Эконометрика

Модель линейной регрессии

Шишкин Владимир Андреевич

Пермский государственный национальный исследовательский университет

Случайные величины

Вероятностью $\mathrm{P}(A)$ события A называется численная мера степени объективной возможности появления этого события.

Случайная величина X — функция, заданная на множестве элементарных событий.

Закон распределения случайной величины X — всякое соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями.

Случайные величины

Характеристики случайных величин

- Математическое ожидание:

$$\mathsf{E}\,(X) = \sum_{i=1}^{n} x_i p_i \qquad \mathsf{E}\,(X) = \int_{-\infty}^{\infty} xp(x)\,dx$$

- Дисперсия и среднее квадратическое отклонение (стандартное отклонение):

$$\mathsf{D}\,(x) = \mathsf{E}\left(X - \mathsf{E}\,(X)\right)^2 \qquad \sigma_X = \sqrt{\mathsf{D}\,(X)}$$

- x_q — квантиль уровня q, если

$$F(x_q) = \mathsf{P}(x \leqslant x_q) = q$$

- Начальные ν_k и центральные μ_k моменты k-го порядка:

$$\nu_k = \mathsf{E}\,(X^k) \qquad \mu_k(X) = \mathsf{E}\left(X - \mathsf{E}\,(X)\right)^k$$

Случайные величины

Характеристики многомерных случайных величин

- Ковариация (ковариационный момент):

$$\operatorname{Cov}(X,Y) = \mathsf{E}\Big(\big(X - \mathsf{E}(X)\big)\big(Y - \mathsf{E}(Y)\big)\Big)$$

- Коэффициент корреляции:

$$\operatorname{Cor}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Оценки параметров распределений

Оценка $\tilde{\theta}_n$ параметра θ

Всякая функция результатов наблюдений над случайной величиной X, с помощью которой судят о величине θ.

- *Несмещённая* оценка: $\mathsf{E}\,(\tilde{\theta}_n) = \theta$.
- *Состоятельная* оценка: для любого $\varepsilon > 0$

$$\lim_{n\to\infty} \mathsf{P}\big(|\tilde{\theta}_n - \theta| \leqslant \varepsilon\big)$$

- *Эффективная* оценка — несмещённая оценка, имеющая наименьшую дисперсию среди всех несмещённых оценок, вычисленных по выборкам одного и того же объёма.

Оценки параметров нормального распределения

Точечные оценки

- Математическое ожидание (среднее арифметическое):

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Дисперсия (исправленная выборочная дисперсия):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- Стандартное отклонение:

$$s = \sqrt{s^2}$$

Оценки параметров нормального распределения

Интервальные оценки

Интервальная оценка параметра θ

Числовой интервал $(\tilde{\theta}_n^{(1)}, \tilde{\theta}_n^{(2)})$, который с заданной вероятностью γ накрывает неизвестное значение параметра θ.

- Доверительный интервал для математического ожидания на уровне значимости α:

$$\left(\overline{x} + t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n-1}}, \overline{x} + t_{1-\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n-1}} \right)$$

- Доверительный интервал для генеральной дисперсии на уровне значимости α:

$$\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \right)$$

Проверка статистических гипотез

Статистическая гипотеза

Любое предположение о виде или параметре неизвестного закона распределения.

Гипотеза H_0	Принимается	Отвергается
Верна	Правильно	Ошибка первого рода
Не верна	Ошибка второго рода	Правильно

Уровень значимости критерия — вероятность α допустить ошибку первого рода.
Мощность критерия — вероятность $(1 - \beta)$ не допустить ошибку второго рода.

Проверка статистических гипотез для величин, имеющих нормальное распределение

$H_0 : \overline{x} = \mu$:

t-критерий Стьюдента

$$\frac{\overline{x} - \mu}{s}\sqrt{n} \sim t_{n-1}$$

T^2-критерий Хотеллинга

$x = (x_1, \ldots, x_m)$

$$n(\overline{x} - \mu)^\top S^{-1}(\overline{x} - \mu) \sim \frac{m(n-1)}{n-m} F_{m,n-m}$$

Проверка статистических гипотез для величин, имеющих нормальное распределение

$H_0 : \overline{x}_1 = \overline{x}_2$:

t-критерий Стьюдента

$$\frac{(\overline{x}_1 - \overline{x}_2)\sqrt{n_1 n_2}}{s_* \sqrt{n_1 + n_2}} \sim t_{n_1+n_2-2} \qquad s_* = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

T^2-критерий Хотеллинга

$$n_1 n_2 (\overline{x}_1 - \overline{x}_2)^\top S_*^{-1} (\overline{x}_1 - \overline{x}_2) \sim \frac{(n_1+n_2-2)m}{n_1+n_2-m-1} F_{m,n_1+n_2-m-1}$$

$$S_* = \frac{1}{n_1+n_2-2}(K_1^\top K_1 + K_2^\top K_2) \qquad K_{ij} = x_{ij} - \overline{x}_j$$

Парная линейная регрессионная модель

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

x — объясняющая переменная (регрессор)

y — объясняемая переменная

ε — ошибка модели

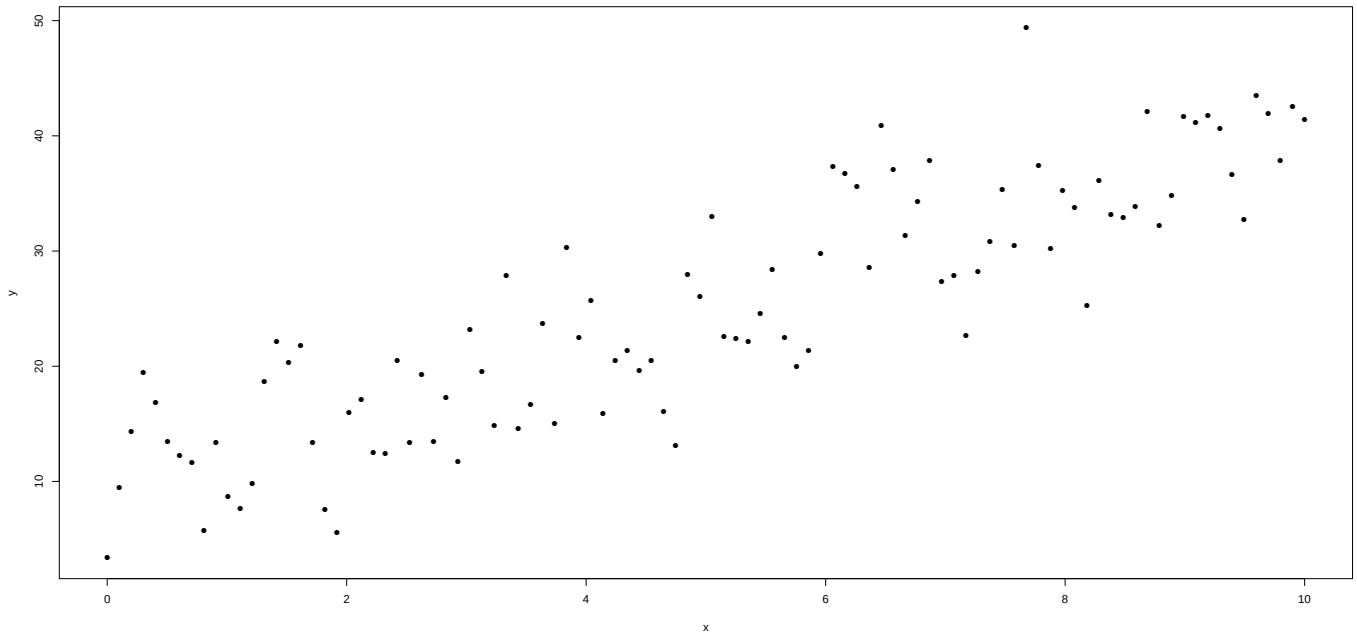
β_1 — величина изменения y при изменении x на единицу

β_0 — величина y при нулевом значении x

Парная линейная регрессия

Исходные данные

(x_i, y_i), $i = 1, 2, \ldots, n$.



Парная линейная регрессия

Задача подгонки кривой

Подобрать функцию $f(x)$ из параметрического семейства $f(x \mid \beta)$ вида

$$f(x \mid \beta) = \beta_0 + \beta_1 x$$

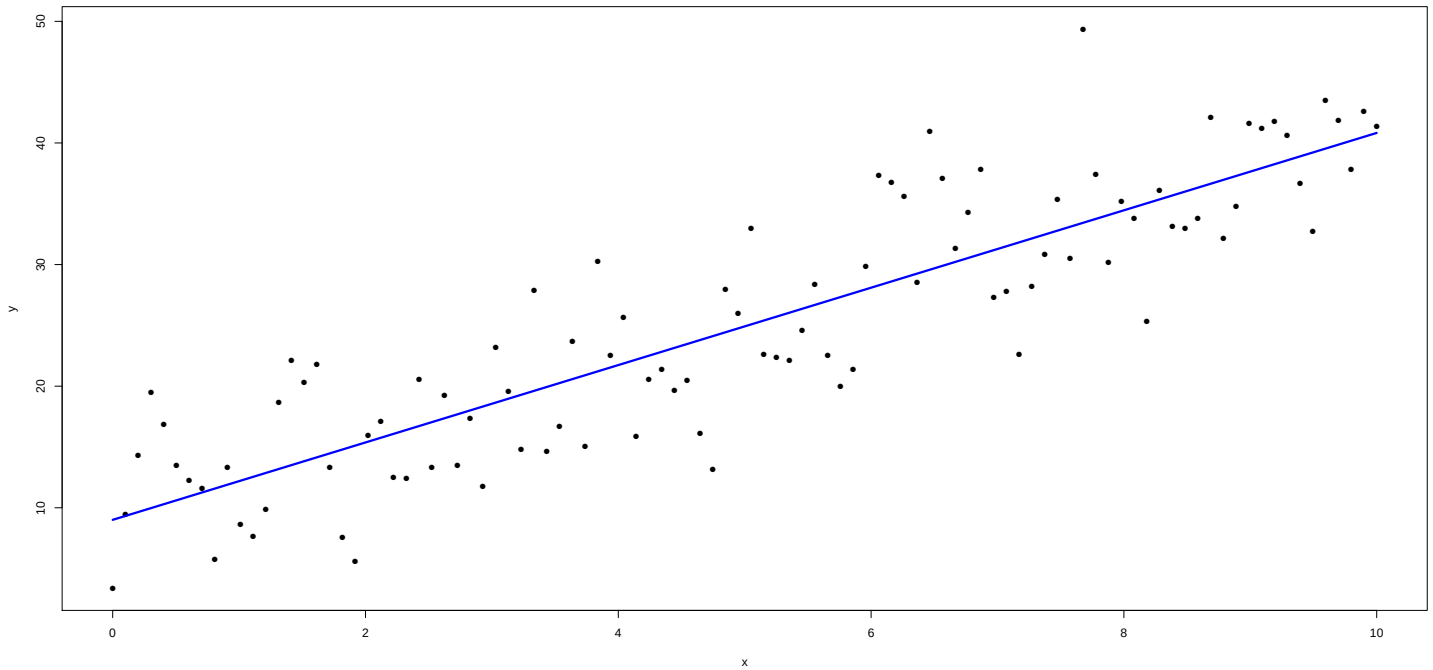
наилучшим образом описывающую зависимость y от x:

$$y_i \approx f(x_i), \quad i = 1, \ldots, n.$$

Оценка близости

- сумма квадратов отклонений $\sum_{i=1}^{n} \big(y_i - f(x_i \mid \beta)\big)^2$;
- сумма модулей $\sum_{i=1}^{n} \big|y_i - f(x_i \mid \beta)\big|$;
- ...

Парная линейная регрессия



Парная линейная регрессия

Оценка коэффициентов регрессии

Метод наименьших квадратов

$$\text{ESS} = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2 \to \min$$

ESS — сумма квадратов ошибок (Errors' Sum of Square)

Парная линейная регрессия

Оценка коэффициентов регрессии

Метод наименьших квадратов

$$\text{ESS} = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2 \to \min$$

Необходимые условия существования экстремума:

$$\frac{\partial \text{ESS}}{\partial \beta_0} = 0, \quad \frac{\partial \text{ESS}}{\partial \beta_1} = 0$$

Парная линейная регрессия

Оценка коэффициентов регрессии

Метод наименьших квадратов

$$\text{ESS} = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2 \to \min$$

Необходимые условия существования экстремума:

$$\frac{\partial \text{ESS}}{\partial \beta_0} = 0, \quad \frac{\partial \text{ESS}}{\partial \beta_1} = 0$$

Стандартная система нормальных уравнений:

$$b_0 n + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \quad b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

Парная линейная регрессия

Оценка коэффициентов регрессии

Метод наименьших квадратов

$$\text{ESS} = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2 \to \min$$

Необходимые условия существования экстремума:

$$\frac{\partial \text{ESS}}{\partial \beta_0} = 0, \quad \frac{\partial \text{ESS}}{\partial \beta_1} = 0$$

Стандартная система нормальных уравнений:

$$b_0 + b_1 \overline{x} = \overline{y}, \quad b_0 \overline{x} + b_1 \overline{x^2} = \overline{yx}$$

Парная линейная регрессия

Оценка коэффициентов регрессии

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad b_1 = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - \overline{x}^2}$$

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} x_i b_1 \qquad b_0 = \overline{y} - b_1\overline{x}$$

$$b_1 = \frac{\operatorname{cov}(x, y)}{s_x^2}$$

Парная линейная регрессия

Оценка коэффициентов регрессии (матричная форма записи)

Систему уравнений

$$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_1$$
$$\dots$$
$$y_n = \beta_0 + \beta_1 x_n + \varepsilon_1$$

можно записать в виде

$$y = X\beta + \varepsilon$$

где

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Парная линейная регрессия

Оценка коэффициентов регрессии (матричная форма записи)

$$y = X\beta + \varepsilon$$

$$X^\top y = X^\top X\beta + X^\top \varepsilon$$

$$X^\top y = X^\top X b$$

$$\boxed{b = (X^\top X)^{-1} X^\top y}$$

Парная линейная регрессионная модель

Основные гипотезы

1. $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, 2, \ldots, n$ — модель правильно специфицирована
2. x — детерминированная величина, ветор $(x_1, \ldots, x_n)$ не коллинеарен вектору $(1, 1, \ldots, 1)^\top$
3. $\mathsf{E}\,\varepsilon_i = 0$, $\mathsf{E}\,(\varepsilon_i^2) = \mathsf{D}\,(\varepsilon_i) = \sigma^2$
4. $\mathsf{E}\,(\varepsilon_i \varepsilon_j) = 0$ при $i \neq j$
5. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Теорема Гаусса–Маркова

В предположениях модели 1–4 оценки b_0, b_1, полученные по методу наименьших квадратов, имеют наименьшую дисперсию в классе всех линейных несмещённых оценок.

Парная линейная регрессионная модель

Дисперсия ошибок

Прогноз значения y_i в точке x_i (модельное значение):

$$\hat{y}_i = b_0 + b_1 x_i, \quad i = 1, 2, \ldots, n$$

Остатки модели:

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n$$

Несмещённая оценка дисперсии ошибок σ_ε^2:

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$$

Парная линейная регрессионная модель

Оценки дисперсии параметров модели

$$s_{b_1}^2 = \frac{s_e^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

$$s_{b_0}^2 == \frac{s_e^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \overline{x})^2}$$

$$\operatorname{cov}(b_0, b_1) = -\frac{s_e^2 \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

Парная линейная регрессионная модель

Проверка гипотезы $H_0 : \beta = \beta^*$

$$t_{i,\text{н}} = \frac{b_i - \beta_i^*}{s_{b_i}} \sim t_{n-2}$$

t-критерий Стьюдента

Если

$$\frac{|b_i|}{s_{b_i}} > t_{1-\frac{\alpha}{2},n-2},$$

то коэффициент b_i линейной модели значим при уровне значимости α (отклоняется нулевая гипотеза $\beta_i = 0$).

Доверительный интервал для β_i:

$$b_i + t_{\frac{\alpha}{2},n-2} \cdot s_{b_i} \leqslant \beta_i \leqslant b_i + t_{1-\frac{\alpha}{2},n-2} \cdot s_{b_i}$$

Парная линейная регрессионная модель

Анализ вариации зависимой переменной в регрессии

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y})^2}_{\text{ESS}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}_{\text{RSS}}$$

TSS — вся дисперсия y (total sum of squares)

ESS — не объяснённая дисперсия (error sum of squares)

RSS — объяснённая моделью часть дисперсии y (regression sum of squares)

Коэффициент детерминации

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

Парная линейная регрессионная модель

Проверка значимости модели

F-критерий Фишера

Если выполняется условие

$$F_{\text{н}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2} > F_{1-\alpha,1,n-2},$$

то модель значима на уровне значимости α.

$$F_{\text{н}} = (n-2)\frac{R^2}{1-R^2}$$