Типы данных для многомерных моделей. Многомерный корреляционный анализ: парный, частный и множественный коэффициенты корреляции.

В статистике многомерные данные — это данные, в которых для каждого объекта измеряется несколько признаков.

Коротко:

1. **Матрица «объект-свойство» (или «наблюдение-свойство»)**
   Это таблица, где:
   - **Строки** — это объекты или наблюдения (например, студенты, предприятия, годы).
   - **Столбцы** — это признаки (например, рост, вес, цена, объём продаж).
   - Размер матрицы: $n \times p$, где n — число объектов, p — число признаков.

   Пример:

Студент	Рост (см)	Вес (кг)	Оценка
1	175	70	5
2	180	80	4
3	170	65	5

   Такая матрица используется в регрессии, кластерном анализе, PCA.
2. **Матрица парных сравнений**
   Это квадратная таблица $n \times n$ или $p \times p$, где элемент показывает меру сходства, различия или связи между объектами или признаками.
   Например, в экспертных оценках: 1 — объект i лучше объекта j, 0 — хуже.

Из лекции

Исходные статистические данные могут быть представлены в одной из двух основных форм.

Первая, наиболее распространенная, форма представления исходных данных - **матрица (или таблица) типа «объект-свойство».** Она возникает в ситуации, когда на каждом из n объектов исследуемой совокупности регистрируются значения целого набора признаков, количество которых p. Таким образом, исходные статистические данные могут быть представлены в виде матрицы размерности $[n \times p]$ (пространственная выборка):

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{pp} \end{pmatrix} \tag{1}$$

где x_{ij} - значение признака j на i-ом объекте. При этом i-я строка матрицы характеризует объект i по всем p признакам, а j-й столбец — признак X_j на всех n объектах.

Таким образом, матрица X представляет собой выборку объема n из p-мерной генеральной совокупности.

Аналогично (1) могут быть представлены и данные проведения серии из n опытов на одном и том же объекте с измерением одних и тех же p показателей. В этом случае будем иметь подобную (1) **матрицу типа «наблюдение-свойство»**, или временную выборку. В этой матрице i-я строка будет представлять характеристики i-го опыта, а j-й столбец — изменение показателя X_j при переходе от одного опыта к другому.

Вторая форма представления исходных данных возникает в ряде ситуаций, когда статистические данные получают с помощью специальных опросов, анкет, экспертных оценок. При этом возможны случаи, когда элементом первичного наблюдения является не состояние i-го объекта, а некоторая характеристика γ_{ij} попарного сравнения двух объектов (или признаков), соответственно с номерами i и j. Характеристика γ_{ij} может выражать меру различия или сходства, меру связи или взаимодействия, отношения предпочтения (например, полагают $\gamma_{ij} = 1$, если объект i не хуже объекта j, и $\gamma_{ij} = 0$ в противном случае), меру взаимной коррелированности и т.д.

В этом случае исследователь располагает в качестве массива исходных статистических данных **матрицей парных сравнений** размера $[n \times n]$ (если речь идет о попарном сравнении n объектов) или размера $[p \times p]$ (при попарном сравнении p признаков):

$$\Upsilon = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1k} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2k} \\ \dots & \dots & \dots & \dots \\ \gamma_{k1} & \gamma_{k2} & \cdots & \gamma_{kk} \end{pmatrix}, \qquad k = n \quad \text{или} \quad k = p. \tag{2}$$

Если исследуется n объектов по p показателям в динамике, то есть имеется последовательность из T матриц типа (1) или (2) для моментов времени $t = 1, 2, \dots, T$, то в этом случае речь идет о пространственно-временной выборке.

3.2. Корреляционный анализ многомерной генеральной совокупности

Корреляционный анализ, разработанный К.Пирсоном и Дж.Юлом, является одним из методов статистического анализа, позволяющим оценить *степень зависимости* нескольких признаков.

Корреляционный анализ — это исследование наличия или отсутствия статистической связи между отдельными факторами или переменными. Он представляет собой первый этап статистического исследования зависимостей между анализируемыми переменными и отвечает, по сути дела, на вопрос: *существует ли такая зависимость* или анализируемые признаки независимы?

В связи с основной своей задачей корреляционный анализ решает следующие проблемы:

- как выбрать подходящий *измеритель* статистической связи (коэффициент корреляции: парный, частный, множественный, ранговый или какую-либо другую величину)?
- как оценить его числовое значение по имеющимся выборочным данным?
- как проверить гипотезу о том, что числовое значение измерителя связи действительно свидетельствует о наличии статистической связи (проверка на статистичеки значимое отличие от нуля)?
- как определить структуру связей между компонентами многомерного признака?

2. Многомерный корреляционный анализ

Корреляционный анализ в многомерном случае изучает связи между несколькими переменными. Используются три вида коэффициентов корреляции:

2.1. Парный коэффициент корреляции

- **Что это:** Мера линейной связи между **двумя** переменными X_k и X_j.
- **Как считается:** По формуле ковариации, делённой на произведение стандартных отклонений.
- **Диапазон:** от -1 до $+1$.
- **Особенность:** Не учитывает влияние других переменных. Может быть искажён из-за их косвенного влияния.

2.2. Частный коэффициент корреляции

- **Что это:** Мера линейной связи между двумя переменными **при исключении влияния всех остальных переменных.**
- **Обозначение:** $r_{kj|1,2,\dots}$ (после черты — фиксируемые переменные).
- **Как считается:** Через алгебраические дополнения корреляционной матрицы R.

$$r_{kj|1,2,\dots} = \frac{R_{kj}}{\sqrt{R_{kk}R_{jj}}},$$

где R_{kj} — алгебраическое дополнение элемента r_{kj} матрицы R.

- **Зачем нужен:** Показывает «чистую» связь двух переменных, без вмешательства других.

2.3. Множественный коэффициент корреляции

- **Что это:** Мера линейной связи **одной переменной X_j со всеми остальными переменными** в наборе.
- **Обозначение:** r_j.
- **Как считается:**

$$r_j = \sqrt{1 - \frac{|R|}{R_j}},$$

где $|R|$ — определитель корреляционной матрицы, R_j — алгебраическое дополнение элемента r_{jj}.

- **Квадрат этого коэффициента (R_j^2) называется коэффициентом детерминации** — показывает, какая доля изменчивости переменной X_j объясняется остальными переменными.

Для определения тесноты *линейной* взаимосвязи признака X_j с остальными $p-1$ признаками используется коэффициент множественной корреляции r_j:

$$r_j = r_{j(1,2,\dots,j-1,j+1,\dots,p)} = \sqrt{1 - \frac{|R|}{R_{jj}}}, \quad R_{jj} = (-1)^{j+j} \cdot M_{jj}, \tag{18}$$

где R — определитель матрицы корреляций R , R_{jj} — алгебраическое дополнение элемента r_{jj}, M_{jj} — минор, определитель матрицы, полученной из матрицы R вычеркиванием j-й строки и j-го столбца.

\

3.2.1. Парные и частные коэффициенты корреляции

Основная задача корреляционного анализа состоит в оценке корреляционной матрицы генеральной совокупности по выборке и определении на ее основе оценок парных, частных и множественных коэффициентов корреляции.

Парный коэффициент корреляции измеряет степень *линейной* зависимости между переменными X_k и X_j на фоне влияния остальных $p-2$ показателей системы.

При этом учитывается, что связь каждой пары признаков находится под воздействием связей всех других признаков между собой и с признаками из данной пары.

Формулы для вычисления парных коэффициентов корреляции и вида матрицы корреляций приведены в (8)и(9).

При исследованиии многомерной совокупности вместо исходных данных - матрицы X - часто переходят к их безразмерным характеристикам - стандартизованным значениям. При этом получают матрицу X^* с элементами (10)

Для стандартизованных данных матрица парных корреляций рассчитывается достаточно просто:

$$R = \frac{1}{n} X^{*T} X^* \tag{12}$$

Если факторы в рассматриваемой модели коррелируют друг с другом, то на величине парного коэффициента корреляции сказывается опосредованное влияние других переменных. Для исследования «чистой» связи двух переменных необходимо исключить влияние всех других. Для этого значения «исключаемых» $p-2$ переменных фиксируются на некотором уровне.

Частный коэффициент корреляции характеризует тесноту линейной зависимости между двумя переменными при исключении влияния всех остальных показателей, входящих в модель. Он определяется через матрицу корреляций (8)) или (12)) по формуле:

$$r_{kj/1,2\ldots} = -\frac{R_{kj}}{\sqrt{R_{kk} R_{jj}}}, \tag{13}$$

где R_{kj} - алгебраическое дополнение элемента r_{kj} корреляционной матрицы R. При этом $R_{kj} = (-1)^{k+j} M_{kj}$, где M_{kj} - минор, определитель матрицы, получаемой из матрицы R, путем вычеркивания k-й строки и j-го столбца. В нижнем индексе частного коэффициента корреляции после вертикальной черты указываются фиксированные признаки. Число фиксируемых признаков l определяет порядок частного коэффициента корреляции.

Для случая трех признаков матрица корреляций будет иметь вид:

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix}.$$

Алгебраические дополнения элементов (напрмер, для вычисления частного коэффициента корреляции $r_{12/3}$) можно вычислить по формулам:

$$R_{12} = (-1)^{1+2} \cdot M_{12} = -\begin{vmatrix} r_{21} & r_{23}, \\ r_{31} & 1 \end{vmatrix} = -(r_{12} - r_{13} r_{23}),$$

$$R_{11} = (-1)^{1+1} \cdot M_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2,$$

$$R_{22} = (-1)^{2+2} \cdot M_{12} = \begin{vmatrix} 1 & r_{13}, \\ r_{31} & 1 \end{vmatrix} = 1 - r_{13}^2.$$

Здесь учтена симметричность матрицы R ($r_{ik} = r_{kj}$).

Тогда формула (13) дает следующее выражение для частного коэффициента корреляции $r_{12/3}$:

$$r_{12/3} = \frac{-R_{12}}{\sqrt{R_{22} R_{11}}} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

Последовательно присоединяя к «мешающим» переменным все новые признаки из рассматриваемого набора, можно получить рекуррентные соотношения для подсчета частных коэффициентов корреляции $r_{ij/(l+1)}$ порядка $l+1$ (т. е. при исключении опосредованного влияния $l+1$ мешающих переменных) по частным коэффициентам корреляции порядка l:

$$r_{ij/(l+1)} = \frac{-R_{12}}{\sqrt{R_{22} R_{11}}} = \frac{r_{ij/(l)} - r_{i\ l+1/(l)} r_{j\ l+1/(l)}}{\sqrt{(1 - r^2_{i\ l+1/(l)})(1 - r^2_{j\ l+1/(l)})}} \tag{14}$$

Парный и частный коэффициенты корреляции изменяются в пределах от -1 до $+1$, причем, чем ближе коэффициент корреляции к ± 1, тем сильнее зависимость между переменными. Если коэффициент корреляции больше 0, то связь положительная, а если меньше нуля - отрицательная.

Если частный коэффициент корреляции меньше, чем соответствующий парный коэффициент корреляции, то взаимозависимость между двумя величинами обусловлена частично (или целиком при равенстве нулю частного коэффициента корреляции) воздействием на эту пару остальных, фиксируемых, случайных величин. Если же, наоборот, частный коэффициент корреляции больше соответствующего парного, то фиксируемые величины ослабляют, затушевывают связь.

3.2.3. Множественный коэффициент корреляции и коэффициент детерминации

Для определения тесноты *линейной* взаимосвязи признака X_j с остальными $p-1$ признаками используется коэффициент множественной корреляции r_j:

$$r_j = r_{j(1,2,\ldots,j-1,j+1,\ldots,p)} = \sqrt{1 - \frac{|R|}{R_{jj}}}, \quad R_{jj} = (-1)^{j+j} \cdot M_{jj}, \tag{18}$$

где R — определитель матрицы корреляций R , R_{jj} — алгебраическое дополнение элемента r_{jj}, M_{jj} — минор, определитель матрицы, полученной из матрицы R вычеркиванием j-й строки и j-го столбца.

С помощью множественного коэффициента корреляции по мере приближения r_j к единице делают вывод о тесноте связи признака X_j с остальными признаками (но не о направлении связи).

Квадрат множественного коэффициента корреляции называется **коэффициентом детерминации**:

$$R_j^2 = r_j^2 = 1 - \frac{|R|}{R_{jj}}. \tag{19}$$

Коэффициент детерминации показывает, какая доля вариации исследуемой переменной X_j объясняется вариацией остальных переменных.

Для проверки значимости множественного коэффициента корреляции (или его квадрата - коэффициента детерминации) используется критерий Фишера.

Проверяемая гипотеза $H_0 : \rho_{j/\ldots} = 0$ - генеральный множественный коэффициент корреляции равен нулю.

По выборочным данным оценивается значение F- статистики:

$$F_{\text{в}} = \frac{r_j^2/(p-1)}{(1-r_j^2)/(n-p)}, \tag{20}$$

которая при выполнении гипотезы H_0 имеет распределение Фишера с $k_1 = p-1$ и $k_2 = n-p$ степенями свободы.

Множественный коэффициент корреляции считается значимым, т.е. имеет место линейная статистическая зависимость между переменной X_j и остальными факторами, если выполнено условие:

$$F_{\text{в}} > F_{\text{кр}}(\alpha, k_1, k_2), \tag{21}$$

где $F_{\text{кр}}$ определяется по таблицам F-распределения.

Если условие (21) не выполняется, то статистическая линейная связь между переменной X_j и остальными факторами считается отсутствующей при уровне значимости α.

Множественный коэффициент корреляции и коэффициент детерминации служат для отбора факторов в модель (например, множественной регрессии), обнаружения мультиколлинеарности факторов.