

© Сергей Владимирович Поляков

*канд. техн. наук, доцент кафедры информационных технологий,
Владимирского филиала РАНХиГС*

ВЫЯВЛЕНИЕ ГРУБЫХ ОШИБОК В ИСХОДНЫХ ДАННЫХ ИССЛЕДУЕМОГО ПРОЦЕССА

Detection of gross errors in the source data of the process under study

Аннотация: приведены источники грубых ошибок. Рассмотрены статистические методы выявления грубых ошибок. Даны примеры.

Abstract: The sources of gross errors. Reviewed statistical methods to identify gross errors. The examples.

Ключевые слова: грубая ошибка, источники, методика оценки.

Keywords: gross error, sources, methods of assessment.

Построение математических моделей исследуемых процессов, например, в социально-экономической сфере, связано с получением и последующей математической обработкой полученных данных или наблюдений. Под наблюдением надо понимать как сам результат, так и его регистрацию. Предположим, что проводится серия однородных испытаний. Наблюдения по этим испытаниям будут различаться из-за наличия случайных факторов, следовательно, каждый истинный результат в реальных условиях превращается в случайную величину. Даже при использовании очень точных методов и наилучших условий получения экспериментальных данных эти данные могут содержать ошибки (погрешности), характер и причины которых могут быть различными. Поэтому анализ и оценка исходных данных, используемых для построения математической модели, является важной и обязательной задачей.

Различают ошибки трех видов [1].

1. Систематическая ошибка, т.е. ошибка, повторяющаяся и одинаковая во всей серии наблюдений. Эта ошибка связана обычно с неправильным ведением эксперимента.

2. Грубая ошибка, т.е. ошибка, связанная с резким нарушением условий испытания при отдельном наблюдении. Если систематическая ошибка характеризуется в первую очередь своей неизменностью во всей серии испытаний, то грубая ошибка присутствует обычно не более чем в одном-двух испытаниях и характерна именно своим отличием по величине от прочих рядовых ошибок.

3. Случайная ошибка, включающая все остальные виды ошибок.

Рассмотрим более подробно способы выявления грубых ошибок в процессе оценки связи факторов исследуемого процесса.

На первом этапе анализ массива исходных данных, описывающих процесс исследуемой предметной области, заключается в выявлении грубых ошибок (промахов, выбросов). Грубая ошибка (промах, выброс) – это ошибка результата отдельного наблюдения, входящего в массив, которая для данных условий резко отличается от остальных наблюдений этого массива. Реже к такому результату наблюдения применяется термин «аномальное значение».

Можно выделить следующие источники грубой ошибки.

1. Ошибка оператора (неправильная запись результата наблюдения). Данные, включенные в массив как неумышленная ошибка оператора, можно считать грубой ошибкой.

2. Ошибка, связанная с условиями измерений (например, резкие изменения условий получения результата измерения). В этом случае такие данные массива надо называть промахом.

3. Умышленное искажение показаний наблюдений с определенной целью. Такие данные массива нельзя называть грубой ошибкой или промахом, т.к. они внесены в массив умышленно; их надо считать выбросом.

4. Резкие отличия показаний различных объектов исследования. Например, показания ряда однотипных предприятий, одно из которых использует устаревшую технологию, в связи с чем его данные будут резко отличаться от данных других предприятий, входящих в массив. В этом случае такие показания надо считать выбросом, а не грубой ошибкой или промахом. При этом целесообразно выявлять причины таких отклонений.

Наличие грубой ошибки, промаха или выброса (в дальнейшем, для краткости, будем использовать один термин – выброс) в исходном массиве данных может в дальнейшем резко исказить математическую модель, поэтому рекомендуется любую совокупность наблюдений, исходных данных проверять на наличие выбросов. Проверка исходных данных на наличие выбросов проводится с помощью статистических критериев.

Для одномерного массива применяются следующие статистические критерии [2] оценки выбросов: критерий Диксона, критерий Шовине, критерий Романовского, критерий Граббса, критерий трех сигм, критерий Ирвина. Эти критерии используются для оценки выброса в условиях различного количества наблюдений, начиная от $n < 10$ (критерий Диксона, критерий Шовине) до $n > 20...50$ (критерий трех сигм, критерий Ирвина).

Оценка отсутствия или наличия выброса заключается в том, что выдвигается гипотеза H_0 – выброса нет. Далее, пользуясь определенным статистическим критерием, опровергают выдвинутую гипотезу H_0 или подтверждают по отношению к проверяемому наблюдению x_i . Если гипотеза H_0 отвергается, то проверяемое наблюдение x_i считается выбросом и его желательно исключить из дальнейшего исследования, связанного с использованием исходного массива данных.

Однако перечисленные критерии позволяют выявлять выбросы в одномерном массиве данных, заданном рядом значений $x_1, x_2, \dots, x_n$. И проверке подвергается наибольшее x_{max} и/или наименьшее x_{min} наблюдение из заданного ряда. Если в качестве исходных данных имеем два взаимосвязанных массива, то предполагаемый выброс или грубую ошибку можно заметить на диаграмме рассеяния, как это показано на рисунке 1. Но если применить для оценки предполагаемого выброса один из рассмотренных выше и соответствующих данному случаю статистических критериев, то эта оценка подтвердит гипотезу H_0 о том, что выброса нет.

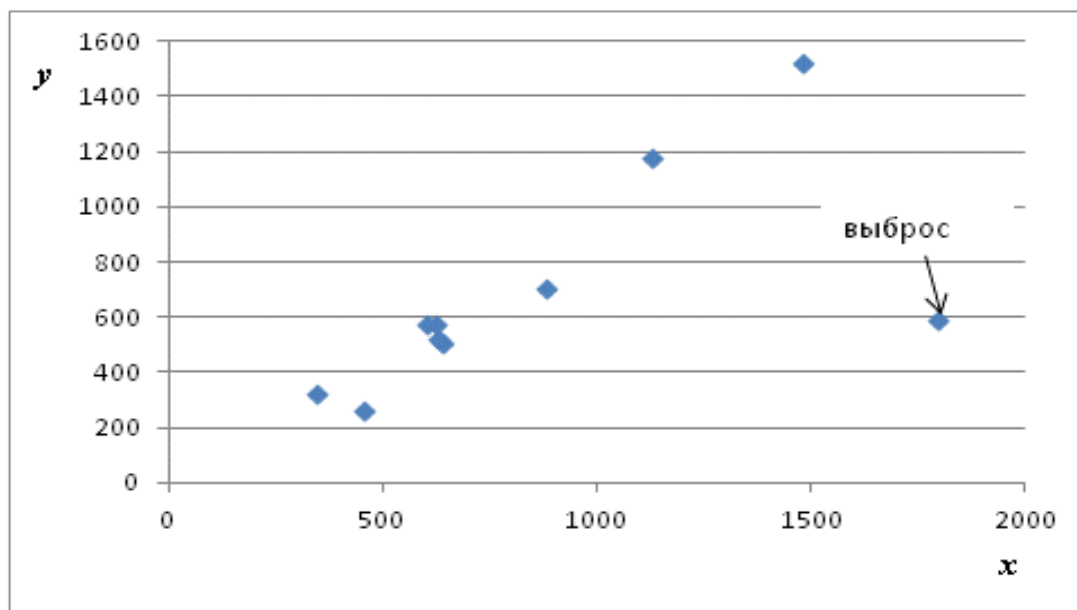


Рис. 1. Диаграмма рассеяния исходного массива данных $y(x)$ с резким отклонением одной из координат от общей закономерности

На рисунке 2 дан пример другой диаграммы двух взаимосвязанных массивов с предполагаемым выбросом, гипотезу о наличии которого необходимо проверить. В этом случае отклонение предполагаемого выброса от общей закономерности связи x и y менее очевидна, чем в предыдущем примере.

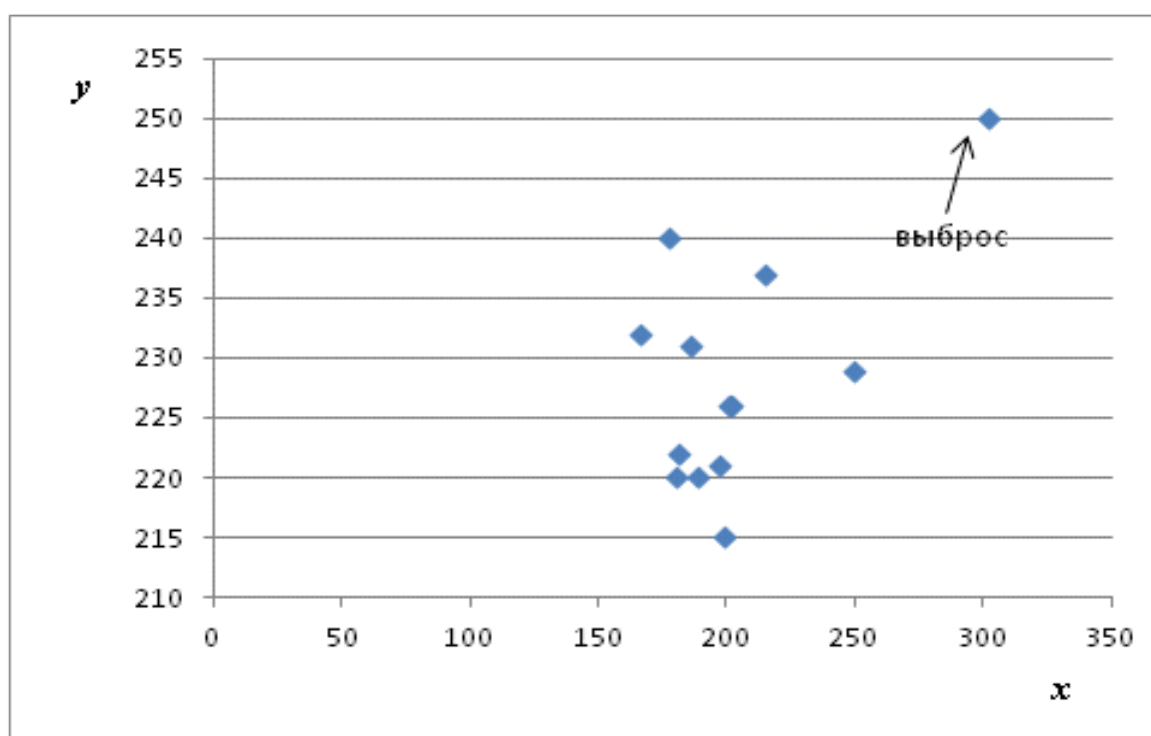


Рис. 2. Диаграмма рассеяния исходного массива данных $y(x)$ с предполагаемым выбросом

Для оценки выброса двух взаимосвязанных массивов X и Y необходимо использовать критерии, характеризующие связи этих массивов. Такими критериями могут быть коэффициент корреляции и регрессия Y по X .

Линейный коэффициент корреляции r_{xy} характеризует тесноту и направление связи двух факторов X и Y и вычисляется по формуле [3]

$$r_{xy} = b_1 \cdot \frac{\sigma_x}{\sigma_y} = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{\sigma_x \cdot \sigma_y}, \quad (1)$$

с учетом того, что линейная модель связи факторов x и y имеет вид $y_{теор} = b_0 + b_1x$, а средние квадратические отклонения вычисляются как

$$\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}; \quad \sigma_y = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

Для выявления выброса строится диаграмма рассеяния взаимосвязанных массивов X и Y . По диаграмме визуально определяется предполагаемый выброс с координатами $(x_\theta; y_\theta)$. Вычисляется коэффициент корреляции по исходному массиву данных r_{xy} и коэффициент корреляции r_{xy1} по исходным данным, но без учета предполагаемого выброса $(x_\theta; y_\theta)$.

Если выполняется условие

$$|r_{xy} - r_{xy1}| > 0,15, \quad (2)$$

то проверяемую координату $(x_\theta; y_\theta)$ можно считать выбросом или грубой ошибкой и она должна быть исключена из дальнейшего рассмотрения (например, построения математической модели исследуемого процесса).

Для повышения надежности полученного вывода необходимо подтверждать статистическую значимость вычисляемых коэффициентов корреляции, которая оценивается с помощью t -критерия Стьюдента. С этой целью вычисляется t -критерий Стьюдента по формуле

$$t_r = r \cdot \frac{\sqrt{n-2}}{\sqrt{1-r^2}} \quad (3)$$

и его величина сравнивается с табличным значением $t_{табл}$ для заданного уровня значимости α (обычно 5 %) и степени свободы $(n-2)$.

Если $t_r > t_{табл}$, то проверяемый коэффициент корреляции статистически значим и связь между исходными массивами данных X и Y можно считать доказанной.

Второй критерий оценки выброса в исходном двумерном массиве данных - регрессия Y по X .

Рассмотрим последовательность действий по оценке предполагаемого выброса в исходном двумерном массиве с помощью линейной регрессии.

1. По исходному двумерному массиву строится диаграмма рассеяния с целью выявления координаты предполагаемого выброса $(x_\theta; y_\theta)$.

2. Строится линейная регрессия $y_{теор} = b_0 + b_1x$ по исходным данным и линейная регрессия $y'_{теор} = b'_0 + b'_1x$ по исходным данным, но без предполагаемого выброса $(x_\theta; y_\theta)$.

3. Вычисляются остаточные компоненты для полученных в п. 2 уравнений регрессии: $\varepsilon = y - y_{теор}$ и $\varepsilon' = y' - y'_{теор}$.

4. Рассчитываются суммы квадратов остаточных компонентов: $S^2 = \sum_{i=1}^n \varepsilon_i^2$ и

$$S'^2 = \sum_{i=1}^{n-1} \varepsilon_i'^2 \text{ и их отношение } R = \frac{S^2}{S'^2}.$$

5. Оценивается статистическая значимость отношения R с помощью F -критерия Фишера. Если $R > F_{табл}$, то предполагаемый выброс считается существенным и

влияющим на искажение характеристики связи исходных факторов двумерного массива X и Y . Тогда координата $(x_6; y_6)$ должна быть исключена из дальнейшего расчета. Другой вариант решения заключается в анализе ситуаций, которые могли привести к выбросу. Отметим, что при оценке $F_{табл}$ берутся следующие степени свободы: степень свободы числителя $n_1 = 2$ (число параметров при переменных x), степень свободы знаменателя $n_2 = n - n_1 - 1 = n - 3$. Вероятность или значимость ошибки можно принять $\alpha = 5\%$.

Рассмотрим пример оценки выброса для двумерного массива, данные которого приведены в таблице 1.

Таблица 1

Величина прожиточного минимума пенсионера и средний размер назначенной пенсии в Российской Федерации в 2012-2017 годах

Год	Величина прожиточного минимума пенсионера, руб.	Средний размер назначенных пенсий, руб.
	x	y
2012	5123	8307
2013	5998	9201,1
2014	6617	10095,3
2015	7965	10961,5
2016	8081	12177,6
2017 (1-й квартал)	8178	17538,7

Источник: URL: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/population/level/ (дата обращения: 04.09.2017).

Диаграмма рассеяния размера прожиточного минимума пенсионера (x) и среднего размера назначенной пенсии (y), построенная по данным таблицы 1, приведена на рисунке 3. На диаграмме можно выделить предполагаемый выброс с координатами $(x_6 = 8178; y_6 = 17538,7)$, который соответствует данным таблицы 1 за 2017 год.

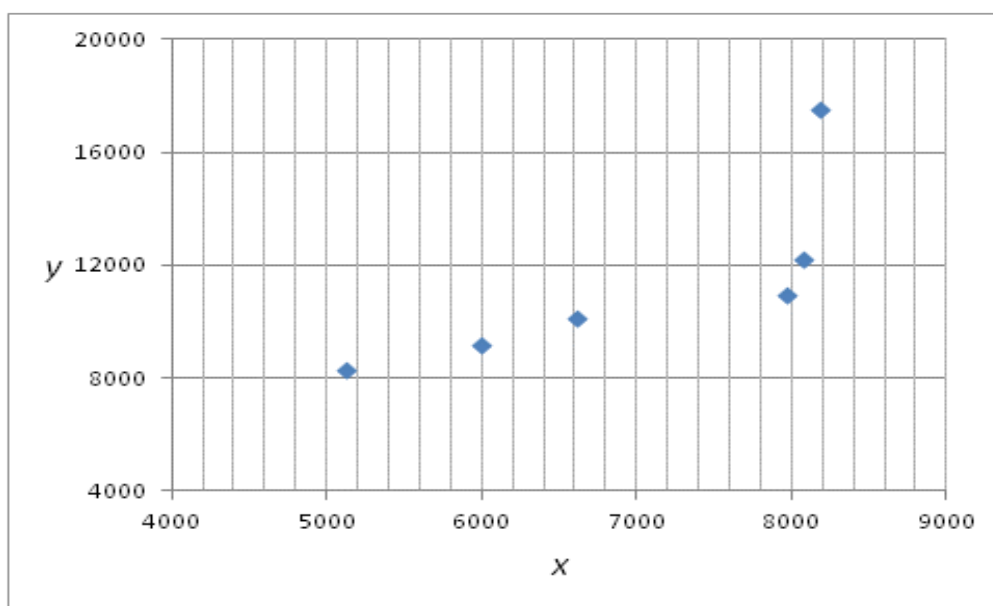


Рис. 3. Диаграмма рассеяния исходных данных

По исходным данным получены следующие линейные коэффициенты корреляции с учетом выброса и без него: $r_{xy} = 0,77$, $r_{xy1} = 0,97$. Так как $|r_{xy} - r_{xy1}| = 0,2 > 0,15$, т.е. условие (2) выполняется, то проверяемую координату ($x_8 = 8178$; $y_8 = 17538,7$) можно считать выбросом и она должна быть исключена из дальнейшего рассмотрения. Оценка статистической значимости линейных коэффициентов корреляции подтверждает сделанный вывод. А именно, расчетное значение t -критерия Стьюдента для данных с учетом выброса составило $t_r = 2,38$, что оказалось меньше табличного $t_{табл} = 4,18$ при уровне значимости 5 %. Это означает, что линейный коэффициент корреляции $r_{xy} = 0,77$ статистически не значим. Соответственно для линейного коэффициента корреляции $r_{xy1} = 0,97$ имеем $t'_r = 6,47$, что больше его табличного значения $t'_{табл} = 3,50$. Таким образом, зависимость между размером прожиточного минимума пенсионера (x) и среднего размера назначенной пенсии (y), представленных таблицей 1, статистически значима без данных с параметрами ($x_8 = 8178$; $y_8 = 17538,7$) за 2017 год.

Использование линейной регрессии для оценки предполагаемого выброса в исходном двумерном массиве дали следующие результаты:

$$y_{теор} = b_0 + b_1x = -2465,33 + 1,98x;$$

$$y'_{теор} = b'_0 + b'_1x = 2425,52 + 1,14x;$$

$$S^2 = \sum_{i=1}^n \varepsilon_i^2 = 22596619;$$

$$S'^2 = \sum_{i=1}^{n-1} \varepsilon_i'^2 = 606822,7;$$

$$R = 37,2; F_{табл} = 9,6.$$

Поскольку $R = 37,2 > F_{табл} = 9,6$, то проверяемая координата ($x_8 = 8178$; $y_8 = 17538,7$) в исходных данных может считаться выбросом, что подтверждает предыдущую оценку, выполненную с помощью линейного коэффициента корреляции.

Таким образом, данные величины прожиточного минимума пенсионера и среднего размера назначенной пенсии по Российской Федерации в 2012-2017 годах, представленные в таблице 1, содержат выброс с параметрами $x_8 = 8178$ и $y_8 = 17538,7$. Это может означать, что повышение среднего размера назначенной пенсии до 17538,7 рубля не соответствует предшествующей 2017 году закономерности связи прожиточного минимума пенсионера и среднего размера назначенной пенсии. Если бы эта закономерность распространялась и на 2017 год, то средний размер назначенной пенсии составил бы 11772,9 рубля.

Список литературы

1. Пустыльник Е. И. Статистические методы анализа и обработки наблюдений. Физико-математическая библиотека инженера. – М. : Наука, 1968.
2. Дубров А. М., Мхитарян В. С., Трошин Л. И. Многомерные статистические методы. Учебник. – М. : Финансы и статистика, 2003.
3. Поляков С. В. Использование Excel в моделировании, оценке деятельности и управлении (учебно-методическое пособие). – Владимир : Владимирский филиал РАНХиГС, 2015.