

Н. Баринов, FRICS, Санкт-Петербург  
М. Зельдин, FRICS, Санкт-Петербург  
Н. Ситников, Лондон

## **Линеаризация нелинейных связей в регрессионной модели, или еще раз об оцифровке влияющих переменных**

Построение моделей множественной регрессии нельзя считать новинкой в оценке недвижимости в России (см., например, [1,2]). Вместе с тем широкому их применению мешает, среди прочего, нелинейный характер влияния объясняющих переменных (ценообразующих факторов) на моделируемую оценщиком зависимую величину (рыночную стоимость, рыночную арендную плату и т.п.).

Трудности заключаются в том, что оценщик, строя регрессионную модель, как правило, имеет представление об общем характере нелинейной зависимости, однако не располагает информацией, достаточной для описания этой зависимости с необходимой точностью. Попытки построения моделей с линейными связями в ряде случаев не приводят к желаемым результатам.

Вместе с тем, наблюдаемые на практике нелинейные связи между зависимой и влияющей переменной не препятствуют использованию линейных<sup>1</sup> (аддитивных) регрессионных уравнений (моделей).

Задача заключается в нахождении *преобразования влияющей переменной*, сводящего *нелинейную* зависимость от влияющей переменной к *линейной*. Будем называть такие преобразования *линеаризующими*. Суть преобразования заключается в соответствующей оцифровке множества возможных значений объясняющей переменной, нелинейно влияющей на исследуемую функцию (зависимую переменную).

При успешном нахождении такого преобразования линейное регрессионное уравнение приводится к собственно линейному с заметным улучшением качества построения модели, в т.ч. показателей ее точности.

Методы и проблемы оцифровки влияющих переменных в оценочных задачах обсуждались ранее в [3]. Тем не менее, вопросы корректного учета нелинейных связей при построении аддитивных уравнений регрессии остаются сложными для восприятия оценщиками и требуют детального рассмотрения.

В предлагаемой публикации предпринята попытка наглядного и, вместе с тем, математически корректного разъяснения сути линеаризующих преобразований при построении регрессионных моделей.

Для облегчения восприятия материала рассмотрены преобразования одномерной (парной) зависимости последовательно для случаев детерминированной модели (при аналитическом и дискретном задании функции) и статистической модели с дискретным заданием функции, наиболее распространенной в оценочной практике.

Все полученные результаты могут быть естественным образом обобщены на случай множественной регрессии.

---

<sup>1</sup> Линейным назовем аддитивное уравнение регрессии, *линейное относительно коэффициентов* регрессии независимо от вида связей с влияющими переменными. Уравнение, линейное относительно своих *коэффициентов и влияющих переменных*, будем называть *собственно линейным*.

## 1. Детерминированная модель

### 1.1 Функция задана аналитически

Рассмотрим модель, заданную уравнением вида  $y = f(x)$ , где  $f(x)$  – монотонная<sup>2</sup> нелинейная функция. Необходимо подобрать преобразование  $z = z(x)$  такое, чтобы функция  $y = f(z(x))$  стала линейной относительно новой переменной  $z$ , т.е.  $f(z(x)) \equiv g(z) = a + bz$ .

Графически эта задача может быть пояснена следующим рисунком:

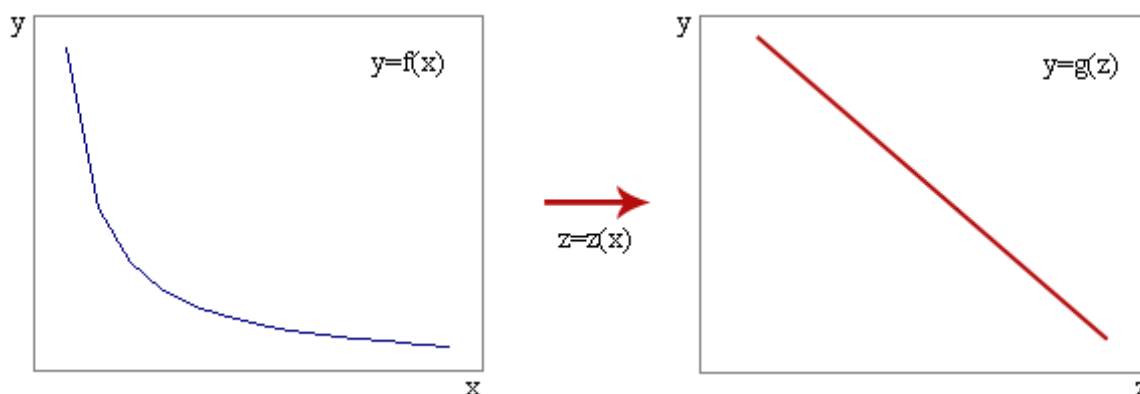


Рис. 1

Универсального преобразования, решающего поставленную задачу, не существует. В каждом случае преобразование  $z = z(x)$  выбирается в зависимости от известного вида функции  $f(x)$ . Ниже в качестве примера приведены линеаризующие преобразования для некоторых элементарных функций<sup>3</sup>:

Исходная функция $y = f(x)$	Преобразование $z = z(x)$
$y = ax^2 + b$	$z = x^2$
$y = \frac{a}{x} + b$	$z = \frac{1}{x}$
$y = a \ln x + b$	$z = \ln x$

### 1.2 Функция задана дискретно

В отличие от непрерывного случая, где нет универсального преобразования, линеаризирующего нелинейную модель, для дискретно заданной функции такое преобразование существует.

Покажем это.

Рассмотрим нелинейную функцию  $y = f(x)$ , заданную на конечном множестве точек  $(x_i, y_i)$ ,  $i \in [1, N]$ . Найдем такое преобразование  $z(x_i) = z_i$ , в результате которого нелинейная функция  $y = g(z) \equiv f(z(x))$ , заданная на множестве точек  $\{(z_i, y_i)\}$ , становится линейной.

<sup>2</sup> Монотонная функция – функция, приращение которой не меняет знака, то есть всегда либо неотрицательно, либо неположительно. Если, в дополнение, приращение не равно нулю, функция называется строго монотонной.

<sup>3</sup> при  $x > 0$ .

Выберем наугад пару точек  $(z_1, y_1)$  и  $(z_2, y_2)$ .

Потребуем, чтобы зависимость между  $y$  и новой переменной  $z$  стала линейной, т.е. чтобы нашлась прямая  $y = a_{12}z + b_{12}$ , содержащая обе точки  $(z_i, y_i)$ ,  $i=1, 2$ .

Для каждой из выбранных точек справедливо  $\begin{cases} y_1 = a_{12}z_1 + b_{12} \\ y_2 = a_{12}z_2 + b_{12} \end{cases}$ , откуда легко найти неизвестные коэффициенты:

$$a_{12} = \frac{y_2 - y_1}{z_2 - z_1} \quad \text{и} \quad b_{12} = \frac{y_1}{y_2 - y_1} + \frac{z_1}{z_2 - z_1}.$$

Взяв другую пару точек  $(z_2, y_2)$ ,  $(z_3, y_3)$ , содержащую одну из точек первой пары, вычислим аналогичным образом коэффициенты прямой  $y = a_{23}z + b_{23}$ :

$$a_{23} = \frac{y_3 - y_2}{z_3 - z_2} \quad \text{и} \quad b_{23} = \frac{y_2}{y_3 - y_2} + \frac{z_2}{z_3 - z_2}.$$

Возможны следующие варианты взаимного расположения прямых  $y = a_{12}z + b_{12}$  и  $y = a_{23}z + b_{23}$  на плоскости: параллельность, пересечение и совпадение. Так как рассматриваемые прямые имеют общую точку  $(z_2, y_2)$ , очевидно, что эти прямые не параллельны, следовательно, они пересекаются.

Пересекающиеся прямые совпадают, когда совпадают их угловые коэффициенты, т.е.  $a_{12} = a_{23}$ .

Отсюда, необходимым и достаточным условием расположения всех точек  $(z_i, y_i)$  на одной прямой является выполнение равенств:

$$a_{12} = a_{23} = \dots = a_{N-2, N-1} = a_{N-1, N}, \quad \text{где} \quad a_{i-1, i} = \frac{y_i - y_{i-1}}{z_i - z_{i-1}}. \quad (1)$$

Необходимые и достаточные условия можно также сформулировать следующим образом:

Если система алгебраических уравнений

$$\begin{cases} a_{12} - a_{23} = 0 \\ a_{23} - a_{34} = 0 \\ \vdots \\ a_{N-2, N-1} - a_{N-1, N} = 0 \end{cases} \quad (2)$$

имеет нетривиальное решение, то обязательно найдется прямая, соединяющая все точки  $(z_i, y_i)$ , причем эта прямая определяется *не единственным образом*.

Действительно, в системе уравнений (2) имеем  $N$  неизвестных и  $N-2$  уравнений, т.е. система имеет две степени свободы, которые и характеризуют расположение прямой на плоскости.

Результат описанного выше преобразования схематически<sup>4</sup> показан на рис. 2

<sup>4</sup> Исходная функция  $y = f(x)$ , как и на рис.1, показана гладкой, тогда как фактически речь пойдет о кусочно-линейной аппроксимации (см. рис. 3-7). При этом под монотонностью такой аппроксимирующей функции будем понимать сохранение знака наклона отрезков, соединяющих точки, на которых определена функция.

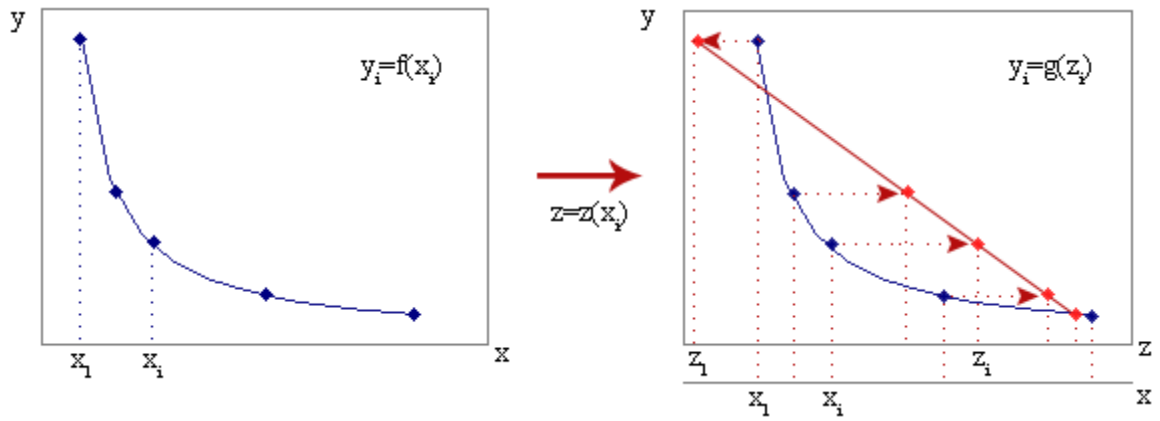


Рис. 2

Видно, что значения влияющей переменной  $x$  при преобразовании в шкалу значений  $z$  изменяются на разную величину и с разными направлениями (знаками).

В качестве примера рассмотрим нелинейную дискретную функцию  $y = f(x)$ , заданную на множестве, состоящем из пяти точек  $(x_1, y_1)$ ,  $(x_2, y_2)$ ,  $(x_3, y_3)$ ,  $(x_4, y_4)$ ,  $(x_5, y_5)$ . Будем считать, что  $x_i \neq x_j$  для всех  $i \neq j$ .

Преобразование  $z(x_i) = z_i$ , после которого точки  $(z_1, y_1)$ ,  $(z_2, y_2)$ ,  $(z_3, y_3)$ ,  $(z_4, y_4)$ ,  $(z_5, y_5)$  лягут на одну прямую, найдем из системы (2):

$$\begin{cases} a_{12} - a_{23} = 0 \\ a_{23} - a_{34} = 0 \\ a_{34} - a_{45} = 0 \end{cases}$$

Согласно выражению (1), полученную систему можно записать:

$$\begin{cases} \frac{y_2 - y_1}{z_2 - z_1} - \frac{y_3 - y_2}{z_3 - z_2} = 0 \\ \frac{y_3 - y_2}{z_3 - z_2} - \frac{y_4 - y_3}{z_4 - z_3} = 0 \\ \frac{y_4 - y_3}{z_4 - z_3} - \frac{y_5 - y_4}{z_5 - z_4} = 0 \end{cases}$$

Так как все  $x_i$  различны, то и соответствующие им  $z_i$  будут различны, поэтому справедлива следующая запись:

$$\begin{cases} (y_2 - y_1)(z_3 - z_2) - (y_3 - y_2)(z_2 - z_1) = 0 \\ (y_3 - y_2)(z_4 - z_3) - (y_4 - y_3)(z_3 - z_2) = 0 \\ (y_4 - y_3)(z_5 - z_4) - (y_5 - y_4)(z_4 - z_3) = 0 \end{cases}$$

Полученная система линейных уравнений может быть представлена в матричной форме:

$$\begin{pmatrix} y_3 - y_2 & y_1 - y_3 & y_2 - y_1 & 0 & 0 \\ 0 & y_4 - y_3 & y_2 - y_4 & y_3 - y_2 & 0 \\ 0 & 0 & y_5 - y_4 & y_3 - y_5 & y_4 - y_3 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Очевидно, переменные  $z_4$  и  $z_5$  могут принимать любые значения, а оставшиеся  $z_3$ ,  $z_2$  и  $z_1$  определяются по формулам, включающим уже известные значения  $z$ :

$$\begin{aligned} z_3 &= \frac{y_5 - y_3}{y_5 - y_4} z_4 + \frac{y_3 - y_4}{y_5 - y_4} z_5, & z_2 &= \frac{y_4 - y_2}{y_4 - y_3} z_3 + \frac{y_2 - y_3}{y_4 - y_3} z_4, \\ z_1 &= \frac{y_3 - y_1}{y_3 - y_2} z_2 + \frac{y_1 - y_2}{y_3 - y_2} z_3 \end{aligned} \quad (*)$$

Рассмотрим пять конкретных точек  $(x_i, y_i)$ : (30;70,7), (50;51,2), (75;31,2), (200;15,7), (400;12,8). Из расположения этих точек на координатной плоскости видна (рис. 3а) нелинейная зависимость  $y = f(x)$ .

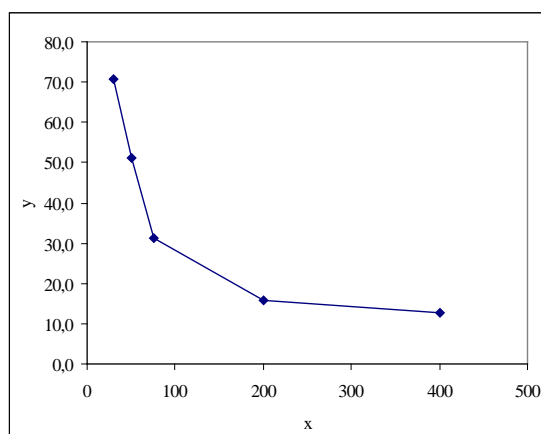


Рис. 3а

Пронумеруем точки следующим образом:

Точка	$(x_1, y_1)$	$(x_2, y_2)$	$(x_3, y_3)$	$(x_4, y_4)$	$(x_5, y_5)$
	(30;70,7)	(50;51,2)	(75;31,2)	(200;15,7)	(400;12,8)
Преобразование	$z_1$	$z_2$	$z_3$	$z_4$ (с)	$z_5$ (с)

Нам необходимо задать значения «свободных»<sup>5</sup> переменных  $z_4$  (с) и  $z_5$  (с)

Рассмотрим несколько вариантов:

<sup>5</sup> «Свободными» переменными могут быть не только последние две точки  $(x_4, y_4)$ ,  $(x_5, y_5)$ , но любые другие две точки. Покажем это.

Предположим, что произвольным образом задаются преобразования для 1-й и 3-й точек (т.е.  $z_1$  и  $z_3$ ):

Точка	$(x_1, y_1)$	$(x_2, y_2)$	$(x_3, y_3)$	$(x_4, y_4)$	$(x_5, y_5)$
Преобразование	$z_1$ (с)	$z_2$	$z_3$ (с)	$z_4$	$z_5$

Перенумеруем точки так, чтобы первая и третья исходные точки в новых обозначениях стали бы предпоследней и последней соответственно:

Новое обозначение	$(\tilde{x}_1, \tilde{y}_1)$	$(\tilde{x}_2, \tilde{y}_2)$	$(\tilde{x}_3, \tilde{y}_3)$	$(\tilde{x}_4, \tilde{y}_4)$	$(\tilde{x}_5, \tilde{y}_5)$
Исходное обозначение	$(x_4, y_4)$	$(x_2, y_2)$	$(x_5, y_5)$	$(x_1, y_1)$	$(x_3, y_3)$
Преобразование	$z_4$	$z_2$	$z_5$	$z_1$ (с)	$z_3$ (с)

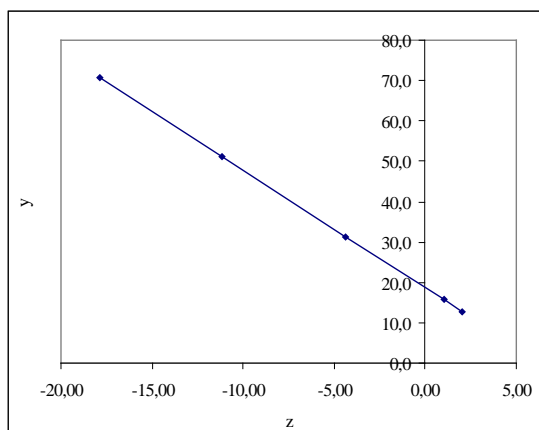
В новых обозначениях задача решается аналогично записанной в исходных.

Вариант 1. Возьмем в качестве значений свободных переменных  $z_4 = 1$ ,  $z_5 = 2$  и из формул (\*) найдем оставшиеся переменные:

$$z_3 = \frac{12,8 - 31,2}{12,8 - 15,7} \cdot 1 + \frac{31,2 - 15,7}{12,8 - 15,7} \cdot 2 = -4,34, \quad z_2 = -4,34 \cdot \frac{15,7 - 51,2}{15,7 - 31,2} + \frac{51,2 - 31,2}{15,7 - 31,2} \cdot 1 = -11,17,$$

$$z_1 = -11,17 \cdot \frac{31,2 - 70,7}{31,2 - 51,2} - 4,34 \cdot \frac{70,7 - 51,2}{31,2 - 51,2} = -17,86.$$

Заметим, что выбор значений свободных переменных нельзя признать удобным, т.к. привел к тому, что некоторые из значений преобразованных переменных  $z_i$  приняли отрицательные значения (рис. 3б) тогда как все исходные были положительными.



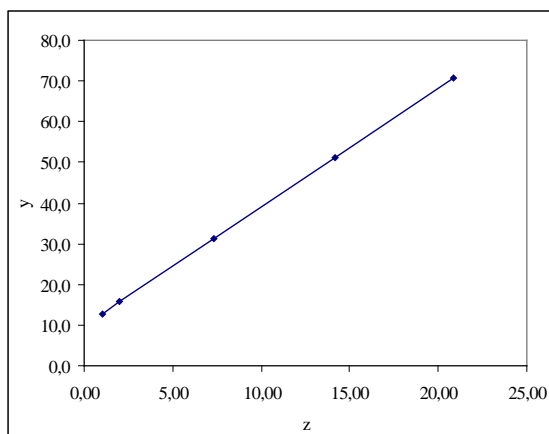
**Рис. 3б**

Вариант 2. Изменим значения свободных переменных на  $z_4 = 2$ ,  $z_5 = 1$  и вновь найдем оставшиеся переменные из формул (\*):

$$z_3 = \frac{12,8 - 31,2}{12,8 - 15,7} \cdot 2 + \frac{31,2 - 15,7}{12,8 - 15,7} \cdot 1 = 7,34, \quad z_2 = 7,34 \cdot \frac{15,7 - 51,2}{15,7 - 31,2} + \frac{51,2 - 31,2}{15,7 - 31,2} \cdot 2 = 14,17,$$

$$z_1 = 14,17 \cdot \frac{31,2 - 70,7}{31,2 - 51,2} + 7,34 \cdot \frac{70,7 - 51,2}{31,2 - 51,2} = 20,86.$$

И этот выбор значений свободных переменных не является удобным. Хотя нам удалось перевести все точки в квадрант положительных значений, преобразованная функция стала возрастающей (рис. 3в), в то время как исходная функция является убывающей.



**Рис. 3в**

Вариант 3. Адекватным<sup>6</sup> преобразованием для заданного набора точек будет, например, следующее:

Точка	$(x_1, y_1)$	$(x_2, y_2)$	$(x_3, y_3)$	$(x_4, y_4)$	$(x_5, y_5)$
	(30;70,7)	(50;51,2)	(75;31,2)	(200;15,7)	(400;12,8)
Преобразование	$z_1$	$z_2$	$z_3$	$z_4(c)$	$z_5(c)$
	0,05	0,45	0,86	1,18	1,24

В результате адекватного преобразования получаем зависимость  $y = g(z) \equiv f(z(x))$  (рис. 3г) с аналогичным исходной зависимости знаком приращения и квадрантом значений аргументов и функции.

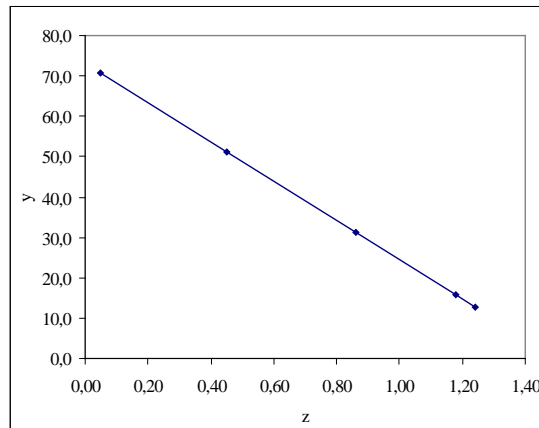


Рис. 3г

### Как определить адекватное преобразование?

Введем следующие обозначения:

$$z_3 = f_3(z_4, z_5) = 6,34z_4 - 5,34z_5$$

$$z_2 = f_2(z_3, z_4) = f_2(f_3(z_4, z_5), z_4) = 13,24z_4 - 12,24z_5$$

$$z_1 = f_1(z_2, z_3) = f_1(f_2(f_3(z_4, z_5), z_4), f_3(z_4, z_5)) = 19,97z_4 - 18,97z_5$$

Условие сохранения квадранта значений переменной требует, чтобы

$$z_3 = f_3(z_4, z_5) > 0$$

$$z_2 = f_2(f_3(z_4, z_5), z_4) > 0$$

$$z_1 = f_1(f_2(f_3(z_4, z_5), z_4), f_3(z_4, z_5)) > 0$$

Условие сохранения знака приращения функции требует, чтобы

$$0 < z_1 < z_2 < z_3 < z_4 < z_5.$$

Преобразование является адекватным, когда  $z_4$  и  $z_5$  удовлетворяют условиям:

$$\begin{cases} z_4 < z_5 \\ f_3(z_4, z_5) < z_4 \\ f_2(f_3(z_4, z_5), z_4) < f_3(z_4, z_5) \\ f_1(f_2(f_3(z_4, z_5), z_4), f_3(z_4, z_5)) < f_2(f_3(z_4, z_5), z_4) \\ f_1(f_2(f_3(z_4, z_5), z_4), f_3(z_4, z_5)) > 0 \end{cases}$$

<sup>6</sup> Адекватным считаем такое линеаризующее преобразование, которое сохраняет знак приращения функции (убывающая, возрастающая) и диапазон изменения (квадрант) значений влияющей переменной ( $x > 0, z > 0$ ).

или

$$\begin{cases} z_4 < z_5 \\ 6,34z_4 - 5,34z_5 < z_4 \\ 13,24z_4 - 12,24z_5 < 6,34z_4 - 5,34z_5 \\ 19,97z_4 - 18,97z_5 < 13,24z_4 - 12,24z_5 \\ 19,97z_4 - 18,97z_5 > 0 \end{cases},$$

или

$$\begin{cases} z_4 < z_5 \\ 1,05z_4 > z_5 \end{cases}$$

В общем случае, для любых других пяти точек:

$$\begin{cases} z_4 < z_5 \\ \frac{ace + ad + be}{acf + bf} z_4 + z_5 > 0 \end{cases}, \quad \text{где} \quad \begin{aligned} z_3 &= ez_4 + fz_5 \\ z_2 &= cz_4 + dz_5 \\ z_1 &= az_4 + bz_5 \end{aligned}$$

Ниже на рисунке серым цветом выделена область значений  $z_4$  и  $z_5$ , соответствующая адекватным преобразованиям зависимости, заданной пятью точками:

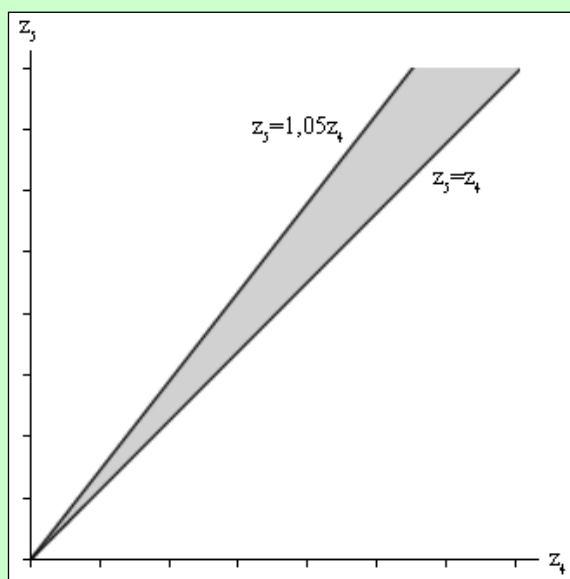


Рис. 3д

Видно, что условия адекватного преобразования<sup>7</sup> могут быть выполнены при задании не единственной пары значений «свободных» переменных. Это означает, что наборов линеаризованных переменных  $z_i$ , и, следовательно, линеаризующих прямых  $y = g(z) \equiv f(z(x))$  может быть достаточно много.

Полученные соотношения могут быть использованы при построении регрессионных моделей, однако для этого необходимо провести дополнительные вычисления.

<sup>7</sup> когда оно существует



## 2. Статистическая модель

В детерминированной модели предполагалось, что существует однозначное соответствие между значениями влияющей и зависимой переменными, т.е. для каждого  $x_i$  существовало только одно значение  $y_i$  такое, что  $y_i = f(x_i)$ , и наоборот.

Предположим теперь, что взаимно-однозначное соответствие не выполняется, т.е. всякому  $x_i$  ставится в соответствие случайная величина  $y$ , которая принимает не единственное значение.

Рассмотрим  $f(x)$  — нелинейную аппроксимирующую функцию дискретного аргумента, заданную на множестве точек  $\{(x_i, y_i)\}$ , причем обязательно найдутся индексы  $l, m$  при которых  $y_l \neq y_m$ ,  $x_l = x_m$ . Будем считать, что существует всего  $k$  различных значений  $x_i$ .

Аппроксимирующую функцию построим методом наименьших квадратов<sup>8</sup> из условия:

$$\min TSS = \min \sum_{i=1}^N (y_i - f(x_i))^2, \quad (3)$$

где  $TSS$  — сумма квадратов остатков модели.

Найдем такое преобразование  $z(x_i) = z_i$ , в результате которого аппроксимирующая функция  $g(z) \equiv f(z(x))$ , заданная на множестве точек  $\{(z_i, y_i)\}$ , становится линейной.

Запишем условие (3) для функции  $g$ , заданной в виде  $g(z) = a_0 + a_1 z$ :

$$\min_{a_0, a_1} TSS(a_0, a_1) = \min_{a_0, a_1} \sum_{i=1}^N (y_i - a_0 - a_1 z_i)^2. \quad (4)$$

Для точек, по которым строится аппроксимирующая функция  $g$ , введем новые обозначения  $(z_i, y_{i,j})$ , где  $i \in [1, k]$ ,  $j \in [1, n_i]$ ,  $\sum_{i=1}^k n_i = N$ . Тогда множество  $M = \{(z_i, y_i)\}$  можно разбить на  $k$  непересекающихся подмножеств  $M_i = \{(z_i, y_{i,1}), (z_i, y_{i,2}), \dots, (z_i, y_{i,n_i})\}$ :

$$M = \bigcup_{i=1}^k M_i.$$

В каждом подмножестве  $M_i$  вычислим среднее значение  $\bar{y}_{n_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{i,j}$ .

Тогда каждая из точек  $(z_i, y_{i,j}) \in M_i$  представима в виде  $(z_i, \bar{y}_{n_i} + \Delta_j) \in M_i$ , где  $\Delta_j = y_{i,j} - \bar{y}_{n_i}$ .

Используя новые обозначения, перепишем формулу (4) в виде:

$$\begin{aligned} \min_{a_0, a_1} TSS(a_0, a_1) &= \min_{a_0, a_1} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{i,j} - a_0 - a_1 z_i)^2 = \sum_{i=1}^k \min_{a_0, a_1} \sum_{j=1}^{n_i} (y_{i,j} - a_0 - a_1 z_i)^2 = \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} \min_{a_0, a_1} (y_{i,j} - a_0 - a_1 z_i)^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} \min_{a_0, a_1} (\bar{y}_{n_i} + \Delta_j - a_0 - a_1 z_i)^2. \end{aligned} \quad (5)$$

<sup>8</sup> Метод наименьших квадратов — статистический метод определения параметров ...путем минимизации критериев суммы квадратов отклонений между фактическими и расчетными данными. *Словарь бизнес-терминов. 2000*: <http://dic.academic.ru/dic.nsf/business/7781>;

В формуле (5) преобразуем выражение, стоящее под знаком суммы:

$$\begin{aligned}\min_{a_0, a_1} (\bar{y}_{n_i} + \Delta_j - a_0 - a_1 z_i)^2 &= \min_{a_0, a_1} \left( (\bar{y}_{n_i} - a_0 - a_1 z_i)^2 + 2\Delta_j (\bar{y}_{n_i} - a_0 - a_1 z_i) + \Delta_j^2 \right) = \\ &= \min_{a_0, a_1} \left( (\bar{y}_{n_i} - a_0 - a_1 z_i)^2 + 2\Delta_j (\bar{y}_{n_i} - a_0 - a_1 z_i) \right) + \Delta_j^2,\end{aligned}$$

и получим

$$\begin{aligned}\min_{a_0, a_1} TSS(a_0, a_1) &= \sum_{i=1}^k \sum_{j=1}^{n_i} \left( \min_{a_0, a_1} \left( (\bar{y}_{n_i} - a_0 - a_1 z_i)^2 + 2\Delta_j (\bar{y}_{n_i} - a_0 - a_1 z_i) \right) + \Delta_j^2 \right) = \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} \min_{a_0, a_1} \left( (\bar{y}_{n_i} - a_0 - a_1 z_i)^2 + 2\Delta_j (\bar{y}_{n_i} - a_0 - a_1 z_i) \right) + \sum_{i=1}^k \sum_{j=1}^{n_i} \Delta_j^2.\end{aligned}\quad (6)$$

Таким образом,  $TSS(a_0, a_1)$  может быть представлено в виде суммы  $SS_1(a_0, a_1)$  и  $SS_2$ , где

$$SS_1(a_0, a_1) = \sum_{i=1}^k \sum_{j=1}^{n_i} \min_{a_0, a_1} \left( (\bar{y}_{n_i} - a_0 - a_1 z_i)^2 + 2\Delta_j (\bar{y}_{n_i} - a_0 - a_1 z_i) \right), \quad SS_2 = \sum_{i=1}^k \sum_{j=1}^{n_i} \Delta_j^2.$$

От параметров аппроксимирующей функции зависит только слагаемое  $SS_1(a_0, a_1)$ , поэтому задача сводится к нахождению его минимума, который, очевидно, достигается при  $\bar{y}_{n_i} = a_0 + a_1 z_i$ . Последнее означает, что все точки  $(z_i, \bar{y}_{n_i})$  должны принадлежать одной прямой.

Необходимым и достаточным условием расположения точек  $(z_i, y_i)$  на одной прямой является выполнение равенств (1)–(2), т.е. в данном случае:

$$\begin{cases} \tilde{a}_{12} - \tilde{a}_{23} = 0 \\ \tilde{a}_{23} - \tilde{a}_{34} = 0 \\ \vdots \\ \tilde{a}_{k-2, k-1} - \tilde{a}_{k-1, k} = 0 \end{cases}, \quad \text{где} \quad \tilde{a}_{i-1, i} = \frac{\bar{y}_{n_i} - \bar{y}_{n_{i-1}}}{z_i - z_{i-1}}. \quad (2^*)$$

Решение системы уравнений (2\*) всегда существует, поэтому обязательно найдется преобразование  $z(x_i) = z_i$  такое, что  $SS_1(\tilde{a}_0, \tilde{a}_1) = 0$ , и новая аппроксимирующая линейная функция  $g(z_i) = \tilde{a}_0 + \tilde{a}_1 z_i$  дает минимум  $TSS(a_0, a_1)$ :

$$\min_{\tilde{a}_0, \tilde{a}_1} TSS(\tilde{a}_0, \tilde{a}_1) = SS_1(\tilde{a}_0, \tilde{a}_1) + SS_2 = SS_2 = \sum_{i=1}^k \sum_{j=1}^{n_i} \Delta_j^2. \quad (7)$$

Результат описанного выше преобразования схематически показан на рис. 4.

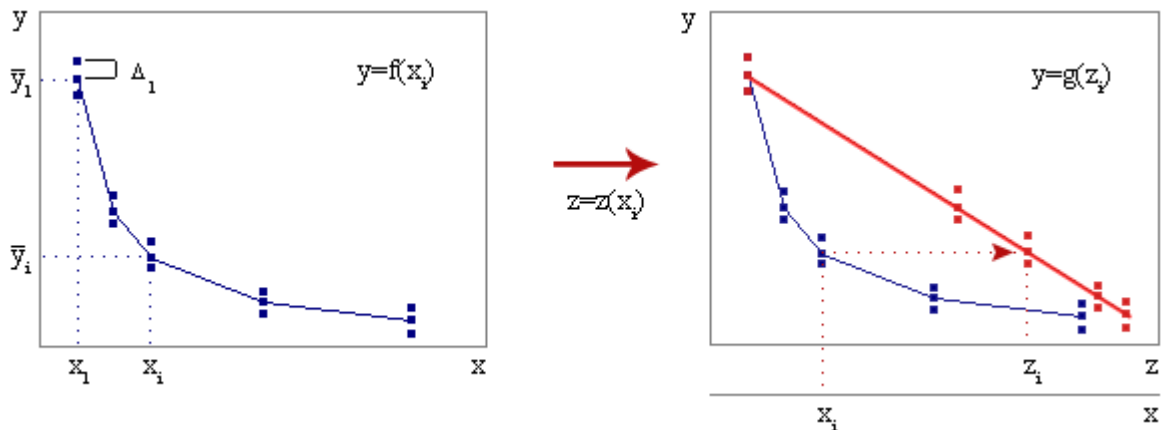


Рис. 4

Рассмотрим пример линейризирующего преобразования при нахождении регрессионной зависимости.

Пусть задана выборка, состоящая из 30 точек  $\{(x_i, y_i)\}$  и описывающая наблюдаемую нелинейную зависимость<sup>9</sup> между влияющей и зависимой переменными:

(30;77,7), (30;72), (30;69), (30;67,5), (30;72,3), (30;65,5),  
 (50;48,9), (50;47), (50;50), (50;53,5), (50;51), (50;56,6),  
 (75;31,4), (75;32,4), (75;35), (75;33), (75;27,3), (75;28,3),  
 (200;20), (200;19,4), (200;16), (200;15), (200;13,6), (200;10),  
 (400;10), (400;11,6), (400;12,3), (400;13,4), (400;14,2), (400;15).

По этой выборке построим сначала линейную аппроксимирующую функцию вида  $h(x) = a_0 + a_1x$ . В *MS Excel* с помощью встроенной функции **ЛИНЕЙНО** получены следующие значения регрессионных статистик:

Коэффициент уравнения	$a_1$	-0,13	55,60	$a_0$	Свободный член уравнения
СКО коэффициента уравнения	$S_{a1}$	0,02	3,76	$S_{a0}$	СКО свободного члена
Коэффициент детерминации	$R^2$	<b>0,63</b>	13,90	$S_{ост.}$	Остаточное СКО= $(TSS/n-k-1)^{0,5}$
Статистика Фишера	$F_{расч.}$	48,26	28	$n - k - 1$	Число степеней свободы
Регрессионная сумма квадратов	$SS_{регр.}$	9321,28	5408,65	$SS_{ост.}$	Остаточная сумма квадратов (TSS)

Значение коэффициента детерминации  $R^2$ , а также рис. 5а свидетельствуют о том, что построенная аппроксимирующая функция  $h(x) = 55,6 - 0,13x$  не годится для описания существующей зависимости.

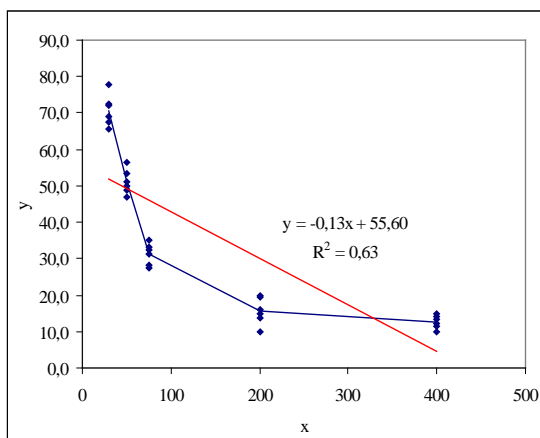


Рис. 5а

Найдем преобразование  $z(x_i) = z_i$ , посредством которого аппроксимирующая функция  $g(z) = \tilde{a}_0 + \tilde{a}_1z$ , заданная на множестве точек  $\{(z_i, y_i)\}$ , будет линейной.

Для этого разобьем исходное множество точек на пять подмножеств в соответствие с пятью значениями  $x$  (30, 50, 75, 200, 400):

$$\begin{aligned}
 M_1 &= \{(z_1;77,7), (z_1;72), (z_1;69), (z_1;67,5), (z_1;72,3), (z_1;65,5)\}, \\
 M_2 &= \{(z_2;48,9), (z_2;47), (z_2;50), (z_2;53,5), (z_2;51), (z_2;56,6)\}, \\
 M_3 &= \{(z_3;31,4), (z_3;32,4), (z_3;35), (z_3;33), (z_3;27,3), (z_3;28,3)\}, \\
 M_4 &= \{(z_4;20), (z_4;19,4), (z_4;16), (z_4;15), (z_4;13,6), (z_4;10)\}, \\
 M_5 &= \{(z_5;10), (z_5;11,6), (z_5;12,3), (z_5;13,4), (z_5;14,2), (z_5;15)\}.
 \end{aligned}$$

<sup>9</sup> Подобным образом может зависеть, например, арендная плата за торговые помещения от расстояния (в определенном диапазоне изменения) до границ мощного локального центра влияния (при прочих равных).

В каждом подмножестве  $M_i$  вычислим *среднее значение* случайной величины  $y_i$  и сформируем пять точек  $(z_i, \bar{y}_{n_i})$ :  $(z_1; 70,7)$ ,  $(z_2; 51,2)$ ,  $(z_3; 31,2)$ ,  $(z_4; 15,7)$ ,  $(z_5; 12,8)$ .

Для этих пяти точек сформулируем условия (2\*) в виде:

$$\begin{cases} \tilde{a}_{12} - \tilde{a}_{23} = 0 \\ \tilde{a}_{23} - \tilde{a}_{34} = 0, \\ \tilde{a}_{34} - \tilde{a}_{45} = 0 \end{cases} \text{ где } \tilde{a}_{i-1,i} = \frac{\bar{y}_{n_i} - \bar{y}_{n_{i-1}}}{z_i - z_{i-1}}. \quad (**)$$

Способ решения полученной системы уравнений детально описан выше для случая детерминированной модели.

Адекватным преобразованием<sup>10</sup> для точек  $(z_i, \bar{y}_{n_i})$  будут значения:  $z_1 = 0,05$ ,  $z_2 = 0,45$ ,  $z_3 = 0,86$  при заданных «произвольных»  $z_4 = 1,18$  и  $z_5 = 1,24$ .

Для аппроксимирующей функции  $g(z) = \tilde{a}_0 + \tilde{a}_1 z$ , заданной на исходном множестве из 30 точек  $\{(z_i, y_i)\}$ , регрессионные статистики<sup>11</sup> имеют следующие значения:

$a_1$	-48,611	73,028	$a_0$
$S_{a1}$	1,28	1,13	$S_{a0}$
$R^2$	<b>0,98</b>	<b>3,17</b>	$S_{ост.}$
$F_{расч.}$	1437,83	28	$n - k - 1$
$SS_{регр.}$	14448,56	281,37	$SS_{ост.}$

Видно (рис. 5б), что аппроксимирующая функция  $g(z) = 73,028 - 48,611z$  практически полностью объясняет наблюдаемую зависимость.

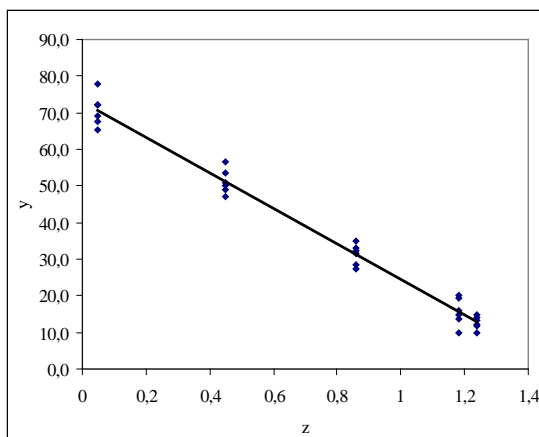


Рис. 5б

Ряд практических задач, в т.ч. построение многомерных регрессионных моделей в оценочной деятельности, *не требует аналитического представления* линеаризующего преобразования, достаточно нахождения численного решения системы уравнений (\*\*).

Преобразование, близкое к численному решению системы (\*\*), может быть получено с помощью инструмента **Solver** (Поиск решения) в MS Excel.<sup>12</sup> Покажем это.

Как и прежде разобьем исходную выборку на подмножества  $M_i$ , для всех точек которого ищется адекватное преобразование  $z_i$ .

<sup>10</sup> См. выше в разделе Детерминированная модель

<sup>11</sup> Получены в MS Excel с помощью встроенной функции **ЛИНЕЙН()**. Отображены два знака после запятой.

<sup>12</sup> См., например, Руководство пользователя . Microsoft® Excel. Версия 5.0 (или более поздняя)

Применение численного метода оптимизации предполагает наличие начального приближения, которое можно задать произвольным образом (например,  $z_1 = 1$ ,  $z_2 = 2$ ,  $z_3 = 3$ ,  $z_4 = 4$ ,  $z_5 = 5$ ), но обязательно *сохраняя порядок следования* (нарастания значений) меток<sup>13</sup> аналогичным порядку нарастания значений  $x_i$ .

$$\begin{aligned} M_1 &= \{(1;77,7), (1;72), (1;69), (1;67,5), (1;72,3), (1;65,5)\}, \\ M_2 &= \{(2;48,9), (2;47), (2;50), (2;53,5), (2;51), (2;56,6)\}, \\ M_3 &= \{(3;31,4), (3;32,4), (3;35), (3;33), (3;27,3), (3;28,3)\}, \\ M_4 &= \{(4;20), (4;19,4), (4;16), (4;15), (4;13,6), (4;10)\}, \\ M_5 &= \{(5;10), (5;11,6), (5;12,3), (5;13,4), (5;14,2), (5;15)\}. \end{aligned}$$

По полученным точкам, с помощью функции **ЛИНЕЙНО** построим аппроксимирующую функцию  $s(z) = a_0 + a_1 z$  (рис. 5в), характеристиками которой являются следующие значения статистик:

$a_1$	-15,13	81,70	$a_0$
$S_{a1}$	0,77	2,54	$S_{a0}$
$R^2$	<b>0,93</b>	5,94	$S_{ост.}$
$F_{расч.}$	389,08	28	$n - k - 1$
$SS_{резр.}$	13741,07	988,86	$SS_{ост.}$

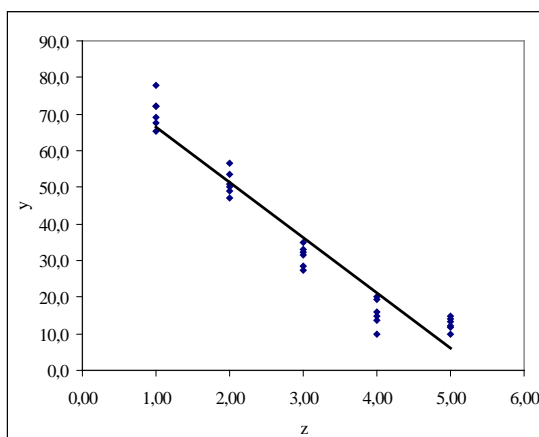


Рис. 5в

Это – *промежуточный результат* нахождения линейризующей функции, соответствующий *начальному приближению* значений  $z_i$ , произвольно заданному лишь с учетом ограничений на неотрицательность и монотонность следования меток.

Теперь с помощью инструмента **Solver**, подберем значения  $z_i$ , минимизирующие сумму квадратов остатков  $TSS(a_0, a_1)$ , что эквивалентно максимизации коэффициента детерминации  $R^2$ .

Не забудем учесть условия адекватности преобразования<sup>14</sup>, для чего в диалоговом окне инструмента **Solver** введем ограничения (условия неотрицательности и монотонности) на допустимые значения  $z_i$  (рис. 6).

<sup>13</sup> В работе [3] показано, что при отсутствии такого ограничения в результате преобразования координат может быть нарушен порядок следования (нарастания значений) меток и получены зависимости, противоречащие экономическим гипотезам, отражающим закономерности ценообразования на рынке.

<sup>14</sup> См. выше в разделе Детерминированная модель

	A	B	C	D	E
1	y	z	a1	a0	
2	77,7	1,00	-15,13	81,70	
3	72,0	1,00	0,77	2,54	
4	69,0	1,00	0,98	5,94	
5	67,5	1,00	389,08	28	
6	72,3	1,00	13741,07	988,86	
7	65,5	1,00			
8	48,9	2,00			
9	47,0	2,00			
10	50,0	2,00			
11	53,5	2,00			
12	51,0	2,00			
13	56,6	2,00			
14	31,4	3,00			
15	32,4	3,00			
16	35,0	3,00			
17	33,0	3,00			
18	27,3	3,00			
19	28,3	3,00			
20	20,0	4,00			
21	19,4	4,00			
22	16,0	4,00			
23	15,0	4,00			
24	13,6	4,00			
25	10,0	4,00			
26	10,0	5,00			
27	11,6	5,00			
28	12,3	5,00			
29	13,4	5,00			
30	14,2	5,00			
31	15,0	5,00			

**Solver Parameters**

Set Target Cell:

Equal To: ☒ Max ☐ Min ☐ Value of:

By Changing Variable Cells:

Subject to the Constraints:

Рис. 6

В результате оптимизации получена линейная аппроксимирующая функция  $g(z) = \tilde{a}_0 + \tilde{a}_1 z$  (рис. 5г) со значениями<sup>15</sup>  $z_1 = 1$ ,  $z_2 = 2.18$ ,  $z_3 = 3.39$ ,  $z_4 = 4.33$ ,  $z_5 = 4.51$  и следующими регрессионными статистиками:

$a_1$	-16,507	87,174	$a_0$
$S_{a1}$	0,44	1,46	$S_{a0}$
$R^2$	0,98	3,17	$S_{ост.}$
$F_{расч.}$	1437,83	28	$n - k - 1$
$SS_{пер.}$	14448,56	281,37	$SS_{ост.}$

Сравнивая регрессионные статистики, полученные аналитической и оптимизационной процедурами решения системы (\*), можно видеть, что обе процедуры дают практически одинаковые<sup>16</sup> результаты:  $R^2 = 0,98$ ;  $S_{ост.} = 3,17$ ;  $SS_{пер.} = 14448,56$ ;  $SS_{ост.} = 281,37$ .

<sup>15</sup> Приведены значения, округленные до двух знаков после запятой.

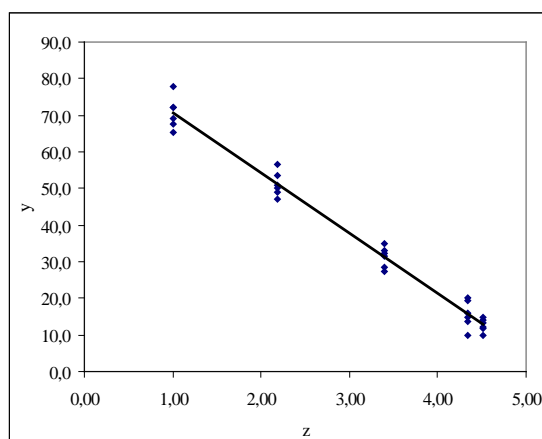


Рис. 5г

Коэффициенты регрессионных уравнений, полученных в результате сравниваемых процедур, различаются, однако это слабо отражается на значениях моделируемой функции. Например, для значения аргумента  $x_3=75$ :

при аналитическом решении-  $z_3 = 0,860$ ,  $g(z) = 73,028 - 48,611z = 31,22$

при «оптимизационном» решении -  $z_3 = 3,389$ ,  $g(z) = 87,174 - 16,507z = 31,23$

В то же время, прогнозное значение модели с линейной зависимостью от  $x_i$  (рис. 5а):  $h(x) = 55,604 - 0,128x = 46,0$  существенно отличается от среднего значения  $\bar{y}_3 = 31,23$ .

Если теперь рассмотреть зависимости  $z = f(x)$  преобразованной переменной от исходной (рис. 7), можно видеть монотонный характер огибающих этих зависимостей, как бы «зеркальных» исходной зависимости  $y = f(x)$  (рис.4) относительно оси абсцисс. В условиях малого объема рыночных данных обеспечение монотонного характера полученных зависимостей и соответствие их характера (с учетом «зеркальности») закономерностям, наблюдаемым на рынке, может служить дополнительными признаками адекватности проведенных преобразований.

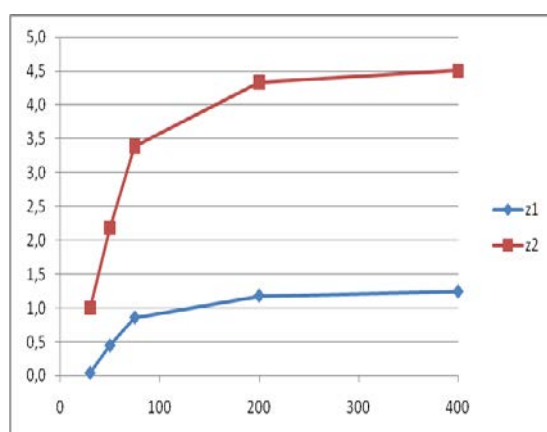


Рис. 7

<sup>16</sup> Строго говоря, сравниваемые преобразования не являются тождественными, т.к. в первом случае значения  $z_i$  определены по пяти *средним значениям* случайной величины  $\bar{y}_i$ , а во втором – по минимуму квадратов отклонений *тридцати значений* самой величины  $y_i$

*Итак, аналитическим расчетом либо оптимизационной процедурой с корректным использованием инструмента **Solver** решается задача учета в аддитивной регрессионной модели нелинейной монотонной зависимости  $y = f(x)$ , заданной на конечном множестве точек, путем адекватного преобразования  $z = z(x)$  исходной влияющей переменной.*

*При этом существенно улучшаются показатели точности регрессионной модели по сравнению с моделью без такого преобразования (линейной по  $x$ ).*

*Также очевидно, что адекватных преобразований может быть несколько в зависимости от заданных значений «свободных» переменных (начальных условий оптимизации).*

Полученные выше результаты легко обобщаются на случай, когда объясняющая переменная  $x_i$  не имеет повторяющихся значений.

Как и прежде, ищется преобразование  $z(x_i) = z_i$ , вместе с линейной функцией  $g(z) = a_0 + a_1 z$ , заданной на множестве  $\{(z_i, y_i)\}$ .

В этом случае возникает дополнительная задача разбиения исходного множества  $M = \{(z_i, y_i)\}$  на подмножества  $M_i$  так, чтобы каждое подмножество (кластер) содержало близкие значения  $z_i$ , а значения  $z_i$  разных кластеров отличались друг от друга существенно.

Эта задача обычно решается экспертом на основе анализа имеющейся выборки данных либо методами кластерного анализа.

Затем в каждом подмножестве  $M_i$  вычисляются средние значения случайных величин  $\bar{z}$ ,  $\bar{y}$ . Каждая координата точки  $(z_i, y_i) \in M_i$  представляется как сумма среднего значения случайной величины, вычисленного на данном подмножестве  $M_i$ , и отклонения  $\Delta_i^z = z_i - \bar{z}$ .

В итоге, преобразуя формулы (5)–(7), получаем требуемый результат.

Таким образом, от предыдущего случая данная ситуация отличается лишь тем, что набор  $y_{i,j}$  приписывается не единственному значению  $z_i$ , а среднему  $\bar{z}_i$  по подмножеству  $M_i$ .

Приемы и нюансы практического применения оптимизационного инструмента **Solver** при построении регрессионных моделей на малых объемах данных (выборках) требуют отдельного рассмотрения, выходящего за рамки данной публикации.

## **Литература**

1. Сивец С.А., Левыкина И.А. Эконометрическое моделирование в оценке недвижимости. – Запорожье: Полиграф, 2003.
2. Грибовский С.В., Сивец С.А. Математические методы оценки стоимости недвижимого имущества / под. ред. С.В. Грибовского, М.А. Федотовой. - М: Финансы и статистика, 2008.
3. Анисимова И.Н., Баринов Н.П., Грибовский С.В. Учет разнотипных ценообразующих факторов в многомерных регрессионных моделях оценки недвижимости - Вопросы оценки, №2, 2004. <http://www.appraiser.ru/default.aspx?SectionId=41&Id=1575>
4. Руководство пользователя. Microsoft® Excel. Версия 5.0, Корпорация Microsoft, 1993.



Опубликовано: Материалы IV Поволжской научно-практической конференции «Статистические методы массовой и индивидуальной оценки», Нижний Новгород, 31.03–01.04.2011

Размещено: <http://www.appraiser.ru/default.aspx?SectionId=35&Id=3605>