ДОКЛАДЫ АН ВШ РФ
2016 октябрь–декабрь № 4 (33)

ТЕХНИЧЕСКИЕ НАУКИ

УДК 519.237.5

МАКСИМАЛЬНО ПРАВДОПОДОБНОЕ ОЦЕНИВАНИЕ НЕЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ С ОШИБКОЙ БЕРКСОНА

А.Ю. Тимофеева
Новосибирский государственный технический университет

В экспериментах, направленных на выявление и оценку зависимости между переменными, неизбежны ошибки измерения. Ошибки Берксона искажают значения объясняющей переменной уже после ее измерения в процессе ее воздействия на отклик. В случае нелинейной зависимости наличие таких ошибок приводит к смещению классических оценок регрессии. В работе рассмотрены известные методы, направленные на устранение смещения: итерационный взвешенный метод наименьших квадратов, разработанный специально для оценки полиномиальных зависимостей, и метод минимального расстояния. Автором предложен собственный метод, основанный на максимально правдоподобном оценивании с использованием аппроксимации радиальными сплайнами заданной нелинейной функции, описывающей зависимость. Сравнение этого метода с известными подходами в ходе вычислительных экспериментов показало, что он в разы превосходит по точности оценивания метод минимального расстояния. При этом он сопоставим по точности с итерационным взвешенным методом наименьших квадратов, однако обладает тем преимуществом, что применим для оценивания не только полиномов, а любых нелинейных регрессий. Предложенный метод использован в задаче анализа показателей деятельности вузов. Для иллюстрации выбрана зависимость между уровнем безработицы населения и долей трудоустроенных выпускников вузов. Наличие ошибки Берксона объясняется тем, что информация об объясняющей переменной представлена только в среднем по региону, в то время как при воздействии на выпускников вуза имеют место индивидуальные отклонения. Оценка полиномиальной регрессии показала, что при высоком уровне безработицы в регионе пороговое значение показателя трудоустройства недостижимо и должно быть скорректировано.

Ключевые слова: модель с ошибками в переменных, ошибка Берксона, нелинейная регрессия, максимально правдоподобное оценивание, итерационный взвешенный метод наименьших квадратов, метод минимального расстояния, показатель деятельности вузов.

DOI: 10.17212/1727-2769-2016-4-88-98

Введение и постановка проблемы

Часто в научных исследованиях возникает задача восстановления некоторой, в общем случае нелинейной, зависимости между изучаемыми переменными по наблюдаемым данным. Предполагается, что в точности функциональную зависимость наблюдать мы не можем в силу погрешностей измерения ε_i выходной переменной Y_i. Поэтому дело приходится иметь со следующей моделью:

$$Y_i = g(X_i; \theta) + \varepsilon_i, \quad (1)$$

где $g(X_i; \theta)$ – некоторая функция входной переменной X_i, определенная с точностью до вектора неизвестных параметров θ, $i = 1,...,n$, n – число наблюдений. В классической регрессионной постановке значения входной переменной являются, как правило, управляемыми и точно измеренными без погрешностей.

Исследование выполнено при финансовой поддержке Совета по грантам Президента РФ для государственной поддержки молодых российских ученых, проект МК-5385.2016.6.

В отличие от этого модели с ошибками в переменных предполагают наличие некоторых погрешностей в измерении не только отклика, но и входных факторов. В зависимости от того, какого рода эти погрешности, разделяют модели с классической ошибкой и с ошибкой Берксона [1].

Случай с ошибкой Берксона [2] предполагает, что исследователь в активном эксперименте может устанавливать величину входной переменной Z_i, но при воздействии на отклик эта величина искажается из-за случайной погрешности δ_i, следовательно, истинные значения входной переменной определяются как

$$X_i = Z_i + \delta_i . \tag{2}$$

Переменная Z_i часто называется суррогатной переменной (или прокси). Здесь будем считать Z_i детерминированной величиной.

Такие модели находят широкое применение в эпидемиологии [3], где чаще всего встречаются схемы активно-пассивного эксперимента. При выявлении характера зависимости тяжести некоторого заболевания (например, легких) Y_i среди жителей города от степени загрязнения (например, воздуха) фактическое содержание вредных веществ не может быть измерено точно для каждого объекта (индивида), но известно их среднее содержание в некотором регионе (области). Это и будет суррогатная переменная. При этом истинное содержание X_i вредных веществ отклоняется от величины Z_i на некоторую погрешность.

Задача состоит в оценивании вектора неизвестных параметров θ по имеющимся значениям суррогатной переменной Z_i и наблюдаемым в ходе эксперимента реализациям y_i случайных величин Y_i, $i = 1,...,n$, в предположении, что имеет место модель (1)–(2).

В модели дополнительно предполагаются нулевые математические ожидания ошибок (погрешности компенсируются), тогда средние значения суррогатной и истинной переменной совпадают. Кроме того допускается отсутствие корреляции между ε_i и δ_i, т. е. погрешности приборов, измеряющих входные и выходные факторы, не взаимосвязаны. Испытания предполагаются независимыми, в силу чего независимы и ошибки в разных экспериментах. Кроме того предполагается конечная дисперсия ошибок. Обобщим предположения относительно ошибок:

$$\begin{gathered} E(\varepsilon_i) = E(\delta_i) = 0 ,\ D(\varepsilon_i) = \sigma_\varepsilon^2 ,\ D(\delta_i) = \sigma_\delta^2 ,\ \forall i , \\ \operatorname{cov}(\varepsilon_i, \varepsilon_j) = \operatorname{cov}(\delta_i, \delta_j) = 0 ,\ \forall i \neq j ,\ \operatorname{cov}(\varepsilon_i, \delta_j) = 0 ,\ \forall i, j . \end{gathered} \tag{3}$$

И еще одно предположение о нормальности распределения ошибок вводится для упрощения процедуры оценивания параметров и статистических выводов.

Известно, что в линейном случае оценивание модели (1)–(3) можно проводить методом наименьших квадратов (МНК), что не приводит к смещению оценок. В нелинейном случае при использовании МНК возникает систематическое смещение, поэтому предложены специальные методы оценивания.

1. Обзор методов оценивания нелинейной модели с ошибкой Берксона

Методы оценивания нелинейной модели с ошибкой Берксона получили развитие относительно недавно. Обзор подходов можно найти в монографии [3], где предложен аппроксимационный метод, названный калибровкой регрессии (regression calibration). Суть метода заключается в замене ненаблюдаемой переменной X_i ее математическим ожиданием при заданном Z_i. Вычисление такого

математического ожидания требует построения регрессии, что в силу латентности X_i возможно только при наличии дополнительной информации. В качестве такой дополнительной информации может быть использована инструментальная переменная. Непараметрический подход к оцениванию моделей с ошибкой Берксона на основе инструментальных переменных предложен в [4]. Однако инструментальные переменные не всегда доступны и порой их очень сложно подобрать, а сбор дополнительной информации в виде повторных наблюдений приводит к увеличению затрат на исследование. Поэтому здесь сосредоточимся на методах, не требующих привлечения какой-либо внешней информации.

В работе [5] специально для полиномиальных зависимостей предложен итерационный взвешенный метод наименьших квадратов (IRLS, iterative reweighted least squares), использующий два первых условных момента Y_i при заданном Z_i. Показано, что этот метод дает состоятельные оценки, а также доказана принципиальная возможность оценивания как вектора параметров θ, так и неизвестных дисперсий ошибок σ_ε^2, σ_δ^2. Более общий подход, использующий те же идеи, но для любой формы зависимости, описан в [6]. Такой метод оценивания назван методом минимального расстояния (MDE, minimum distance estimator). Для оценивания моделей с ошибкой Берксона в общем случае при любом заданном распределении ошибок в [7] предлагается использовать имитационный подход для упрощения расчета интегралов. При нормальном распределении ошибок для моментов легко получить аналитические выражения.

К сожалению, в работах [6, 7] не приведены результаты вычислительных экспериментов, но с учетом того, что MDE использует информацию только о моментах, можно предположить, что, как и оценки метода моментов, MDE-оценки будут обладать большой дисперсией. В этой связи представляется более перспективным использование метода максимального правдоподобия (ММП), потенциально позволяющего получить оценки с меньшей дисперсией. На его основе автором разработан новый подход к оцениванию модели с ошибкой Берксона.

2. ММП-оценки на основе аппроксимации радиальными сплайнами

В рамках предлагаемого подхода будем исходить из нормальности распределения ошибок. Значения входного фактора в рассматриваемой постановке детерминированы, следовательно, логарифмическая функция правдоподобия будет зависеть только от распределения наблюдаемых значений отклика:

$$\ln L = \sum_{i=1}^{n} \ln f_{Y_i}(y_i),$$

где $f_{Y_i}(y_i)$ – значение функции плотности случайной величины Y_i в точке y_i.

Построение функции плотности распределения Y_i для произвольной функции g является достаточно трудной задачей. Поэтому здесь предлагается осуществлять аппроксимацию функции $g(x; \theta)$ с помощью сплайна. Для этой цели выбраны сплайны с линейными радиальными базисными функциями [8]. С их помощью любую гладкую функцию можно задать как

$$g(x;\theta) \approx R\left(x;\, \beta,\, q_1,\, q_2,\, \ldots,\, q_K\right) = \beta_0 + \beta_1 x + \sum_{j=1}^{K} \beta_{j+1} \left\| x - q_j \right\|,$$

где $R(x;\, \beta,\, q_1,\, q_2, \ldots,\, q_K)$ – линейный радиальный сплайн, $\beta = (\beta_0, \beta_1, ..., \beta_{K+1})$ – вектор коэффициентов сплайна, $q_1 < q_2 < \ldots < q_K$ – узловые точки, принимающие

значения из области значений x, K – число узлов, $\|\cdot\|$ – норма. Здесь в качестве нормы будет рассматриваться метрика L_1, тем самым сплайн будет представлять собой кусочно-линейную аппроксимацию. Обозначим угловые коэффициенты линий на участках $q_j \le x < q_{j+1}$ как k_j, $j = 0,\ldots,K$. В данном контексте q_0 и q_{K+1} доопределяются $x_{\min}$ и $x_{\max}$ соответственно. Введенные угловые коэффициенты рассчитываются из условия стыковки аппроксимирующей кривой и исходной гладкой кривой в точках узлов:

$$k_j = \frac{g(q_{j+1};\,\theta) - g(q_j;\,\theta)}{q_{j+1} - q_j}.$$

На основе этого коэффициенты сплайна вычисляются следующим образом:

$$\beta_0 = \frac{g(q_0;\,\theta) - k_0 q_0 + g(q_K;\,\theta) - k_K q_K}{2},$$

$$\beta_1 = \frac{k_0 + k_K}{2}, \quad \beta_{j+1} = \frac{k_j - k_{j-1}}{2}.$$

Следовательно, искомый вектор параметров θ и узловые точки определяют вектор коэффициентов сплайна β. С учетом аппроксимации сплайнами

$$Y_i \approx R\left(Z_i + \delta_i;\beta, q_1, q_2, \ldots, q_K\right) + \varepsilon_i,\ i = 1,\ldots,n,$$

распределение Y_i представляет собой свертку распределений ошибки отклика и нелинейной функции от ошибки регрессора.

Для удобства представим Y_i как сумму двух случайных величин $Y_i = \omega_i + e_i$:

$$\omega_i = \beta_1(Z_i + \delta_i) + \sum_{j=1}^{K} \beta_{j+1} \left| Z_i + \delta_i - q_j \right|,\ e_i = \beta_0 + \varepsilon_i.$$

Исходя из введенных ранее предположений, $e_1,\ldots,e_n$ – независимые одинаково распределенные случайные величины, имеющие нормальное распределение с вектором параметров $\psi_0 = (\beta_0, \sigma_\varepsilon)$.

Распределение ω_i будет зависеть от расположения $(Z_i + \delta_i)$ относительно узловых точек. Для того чтобы учесть этот факт, введем ряд гипотез, заключающихся в попадании δ_i в заданный интервал. Разобьем область значений δ_i на $K+1$ непересекающихся интервалов. Для удобства здесь доопределим узлы $q_0 = -\infty$, $q_{K+1} = +\infty$. Тогда вероятность справедливости гипотезы $H_{ij} = \{q_j - Z_i < \delta_i < q_{j+1} - Z_i\}$ будет вычисляться следующим образом:

$$P(H_{ij}) = \Phi(q_{j+1} - Z_i;\,\sigma_\delta) - \Phi(q_j - Z_i;\,\sigma_\delta),\ j = 0,\ldots,K,$$

где $\Phi(t;\,\sigma)$ – функция нормального распределения с нулевым параметром сдвига и параметром масштаба σ.

Искомая функция плотности выражается по формуле полной вероятности:

$$f_{Y_i}(u) = \sum_{j=0}^{K} f_{Y_i|H_{ij}}(u) P(H_{ij}). \qquad (4)$$

Условная плотность $f_{Y_i|H_{ij}}$ представляет собой плотность свертки условных распределений $f_{Y_i|H_{ij}}(u) = \left(f_{\omega_i|H_{ij}} * f_{e_i|H_{ij}}\right)(u)$.

Распределение случайной величины e_i не зависит от введенных гипотез. Условная плотность распределения $f_{\omega_i|H_{ij}}(u) = f_{h_{ij}}(u)$ будет определяться поведением случайной величины h_{ij}, которую можно представить как

$$h_{ij} = \beta_1\left(Z_i + \tilde{\delta}_{ij}\right) + \sum_{l=1}^{K}(-1)^{I(l\le j)}\beta_{l+1}\left(q_l - Z_i - \tilde{\delta}_{ij}\right),$$

где $\tilde{\delta}_{ij} = \delta_i \mid H_{ij}$ имеет усеченное нормальное распределение с нулевым параметром сдвига, параметром масштаба, равным σ_δ, и интервалом усечения $q_j - Z_i < \delta_i < q_{j+1} - Z_i$, $I(l \le j)$ – индикаторная функция, возвращающая единицу при $l \le j$ и ноль – в противном случае, $(-1)^0 = 1$. Следовательно, случайная величина h_{ij} имеет усеченное нормальное распределение с вектором параметров $\psi_{ij}^1 = (\mu_{ij}, \sigma_j, a_j, a_{j+1})$, вычисляемых следующим образом:

$$\mu_{ij} = \beta_1 Z_i + \sum_{l=1}^{K}(-1)^{I(l\le j)}\beta_{l+1}(q_l - Z_i),\ j = 0,\ldots,K,$$

$$\sigma_j = \left|\beta_1 - \sum_{l=1}^{K}(-1)^{I(l\le j)}\beta_{l+1}\right|\sigma_\delta,\ a_j = \beta_1 q_j + \sum_{l=1}^{K}\beta_{l+1}\left|q_l - q_j\right|,\ j = 0,\ldots,K+1.$$

Если $\beta_1 - \sum_{l=1}^{K}(-1)^{I(l\le j)}\beta_{l+1} < 0$, то правая и левая границы меняются местами.

Сначала остановимся на самом простом случае, когда $\beta_1 - \sum_{l=1}^{K}(-1)^{I(l\le j)}\beta_{l+1} = 0$ для какого-либо j. Очевидно, что тогда единственной стохастической характеристикой останется ε_i. После упрощения для таких интервалов условная плотность будет равна

$$f_{Y_i|H_{ij}}(u) = \varphi\left(u - \beta_0 - \sum_{l=1}^{K}(-1)^{I(l\le j)}\beta_{l+1}q_l; \sigma_\varepsilon\right),\ \forall i,j : \beta_1 - \sum_{l=1}^{K}(-1)^{I(l\le j)}\beta_{l+1} = 0,$$

где $\varphi(t; \sigma)$ – функция плотности нормального распределения с нулевым параметром сдвига и параметром масштаба σ.

В остальных случаях распределение величины $f_{Y_i|H_{ij}}(u)$ представляет собой свертку усеченного нормального распределения с вектором параметров ψ_{ij}^1 и нормального распределения с параметрами $\psi_0 = (\beta_0, \sigma_\varepsilon)$. По формуле свертки условная плотность Y_i определяется как

$$f_{Y_i|H_{ij}}(u) = \varphi\left(u - \beta_0 - \mu_{ij}; \sqrt{\sigma_j^2 + \sigma_\varepsilon^2}\right) \frac{\Phi_1\left(u;\, \psi^0,\, \psi_{ij}^1\right)}{\Phi\left(a_{j+1} - \mu_{ij};\, \sigma_j\right) - \Phi\left(a_j - \mu_{ij};\, \sigma_j\right)},$$

$$\forall i, j : \beta_1 - \sum_{l=1}^{K} (-1)^{I(l \le j)} \beta_{l+1} \neq 0,$$

где

$$\Phi_1\left(u;\, \psi^0,\, \psi_{ij}^1\right) = \Phi\left((a_{j+1} - u + \beta_0)\sigma_j^2 + (a_{j+1} - \mu_{ij})\sigma_\varepsilon^2;\, \sigma_j \sigma_\varepsilon \sqrt{\sigma_j^2 + \sigma_\varepsilon^2}\right) -$$

$$\Phi\left((a_j - u + \beta_0)\sigma_j^2 + (a_j - \mu_{ij})\sigma_\varepsilon^2;\, \sigma_j \sigma_\varepsilon \sqrt{\sigma_j^2 + \sigma_\varepsilon^2}\right).$$

После упрощения оказывается, что знаменатель последнего множителя в этом выражении совпадает с $P(H_{ij})$, поэтому он сокращается при подстановке в (4). Нужно, однако, учесть, что границы усечения a_j и a_{j+1} меняются местами при $\beta_1 - \sum_{l=1}^{K} (-1)^{I(l \le j)} \beta_{l+1} < 0$, поэтому возьмем числитель последнего множителя по модулю. Тогда итоговое выражение для функции плотности примет вид

$$f_{Y_i}(u) = \sum_{j:\beta_1 - \sum\limits_{l=1}^{K} (-1)^{I(l \le j)} \beta_{l+1} = 0} \varphi\left(u - \beta_0 - \sum_{l=1}^{K} (-1)^{I(l \le j)} \beta_{l+1} q_l;\, \sigma_\varepsilon\right) P(H_{ij}) +$$

$$+ \sum_{j:\beta_1 - \sum\limits_{l=1}^{K} (-1)^{I(l \le j)} \beta_{l+1} \neq 0} \varphi\left(u - \beta_0 - \mu_{ij};\, \sqrt{\sigma_j^2 + \sigma_\varepsilon^2}\right) \left| \Phi_1\left(u;\, \psi^0;\, \psi_{ij}^1\right) \right|.$$

Тем самым логарифмическая функция правдоподобия выражается через вектор коэффициентов сплайна β, при заданных узловых точках однозначно соответствующий вектору неизвестных параметров θ; σ_δ, σ_ε. Путем максимизации функции правдоподобия по неизвестным параметрам получаются ММП-оценки.

Предложенный метод оценивания (далее MLE_{RS}, maximum likelihood estimator based on radial splines) реализован в среде R [9]. Оптимизация осуществлялась методом Нелдера–Мида. В качестве начального приближения вектора θ задавались МНК-оценки. Начальное значение $\sigma_\delta^2 = 0$. Начальное приближение σ_ε^2 определялось как средний квадрат остатков модели, оцененной по МНК. Для сравнения в среде R реализованы описанные выше методы IRLS и MDE.

3. Результаты вычислительных экспериментов

Для исследования работы алгоритмов и сравнения различных методов оценивания проведены вычислительные эксперименты на основе модельного примера из [5]. Предполагается, что $Y_i = 3 + 2X_i + X_i^2 + \varepsilon_i$, $X_i = Z_i + \delta_i$.

Значения Z_i фиксировались во всех экспериментах, $Z_i \sim N(0,1)$. Объем выборок задавался равным 1000. Ошибки ε_i и δ_i моделировались как независимые

нормально распределенные случайные величины с дисперсиями $\sigma_\varepsilon^2 = \sigma_\delta^2 = 0,5$. Значение дисперсии ошибки входного фактора выбрано высоким, поскольку и в практическом приложении приходится сталкиваться с большими погрешностями.

Результаты экспериментов усреднялись по 300 повторениям.

Рассматривалось два варианта расположения узловых точек:

– равновероятно: выбирались как квантили эмпирического распределения Z_i порядка $\frac{j}{K+1}$, $j = 1,\ldots,K$;

– через равные интервалы: узловые точки q_j задавались соотношением

$$\min Z_i + j\frac{\max Z_i - \min Z_i}{K+1},\ j = 1,\ldots,K.$$

Число узлов K выбиралось равным 10 и 30. В таблице приведены средние значения оценок параметров (в скобках их стандартные отклонения), полученные с помощью метода минимального расстояния (MDE), итерационного взвешенного метода наименьших квадратов (IRLS) и предложенного подхода, основанного на методе максимального правдоподобия с аппроксимацией сплайнами (MLE_{RS}). Видно, что метод максимального правдоподобия в отличие от метода минимального расстояния обеспечивает в 2…4 раза меньшее среднеквадратическое отклонение оценок. С помощью этого метода удается хорошо оценить дисперсию ошибок, в то время как MDE и IRLS дают больший разброс, особенно дисперсии ошибки отклика. В целом результаты IRLS сопоставимы по качеству с MLE_{RS}, но этот подход ограничен, так как предназначен только для оценивания полиномов.

Сравнение точности оценивания

The comparison of estimation accuracy

Параметры / Parameters	**MDE**	**IRLS**	**MLE_{RS}**			
			Равновероятные		Равноотстоящие	
			$K = 10$	$K = 30$	$K = 10$	$K = 30$
$\theta_0 = 3$	3,04 (0,152)	3,001 (0,102)	3,000 (0,066)	3,024 (0,067)	2,928 (0,066)	2,990 (0,064)
$\theta_1 = 2$	1,987 (0,15)	1,994 (0,082)	2,018 (0,08)	1,993 (0,079)	2,001 (0,078)	2,002 (0,078)
$\theta_2 = 1$	0,99 (0,204)	1,002 (0,061)	0,912 (0,058)	0,959 (0,065)	1,007 (0,06)	1,006 (0,06)
$\sigma_\delta^2 = 0,5$	0,509 (0,183)	0,497 (0,09)	0,49 (0,042)	0,493 (0,042)	0,496 (0,041)	0,496 (0,041)
$\sigma_\varepsilon^2 = 0,5$	0,789 (0,76)	0,602 (0,524)	0,454 (0,056)	0,496 (0,056)	0,502 (0,057)	0,502 (0,057)
MSE	0,12 (0,217)	0,024 (0,022)	0,04 (0,029)	0,023 (0,021)	0,022 (0,018)	0,018 (0,019)
MAE	0,19 (0,114)	0,108 (0,055)	0,125 (0,044)	0,092 (0,042)	0,105 (0,044)	0,083 (0,04)

Для характеристики точности восстановления регрессионной кривой использовались следующие показатели:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|g(Z_i;\theta) - g\left(Z_i;\hat{\theta}\right)\right|,\ \mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(g\left(Z_i;\theta\right) - g\left(Z_i;\hat{\theta}\right)\right)^2,$$

где $\hat{\theta}$ – вектор оценок параметров, полученный разными методами.

В первую очередь из таблицы видно, что метод максимального правдоподобия дает наилучшие предсказания по сравнению с остальными методами. При этом наилучшая точность достигается при максимальном числе узлов и их расположении через равные интервалы.

Следовательно, исходя из результатов вычислительных экспериментов, можно рекомендовать использовать метод максимального правдоподобия с аппроксимацией радиальными сплайнами с достаточно большим числом узлов (примерно 30 наблюдений на узел), расположенных через равные интервалы.

4. Применение в задаче анализа показателей деятельности вузов

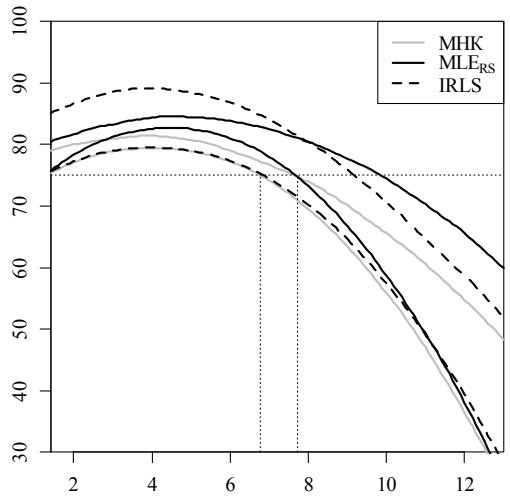
Как отмечено ранее, модель с ошибкой Берксона имеет место, если исследователь располагает только усредненными значениями входного фактора, в то время как в реальности истинные значения случайно от них отклоняются на индивидуальном уровне. Такой постановке соответствует модель, описывающая зависимость показателей эффективности деятельности вузов от характеристик региона. Рассмотрим ее на примере показателя трудоустройства. Во многом его критика связана с тем, что этот показатель больше связан с ситуацией на региональном рынке труда, чем определяется эффективностью работы вуза.

В этой связи возникает задача восстановления зависимости между долей трудоустроенных выпускников и уровнем безработицы на региональном рынке труда. Наличие ошибки Берксона объясняется тем, что уровень безработицы как обусловливающий фактор для выпускников конкретного вуза определяется и их специализацией. Например, в регионе может быть переизбыток одних специалистов и недостаток других, в то время как в распоряжении имеется информация только о среднем уровне безработицы.

В качестве информационной базы использовались данные мониторинга эффективности деятельности образовательных организаций высшего образования за 2015 г., полученные для каждого отдельного вуза [10]. Кроме того взяты официальные данные Росстата [11] об уровне безработицы населения по субъектам РФ в среднем за 2014 г. Из рассмотрения исключены аномальные наблюдения: регионы с уровнем безработицы больше 15 % и вузы с показателем трудоустройства меньше 20 %. Всего в выборке 564 вуза, филиалы вузов не включены в анализ. Следует отметить, что вузы отличаются по широте специализации. Так из 28 направлений по ОКСО в 11 % вузов представлено только два (узкоспециализированные), в 34 % реализуется не более пяти направлений, в 65 % – не более десяти направлений. Поэтому ошибку Берксона не удается нивелировать благодаря многопрофильности вузов.

На основе предварительного анализа зависимости с помощью МНК в качестве g выбран полином второй степени, поскольку квадратичный эффект значим на 10^{-12} уровне и обеспечивает двукратный рост F-статистики по сравнению с линейной моделью. С целью построения интервальных оценок применялось размножение выборки. Для этого на первом шаге из исходной выборки случайно и независимо извлекалась подвыборка объемом 500. На втором шаге производилось оценивание модели (1)–(3) с помощью МНК, IRLS и MLE_{RS}. При использовании MLE_{RS} выбрано 15 равноотстоящих узлов. В результате сохранялся вектор оценок параметров полинома и дисперсий ошибок, полученный каждым методом. Шаги 1–2 повторялись 500 раз, и получена выборка оценок параметров. Что касается дисперсии ошибки входного фактора, то методом IRLS получено среднее значение оценки 4,76 со стандартным отклонением 5,27, а методом MLE_{RS} – в среднем оценка равна 7,32 с отклонением 3,6. При сравнении с дисперсией входной переменной 5,14 становится понятно, что размер погрешности очень велик.

Для каждого выборочного вектора оценок параметров построены прогнозные значения показателя трудоустройства. Далее найдены квантили порядка 2,5 и 97,5 % их эмпирического распределения, они изображены на рисунке.



Зависимость показателя трудоустройства выпускников от уровня безработицы

Dependence of the indicator of graduate employment from the unemployment rate

Видно, что МНК-оценка значительно отличается от других оценок, поскольку дисперсия ошибки Берксона велика. При этом IRLS дает очень широкий доверительный интервал. Тем самым предложенный автором метод MLE_{RS} обеспечивает наиболее пригодный для интерпретации результат. Пороговое значение по показателю трудоустройства для большинства федеральных округов установлено в 75 %. В соответствии с прогнозом МНК и IRLS оно не достигается в регионах с уровнем безработицы, превышающим 6,8 %, однако метод MLE_{RS} дает более высокое значение в 7,7 %. Графически это представлено на рисунке пунктирными линиями. Тем самым методы дают разные результаты с точки зрения их применения на практике, например, в качестве обоснования снижения пороговых значений для определенных регионов. В дальнейшем анализе предполагается учесть специализацию вузов как важный фактор трудоустройства их выпускников.

Заключение

Таким образом, в работе предложен новый метод оценивания нелинейных моделей с ошибкой Берксона, заключающийся в максимизации функции правдоподобия, построенной путем аппроксимации нелинейной функции радиальными сплайнами. Метод основан на предположении о нормальности распределения ошибок. В ходе вычислительных экспериментов произведено сравнение предложенного метода с известными подходами и показано, что он имеет преимущества как в точности оценивания параметров, так и в точности восстановления значений отклика. Хотя итерационный взвешенный метод наименьших квадратов лишь немного хуже, но он ограничивается только оцениванием полиномиальных моделей. Рассмотренные методы применены для решения практической задачи анализа взаимосвязи между долей трудоустроенных выпускников вуза и региональным уровнем безработицы.

ЛИТЕРАТУРА

1. **Fuller W.A.** Measurement error models. – New York: John Wiley and Sons, 1987.
2. **Кендалл М., Стьюарт А.** Статистические выводы и связи. – М.: Наука, 1973. – 899 с.
3. Measurement error in nonlinear models: a modern perspective / R.J. Carroll, D. Ruppert, L.A. Stefanski, C.M. Crainiceanu. – 2nd ed. – New York: Chapman & Hall, 2006.
4. **Schennach S.M.** Regressions with Berkson errors in covariates – a nonparametric approach // The Annals of Statistics. – 2013. – Vol. 41, N 3. – P. 1642–1668.
5. **Huwang L., Huang Y.H.S.** On errors-in-variables in polynomial regression-Berkson case // Statistica Sinica. – 2000. – Vol. 10, N 3. – P. 923–936.
6. **Wang L.** Estimation of nonlinear Berkson-type measurement error models // Statistica Sinica. – 2003. – Vol. 13, N 4. – P. 1201–1210.
7. **Wang L.** Estimation of nonlinear models with Berkson measurement errors // The Annals of Statistics. – 2004. – Vol. 32, N 6. – P. 2559–2579.
8. **Mai-Duy N., Tran-Cong T.** Approximation of function and its derivatives using radial basis function networks // Applied Mathematical Modelling. – 2003. – Vol. 27, N 3. – P. 197–220.
9. The R Project for Statistical Computing [Electronic resource]. – URL: http://www.R-project.org/ (accessed: 29.11.2016).
10. Информационно-аналитические материалы по результатам проведения мониторинга эффективности образовательных организаций высшего образования [Электронный ресурс]. – URL: http://indicators.miccedu.ru/monitoring/2015/ (дата обращения: 24.11.2016).
11. Безработица в России по регионам [Электронный ресурс]. – URL: https://person-agency.ru/statistic-regions.html (дата обращения: 24.11.2016).

MAXIMUM LIKELIHOOD ESTIMATION OF NONLINEAR REGRESSION MODELS WITH BERKSON MEASUREMENT ERRORS

Timofeeva A.Yu.
Novosibisk State Technical University, Novosibirsk, Russia

In experiments designed to identify and estimate the relationship between some variables, measurement errors are unavoidable. Berkson errors distort the values of the explanatory variable after its measurement in the process of its effect on the response. In the case of the nonlinear dependence the presence of these errors leads to a bias in the classical regression estimates. The paper describes the known methods aimed at the bias elimination, namely the iterative reweighted least squares method developed specifically for the estimating of polynomial relationship and the minimum distance estimator. The author suggests her own method based on maximum likelihood estimation using the radial basis function approximation of the given nonlinear function describing the relationship. The comparison of this method with the known approaches in numerical experiments showed that it exceeds several times the estimation accuracy of the minimum distance estimator. Thus it is comparable in accuracy to the iterative weighted least squares method, but it has the advantage that it is applicable to estimate not only polynomials, but any nonlinear regression. The proposed method is applied to the problem of indicator analysis for evaluating the activity of universities. As an illustration a relationship between the unemployment rate and the share of employed graduates is selected. There are Berkson errors because information about the explanatory variable is represented only by a regional average, while individual variations occur in the case of high school graduates. The estimation of the polynomial regression has shown that under a high regional unemployment rate the threshold value of the indicator is unachievable and should be adjusted.

Keywords: Errors-in-variables model; Berkson-type measurement error; nonlinear regression; maximum likelihood estimation; iterative weighted least squares; minimum distance estimato; indicator for evaluating the activity of universities.

DOI: 10.17212/1727-2769-2016-4-88-98

REFERENCES

[1] **Fuller W. A.** *Measurement error models*, New York, John Wiley and Sons, 1987.
[2] **Kendall M., Stuart A.** *The Advanced Theory of Statistics: Inference and relationship*. London, Charles Griffin and Co., Ltd., 1961, 676 p. (Russ. ed.: Kendall M., St'iuart A. *Statisticheskie vyvody i sviazi*. Moscow, Nauka Publ., 1973. 899 p.).
[3] **Carroll R. J., Ruppert D., Stefanski L. A., Crainiceanu C.M.** *Measurement error in nonlinear models: a modern perspective*, New York, Chapman & Hall, 2006.
[4] **Schennach S. M.** Regressions with Berkson errors in covariates – A nonparametric approach. *The Annals of Statistics*, 2013, no. 3, pp. 1642-1668.
[5] **Huwang L., Huang Y. H. S.** On errors-in-variables in polynomial regression-Berkson case. *Statistica Sinica*, 2000, no. 3, pp. 923-936.
[6] **Wang L.** Estimation of nonlinear Berkson-type measurement error models. *Statistica Sinica*, 2003, no. 4, pp. 1201-1210.
[7] **Wang L.** Estimation of nonlinear models with Berkson measurement errors. *Annals of Statistics*, 2004, no. 6, pp. 2559-2579.
[8] **Mai-Duy N., Tran-Cong T.** Approximation of function and its derivatives using radial basis function networks. *Applied Mathematical Modelling*, 2003, no. 3, pp. 197-220.
[9] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, 2013. http://www.R-project.org/.
[10] Informatsionno-analiticheskie materialy po rezul'tatam provedeniia monito-ringa effektivnosti obrazovatel'nykh organizatsii vysshego obrazovaniia (The information-analytical materials on the results of monitoring the effectiveness of the educational institutions of higher education) Available at: http://indicators.miccedu.ru/monitoring/2015/ (accessed 7 September 2016)
[11] Bezrabotitsa v Rossii po regionam (Unemployment in Russia by region) Available at: https://person-agency.ru/statistic-regions.html (accessed 7 September 2016)

СВЕДЕНИЯ ОБ АВТОРАХ

Тимофеева Анастасия Юрьевна – родилась в 1984 году, канд. экон. наук, доцент, кафедра экономической информатики, НГТУ. Область научных интересов: развитие методов статистического анализа объектов стохастической природы, в том числе социально-экономических явлений. Опубликовано 50 научных работ. (Адрес: 630073, Россия, г. Новосибирск, пр. Карла Маркса, д. 20. E-mail: a.timofeeva@corp.nstu.ru).

Timofeeva Anastasiia Yurievna (b. 1984) – Candidate of Sciences (Econ.), associate professor, Department of Computer Science in Economics, Novosibirsk State Technical University. Her research interests are currently focused on the development of methods for the statistical analysis of stochastic objects including socioeconomic phenomena. She is the author of 50 scientific papers. (Address: 20, Karl Marx Av., Novosibirsk, 630073, Russia. E-mail: a.timofeeva@corp.nstu.ru).

Статья поступила 09 сентября 2016 г.
Received September 9, 2016

To Reference:

Timofeeva A.Yu. Maksimal'no pravdopodobnoe otsenivanie nelineinykh regressionnykh modelei s oshibkoi berksona [Maximum-likelihood estimation of nonlinear regression models with berkson measurement errors]. *Doklady Akademii nauk vysshei shkoly Rossiiskoi Federatsii – Proceedings of the Russian higher school Academy of sciences*, 2016, no. 4 (33), pp. 88–98. doi: 10.17212/1727-2769-2016-4-88-98