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 $\bar{\beta}$

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

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

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

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 θ - скалярный параметр

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

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

$$M\left(\hat{\theta}_{опт} - \theta\right)^2 = \min_{\hat{\theta} \in M} M\left(\hat{\theta} - \theta\right)^2 \qquad (1)$$

б) Пусть $\beta = (\beta_0, \beta_1, ..., \beta_{\hat{e}})$ - вектор коэффициентов КЛММР

$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{\hat{e}})$ - вектор оценок коэффициентов КЛММР

Оптимальность (качество) векторной оценки – как это понимать?

Рассмотрим функционал $C^T \beta$

(1)

Пример 1: $C^T = (0, 0, ..., 0, \underset{(l)}{1}, 0, ..., 0)$, *где* $l \in (0, 1, ..., k)$

$C^T \beta = \beta_l$

Пример 2: $C^T = (1, x_1, ..., x_k)$

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

Вывод: Функционал (1) позволяет говорить об оптимальности вектора θ в смысле (1)

Теорема Гаусса-Маркова

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

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

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

$$M\left(C^T\hat{\theta}_{мнк} - \theta_c\right)^2 \le M\left(B^T Y - \theta_c\right) \quad \forall \textit{ любой оценки } B^T Y \in M$$

Доказательство

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

2) $MC^T\hat{\theta}_{МНК} = C^TM\hat{\theta}_{МНК} = C^T\theta = \theta_C - \textit{несмещенная}$

3) $MB^TY = \left|\textit{должно быть}\right| = \theta_C \Rightarrow MB^TY = B^TX\theta = C^T\theta \Rightarrow B^TX = C^T \qquad (2)$

$M\left(B^TY - C^T\theta\right)^2 = \sigma^2\left(B^TB\right) \qquad (3)$

$M\left(C^T\theta_{МНК} - C^T\theta\right)^2 = M\left[C^T\left(\hat{\theta}_{МНК} - \theta\right)\right]^2 = \sigma^2C^T(X^TX)^{-1}C = \sigma^2B^TX(X^TX)^{-1}X^TB \qquad (4)$

$M\left(B^TY - C^T\theta\right)^2 - M\left(C^T\hat{\theta}_{МНК} - C^T\theta\right)^2 = \sigma^2B^T\left(E_n - X(X^TX)^{-1}X^T\right)B = \sigma^2B^TZB \geq 0 \qquad (5)$

Ч.Т.Д.

Свойства оценок нормальной КЛММР

$$\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon \in N\left(0, \sigma^2 E_n\right) \\ x_1, \ldots, x_k - \textit{неслучайные переменные} \\ \textit{ранг } X = k+1 \end{cases}$$

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

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

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

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

Откуда, в частности,

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

Несмещенная оценка для σ^2

$$\hat{S}^2 = \frac{1}{n-k-1} (Y - X\bar{b})^T (Y - X\bar{b})$$
$$\hat{\Sigma}_b = \hat{S}^2_{\text{íñò}} (\tilde{O}^{\dot{O}} \tilde{O})^{-1}$$

Анализ вариации результативного признака

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

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

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

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

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

где σ_Y^2 – дисперсия результативного признака;

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

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

Выборочный коэффициент детерминации

Оценка коэффициента детерминации рассчитывается по формуле:

$$\hat{R}^2_{y/x_1,\dots,x_n} = \frac{Q_{\hat{o}\,\grave{a}\hat{e}\grave{o}}}{Q_{\hat{i}\,\acute{a}\grave{u}}} = 1 - \frac{Q_{\hat{i}\,\tilde{n}\grave{o}}}{Q_{\hat{i}\,\acute{a}\grave{u}}} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{Y})^2}$$

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

$$Q_{\hat{i}\,\tilde{n}\grave{o}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \equiv \sum_{i=1}^{n} e_i^2$$, где $e_i = y_i - \hat{y}_i$ - оценка регрессионного остатка для i-го наблюдения.

Несмещенная оценка коэффициента детерминации имеет вид:

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