Спецкурс "Регрессия"

Шкляев А.В.

15 января 2019 г.

1 Простая линейная регрессия для нормальной модели

1.1 Оценка для y

Напомню, что мы рассматриваем модель $y_i = ax_i + b + \varepsilon_i$, где ε_i н.о.р. $\mathcal{N}(0, \sigma^2)$.
Мы получили оценки максимального правдоподобия

$$\widehat{a} = \frac{\overline{xy} - \overline{y}\ \overline{x}}{\overline{x^2} - \overline{x}^2},\ \widehat{b} = \overline{y} - \widehat{a}\overline{x},\ \widehat{\sigma}^2 = \overline{(y - \widehat{a}x - \widehat{b})^2}$$

и показали, что

$$(\widehat{a}, \widehat{b}) \sim \mathcal{N}\left((a, b), \frac{\sigma^2}{nS_x^2}\Sigma\right),\ \Sigma = \begin{pmatrix} 1 & -\overline{x}, \\ -\overline{x} & \overline{x^2}. \end{pmatrix}.$$

Разберемся с распределением $\widehat{\sigma}^2$. Для этого нам понадобятся следующее утверждение:

Лемма 1. *Пусть $\vec{Z} = (Z_1, ..., Z_n) \sim \mathcal{N}(0, \sigma^2 E)$, L — некоторое подпространство размерности m в $\mathbb{R}^n$, $L^\perp$ — его ортогональное дополнение. Пусть $U = \mathrm{proj}_L \vec{Z}$, $V = \mathrm{proj}_{L^\perp} \vec{Z}$. Тогда U, V независимы, $||U||^2/\sigma^2 \sim \chi^2_m$, $||V||^2/\sigma^2 \sim \chi^2_{n-m}$.*

Доказательство. Рассмотрим такую ортонормированную систему координат, что первые m базисных векторов образуют базис L, а остальные — базис $L^\perp$. Тогда в новых координатах вектор Z будем иметь вид $W = CZ$, где C — матрица перехода от исходного базиса к новому. Таким образом, вектор W также имеет распределение $\mathcal{N}(0, \sigma^2 E)$, поскольку C — ортогональная матрица.

Следовательно, векторы $\widetilde{U} = (W_1, ..., W_m, 0, 0, ..0)$ и $\widetilde{V} = (0, ..., 0, W_{m+1}, ..., W_n)$ независимы и квадраты их длин являются χ^2_m и χ^2_{n-m} случайными величинами. Однако, тогда $U = C^{-1}\widetilde{U}$, $V = C^{-1}\widetilde{V}$ независимы, а их квадраты длин в точности совпадают с квадратами длин U и V. □

Теперь заметим, что поскольку оценки $\widehat{a}$, $\widehat{b}$ минимизируют

$$\sum_{i=1}^n (y_i - ax_i - b)^2,$$

то $(\widehat{a}x_1 + \widehat{b}, ..., \widehat{a}x_n + \widehat{b})$ — это проекция $(y_1, ..., y_n)$ на двумерное подпространство L, порожденное векторами $(1, 1, ..., 1)$, $(x_1, ..., x_n)$. При этом вектор $Y_1, ..., Y_n$ имеет распределение $\mathcal{N}(a\vec{x} + b, \sigma^2 E)$, а значит проекция $\vec{Y}$ на $L^\perp$ совпадает с проекцией $\vec{Y} - a\vec{x} - b$ на $L^\perp$. В силу Леммы 1 квадрат длины этой проекции

$$RSS = \sum_{i=1}^n (y_i - \widehat{a}x_i - \widehat{b})^2$$

при делении на σ^2 имеет χ^2_{n-2} распределение.

Итак,

$$\frac{n\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Более того, в силу леммы RSS не зависит от пары $(\widehat{a}, \widehat{b})$, являющейся координатами проекции $\vec{Y}$ на L. Тем самым, мы можем утверждать, что

$$\sqrt{n}S_x \frac{\widehat{a} - a}{\widehat{\sigma}\sqrt{n/(n-2)}} \sim t_{n-2},\ \frac{\sqrt{n}S_x}{\sqrt{\overline{x^2}}} \frac{\widehat{b} - b}{\widehat{\sigma}\sqrt{n/(n-2)}} \sim t_{n-2},$$

откуда имеем доверительные интервалы

$$a \in \left(\widehat{a} + t_{\alpha/2}\widehat{\sigma}\sqrt{\frac{S_x^2}{n-2}}, \widehat{a} + t_{1-\alpha/2}\widehat{\sigma}\sqrt{\frac{S_x^2}{n-2}}\right),\ b \in \left(\widehat{b} + t_{\alpha/2}\widehat{\sigma}\sqrt{\frac{S_x^2}{(n-2)\overline{x^2}}}, \widehat{b} + t_{1-\alpha/2}\widehat{\sigma}\sqrt{\frac{S_x^2}{(n-2)\overline{x^2}}}\right).$$

1.2 Нормальная модель: прогноз

Осуществим на основе наших наблюдений прогноз $\widehat{y}_*$ значения y_* в точке x_*, которой не было среди x_i. Естественным образом, зададим оценку соотношением $\widehat{y}_* = \widehat{a}x_* + \widehat{b}$. Тогда y_* имеет нормальное распределение с параметрами

$$m_* = ax_* + b,\ \sigma_*^2 = x_*^2 \mathbf{D}\widehat{a} + \mathbf{D}\widehat{b} + 2x_* \operatorname{cov}(\widehat{a}, \widehat{b}) = \frac{\sigma^2}{nS_x^2}\left(x_*^2 - 2x_*\overline{x} + \overline{x^2}\right) = \frac{\sigma^2}{nS_x^2}\left(\frac{1}{n}\sum_{i=1}^n (x_i - x_*)^2\right).$$

Соответственно, погрешность прогноза по сравнению с реальными данными будет иметь распределение

$$\mathcal{N}\left(0, \sigma^2\left(\frac{\sum_{i=1}^n (x_i - x_*)^2}{n^2 S_x^2} + 1\right)\right).$$

При этом $\widehat{y}_*$ есть функция от $\widehat{a}$, $\widehat{b}$ и x_i, а значит не зависит от RSS. Тем самым, можно заметить, что

$$\frac{\widehat{y}_* - y_*}{\sqrt{\frac{RSS}{(n-2)}\left(\frac{\sum_{i=1}^n (x_i - x_*)^2}{n^2 S_x^2} + 1\right)}} \sim t_{n-2},$$

откуда мы можем построить доверительный интервал

$$y_* \in \left(\widehat{y}_* + t_{\alpha/2}\sqrt{\frac{RSS}{(n-2)}\left(\frac{\sum_{i=1}^n (x_i - x_*)^2}{n^2 S_x^2} + 1\right)},\ \widehat{y}_* + t_{1-\alpha/2}\sqrt{\frac{RSS}{(n-2)}\left(\frac{\sum_{i=1}^n (x_i - x_*)^2}{n^2 S_x^2} + 1\right)}\right).$$

1.3 Множественная регрессия в нормальной модели

Обобщим задачу, рассмотренную в предыдущем разделе. Пусть

$$\begin{cases} Y_1 = a_1 X_{1,1} + ... + a_m X_{1,m} + \varepsilon_1, \\ Y_2 = a_1 X_{2,1} + ... + a_m X_{2,m} + \varepsilon_2, \\ \qquad ... \\ Y_n = a_1 X_{n,1} + ... + a_m X_{n,m} + \varepsilon_n, \end{cases}$$

где $a_1, ..., a_m$ — неизвестные коэффициенты, ε_j — случайные величины, являющиеся условно независимыми и одинаково распределенными $\mathcal{N}(0, \sigma^2)$ при условии $X_{1,1} = x_{1,1}, ..., X_{n,m} = x_{n,m}$. Как и прежде, эти условия позволяют нам рассматривать $x_{i,j}$ как константы. Мы будем предполагать, что матрица X имеет линейно независимые столбцы.

Как и в предыдущем случае мы можем выписать функцию правдоподобия

$$L(y_1, ..., y_n; a_1, ..., a_m, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \langle\vec{a}, X_{i,\cdot}\rangle)^2\right)$$

и максимизировать ее для поиска ОМП. Задача разбивается на две части:
1) Поиск наилучшего вектора $\vec{a}$.
Мы ищем вектор $\vec{a}$ такой, что $||y - X\vec{a}||$ минимально.
2) Поиск σ^2 так, чтобы $\ln L$ было максимально.
Мы ищем σ^2 такое, что

$$\ln L = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \langle\vec{a}, X_{i,\cdot}\rangle)^2 + n\ln\sigma.$$

Вторая задача достаточно стандартна, дифференцированием мы получаем ответ

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle\vec{a}, X_{i,\cdot}\rangle)^2$$

Остается найти оценку $\vec{a}$. Это задача поиска ближайшего к $\vec{y}$ вектора в плоскости $X_{\cdot,1}a_1 + ...X_{\cdot,m}a_m$. Здесь $X_{\cdot,i}$ — базис в указанной плоскости, а a_i — коэффициенты разложения по этому базису. Наилучшие $(a_1, ..., a_m)$ определяются тем, что разность $\vec{y}$ и $X\vec{a}$ есть перпендикуляр к плоскости:

$$(\vec{y} - X\vec{a}, X_{\cdot,i}) = 0$$

Иначе говоря,

$$X^t\vec{y} = (X^tX)\vec{a}.$$

Таким образом, $\vec{a}$ определяется как

$$(X^tX)^{-1}X^t\vec{y}.$$

Пример 1. Простая линейная регрессия соответствует случаю множественной регресией с матрицей X вида $X_{i,1} = x_i$, $X_{i,2} = 1$. Соответственно,

$$X^tX = \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & 1 \end{pmatrix}.$$

Обратная матрица будет иметь вид

$$(X^tX)^{-1} = \frac{1}{nS_x^2}\begin{pmatrix} n & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}.$$

Следовательно,

$$(X^tX)^{-1}X^t\vec{y} = \frac{1}{n^2S_x^2}\begin{pmatrix} n & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}\begin{pmatrix} x_1y_1 + ... + x_ny_n \\ y_1 + ... + y_n \end{pmatrix} = \frac{1}{\overline{x^2} - \overline{x}^2}\begin{pmatrix} \overline{xy} - \overline{x}\ \overline{y} \\ \overline{x^2}\ \overline{y} - \overline{xy}\ \overline{x} \end{pmatrix}.$$

Оценка совпадает с той, которая была получена в этой модели.

1.4 Распределение оценок

Распределение вектора $\vec{a} = (X^tX)^{-1}X^t\vec{y}$ можно выписать, пользуясь простым утверждением

Лемма 2. *Если $Z \sim \mathcal{N}(\vec{\mu}, \Sigma)$, а C — матрица, то $CZ \sim \mathcal{N}(C\vec{\mu}, C\Sigma C^t)$.*

Доказательство. Мы будем пользоваться определением нормальности в терминах характеристических функций: $Z \sim \mathcal{N}(\vec{\mu}, \Sigma)$, если

$$\psi_Z(\vec{s}) = \mathbf{E}\exp(i\langle Z, \vec{s}\rangle) = \exp\left(i\langle \vec{s}, \vec{\mu}\rangle - \frac{1}{2}\vec{s}^t\Sigma\vec{s}\right).$$

Тогда

$$\psi_{CZ}(\vec{s}) = \mathbf{E}\exp(i\langle CZ, \vec{s}\rangle) = \mathbf{E}\exp(i\langle Z, C^t\vec{s}\rangle) = \exp\left(i\langle C^t\vec{s}, \vec{\mu}\rangle\right)\exp\left(-\frac{1}{2}\vec{s}^t C\Sigma C^t\vec{s}\right).$$

Глядя на формулу характеристической функции, мы видим что полученное распределение нормально с указанными параметрами. □

В силу Леммы 2 и соотношения

$$\vec{Y} \sim \mathcal{N}(X\vec{a}, \sigma^2 E),$$

имеем

$$\widehat{a} = (X^tX)^{-1}X^t\vec{Y} \sim \mathcal{N}\left((X^tX)^{-1}X^tX\vec{a}, \sigma^2(X^tX)^{-1}X^t((X^tX)^{-1}X^t)^t\right) = \mathcal{N}\left(\vec{a}, \sigma^2(X^tX)^{-1}\right).$$

Таким образом, оценки несмещенные и имеют нормальное распределение.

Как и в предыдущем случае, пользуясь Леммой 1, мы получаем, что

$$RSS = ||\vec{Y} - X\vec{a}||^2$$

не зависит от $\vec{a}$ и после деления на σ^2 имеет χ^2_{n-m} распределение.

Тем самым, мы можем строить построить несмещенную оценку $RSS/(n-m)$ для σ^2 (отличающуюся от ОМП), а также может построить доверительные интервалы для исследуемых параметров:

$$a_i \in \left(\hat{a}_i - \frac{t_{1-\alpha/2}\sqrt{n-m}}{\sqrt{(X^tX)^{-1}_{i,i}}\sqrt{RSS}}, \hat{a}_i + \frac{t_{1-\alpha/2}\sqrt{n-m}}{\sqrt{(X^tX)^{-1}_{i,i}}\sqrt{RSS}}\right),$$

$$\sigma^2 \in \left(\frac{RSS}{y_{1-\alpha/2}}, \frac{RSS}{y_{\alpha/2}}\right),$$

где $t_{\cdot}$ — квантиль распределения Стьюдента с $n-m$ степенями свободы, $y_{\cdot}$ — квантиль распределения χ^2_{n-m}.