**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 .

Если обнаружено сильное (значимое) влияние факторов на Y, то пытаются найти вид их связи в следующей форме: .

*–* влияние факторов

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* – 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 :

(вообще изначально мы рассматриваем случайные величины такие, у которых )

1. **Что такое наилучшая линейная оценка**

– замкнутое линейное подпространство,

Случайная величина – *наилучшее линейное приближение* случайной величины , если

1. **Лемма о перпендикуляре**

– наилучшее линейное приближение ⬄

* (это на самом деле СЛАУ)

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

*Простой (парный) коэффициент корреляции* невырожденных (не const, иначе ) случайных величин – число .

* измеряет долю изменчивости , которую можно объяснить линейным влиянием
* измеряет ту часть изменчивости , которую не удалось объяснить линейным влиянием и необходимо привлечь другие факторы

*Свойства:*

1. – не коррелированы
2. Если
3. Если

Если

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

Пытаемся объяснить поведение Y с помощью нескольких факторов (совокупное влияние всех факторов вместе)

Пусть –наилучшее линейное приближение Y

*Множественный коэффициент корреляции* Y и набора случайных величин – число

* - показывает, какую долю изменчивости Y можно объяснить линейным влиянием выбранных факторов
* – то, что вызвано неучтенными факторами

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

Изучаем зависимость Y от факторов (чистое влияние одного фактора)

Выберем некоторый фактор

– набор остальных факторов.

– наилучшее линейное приближение Y через C

– наилучшее линейное приближение через C

*Частный коэффициент корреляции* Y и , когда устранено влияние всех остальных факторов, - число

*Свойства:*

– 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, - объясняющие переменные

Представление: .

Необходимо найти функцию наилучшим образом приближающую Y с помощью факторов.

Если расстояние между сл.в. измеряется в среднем квадратическом, то наилучшее приближение задается по правилу: , 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 . При этом предполагается, что

*Ограничения:*

1. Модель линейна по параметрам, то есть
2. Факторы измерены точно, то есть это не случайные величины
3. , то есть нет систематических ошибок
4. – дисперсия одинакова для всех j – условие гомоскедастичности
5. – ошибки некоррелированы
6. – имеет нормальное распределение

имеет многомерное нормальное распределение со средним и матрицей ковариации

1. **Описание МНК для оценки параметров**

*–* параметры модели.

Для оценки параметров модели решаем следующую экстремальную задачу:

Имеем невырожденную ситуацию, если векторы линейно независимы (экв матрица X имеет ранг d+1).

Задача на минимум решается с помощью необходимых условий на экстремум:

После преобразований получаем: – система нормальных уравнений. Если невырождена, то отсюда следуют оценки параметров.

(фактически решается задача о наилучшей линейной оценке Y в линейном пространстве, порожденном случайными векторами )

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

Предсказанные значения:

Остатки: ( наследуют многие свойства )

(средние у = 0)

Если бы были известны, то

Предлагается следующая оценка для :

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

Оценка

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 , то оценка состоятельная – (оценка, сходящаяся по вероятности к своему параметру, количество наблюдений стремится к бесконечности)

Оценка

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 и выполнены основные ограничения 1-6). Рассмотрим проверку гипотезы:

*-*неслучайная матрица размером , a - неслучайный вектор размером p.

Описание процедуры проверки:

1. Оцениваем модель без учета ограничений (2) и находим сумму квадратов остатков
2. Оцениваем модель с учетом ограничений (2) и находим сумму квадратов остатков
3. При верной случ.в. независимы и имеют -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 имеет распределение Снедекора-Фишера с (p, N-(M+1)) степенями свободы.

1. При заданном ищем по таблицам
2. Если реально наблюдаемое значение статистики , то гипотеза H0 отвергается.
3. В противном случае H0 не противоречит экспериментальным данным.
4. **Для чего используется Т-критерий**

Статистика (имеет распределение Стьюдента с N-(m+1) степенями свободы при верной H0).

*–* элемент матрицы (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: .

Если 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

Если близко к 1, то модель хорошая. – оценка квадрата множественного коэффициента корреляции (смещенная оценка. Вначале растет, потом убывает).

F-критерий однозначно записывается через

TSS – полная сумма квадратов

ESS – сумма квадратов остатков

RSS – уклонение за счет влияния факторов, объясненная сумма квадратов (объясняемая с помощью регрессии)

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

Пусть имеется следующий набор измерений:

k – уровень фактора, j – номер измерения

Предполагаем:

1. – неслучайные веществ числа
2. - независимы

Модель в новом виде:

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

Формально проверяем гипотезу

*Замечание*: есть модификация модели, где (отклонение от среднего уровня) – случайная величина, независимая от ошибок и

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

Формально:

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

Пусть имеется следующая модель измерений:

; – 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. – некоторые константы
2. – независимы

Введем следующие обозначения:

- эффект столбца; - эффект строки

– модель без учета взаимодействия факторов

Перепишем модель:

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

Формально:

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 имеет нормальное двумерное распределение со средними и матрицей ковариации ∑.

Пусть мы имеем две совокупности, которые различаются средними:

, но имеют одну и ту же матрицу ковариаций.

Задача: отнести вновь поступивший объект с хар-ками 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: , у каждого объекта p количественных характеристик измерение i-ой характеристики у j-ого объекта - – измерения всех характеристик объекта j.

– 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 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  нужно объединить в один класс, если расстояние между измерениями 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).