

Семинар 12.

Ошибки спецификации модели.

1. (Включение лишних переменных) Пусть процесс, порождающий данные, имеет вид:

$$y = X\beta + \varepsilon. \quad (1)$$

Модель, которую мы оцениваем:

$$y = X\beta + Z\gamma + \varepsilon. \quad (2)$$

Здесь X — $n \times k$ матрица, Z — $n \times l$ матрица, y — $n \times 1$ вектор, β — $k \times 1$ вектор, γ — $l \times 1$ вектор, ε — $n \times 1$ вектор.

- (а) Будет ли МНК-оценка вектора параметров β несмещённой?
- (б) Что произойдёт с оценкой ковариационной матрицы $\widehat{\text{Var}}(\hat{\beta})$?
- (в) Будет ли несмещённой МНК-оценка дисперсии случайной ошибки σ^2 ?

Решение:

- (а) Вычислим оценку вектора β по модели (2) (достаточно вспомнить формулу из Задачи 3 из КР-1).

$$\hat{\beta} = (X'M_zX)^{-1}X'M_zy, \text{ где } M_z = I - Z(Z'Z)^{-1}Z'.$$

Проверим, является ли данная МНК-оценка несмещённой:

$$\begin{aligned} \mathbb{E}(\hat{\beta}) &= \mathbb{E}[(X'M_zX)^{-1}X'M_zy] = \mathbb{E}[(X'M_zX)^{-1}X'M_z(X\beta + \varepsilon)] = \\ &= (X'M_zX)^{-1}X'M_zX\beta + (X'M_zX)^{-1}X'M_z\mathbb{E}(\varepsilon) = \beta. \end{aligned}$$

Следовательно, при включении лишних переменных МНК-оценка вектора параметров β остаётся несмещённой.

- (б) Рассчитаем ковариационную матрицу для оценки $\hat{\beta}$:

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \text{Var}[(X'M_zX)^{-1}X'M_zy] = \text{Var}[(X'M_zX)^{-1}X'M_z(X\beta + \varepsilon)] = \\ &= (X'M_zX)^{-1}X'M_z\text{Var}(X\beta + \varepsilon)M_zX(X'M_zX)^{-1} = \\ &= (X'M_zX)^{-1}X'M_z\text{Var}(\varepsilon)M_zX(X'M_zX)^{-1} = \sigma^2(X'M_zX)^{-1}X'M_zM_zX(X'M_zX)^{-1} = \\ &= \sigma^2(X'M_zX)^{-1}. \end{aligned}$$

Для истинной модели ковариационная матрица для МНК-оценки вектора параметров β имеет вид:

$$\text{Var}(\hat{\beta}^{true}) = \sigma^2(X'X)^{-1}.$$

Сравним данные ковариационные матрицы, рассчитав разницу между ними

$$\text{Var}(\hat{\beta}^{true}) - \text{Var}(\hat{\beta}).$$

Вместо разности выше рассмотрим разность

$$\begin{aligned} \left[\text{Var}(\hat{\beta}) \right]^{-1} - \left[\text{Var}(\hat{\beta}^{true}) \right]^{-1} &= \frac{1}{\sigma^2}(X'M_zX) - \frac{1}{\sigma^2}(X'X) = \frac{1}{\sigma^2}(X'M_zX - X'X) = \\ &= \frac{1}{\sigma^2}(X'(I - P_z)X - X'X) = \frac{1}{\sigma^2}(X'X - X'P_zX - X'X) = -\frac{1}{\sigma^2}(X'P_zX). \end{aligned}$$

Здесь $(X'P_zX)$ — положительно полуопределенная матрица. Тогда

$$\text{Var}(\hat{\beta}) - \text{Var}(\hat{\beta}^{true})$$

является положительно полуопределенной матрицей, что означает, что дисперсии оценок параметров не могут уменьшиться.

(в) Нам известно, что оценка дисперсии ошибки в модели (2) равна:

$$\hat{\sigma}^2 = \frac{RSS}{n - k - l}.$$

Проверим, будет ли она несмещённой:

$$\mathbb{E} \left(\frac{RSS}{n - k - l} \right) = \frac{1}{n - k - l} \mathbb{E}(RSS).$$

Запишем RSS в модели (2) в матричном виде:

$$RSS = y'M^*y,$$

где M^* — матрица оператор ортогонального проектирования на подпространство, образованное X и Z , то есть $M^*X = M^*Z = 0$.

$$RSS = y'M^*y = (X\beta + \varepsilon)'M^*(X\beta + \varepsilon) = \varepsilon'M^*\varepsilon \text{ так как } M^*X = 0.$$

Обозначим через $X^* = [X \ Z]$ матрицу размерности $n \times (k + l)$, содержащую все объясняющие показатели. Тогда

$$\mathbb{E}(RSS) = \mathbb{E}(\varepsilon'M^*\varepsilon) = \mathbb{E}(\text{tr}(\varepsilon'M^*\varepsilon)) = \mathbb{E}(\text{tr}(\varepsilon'M^*\varepsilon)) = \mathbb{E}(\text{tr}(\varepsilon\varepsilon'M^*)) =$$

$$\begin{aligned}
 &= \text{tr}(M^* \mathbb{E}(\varepsilon \varepsilon')) = \sigma^2 \text{tr}(M^*) = \sigma^2 \text{tr}(I - X^*(X^{*'}X^*)^{-1}X^{*'}) = \\
 &= \sigma^2 \text{tr}(I_n) - \sigma^2 \text{tr}(X^*(X^{*'}X^*)^{-1}X^{*'}) = \sigma^2 n - \sigma^2 \text{tr}(X^{*'}X^*(X^{*'}X^*)^{-1}) = \\
 &= \sigma^2 n - \sigma^2 \text{tr}(I_{k+l}) = \sigma^2(n - k - l).
 \end{aligned}$$

Для вывода $\mathbb{E}(RSS)$ мы воспользовались тем, что $\varepsilon' M^* \varepsilon$ — скаляр, который можно рассматривать как матрицу размерности 1×1 , след которой и есть этот скаляр. Затем использовали свойство следа $\text{tr}(A \cdot B) = \text{tr}(B \cdot A)$. Таким образом, получаем:

$$\mathbb{E}\left(\frac{RSS}{n - k - l}\right) = \frac{1}{n - k - l} \mathbb{E}(RSS) = \sigma^2,$$

то есть оценка дисперсии ошибки является несмещённой при включении в модель лишних переменных.

2. (Исключение существенных переменных) Прodelали то же самое, что и в задании 1 (см. лекцию).
3. (Исключение существенных переменных) Дана стандартная модель парной регрессии

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

- (а) Чему равна МНК-оценка коэффициента β_2 при ограничении $\beta_1 = 0$.
- (б) Чему равна дисперсия оценки в пункте (а)? Покажите, что она меньше, чем $\sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2$ — дисперсия МНК-оценки β_2 в регрессии без ограничения. Противоречит ли это теореме Гаусса–Маркова?

Решение:

- (а) GDP (истинный процесс): $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \quad i = 1, \dots, n.$

Модель, которую оцениваем: $y_i = \beta_2 x_i + \varepsilon_i, \quad i = 1, \dots, n.$

Найдем МНК-оценку для нашей модели:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}.$$

Как известно, при пропуске существенных переменных (в нашем случае пропущена константа) МНК-оценки смещены. Убедимся в этом:

$$\begin{aligned}
 \mathbb{E}(\hat{\beta}_2) &= \mathbb{E}\left(\frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}\right) = \mathbb{E}\left(\frac{\sum_{i=1}^n (\beta_1 + \beta_2 x_i + \varepsilon_i) x_i}{\sum_{i=1}^n x_i^2}\right) = \\
 &= \beta_1 \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2} + \beta_2 \neq \beta_2.
 \end{aligned}$$

Таким образом, МНК–оценка действительно смещённая.

(б) Вычислим дисперсию данной оценки:

$$\begin{aligned}\text{Var}(\hat{\beta}_2) &= \text{Var}\left(\frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}\right) = \text{Var}\left(\frac{\sum_{i=1}^n \varepsilon_i x_i}{\sum_{i=1}^n x_i^2}\right) = \frac{1}{(\sum_{i=1}^n x_i^2)^2} \sum_{i=1}^n x_i^2 \text{Var}(\varepsilon_i) = \\ &= \frac{1}{(\sum_{i=1}^n x_i^2)^2} \sum_{i=1}^n x_i^2 \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.\end{aligned}$$

Сравним данную дисперсию с дисперсией МНК–оценки параметра β_2 для истинной модели (обозначим эту оценку как β_2^{true}) которая, как нам известно, имеет вид:

$$\text{Var}(\beta_2^{true}) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Сравним знаменатели:

$$\sum_{i=1}^n x_i^2 - \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i^2 + n(\bar{x})^2 = n(\bar{x})^2 \geq 0.$$

Следовательно, МНК–оценка в модели с пропущенной константой имеет меньшую дисперсию. Однако это не противоречит теореме Гаусса–Маркова. Согласно теореме Гаусса–Маркова МНК–оценка β_2^{true} в истинной модели действительно имеет минимальную дисперсию, однако, в классе линейных по y и несмещённых оценок, но МНК–оценка β_2 в модели с пропущенной константой является смещённой.