Нелинейная парная регресия

Предположим, вы считаете, что переменная y связана с переменной x следующим соотношением:

$$y = \alpha + \beta\, x^{\gamma} + u, \tag{5.28}$$

и хотите получить оценки α, β и γ, имея значения y и x. Уравнение (5.28) не может быть преобразовано в уравнение линейного вида, поэтому в этом случае невозможно применение обычной процедуры оценивания регрессии.

Тем не менее, для получения оценок параметров мы по-прежнему можем применить принцип минимизации суммы квадратов отклонений.

Процедуру лучше всего описать как последовательность шагов.

1. Принимаются некоторые правдоподобные исходные значения параметров.

2. Вычисляются предсказанные значения y по фактическим значениям x с использованием этих значений параметров.

3. Вычисляются остатки для всех наблюдений в выборке и, следовательно, S – сумма квадратов остатков.

4. Вносятся небольшие изменения в одну или более оценку параметров.

5. Вычисляются новые предсказанные значения y, остатки и S.

6. Если S меньше, чем прежде, то новые оценки параметров лучше прежних, и их следует использовать в качестве новой отправной точки.

7. Шаги 4, 5 и 6 повторяются вновь до тех пор, пока не окажется невозможным внести такие изменения в оценки параметров, которые привели бы к уменьшению S.

8. Делается вывод о том, что величина S минимизирована, и конечные оценки параметров являются оценками по методу наименьших квадратов.

Пример

Вернемся к примеру с бананами, рассмотренному в разделе 4.1, где y и x связаны следующей зависимостью:

$$y = \alpha + \frac{\beta}{x} + u. \tag{5.29}$$

Для большей простоты предположим, что мы знаем, что $\alpha = 12$; следовательно, нам нужно определить только один неизвестный параметр. Предположим, мы поняли, что зависимость имеет вид (5.29), однако не можем догадаться, что следует применить преобразования, рассмотренные в разделе 5.1. Вместо этого мы применяем нелинейную регрессию.

На рис. 5.2 показаны значения S, которые будут получены при любом возможном выборе b при значениях y и x. Предположим, что мы начнем, приняв b равным -6. Уравнение при этом примет вид:

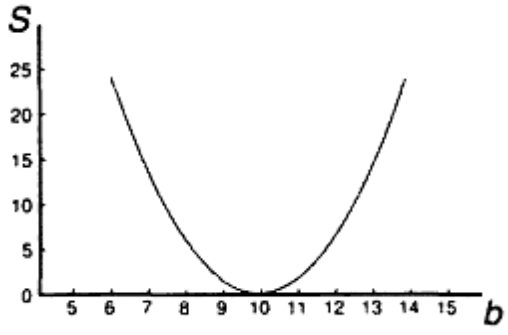

Рис. 5.2. Сумма квадратов отклонений как функция b

$$y = 12 - 6/x. \qquad (5.30)$$

Вычислим предсказанные значения y и остатки, и на основании последних вычислим значение S = 24,02. Затем подставим b = –7. Теперь величина S равна 13,40, т. е. она уменьшилась. Методом деления шага пополам получим, что оценка находится между –9,92 и –9,94. Ясно, что в результате дальнейшего продолжения итерационного процесса могла быть получена более высокая точность.

Заметим, что, хотя полученная оценка очень близка к истинному значению, она не совпадает с оценкой, полученной для уравнения (5.10). В принципе оба набора результатов должны быть одинаковыми, так как и тот и другой минимизируют сумму квадратов отклонений. Расхождение вызвано тем, что мы были не совсем честны в нелинейном случае. Мы предположили, что a равно истинному значению 12, а не оценили его. Если бы мы действительно не смогли найти преобразование, которое позволяет использовать линейный регрессионный анализ, то нам бы пришлось использовать нелинейный метод и искать наилучшие значения a и b одновременно, и тогда мы получили бы оценку a, равную 12,08, и оценку b, равную -10,08, как и в уравнении (5.10).

5.5. Выбор функции: тесты Бокса—Кокса

Возможность построения нелинейных моделей, как с помощью их приведения к линейному виду, так и путем использования нелинейной регрессии, значительно повышает универсальность регрессионного анализа, но и усложняет задачу исследователя. Нужно спросить себя, будете ли вы начинать с линейной зависимости или с нелинейной и если с последней, то какого типа.

Если вы ограничиваетесь парным регрессионным анализом, то можете построить график наблюдений y и x как диаграмму разброса, и это поможет вам принять решение. В примере в разделе 5.2 было очевидно, что зависимость является нелинейной, и не потребовалось бы большого труда, чтобы убедиться, что уравнение вида (5.3) дает почти точное соответствие. Однако обычно все оказывается не так просто. Часто несколько разных нелинейных функций приблизительно соответствуют наблюдениям, если они лежат на некоторой кривой. Однако в случае множественного регрессионного анализа не всегда возможно даже построить график.

При рассмотрении альтернативных моделей с одним и тем же определением зависимой переменной процедура выбора достаточно проста. Наиболее разумным является оценивание регрессии на основе всех

вероятных функций, которые можно вообразить, и выбор функции, в наибольшей степени объясняющей изменения зависимой переменной. Если две или более функции подходят примерно одинаково, то вы должны представить результаты для каждой из них.

Из примера в разделе 5.1 видно, что линейная функция объясняет 64% дисперсии y, а гиперболическая функция (5.3) — 99,9%. В этом примере мы без колебаний выбираем последнюю. Однако если разные модели используют разные функциональные формы, то проблема выбора модели становится более сложной, так как нельзя непосредственно сравнить коэффициенты R2 или суммы квадратов отклонений. В частности — и это наиболее общий пример для данной проблемы, — нельзя сравнить эти статистики для линейного и логарифмического вариантов модели.

Например, линейная регрессия между расходами на жилье и личным располагаемым доходом для США имела коэффициент $R^2 = 0{,}985$, а сумма квадратов отклонений (СКО) была равна 385,2. Для двойной логарифмической версии модели, когда логарифмы берутся по обеим осям, соответствующие значения будут равны 0,9915 и 0,02. Во втором случае, СКО значительно меньше, но это ничего не решает. Значения $\log y$ значительно меньше соответствующих значений y, поэтому неудивительно, что остатки также значительно меньше. Величина R^2 безразмерна, однако в двух уравнениях она относится к разным понятиям. В одном уравнении она измеряет объясненную регрессией долю дисперсии y, а в другом — объясненную регрессией долю дисперсии $\log y$.

Если, с учетом выше сказанного, для одной модели коэффициент R^2 значительно больше, чем для другой, то вы сможете сделать оправданный выбор без особых раздумий, однако, если значения R^2 для двух моделей приблизительно равны, то проблема выбора существенно усложняется.

В этом случае следует использовать стандартную процедуру, известную под названием теста Бокса—Кокса (Box, Cox, 1964). Если вы хотите только сравнить модели с использованием y и $\log y$ в качестве зависимой переменной, то можно использовать вариант теста, разработанный Полом Зарембкой (Zarembka, 1968). Данный тест предполагает такое преобразование масштаба наблюдений y, при котором обеспечивалась бы возможность непосредственного сравнения СКО в линейной и логарифмической моделях. Процедура включает следующие шаги:

1. Вычисляется среднее геометрическое значений y в выборке. (Оно совпадает с экспонентой среднего арифметического $\log y$, поэтому если вы уже оценили логарифмическую регрессию и регрессионная программа выдает вам распечатку среднего значения зависимой переменной, то необходимо вычислить лишь экспоненту от этого значения.)

2. Пересчитываются наблюдения y путем деления на это значение, то есть

$$y_i^* = y_i / (\textit{Среднеегеометрическое } y)$$

где y_i^* — пересчитанное значение для i-го наблюдения.

3. Оценивается регрессия для линейной модели с использованием y^* вместо y в качестве зависимой переменной и для логарифмической модели с использованием $\log y^*$ вместо $\log y$; во всех других отношениях модели должны оставаться неизменными. Теперь значения СКО для двух регрессий сравнимы, и, следовательно, модель с меньшей суммой квадратов отклонений обеспечивает лучшее соответствие.

4. Для того чтобы проверить, не обеспечивает ли одна из моделей значимо лучшее соответствие, можно вычислить величину $(T/2\log Z)$, где T – число наблюдений, отношение значений СКО в пересчитанных регрессиях, и взять ее абсолютное значение (т. е. игнорировать знак «минус», если он имеется). Эта статистика имеет распределение χ^2 с одной степенью свободы. Если она превышает критическое значение χ^2 при выбранном уровне значимости, то делается вывод о наличии значимой разницы в качестве оценивания.

Пример

Тест будет выполнен как для данных о расходах на продукты питания, так и для данных о расходах на жилье в США. Логарифмические регрессии для этих двух видов благ [уравнение (5.18)] показали, что средние значения log y составляют 4,8422 для расходов на питание и 4,6662 для расходов на жилье. Масштабирующие множители равны $e^{4,8422}$ и $e^{4,6662}$ соответственно. В табл. 5.2 приведены значения СКО для линейной и двойной логарифмической регрессии, при этом использованы пересчитанные данные для двух видов благ.

Таблица 5. 2

	Расходы на питание	Расходы на жилье
Линейная регрессия	0,0119	0,0341
Логарифмическая регрессия	0,0119	0,0221

Из табл. 5.2 видно, что для регрессии расходов на питание соответствие одинаково хорошо в обоих случаях. В случае расходов на жилье логарифмическая регрессия дает более точное соответствие. Логарифм отношения значений СКО для двух регрессий равен здесь 0,4337, и, следовательно, после умножения на 12,5 тестовая статистика составляет 5,42. Критический уровень χ^2 с одной степенью свободы составляет 3,84 при 5-процентном уровне значимости и 6,64 – при однопроцентном уровне (см. табл. А.4), так что в данном случае соответствие будет значимо различным

для двух регрессий только при 5-процентном уровне. Эти результаты могут показаться несколько неожиданными, так как можно предположить, что с точки зрения теории модель с логарифмами является более совершенной. Однако период выборки настолько мал, что кривизна функции Энгеля, вероятно, не успеет проявиться, поэтому линейная функция может обеспечить почти столь же хорошее соответствие, как и нелинейная функция.

Следует заметить, что регрессии, пересчитанные по методу Зарембки, могут быть использованы только для того, чтобы решить, какую предпочесть модель. Не надо обращать внимание на коэффициенты, важны только значения СКО. Коэффициенты следует определять непосредственно из непересчитанного варианта выбранной модели.