О РОБАСТНОМ ОЦЕНИВАНИИ В ЭКОНОМЕТРИЧЕСКОМ МОДЕЛИРОВАНИИ

Стебунова О.И.
Оренбургский государственный университет, г. Оренбург

В настоящее время существенно расширилось использование эконометрических методов в исследованиях социально-экономических процессов и явлений. Кроме того, реализация образовательных стандартов подготовки магистров в области экономики и управления требует применения продвинутого математического аппарата. Так, при построении классической регрессионной модели предполагается выполнение ряд предположений, например, такие как независимость наблюдений, постоянство дисперсии регрессионных остатков, которые должны иметь нормальное распределение. Если все эти предпосылки выполняются, то полученные оценки коэффициентов регрессионной модели обладают свойствами несмещенности, состоятельности и эффективности. Даже при средних размерах объемов выборки для распределений нормального, пуассоновского и гамма-распределений можно использовать асимптотически нормальные приближения. Однако даже несколько резко выделяющихся наблюдений могут изменить значения оценок статистических характеристик исходной совокупности. Как известно, наличие выбросов в статистических данных, во-первых, искажает структуру совокупности, во-вторых, вносит смещение в интегральные параметры статистического оценивания. Причины появления выбросов могут быть различными: специфика объекта, случайный разброс, неправильное причисление данных к совокупности, ошибки при регистрации и обработке исходной информации. Поэтому первым шагом эконометрического моделирования является диагностика или выявление грубых ошибок в статистической совокупности, осуществляемое, например, с помощью T - критерия Смирнова-Граббса, критерия Титьена-Мура, T - статистика Хоттелинга и других [1, 2]. Однако в многомерном случае удаление объекта из исследуемой совокупности из-за ошибки для одного признака зачастую неприемлемо, так как сокращается выборкам, а так же может быть утеряна значимая информация по другим признакам. Поэтому на практике для получения устойчивых интегральных статистических характеристик применяются методы оценивания по усеченной совокупности данных, в которой устранены грубые ошибки: подходы Пуанкаре и Винзора [2].

Для определения параметров регрессионной модели без анализа резковыделяющихся наблюдений, рекомендуется использовать робастные методы, нечувствительные к выбросам и зашумленности (загрязнениям) распределения регрессионных остатков. В литературе существуют различные непараметрические методы построения оценок коэффициентов регрессионной модели, в статье рассмотрены M - оценки, которые разработаны Хьюбертом [3].

Пусть на основе предварительного анализа установлено, что эндогенная переменная (результативный признак) y зависит от предопределенных

(объясняющих переменных) $x_1, x_2, ..., x_k$. Результаты наблюдений результирующего признака и объясняющих переменных представлены вектором $Y_{n\times 1} = (y_1 \quad y_2 \quad ... \quad y_n)^T$ и матрицей X типа «объект-свойство»:

$$X_{n\times k} = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1k} \\ x_{21} & x_{22} & ... & x_{2k} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{nk} \end{pmatrix}$$

где yi – наблюденное значение результативного признака для *i*-го объекта;

x_{ij} – значение *j*-го признака на *i*-м объекте наблюдения $i = \overline{1,n}$, $j = \overline{1,k}$; столбец из "1" можно считать столбцом "наблюденных" значений для признака $x_0 = 1$.

Для моделирования взаимосвязи между результативным признаком и объясняющими переменными предлагается использовать линейную регрессионную модель вида (1):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + z_i, \quad i = \overline{1,n}, \tag{1}$$

где $\beta = (\beta_0 \; \beta_1 ... \beta_k)^T$ - вектор коэффициентов регрессионной модели;

z_i - регрессионный остаток, характеризующий влияние неучтенных факторов на величину результативного признака для *i*–го объекта.

Как известно, для оценки коэффициентов регрессионной модели применяется метод наименьших квадратов, суть которого в минимизации суммы квадратов регрессионных остатков:

$$F(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = (Y - X\beta)^T (Y - X\beta) \to \min. \tag{2}$$

Относительно неизвестных коэффициентов имеется квадратичный функционал, для нахождения минимума воспользуемся необходимым условием существования экстремума. В итоге получаем систему уравнений $\nabla F(\beta_0, \beta_1, ..., \beta_K) = 2X^T X\beta - 2X^T Y = 0$. Апостериорная оценка коэффициентов линейной модели множественной регрессии может быть представлена в виде: $\hat{\beta}_{МНК} = (X^T X)^{-1} X^T Y$. Однако резко выделяющиеся наблюдения нарушают основные предположения МНК о том, что регрессионные остатки являются независимыми одинаково распределенными случайными величинами. Поэтому

применение стандартного подхода к оценки коэффициентов может привести к большим ошибкам, в результате которых модель может не иметь смысла.

Простой способ построения робастной регрессии состоит в применении метода наименьших квадратов к цензурированной выборке. Для этого из статистической совокупности исключается некоторая доля объектов, имеющих слишком большие значения регрессионных остатков ε_i. Итерации продолжаются до тех пор, пока удается выделять объекты с большими значениями регрессионных остатков. Максимальная доля отсеиваемых объектов задается исходя из содержания задачи. Например, если выбросы действительно обусловлены грубыми ошибками измерений, то на гистограмме распределения регрессионных остатков соответствующие точки легко отделяются с помощью статистических критериев.

Еще один способ получения апостериорных оценок коэффициентов регрессионной модели заключается в том, чтобы вместо квадратичного функционала (2) ввести ограниченную сверху функцию ρ, которая в окрестности нуля ведет себя как квадратичная, а на бесконечности стремится к горизонтальной асимптоте [3]:

$$Q(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} \rho\left(y_i - \sum_{j=1}^{k} \beta_j x_{ij} \right) \to \min \ . \qquad (3)$$

где ρ - некоторая выпуклая функция.

Например, если использовать в качестве ρ экспоненциальную функцию $\rho(b) = -\exp\left(-\frac{\lambda \cdot b^2}{2} \right)$, предложенную Мешалкиным, то получаем [3]:

$$Q(\beta_0, \beta_1, ..., \beta_k) = -\exp\left[-\frac{\lambda \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2}{2} \right] \to \min \ . \qquad (4)$$

Дифференцируя (4) по β, получаем систему уравнений (5):

$$\frac{\partial Q}{\partial \beta_l} = \lambda \cdot \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} \beta_j \cdot x_{ij} \right) (-x_{il}) \cdot \exp\left[-\frac{\lambda \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2}{2} \right] = 0, \qquad (5)$$

Как видно, задача минимизации функционала $Q(\beta_0, \beta_1, ..., \beta_k)$ уже не может быть решена средствами линейной алгебры, и приходится применять численные методы оптимизации.

Полученные апостериорные оценки коэффициентов для некоторой функции ρ называются M - *оценкой*. Открытым остается вопрос о выборе функции ρ, в литературе наряду с экспоненциальным семейством функций, предлагается использовать функцию $\rho(b) = |b|$ и $\rho(b) = (b)^{\nu}$, *где* $1 < \nu < 2$, предложенную Форсайт [2, 3].

Более полное описание существующих методов робастной регрессии (метод модифицированных остатков, метод модифицированных весов, метод псевдонаблюдений) приведено в литературе [2, 3]. Например моделирования спроса на рабочую силу производственных предприятий [4], продемонстрируем влияние наличия выбросов на оценки коэффициентов регрессионной модели.

Как видно, из таблицы 1, двенадцатое наблюдение для этих статистических данных является аномально большим. Оценка линейной регрессионной модели, полученная на основе статистической совокупности без аномального наблюдения, имеет вид (6):

$$\hat{y} = \underset{(19,64)}{287,72} - \underset{(12,43)}{167,72}\, x_1 + \underset{(0,009)}{0,382}\, x_2 - \underset{(0,007)}{0,114}\, x_3, \quad R^2 = 0,953, \quad F = 2716,02 \quad , (6)$$

где y - численность рабочих, чел.;

x_1 - суммарные расходы на заработную плату, млн. руб.;

x_2 - объем производства, млн. руб.;

x_3 - стоимость основных фондов, млн. руб.

Таблица 1 – Фрагмент статистической совокупности данных о спросе на рабочую силу

n/n	Численность рабочих, чел.	Суммарные расходы на заработную плату, млн. руб.	Объем производства, млн. руб.	Стоимость основных фондов, млн. руб.
1	355	6962,00	10296,08	1751
2	263	5908,71	1478,16	2910
3	26	4379,21	377,53	1357
5	141	5158,03	642,03	969
…	173	4529,49	341,69	1643
12	**1468**	**8012,55**	**82540,63**	**2223**
…	……	……	……	……..
109	67	3813,11	578,9	1829
110	90	5260,96	1385,21	1622

Полученная модель (6) значима, значимы все коэффициенты. Все оценки коэффициентов имеют ожидаемый знак: более высокая заработная плата при прочих равных условиях приводит к снижению численности работников, в то время как больший объем производства требует большого количества труда.

Оценка регрессионной модели (7) по данным с аномальным наблюдением свидетельствует о том, что МНК-оценка коэффициентов даже при одном аномальном наблюдении не имеет никакого смысла не только по своим значениям, но и по знакам этих компонент.

$$\hat{y}' = \underset{(9,82)}{1077,13} + \underset{(115,49)}{1,345}\,x_1 - \underset{(0,489)}{116,8}\,x_2 + \underset{(0,145)}{0,987}\,x_3 . \qquad (7)$$

$$\hat{y}^{PO} = \underset{(19,64)}{297,52} - \underset{(12,43)}{167,72}\,x_1 + \underset{(0,009)}{0,382}\,x_2 - \underset{(0,007)}{0,114}\,x_3 , \; R^2 = 0,953, \; F = 2716,02. \qquad (8)$$

Используя метод робастной регрессии (количество итераций равно 6), построена оценка параметров линейной регрессионной модели (8), которая близка к оценке без аномального наблюдения, что отражает целесообразность использования робастных методов оценивания в эконометрическом моделировании социально-экономических процессов.

Список литературы

1 Математическое моделирование: исследование социальных, экономических и экологических процессов (региональный аспект) [Электронный ресурс]: учебное пособие / О. И. Бантикова, В.И. Васянина, Ю.А. Жемчужникова, А.Г. Реннер, Е.Н. Седова, О.И. Стебунова, Л.М. Туктамышева, О.С Чудинова; под ред. А.Г. Реннера. - ОГУ, 2012. – Режим доступа: http://rucont.ru/efd/202382?cldren=02.

2 ***Большаков А.А.,*** *Методы обработки многомерных данных и временных рядов: учебное пособие для вузов / А.А. Большаков, Р.Н. Каримов. – М.: Горячая линия – Телеком, 2007. – 522с.*

3 ***Шурыгин А.М.*** *Прикладная стохастика: робастность, оценивание, прогноз / А.М. Шурыгин – М.: Финансы и статистика, 2000 – 224 с.*

4. Математические методы моделирования социально-экономических процессов (региональный аспект) [Текст] / А. Г. Реннер, О.И. Бантикова, О.С. Бравичева, О.И. Стебунова, Л.М. Туктамышева. - Самара: Изд-во СамНЦ РАН, 2008. – 182с.