# Прикладной многомерный статистический анализ

## Теоретический минимум

**Лектор:** Хохлов Юрий Степанович

2016 г.

**Вопрос №1** Основные задачи многомерного статистического анализа:

1. **Корреляционный анализ** изучает наличие и силу связи между случайными величинами. Используются коэффициенты корреляции.

2. **Регрессионный анализ.** Выделяются объясняемая переменная $Y$(отклик) и несколько объясняющих факторов $X_1, \ldots, X_m$. Если обнаружено влияние факторов $X_1, \ldots, X_m$ на $Y$, то пытаются найти вид их связи, в следующем формате

$$Y = g(X_1, \ldots, X_m) + \varepsilon$$

   где, $g(X_1, \ldots, X_m)$ - влияние факторов, а $\varepsilon$ - то, что не удалось объяснить.

3. **Методы снижения размерности.** Обычно размерность пространства факторов $d$ - велика. Пытаются найти небольшое количество (возможно новых) факторов, которые достаточно хорошо представляют изменения в рамках исходной совокупности. Для этих целей применяют факторный анализ, главные компоненты и т.д.

4. **Дисперсионный анализ.** Поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях. ~~Суть дисперсионного анализа сводится к изучению влияния одной или нескольких независимых переменных, обычно именуемых факторами, на зависимую переменную.~~

5. **Дискриминантный анализ.** Предположим, что наши данные неоднородны. Например они выбраны из 2-ух совокупностей с разным средним. В таком случае, основной задачей является нахождение процедуры позволяющей разделить все наблюдения по признаку принадлежности к одной из совокупностей.

6. **Кластерный анализ.** Схожий с дискриминантным анализ, но отсутствуют знания о классах. Визуально видно, что данные как-то группируются в несколько классов. Основной задачей является нахождение некоторого правила, объединяющего точки в один класс.

**Вопрос №2** Гильбертово пространство случайных величин.

Линейное пространство случайных величин 1) на котором задано скалярное произведение, 2) доказана сходимость и относительная сходимость в среднем квадратическом, 3) которое полно отосительно этой сходимости называется Гильбертовым пространством случайных величин.

Пусть $L_2$ есть пространство случайных величин $\xi$,таких, что

$$E(|\xi|^2) < \infty$$

Определим на $L_2$ следующий функционал: $\forall \xi_1, \xi_2 \in L_2$, по определению

$$(\xi_1, \xi_2) := E(\xi_1 \cdot \xi_2) \tag{1}$$

Обладающий следующими свойствами:

1. $(\xi, \xi) \geq 0$ и $(\xi, \xi) = 0 \Leftrightarrow \xi = 0$
2. $(\xi_1, \xi_2) = (\xi_2, \xi_1)$
3. $(C_1\xi_1 + C_2\xi_2, \xi_3) = C_1(\xi_1, \xi_3) + C_2(\xi_2, \xi_3)$

В таком случае этот функционал является скалярным произведением.

Нормой случайной величины $\xi$ из $L_2$ называется число

$$||\xi|| = \sqrt{(\xi, \xi)} = \sqrt{E(|\xi^2|)} \tag{2}$$

Последовательность случайных величин $\xi_n$ из $L_2$ сходится в среднем квадратическом к случайной величине $\xi_0$, если норма $\xi_n - \xi_0$ стремится к нулю при $n \to \infty$:

$$||\xi_n - \xi_0||^2 = E(|\xi_n - \xi_0|^2) \to 0$$

Если

$$E(|\xi_n - \xi_m|^2) \to 0, \text{когда } m, n \to \infty$$
$$\exists \xi_0 \in L_2 : E(|\xi_n - \xi_0|^2) \to 0$$

то пространство $L_2$ является Гильбертовым пространством.

**Вопрос №3** Что такое наилучшая линейная оценка(приближение)?

Пусть $H \subset L_2$ замкнутое линейное подпространство, а $\eta \in L_2$ случайная величина, для которой необходимо найти линейное приближение в $H$. Тогда случайная $\hat{\eta}$ есть **наилучшее линейное приближение** $\eta$ в пространстве $H$, если

1. $\hat{\eta} \in H$
2. $\forall \xi \in H$ :

$$||\eta - \hat{\eta}||^2 \leq ||\eta - \xi||^2 \Leftrightarrow E(|\eta - \hat{\eta}|^2) \leq E(|\eta - \xi|^2)$$

**Вопрос №4** Лемма о перпендикуляре.

Если $\hat{\eta}$ есть наилучшее линейное приближение $\eta$ в пространстве $H$ тогда:

1. $\hat{\eta} \in H$
2. $\forall \xi \in H$ :

$$(\eta - \hat{\eta}, \xi) = E((\eta - \hat{\eta}) \cdot \xi) = 0 \tag{3}$$

**Впорос №5** Простой коэффициент корреляции.

Простым или парным коэффициентом корреляции невырожденных случайных величин $\xi_1$ и $\xi_2$ называется число:

$$\rho(\xi_1, \xi_2) := \frac{cov(\xi_1, \xi_2)}{\sqrt{D(\xi_1) \cdot D(\xi_2)}}$$

Измеряет зависимость двух величин.

$|\rho(\xi_1, \xi_2)|^2$ измеряет долю изменчивости $\xi_2$, которую можно объяснить линейным влиянием $\xi_1$.

$1 - |\rho(\xi_1, \xi_2)|^2$ измеряет ту часть изменчивости $\xi_2$, которую не удалось объяснить линейным влиянием $\xi_1$ и необходимо привлечь дополнительные факторы.

**Вопрос №6** Множественный коэффициент корреляции.

Множественный коэффициент корреляции пытается объяснить поведение $y$ с помощью нескольких факторов $x_1, x_2, \ldots, x_m; m \geq 2$. Пусть $\hat{y} = \alpha + \beta_1 x_1 + \ldots + \beta_m x_m$ - наилучшее линейное приближение $y$. Тогда, **множественным коэффициентом корреляции** случайной величины $y$ и набора $x_1, \ldots, x_m$ называется число

$$\rho_{y.x_1,\ldots,x_m} := \rho(y, \hat{y})$$

$\rho^2_{y.x_1,\ldots,x_m}$ показывает, какую долю изменчивости $y$ можно объяснить линейным влиянием выбранных факторов.

**Вопрос №7** Частный коэффициент корреляции.

Пусть изучаем зависимость $y$ от факторов $x_1, \ldots, x_m$. Выделим некоторый фактор $x_k$. Пусть $C$ - набор всех остальных факторов, а $y^C$ - наилучшее линейное приближение $y$ через все $x_i$, кроме $x_k$ - $C$. $x_k^C$ - наилучшее приближение самого $x_k$ через $C$. Тогда ошибки будут равны

$$z_y = y - y^C, z_{x_k} = x_k - x_k^C$$

**Частным коэффициентом корреляции** случайной величины $y$ и $x_k$, когда устранено влияние всех остальных факторов называется:

$$\rho_{yx_k.C} = \rho(z_y, z_{x_k})$$

$\rho^2_{yx_k.C}$ - показывает какую долю необъяснённой дисперсии удалось объяснить введением нового фактора. Частный коэффициент корреляции измеряет чистое влияние фактора $x_k$ на $y$.

**Вопрос №8** Модель и основные ограничения множественной линейной регрессии.

**Модель:** проводится $N$ одновременных измерений величины $Y$ и факторов $X_1, \ldots, X_d$. При этом предполагается, что

$$Y_j = g(X_{j1}, \ldots, X_{jd}) + \varepsilon_j$$

**Ограничения:**

1. Модель линейна по параметрам, т.е.:

$$Y_j = \alpha + \beta_1 \cdot X_{j1} + \ldots + \beta_d \cdot X_{jd} + \varepsilon_j$$

2. Факторы $X_{jk}$ измерены точно, т.е. это не случайные величины.

3. $E(\varepsilon_j) = 0$ для любого $j$. Т.е. иксы в среднем правильно описывают поведение $Y$ и нет систематических ошибок.

4. Дисперсия $D(\varepsilon_j) = \sigma^2 \forall j$ одинакова для всех $j$. Условие гомоскедастичности.

5. $cov(\varepsilon_j, \varepsilon_k) = 0$, когда $j \neq k$. Т.е. ошибки не коррелируют.

6. $\varepsilon_j$ имеет нормальное распределение.

**Вопрос №9** Описание метода наименьших квадратов для оценки параметров.

Пусть имеются следующие параметры модели:

$$\Theta_0, \Theta_1, \ldots, \Theta_m \text{ и } \sigma^2 = D(\varepsilon_j)$$

Тогда для оценки параметров $\Theta_j$ необходимо решить следующую экстремальную задачу:

$$Q(\Theta) = ||Y - X \cdot \Theta||^2 = \sum_{j=1}^{N} [Y_j - \Theta_0 - \Theta_1 X_{j1} - \ldots - \Theta_m X_{jm}]^2 \to \min_{\Theta}$$

Необходимое условие экстремума:

$$\frac{\partial Q}{\partial \Theta_k} = 0, k = \overline{0, m}$$

После несложных преобразований получаем систему нормальных уравнений:

$$X^T \cdot X \cdot \Theta = X^T \cdot Y$$

Отсюда оценка параметра $\Theta$ по методу наименьших квадратов будет:

$$\hat{\Theta} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$$

Она является линейной,несмещенной и по теореме Гаусса - Маркова является оптимальной в среднем квадратическом в классе всех линейных и несмещенных оценок.

**Вопрос №10** Явный вид оценок параметров по МНК.

Оценка парамерта $\Theta$:

$$\hat{\Theta} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$$

Оценка для среднего квадратического вектора остатков $\sigma^2 = D(\varepsilon_j) = E(\varepsilon_j^2)$ будет:

$$S^2 := \frac{1}{N - (m+1)} \sum_{j=1}^{N} e_j^2$$

где $e = Y - \hat{Y}$.

**Вопрос №11** Общая схема проверки гипотезы о параметре.

Статистической гипотезой называется утверждение о распределении генеральной совокупности, соответствующее некоторым представлениям об изучаемом явлении. В частном случае это может быть утверждение о значениях параметров нормально распределенной генеральной совокупности. Статистические гипотезы обычно рассматривают, генеральные совокупности, одна из которых может представлять собой теоретическую модель, а о второй судят по выборке из нее. В других случаях обе генеральные совокупности представлены выборками. Изначально формулируются 2 гипотезы $H_0$ и $H_1$. Нулевая гипотеза гласит:

*Между двумя генеральными совокупностями нет ожидаемого различия*

Соответственно, альтернативная гипотеза $H_1$ заявляет об обратном.

Схема проверки:

1. Определяется уровень значимости $\alpha = (0.1, 0.05, 0.001)$.

2. По выборочным данным вычисляется значение некоторой новой случайной величины $K_{\text{набл}}$, которая имеет известное стандартное распределение. Например, $T$ - распределение или $F$ - распределение.

3. По таблицам соответствующего распределения(нормального, $T$ - распределения и т.д.), находится значение критической константы при соответствующем уровне значимости - $k(\alpha)$.

4. Если реально полученное наблюдаемое значение $K_{\text{набл}}$ статистики $K$ больше, по модулю, чем $k(\alpha)$, то гипотеза $H_0$ отвергается.

5. Если выяснилось обратное, то говорят, что $H_0$ не противоречит экспериментальным данным.

**Вопрос №12** Для чего используется T-критерий.

T-критерий или критерий Стьюдента используется для проверки гипотез, где выборки имеют распределение близкое к нормальному. В случае одной выборки применяется для проверки какого-то утверждения, например:

$$E(X) = m$$

В случае двух выборок проверяются различия между ними.

В нашем курсе $T$ - критерий использовался для проверки гипотез:

1. О том, что случайные величины $X$ и $Y$ независимы, тогда и только тогда, когда $\rho(X, Y) = 0$: $T := \dfrac{R}{\sqrt{1 - R^2}} \cdot \sqrt{N - 2}$

2. О том, что повторная выборка из одномерного нормального распределения $N(a, \dfrac{\vartheta^2}{N})$ имеет такое же мат.ожидание, как и первая выборка: $T := \dfrac{\overline{X} - a}{S_1}\sqrt{N}$

3. О значимости влияния отдельного фактора в присутствии всех остальных: $T := \dfrac{\hat{\theta}_k}{S_k}$

**Вопрос №13** Основное различие T-критерия и F-критерия в задаче проверки значимости влияния фактора

Пусть есть такая линейная модель регрессии:

$$Y = X \cdot \Theta + \varepsilon$$

$$Y_j = \Theta_0 + \Theta_1 \cdot X_{j1} + \ldots + \Theta_n \cdot X_{jn} + \varepsilon_j$$

Проверяется, что

- $H_0 : \Theta_k = 0$ (фактор не значим), против
- $H_1 : \Theta_k \neq 0$ (фактор значим).

В данной задаче:

1. $T$ - критерий проверяет значимость влияния отдельного фактора в присутствии всех остальных факторов.
2. $F$ - критерий оценивает чистое влияние отдельного фактора, когда все остальные факторы устранены.

Очевидно, что если фактор всего один, то оба критерия эквивалентны.

**Вопрос №14** Адекватность модели. Постановка задачи.

Модель регрессии $Y = X \cdot \Theta + \varepsilon$ считается адекватной, если предложеный набор факторов $X_1, \ldots, X_n$ оказывает совместно значимое влияния на $Y$.

Формально, проверяется гипотеза:

- $H_0 : \Theta_1 = \Theta_2 = \ldots \Theta_m = 0$, против
- $H_1 : \Theta_k \neq 0$.

Если $H_0$ отвергается, то модель считается адекватной. В противном случае, выбранный набор фактор не оказывает значимого влияния $Y$ и модель не адекватна. Необходимо выбрать другие факторы.

**Вопрос №15** Коэффициент детерминации и что он измеряет.

Можно показать, что справедливо следующее тождество:

$$\sum_{j=1}^{N} [Y_j - \overline{Y}]^2 = \sum_{j=1}^{N} [Y_j - \hat{Y}_j]^2 + \sum_{j=1}^{N} [\hat{Y}_j - \overline{Y}]^2$$

Для сокращения, эту формулу записывают по другому:

$$TSS = ESS + RSS$$

где

1. $TSS$ - полная сумма квадратов.

2. $ESS$ - сумма квадратов остатков.

3. $RSS$ - объясненная сумма квадратов.

Коэффициентом детерминизации называется число:

$$R^2 := \frac{RSS}{TSS}$$

$R^2$ - это оценка квадрата множественного коэффициента корреляции. Иными словами это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными. Чем ближе он к 1, тем лучше выбранная модель.

**Вопрос №16** Основная задача в однофакторном дисперсионном анализе. Пусть имеется следующий набор измерений:

$$Y_{kj} = \mu_k + \varepsilon_{kj}; \quad k = \overline{1,m}, j = \overline{1,N} \tag{1}$$

где, $k$ - уровень фактора, $j$ - номер измерения. Предполагается:

1. $\mu_k$ - неслучайные вещественные числа;

2. $\varepsilon_{kj}$ - имеют многомерное нормальное распределение $N(0,\sigma^2)\forall k,j$;

3. $\varepsilon_{kj}$ - независимы для любых $k,j$. Запишем модель в новом виде:

$$\mu = \frac{1}{m}\sum_{k=1}^{m}\mu_k; \quad \alpha_k = \mu_k - \mu \Rightarrow \sum \alpha_k = 0$$

Отсюда, $Y_{jk} = \mu + \alpha_k + \varepsilon_{jk}$, что эквивалентно (1).

**Основная задача:** Есть ли различия в поведении $Y$ на различных уровнях. Более формально можно записать:

Проверяется гипотеза:

1. $H_0 : \alpha_k = 0 : \quad \forall k$, против

2. $H_1 : \exists k : \alpha_k \neq 0$

**Вопрос №17** Основная задача в двухфакторном дисперсионном анализе. Пусть имеется следующая модель измерений (для каждого сочетания факторов по одному измерению):

$$Y_{kj} = \mu_{kj} + \varepsilon_{kj}; \quad k = \overline{1,m}; j = \overline{1,N}; n = m \cdot N \tag{1}$$

Предполагается, что на исследование влияют 2 фактора:

1. $k$ - уровень первого фактора;

2. $j$ - уровень второго фактора;

Преполагается, что :

1. $\mu_{kj}$ - константа;

2. $\varepsilon_{kj}$ - случайные, независимые;

3. $\varepsilon_{kj}$ - имеет многомерное нормальное распределение $N(0, \sigma^2)$

Введем следующие обозначения:

$$\mu_{\bullet\bullet} = \frac{1}{n}\sum_{kj}\mu_{kj}; \quad \mu_{\bullet j} = \frac{1}{m}\sum_{k}\mu_{kj}; \quad \mu_{k\bullet} = \frac{1}{N}\sum_{j}\mu_{kj}$$

Тогда, $\tau_j := \mu_{\bullet j} - \mu_{\bullet\bullet}$ - эффект столбца, а $\zeta_k := \mu_{k\bullet} - \mu_{\bullet\bullet}$ - эффект строки.

**Основная задача:** Есть ли влияние факторов - эффект столбца или эффект строки?

Более формально, для эффекта строки, проверяется разница в средних по строкам. Ещё более формально, проверяется:

- $H_0 : \zeta_1, \zeta_2, \ldots, \zeta_m = 0$, против
- $H_1 : \exists k : \zeta_k \neq 0$.

**Вопрос №18** Основная задача дискриминантного анализа. Пусть каждый изучаемый объект характеризуется парой чисел $X = (X_1, X_2)$, преполагается, что $(X_1, X_2)$ имеет двумерное нормальное распределение со средним $\mu = \dfrac{\mu_1}{\mu_2}$ и некоторой матрицей ковариаций $\sum$.

Пусть имеется две совокупности, которые различаются средними $\mu^{(1)} = (\mu_1^{(1)}, \mu_2^{(1)})$ и $\mu^{(2)} = (\mu_1^{(2)}, \mu_2^{(2)})$, но имеют одну и ту же матрицу ковариаций $\sum$.

**Основная задача:** Отнести вновь поступивший объект $(X_1, X_2)$ к одной из этих совокупностей.

**Вопрос №19** Кластерный анализ: постановка задачи.

Пусть из некоторого множества $\pi$ *(генеральной совокупности)* отобрано $n$ объектов $I = (I_1, \ldots, I_n)$. У каждого из этих объектов измеряется несколько характеристик $C = (C_1, \ldots, C_p)^T$. Преполагается, что эти характеристи являются *количественными*. Пусть, $x_{ij}$ есть результат измерения

$i$-й характеристики у объекта $I_j$. Тогда $X_j = (x_{1j}, \ldots, x_{pj})^T$ есть измерения всех его характеристик у объекта $I_j$. В итоге имеем матрицу измерений:

$$X = (X_1, \ldots, X_n).$$

Далее каждый набор $X_j$ рассматривается как вектор в пространстве $\mathbb{R}^p$. Пусть $m < n$, тогда **основная задача:** на основе измерений $X$ разбить множество объектов $I$ на $m$ классов*(кластеров)* $\pi_1, \ldots, \pi_m$, так чтобы:

1. каждый объект $I_j$ принадлежал одному и только одному классу;
2. объекты внутри одного класса были бы *"сходными"*;
3. объекты из разных классов были бы *"несходными"*.

Интуитивно ясно, что объекты $I_j$ и $I_k$ нужно отнести в один класс, если расстояние между $X_j$ и $X_k$ будет *достаточно малым*, а между точками из разных классов *достаточно большим*.

**Вопрос №20** Кластерный анализ: последовательное построение ~~факторов~~ (кластеров?).

1. Сначала все объекты рассматривают как отдельные кластеры.
2. Выбирают 2 порога $s$ и $r$.
3. Если все кластеры находятся на расстоянии более, чем $s$, то процедура заканчивается.
4. Если есть кластеры, которые ближе друг к другу, чем $s$, то находим два наиболее близких и объединяем их.
5. Находим расстояния внутри кластеров и расстояния между кластерами.
6. Процедура продолжается до тех пор, пока расстояния внутри всех класетров не более чем $r$, а между кластерами не более чем $s$.