Конспектировал: Ефим Лубошников.

1. Свойства математического ожидания

Вспомним, какими свойствами обладает математическое ожидание случайной величины и случайного вектора. Обозначим: y_1, y_2 — случайные величины, y, z — случайные вектора, a — скалярная константа, A, B — матрицы констант.

Случайная величина	Случайный вектор
$\mathrm{E}(ay_1) = a\,\mathrm{E}(y_1)$	$\mathrm{E}(AyB) = A\,\mathrm{E}(y)B$
$\mathrm{Var}(y_1) = \mathrm{E}(y_1^2) - \mathrm{E}(y_1)^2$	$\mathrm{Var}(y) = \mathrm{E}(yy^T) - \mathrm{E}(y)\,\mathrm{E}(y^T)$
$\mathrm{Var}(ay_1) = a^2\,\mathrm{Var}(ay_1)$	$\mathrm{Var}(Ay) = A(\mathrm{E}(yy^T) - \mathrm{E}(y)\,\mathrm{E}(y^T))A^T$
$\mathrm{Var}(a + y_1) = \mathrm{Var}(y_1)$	$\mathrm{Var}(a + y) = \mathrm{Var}(y)$
$\mathrm{Cov}(y_1, y_2) = \mathrm{E}(y_1y_2) - \mathrm{E}(y_1)\,\mathrm{E}(y_2)$	$\mathrm{Cov}(y, z) = \mathrm{E}(yz^T) - \mathrm{E}(y)\,\mathrm{E}(z^T)$
	$\mathrm{Cov}(Ay, Bz) = A\,\mathrm{Cov}(y, z)B^T$
	$\mathrm{Cov}(y, z) = \mathrm{Cov}(z, y)^T$
	Если $y, z \in \mathbb{R}^n$, то справедливо:
	$\mathrm{Cov}(y + z, w) = \mathrm{Cov}(y, w) + \mathrm{Cov}(z, w)$
	$\mathrm{Var}(y + z) = \mathrm{Var}(y) + \mathrm{Var}(z) + \mathrm{Cov}(y, z) + \mathrm{Cov}(z, y)$

Так выглядит дисперсия случайного вектора $y \in \mathbb{R}^n$:

$$\mathrm{Var}(y) = \begin{bmatrix} \mathrm{Var}(y_1) & \mathrm{Cov}(y_1, y_2) & \dots & \mathrm{Cov}(y_1, y_n) \\ \mathrm{Cov}(y_2, y_1) & \mathrm{Var}(y_2) & \dots & \mathrm{Cov}(y_2, y_n) \\ \vdots & \vdots & \dots & \vdots \\ \mathrm{Cov}(y_n, y_1) & \mathrm{Cov}(y_n, y_2) & \dots & \mathrm{Var}(y_n) \end{bmatrix}$$

2. Предпосылки для теоремы Гаусса-Маркова

Рассмотрим задачу регрессии $y = X\beta + u$ в случае двух регрессоров: $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + u_i$ Предпосылки:

1. Предполагаем β — неизвестная константа.
2. X должен иметь полный ранг, чтобы оценки МНК существовали. X может быть:
   а) известная константа, $y_t = \beta_1 + \beta_2 t + u_t$
   б) наблюдение - случайная величина, временной ряд.
3. Гомоскедастичность: $\mathrm{Var}(u_i|X) = \sigma^2 I$
4. Отсутствие автокорреляции: $\mathrm{Cov}(u_i, u_j|x) = 0, i \neq j$
5. $\mathrm{E}(u|X) = 0$

Обозначим:

- $\hat{y}$ — предсказанный y на тренировочной выборке.
- $\hat{y}_{test}$ — предсказанный y на тестовой выборке.

2.1. Упражнение "Мега-матрица"

Задание: Заполнить матрицу Var, Cov для соответсвующих элементов матрицы:

	y	$\hat{y}$	u	$\hat{u}$	$\hat{\beta}$
y					
$\hat{y}$					
u					
$\hat{u}$					
$\hat{\beta}$					
y_{test}					
$\hat{y}_{test}$					
$\hat{u}_{test}$					

Решение:
Будем пользоваться результатами, полученными ранее:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$
$$\hat{y} = X\hat{\beta}$$
$$\hat{u} = y - \hat{y}$$
$$y_{test} = X_{test}\beta + u_{test}$$
$$\hat{y}_{test} = X_{test}\hat{\beta}$$

Предварительно найдём математические ожидания:

$$\mathrm{E}(y|X) = \mathrm{E}(X\beta + u|X) = X\beta + \mathrm{E}(u|X) = X\beta$$

$$\mathrm{E}(\beta|X) = \mathrm{E}((X^T X)^{-1} X^T y|X) = \mathrm{E}((X^T X)^{-1} X^T X\beta|X) = (X^T X)^{-1} X^T X\beta = \beta$$

Получим элементы мега-матрицы:

$$\mathrm{Var}(y|X) = \mathrm{Var}(X\beta + u|X) = \mathrm{Var}(u|X) = \sigma^2 I$$

$$\mathrm{Var}(\beta|X) = \mathrm{Var}((X^T X)^{-1} X^T y|X) = (X^T X)^{-1} X^T \mathrm{Var}(y|X)((X^T X)^{-1} X^T)^T = \\ (X^T X)^{-1} X^T \sigma^2 X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$$

$$\mathrm{Cov}(\hat{\beta}, \hat{u}|X) = \mathrm{Cov}((X^T X)^{-1} X^T y, y - \hat{y}|X) = \mathrm{Cov}((X^T X)^{-1} X^T y, y|X) - \mathrm{Cov}((X^T X)^{-1} X^T y, \hat{y}|X) = \\ (X^T X)^{-1} X^T \mathrm{Var}(y|X) - \mathrm{Cov}(\beta, X\beta|X) = \sigma^2 (X^T X)^{-1} X^T - \sigma^2 (X^T X)^{-1} X^T = 0$$

На контрольной работе - вывести 2 элемента мега-матрицы.

2.2. Математическое ожидание RSS

RSS можно представить как:

$$RSS = \sum_{i=1}^{n} \hat{u_i}^2 = \hat{u}^T\hat{u} = ((I-H)y)^T((I-H)y) = y^T(I-H)^T(I-H)y = y^T(I-H)y$$

Далее, найдем математическое ожидание RSS при фиксированном X:

$$\mathrm{E}(RSS|X) = \mathrm{E}(y^T(I-H)y|X) = \mathrm{E}(tr(y^T(I-H)y|X)) = \mathrm{E}(tr(I-H)yy^T|X)) = tr((I-H)\,\mathrm{E}(yy^T)) =$$
$$\mathrm{trace}(I-H)(\mathrm{Var}(y|X) + \mathrm{E}(y|X)\,\mathrm{E}(y^T|X)) = tr(I-H)(\sigma^2 I + X\beta\beta^T X^T)) = tr[(I-H)(\sigma^2 I + X\beta\beta^T X^T))] =$$
$$\mathrm{trace}(I-H)\sigma^2 + tr((I-H)(X\beta\beta^T X^T)) = tr(I-H)\sigma^2 = (n-k)\sigma^2 = (n-k)\sigma^2$$

Слагаемое $(I-H)X = 0$, так как $(I-H)$ — проекция на ортогональное дополнение к X. Можем сделать вывод, что оценка $\hat{\sigma}^2 = \frac{RSS}{n-k}$ — несмещённая.

3. Теорема Гаусса-Маркова

Теорема 3.1. Если:

- Выполняются предпосылки про β, u, X;
- Помимо МНК-оценки $\hat{\beta} = (X^TX)^{-1}X^Ty$ существует альтернативная несмещённая оценка $\hat{\beta}_{alt} = A_{alt}^T y$

то можно утверждать, что:

- $\mathrm{Var}(\hat{\beta}_{alt_j}|X) \geq \mathrm{Var}(\hat{\beta}_j|X)$
- Матрица $Delta = \mathrm{Var}(\hat{\beta}_{alt}|X) - \mathrm{Var}(\hat{\beta}|X)$ являетеся неотрицатиельно положительно определенной.

Замечание.
Первая формулировка является частным и упрощенным вариантом второй формулировки.

Доказательство. $\hat{\beta} = ((X^TX)^{-1}X^T)^Ty = A^Ty$
$\hat{\beta}_1$ — первая строка из A^T, умноженная на y. Также $\hat{\beta}_1 = \langle$первый столбец в $A, y\rangle$. Матрица A задаёт веса, с которыми надо брать y, чтобы получить $\hat{\beta}$. Рассмотрим, каким образом собрана матрица A:

$$A = X \cdot (X^TX)^{-1}$$

Для этого обозначим столбцы матрицы A векторами c_1, c_1, ..., c_k:

$$A = \begin{pmatrix} \vdots & \vdots & \dots & \vdots \\ c_1 & c_2 & \dots & c_k \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \cdot \begin{pmatrix} * & \dots \\ * & \dots \\ \vdots & \dots \\ * & \dots \end{pmatrix}$$

Первый столбец матрицы A формируется, как линейная комбинация столбцов c_i:
$a_1 = * \cdot c_1 + * \cdot c_2 + ... + * \cdot c_k$

Веса, с которыми надо брать y, чтобы получить $\hat{\beta}$, являются линейной комбинацией столбцов матрицы X.

Дисперсия $\mathrm{Var}(\hat{\beta}_1|X) = \mathrm{Var}(a_1^T y|X) = a_1^T \sigma^2 I a_1 = \sigma^2 a_1^T a_1 = \sigma^2 \cdot \left\| a_1 \right\|^2$
$\mathrm{Var}(\hat{\beta}_1^{alt}|X) = \sigma^2 \left\| a_1^{alt} \right\|^2 \geq \mathrm{Var}(\hat{\beta}_1|X)$

Краткая схема доказательства:

1. $a_1 \in \mathrm{Lin}(\text{столбцы } X)$
   Это следует из $A = X(X^T X)^{-1}$

2. $a_1^{alt} - a_1 \perp \mathrm{Lin}(\text{столбцы } X)$
   Это следует из $(a_1^{alt})^T X\beta = \beta_1 = a_1^T X\beta$

Рассмотрим доказательство для общего случая:

Введём матрицу $\Delta = \mathrm{Var}(\hat{\beta}^{alt}|X) - \mathrm{Var}(\hat{\beta}|X)$. Нам нужно доказать, что матрица Δ неотрицательно определённая. Для этого для любого вектора w нужно, чтобы $w^T \Delta w \geq 0$. Или, эквивалентно:

$$\mathrm{Var}(\omega^T \hat{\beta}_{alt}|X) \geq \mathrm{Var}(\omega^T \hat{\beta}|X)$$

$$\mathrm{Var}(a_{alt}^T Y|X) \geq \mathrm{Var}(a^T y|X)$$

Далее доказательство, как и для частного случая:

1. $a \in \mathrm{Lin}(\text{столбцы } X)$

2. $a^T - a \perp \mathrm{Lin}(\text{столбцы } X)$

□

4. Теоремы без доказательства

Если верны предпосылки о β, X, u, и известно, что $u \sim N(0; \sigma^2 I)$, то:

1. $\hat{\beta}_j|X \sim N(\beta_j, \mathrm{Var}(\hat{\beta}_j|X))$

2. $\frac{\hat{\beta}_j - \beta_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j|X)}} = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k}$

3. Если существуют две вложенные модели: «short» — частный случай «long», то:

$$\frac{(RSS_s - RSS_l)/(k_l - k_s)}{RSS_l/(n - k_l)} \sim F_{k_l - k_s, n - k_l}$$

4.1. Задача

Дано: 200 наблюдений.
Модель A: $\hat{y}_i = \underset{(1.1)}{2.7} - \underset{(0.6)}{3.2} x_i, \quad R^2 = 0.6$
Модель B: $\hat{y}_i = \underset{(1.2)}{3.2} - \underset{(0.7)}{2.8} x_i + \underset{(1.1)}{4.2} z_i + \underset{(0.5)}{0.1} x_i^2, \quad R^2 = 0.7$
В скобках ниже указаны стандартные ошибки, standard errors, $se(\hat{\beta})$

Задание:

1. Проверьте гипотезу при уровне значимости $\alpha = 5\%$

   H_0: верна модель А.
   H_A: модель А неверна, но верна модель B.

2. Постройте 95% доверительный интервал для β_x.

Решение

Из теоремы 2:

$$\frac{\hat{\beta}_x - \beta_x}{se(\hat{\beta}_x)} \sim t_{200-2}$$

Наблюдений много, поэтому t-распределение с 198 степенями свободы очень похоже на нормальное.
$t_{198} = N(0;1)$
$-t_{cr} \leq \frac{\hat{\beta}_x - \beta_x}{se(\hat{\beta}_x)} \leq t_{cr}$
Дробь с вероятностью 95% лежит от -1.96 до 1.96.

$$\beta_x \in [\hat{\beta}_x - t_{cr} se(\hat{\beta}_x); \hat{\beta}_x + t_{cr} se(\hat{\beta}_x)]$$

$$\beta_x \in [-3.2 - 1.96 \cdot 0.6; -3.2 + 1.96 \cdot 0.6] \approx [-4.4; -2.0]$$

Замечаем, что $\beta_x = 0$ не попадает в $[-4.4; -2.0]$.
$H_0 : \beta_x = 0$
$H_A : \beta \neq 0, \alpha = 5\%$
H_0 отвергается, $\hat{\beta}_x$ значительно отличается от нуля.

$$R^2 = 1 - \frac{RSS}{TSS}$$

Воспользуемся третьей теоремой:

$$\frac{\frac{RSS_s - RSS_l}{k_l - k_s}}{\frac{RSS_l}{n - k_l}} \sim F_{k_l - k_s, n - k_l}$$

$$\frac{\frac{RSS_s - RSS_l}{4-2}}{\frac{RSS_l}{200-4}} = \frac{\frac{RSS_s - RSS_l}{TSS}/2}{\frac{RSS_l}{TSS}/196} = \frac{(0.4 - 0.3) \cdot 196}{2 \cdot 0.3} \approx 3.04$$

$$F_{cr} = 3.04$$

Сравним F_{cr} с $F_{2,196}$, видим, что гипотеза H_0 отвергается, поэтому выбираем модель B для оценки.