

**Анализ вариации результативного признака. Статистические свойства МНК-оценок.**

# Статистические свойства оценок коэффициентов, полученных МНК

$$\bar{b}_{\text{МНК}} \equiv \bar{b} = (X^T X)^{-1} X^T Y$$

$$b = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\bar{\beta} + \bar{\varepsilon}) = \bar{\beta} + (X^T X)^{-1} X^T \bar{\varepsilon}.$$

1) МНК – оценка  $\bar{b}$  является несмещенной оценкой вектора  $\bar{\beta}$

$$b = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\bar{\beta} + \bar{\varepsilon}) = \bar{\beta} + (X^T X)^{-1} X^T \bar{\varepsilon}$$

$$M\bar{b} = M(\bar{\beta} + (X^T X)^{-1} X^T \bar{\varepsilon}) = \bar{\beta} + (X^T X)^{-1} X^T M\bar{\varepsilon} = \bar{\beta}$$

2) Свойство состоятельности в данном случае определяется структурой матрицы  $X$ : наименьшее собственное число матрицы  $X^T X$  стремится к  $\infty$ , при  $n \rightarrow \infty$ .

3) Оценки считаются эффективными, если они характеризуются наименьшей дисперсией. Эффективность МНК-оценок доказывается в предположении о нормальности регрессионных остатков ММП.

# Оптимальность МНК-оценок КЛММР

а) Пусть  $\theta$  - скалярный параметр

*df 1* Оценка  $\hat{\theta}_1$  скалярной величины  $\theta$  точнее (лучше, эффективнее), чем оценка  $\hat{\theta}_2$ , если  $M(\hat{\theta}_1 - \theta)^2 < M(\hat{\theta}_2 - \theta)^2$

*df 2* Оценка параметра  $\theta$  является оптимальной (эффективной) в классе  $M$  (обозначим ее  $\hat{\theta}_{opt}$ ), если

$$M(\hat{\theta}_{opt} - \theta)^2 = \min_{\hat{\theta} \in M} M(\hat{\theta} - \theta)^2 \quad (1)$$

б) Пусть  $\beta = (\beta_0, \beta_1, \dots, \beta_e)$  - вектор коэффициентов КЛММР  
 $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_e)$  - вектор оценок коэффициентов КЛММР

Оптимальность (качество) векторной оценки – как это понимать?

Рассмотрим функционал  $C^T \beta$

$$(1)$$

Пример 1:  $C^T = (0, 0, \dots, 0, 1, 0, \dots, 0)$ , где  $l \in (0, 1, \dots, k)$

$$(l)$$

$$C^T \beta = \beta_l$$

Пример 2:  $C^T = (1, x_1, \dots, x_k)$

$$C^T \beta = \beta_0 \cdot 1 + \beta_1 x_1 + \dots + \beta_k x_k = y(x)$$

Вывод: Функционал (1) позволяет говорить об оптимальности вектора  $\theta$  в смысле (1)

# Теорема Гаусса-Маркова

Рассматривается задача статистического оценивания заданной (с помощью  $C$ ) линейной функции  $\theta_C = C^T \theta$  от известных параметров МЛММР. Пусть  $M = \{B^T Y\}$  - класс линейных относительно  $(y_1, y_2, \dots, y_n)$  и несмещенных оценок параметра  $\theta_C$ .

( $B = (b_1, \dots, b_n)^T$  - вектор коэффициентов, с помощью которого формируется класс  $M$ ).

Тогда оценка  $C^T \hat{\theta}_{МНК}$  является оптимальной в классе  $M$  оценкой параметра  $\theta_C$  в смысле (1), т.е.

$$M(C^T \hat{\theta}_{МНК} - \theta_c)^2 \leq M(B^T Y - \theta_c) \quad \forall \text{ любой оценки } B^T Y \in M$$

# Доказательство

$$1) C^T \hat{\theta}_{МНК} \in M : (C^T \hat{\theta}_{МНК} = \underbrace{C^T}_{B^T} \underbrace{(X^T X)^{-1}}_{X^T} X^T Y \sim B^T Y)$$

$$2) M C^T \hat{\theta}_{МНК} = C^T M \hat{\theta}_{МНК} = C^T \theta = \theta_C - \text{несмещенная}$$

$$3) M B^T Y = |\text{должно быть}| = \theta_C \Rightarrow M B^T Y = B^T X \theta = C^T \theta \Rightarrow B^T X = C^T \quad (2)$$

$$M (B^T Y - C^T \theta)^2 = \sigma^2 (B^T B) \quad (3)$$

$$M (C^T \hat{\theta}_{МНК} - C^T \theta)^2 = M [C^T (\hat{\theta}_{МНК} - \theta)]^2 = \sigma^2 C^T (X^T X)^{-1} C = \sigma^2 B^T X (X^T X)^{-1} X^T B \quad (4)$$

$$M (B^T Y - C^T \theta)^2 - M (C^T \hat{\theta}_{МНК} - C^T \theta)^2 = \sigma^2 B^T (E_n - X (X^T X)^{-1} X^T) B = \sigma^2 B^T Z B \geq 0 \quad (5)$$

ч.т.д.

# Свойства оценок нормальной КЛММР

$$\left\{ \begin{array}{l} Y = X\beta + \varepsilon \\ \varepsilon \in N(0, \sigma^2 E_n) \\ x_1, \dots, x_k - \text{нечслучайные переменные} \\ \text{ранг } X = k + 1 \end{array} \right.$$

Теорема: 1) Оценки  $\hat{\beta}_{МНК} \in N(\beta, \Sigma_{\hat{\beta}}) \equiv N(\beta, \sigma^2 (X^T X)^{-1})$

2)  $(n - k - 1) \frac{\hat{\sigma}^2}{\sigma^2} \in \chi^2(n - k - 1)$

3) Оценки  $\hat{\beta}_{МНК}$  и  $\hat{\sigma}^2$  статистически независимы

Следствие:

$$t = \frac{\hat{\beta}_i - \beta_i}{S_{\hat{\beta}_i}} \in t(n - k - 1)$$

# Ковариационная матрица вектора оценок коэффициентов

Оценим ковариационную матрицу случайного вектора  $\bar{b}$ , при условии выполнения 1, 3, 4, 5 условий Гаусса-Маркова

$$\begin{aligned} \sum_{\bar{b}} &= M[(\bar{b} - M\bar{b})(\bar{b} - M\bar{b})^T] = M[((X^T X)^{-1} X^T \bar{\varepsilon})((X^T X)^{-1} X^T \varepsilon)^T] = \\ &= M[(X^T X)^{-1} X^T \bar{\varepsilon} \bar{\varepsilon}^T X (X^T X)^{-1}] = (X^T X)^{-1} X^T M(\bar{\varepsilon} \bar{\varepsilon}^T) X (X^T X)^{-1} = \\ &= (X^T X)^{-1} X^T \sigma^2 \varepsilon_n X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1} \end{aligned}$$

Откуда, в частности,

$$D_{bj} = \sigma^2 [(X^T X)^{-1}]_{jj}$$

Несмещенная оценка для  $\sigma^2$

$$\begin{aligned} \hat{S}^2 &= \frac{1}{n-k-1} (Y - X\bar{b})^T (Y - X\bar{b}) \\ \hat{\Sigma}_b &= \hat{S}_{i\tilde{n}\tilde{o}}^2 (\tilde{O}^{\tilde{O}} \tilde{O})^{-1} \end{aligned}$$



# Анализ вариации результативного признака

В качестве характеристики степени рассеивания случайной величины  $Y$  относительно функции регрессии используется в случае нелинейной связи корреляционное отношение:

$$\rho_{Y/X_1, \dots, X_k}^2 = 1 - \frac{M(Y - f_Y(X_1, X_2, \dots, X_k))^2}{\sigma_Y^2} = \frac{M(f_Y(X_1, X_2, \dots, X_k) - MY)^2}{\sigma_Y^2},$$

которое характеризует качество подгонки функции регрессии под выборочные данные. В случае линейной регрессии  $\rho_{Y/X_1, \dots, X_k}^2$  называется коэффициентом детерминации  $R_{Y/X_1, \dots, X_k}^2 \equiv R^2$ .

Коэффициент детерминации строится из тех соображений, что общая дисперсия результативного признака складывается из факторной и остаточной дисперсий:

$$\sigma_Y^2 = \sigma_{\text{факт}}^2 + \sigma_{\text{ост}}^2,$$

где  $\sigma_Y^2$  – дисперсия результативного признака;

$\sigma_{\text{факт}}^2 = M(f_Y(X_1, \dots, X_k) - MY)^2$  – факторная дисперсия;

$\sigma_{\text{ост}}^2 = M(Y - f_Y(X_1, \dots, X_k))^2$  – остаточная дисперсия.

# Выборочный коэффициент детерминации

Оценка коэффициента детерминации рассчитывается по формуле:

$$\hat{R}_{y/x_1, \dots, x_n}^2 = \frac{Q_{\hat{y}}}{Q_{\hat{y}}} = 1 - \frac{Q_{\text{ост}}}{Q_{\hat{y}}} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{Y})^2}$$

$$Q_{\hat{y}} = \sum_{i=1}^n (\hat{y}_i - \bar{Y})^2$$

$$Q_{\text{ост}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \equiv \sum_{i=1}^n e_i^2, \text{ где } e_i = y_i - \hat{y}_i - \text{ оценка регрессионного}$$

остатка для  $i$ -го наблюдения.

Несмещенная оценка коэффициента детерминации имеет вид:

$$\hat{R}_{y/x_1, \dots, x_n}^{*2} \approx 1 - (1 - \hat{R}_{y/x_1, \dots, x_n}^2) \frac{n-1}{n-k-1}$$