Выявление закономерностей во множествах данных на основе моделей дисперсионного анализа

Баклушин Станислав Юрьевич
Крымский федеральный университет имени В.И. Вернадского
Таврическая академия
факультет математики и информатики
кафедра информатики (группа 602-И)
e-mail: Stasbaclushyn@mail.ru

Рассматриваются вопросы применения моделей и методов дисперсионного анализа в задачах изучения и анализа многомерных данных. Рассмотрены основные модели одномерного и многомерного дисперсионного анализа и методы проверки гипотез. Приведены иллюстрирующие примеры и прикладные задачи анализа в области образования.

Одним из мощных и широко распространенных методов многомерного статистического анализа является дисперсионный анализ. Более того, он занимает особое место в статистических исследованиях, поскольку практически все модели анализа данных опираются на изучение вариации признаков и численные оценки этой вариации (дисперсии). Концепция дисперсионного анализа была предложена Р. Фишером в 1920-х годах. Несмотря на глубокую проработанность теоретических обоснований и широкое использование в прикладных задачах, методы дисперсионного анализа остаются актуальными в современных исследованиях, использующих процедуры статистического анализа больших массивов данных.

Сочетание ANOVA (Analysis of Variances) само означает анализ вариаций (дисперсий). Для этой группы методов характерно наличие одной зависимой переменной, на вариацию которой оказывают влияние одна или несколько независимых переменных – факторов. При этом факторы могут иметь неколичественный характер; предполагается возможность их изменения в нескольких уровнях. Зависимая переменная обязана быть количественной и нормально распределенной. В этом заключается важное отличие метода дисперсионного анализа от Т-теста сравнения средних, который не опирается на анализ

причины вариации между сравниваемыми группами и работает только для количественных нормально распределенных данных.

Для многомерного статистического анализа характерно сочетание MANOVA (Multicriterian Analysis of Variances). В этом случае предполагается наличие группы зависимых переменных, варьирующих под влиянием ряда факторов. В этом случае возрастает сложность вычислений, но содержание и основные принципы не меняются: проверяются гипотезы о равенстве средних в подгруппах, соответствующих плану эксперимента.

Модель дисперсионного анализа есть математическое соотношение, представляющее каждую переменную в виде суммы среднего значения и ошибки; среднее значение представляется в виде суммы генерального среднего и «эффекта» от каждого фактора и комбинаций факторов. Возникающие статистические задачи связаны с оценкой этих эффектов и проверкой гипотез о них.

Теоретической основой моделирования является общая линейная модель, выражающая значения зависимой переменной через линейную комбинацию факторов, включенных в анализ, с учетом ошибки. С помощью метода наименьших квадратов определяются оценки параметров модели, а также оценка дисперсии как средний квадрат ошибки, или остаточная сумма квадратов. Получают также разложение общей суммы квадратов на компоненты, связанные с дифференциальными эффектами и эффектами взаимодействий. Для проверки гипотез используется F-критерий Фишера.

В рамках одномерной модели рассматриваются однофакторный ANOVA (One way ANOVA), двухфакторный ANOVA с пересекающимися факторами, а также ANOVA с повторными наблюдениями (Repeated Measures ANOVA). Применение этих моделей для выявления закономерностей данных проиллюстрировано примером анализа сводных данных оценки знаний студентов факультета.

Задачи и методы многомерного дисперсионного анализа (MANOVA) опираются на обобщенную линейную модель, которая образуется из взятых вместе общих линейных моделей для всех зависимых переменных, включенных в анализ. В этом случае рассматриваются остаточные суммы квадратов и остаточные суммы произведений, на основе которых определяются несмещенные оценки дисперсий.

Наиболее важное значение отводится матрице R1 остаточных сумм квадратов и произведений. Водится также матрица (R1 – R0) сумм квадратов и произведений, обусловленных отклонением от гипотезы.

Разложение R1 = R0 – (R1 – R0) является обобщением многомерного дисперсионного анализа. Таким образом, отклонение от гипотезы может быть определено сравнением матриц R0 и (R1 – R0).

Для проверки гипотезы могут быть использованы различные функции, зависящие от корней $\lambda_1,\lambda_2,\ldots,\lambda_p$ характеристического уравнения det(R1 – λR0)=0. В числе критериев наиболее часто используются следующие:

1) критерий наибольшего характеристического корня V=max$\lambda_1,\lambda_2,\ldots,\lambda_p$ предложен С. Н. Роем (Roy S. N, 1965) на основе эвристического метода, поскольку этот корень отражает максимальное отклонение от гипотезы;

2) Λ-критерий Уилкса: ?=$\lambda_1,\lambda_2,\ldots,\lambda_p$=|R0|/|R1| предложен Уилксом (Wilks S. S., 1961). Статистика Λ при некоторых сочетаниях параметров имеет F распределение. В общем случае Λ имеет приближенно распределение хи-квадрат с ps степенями свободы. Лучшая аппроксимация для Λ-критерия предложена Рао (Rao S. R.) и сводится к F-распределению;

3) критерий следа Лоули и Хотеллинга (Lawley D. N., Hotelling H., 1951) T^2=$\Sigma\lambda_1$. Этот критерий считается удобным для практических вычислений, так как не требует нахождения собственных значений. Здесь также можно при определенном сочетании параметров использовать хи-квадрат распределение или F распределение.

Рассматривается ряд примеров дисперсионного анализа для одномерного и многомерного случая, разработанных автором для задач анализа в области образования и основанных на статистических данных успеваемости, результатах опросов. Эти примеры могут быть использованы в качестве сценариев статистического моделирования в лабораторном практикуме по анализу данных.

Так, например, рассмотрены модели дисперсионного анализа, данные для которых взяты из сводной ведомости успеваемости студентов факультета по итогам сессии. Генеральная совокупность (весь контингент бакалавриата) разбита на группы по направлениям подготовки (коды М, ПМ, ПМИ) и курсам (1, 2, 3, 4). Оцениваемые отклики в модели – относительное значение успеваемости (ABS), относительное значение качества (QUA). В качестве возможных факторов, определяющих различие средних в группах рассматриваются направления и курсы.

По отдельности рассмотрены две одномерные модели с зависимыми переменными ABS и QUA. Наблюдения образуют полные факторные планы 3x4. Получены оценки дифференциальных эффектов по итогам одной сессии.

Гипотеза о равенстве средних принимается по значению F-критерия. Интерпретация принятия гипотезы такова: относительные значение успеваемости и качества не показывают статистически значимой зависимости от факторов.

Рассмотрена также модель многомерного дисперсионного анализа на базе полного двухфакторного плана с двумя зависимыми переменными ABS и QUA. Значение Λ-критерия Уилкса также не отклонило гипотезу о равенстве средних, что согласуется с результатами и интерпретацией одномерного анализа.

Список литературы

[1] Аренс, Х. *Многомерный дисперсионный анализ.* – М.: Финансы и статистика, 1985. – 230с.

[2] Афифи, А. *Статистический анализ: Подход с использованием ЭВМ.* – М.: Мир, 1982. – 488с.

[3] Тюрин, Ю.Н. *Анализ данных на компьютере: учеб. пособие для студентов вузов* . – М.: Форум, 2008. – 366с.