Математическая статистика. ДЗ 15.

Прохоров Юрий, 776

Задача 1

Рассматривается модель линейной регрессии

$$y_i = \theta x_i + \varepsilon_i, \qquad i = \overline{1, n},$$

где x_i — известные действительные числа, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ — независимый нормальный шум с неизвестной дисперсией σ^2, y_i — наблюдаемые величины.

На уровне значимости α проверить гипотезу

$$H_0: \quad \theta = \theta_0$$

Решение:

Построим некоторую статистику, имеющую известное распределение.
Если бы σ^2 было известно, то можно бы было взять

$$T(\mathbf{y}) = \frac{1}{\sigma^2}\|\mathbf{y} - \theta\mathbf{x}\|^2 = \sum_{i=1}^{n}\left(\frac{\varepsilon_i}{\sigma}\right)^2 \sim \chi^2(n)$$

Оценка методом наименьших квадратов:

$$\widehat{\theta} = \widehat{\theta}^{\text{МНК}} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}$$

В случае нормальной модели:

- случайные векторы $(\mathbf{y} - \widehat{\theta}\mathbf{x})$ и $\widehat{\theta}$ независимы;
- $\widehat{\theta} \sim \mathcal{N}\big(\theta,\, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\big)$;
- $\dfrac{1}{\sigma^2}\big\|\mathbf{y} - \widehat{\theta}\mathbf{x}\big\|^2 \sim \chi^2(n-1)$.

Отсюда следует, что

$$\frac{\widehat{\theta} - \theta}{\sqrt{\sigma^2(\mathbf{x}^T\mathbf{x})^{-1}}} \sim \mathcal{N}(0,1) \qquad \Longrightarrow \qquad T(\mathbf{y}) = \frac{\dfrac{\widehat{\theta} - \theta}{\sqrt{\sigma^2(\mathbf{x}^T\mathbf{x})^{-1}}}}{\sqrt{\dfrac{1}{n-1}\dfrac{1}{\sigma^2}\big\|\mathbf{y} - \widehat{\theta}\mathbf{x}\big\|^2}} = \sqrt{n-1}\frac{\widehat{\theta} - \theta}{\sqrt{(\mathbf{x}^T\mathbf{x})^{-1} \cdot \big\|\mathbf{y} - \widehat{\theta}\mathbf{x}\big\|^2}} \sim \mathrm{St}(n-1)$$

Пусть $\theta = \theta_0$. Критическую область выберем симметрично:

$$\Omega_{\text{кр}} = \mathbb{R} \setminus [z_1, z_2], \qquad -z_1 = z_2 = \lambda_{\frac{1-\alpha}{2}} \text{ — квантили распределения } \mathrm{St}(n-1)$$

Итого, решающее правило,

$$\delta(\mathbf{y}) = \begin{cases} \text{принимается } H_0 & , \quad T(\mathbf{y}) \in \left[-\lambda_{\frac{1-\alpha}{2}},\, \lambda_{\frac{1-\alpha}{2}}\right], \\ \text{отвергается } H_0 & , \quad \text{иначе} \end{cases}$$

$$T(\mathbf{y}) = \sqrt{n-1}\frac{(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} - \theta_0}{\sqrt{(\mathbf{x}^T\mathbf{x})^{-1} \cdot \big\|\mathbf{y} - (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} \cdot \mathbf{x}\big\|^2}}$$

Задача 2

Для наблюдений

$$y_i = z_i^2 + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{U}[0, 0.7], \qquad z_i = 2 + 0.1(i-1), \qquad i = \overline{1, n}$$

строится регрессионная модель

$$y = \sum_{j=1}^{k} \theta_j f_j(z), \qquad f_j(z) \text{ — полиномы Чебышева (1-го рода)}$$

Смоделировать выборку при разных значениях k.

Решение:

Нам нужно, чтобы ошибки имели нулевое матожидание, поэтому перепишем:

$$y = z^2 + 0.35 + \eta, \qquad \eta \sim \mathcal{U}[-0.35,\, 0.35]$$

Обозначим

$$X^{(k)} = \begin{bmatrix} f_1(z_1) & \cdots & f_k(z_1) \\ \vdots & \ddots & \vdots \\ f_1(z_n) & \cdots & f_k(z_n) \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \theta^{(k)} = \big(\theta_1, \ldots, \theta_k\big)$$

Оценка параметров по МНК:

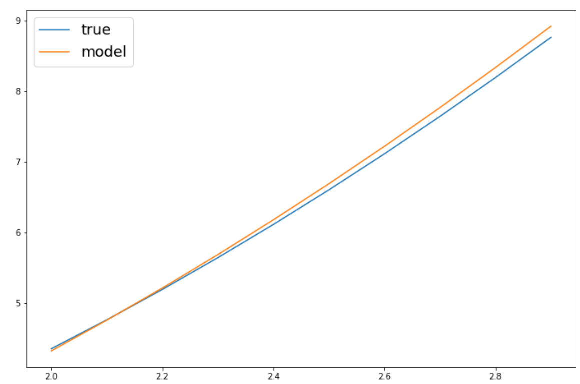
$$\widehat{\theta} = (X^T X)^{-1} X^T \mathbf{y}$$

Известно, что

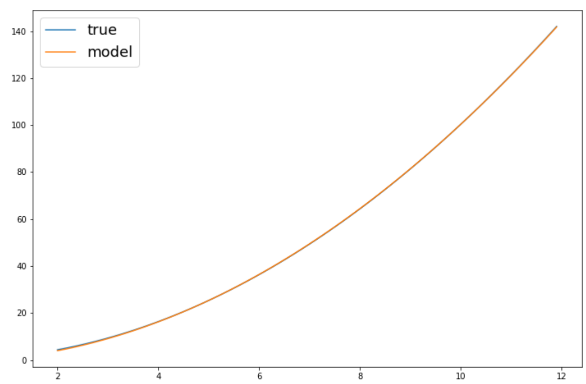
$$f_1(z) = z, \qquad f_2(z) = 2z^2 - 1, \qquad \ldots$$

поэтому оптимальные значения параметров отличаются от истинных и зависят от n.

1. $k = 2$

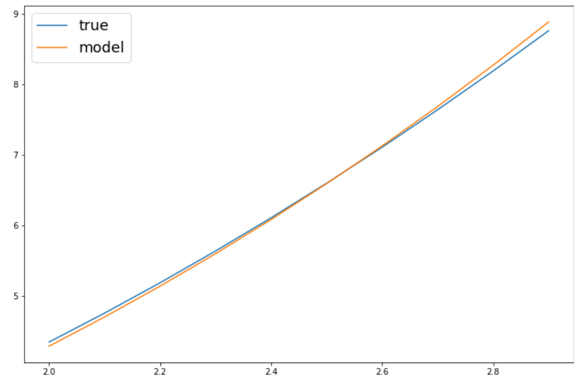


$n = 10$

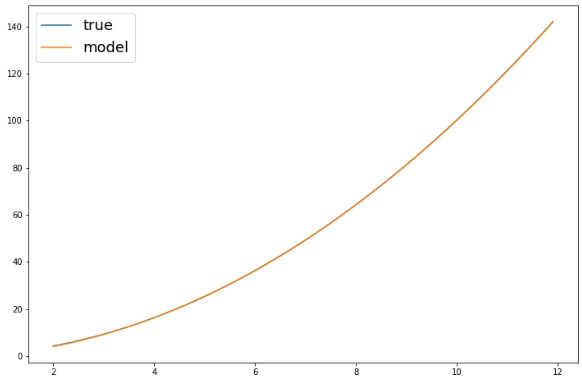


$n = 100$

2. $k = 3$

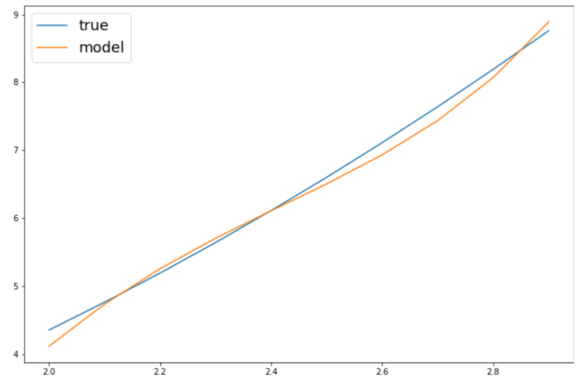


$n = 10$

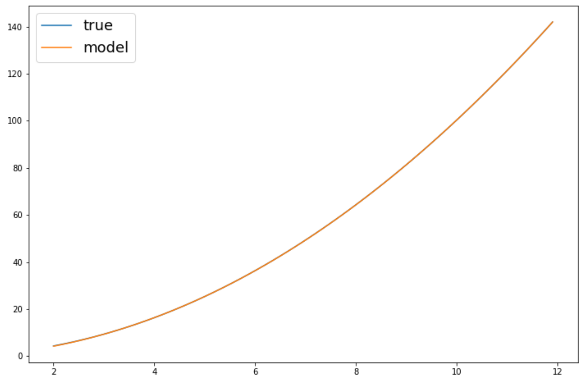


$n = 100$

3. $k = 4$



$n = 10$



$n = 100$