Содержание

ЧАСТЬ I

Логика логистической регрессии

Многие социальные явления являются по своей природе бинарными, а не непрерывными или количественными – произошло событие или оно не произошло, человек купил товар или не купил.

Бинарные дискретные явления обычно принимают форму дихотомического индикатора. Хотя эти два значения можно представлять любыми числами, использование зависимых переменных со значениями 1 и 0 имеет свои преимущества. Среднее значение такой переменной равно доле случаев со значением 1 и может интерпретироваться как вероятность.

1. Проблемы обычной регрессии с бинарной зависимой переменной

На первый взгляд, бинарная зависимая переменная со значениями 0 и 1 кажется подходящей для использования в множественной регрессии. Регрессионные коэффициенты имеют полезную интерпретацию – они показывают увеличение или уменьшение прогнозируемой вероятности возникновения события в силу изменения того или иного предиктора на единицу своего измерения.

Сама зависимая переменная принимает только значения 0 и 1, но предсказанные значения для регрессии принимают вид усредненных пропорций или вероятностей, зависящих от значений предикторов. Чем выше прогнозируемое значение или условное среднее значение, тем больше вероятность того, что с человеком, обладающим конкретными значениями характеристик (независимых переменных), произойдет интересующее событие. Линейная регрессия предполагает, что условные пропорции или вероятности задают прямую линию для значений *X*.

Например, в ходе опроса мы спросили респондентов, курят ли они. Присваиваем тем, кто курит, значение 1, а тем, кто не курит, – значение 0 и получаем бинарную зависимую переменную.

Если взять курение (*S*) как функцию от количества лет, потраченных на образование (*E*), и дамми-переменную для пола (*G*), где женщины получают код 1, то получится уравнение регрессии:

$$S = 0{,}661 - 0{,}029 * E + 0{,}004 * G.$$

Регрессионный коэффициент для образования показывает, что при увеличении количества лет, потраченных на образование, на 1 год вероятность курения снижается на 0,029. Респонденты мужского пола без образования имеют спрогнозированную вероятность курения 0,661 (константа). У мужчины с 10-летним образованием спрогнозированная вероятность курения составляет 0,371 (0,661 – 0,029 * 10 + 0,004 * 0). Коэффициент при дамми-переменной показывает, что у женщин вероятность курения на 0,004 выше, чем у мужчин. Для женщин без образования спрогнозированная вероятность курения составляет 0,665 (0,661 – 0,029 * 0 + 0,004 * 1).

Несмотря на несложную интерпретацию коэффициентов для множественной регрессии с бинарной зависимой переменной, регрессионные оценки сталкиваются с двумя проблемами. Первая проблема носит концептуальный характер, а вторая – статистический. В совокупности проблемы оказываются достаточно серьезными, чтобы потребовать альтернативу обычной регрессии с бинарной зависимой переменной.

1.1. Проблема, связанная с формой функциональной зависимости

Концептуальная проблема линейной регрессии с бинарной зависимой переменной связана с тем, что вероятности имеют максимальные и минимальные значения 1 и 0. По определению, вероятности и пропорции не могут превышать 1 или падать ниже 0. Однако линия линейной регрессии может простираться вверх к $+\infty$, поскольку значения предикторов могут увеличиваться бесконечно, и простираться вниз к $-\infty$, поскольку значения предикторов могут уменьшаться бесконечно. В зависимости от коэффициента наклона линии и наблюдаемых значений *X* модель может дать прогнозные значения зависимой переменной выше 1 и ниже 0. Такие значения не имеют смысла и малопригодны для прогнозирования.

Несколько графиков могут проиллюстрировать проблему. Обычная диаграмма рассеяния, представляющая собой зависимость между двумя непрерывными переменными, показывает облако точек, как на рисунке слева. Здесь линия, проходящая через центр облака точек, минимизирует сумму квадратов отклонений. Мы видим, когда *X* принимает более высокие или более низкие значения, то же самое происходит и с *Y*.

Однако диаграмма рассеяния, представляющая собой зависимость между непрерывной независимой переменной и бинарной зависимой переменной,

уже не представляет из себя облако точек. Вместо этого показаны два параллельных набора точек. Подгонка с помощью прямой линии кажется здесь менее уместной. Любая линия (за исключением линии с коэффициентом наклона 0) в конечном итоге превысит 1 и опустится ниже 0.

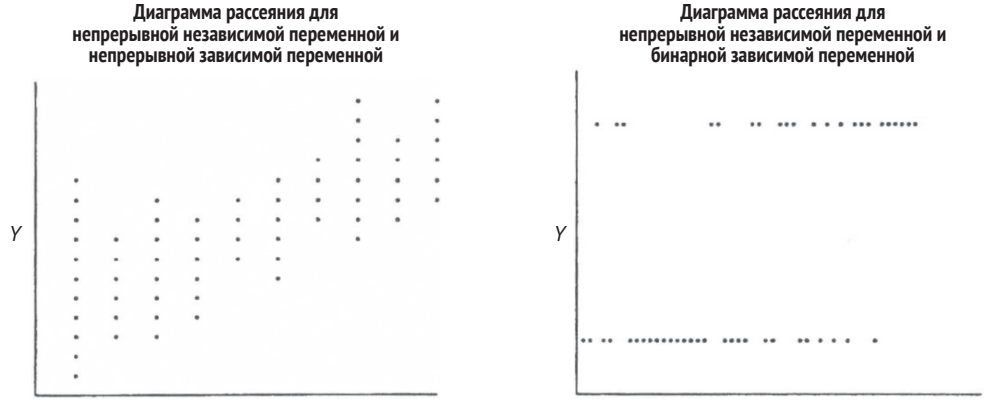


Рис. 1 ❖ Диаграммы рассеяния

Некоторые области двух параллельных наборов точек могут содержать больше наблюдений, чем остальные, и с помощью джиттеринга мы можем взглянуть на плотность наблюдений вдоль двух линий. Прибегнем к джиттерингу – технике визуализации данных, которая используется для усиления разброса между точками при их наложении друг на друга. Он уменьшает перекрытие точек на диаграмме рассеяния, добавляя случайную вариацию к каждому наблюдению. На рис. 2 подвергнутое джиттерингу распределение бинарной зависимой переменной (курит или не курит) по количеству лет, потраченных на образование, указывает на небольшую зависимость. Респонденты с бóльшим количеством лет, потраченным на образование, курят реже, чем респонденты с меньшим количеством лет, потраченных на образование. Однако рисунок отличается от графика зависимости между непрерывными переменными.

Риск получения спрогнозированных вероятностей ниже 0 или выше 1 зависит не только от диапазона значений предиктора, но еще и от соотношения значений 0 и 1 бинарной зависимой переменной. При соотношении 50:50 спрогнозированные значения попадают в центр распределения вероятностей. В предыдущем примере с курением (где соотношение 28:72) самое низкое спрогнозированное значение 0,081 мы получаем для мужчин с максимальным количеством лет, потраченным на образование, т. е. проучившиеся 20 лет (0,661 – 0,029 ∗ 20 + 0,004 ∗ 0), а самое большое спрогнозированное значение 0,665 мы получаем для женщин с минимальным образованием, т. е. проучившиеся 0 лет (0,661 – 0,029 ∗ 0 + 0,004 ∗ 0). Теперь возьмем зависимую переменную с еще большей диспропорцией. Мы спрашиваем респондентов, стремятся они сохранить/защитить окружающую среду или нет. 10 % ответили «да», мы их кодируем единицами, а остальных кодируем нулями.

Регрессия по полу и образованию дает:

$B = -0{,}024 + 0{,}008 * E - 0{,}006 * G.$

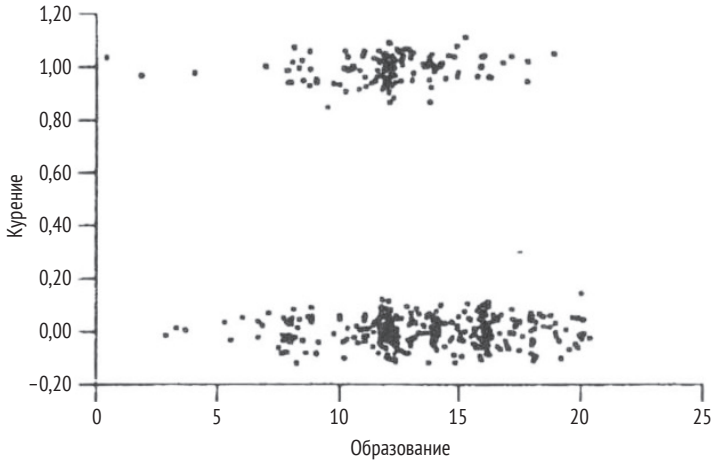


Рис. 2 ❖ Диаграмма рассеяния (с джиттерингом) для бинарной зависимой переменной (зависимость факта наличия/отсутствия курения от количества лет, потраченных на образование)

Константа показывает здесь бессмысленную вероятность – у мужчины без образования спрогнозированная вероятность принадлежности к группе «защитников природы» будет ниже нуля. Предположение о линейности, хоть и является общей проблемой, в этой конкретной модели оказывается особенно неуместным.

Одно из решений данной задачи предполагает, что любое значение, равное или превышающее 1, должно быть усечено до максимального значения 1. Линия регрессии будет прямой до этого максимального значения, но последующие изменения в *X* не будут иметь никакого влияния на зависимую переменную. То же самое можно было бы сказать и о малых значениях, для которых можно применить усечение в 0. Такой подход привел бы к внезапным разрывам нашей зависимости, в результате чего в определенных точках влияние *X* на *Y* немедленно стало бы нулевым (см. рис. 3).

Однако, помимо усеченной линейности, больший смысл может иметь другой, S-образный вид функциональной зависимости. Мы можем задать «пол» и «потолок», при этом предположить, что влияние предиктора на зависимую переменную (при изменении предиктора на единицу своего измерения) будет меньше в районе «пола» и «потолка», чем посередине. Таким образом, вводим нелинейность. В центре нашей зависимости нелинейная кривая может аппроксимировать линейность, но вместо бесконечного движения вверх или вниз нелинейная кривая, приближаясь к 0 или 1 по оси *Y*, медленно и плавно изгибается по оси *X*. По мере приближения к 0 или 1 по оси *Y* требуется все большее изменение предиктора, чтобы оказать такое же влияние на зависимую переменную, что и меньшее изменение предиктора в середине

кривой. Для изменения вероятности возникновения события с 0,95 до 0,96 требуется большее изменение *X*, чем для изменения вероятности события с 0,45 до 0,46.

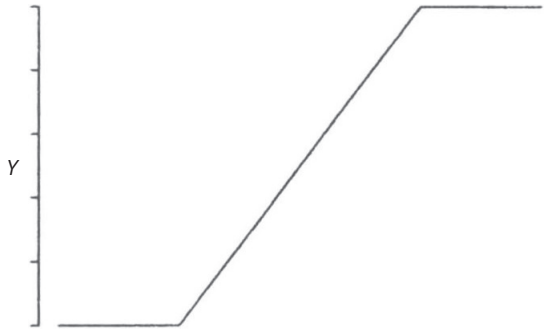


Рис. 3 ❖ Усеченная линейная зависимость

Итак, главный принцип нашей зависимости заключается в том, что одно и то же изменение предиктора оказывает меньшее влияние на зависимую переменную в районе «потолка» и «пола» и нашему предиктору потребуется гораздо большее изменение, чтобы оказать то же самое влияние на зависимую переменную в районе «потолка» и «пола».

Проиллюстрируем нелинейную зависимость на нескольких примерах. Если доход увеличивает вероятность владения домом, то увеличение дохода на 10 тысяч долларов с 40 000 до 50 000 долларов увеличит эту вероятность в большей степени, нежели увеличение дохода с 200 000 до 210 000 долларов. Без сомнения, люди с высоким доходом уже имеют высокую вероятность владения жильем, а увеличение на 10 000 долларов незначительно увеличило бы их и без того высокую вероятность владения домом. То же самое можно сказать и об увеличении дохода с 0 до 10 000 долларов: поскольку обе суммы вряд ли будут достаточными для покупки дома, увеличение дохода мало повлияет на возможность обладать собственностью. Однако в середине диапазона дополнительные 10 000 долларов могут иметь значение в плане прогнозирования наличия/отсутствия дома.

Аналогичным образом увеличение возраста на 1 год влияет на вероятность вступления в брак гораздо сильнее в раннем молодом возрасте, нежели в совсем юном или зрелом возрасте. Мало кто женится в возрасте до 15 лет, даже если он и становится на 1 год старше, и мало кто, будучи холостым в 50 лет, женится в 51 год. Однако изменение возраста с 21 до 22 лет может привести к существенному увеличению вероятности вступления в брак.

Подобные рассуждения применимы и во многих других случаях: влияние количества сверстников с девиантным поведением на вероятность совершения тяжкого преступления, влияние продолжительности рабочего времени женщин на вероятность рождения ребенка, влияние употребления алкоголя на преждевременную смерть – скорее всего, сильнее в средних значениях диапазонов независимых переменных, чем в крайних случаях.

Более подходящая нелинейная зависимость будет выглядеть так, как показано на рис. 4, на котором кривая выравнивается и приближается к «потолку», определяемому значением 1 по оси *Y*, и к «полу», определяемому значением 0 по оси *Y*. Для аппроксимации кривой потребуется последовательность прямых линий, каждая из которых будет иметь разные коэффициенты наклона. Линии ближе к потолку и полу будут иметь меньшие коэффициенты наклона, чем в середине. Тем не менее постоянно меняющаяся кривая более плавно и адекватно представляет зависимость. Концептуально S-образная кривая имеет больший смысл, чем прямая.

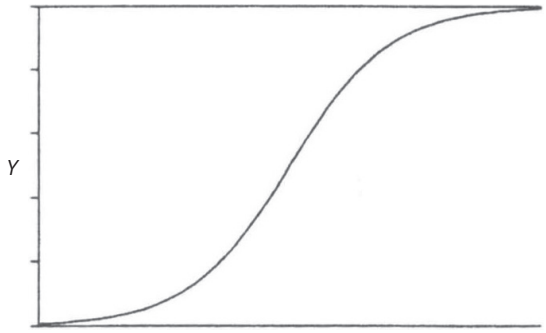


Рис. 4 ❖ S-образная кривая

В пределах диапазона выборочных значений линия линейной регрессии может аппроксимировать криволинейную зависимость, принимая среднее значение различных коэффициентов наклона, подразумеваемых кривой. Тем не менее линейность все еще неадекватно представляет фактические зависимости в середине и переоценивает зависимости в экстремальных значениях (если только у предиктора не отсутствуют значения в области, где кривая почти линейна). На рис. 5 мы сравниваем S-образную кривую с прямой линией. Разрыв между ними иллюстрирует природу ошибки и потенциальную неточность линейной регрессии.

Помимо нелинейности, «потолок» и «пол» создают еще одну концептуальную проблему в обычной регрессии с бинарной зависимой переменной. Регрессия обычно предполагает аддитивность, т. е. влияние одной переменной на зависимую переменную остается неизменным независимо от уровней других независимых переменных. Модели могут включать отобранные произведения членов для учета неаддитивности, но бинарная зависимая переменная, вероятно, нарушает предположение об аддитивности для всех комбинаций независимых переменных. Если значение одной независимой переменной достигает достаточно высокого уровня, чтобы сдвинуть вероятность зависимой переменной к 1 (или к 0), то влияние остальных переменных не может увеличиться. Таким образом, «потолок» и «пол» делают влияние всех независимых переменных по своей природе неаддитивным и интерактивным.

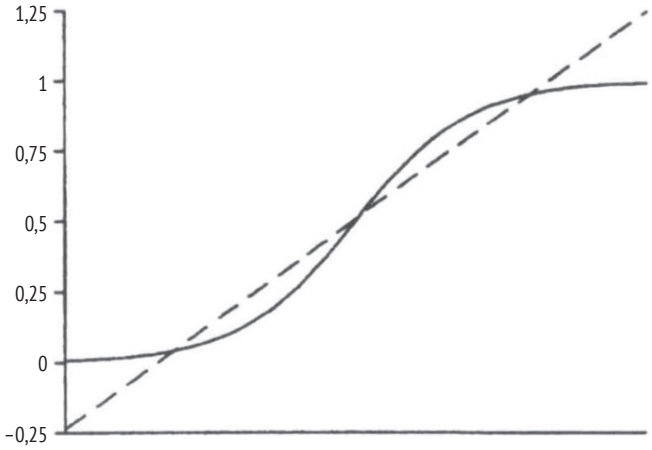


Рис. 5 ❖ Сравнение линейной и криволинейной зависимостей

Вернемся к примеру с курением. Люди с 20-летним образованием имеют настолько низкую вероятность курения, что между мужчинами и женщинами может существовать лишь небольшая разница. Иными словами, пол может слабо влиять на курение при высоком уровне образования. Напротив, более выраженная разница между полами, возможно, проявляется, когда уровень образования ниже, а вероятность курения выше. Хотя влияние пола на курение, вероятно, зависит от уровня образования, аддитивные регрессионные модели ошибочно предполагают, что влияние пола на курение одинаково для всех уровней образования (а влияние образования одинаково для обоих полов).

1.2. Проблема достоверности статистического вывода

Даже если в некоторых случаях прямая линия аппроксимирует нелинейные зависимости, возникают некоторые проблемы, которые снижают эффективность оценок, несмотря на то что оценки остаются несмещенными. Проблемы связаны с тем, что регрессия с бинарной зависимой переменной нарушает предположения о нормальности и гомоскедастичности. Обе эти проблемы возникают по причине существования только двух наблюдаемых значений для зависимой переменной. Линейная регрессия предполагает, что в генеральной совокупности ошибки модели подчиняются нормальному распределению, а дисперсия ошибок для каждого значения X одинакова. Одним словом, у нас должно быть нормальное распределение ошибок с одинаковой дисперсией.

Однако в случае с бинарной зависимой переменной только два значения Y и только два остатка существуют для любого отдельного значения X. Для любого значения X_i прогнозируемая вероятность равна $b_0 + b_1X_i$. Следовательно, остатки принимают значение

$1 - (b_0 + b_1X_i)$, когда Y_i равно 1,

и

$0 - (b_0 + b_1X_i)$, когда Y_i равно 0.

Даже в генеральной совокупности распределение ошибок для любого значения X не может быть нормальным, если распределение имеет только два значения.

Кроме того, у нас нарушается предположение о гомоскедастичности или постоянной дисперсии, поскольку ошибка регрессии изменяется со значением X. Чтобы проиллюстрировать это графически, посмотрите на рис. 6, на котором показана зависимость между X и бинарной зависимой переменной. Подгонка с помощью прямой линии, идущей от нижнего левого к верхнему правому углу рисунка, будет определять остатки как вертикальное расстояние от точек до линии. Вблизи нижних и верхних экстремальных значений X, где линия приближается к полу 0 и потолку 1, остатки относительно невелики. Вблизи средних значений X, где линия находится на полпути между потолком и полом, остатки относительно велики. В результате дисперсия ошибок не является постоянной.

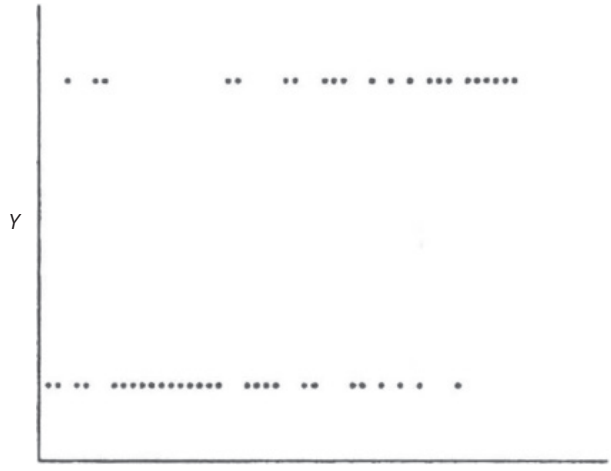


Рис. 6 ❖ Диаграмма рассеяния для бинарной зависимой переменной

Если нарушение предположения о нормальности создает мало проблем при работе на больших выборках, гетероскедастичность имеет более серьезные последствия. Выборочные оценки коэффициентов регрессии являются несмещенными, но они больше не характеризуются наименьшей дисперсией, и выборочные оценки стандартных ошибок будут смещенными. Таким образом, даже при больших выборках стандартные ошибки при наличии гетероскедастичности будут некорректными, а тесты значимости будут невалидными (непригодными). Технически метод взвешенных наименьших

квадратов может решить эту проблему, однако не решаются концептуальные проблемы нелинейности и неаддитивности, что более важно. Поэтому использование линейной регрессии с бинарной зависимой переменной остается неуместным.

2. Знакомство с логистической функцией

Ранее мы выяснили, что обычная линейная регрессия с бинарной зависимой переменной сталкивается с рядом трудностей. На помощь нам приходит модель логистической регрессии. В рамках такой модели мы строим модель вероятности того, что бинарная зависимая переменная примет значение 1 при заданных значениях независимых переменных. Для моделирования вероятности бинарной зависимой переменной подбирают специальную монотонно возрастающую логистическую функцию (логистический сигмоид), которая может принимать значения только от 0 до 1. Она имеет вид $y = \frac{1}{1+e^{-x}}$, где x – это логит. Внимательно посмотрим на нее (рис. 7).

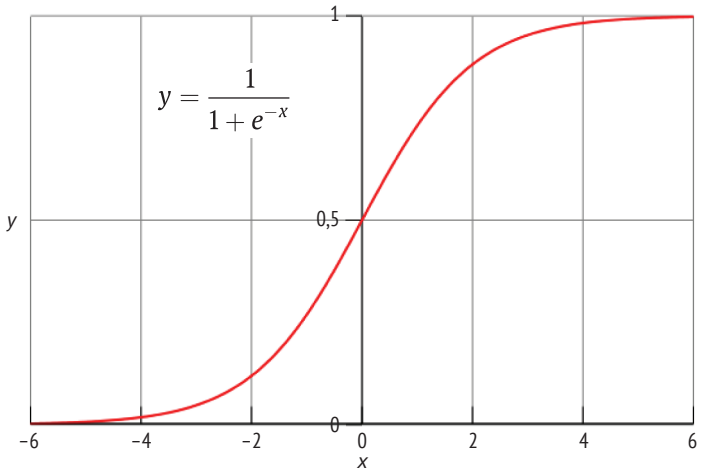


Рис. 7 ❖ Логистическая функция

Мы сразу замечаем, что независимо от того, какое значение может принимать аргумент x (неограниченная область определения функции), сигмоид ограничен двумя горизонтальными асимптотами, к которым стремится при стремлении аргумента к $\pm\infty$ (получаем ограниченный диапазон значений функции). Обычно этими асимптотами являются $y = 0$ в $-\infty$ и $y = 1$ в $+\infty$. Наличие ограниченного диапазона вещественных чисел от 0 до 1 нужно нам для прогнозирования вероятности.

Мы видим, что наша функция симметрична относительно точки перегиба $(x, y) = (0, 0.5)$, которая является серединой диапазона значений функции.

Это означает, что в дополнение к тому, что *y* ограничена вещественными числами от 0 до 1, значения *y* также будут симметрично распределены по обе стороны от точки перегиба. Это позволяет нам определить $y = 0{,}5$ в качестве точки, в которой прогнозируемый класс меняется с 0 на 1 (или наоборот), и вероятность того, что наблюдение принадлежит классу 1 (положительному классу), составляет точно 50 %. Эта средняя точка / точка перегиба будет решающей границей и будет использоваться для прогнозирования классов.

А еще мы видим, что единичное изменение по оси *x* вблизи «пола» и «потолка» приводит к меньшему изменению вероятности *y*, нежели единичное изменение по оси *x* посередине кривой.

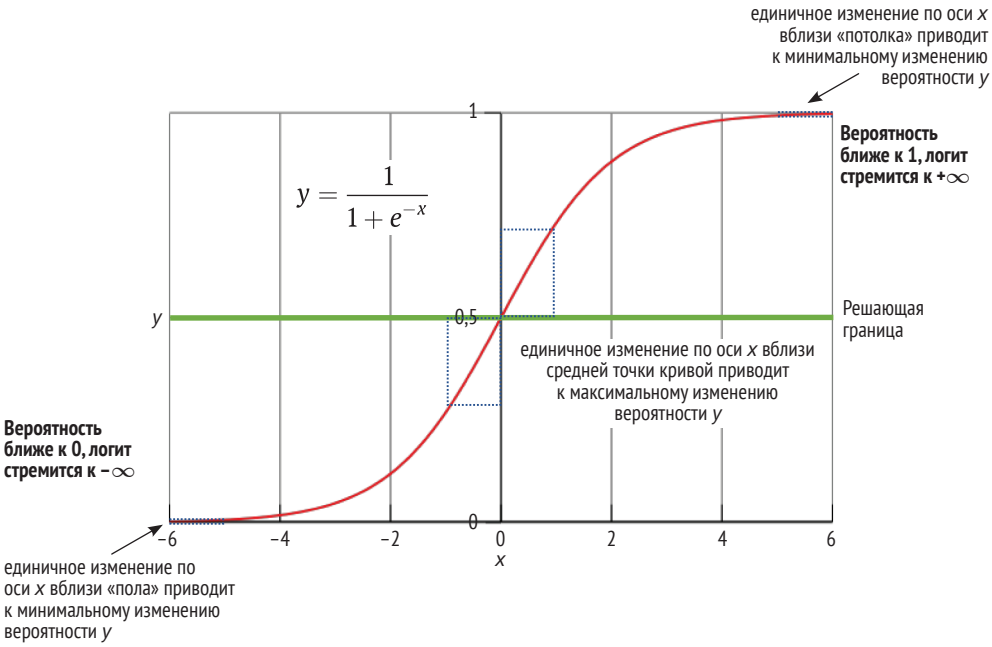


Рис. 8 ❖ Свойства логистической функции

Теперь нам предстоит выяснить, что представляет из себя логит, использующийся в уравнении логистической функции.

3. Преобразование вероятностей в логиты

Логит тесно связан с понятиями «вероятность» и «шанс».

Вероятность – это объективная мера появления некоторого события, измеряемая от 0 до 1. На практике оценкой вероятности служит относительная частота появления события. Значение вероятности 0 означает невозможность появления события. Значение вероятности 1 означает, что событие непременно произойдет.

Вероятность (probability): Y_i

Шансы – это отношение вероятности того, что событие произойдет, к вероятности того, что событие не произойдет. Можно еще сказать так: шансы – это отношение вероятности наступления события к вероятности ненаступления события. Вероятность наступления события часто называют просто вероятностью события, и когда вы встречаете фразы «вычислить вероятность», «оценить влияние предикторов на вероятности» в контексте логистической регрессии, то речь идет именно о вероятности события. С ростом вероятности растут шансы, и наоборот. Значение шансов 1 соответствует ситуации, когда вероятности наступления события и ненаступления события равны.

Шансы (odds): $\frac{P_i}{1-P_i}$

Наконец, логит – это натуральный логарифм шансов.

Логит (logit), логарифм шансов (log odds),

прологарифмированные шансы (logged odds): $\ln\left(\frac{P_i}{1-P_i}\right)$

Поупражняемся вычислять шансы и логиты.

Например, если P_i для первого наблюдения равно 0,2, то шансы равны 0,25, или 0,2/0,8, а логит равен –1,386, т. е. натуральному логарифму шансов.

Если P_i для второго наблюдения равно 0,7, то шансы равны 2,33, или 0,7/0,3, а логит равен 0,847.

Если P_i для третьего наблюдения равно 0,9, то шансы равны 9, или 0,9/0,1, а логит равен 2,197.

Хотя формула преобразования вероятностей в логиты проста, требуется некоторое объяснение, чтобы проиллюстрировать ее полезность. Оказывается, она прекрасно описывает зависимость между предикторами и распределением вероятностей, определяемым бинарной зависимой переменной. Формула включает два шага: на первом шаге мы берем отношение вероятности, что событие произойдет, к вероятности, что событие не произойдет, $\frac{P_i}{1-P_i}$, и получаем шансы возникновения события; на втором шаге берем натуральный логарифм шансов и получаем логит. Давайте подробнее раскроем смысл шансов, которые мы получаем на первом шаге формулы.

3.1. Смысл шансов

Итак, вычисление логита начинается с преобразования вероятностей в шансы. Вероятности варьируют от 0 до 1. И вероятность, и шансы имеют нижний предел, равный нулю, и оба выражают растущую вероятность события по мере увеличения положительных чисел, но в остальном они различаются.

В отличие от вероятности, шансы не имеют верхней границы, или «потолка». Когда вероятность становится ближе к 1, числитель в формуле шансов

становится больше относительно знаменателя, и шансы постоянно растут. Таким образом, шансы значительно увеличиваются, когда вероятности незначительно изменяются вблизи их верхней границы 1. Например, вероятности 0,99, 0,999, 0,9999, 0,99999 и т. д. дают шансы 99, 999, 9999, 99999 и т. д. Незначительные изменения вероятностей приводят к огромным изменениям шансов и показывают, что шансы бесконечно увеличиваются, по мере того как вероятности становятся все ближе и ближе к 1.

Чтобы проиллюстрировать взаимосвязь между вероятностями и шансами, рассмотрим значения:

P_i	0,01	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,99
$1 - P_i$	0,99	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,01
Шансы	0,01	0,111	0,25	0,429	0,667	1	1,5	2,33	4	9	99

Обратите внимание, что когда вероятность равна 0,5, шансы равны 1 или одинаковы.

Часто шансы выражают как отношение числа к единице. Например, если вероятность просрочки 90+ (событие произошло) равна 0,8, тогда вероятность отсутствия просрочки 90+ (событие не произошло) равна $1 - 0{,}8 = 0{,}2$, шансы наличия просрочки 90+ равны $0{,}8/0{,}2 = 4$. Это означает, что шансы наличия просрочки 90+ составляют 4 к 1. Шансы отсутствия просрочки 90+ будут равны $0{,}2/0{,}8 = 0{,}25$. Выглядит немного странно, но действительно шансы отсутствия просрочки 90+ будут равны 1 к 4. Шансы наличия просрочки 90+ и шансы отсутствия просрочки 90+ являются обратными величинами по отношению друг к другу, т. е. $1/4 = 0{,}25$ и $1/0{,}25 = 4$.

Шансы больше 1 означают, что вероятность возникновения события больше вероятности отсутствия события. Шансы 9 означают, что вероятность возникновения события в 9 раз больше вероятности отсутствия события. Например, вероятность наличия просрочки равна 0,9, а вероятность отсутствия просрочки равна 0,1, $0{,}9/0{,}1 = 9$. Шансы меньше 1 означают, что вероятность возникновения события меньше вероятности отсутствия события. Шансы 0,111 означают, что вероятность возникновения события в 0,111 раза меньше вероятности отсутствия события. Например, вероятность наличия просрочки равна 0,1, а вероятность отсутствия просрочки равна 0,9, $0{,}1/0{,}9 = 0{,}111$. Шансы, равные 1, означают, что вероятности возникновения события и отсутствия события равны.

Манипуляции с формулой вычисления шансов дает более глубокое понимание их связи с вероятностями. Определив шансы O_i как отношение вероятности события к единице минус вероятность события, мы можем с помощью простой алгебры выразить вероятность в терминах шансов: $O_i = \dfrac{P_i}{1 - P_i}$ подразумевает, что $P_i = \dfrac{O_i}{1 + O_i}$.

$O_i = P_i/(1 - P_i)$,
$P_i = O_i * (1 - P_i)$,

$P_i = O_i - O_i * P_i$,
$P_i + O_i * P_i = O_i$,
$P_i(1 + O_i) = O_i$,
$P_i = O_i/(1 + O_i)$.

Вероятность равна шансам, поделенным на единицу плюс шансы. В нашем примере с просрочкой 90+ мы, зная шансы $\frac{P_i}{1-P_i} = \frac{0,8}{0,2} = 4$, можем легко получить вероятность по формуле $\frac{O_i}{1+O_i} = \frac{4}{1+4} = 0,8$.

Из формулы $\frac{O_i}{1+O_i}$ становится понятно, что вероятность никогда не может быть равна или превышать единицу: независимо от того, насколько большими стали шансы в числителе, они всегда будут на единицу меньше шансов в знаменателе. Конечно, когда шансы станут большими, разница между шансами и шансами плюс единица станет относительно небольшой, и вероятность приблизится к 1 (но не достигнет ее). И наоборот, вероятность никогда не может упасть ниже 0. Если шансы равны 0 или превышают его, то вероятность должна быть равна 0 или превышать его. Вероятность становится все ближе к 0, по мере того как шансы становятся все ближе к 0.

Не путайте шансы с **отношениями шансов (odds ratio)**. Шансы – это отношение вероятностей, тогда как отношение шансов – это именно отношение шансов, или отношение отношений вероятностей. Например, согласно опросу, 29,5 % мужчин и 13,1 % женщин владеют оружием. Поскольку шансы владеть оружием для мужчин равны 0,418 (0,295/0,705), это означает, что 4 мужчины с оружием приходятся на 10 мужчин без оружия. Шансы владеть оружием для женщин равны 0,151 (0,131/0,869), это означает, что примерно 1,5 женщины с оружием приходятся на 10 женщин без оружия. Отношение шансов равно 0,418/0,151 = 2,77. Это означает, что шансы владеть оружием у мужчин почти в три раза выше, чем у женщин.

3.2. Смысл логитов

Использование натурального логарифма шансов исключает минимальное значение 0 («пол») так же, как преобразование вероятностей в шансы исключает максимальное значение 1 («потолок»). Когда мы берем:

- натуральный логарифм шансов выше 0, но ниже 1, мы получаем отрицательные числа;
- натуральный логарифм шансов, равный 1, мы получаем 0;
- натуральный логарифм шансов выше 1, мы получаем положительные числа.

Напомним, что логарифм 0 и отрицательных чисел не существует.

Таким образом, первое свойство логита состоит в том, что, в отличие от вероятности, он не имеет верхней или нижней границы. Шансы устраняют верхнюю границу вероятностей, а прологарифмированные шансы устраняют нижнюю границу вероятностей. Давайте убедимся в этом. Если $P_i = 1$, логит не определен, потому что шансы 1/0 не существуют. По мере того как вероятность приближается к 1, логит движется к $+\infty$. Если $P_i = 0$, логит не определен, потому что логарифм шансов 0/1 или 0 не существует. По мере того как вероятность приближается к 0, логит движется к $-\infty$. Таким образом, логит варьирует от $-\infty$ до $+\infty$. Проблема нижней и верхней границ для вероятностей (или нижней границы для шансов) отпадает.

Второе свойство логита заключается в том, что логит-преобразование симметрично относительно вероятности 0,5. Когда $P_i = 0{,}5$, логит равен 0 (0,5/0,5 = 1, а логарифм 1 равно 0). Вероятности ниже 0,5 (P_i меньше $1 - P_i$) приведут к отрицательным логитам, потому что шансы падают ниже 1, но выше 0. А из курса школьной математики мы помним, что логарифм чисел выше 0, но ниже 1 дает отрицательное число. Вероятности выше 0,5 (P_i выше $1 - P_i$) приведут к положительным логитам, потому что шансы превышают 1. Опять же из курса школьной математики помним, что логарифм чисел выше 1 дает положительное число.

Кроме того, вероятности, которые находятся на одинаковом расстоянии от 0,5 выше или ниже (например, 0,6 и 0,4, 0,7 и 0,3, 0,8 и 0,2), имеют одинаковые логиты, но с разными знаками (например, логиты для вероятностей, перечисленных выше, равны 0,405 и –0,405, –0,847 и –0,847, 1,386 и –1,386). Удаленность логита от 0 отражает удаленность вероятности от 0,5 (опять же отметим, что логиты не имеют границ).

Третье свойство логита заключается в том, что одно и то же изменение вероятности приводит к различным изменениям в логитах. Простой принцип заключается в том, что по мере приближения P_i к 0 и 1 одно и то же изменение вероятности приводит к большему изменению логита. Вы можете увидеть это на примере:

P_i	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
$1 - P_i$	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1
Шансы	0,111	0,25	0,429	0,667	1	1,5	2,33	4	9
Логит	–2,2	–1,39	–0,847	–0,405	0	0,405	0,847	1,39	2,2
Δ логит		0,81	0,543	0,442	0,405	0,405	0,442	0,543	0,81

Изменение вероятности на 0,1 с 0,5 до 0,6 (или с 0,5 до 0,4) приводит к изменению логита на 0,405, тогда как такое же изменение вероятности на 0,1 с 0,8 до 0,9 (или с 0,2 до 0,1) приводит к изменению логита на 0,81.

Для одного и того же изменения вероятности изменение логита в крайних значениях вероятности будет в два раза больше изменения логита для среднего значения вероятности. Повторим, что общий принцип заключается в том, что небольшое изменение вероятности приводит к большому изменению логита, когда вероятности находятся вблизи границ 0 и 1.

4. Линеаризация нелинейности

В итоге мы можем рассматривать логит-преобразование как линеаризацию нелинейной зависимости между *X* и вероятностью *Y*. Мы ожидаем, что одно и то же изменение *X* окажет меньшее влияние на вероятность *Y* в районе «пола» или «потолка», чем в районе средней точки. Поскольку логит расширяет или растягивает вероятности в районе крайних значений *Y* в сравнении со значениями *Y*, близкими к средней точке, одно и то же изменение *X* приводит к схожим эффектам во всем диапазоне логит-преобразования вероятности *Y*. Другими словами, без «пола» и «потолка» логит может линейно соотноситься с изменениями *X*. Теперь можно вычислить линейную зависимость между *X* и логит-преобразованием. Логит-преобразование линеаризует нелинейную связь между *X* и исходными вероятностями.

И наоборот, линейная связь между *X* и логитом подразумевает нелинейную связь между *X* и исходными вероятностями. Изменение логита на единицу приводит к меньшим изменениям вероятности в крайних значениях, чем в среднем значении. Подобно тому, как мы переводим вероятности в логиты, мы можем переводить логиты в вероятности.

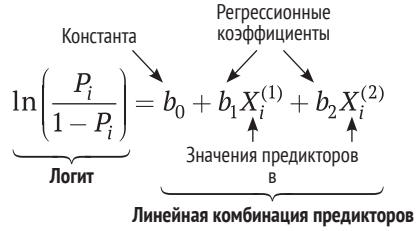
Логит	−3	−2	−1	0	1	2	3
P_i	0,047	0,119	0,269	0,5	0,731	0,881	0,953
ΔP_i	0,072	0,15	0,231	0,231	0,15	0,072	

Изменение логарифма на одну единицу приводит к большему изменению вероятности вблизи среднего значения, чем вблизи крайних значений. Другими словами, линейность логитов определяет теоретически значимую нелинейную связь с вероятностями.

4.1. Получение вероятностей из логитов

Итак, линейная связь между предикторами и логитом подразумевает нелинейную связь между предикторами и вероятностями. Линейную зависимость предикторов от спрогнозированного логита можно выразить в виде уравнения:

Модель логистической регрессии, выраженная через логит



Именно оно и является главным уравнением в методе логистической регрессии. В левой части уравнения у нас – логит, а в правой части – линейная

комбинация предикторов, состоящая из свободного члена или константы (b_0), регрессионных коэффициентов (b_1, b_2) и предикторов ($X_i^{(1)}$, $X_i^{(2)}$). Именно с помощью коэффициентов и значений предикторов в i-м наблюдении мы прогнозируем логит для i-го наблюдения. Константу и коэффициенты мы получаем, решая задачу минимизации логистической функции потерь с помощью того или иного метода оптимизации (например, с помощью градиентного спуска). Константу еще называют смещением, а коэффициенты – весами.

Итак, мы выразили модель логистической регрессии через логит. А теперь выразим модель логистической регрессии через вероятность. Для этого из школьного курса математики вспомним экспоненцирование. К примеру, у нас есть выражение $\log_5(2x + 3) = 3$. Проэкспоненцируем обе части, получаем $2x + 3 = 5^3$.

Теперь возьмем наше уравнение $\ln\left(\frac{P_i}{1-P_i}\right) = b_0 + b_1X_i^{(1)} + b_2X_i^{(2)}$. Проэкспоненцировав обе части уравнения $\ln\left(\frac{P_i}{1-P_i}\right) = b_0 + b_1X_i^{(1)} + b_2X_i^{(2)}$, мы видим, что в левой части мы лишаемся логарифма и от логита переходим к шансам:

$$\frac{P_i}{1-P_i} = e^{b_0} * e^{b_1X_i^{(1)}} * e^{b_2X_i^{(2)}}$$

или

$$\frac{P_i}{1-P_i} = e^{b_0+b_1X_i^{(1)}+b_2X_i^{(2)}}$$

(вспоминаем, что $e^X * e^Y = e^{X+Y}$).

А от шансов легко перейти к вероятностям:

$$\frac{P_i}{1-P_i} = e^{b_0+b_1X_i^{(1)}+b_2X_i^{(2)}},$$

$$P_i = e^{b_0+b_1X_i^{(1)}+b_2X_i^{(2)}} * (1-P_i),$$

$$P_i = 1 * (e^{b_0+b_1X_i^{(1)}+b_2X_i^{(2)}}) - P_i * (e^{b_0+b_1X_i^{(1)}+b_2X_i^{(2)}}),$$

$$P_i + P_i * (e^{b_0+b_1X_i^{(1)}+b_2X_i^{(2)}}) = (e^{b_0+b_1X_i^{(1)}+b_2X_i^{(2)}}),$$

$$P_i * (1 + e^{b_0+b_1X_i^{(1)}+b_2X_i^{(2)}}) = (e^{b_0+b_1X_i^{(1)}+b_2X_i^{(2)}}),$$

$$P_i = e^{b_0+b_1X_i^{(1)}+b_2X_i^{(2)}}/(1 + e^{b_0+b_1X_i^{(1)}+b_2X_i^{(2)}}).$$

Поскольку логит равен $b_0 + b_1X_i^{(1)} + b_2X_i^{(2)}$, мы можем заменить эту длинную формулу на L_i. Тогда получаем:

Модель логистической регрессии, выраженная через вероятность

$$P_i = e^{L_i}/(1 + e^{L_i}).$$

Учитывая, что e^{L_i} – это шансы, наша формула эквивалентна уравнению $P_i = \frac{O_i}{1+O_i}$, представленному ранее.

Теперь преобразуем логиты в экспоненты логитов, а затем экспоненты логитов преобразуем в вероятности.

L	–6,91	–4,61	–2,30	–1,61	–0,223	0	0,223	1,61	2,30	4,61	6,91
e^L	0,001	0,01	0,1	0,2	0,8	1	1,25	5	10	100	1000
$1 + e^L$	1,001	1,01	1,1	1,2	1,8	2	2,25	6	11	101	1001
P	0,001	0,01	0,091	0,167	0,444	0,5	0,556	0,883	0,909	0,99	0,999

Сначала обратите внимание, что экспоненты отрицательных логитов лежат между 0 и 1, а экспоненты положительных логитов превышают единицу. Отметим также, что отношение экспоненты логита к экспоненте логита плюс 1 всегда будет ниже единицы – знаменатель будет всегда превышать числитель на 1. Однако по мере увеличения экспоненты логита разница между числителем и знаменателем уменьшается (другими словами, лишняя единица в знаменателе становится все меньше относительно значения в числителе). Кроме того, отношение экспоненты логита к экспоненте логита плюс 1 никогда не может опускаться ниже нуля, поскольку экспоненты как отрицательных, так и положительных чисел оказываются положительными, а отношение двух положительных чисел всегда оказывается положительным. Учитывая границы вероятностей, пример показывает, что чем больше L, тем больше e^L и больше P.

Кроме того, это преобразование демонстрирует нелинейность. При изменении X на единицу L изменяется на постоянную величину, а P – нет. Экспоненты в формуле вычисления P_i делают связь нелинейной. Рассмотрим пример. Если $L_i = 2 + 0{,}3X_i$, то логит изменится на 0,3 при изменении X на единицу независимо от значения X. Если X меняется с 1 на 2, то L меняется с 2 + 0,3 на 2 + 0,3 * 2, т. е. с 2,3 на 2,6. Если X меняется с 11 на 12, то L меняется с 2 + 0,3 * 11 на 2 + 0,3 * 12 , т. е. с 5,3 на 5,6. В обоих случаях измение было одинаковым и было равно 0,3. Это определяет линейность.

Возьмем те же самые значения X и значения L, которые мы получаем при данных значениях X, и запишем изменения, которые они подразумевают, в вероятностях.

X	1	2	11	12
L	2,3	2,6	5,3	5,6
ΔL	0,3		0,3	
e^L	9,97	13,46	200,3	270,4
$1 + e^L$	10,97	14,46	201,3	271,4
P	0,909	0,931	0,995	0,996
ΔP	0,022		0,001	

Одно и то же изменение L, вызванное изменением X на единицу, приводит к большему изменению вероятностей для более низких значений X и P, чем для более высоких значений. Аналогичную картину можно увидеть на другом конце распределения вероятностей.

Эта нелинейность между логитом и вероятностью создает фундаментальную проблему интерпретации. Мы можем подытожить влияние X на логит просто в виде одного линейного коэффициента, но мы не можем сделать то же самое с вероятностями: влияние X на вероятность меняется в зависимости от значения X и значения вероятности. Сложности, возникающие при интерпретации влияния на вероятность, требуют отдельного разбора содержательного смысла коэффициентов логистической регрессии. Однако решение проблем интерпретации выглядит проще, если полностью рассмотреть логику логит-преобразования.

4.2. Альтернативная формула

Формулу вычисления вероятности как функции независимых переменных и коэффициентов можно записать в несколько более простом, но менее интуитивном виде:

$$P_i = 1/(1 + e^{-(b_0+b_1X_i^{(1)}+b_2X_i^{(2)})}),$$

$$P_i = 1/(1 + e^{-L_i}).$$

Давайте подробнее рассмотрим, как мы от формулы $P_i = e^{L_i}/(1 + e^{L_i})$пришли к формуле $P_i = 1/(1 + e^{-L_i})$.

Из школьного курса математики вспомним, что $e^{-X} = 1/e^X$, $e^X = 1/e^{-X}$, и примем, что $b_0 + b_1X_i^{(1)} + b_2X_i^{(2)}$ равно L_i.

$$P_i = e^{L_i}/(1 + e^{L_i}),$$

$$P_i = (1/e^{-L_i})/(1 + e^{L_i}/1),$$

$$P_i = (1/e^{-L_i}) * (1/1 + e^{L_i}),$$

$$P_i = 1/[(e^{-L_i}) * (1 + e^{L_i})],$$

$$P_i = 1/(e^{-L_i} + e^{L_i} * e^{-L_i}).$$

Поскольку $e^X * e^Y = e^{X+Y}$, а $e^{X-X} = e^0 = 1$, то формула сокращается до следующего вида:

$$P_i = 1/(e^{-L_i} + 1) = 1/(1 + e^{-L_i}).$$

Теперь вероятность равна единице, поделенной на единицу плюс экспонента логита, взятого с отрицательным знаком. Эта формула дает нам точно такой же результат, что и ранее приведенная формула $P_i = e^{L_i}/(1 + e^{L_i})$.

Убедимся, что обе формулы позволяют перевести логиты в вероятности и дают одни и те же результаты. Если логит равен –2,302, то мы решаем $P_i = e^{-2,302}/(1 + e^{-2,302})$ или $P_i = 1/(1 + e^{-(-2,302)})$. Экспонента числа –2,302 равна приблизительно 0,1, а экспонента числа –2,302, взятого с отрицательным знаком, т. е. экспонента числа 2,302, равна 9,994. Таким образом, вероятность равна $\frac{0,1}{1,1} = 0,091$, и по альтернативной формуле вероятность тоже равна $\frac{1}{1+9,994} = 0,091$.

5. Выводы

Итак, мы рассмотрели, как логит преобразует зависимую переменную, имеющую естественную нелинейную связь с набором независимых переменных, в зависимую переменную, линейно связанную с набором независимых переменных. Таким образом, модели логистической регрессии оценивают линейные детерминанты прологарифмированных шансов или логитов, а не нелинейные детерминанты вероятностей. Получение этих оценок связано со сложностями, которые мы рассмотрим позже. Однако это помогает рассматривать логистическую регрессию в простых терминах как регрессию зависимой переменной, которая превращает нелинейные связи в линейные.

При линеаризации нелинейных связей логистическая регрессия также сдвигает интерпретацию коэффициентов от изменений вероятностей к менее интуитивно понятным изменениям логитов. Однако потеря интерпретируемости в случае с коэффициентами логистической регрессии компенсируется выигрышем с т. з. компактности анализа: линейную связь с логитом можно подытожить с помощью одного коэффициента, однако нелинейную связь с вероятностью нельзя так просто обобщить.