# 25. Общая схема оценки параметров линейной множественной регрессии методом наименьших квадратов.

### 3.3.1. Линейная модель множественной регрессии

Линейная модель множественной регрессии имеет вид

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon,$$

где $Y$ - объясняемая переменная; $X_i$ - объясняющие переменные; $\beta_i$ - параметры модели; $\varepsilon$ - ошибка модели (случайные отклонения); $n$ - число наблюдений, $p$ - число факторов.

Исходными данными для оценки параметров $\beta_0, \beta_1, \ldots, \beta_p$ являются наблюдаемые значения зависимой переменной $Y$ (вектор-столбец $[n \times 1]$) и независимые переменные, объединенные в матрицу $X$ (ее размерность $[n \times (p+1)]$):

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{41} & x_{42} & \ldots & x_{4p} \end{pmatrix},$$

где $n$ - число объектов (наблюдений); $p$ - число факторов. Столбец из единиц введен для отражения в модели свободного члена уравнения регрессии $\beta_0$ Если ввести обозначения для вектора-столбца параметров $\beta$ и вектора случайных ошибок $\varepsilon$:

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

то модель множественной регрессии можно записать в матричном виде:

$$Y = X\beta + \varepsilon$$

.

Оценкой модели по выборке является уравнение:

$$\widehat{Y} = Xb + e,$$

где вместо неизвестных коэффициентов $\beta$ и ошибок $\varepsilon$ используются их оценки - вектор $b$ и фактическая ошибка - вектор $e$.

***Метод наименьших квадратов (МНК).***

**Теорема Гаусса-Маркова**. Предположим, что:
1. $Y = X\beta + \varepsilon$;
2. $X$ - детерминированная $[n \times (p+1)]$ матрица, имеющая максимальный ранг $p+1$;
3. $M(\varepsilon) = 0; \quad D(\varepsilon) = \sigma^2 I_n$.

Тогда оценка метода наименьших квадратов $b = (X^T X)^{-1} X^T Y$ является наиболее эффективной (в смысле наиманьшей дисперсии) оценкой в классе линейных (по $Y$) несмещенных оценок.

Целью МНК является выбор вектора оценок $b$, минимизирующего сумму квадратов остатков $e_i$. Обозначим $\widehat{Y} = Xb + e$ - построенную нами модель (здесь $b_i$ - оценки коэффициентов $\beta_i$, $e$ - ошибка модели). Тогда в математической форме исходное положение МНК будет выглядеть одним

из нижеуказанных способов:

$$\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \longrightarrow \min$$

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_{i1} - b_2 X_{i2} - \ldots - b_p X_{ip})^2 \longrightarrow \min$$

$$\sum_{i=1}^{n} e_i^2 \longrightarrow \min$$

Рассмотрим более подробно условия применимости МНК.

1. Матрица $X$ - детерминированная - матрица, состоящая из констант (в отличие от стохастической матрицы $X$, которая содержит случайные величины);

2. Математическое ожидание значений ошибки модели для всех наблюдений равно нулю

$$M(\varepsilon_i) = 0, \quad M(\varepsilon) = \overline{0}$$

3. Значение дисперсии ошибки является постоянной величиной для всех наблюдений $i = 1, \ldots n$ (ошибки *гомоскедастичны*)

$$D(\varepsilon_i) = \sigma_\varepsilon^2 I = const.$$

4. Значения ошибки, для различных наблюдений независимы между собой, т.е. ковариация ошибок $\varepsilon_i, \varepsilon_j (i \neq j)$ равна нулю (отсутствие *автокорреляции*)

$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0.$$

5. Значения независимых факторов модели и ошибки, рассматриваемые для одного и того же наблюдения, являются независимыми, т.е их ковариация равна нулю

$$\mathrm{cov}(X_{ij}, \varepsilon_j) = 0 \text{ для } i = 1, \ldots n.$$

6. Матрица $(X^T X)$ является невырожденной. Это означает, что факторы $X_{ij}, i = 1, \ldots n$ независимы между собой (их выборочные парные коэффициенты корреляции не превосходят некоторого порога $r^*$)

$$|r_{ij}| \leq r^*.$$

7. Случайные ошибки распределены по стандартному нормальному закону. Это последнее условие уже не связано с применимостью МНК, однако оно необходимо, поскольку позволяет получить статистические оценки значимости регрессионного уравнения и регрессионных коэффициентов.

Свойства 2 и 5 для ФАКТИЧЕСКОЙ ошибки $e$ выполняются априори при использовании МНК. В то время, как другие свойства (3, 4) должны проверяться для модели после ее построения. Нарушения условия (6) приводит к *мультиколлинеарности* (см. ниже).

**Статистические свойства МНК-оценок.**

1. Несмещенная оценка дисперсии ошибок модели $\sigma_\varepsilon^2$ определяется:

$$S_\varepsilon^2 = \frac{e^T e}{n - p - 1} = \frac{\sum e_i^2}{n - p - 1}$$

2. Дисперсия $S_{bi}^2$ коэффициента регрессии $b_i$ равна $i$-ому диагональному элементу матрицы ковариаций для коэффициентов:

$$\mathrm{cov}(b_i, b_j) = \sigma_\varepsilon^2 (X^T X)^{-1}$$

3. При выполнении предпосылок МНК оцеки коэффициентов регрессии являются несмещенными (математическое ожидание оценки равно оцениваемому параметру $M(b_i) = \beta_i$), эффективными (дисперсия $b_i$ минимальна) и состоятельными (при $n \to \infty$ она стремится по вероятности к оцениваемому параметру):

$$\lim_{n \to \infty} P(|b_i - \beta_i| < \varepsilon) = 1.$$

Иначе говоря, состоятельной называется такая оценка, которая дает точное значение для большой выборки независимо от входящих в нее конкретных наблюдений.