Спецкурс "Регрессия"

Шкляев А.В.

21 января 2019 г.

1 Нормальная модель

1.1 Некоторые замечания

На прошлой лекции мы получили выражения для оценок в общей линейной модели $\vec{Y} = X\vec{a} + \vec{\varepsilon}$, где

$$\widehat{a} = (X^tX)^{-1}X^t\vec{Y},\ \widehat{\sigma}^2 = \frac{RSS}{n} = \frac{||\vec{Y} - X\hat{a}||^2}{n}.$$

Отметим также, что

$$RSS = ||\vec{Y}(E - X(X^tX)^{-1}X^t)||^2 = \vec{Y}^t(E - X(X^tX)^{-1}X^t)^t(E - X(X^tX)^{-1}X^t)\vec{Y} = \\ \vec{Y}^t(E - 2X(X^tX)^{-1}X^t + X(X^tX)^{-1}X^t) = \vec{Y}^t(E - X(X^tX)^{-1}X^t)\vec{Y}.$$

Также как и прежде можно оценить $\mathbf{E}(Y^*|X^*) = \langle\vec{X}^*, \vec{a}\rangle$. При этом

$$(\vec{X}^*)^t(X^tX)^{-1}X^t\vec{Y} = (\vec{X}^*)^t\vec{a} + (\vec{X}^*)^t(X^tX)^{-1}X^t\vec{\varepsilon},$$

откуда

$$\widehat{Y}^* \sim \mathcal{N}(\langle X^*, \vec{a}\rangle, \sigma^2(E + \Sigma)),\ \Sigma^2 = (\vec{X}^*)^t(X^tX)^{-1}X^tX(X^tX)^{-1}\vec{X}^* = (\vec{X}^*)^t(X^tX)^{-1}\vec{X}^*.$$

Следовательно, $\widehat{Y}^*$ несмещенная оценка $\mathbf{E}(Y^*|X^*)$.

Сделаем также следующее замечание. Если матрица X неполного ранга (в нашей задаче это означает, что столбцы матрицы линейно зависимы), матрица X^tX необратима и указанные выше формулы для подсчета коэффициентов неприменимы. В этом случае, как нетрудно понять, решение $\widehat{a}$ не единственно, а определяется соотношением

$$X^t\vec{y} = X^tX\widehat{a}. \tag{1}$$

При этом мы по-прежнему находим проекцию $\vec{Y}$ на пространство $L = \{X\vec{a}, \vec{a} \in \mathbb{R}^n\}$, просто пространство L плохо (избыточно) параметризовано. Мы не можем найти оценки параметров $\vec{a}$, обладающие сколько-то хорошими свойствами, но при любом выборе $\vec{a}$ мы получим одно и то же RSS — квадрат расстояния между $\vec{y}$ и L. Следовательно, оценка σ^2 будет обладать теми же свойствами, что и прежде.

1.2 Выбор матрицы плана

Предположим, что мы можем выбрать матрицу X. Как выбрать ее таким образом, чтобы снизить дисперсии оценок $\hat{a}_i$? Мы знаем явные выражения $\mathbf{D}\hat{a}_i = (X^tX)^{-1}_{i,i}\sigma^2$.
Минимизируем диагональные элементы, наложив ограничения на столбцы матрицы X вида $||x_{\cdot,i}|| \leq c_i$.

При этом

$$A = X^tX = \begin{pmatrix} ||x_{\cdot,1}||^2 & \vec{b}^t \\ \vec{b} & F \end{pmatrix},\ F = \widetilde{X}^t\widetilde{X},\ \widetilde{X}_{i,j} = \langle x_{\cdot,i+1}, x_{\cdot,j+1}\rangle,\ \vec{b}_i = \langle x_{\cdot,1}, x_{\cdot,i+1}\rangle.$$

Следовательно,

$$\mathbf{D}\widehat{a}_1 = A^{-1}_{i,i} = \det F/\det A.$$

Свяжем $\det(F)$ и $\det(A)$, используя матрицу

$$C = \begin{pmatrix} 1 & 0 \\ -F^{-1}b & E \end{pmatrix},\ AC = \begin{pmatrix} ||x_{\cdot,1}||^2 & \vec{b}^t \\ \vec{b} & F \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -F^{-1}b & E \end{pmatrix} = \begin{pmatrix} ||x_{\cdot,1}||^2 - \vec{b}^tF^{-1}\vec{b} & \vec{b}^t \\ 0 & F \end{pmatrix}.$$

Поскольку $\det(C) = 1$, то

$$\det(A) = \det(AC) = \left(||x_{\cdot,1}||^2 - \vec{b}^tF^{-1}\vec{b}\right)\det(F).$$

Отсюда

$$\mathbf{D}\widehat{a}_1 = \frac{1}{||x_{\cdot,1}||^2 - \vec{b}^tF^{-1}\vec{b}}$$

будет минимальной при минимальной $\vec{b}^tF^{-1}\vec{b}$. При этом матрица F является положительно определенной, поскольку

$$\vec{u}^tF\vec{u} = \vec{u}^t\widetilde{X}^t\widetilde{X}\vec{u} = \langle\widetilde{X}\vec{u}, \widetilde{X}\vec{u}\rangle > 0$$

при $\vec{u} \neq 0$ (так как $\widetilde{X}$ полного ранга). Следовательно, F^{-1} также положительно определена и $\vec{b}^tF^{-1}\vec{b} \geq 0$, где равенство возможно только при $\vec{b} = 0$. Значит минимум $\mathbf{D}\widehat{a}_{i,i}$ достигается при максимальной доступной норме $||x_{\cdot,1}|| = c_1$ и при $\vec{b} = 0$, что в силу определения $\vec{b}$ означает ортогональность $x_{\cdot,1}$ и $x_{\cdot,i}$ при $i > 1$.

Аналогичные рассуждения приводят нас к выводу о том, что все $x_{\cdot,i}$ должны быть ортогональными и иметь максимальную возможную норму.

1.3 Общая модель без предположения нормальности

Напомню, что идеологически мы рассматриваем задачу

$$\mathbf{E}(Y - f(\vec{X}))^2 \to \min \tag{2}$$

в классе функций $f(\vec{u}) = \langle\vec{u}, \vec{a}\rangle$. В случае, если мы не знаем распределения (X, Y), но знаем, что $Y_i = \varepsilon_i + \langle X_{i,\cdot}, \vec{a}\rangle$, где ε_i н.о.р., то естественно заменить задачу о минимизации (2) на минимизацию

$$\sum_{i=1}^{n}(Y_i - \langle X_{i,\cdot}, \vec{a}\rangle)^2.$$

Вторая задача приводит нас к той же оценке МНК, что и в нормальной модели. При этом решение задачи минимизации $||\vec{y} - X\vec{a}||$ не связано с распределениями (да и вообще с теорией вероятностей), поэтому решением ее как и прежде будет

$$\widehat{a} = (X^tX)^{-1}X^t\vec{Y}.$$

Более того, это как и прежде будет несмещенная оценка $\vec{a}$ с ковариационной матрицей $\sigma^2(X^tX)^{-1}$. Аналогичным образом

$$\widehat{\sigma}^2 = RSS/(n-2),\ RSS = ||\vec{Y} - X\widehat{a}||^2,$$

будет несмещенной оценкой σ^2.

Такая оценка называется оценкой методом наименьших квадратов или МНК-оценкой. Как мы видели, в нормальной модели эта оценка была оценкой максимального правдоподобия. В общей модели это не так, но можно сформулировать следующую теорему:

Теорема 1 (Теорема Гаусса-Маркова)**.** *Пусть $\vec{c}$ — детерминированный вектор. Тогда величина $\langle \vec{c}, \widehat{a} \rangle$ есть несмещенная оценка $\langle \vec{a}, \vec{a} \rangle$ с наименьшей дисперсией среди всех несмещенных оценок вида $d_1(X)Y_1 + ... + d_n(X)Y_n$.*

Доказательство. Как и прежде будем рассматривать фиксированные $x_{i,j}$.
Наша оценка имеет вид

$$\langle \vec{c}, \widehat{a} \rangle = \langle \vec{c}, (X^t X)^{-1} X^t \vec{Y} \rangle = \vec{Y}^t X (X^t X)^{-1} \vec{c}.$$

Для произвольной несмещенной оценки вектор $\vec{d}(X) = (d_1(X), ..., d_n(X))$ удовлетворяет условиям

$$\mathbf{E} \left(\sum_{i=1}^{n} d_i(X) Y_i \right) = \vec{d}(X)^t X \vec{a} = \vec{c}^t \vec{a}$$

при всех $\vec{a}$, то есть $X^t \vec{d}(X) = \vec{c}$. Среди таких векторов мы должны выбрать вектор минимальной длины $||\vec{d}||$, поскольку

$$\mathbf{D} \left(\sum_{i=1}^{n} d_i(X) Y_i \right) = \sigma^2 \sum_{i=1}^{n} d_i(X)^2.$$

Минимальную длину среди таких $\vec{d}$ имеет перпендикуляр из нуля на плоскость $X^t \vec{d} = \vec{c}$. Убедимся, что вектор $\vec{d_0} = X(X^t X)^{-1} \vec{c}$ ортогонален всем векторам внутри этой плоскости. Векторы внутри плоскости удовлетворяют соотношению $X^t \vec{d} = 0$, при этом

$$\langle \vec{d_0}, \vec{d} \rangle = \vec{c}^t (X^t X)^{-1} X \vec{d} = 0.$$

Итак, $\vec{d_0}$ ортогонален плоскости $X^t \vec{d} = \vec{c}$ и лежит в ней. Значит он имеет минимальную длину из векторов этой плоскости, откуда и вытекает теорема Гаусса-Маркова. □