Корреляционная матрица.

Пусть X(N×p) - стандартизованная матрица данных.

Тогда КК равен $r_{ij} = \frac{1}{N}\sum_{k=1}^{N} x_{ki} x_{kj}, \quad i = \overline{1, p}, \quad j = \overline{1, p}$

Матрица, состоящая из таких коэффициентов корреляции, называется корреляционной матрицей и обозначается R(p×p).

$$R(p \times p) = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ & \dots & & \\ r_{p1} & r_{p2} & \dots & r_{pp} \end{bmatrix}$$

Матрица R получается умножением матриц X и X^T.

$R(p\times p) = \frac{1}{N} X^T(p\times N)\ X(N\times p)$

Свойства корреляционной матрицы:

1. R- симметричная матрица.
2. Элементы диагонали $r_{ii} = \cos(\alpha_{ii}) = 1$.

Положительная полуопределенность корреляционной матрицы.

Определение 1: Матрица A(p×p) - называется положительно полуопределенной, если для любого p - мерного вектора $\alpha=(\alpha_1, \alpha_2, \dots, \alpha_p)$ скалярное произведение $(A\alpha, \alpha)\geq 0$.

Определение 2: Матрица A(p×p) - называется положительно определенной, если для любого ненулевого p - мерного вектора $\alpha=(\alpha_1, \alpha_2, \dots, \alpha_p)$ скалярное произведение $(A\alpha, \alpha)>0$.

Свойство КМ:

3. корреляционная матрица - положительно полуопределенная.

Т.е. надо доказать, что $(R\alpha, \alpha)\geq 0$.

$$R(p\times p) * \alpha(p\times 1) = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ \dots & \dots & & \dots \\ r_{p1} & r_{p2} & \dots & r_{pp} \end{bmatrix} * \begin{bmatrix} \alpha_1 \\ \dots \\ \alpha_P \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{p} r_{1i}\alpha_i \\ \dots \\ \sum_{i=1}^{p} r_{pi}\alpha_i \end{bmatrix}$$

и

$$(R\alpha, \alpha) = \sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_i \alpha_j r_{ij} = \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_i \alpha_j (X^i, X^j)$$

$$\Rightarrow R(\alpha,\alpha)=\frac{1}{N}\left(\sum_{i=1}^{p}\alpha_i X_i^i\right)^2\geq 0$$

$$\left(\sum_{i=1}^{p}c_i\xi^i\right)^2=\sum_{i=1}^{p}\sum_{j=1}^{p}c_i c_j\left(\xi^i,\xi^j\right)$$

Замечание 1:

Линейная комбинация равняется нулю ⇔ столбцы линейно зависимы (при ненулевом α). А столбцы есть признаки. Так как все признаки измеряются, т.е. имеется случайная погрешность измерения и N>>p ,то линейная зависимость столбцов практически невозможна ⇒ на практике R можно считать положительно определенной.

Свойства корреляционной матрицы в терминах собственных чисел и собственных векторов.

Рассмотрим равенство

$$R\alpha=\lambda\alpha,$$

где λ - число, $\alpha=(\alpha_1, \alpha_2, \ldots, \alpha_p)$ - неизвестный вектор, R(p×p) - корреляционная матрица.

$R\alpha-\lambda\alpha=0$, отсюда следует, что $(R-\lambda E)\alpha=0$, где E - единичная матрица размера (p×p).

Эта система p линейных однородных уравнений с p неизвестными имеет нетривиальное решение, если определитель $|R-\lambda E|=0$, где определитель матрицы R–λE - полином степени p относительно α. Он имеет p корней, они могут быть как различными, так и одинаковыми.

$$\gamma_0\lambda^n+\gamma_1\lambda^{n-1}+\ldots+\gamma_n=\gamma_0\prod_{i=1}^{l}(\lambda-\lambda_i)^{k_i},\qquad \sum_{i=1}^{n}k_i=n$$

где k_i- кратность корня λ_i,

λ_i – корень полинома.

Если некоторый корень λ_i полинома имеет кратность k_i , то нам удобнее считать, что данный полином имеет k_i корней, равных λ_i. И далее будем считать, что полином n-ой степени имеет ровно n корней, только некоторые из них могут быть равными между собой.

Определение 1: число λ_i, которое является решением уравнения $|R-\lambda E|=0$ называется собственным числом (собственным значением) матрицы R, i=1,...,p. В общем случае эти числа могут быть как действительными, так и комплексными.

Свойство КМ.

4. Все собственные числа симметричной, положительно полуопределенной матрицы, а следовательно (в силу свойств 1-3), все собственные числа корреляционной матрицы являются действительными неотрицательными

числами, а если матрица R положительно определена, то все ее собственные числа положительны.

Замечание 2:

В замечании 1 мы говорили, что на практике R является положительно определенной, следовательно, в силу свойства 4), все ее собственные числа положительны. Отсутствие среди собственных чисел матрицы R равных нулю означает, что определитель такой матрицы не равен нулю, а следовательно матрица имеет обратную.

Рассмотрим $(R-\lambda E)\alpha=0$ и выберем $\lambda=\lambda_i$. Как известно из линейной алгебры, множество решений однородной системы уравнений есть линейное подпространство, размерность которого равна разности между числом уравнений и рангом матрицы, т.е. $p-(p-k_i)=k_i$

Система уравнений $R\alpha=\lambda\alpha$ среди прочих имеет и такие решения,

для которых $|\alpha|^2=\sum_{j=1}^{p}\alpha_j^2=1$.

Определение: Любой вектор α^i единичной длины, являющийся решением системы $R\alpha=\lambda\alpha$, называется собственным вектором матрицы R, соответствующим собственному числу λ_i.

Свойство КМ:

5. Собственные векторы симметричной матрицы, а следовательно и собственные векторы матрицы R, соответствующие различным собственным числам, ортогональны между собой.

Пусть сначала все собственные числа матрицы R различны, тогда имеем p ортогональных между собой собственных векторов. Поскольку они p-мерные, то система векторов $\alpha^1, \alpha^2, \ldots, \alpha^p$ образует базис в пространстве всех p - мерных векторов. Рассмотрим случай, когда некоторые собственные числа кратные. Пусть λ_i имеет кратность $k_i>1$.

Как известно из линейной алгебры, множество собственных векторов, соответствующих такому собственному числу, заполняет единичную сферу в k_i - мерном пространстве. Значит, из этого множества векторов можно всегда выбрать k_i - ортогональных между собой векторов. Кроме того, все остальные собственные векторы, которые соответствуют различным собственным числам, также ортогональны между собой (в силу свойства 5).

6. Из множества собственных векторов корреляционной матрицы (поскольку она симметричная) можно выбрать p – векторов, образующих ортогональный базис пространства p - мерных векторов. В этом базисе каждому собственному числу соответствует столько векторов, какова кратность этого собственного числа.

Регрессионный анализ.

Пусть имеется матрица данных:

$$Z = \begin{bmatrix} z_{11} & \dots & z_{1p} \\ \dots & \dots & \dots \\ z_{N1} & \dots & z_{Np} \end{bmatrix} \quad N >> p$$

Определение: Уравнение, связывающее один из признаков зависимостью от других признаков, называется уравнением регрессии.

$Y = \begin{bmatrix} y_1 \\ \dots \\ y_N \end{bmatrix}$ - вектор-столбец матрицы Z, назовем его целевым признаком

И пусть векторы:

$$X^1 = \begin{bmatrix} x_{11} \\ \dots \\ x_{N1} \end{bmatrix}, \; X^2 = \begin{bmatrix} x_{12} \\ \dots \\ x_{N2} \end{bmatrix}, \dots, \; X^m = \begin{bmatrix} x_{1m} \\ \dots \\ x_{Nm} \end{bmatrix} \qquad m \le p-1$$

- независимые переменные.

$y=f(x_1, ..., x_m, \alpha)$- где α - вектор столбец неизвестных коэффициентов.

Замечание: В корреляционном анализе матрица X- стандартизованная матрица данных, здесь X- просто матрица данных.

Уравнение линейной множественной регрессии

$y=\alpha_1 x^1+\alpha_2 x^2+...+\alpha_m x^m+\varepsilon$ (*)

$X^i = \begin{bmatrix} x_{1i} \\ \dots \\ x_{Ni} \end{bmatrix}$ - вектор независимых переменных,

i =1,..., m, $m \le p-1$

α- вектор неизвестных параметров.

ε- вектор- случайная помеха.

(*)- векторное равенство.

$y_k=\alpha_1 x_{k1}+\alpha_2 x_{k2}+...+\alpha_m x_{km}+\varepsilon_k$, k =1,..., N

ε- случайная компонента, комплексно характеризующая наличие случайных ошибок, неучтенных признаков и т.д.

Введем:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{Nm} \end{bmatrix}$$

$y=X\alpha+\varepsilon$ (**)

Постулаты регрессионного анализа.

В уравнении регрессии фигурируют матрица X, вектор неизвестных параметров α и вектор случайной помехи ε. Поэтому предположения регрессионного анализа касаются этих трех элементов.

1. На α нет ограничений, $\alpha \in R^m$.
2. Вектор ε- случайный . Отсюда следует, что вектор y- случайный.
3. Математическое ожидание всех компонент вектора ε равно 0.
   $M(\varepsilon_k)=0$, k=1,...,N.
4. Ковариация $\text{cov}(\varepsilon_k,\varepsilon_j)=\begin{cases} 0, \text{если } k \neq j \\ \sigma, \text{если } \kappa = j \end{cases}$, k=1,...,N, j=1,...,N,

   т.е. у различных объектов (для различных наблюдений) случайные помехи не коррелированы, а дисперсия конечна и одинакова для всех наблюдений. Т.е. условия проведения наблюдений одинаковы для всех объектов.
5. Матрица X - детерминирована, не случайна, т.е. значения независимых признаков известны точно.
6. Ранг матрицы X равен m, т.е. матрица X имеет m линейно независимых строк или столбцов.

МНК оценка параметров уравнения регрессии.

Суть МНК состоит в следующем: неизвестные параметры выбираются из условия минимума суммы квадратов отклонений фактических значений от расчетных. Сумму квадратов отклонений фактических значений от расчетных обозначим $Q(\alpha)$.

$$Q(a)=\sum_{k=1}^{N}\left(y_k-\left(\alpha_1 x_{k_1}+\alpha_2 x_{k_2}+...+\alpha_m x_{k_m}\right)\right)^2 \rightarrow \min_{\alpha_1,\alpha_2,...,\alpha_m}$$

$$|b|^2=\sum_{i=1}^{n} b_i^2=\begin{pmatrix} b_1 & b_2 & ... & b_n \end{pmatrix}\begin{pmatrix} b_1 \\ ... \\ b_n \end{pmatrix}$$

$Q(\alpha)=(y-x\alpha)^T(y-x\alpha)=y^Ty-y^T(x\alpha)-(x\alpha)^Ty+(x\alpha)^T(x\alpha)=$

$=y^Ty-2\alpha^Tx^Ty+\alpha^Tx^Tx\alpha$ - должно быть минимально по α, поэтому берем производную по α

$$\frac{dQ}{d\alpha} = -2x^T y + 2x^T x\alpha = 0$$

$2x^T y=2x^T x\alpha$

$x^T y=x^T x\alpha$

a - оценка α

$a = \left(X^T X\right)^{-1} X^T y$ - МНК оценка вектора α

Свойства МНК оценки вектора α.

1. МНК оценка линейная по y.
2. МНК оценка несмещенная.

   Определение: Оценка B параметра β называется несмещенной, если M(B)=β

Докажем несмещенность.

Надо доказать, что $M\left(\left(X^T X\right)^{-1} X^T y\right)= \alpha$.

Мы знаем, что $y=X\alpha+\varepsilon$

В силу случайности ε, это уравнение стохастическое.

$M(y)=M(X\alpha+\varepsilon)=M(X\alpha)+M\varepsilon=X\alpha$

Рассмотрим $M\left(\left(X^T X\right)^{-1} X^T y\right)= \left(X^T X\right)^{-1} X^T My = \left(X^T X\right)^{-1} X^T X\alpha = \alpha$

3. МНК-оценка единственная, если справедлив постулат № 6.
4. В классе линейных по Y несмещенных оценок МНК-оценка обладает минимальной дисперсией, т.е. оценка - эффективная.

Уравнение регрессии со свободным членом.

В силу 3-го постулата регрессии, считается, что эффект неучтенных признаков в среднем равен 0. Это предположение на практике малоправдоподобно. Чаще эффект неучтенных факторов не 0, тогда вместо постулата 3 вводят постулат 3′: $M(\varepsilon)=\alpha_{m+1}=const$, где $\alpha_{m+1} \in R$.

Тогда уравнение регрессии будет иметь вид:

$y_k =\alpha_1 x_{k1}+...+\alpha_m x_{km}+\alpha_{m+1}+\varepsilon'_k$, где $\varepsilon'_k = \varepsilon_k - \alpha_{m+1}$, k=1,...,N

Тогда $M\left(\varepsilon'_k\right)= 0$

Мы оказались в условиях предыдущей системы постулатов, поэтому далее будем считать, что уравнение регрессии имеет вид:

Введем вектор, тогда

$y_k=\alpha_1 x_{k1}+...+\alpha_m x_{km}+\alpha_{m+1}x_{k,m+1}+\varepsilon'_k$, где $X^{m+1}=\begin{bmatrix}1\\...\\1\end{bmatrix}=I$

$\alpha=\begin{bmatrix}\alpha_1\\...\\\alpha_{m+1}\end{bmatrix}$, а $X(N\times(m+1))$

Среднее значение расчетных и фактических данных зависимых переменных.

Уравнение регрессии имеет вид: $y_k=\alpha_1 x_{k1}+...+\alpha_m x_{km}+\varepsilon_k$, где x_{km}– фиктивная переменная.

Вычислим МНК оценку неизвестных параметров:

а- оценка α $a=\left(X^T X\right)^{-1}X^T y$.

Находим вектор расчетных значений зависимой переменной:

$$\hat{y}=Xa$$

Тогда $y=Xa+e$, где $e=y-\hat{y}$, где вектор $e=\begin{bmatrix}e_1\\...\\e_N\end{bmatrix}$.

Вектор e называется вектором оценочных отклонений.

Тогда $e=y-\hat{y}$.

МНК оценка удовлетворяет уравнению:

$-X^Ty+X^TXa=0 \Rightarrow X^T(y-Xa)=0 \Rightarrow X^T(y-\hat{y})=0 \Rightarrow X^Te=0.$

Рассмотрим последнюю строку матрицы X^T. Это единицы

$I^Te=0,$

$\sum_{k=1}^{N}e_k=0,\ \frac{1}{N}\sum_{k=1}^{N}e_k=0 \Rightarrow \bar{e}=0$

Вернемся к равенству $y=\hat{y}+e \Rightarrow y_k=\hat{y}_k+e_k$, просуммируем по k и разделим на N:

$\frac{1}{N}\sum_{k=1}^{N}y_k=\frac{1}{N}\sum_{k=1}^{N}\hat{y}_k+\frac{1}{N}\sum_{k=1}^{N}e_k \Rightarrow \bar{y}=\bar{\hat{y}}+\bar{e},\ \bar{e}=0 \Rightarrow \bar{y}=\bar{\hat{y}}.$

Т.е. среднее расчетное значение и среднее фактическое значение совпадают.