Классическая линейная модель множественной регрессии. МНК

Постановка задачи регрессионного анализа

Ставится задача на основе выборочных данных, представленных в виде вектора $Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$, где y_i – наблюдаемое значение результативного признака Y для i-го объекта и матрицы $X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$, где x_{ij} – наблюдаемое значение j-го объясняющего признака для i-го объекта выборочной совокупности, выявить «зависимость» результативного показателя Y от факторных (объясняющих) признаков $X_1, X_2, \ldots, X_k$.

Функция $f_Y(X_1, X_2, \ldots, X_k)$, описывающая зависимость условного среднего значения результативного признака Y от заданных значений факторных признаков $X_1, X_2, \ldots, X_k$, называется **функцией (уравнением) регрессии**, т.е. $f_Y(X_1, X_2, \ldots, X_k) = MY / X_1, X_2, \ldots, X_k$.

Линейная модель множественной регрессии

$$f_Y(X_1, X_2, \ldots, X_k) \approx \widetilde{Y} = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k ,$$

где $\beta_0, \beta_1, \ldots, \beta_k$ – коэффициенты линейного уравнения регрессии.

Система линейных уравнений $y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2} + \ldots + \beta_k x_{i_k} + \varepsilon_i$, где ε_i – регрессионный остаток, характеризующий расхождение между наблюдаемым значением y_i и «осредненным» значением $\widetilde{y}_i$, $i = \overline{1, n}$;

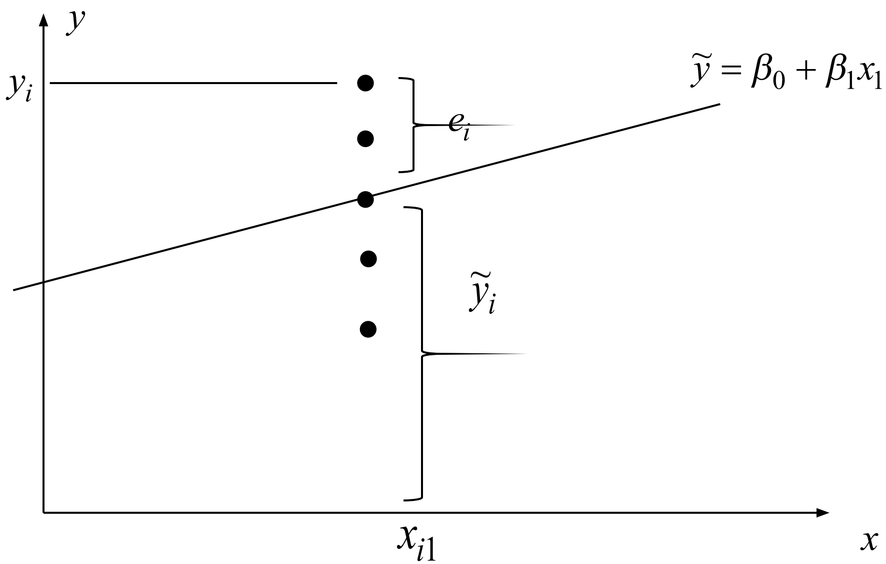
или

$$Y = X\beta + \varepsilon, \ \beta = (\beta_0, \ldots, \beta_k), \ \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T ,$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

называется **линейной моделью множественной регрессии**.

Геометрическая интерпретация функции регрессии



Классическая линейная модель множественной регрессии (КЛММР)

Линейная модель множественной регрессии, удовлетворяющая следующим пяти требованиям, называется **классической линейной моделью множественной регрессии**.

Условия Гаусса–Маркова

1) $X_1, X_2, ..., X_k$ – детерминированные переменные;
2) ранг матрицы X равен $k+1$ – среди признаков нет линейно зависимых;
3) $M\varepsilon_i = 0$, $i = \overline{1,n}$ – нет систематических ошибок в измерении Y;
4) $D\varepsilon_i = M\varepsilon_i^2 = \sigma^2$, $i = \overline{1,n}$ – гомоскедастичность регрессионных остатков (равноточные измерения);
5) $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = M(\varepsilon_i \cdot \varepsilon_j) = 0$ – условие некоррелированности регрессионных остатков, $i \neq j$, $i = \overline{1,n}$, $j = \overline{1,n}$.

Методы оценки коэффициентов КЛММР

1. Минимума отклонения наблюдаемых значений результативного признака от значений, полученных по функции регрессии: $\sum_{i=1}^{n} (y_i - \widetilde{y}_i) \to \min$.

2. Минимум сумму абсолютных величин отклонения значений результативного признака от значений, полученных по функции регрессии: $\sum_{i=1}^{n} / y_i - \widetilde{y}_i / \to \min$.

3. Минимум суммы квадратов отклонений наблюденных значений y_i от "значений" функции регрессии: $\sum_{i=1}^{n} (y_i - \widetilde{y}_i)^2 \to \min$. Критерий положен в основу метода наименьших квадратов.

4. Метод максимального правдоподобия, который может быть применен в тех случаях, когда с точностью до неизвестных значений параметров известен общий вид закона распределения вероятностей имеющихся выборочных данных.

Метод наименьших квадратов

$$F=\sum_{i=1}^{n}(y_i-b_0-b_1x_{i1}-b_2x_{i2}-\ldots-b_kx_{in})^2=(Y-X\overline{b})^T(Y-X\overline{b})=$$

$$=Y^TY-\overline{b}^TX^TY-Y^TX\overline{b}+\overline{b}^TX^TX\overline{b}=Y^TY-2\overline{b}^TX^TY+\overline{b}^TX^TX\overline{b}\to\min$$

$$2X^TX\overline{b}-2X^TY=0$$

В силу предположения о справедливости условий Гаусса-Маркова, в частности ($X=k+1$), матрица X^TX – не вырождена и получим МНК - оценки для вектора $\overline{\beta}$:

$$\overline{b}_{\text{МНК}}\equiv\overline{b}=(X^TX)^{-1}X^TY$$

оценка уравнения регрессии:

$$\hat{y}=\underset{(S_{b0})}{b_0}+\underset{(s_{b1})}{b_1}x_1+\underset{(S_{b2})}{b_2}x_2+\ldots+\underset{(S_{bk})}{b_k}x_k$$