МИНЕСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

КАФЕДРА МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ И АНАЛИЗА ДАННЫХ

Румянцев
Андрей Кириллович

"Робастные оценки параметров регрессии при наличии группирования выборки"

Курсовая работа

Научный руководитель:
зав. кафедрой ММАД,
канд. физ.-мат. наук
Бодягин Игорь Александрович

Минск, 2018

Содержание

1 Введение

Существует несколько подходов для оценки параметров регрессии, но далеко не все устойчивы к возникновениям аномальных наблюдений, то есть таких наблюдений, которые не подчиняются общей модели. В реальной жизни аномальные наблюдения возникают постоянно, поэтому большинство методов просто неприменимо. В прошлом веке в работах Хьюбера была заложена теория робастного оценивания.

Были предложены следующие робастные оценки[1]:

- M-Оценки
- R-Оценки
- L-Оценки

M-оценки – некоторое подобие оценок максимального правдоподобия (ММП-оценки - частный случай), L-оценки строятся на основе линейных комбинаций порядковых статистик, R-оценки – на основе ранговых статистик.

Будет предложен новый способ оценивания параметров регрессии,где используется группирование выборки, то есть такая модель наблюдений линейной множественной регрессии, когда вместо истинных значений зависимой переменной наблюдаются номера классов (интервалов), в которые попадают эти значения[2]. На практике были полностью реализованы описанные оценки и был произведен анализ оценок.

2 Модель функции регрессии с аномальными наблюдениями и оценки ее параметров

Введем модель линейной регрессию:

$$\begin{aligned} y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i, i = \overline{1, N} \\ y_i = f(x_i, \beta) + \varepsilon_i, \\ f(x_i, \beta) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} \end{aligned} \tag{1}$$

Или, в векторной форме:

$$y_i = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_n \end{pmatrix} \times \begin{pmatrix} 1 \\ x_{i1} \\ \cdots \\ x_{in} \end{pmatrix}^T + \varepsilon_i, \tag{2}$$

где y_i – i-е наблюдение из N наблюдений(N-объем выборки), $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ регрессоры, $\{\beta_k, k = \overline{0, n}\}$– параметры регрессии, а ε_i – случайная ошибка i-го эксперемента, распределение которой подчиняется нормальному закону с нулевым математическим ожиданием и дисперсией σ^2.

В нашей задаче считаем параметры $\{\beta_k, k = \overline{0, n}\}$ неизвестными, их нам и требуется найти.

Но мы будем рассматривать не линейную регрессию, заданную формулами (1)-(2), а линейную регрессию с аномальными наблюдениями вида:

$$y_i^{\widetilde{\varepsilon}} = (\xi_i) y_i + (1 - \xi_i)\eta_i, \tag{3}$$

где ξ_i принимает значение, равное 1, с вероятностью $1 - \widetilde{\varepsilon}$ и значение, равное 0, с вероятностью $\widetilde{\varepsilon}$, т.е.:

$$\begin{cases} p(\xi_i = 0) = \widetilde{\varepsilon}, \\ p(\xi_i = 1) = 1 - \widetilde{\varepsilon}. \end{cases}, \tag{4}$$

которая называется функцией линейной регрессии с выбросами. η_i-случайная величина из некоторого вообще говоря неизвестного распределения. Переменную $\widetilde{\varepsilon}$ будем называть процентом аномальных наблюдений. Величины ξ_i, x_i и η_i являются независимыми

Теперь рассмотрим некоторые методы оценки параметров регрессии:

2.1 Метод наименьших квадратов

Предлоположим, что случайные ошибки подчиняются нормальному закону распределения вероятностей:

$$L\{\varepsilon_i\} = N_1(0, \sigma^2), i = \overline{1, n}. \tag{5}$$

Строим логарифмическую функцию правдоподобия. В силу (1) и (2) имеем:

$$L\{y_i\} = N_1(f(x_i; \beta), \sigma^2). \tag{6}$$

Логарифмическая функция правдоподобия выглядит так[3]:

$$l(\beta) = \ln \prod_{i=1}^{n} (\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - f(x_i;\beta))^2}{2\sigma^2}}) = -\frac{1}{2} n \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} R^2(\beta), \tag{7}$$

$$R^2(\beta) = \sum_{i=1}^{n} (\delta y_i)^2 = \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2 \geq 0. \tag{8}$$

Тогда оценка максимального правдоподобия из формул (4)-(5) такова:

$$\hat{\beta} = arg \min_{\beta} R^2(\beta). \tag{9}$$

2.2 M-оценки

Швейцарский статистик П.Хьюбер предложил использовать М-оценки [3], которые являются решениями экстремальных задач вида:

$$\sum_{i=1}^{n} \phi(x_t; \beta) \to \min_{\beta}, \tag{10}$$

где $\phi(\cdot; \beta)$-некоторая функция, определяющая конкретный тип оценок и их точность.
Очевидно, что $\phi(\cdot; \beta) \equiv -\ln p(\cdot; \beta)$ дает обычную оценку максимального правдоподобия, построенная по модели без выбросов (1).
Рассмотрим теперь некоторые способы выбора $\phi(\cdot; \beta)$.

2.2.1 Способы выбора функции для решения экстремальной задачи в М-оценках

Для начала определим:

$$u_i = y_i^{\widetilde{\varepsilon}} - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}). \tag{11}$$

Тогда существует такие методы[4]:

Способы выбора $\phi(\cdot;\beta)$	
Метод	Целевая функция
Метод Наименьших Квадратов	$\phi(\cdot;\beta)_{OLS} = u^2$
Хьюбера	$\phi(\cdot;\beta)_H = \begin{cases} \frac{1}{2}u^2, \lvert u\rvert \le k, \\ k\lvert u\rvert - \frac{1}{2}k^2, \lvert u\rvert > k \end{cases}$
Биквадратный	$\phi(\cdot;\beta)_B = \begin{cases} \frac{k^2}{6}(1-[1-(\frac{u}{k})^2]^3), \lvert u\rvert \le k \\ \frac{k^2}{6}, \lvert u\rvert > k \end{cases}$

3 Моделирование функции регрессии с аномальными наблюдениями

Для начала смоделируем функцию регрессии по методу (3). Для удобства моделируем регрессию с одномерными регрессорами $x_i, i = \overline{1, N}$.
Воспользуемся такими параметрами:

Параметры программы	
Переменная	значение
Размер выборки N	1000
Доля выбросов $\widetilde{\varepsilon}$	0.1
Параметры регрессии β	$(100, 4)$
Регрессоры x_i	$\sim U(-5, 5)$
ε_i	$\sim N(0, 16)$
η_i	$\sim N(100, 100)$

,

$U(-5, 5)$ - равномерное распределение на отрезке $[-5, 5]$.
Получаем такой график:

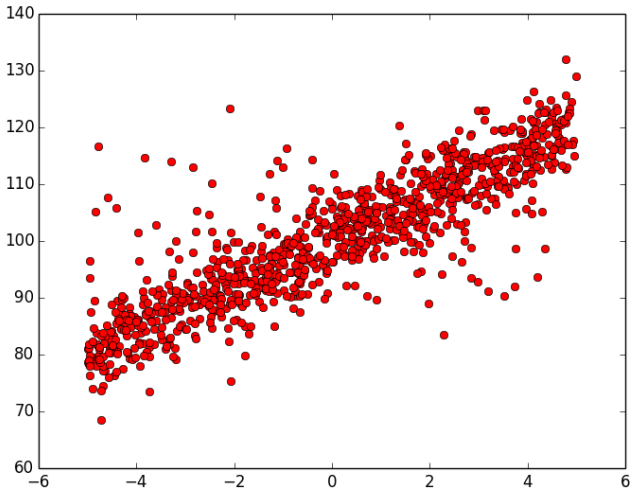

Рис. 1: Вывод графика рассеяния (y_i, x_i)

4 Построение оценки параметров регресии с помощью группирования выборки

Будем работать с моделью регрессии (3), предполагая что имеем регрессию без выбросов (1). Каждый y_i принадлежит нормальному распределению:

$$y_i = f(x_i, \beta) + \varepsilon_i \sim \mathcal{N}(f(x_i, \beta), \sigma^2). \tag{12}$$

Разделим множество значений функции регрессии, т.е множество $\mathcal{R}$, на k полуинтервалов:

$$\mathcal{R} = (-\infty, a_1] \bigcup (a_1, a_2] \bigcup \cdots \bigcup (a_{k-1}, +\infty). \tag{13}$$

Обозначим полученные интервалы: $\nu_0, \ldots, \nu_{k-1}$.
Далее в работе будем считать, что вместо истинных значений зависимых переменных y_i наблюдается только номер класса, к которому это наблюдение попало. Тогда для каждого y_i будем набюдаем лишь номер полуинтервала μ_i, в который он попал.

$$\mu_i = j, \text{если } y_i \text{ отнесли к полуинтервалу } \nu_j. \tag{14}$$

Функцию распределения нормального закона с параметрами μ, σ^2 можно представить как:

$$F(x) = \Phi(\frac{x - \mu}{\sigma}), \tag{15}$$

где $\Phi(x)$ функция распределения стандартного нормального закона, а $\sigma = \sqrt{\sigma^2}$:

$$\Phi(x) = \frac{1}{\sqrt{2}\sigma} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt. \tag{16}$$

Обозначим:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt. \tag{17}$$

Тогда:

$$\Phi(x) = \frac{1}{2}\Big[1 + \operatorname{erf}\Big(\frac{x}{\sqrt{2}}\Big)\Big]. \tag{18}$$

Поэтому:

$$F(x) = \frac{1}{2}\Big[1 + \operatorname{erf}\Big(\frac{x - \mu}{\sqrt{2}\sigma}\Big)\Big]. \tag{19}$$

Тогда при модельных предположениях (12) вероятность попадания y_i в полуинтервал ν_j равна:

$$P\{y_i \in \nu_j\} = F_{y_i}(a_{j+1}) - F_{y_i}(a_j) = \tag{20}$$

$$= \begin{cases} \frac{1}{2}(\mathrm{erf}(\frac{a_{j+1}-f(x_i,\beta)}{\sqrt{2}\sigma}) - \mathrm{erf}(\frac{a_j-f(x_i,\beta)}{\sqrt{2}\sigma})), \ j = \overline{1, k-2} \\ \frac{1}{2}(1 + \mathrm{erf}(\frac{a_1-f(x_i,\beta)}{\sqrt{2}\sigma})), \ j = 0 \\ \frac{1}{2}(1 + \mathrm{erf}(\frac{a_{k-1}-f(x_i,\beta)}{\sqrt{2}\sigma})), \ j = k-1 \end{cases}.$$

Понятно, что:

$$P(\mu_i = j) = P(y_i \in \nu_{\mu_i}). \tag{21}$$

4.1 Построение функции правдоподобия

Составим функцию правдоподобия:

$$l(\beta, \sigma^2, \nu_0, \ldots, \nu_{k-1}) = \ln(\prod_{i=1}^{n} P(\mu_i = j)) = \tag{22}$$

$$= \sum_{i=1}^{n} \ln(P(\mu_i = j)). \tag{23}$$

Известно приближение для функции $\mathrm{erf}(x)$:

$$(\mathrm{erf}x)^2 \approx 1 - \exp(-x^2 \frac{\frac{4}{\pi} + ax^2}{1 + ax^2}), \tag{24}$$

$$a = \frac{8}{3\pi}\frac{3-\pi}{\pi-4}. \tag{25}$$

Оно считается достаточно точным для x близких к 0 и к ∞ [7].
Найдем производную для этого приближения:

$$\mathrm{erf}'(x) = \exp(-x^2 \frac{\frac{4}{\pi} + ax^2}{1 + ax^2}) \frac{-2x\frac{\frac{4}{\pi}+ax^2}{1+ax^2} + (2ax^3)\frac{\frac{4}{\pi}+ax^2}{1+ax^2} - \frac{2ax^3}{1+ax^2}}{2\sqrt{1 - \exp(-x^2\frac{\frac{4}{\pi}+ax^2}{1+ax^2})}}. \tag{26}$$

Будем максимизировать функцию l. Для этого будем искать нули ее производной. Вычисление будем производить с помощью вычислительных методов (будем использовать метод секущих), так как из-за сложного вида

производной вычислить ее аналитически не представляется возможным.

$$\frac{\delta l}{\delta \beta} = \frac{\delta \sum_{i=1}^{n} \ln(P(\mu_i = j))}{\delta \beta} = \frac{\delta \sum_{i=1}^{n} \ln P(y_i \in \nu_{\mu_i})}{\delta \beta} = \quad (27)$$

$$= \frac{\delta \sum_{i=1}^{n} \ln(\frac{1}{2}(\mathrm{erf}(\frac{a_{\mu_i+1} - f(x_i,\beta)}{\sqrt{2}\sigma}) - \mathrm{erf}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma})))}{\delta \beta} = \quad (28)$$

$$= \sum_{i=1}^{n} \Big((1 - (\delta_{\mu_i 0} + \delta_{\mu_i k-1})) \frac{(\mathrm{erf'}(\frac{a_{\mu_i+1} - f(x_i,\beta)}{\sqrt{2}\sigma}) - \mathrm{erf'}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma}))}{(\mathrm{erf}(\frac{a_{\mu_i+1} - f(x_i,\beta)}{\sqrt{2}\sigma}) - \mathrm{erf}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma}))} + \quad (29)$$

$$+ (\delta_{\mu_i 0} + \delta_{\mu_i k-1}) \frac{\mathrm{erf'}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma})}{(1 + \mathrm{erf}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma}))} \Big) (-1) \frac{\delta f(x_i,\beta)}{\delta \beta}) =$$

$$= -\sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_{i1} \\ \dots \\ x_{in} \end{pmatrix} \times \Big((1 - (\delta_{\mu_i 0} + \delta_{\mu_i k-1})) \frac{(\mathrm{erf'}(\frac{a_{\mu_i+1} - f(x_i,\beta)}{\sqrt{2}\sigma}) - \mathrm{erf'}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma}))}{(\mathrm{erf}(\frac{a_{\mu_i+1} - f(x_i,\beta)}{\sqrt{2}\sigma}) - \mathrm{erf}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma}))} + \quad (30)$$

$$+ (\delta_{\mu_i 0} + \delta_{\mu_i k-1}) \frac{\mathrm{erf'}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma})}{(1 + \mathrm{erf}(\frac{a_{\mu_i} - f(x_i,\beta)}{\sqrt{2}\sigma}))} \Big),$$

где δ_{ij} - символ Кронекера.
Доказано, что максимизируя функцию правдоподобия (22), можем получить состоятельную оценку[2] параметров.
Итак, выражение (27) и будем использовать для метода дихотомии, приближая $\mathrm{erf}'(x)$ с помощью выражения (26).

4.2 Метод секущих

Так как мы не можем привести систему $\frac{\delta l}{\delta \beta} = 0$ к виду, удобному для итерации, то нам придется искать ее нули с помощью метода Ньютона. Введем вектор ошибки $\check{\varepsilon}^{(k)} = \beta^* - \beta^{(k)}$. Тогда для его определения имеем:

$$\frac{\delta l(\beta^{(k)} + \check{\varepsilon}^{(k)})}{\delta \beta} = 0. \quad (31)$$

Строя разложение левой части по формуле Тейлора и ограничиваясь лишь линейными членами[8], будем иметь систему:

$$\frac{\delta}{\delta \beta} \frac{\delta l(\beta^{(k)}}{\delta \beta} \Delta \beta^{(k)} = -\frac{\delta l(\beta^{(k)})}{\delta \beta}. \quad (32)$$

Если матрица $\frac{\delta}{\delta \beta} \frac{\delta l(\beta^{(k)}}{\delta \beta}$ невырожденная (а в нашем случае она диагональная), то из этой системы можно единственным образом найти $\Delta \beta^{(k)}$ и построить

приближение:

$$\beta^{(k+1)} = \beta^{(k)} + \Delta\beta^{(k)}. \tag{33}$$

Так как для второй производной l получится довольно сложное выражение, то будем приближать ее с помощью выражения:

$$\frac{\delta}{\delta\beta_j}\frac{\delta l(\beta_1^{(k)},\ldots,\beta_n^{(k)})}{\delta\beta} \approx \frac{\frac{\delta l(\beta_1^{(k)},\ldots,\beta_j^{(k)},\ldots,\beta_n^{(k)})(\beta^{(k)}}{\delta\beta} - \frac{\delta l(\beta_1^{(k)},\ldots,\beta_j^{(k-1)},\ldots,\beta_n^{(k)})(\beta^{(k)}}{\delta\beta}}{\beta_j^{(k)} - \beta_j^{(k-1)}}. \tag{34}$$

Теперь имеем нули производной функции l, а также ее значения на границе отрезка $[a, b]$. Переберем эти значения и таким образом найдем значение вектора $\hat{\beta}$, где она достигает своего максимального значения.

4.3 Переклассификация выборки

На данном этапе для каждого x_i имеем класс μ_i: т.е. пару (x_i, μ_i). Теперь попытаемся переклассифицировать выборку. Для этом будем строить новую выборку такого же объема N. Будем идти по каждому элементу (x_i, μ_i) выборки и для этого наблюдения построим новое:

$$(x_i, \check{\mu}_i), \tag{35}$$

где $\check{\mu}_i$ максимально встречающийся класс близлежайших соседей:

$$\check{\mu}_i = \arg\max_j \sum_{|x_k - x_i| \le \Delta,\ k \ne i} \delta_{\check{\mu}_k j}, \tag{36}$$

где Δ параметр, задающий уровень близости. Чем он выше, тем больше используется соседей для коррекции класса нашего наблюдения.

Итак, переклассифицировав выборку, применим к ней функцию правдоподобия из уравнений (22-23), только используя теперь новые классы $\check{\mu}_i$ вместо μ_i. Аналогично пунктам 4.1-4.2 максимизируем ее и найдем новую оценку параметров $\hat{\beta}$.

5 Реализация оценок на практике

На практике построение оценок является нетривиальной задачей, так как метод секущих имеет свои недостатки. Для метода секущих необходимо, чтобы корни уравнения были отделены, но не существует способа отделения корней в общем случае. Поэтому для решения уравнения нам нужны дополнительные параметры:

- границы отрезка, на котором находятся предполагаемая оценка $\hat{\beta}$,
- расстояние по норме между двумя начальными приближениями $\hat{\beta}^{(0)}, \hat{\beta}^{(1)}$,
- шаг для каждой переменной $\hat{\beta}_i^{(k)}$.

Тогда решая методом секущих уравнение

$$\frac{\delta l(\beta)}{\delta \beta} = 0. \tag{37}$$

на каждом из отрезков и найдя среди полученных приближений $\hat{\beta}$ то, на котором функция правдоподобия (22-23) достигает максимума найдем решение.

6 Заключение

Были описаны некоторые методы робастного оценивания параметров регрессии. Был предложен еще один способ оценивания параметров регрессии для модели регрессии с аномальными наблюдениями при наличии группирования выборки. Оценки из пункта 4 были реализованы на практике. Построенный метод был исследован на точность.

Список литературы

[1] Хьюбер Дж П., *Робастность в статистике:пер. с англ.*. М.:Мир,1984-304с

[2] Е. С Агеева, чл.-корр. НАН Беларуси Ю.С. Харин, *Состоятельность оценки максимального правдопобия параметров множественной регрессии по классифицированным наблюдениям*

[3] Харин Ю.С., Зуев Н.М., Жук Е.Е, *Теория вероятностей, математическая и прикладная статистика: учебник* Минск: БГУ, 2011.-463с

[4] John Fox & Sanford Weisberg, *Robust Regression*, October 8, 2013

[5] А.В. Омельченко, *Робастное оценивание параметров полиномиальной регрессии второго порядка,* Харьковский национальный университет радиоэлектроники, Украина, 2009

[6] Özlem Gürünlü Alma, *Comparison of Robust Regression Methods in Linear Regression*, Int. J. Contemp. Math. Sciences, Vol. 6, 2011, no. 9, 409 - 421

[7] Sergei Winitzki, *A handy approximation for the error function and its inverse*

[8] Мандрик П.А., Репников В.И., Фалейчик Б.В., *Численные методы*