

Глава 12

Критерии в линейной нормальной модели

12.1 Базовая часть

12.1.1 Простая линейная модель

Простой линейной моделью называют $Y_i = aX_i + b + \varepsilon_i$, где ε_i — н.о.р. с нулевым средним, (X_i, Y_i) — наблюдаемые величины. Мы будем считать, что $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, X_i — заданные константы.

Пример 1. Представим себе, что мы наблюдаем за временем поездки T и расстоянием S , которое водитель такси проезжает по маршрутам. Предположим, что мы знаем, что средняя скорость водителя есть некоторая константа v , к которой добавляется случайная погрешность $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Тогда $S_i = vT_i + \varepsilon_i$, а значит

$$\vec{S} = \vec{T}v + \vec{\varepsilon}.$$

Мы можем построить ОМП для v, σ на основе правдоподобия

$$L = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (s_i - t_i v)^2 \right), \quad \ln L = -n/2 \ln(2\pi\sigma^2) - \frac{1}{\sigma^2} \sum_{i=1}^n (s_i - t_i v)^2.$$

Тогда ОМП будет определяться из соотношений

$$\frac{\partial \ln L}{\partial v} = -\frac{1}{\sigma^2} \sum_{i=1}^n t_i (s_i - t_i v) = 0, \quad \frac{\partial \ln L}{\partial \sigma} = \frac{n}{\sigma} - \frac{1}{\sigma^3} \sum_{i=1}^n (s_i - t_i v)^2 = 0,$$

откуда

$$\hat{v} = \frac{\overline{TS}}{\overline{T^2}}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (S_i - T_i \hat{v})^2.$$

Аналогичным образом в простой линейной модели мы получим оценки

$$\hat{a} = \frac{\overline{YX} - \overline{Y} \overline{X}}{S_X^2}, \quad \hat{b} = \overline{Y} - \hat{a} \overline{X}.$$

12.1.2 Общая линейная модель

Будем рассматривать линейную модель $\vec{Y} \sim \mathcal{N}(\vec{\mu}(\vec{a}), \sigma^2 E)$, где E — единичная матрица, $\vec{\mu}$ — некоторый вектор, линейно зависящий от неизвестных параметров a_1, \dots, a_m : $\vec{\mu} = X\vec{a}$, где X — заданная матрица

Рассмотрим поиск ОМП в общей линейной модели:

$$\vec{Y} = X\vec{a} + \vec{\varepsilon},$$

где X — некоторая заданная матрица $n \times k$, $k < n$. Логарифм правдоподобия имеет вид

$$\ln L = -n \ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \|\vec{y} - X\vec{a}\|^2.$$

Можно найти ОМП дифференцированием, а можно найти \hat{a} из условия на то, что оно минимизирует $\|\vec{y} - X\vec{a}\|^2$.

Это равносильно тому, что $\vec{y} - X\hat{a}$ ортогонален всем векторам $X\vec{a}$:

$$\langle \vec{y} - X\hat{a}, X\vec{a} \rangle = \vec{a}^t X^t \vec{y} - \vec{a}^t X^t X \hat{a} = \vec{a}^t (X^t \vec{y} - X^t X \hat{a})$$

при всех \vec{a} . Следовательно,

$$X^t \vec{y} - X^t X \hat{a} = 0, \quad \hat{a} = (X^t X)^{-1} X^t \vec{y}.$$

ОМП $\hat{\sigma}^2$ находится при заданном \hat{a} из соотношения

$$\hat{\sigma}^2 = \frac{1}{n} \|\vec{y} - X\hat{a}\|^2,$$

получаемого при дифференцировании $\ln L$ по σ .

При этом расстояние от \vec{y} до $L = \{X\vec{a}\}$ называется остаточной суммой квадратов и имеет вид

$$RSS = \|\vec{y} - X\hat{a}\|^2 = \|\vec{y}\|^2 - (X\hat{a})^t (X\hat{a}) = \|\vec{y}\|^2 - \vec{y}^t X (X^t X)^{-1} X^t X (X^t X)^{-1} X^t \vec{y} = \|\vec{y}\|^2 - \vec{y}^t X (X^t X)^{-1} X^t \vec{y}.$$

Полученные оценки называются оценками методом наименьших квадратов (МНК).

12.1.3 Распределение оценок МНК

Можно найти распределение статистики \hat{a}

$$\hat{a} = (X^t X)^{-1} X^t (X\vec{a} + \vec{\varepsilon}) = \vec{a} + (X^t X)^{-1} X^t \vec{\varepsilon}.$$

При этом вектор $(X^t X)^{-1} X^t \vec{\varepsilon}$ имеет нормальное распределение (как линейное преобразование нормального вектора) с нулевым средним и матрицей ковариации

$$\text{cov}(\hat{a}, \hat{a}) = \mathbf{E}(\hat{a} - \vec{a})(\hat{a} - \vec{a})^T = (X^t X)^{-1} X^t \mathbf{E} \vec{\varepsilon} \vec{\varepsilon}^t X (X^t X)^{-1} = (X^t X)^{-1} \sigma^2.$$

Здесь мы воспользовались линейностью математического ожидания и тем, что матрица ковариации вектора \vec{X} есть

$$\mathbf{E}(\vec{X} - \mathbf{E}\vec{X})(\vec{X} - \mathbf{E}\vec{X})^T,$$

в чем нетрудно убедиться, расписав ее по определению.

Таким образом,

$$\hat{a} \sim \mathcal{N}(\vec{a}, (X^t X)^{-1} \sigma^2),$$

то есть оценки коэффициентов имеют нормальное распределение.

12.1.4 Лемма об ортогональности

В прошлый раз мы сформулировали лемму Фишера:

Лемма 1. Если $X_i \sim \mathcal{N}(\mu, \sigma^2)$, то величины

$$T_1 = \sqrt{n} \frac{\bar{X} - \mu}{\sigma}, \quad T_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}$$

являются независимыми, причем $T_1 \sim \mathcal{N}(0, 1)$, $T_2 \sim \chi_{n-1}^2$.

Данное утверждение вытекает из следующего более общего утверждения:

Лемма 2. Если $\vec{X} \sim \mathcal{N}(0, \sigma^2 E)$ — многомерный нормальный вектор, а L_1, L_2 — два ортогональных подпространства \mathbb{R}^n , то проекции \vec{X} на L_1 и L_2 независимы и имеют длины, квадраты которых распределены как $\sigma^2 \cdot \chi_{\dim L_1}^2$ и $\sigma^2 \cdot \chi_{\dim L_2}^2$.

Доказательство Леммы 2. Доказательство достаточно бесхитростно. Пусть вектор \vec{Y} — это вектор координат \vec{X} в базисе, первые $k = \dim L_1$ базисных векторов которого образуют пространство L_1 , следующие l базисных векторов, где $l = \dim L_2$, образуют пространство L_2 . Тогда $\vec{Y} = C\vec{X}$, где C — ортогональная матрица перехода. При этом

$$\psi_{\vec{Y}}(\vec{t}) = \mathbf{E} \exp(i\langle \vec{t}, \vec{Y} \rangle) = \mathbf{E} \exp(i\langle \vec{t}, C\vec{X} \rangle) = \mathbf{E} \exp(i\langle C^t \vec{t}, \vec{X} \rangle) = \exp(-\sigma^2 \vec{t}^t C C^t \vec{t} / 2) = \exp(-\sigma^2 \|\vec{t}\|^2 / 2),$$

то есть вектор \vec{Y} также распределен $\mathcal{N}(0, \sigma^2 E)$. При этом проекция \vec{X} на L_1 это $(Y_1, \dots, Y_k, 0, \dots, 0)$, а \vec{X} на L_2 это $(0, \dots, 0, Y_{k+1}, \dots, Y_{k+l}, 0, \dots, 0)$. Следовательно, проекции независимы (независимость векторов — это свойство векторов, не привязанное к выбору базиса: вероятности попадания в некоторые множества равны произведению вероятностей попадания в другие множества). Длины векторов в старом и новом базисе одинаковы, поскольку замена была ортогональной. В новом базисе квадраты длин проекций при делении на σ^2 имеют распределения χ_k^2 и χ_l^2 , поскольку это суммы квадратов н.о.р. $\mathcal{N}(0, 1)$ величин. \square

Доказательство Леммы 1. Будем доказывать для $\mu = 0, \sigma = 1$, общее доказательство получается заменой $(X_i - \mu)/\sigma$ на новую переменную.

То, что $T_1 \sim \mathcal{N}(0, 1)$ очевидно. Остальные утверждения прямо вытекают из Леммы 2, поскольку проекция (X_1, \dots, X_n) на вектор $(1, \dots, 1)$ есть $(\bar{X}, \dots, \bar{X})$, поскольку $(X_1 - \bar{X}, \dots, X_n - \bar{X})$ ортогонален $(1, \dots, 1)$. Соответственно, \bar{X} не зависит от $nS^2 = \sum_{i=1}^n (X_i - \bar{X})^2$ — квадрат длины разности \vec{X} и $(\bar{X}, \dots, \bar{X})$, то есть квадрат длины проекции \vec{X} на ортогональное дополнение к $(1, \dots, 1)$. \square

12.1.5 Гипотезы в линейной модели

Представим себе, что мы интересуемся линейной гипотезой — $H_0 : \vec{Y} = \tilde{X}\vec{b} + \vec{\varepsilon}$, где $L_1 = \{\tilde{X}\vec{b}\}$ — подпространство L размерности m . Тогда максимум правдоподобия в общем случае имеет вид

$$L(x_1, \dots, x_n; \vec{a}, \sigma) = \left(\frac{\sqrt{n}}{\sqrt{2\pi RSS}} \right)^n \exp\left(-\frac{n}{2}\right).$$

При гипотезе оценки другие и максимум правдоподобия имеет вид

$$L(x_1, \dots, x_n; \vec{b}, \sigma) = \left(\frac{\sqrt{n}}{\sqrt{2\pi RSS_0}} \right)^n \exp\left(-\frac{n}{2}\right), \quad RSS_0 = \|\vec{y} - \tilde{X}\hat{\vec{b}}\|^2, \quad \hat{\vec{b}} = (\tilde{X}^t \tilde{X})^{-1} \tilde{X}^t \vec{y}.$$

Тем самым критерий отношения правдоподобий имеет вид

$$\left\{ \left(\frac{RSS_0}{RSS} \right)^{n/2} > c \right\}, \left\{ \frac{RSS_0 - RSS}{RSS} > \tilde{c} \right\}.$$

Остается заметить, что RSS_0 — это квадрат расстояния от \vec{Y} до L_0 , то есть квадрат длины проекции \vec{Y} на L_0^\perp , RSS — квадрат длины проекции на L^\perp , а $RSS_0 - RSS$ — квадрат длины проекции на $L \setminus L_0$. При выполнении гипотезы $\vec{Y} = \tilde{X}\vec{b} + \vec{\varepsilon}$, где $\tilde{X}\vec{b}$ лежит в L_0 , а значит проекция \vec{Y} на L_0^\perp такая же как и $\vec{\varepsilon}$. Аналогичным образом, проекция \vec{Y} на L_1^\perp вне зависимости от справедливости гипотезы такая же как у $\vec{\varepsilon}$. В силу Леммы при выполнении гипотезы $RSS_0 - RSS$ имеет χ_{k-m}^2 распределение и не зависит от RSS , имеющей χ_{n-k}^2 распределение. Следовательно, критерий отношения правдоподобий имеет вид

$$\frac{(RSS_0 - RSS)/(k - m)}{RSS/(n - k)} > f_{1-\alpha},$$

где f — квантиль распределения фишера $F_{k-m, n-k}$.

При этом зачастую удобнее использовать для числителя представление

$$RSS_0 - RSS = \sum_{i=1}^n (\hat{y}_i^* - y_i^*)^2,$$

где $y_i^* = X\hat{B}$ — прогноз в первой модели, \hat{y}_i^* — во второй.

Параллельно мы доказали, что

$$\frac{1}{\sigma^2} RSS \sim \chi_{n-k}^2.$$

12.2 Факультативная часть

12.2.1 Оценки МНК в ненормальной модели

Если $\vec{Y} = X\vec{a} + \vec{\varepsilon}$, где ε н.о.р., $\mathbf{E}\varepsilon = 0$, $\mathbf{D}\varepsilon = \sigma^2$, но при этом ε , вообще говоря, не нормальная. В этом случае оценки МНК рассчитываются по тем же формулам, однако уже не являются ОМП. Однако, верна следующая теорема Гаусса-Маркова:

Теорема 1. Среди линейных оценок $\sum_{i=1}^n b_i(X)Y_i$ оценки МНК являются оптимальными (имеющими минимальную дисперсию среди несмещенных).

А что насчет асимптотических свойств? Здесь проблема в том, что увеличивая n , мы должны изменять набор предикторов X , а значит требуется описать как это происходит с ростом n . Можно наложить условия на матрицу, но более естественным является следующий путь: пусть $(X_{i\cdot}, \varepsilon_i)$ — н.о.р. векторы с некоторыми распределениями, причем $\mathbf{E}(\varepsilon_i|X_{i\cdot}) = 0$, $\mathbf{E}(\varepsilon_i^2|X_{i\cdot}) = \sigma^2$. Тогда оценки МНК оказываются состоятельными и асимптотически нормальными.

Теорема 2. Пусть $\mathbf{E}X_{1,i}^4 < +\infty$, $\mathbf{E}\varepsilon_i^4 < +\infty$, $Q = \mathbf{E}X^t X$ имеет полный ранг. Тогда

$$\sqrt{n}(\hat{a} - \vec{a}) \xrightarrow{d} Z \sim \mathcal{N}(\vec{0}, \Sigma),$$

где $\Sigma = Q^{-1}\sigma^2$.

12.2.2 Факторный анализ

Представим себе $X_{i,j} \sim \mathcal{N}(\mu_i, \sigma^2)$, $j \leq n_i$. Иначе говоря, у нас есть несколько совокупностей, каждая со своим средним, но с одинаковой дисперсией.

Такая задача называется "однофакторным анализом" — есть некий фактор, который возможно оказывает влияние на среднее нашего распределения, для разных выборок этот фактор принимает разные значения. Например, я мог взять несколько грядок с растениями и одну не снабжать удобрениями, вторую снабжать небольшой дозой, а третью обеспечить большой дозой удобрения.

Основная гипотеза здесь заключается в том, что фактор не оказывает влияния: $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$. Эта модель попадает в нашу линейную структуру: объединенная выборка

$$\vec{Y} = (X_{1,1}, \dots, X_{1,n_1}, X_{2,1}, \dots, X_{2,n_2}, \dots, X_{k,1}, \dots, X_{k,n_k}) \sim \mathcal{N}(\vec{\mu}, \sigma^2 E),$$

где $\vec{\mu}$ — вектор из n_1 величин μ_1 , n_2 величин μ_2 и так далее. Если ввести параметры μ_1, \dots, μ_k , то модель легко сводится к линейной с некоторой матрицей X из нулей и единиц. В одной из домашних задач предлагается получить в этой модели критерий.

Двухфакторная модель предлагает рассматривать наблюдения $X_{i,j,k}$, где $i \leq l$ — номер строки, $j \leq m$ — номер столбца, $k \leq n$ — номер наблюдения в каждой категории. Скажем, мы можем проводить эксперимент с различными сочетаниями двух факторов. При этом мы предполагаем, что $X_{i,j,k} \sim \mathcal{N}(\mu_{i,j}, \sigma^2)$,

где $\mu_{i,j} = a_i + b_j + \mu$, $\sum a_i = 0$, $\sum b_j = 0$. Показатели a и b характеризуют влияние фактора строки и фактора столбца. Здесь мы также можем проверять гипотезу $H_0 : a_1 = a_2 = \dots = a_l = 0$ или $b_1 = b_2 = \dots = b_m = 0$. Эта модель также может быть сведена к общей линейной модели.