Множественная линейная регрессия. Постановка задачи. Вероятностный подход. Решение задачи. Прогнозирование.

Напомним, что *регрессионный анализ* может одновременно рассматриваться и как раздел *математической статистики*, и как раздел *эконометрики; регрессионному анализу* посвящена глава в большинстве монографий и учебников по *анализу данных*.

Постановка задачи.

Предположим, что имеется n *объектов*, каждый из которых описывается m *признаками (факторами, предикторами)*, влияющими на значение *переменной отклика*. Линейная регрессионная модель необходима нам для прогноза значений переменной отклика по известным значениям факторов.

Например, весьма актуально получение инструмента для оценки возможной *стоимости квартиры* по её основным характеристикам, таким как

- общая площадь;
- жилая площадь;
- расстояние до центра;
- расстояние до метро;
- этаж;
- этажность дома
- …

В терминах регрессионной модели,

- под *объектами* будем понимать квартиры;
- под *факторами* – перечисленные выше свойства;
- под *переменной отклика* – стоимость квартиры.

Будем нумеровать объекты индексом i $\left(i=1,\dots,n\right)$, а факторы – индексом j $\left(j=1,\dots,m\right)$.

Обозначим через x_{ij} –значение j-го признака i-го объекта $\left(i=1,\dots,n;\ \ j=1,\dots,m\right)$. Таким образом, каждый i-ый объект представим как m-мерный вектор $X_i \equiv \left(x_{i1},\dots,x_{im}\right)$, $i=1,\dots,n$.

Сырые данные для этой задачи могут иметь, например, следуюший вид:

TotalSquare (m2)	LivingSquare (m2)	DistCenter (km)	DistMetro (km)	Price
80	53	17	2,1	14 612 000,00 ₽
76	51	1	0,7	16 931 128,00 ₽
96	72	16	1,3	18 905 472,00 ₽
56	37	16	2,2	14 829 304,00 ₽
75	56	6	2,8	19 214 025,00 ₽
75	56	11	1,1	19 582 950,00 ₽
97	65	12	1,4	19 123 259,00 ₽
30	24	14	2,3	6 035 280,00 ₽
84	63	7	1,9	20 058 696,00 ₽
50	33	11	2,4	13 807 800,00 ₽
55	44	6	1,7	13 087 745,00 ₽
94	71	10	1,7	17 337 266,00 ₽
91	68	5	1,4	17 189 900,00 ₽
32	26	7	1,8	6 405 792,00 ₽
86	65	4	0,3	19 267 698,00 ₽
55	41	2	1,6	13 827 495,00 ₽
65	49	18	1,7	11 242 920,00 ₽
45	34	9	2,2	12 004 470,00 ₽
47	38	4	1,3	10 586 844,00 ₽

Здесь каждая строка соответствует одной квартире.

Обозначим значение *зависимого признака* i-го объекта (т.е. стоимости квартиры) через y_i, $i=1,\ldots,n$.

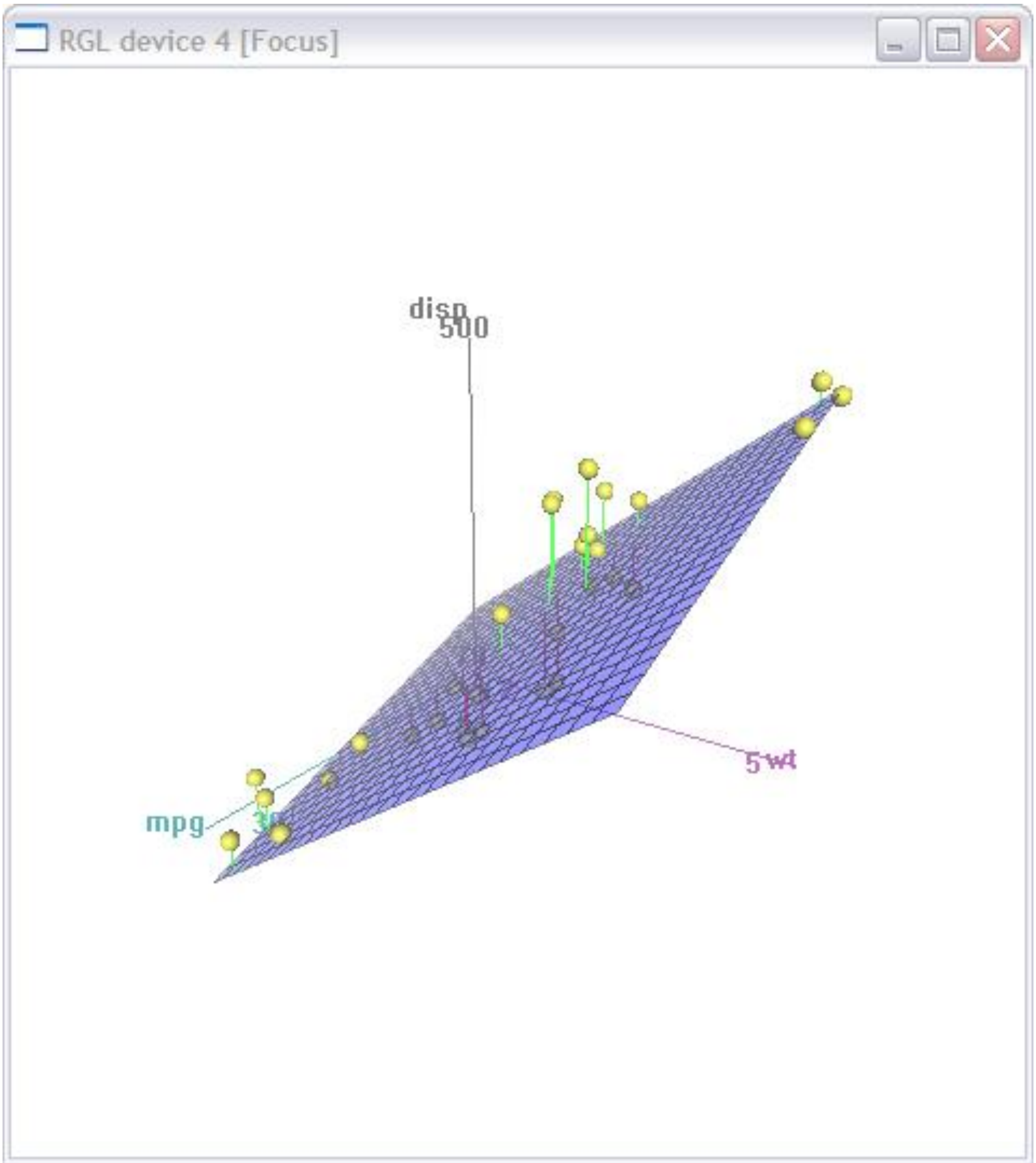
Будем искать зависимость в виде линейной функции m переменных:

$$f\left(x_1,\ldots,x_m\right)=\theta_0+\theta_1 x_1+\ldots+\theta_m x_m, \qquad (1)$$

Здесь x_j, $\left(j=1,\ldots,m\right)$ – независимые переменные (факторы, т.е. характеристики квартиры), зависимая переменная (прогнозируемая цена квартиры), $\theta_0, \theta_1,\ldots, \theta_m$ – *искомые параметры* линейной зависимости. При $m=1$ получаем рассмотренную ранее модель *однофакторной линейной регрессии*.

Геометрическая интерпретация

Геометрически можно интерпретировать задачу следующим образом. Точки с координатами $\left(x_{i1},\ldots,x_{i\,m}, y_i\right)$, $i=1,\ldots,n$, образуют «облако» в пространстве R^{m+1}. Нужно построить такую *гиперплоскость* вида (1), которая бы *наилучшим* образом проходила через это облако точек. Здесь термин «*наилучший*» означает (как и в случае однофакторной линейной регрессии) «*минимизирующий сумму квадратов отклонений фактических значений зависимой переменной от ожидаемых в соответствии с линейной моделью*». Для случая 2-х переменных линейная модель представима как плоскость в 3-хмерном пространстве:



Источник: http://www.statmethods.net/graphs/images/scatter3d.png

Решение задачи. Вероятностный подход.

Понятно, что вследствие наличия случайной компоненты точки «облака», вообще говоря, не будут находиться в одной гиперплоскости.

Обозначим случайную величину («ошибку») i-го измерения через ε_i. Будем считать, что ошибки всех n измерений распределены одинаково, а именно – закон распределения вероятности ошибки будем предполагать *нормальным* (*гауссовым*) и будем считать, что *систематическая* ошибка отсутствует (то есть «разброс» цен при одинаковых параметрах квартир вызван исключительно совокупным влиянием случайных факторов). В этом случае математическое ожидание ошибки равно нулю. Обозначим среднее квадратическое отклонение ошибки через σ. Математическая запись названных условий имеет вид: $M\varepsilon_i = 0$, $D\varepsilon_i = \sigma^2$, $i = 1,...,n$).

Запишем систему относительно искомых величин $\theta_0, \theta_1, ..., \theta_m$:

$$\begin{array}{l} y_1 = \theta_0 + \theta_1 x_{1\,1} + ... + \theta_m x_{1\,m} + \varepsilon_1 \\ y_2 = \theta_0 + \theta_1 x_{2\,1} + ... + \theta_m x_{2\,m} + \varepsilon_2 \\ .... \\ y_i = \theta_0 + \theta_1 x_{i\,1} + ... + \theta_m x_{i\,m} + \varepsilon_i \\ .... \\ y_n = \theta_0 + \theta_1 x_{n\,1} + ... + \theta_m x_{n\,m} + \varepsilon_n \end{array}, \qquad (2)$$

Введём обозначения:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1j} & ... & x_{1m} \\ 1 & x_{21} & x_{22} & ... & x_{2j} & ... & x_{2m} \\ ... & ... & ... & ... & ... & ... & ... \\ 1 & x_{i1} & x_{i2} & ... & x_{ij} & ... & x_{im} \\ ... & ... & ... & ... & ... & ... & ... \\ 1 & x_{n1} & x_{n2} & ... & x_{nj} & ... & x_{nm} \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ ... \\ \varepsilon_n \end{pmatrix}, \quad \Theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ ... \\ \theta_m \end{pmatrix}$$

и перепишем систему (2) в матричной форме:

$$Y = X\Theta + \varepsilon. \qquad (3)$$

Выразим вектор ошибок ε из системы (3):

$$\varepsilon = Y - X\Theta. \tag{4}$$

Запишем сумму квадратов ошибок (минимизируемую функцию):

$$\sum_{i=1}^{n} \varepsilon_i^2 \equiv \varepsilon^T \varepsilon = \left(Y - X\Theta\right)^T \left(Y - X\Theta\right) = \left(Y^T - \Theta^T X^T\right)\left(Y - X\Theta\right) =$$

$$= Y^T Y - Y^T X\Theta - \Theta^T X^T Y + \Theta^T X^T X\Theta = Y^T Y - 2\,\Theta^T X^T Y + \Theta^T X^T X\Theta.$$

Обозначим

$$F\left(\Theta\right) \equiv Y^T Y - 2\,\Theta^T X^T Y + \Theta^T X^T X\Theta. \tag{5}$$

Минимизируем функцию (5) по Θ. Найдём градиент функции $F\left(\Theta\right)$:

$grad\ F\left(\Theta\right) = -2X^T Y + 2X^T X\Theta$ и потребуем выполнения равенства $grad\ F\left(\Theta\right) = 0$. Получим систему:

$$X^T X\Theta = X^T Y. \tag{6}$$

Будем считать, что $\det\left(X^T X\right) \neq 0$, а значит, решение системы (6) запишется в виде:

$$\hat{\Theta} = \left(X^T X\right)^{-1} X^T Y. \tag{7}$$

Найденное решение (вектор параметров) позволяет осуществлять прогноз значений зависимого признака по известным значениям независимых признаков:

Применительно к рассмотренному примеру, получим прогнозируемое значение

$$\widehat{y_i} = X_i \hat{\Theta}, \quad i = n+1, \ldots \tag{8}$$