличение вероятности успеха при изменении значения объясняющей переменной на единицу.

Вслучае, когда в модели имеется более одной объясняющей переменной, тогда интерпретация  аналогична, как в случае одной explanatory variable case. Отношение шансов is exp associated with explanatory variable  keeping otherдругой объясняющей переменной постоянной. Это является аналогичное толкование  ян множественная линейная регрессионная модель.

Если объясняющей переменной является изменение в  единице измерения, тоотношение estimated increase iпредполагаемого увеличения in равноexp, ,,

**Проверка гипотезы:**

Испытание гипотезы для параметров ян в модели логистической регрессии на основе асимптотической теории. Это тест с большой выборкой, основанный на тесте отношения правдоподобия, основанном на статистике, называемой **отклонением.**

Модель с точно *p* параметрами, которые идеально соответствуют выборочным данным, называется **насыщенной моделью.**

Статистика, которая сравнивает логарифмические вероятности of fitted and saturated models , называется **model deviance**. Это определяется как



где  is the log-likelihood and  is the maximum likelihood estimate of  - логарифмическое правдоподобие.

В случае the logistic regression model,  ’s are полностью неограниченными. Поэтому вероятность будет максимальной В  , а максимальное значение (насыщенный модальные) - это



Пусть  оценка максимального правдоподобия , тогда логарифмическое правдоподобие максимально при , и



Предполагая, что функция логистической регрессии верна, распределениестатистикиstatistic  приблизительно распределяется как , когда  оно велико.

Небольшоезначение , что the model is incorrect. Значение торгового центра означает, что the model is хорошо подогнана и не уступает модели saturated. Обратите внимание, что, как правило, приталенная модель будет иметь очень небольшие количества параметров-йв насыщенном модели, учитывающей все параметры. Таким образом, на  уровне значимости.