**Дополнительные вопросы к экзамену**

по курсу «Прикладной многомерный статистический анализ»

1. **Основные задачи многомерного статистического анализа:**

**- корреляционный анализ**

Изучается наличие и, если присутствует, сила связи между случайными величинами. Для этого используют коэффициент корреляции.

**- регрессионный анализ**

Выделяется объясняемая переменная Y (отклик) и несколько (возможно 1) объясняющих факторов $X\_{1}, …, X\_{m}$.

Если обнаружено сильное (значимое) влияние факторов на Y, то пытаются найти вид их связи в следующей форме: $Y=g\left(X\_{1}, ..,X\_{m}\right)+ε$.

$g\left(X\_{1}, ..,X\_{m}\right)$ *–* влияние факторов

$ε- $влияние неучтенных факторов

**- снижение размерности**

Обычно размерность d велика. Пытаются найти небольшое количество факторов (как старых, так и новых, выраженных через старые), которые достаточно хорошо представляют изменчивость в рамках исходящей совокупности. Например: методы факторного анализа, метод главных компонент, ~визуальный метод (?)

**- дисперсионный анализ**

метод в математической статистике, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях

(википедия)

**- дискриминантный анализ**

Предположим, что данные неоднородны: например, выбраны из двух совокупностей с разными средними.

*Основная задача:* найти процедуру (правило), позволяющее разделить все наблюдения по признаку принадлежности к одной из совокупности.

**- кластерный анализ**

Визуально видно, что данные как-то группируются в несколько классов. Заранее неизвестно, сколько классов.

*Задача:* предложить некоторое правило объединения точек в группу

1. **Гильбертово пространство случайных величин**

*Гильбертово пространство* $L\_{2}$ – линейное пространство со скалярным произведением и которое является полным относительно сходимости, порожденной этим скалярным произведением (в данном случае сходимость в среднем квадратическом), то есть если $E\left(\left|ε\_{n}-ε\_{m}\right|^{2}\right)\rightarrow 0, n\rightarrow \infty , m\rightarrow \infty => ∃ε\_{0}\in L\_{2}$: $E\left(\left|ε\_{n}-ε\_{0}\right|^{2}\right)\rightarrow 0, n\rightarrow \infty $

(вообще изначально мы рассматриваем случайные величины такие, у которых $E\left(\left|ε\right|^{2}\right)< \infty $)

1. **Что такое наилучшая линейная оценка**

$H \in L\_{2}$ – замкнутое линейное подпространство, $η\in L\_{2}$

Случайная величина $\hat{η}$ – *наилучшее линейное приближение* случайной величины $η в H$, если

* $\hat{η}\in H$
* $для любого ε\in H: \left‖η-\hat{η}\right‖^{2}= E\left(\left|η-\hat{η}\right|^{2}\right)\leq E\left(\left|η- ε\right|^{2}\right)= \left‖η- ε\right‖^{2}$
1. **Лемма о перпендикуляре**

$\hat{η}$ – наилучшее линейное приближение $η в H$ ⬄

* $\hat{η}\in H$
* $для любого ε\in H:\left(η- \hat{η}, ε\right)=E\left(\left(η- \hat{η}\right)\*ε\right)=0$ (это на самом деле СЛАУ)
1. **Простой коэффициент корреляции и что он измеряет**

*Простой (парный) коэффициент корреляции* невырожденных (не const, иначе $D\left(ε\right)=0$) случайных величин $ε\_{1}, ε\_{2}$ – число $ρ\left(ε\_{1}, ε\_{2}\right)=\frac{cov(ε\_{1}, ε\_{2})}{\sqrt{D(ε\_{1})D(ε\_{2})}}$.

* $\left(ρ\left(ε\_{1}, ε\_{2}\right)\right)^{2}$ измеряет долю изменчивости $ε\_{2}$, которую можно объяснить линейным влиянием $ε\_{1}$
* $1- \left|ρ\left(ε\_{1}, ε\_{2}\right)\right|^{2}$ измеряет ту часть изменчивости $ε\_{2}$, которую не удалось объяснить линейным влиянием $ε\_{1}$ и необходимо привлечь другие факторы

*Свойства:*

1. $\left|ρ\left(ε\_{1}, ε\_{2}\right)\right|^{2}\leq 1$
2. $ρ\left(ε\_{1}, ε\_{2}\right)=0<=>cov\left(ε\_{1}, ε\_{2}\right)=0, т.е. ε\_{1}, ε\_{2}$ – не коррелированы
3. Если $η\_{1}=a\_{1}ε\_{1}+b\_{1}; η\_{2}=a\_{2}ε\_{2}+b\_{2};a\_{1}a\_{2}\ne 0, то ρ\left(η\_{1}, η\_{2}\right)= ρ\left(ε\_{1},ε\_{2}\right)\*sgn(a\_{1}a\_{2})$
4. Если $ρ\left(ε\_{1}, ε\_{2}\right)=+1, то ∃a>0, b\in R^{1}: ε\_{2}=aε\_{1}+b$

Если $ρ\left(ε\_{1}, ε\_{2}\right)=-1, то ∃a<0, b\in R^{1}: ε\_{2}=aε\_{1}+b$

1. **Множественный коэффициент корреляции и что он измеряет**

Пытаемся объяснить поведение Y с помощью нескольких факторов $X\_{1}, …, X\_{m}, m\geq 2$ (совокупное влияние всех факторов вместе)

Пусть $\hat{Y}= α+ β\_{1}X\_{1}+…+ β\_{m}X\_{m}$ –наилучшее линейное приближение Y

*Множественный коэффициент корреляции* Y и набора случайных величин $X\_{1}, …, X\_{m}$ – число $ρ\_{Y\*X\_{1},…,X\_{m}}= ρ(Y, \hat{Y})$

* $ρ\_{Y\*X\_{1},…,X\_{m}}^{2}$ - показывает, какую долю изменчивости Y можно объяснить линейным влиянием выбранных факторов
* $1-ρ\_{Y\*X\_{1},…,X\_{m}}^{2}$ – то, что вызвано неучтенными факторами
1. **Частный коэффициент корреляции и что он измеряет**

Изучаем зависимость Y от факторов $X\_{1}, …, X\_{m}$ (чистое влияние одного фактора)

Выберем некоторый фактор $X\_{k}\in X\_{1}, .., X\_{m}. $

$С$ – набор остальных факторов.

$Y^{C}$ – наилучшее линейное приближение Y через C

$X\_{K}^{C}$ – наилучшее линейное приближение $X\_{k}$ через C

$Z\_{Y}=Y-Y^{C}, Z\_{X\_{K}}=X\_{K}-X\_{K}^{C}$

*Частный коэффициент корреляции* Y и $X\_{k}$, когда устранено влияние всех остальных факторов, - число $ρ\_{YX\_{K}\*C}= ρ(Z\_{Y}, Z\_{X\_{K}})$

*Свойства:*

1. $\left|ρ\_{YX\_{K}\*C}\right|\leq 1$
2. $1- ρ\_{Y\*X\_{1},…,X\_{k}}^{2}=(1- ρ\_{Y\*X\_{1},.., X\_{K-1}}^{2})(1- ρ\_{YX\_{K}\*X\_{1},,.., X\_{K-1}}^{2})$

$ρ\_{YX\_{K}\*X\_{1},,.., X\_{K-1}}^{2}$ – показывает, какую долю необъяснимой дисперсии удалось объяснить введением еще одного фактора. (когда факторов много простой коэф корреляции может давать неверную инфу, как фактор влияет на сл.в.)

1. **Множественная линейная регрессия: модель и основные ограничения**

*Постановка задачи:*

Y – объясняемая переменная, $X\_{1}, …, X\_{d}$- объясняющие переменные

Представление: $Y=f\left(X\_{1}, ..,X\_{d}\right)+ε$.

Необходимо найти функцию $g\left(X\_{1}, ..,X\_{d}\right)$ наилучшим образом приближающую Y с помощью факторов.

Если расстояние между сл.в. измеряется в среднем квадратическом, то наилучшее приближение задается по правилу: $g\left(x\_{1}, ..,x\_{d}\right)=E(Y|X\_{1}=x\_{1},…,X\_{d}=x\_{d})$, тогда g – функция регрессии.

Вычислять мат ожидание очень сложно (распределение может быть не тривиальным), поэтому основная задача: по экспериментальным данным оценить функцию регрессии.

*Модель:*

Проводится N одновременных наблюдений Y и факторов $X\_{1}, …, X\_{d}$. При этом предполагается, что $Y\_{j}=g\left(x\_{1j}, ..,x\_{dj}\right)+ε\_{j}, j=1,…,N$

*Ограничения:*

1. Модель линейна по параметрам, то есть $Y\_{j}= α+β\_{1}X\_{1j}+…+β\_{d}X\_{dj}+ε\_{j}$
2. Факторы $X\_{kj}$ измерены точно, то есть это не случайные величины
3. $E\left(ε\_{j}\right)=0$, то есть нет систематических ошибок
4. $D\left(ε\_{j}\right)=σ^{2}$ – дисперсия одинакова для всех j – условие гомоскедастичности
5. $cov\left(ε\_{j}, ε\_{i}\right)=0, j\ne i$ – ошибки некоррелированы
6. $ε\_{j}$ – имеет нормальное распределение

$=>Y$ имеет многомерное нормальное распределение со средним $Xθ$ и матрицей ковариации $Σ=σ^{2}I\_{N}$

$X=\left(X\_{jk}\right), j=1,…,N;k=0,..,d;X\_{0}=\left(1…1\right)^{T}; θ=\left(α β\_{1}… β\_{d}\right)^{T}$

1. **Описание МНК для оценки параметров**

$θ\_{0}, θ\_{1}, …, θ\_{d}, σ^{2}=D\left(ε\_{j}\right)$ *–* параметры модели.

Для оценки параметров модели решаем следующую экстремальную задачу:

$Q\left(θ\right)= \left‖Y-Xθ\right‖^{2}= \sum\_{j=1}^{N}\left[Y\_{j}-\left(θ\_{0}+ θ\_{1}X\_{j1}+…+θ\_{d}X\_{jd}\right)\right]^{2}\rightarrow \min\_{θ}$

Имеем невырожденную ситуацию, если векторы $X\_{j}, X\_{i}$ линейно независимы (экв матрица X имеет ранг d+1).

Задача на минимум решается с помощью необходимых условий на экстремум:

$\frac{dQ}{dθ\_{k}}=0, k=1,…,d$

После преобразований получаем: $X^{T}Xθ=X^{T}Y$ – система нормальных уравнений. Если $X^{T}X$ невырождена, то отсюда следуют оценки параметров.

(фактически решается задача о наилучшей линейной оценке Y в линейном пространстве, порожденном случайными векторами $\vec{X\_{0}}, …, \vec{X\_{d}}$)

1. **Явный вид оценок параметров по МНК**

$\hat{θ}=\left(X^{T}X\right)^{-1}X^{T}Y$

Предсказанные значения: $\hat{Y}=X\hat{θ}$

Остатки: $e=Y- \hat{Y}$ ($e\_{j}$ наследуют многие свойства $ε\_{j}$)

$σ^{2}=D\left(ε\_{j}\right)=E(ε\_{j}^{2})$(средние у $ε\_{j}$ = 0)

Если бы $ε\_{j}$ были известны, то $\frac{1}{N}\sum\_{j=1}^{N}ε\_{j}^{2}\rightarrow \_{p}E\left(ε\_{j}^{2}\right)=σ^{2} $

Предлагается следующая оценка для $σ^{2}$:

$\hat{σ^{2}}=S^{2}=\frac{1}{N-(m+1)}\sum\_{j=1}^{N}e\_{j}^{2}$

(изменение нормировки нужно для того, чтобы получить несмещенную оценку)

Оценка $\hat{θ}:$

1. Линейная
2. Несмещенная - (математическое ожидание оценки равно оцениваемому параметру)
3. Если выполнены ограничения 1)-5), то оптимальная в среднем квадратическом в классе всех линейных несмещенных оценок
4. Если выполнены ограничения 1-5) и диагональные элементы матрицы $X^{T}X\rightarrow 0$, то оценка состоятельная – (оценка, сходящаяся по вероятности к своему параметру, количество наблюдений стремится к бесконечности)

Оценка $σ^{2}:$

1. Если выполнены ограничения 1-6), то несмещенная и состоятельная
2. **Общая схема проверки гипотезы о параметре (возможно, здесь вообще не это)**

Пусть имеем линейную регрессию $Y=Xθ+ ε$ и выполнены основные ограничения 1-6). Рассмотрим проверку гипотезы:

$H\_{0}:Aθ=a (2) vs H\_{1}:Aθ \ne a, где A$*-*неслучайная матрица размером $p x \left(m+1\right)$, a - неслучайный вектор размером p. $p\leq m+1, rank\left(A\right)=p$

Описание процедуры проверки:

1. Оцениваем модель без учета ограничений (2) и находим сумму квадратов остатков $ESS\_{UR}$
2. Оцениваем модель с учетом ограничений (2) и находим сумму квадратов остатков $ESS\_{R}$
3. При верной $H\_{0}$ случ.в. $\frac{ESS\_{UR}}{σ^{2}} и\frac{ESS\_{R}-ESS\_{UR}}{σ^{2}}$ независимы и имеют $χ^{2}$-распределение с N-(m+1) и p степенями свободы
4. Нужно учесть степени св при оценке ошибки, так как от этого зависит количество

Случайная величина $F= \frac{(ESS\_{R}-ESS\_{UR})/p}{ESS\_{UR}/(N-(m+1))}$ имеет распределение Снедекора-Фишера с (p, N-(M+1)) степенями свободы.

1. При заданном $α$ ищем по таблицам $F\_{α}>0:P\left(F>F\_{α}\right)=α$
2. Если реально наблюдаемое значение статистики $F>F\_{α}$, то гипотеза H0 отвергается.
3. В противном случае H0 не противоречит экспериментальным данным.
4. **Для чего используется Т-критерий**

Статистика $T\_{k}=\frac{\hat{θ\_{k}}}{S\_{k}}$ (имеет распределение Стьюдента с N-(m+1) степенями свободы при верной H0).

$S\_{k}=S\* \sqrt{b\_{kk}}, b\_{kk}$ *–* элемент матрицы $\left(X^{T}X\right)^{-1}, S^{2}= \hat{σ^{2}}$ (см выше)

Т-критерий используется для проверки значимости влияния отдельного фактора. Этот критерий позволяет проверить значимость только 1 фактора, а не нескольких одновременно, т.к. задача решается, когда и другие факторы вместе влияют на результат.

Может быть ситуация, когда один фактор перекрывает другой или они тесно связаны.

1. **Основное различие Т-критерия и F-критерия в задаче проверки значимости влияния фактора**

В случае простой линейной регрессии критерии эквиваленты. Они различаются только для множественной линейной регрессии.

F – критерий оценивает чистое влияние одного фактора, когда устранено влияние всех остальных.

T – критерий проверяет значимость влияния фактора в присутствии всех остальных.

1. **Адекватность модели. Постановка задачи**

Модель адекватна, если предложенный наборов факторов совместно оказывает значимое влияние на Y.

Формально проверяем: $H\_{0}: θ\_{1}=…= θ\_{m}=0 vs H\_{1}: ∃k θ\_{k}\ne 0$.

Если $H\_{0}$ отвергается, то модель адекватна. В противном случае выбранный набор факторов не оказывает существенного влияния и модель неадекватна.

1. **Коэффициент детерминации и что он измеряет**

Коэффициент детерминации - это число $R^{2}=\frac{RSS}{TSS}$

Если $R^{2}$ близко к 1, то модель хорошая. $R^{2}$ – оценка квадрата множественного коэффициента корреляции (смещенная оценка. Вначале растет, потом убывает).

F-критерий однозначно записывается через $R^{2}:F=\frac{N-(m+1)}{m}\*\frac{R^{2}}{1-R^{2}}$

$TSS= \sum\_{j=1}^{N}\left[Y\_{j}-\overbar{Y}\right]^{2}, RSS= \sum\_{j=1}^{N}\left[\hat{Y\_{j}}-\overbar{Y}\right]^{2}, ESS= \sum\_{j=1}^{N}\left[Y\_{j}-\hat{Y\_{j}}\right]^{2}$

$TSS=ESS+RSS$

TSS – полная сумма квадратов

ESS – сумма квадратов остатков

RSS – уклонение $\overbar{Y}$ за счет влияния факторов, объясненная сумма квадратов (объясняемая с помощью регрессии)

1. **Основная задача в однофакторном дисперсионном анализе**

Пусть имеется следующий набор измерений: $Y\_{kj}= μ\_{k}+ε\_{kj}, k=1, .., m;j=1, .., N$

k – уровень фактора, j – номер измерения

Предполагаем:

1. $μ\_{k}$ – неслучайные веществ числа
2. $ε\_{kj} \~ N\left(0, σ^{2}\right) для любых k,j$
3. $ε\_{kj}$ - независимы

Модель в новом виде: $M=\frac{1}{m}\sum\_{k=1}^{m}μ\_{k}, α\_{k}= μ\_{k}-M=> \sum\_{k}^{}α\_{k}=0=>$

$Y\_{jk}=M+α\_{k}+ε\_{jk}$

*Основная задача:* Есть ли различия в поведении Y на разных уровнях.

Формально проверяем гипотезу $H\_{0}: α\_{1}=…=α\_{m}=0 vs H\_{1}: ∃k: α\_{k}\ne 0$

*Замечание*: есть модификация модели, где $α\_{k}$ (отклонение от среднего уровня) – случайная величина, независимая от ошибок $ε\_{kj}$ и $α\_{k} \~ N\left(0, σ\_{α}^{2}\right)$

В таком случае *основная задача*: вносит ли изучаемый фактор вклад в общую дисперсию модели.

Формально: $H\_{0}: σ\_{α}^{2}=0 vs H\_{1}: σ\_{α}^{2}\ne 0 (σ\_{α}^{2}>0)$

1. **Основная задача в двухфакторном дисперсионном анализе**

Пусть имеется следующая модель измерений: $Y\_{kj}= μ\_{kj}+ε\_{kj}, k=1, .., m;$

$j=1, .., N;n=mN$; $ε\_{kj}$ – случайные ошибки

На исследуемую величину влияет два фактора: k – уровень первого фактора, j – уровень второго фактора. (рассматриваем случай, когда для каждого набора (сочетания) факторов имеется только одно измерение).

Предполагаем:

1. $μ\_{kj}$ – некоторые константы
2. $ε\_{kj} \~ N\left(0, σ^{2}\right) для любых k,j$
3. $ε\_{kj}$ – независимы

Введем следующие обозначения: $μ\_{\*\*}=\frac{1}{n}\sum\_{k,j}^{}μ\_{kj}; μ\_{\*j}=\frac{1}{m}\sum\_{k}^{}μ\_{kj}; μ\_{k\*}=\frac{1}{N}\sum\_{j}^{}μ\_{kj}$

$τ\_{j}= μ\_{\*j}-μ\_{\*\*}$ - эффект столбца; $ρ\_{k}= μ\_{k\*}- μ\_{\*\*}$ - эффект строки

$μ\_{kj}= μ\_{\*\*}+ τ\_{j}+ ρ\_{k}$ – модель без учета взаимодействия факторов

Перепишем модель: $Y\_{kj}=μ\_{\*\*}+ τ\_{j}+ ρ\_{k}+ε\_{kj}$

*Основная задача*: есть ли влияние факторов (есть ли эффект строки или столбца). Проверяем, есть ли разница средних по строкам.

Формально: $H\_{0}: ρ\_{1}=…= ρ\_{m}=0 vs H\_{1}: ∃k\in \left[1,m\right]: ρ\_{k}\ne 0$

1. **Основная задача дискриминантного анализа**

Две основные задачи:

1. *Интерпретация*: можно ли по измеряемым характеристикам различить изучаемые совокупности?
2. *Классификация*: Найти одну или несколько функций от измеряемых характеристик, которые позволят разделить изучаемые группы

*Постановка задачи:*

Двумерный случай для простоты: каждый объект характеризуется парой чисел: $X=(X\_{1}, X\_{2})$. X имеет нормальное двумерное распределение со средними $μ=\left(μ\_{1}, μ\_{2}\right)^{T}$ и матрицей ковариации ∑.

Пусть мы имеем две совокупности, которые различаются средними:

 $μ^{(1)}=\left(μ\_{1}^{(1)}, μ\_{2}^{(1)}\right)^{T}; μ^{(2)}= \left(μ\_{1}^{\left(2\right)}, μ\_{2}^{\left(2\right)}\right)^{T}$, но имеют одну и ту же матрицу ковариаций.

Задача: отнести вновь поступивший объект с хар-ками $X\_{1t}, X\_{2t}$ к одной из совокупностей.

Интуитивно нужно построить линейную функцию и сравнить новую с этой прямой.

Задача легко решаема, когда известны средние и матрица ковариации одна и та же для обеих совокупностей. В реальной задаче средние неизвестны, а матрицы ковариации могут отличаться.

В этом случае первый этап – обучение (2 набора измерений, про каждый из которых известно, из какой совокупности).

Два основных подхода:

1. Метод главных компонент
2. Метод канонических корреляций
3. **Кластерный анализ: постановка задачи**

Из генеральной совокупности выбрано n объектов: $I=(I\_{1}, …, I\_{n})$, у каждого объекта p количественных характеристик $C=\left(C\_{1}, …, C\_{p}\right)^{T}, $измерение i-ой характеристики у j-ого объекта - $X\_{ij}. X\_{j}=\left(x\_{1j}, …, x\_{pj}\right)^{T}$ – измерения всех характеристик объекта j.

$Х=\left(X\_{j}\right)$ – матрица измерений.

*Постановка задачи*: Пусть m < n (если m = n, то каждый объект – кластер). Требуется на основе измерений X разбить множество объектов I на m классов (кластеров) так, чтобы:

1. Каждый объект $I\_{j}$ принадлежал одному и только одному кластеру
2. Объекты внутри одного кластера были бы в некотором смысле сходными
3. Объекты из разных кластеров были бы несходными

Для решения задачи используется некоторая *целевая функция.* Она учитывает число кластеров и качество группировки. Интуитивно ясно, что объекты $I\_{i}, I\_{j}$ нужно объединить в один класс, если расстояние между измерениями $X\_{i}, X\_{j}$ будет достаточно мало, а для точек из разных классов оно будет достаточно большим.

В реальной задаче количество кластеров, как правило, неизвестно и они строятся последовательно.

*Меры сходства:*

1. Евклидово расстояние
2. L1-норма
3. Максимальная норма
4. Расстояние Махаланобиса
5. **Кластерный анализ: последовательное построение факторов**

*Общая схема:*

1. Сначала все объекты рассматриваются как отдельные кластеры
2. Выбирают два порога: 0<r<s
3. Если все кластеры находятся на расстоянии большем, чем s, то все заканчивается
4. Если расстояние между какими-то кластерами меньше s, то находим два наиболее близких и объединяем их, если расстояния внутри нового кластера не более r
5. Пересчитываем новые расстояния между кластерами
6. Процедура продолжается до тех пор, пока расстояния внутри всех кластеров будут не более r, а расстояния между кластерами не более(?) s.

Может привести к построению только одного кластера (т.е. разделить не удалось).