Основы системного анализа

Лекция 7 (13-я неделя)

3. Методы анализа экспериментальных данных

3.1. Дисперсионный анализ

3.2. Корреляционный анализ

3.3. Регрессионный анализ

3.3. Регрессионный анализ

Термин «регрессия» был введён Фрэнсисом Гальтоном в конце 19го века. Гальтон обнаружил, что дети родителей с высоким или низким ростом обычно не наследуют выдающийся рост и назвал этот феномен «регрессия к посредственности».

Сначала этот термин использовался исключительно в биологическом смысле.

После работ Карла Пирсона этот термин стали использовать и в статистике.

Регрессионный анализ – статистический метод исследования влияния одной или нескольких независимых переменных $X_1, X_2,.., X_n$ на зависимую переменную Y.

Независимые переменные иначе называют *регрессорами* или *предикторами*, а зависимые переменные – *критериальными*.

В регрессионном анализе, в отличие от корреляционного, только выходные величины Y являются случайными. Входные величины X должны быть неслучайными и некоррелированными между собой.

Аналитические зависимости, полученные по данным эксперимента путем регрессионного анализа называются *эмпирическими* или *аппроксимирующими*. Необходимо иметь в виду, что если теоретические формулы, полученные на основе знания законов процесса, могут быть использованы при произвольных значениях аргументов, то эмпирические являются приближенными и могут применяться лишь в определенных условиях и в ограниченных интервалах аргументов. Один и тот же процесс может быть описан несколькими различными эмпирическими формулами.

Задачи регрессионного анализа:

установления формы зависимости между переменными;

определение функции регрессии;

прогнозирование неизвестных значений зависимой переменной.

Последовательность этапов регрессионного анализа:

- формулировка задачи - на этом этапе формируются предварительные гипотезы о зависимости исследуемых явлений;
- определение зависимых и независимых (объясняющих) переменных;
- сбор статистических данных - данные должны быть собраны для каждой из переменных, включенных в регрессионную модель;

- формулировка гипотезы о форме связи (простая или множественная, линейная или нелинейная);
- определение функции регрессии (заключается в расчете численных значений параметров уравнения регрессии);
- оценка точности регрессионного анализа;
- интерпретация полученных результатов; - полученные результаты регрессионного анализа сравниваются с предварительными гипотезами.
- прогнозирование неизвестных значений зависимой переменной внутри и снаружи интервала исходных данных.

Гипотеза о зависимости исследуемых независимых переменных X_1, X_2,.., X_n на зависимую переменную Y формируется в результате предположения или экспериментальных наблюдений.

Определение зависимых и независимых переменных

В регрессионный анализ не рекомендуется включать факторы, слабо связанные с показателем, но тесно связанные с другими факторами. Не включают в уравнение и факторы, функционально связанные друг с другом (для них коэффициент корреляции равен 1). Включение таких факторов приводит к вырождению системы уравнений для оценок коэффициентов регрессии и к неопределенности решения.

Формулировка гипотезы о форме связи является неформализованной процедурой. Здесь многое зависит от опыта исследователя. Уже отмечалось, что один и тот же процесс может быть описан различными эмпирическими зависимостями. На практике при выборе вида уравнения обычно руководствуются следующим. По данным эксперимента первоначально строят графическую зависимость. Ее сравнивают с различными кривыми, уравнения которых известны, и останавливаются на наиболее вероятной.

Характер и форма зависимости между переменными могут образовывать следующие разновидности регрессии:

- положительная линейная регрессия (выражается в равномерном росте функции);
- положительная равноускоренно возрастающая регрессия;
- положительная равнозамедленно возрастающая регрессия;
- отрицательная линейная регрессия (выражается в равномерном падении функции);
- отрицательная равноускоренно убывающая регрессия;
- отрицательная равнозамедленно убывающая регрессия.

Однако описанные разновидности обычно встречаются не в чистом виде, а в сочетании друг с другом. В таком случае говорят о комбинированных формах регрессии. При выборе формулы нет необходимости ориентироваться на сложные зависимости. Ценность формулы определяется не сложностью, а той погрешностью, которая допускается при ее применении.

Для выбора вида функциональной зависимости можно рекомендовать следующий подход:

- в пространстве параметров графически отображают точки со значениями показателя;

- по расположению точек и на основе анализа сущности взаимосвязи показателя и параметров объекта делают заключение о примерном виде регрессии или её возможных вариантах; после расчета параметров уравнения регрессии оценивают качество аппроксимации, т.е. оценивают степень близости расчетных и фактических значений.

Если расчетные и фактические значения близки во всей области задания, то задачу регрессионного анализа можно считать решенной. В противном случае можно попытаться выбрать другой вид полинома или другую аналитическую функцию, например периодическую.

Определение функции регрессии сводится к выяснению действия на зависимую переменную главных факторов или причин, при неизменных прочих равных условиях, и при условии исключения воздействия на зависимую переменную случайных элементов.

Функция регрессии определяется в виде математического уравнения того или иного типа.

Частным случаем, широко применяемым на практике, является полином первой степени или уравнение линейной регрессии.

Наиболее распространенным методом поиска коэффициентов уравнений регрессии является *метод наименьших квадратов*. Согласно нему наилучшими коэффициентами b_0, b_1, …, b_m в уравнении регрессии считаются те, для которых сумма квадратов разности отклонения численных значений и функциональных данных для конкретной функции минимальна:

$$F(b_0, b_1, \ldots, b_m) = \sum_{i=1}^{n} \left[y_i - f_i\left(x_i, b_0, b_1, \ldots, b_m\right) \right]^2 \to \min .$$

где y_i – численные данные; f_i – значения функции уравнения регрессии.

Условием экстремума функции является равенство нулю частных производных этой функции по варьируемым параметрам, то есть по коэффициентам b_0, b_1, …, b_m уравнения регрессии:

$$\begin{cases} \dfrac{\partial F(b_0, b_1, \ldots, b_m)}{\partial b_0} = 0; \\ \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \\ \dfrac{\partial F(b_0, b_1, \ldots, b_m)}{\partial b_m} = 0. \end{cases}$$

Частные производные функции $F(b_0, b_1, ..., b_m)$ по варьируемым параметрам будут иметь вид

$$\frac{\partial F(b_0, b_1, \ldots, b_m)}{\partial b_i} = -2\sum_{i=1}^{n}\left[y_i - f_i(x_i, b_0, b_1, \ldots, b_m)\right] f'_{b_i}(x_i, b_0, b_1, \ldots, b_m).$$

Решение системы уравнений относительно b_0, b_1, …, b_m дает искомые наилучшие значения параметров уравнения регрессии:

$$\begin{cases} \sum_{i=1}^{n}\left[y_i - f_i(x_i, b_0, b_1, \ldots, b_m)\right] f'_{b_1}(x_i, b_0, b_1, \ldots, b_m) = 0; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum_{i=1}^{n}\left[y_i - f_i(x_i, b_0, b_1, \ldots, b_m)\right] f'_{b_m}(x_i, b_0, b_1, \ldots, b_m) = 0. \end{cases}$$

Для анализа общего качества уравнения линейной регрессии используется обычно коэффициент детерминации R^2, который получается посредством простого возведения в квадрат коэффициента корреляции r. Коэффициент детерминации показывает, в какой мере изменчивость величины Y объясняется поведением величины X.

Например, если коэффициент корреляции совокупных данных равняется 0,8, то коэффициент детерминации R^2 = 0,82 = 0,64 или 64%. Это значение говорит о том, что 64% вариации (изменчивости) величины Y объясняется изменением независимых переменных X_1, X_2,.., X_n. Остальная часть (36%) вариации Y объясняется другими причинами.

Коэффициент детерминации R^2 вычисляется по формуле

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - f_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2},$$

где $\bar{y}$ – среднее значение выборки численных значений функции. При значении $R^2 \geq 0{,}7$ имеется высокая степень связи выявленного уравнения регрессии с найденными экспериментальными данными. Модели с коэффициентом детерминации выше 80 % можно признать достаточно хорошими (коэффициент корреляции превышает 90 %). Значение коэффициента детерминации 1 означает функциональную зависимость между переменными.

Так как в большинстве случаев уравнение регрессии приходится строить на основе выборочных данных, то возникает вопрос об адекватности построения уравнения данным генеральной совокупности. Для этого проводится проверка статистической значимости коэффициента детерминации R^2 на основе F-критерия Фишера:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}$$

где n – число наблюдений, m – число факторов в уравнении регрессии.

Коэффициент детерминации R^2 признается значимым с доверительной вероятностью α, если

$$F > F_a\ (d.\ f_1;\ d.\ f_2\ )\ ,$$

где $F_\alpha(d.f_1;\ d.f_2)$ – квантиль F распределения с $d.f_1 = m$ и $d.f_2 = n - m - 1$ степенями свободы.

Оценка неизвестных значений зависимой переменной сводится к решению задачи одного из типов:

- оценка значений зависимой переменной внутри рассматриваемого интервала исходных данных, т.е. пропущенных значений; при этом решается задача интерполяции;

- оценка будущих значений зависимой переменной, т.е. нахождение значений вне заданного интервала исходных данных; при этом решается задача экстраполяции.

Уравнение регрессии является всего лишь хорошим аналитическим описанием имеющихся данных, а не законом, описывающим взаимосвязи параметров и показателя. Это уравнение применяют для расчета значений показателя в заданном диапазоне изменения параметров. Оно ограниченно пригодно для расчета вне этого диапазона, т.е. его можно применять для решения задач интерполяции и в ограниченной степени для экстраполяции.

Главной причиной неточности прогноза является не столько неопределенность экстраполяции линии регрессии, сколько значительная вариация показателя за счет неучтенных в модели факторов. Ограничением возможности прогнозирования служит условие стабильности неучтенных в модели параметров и характера влияния учтенных факторов модели. Если резко меняется внешняя среда, то составленное уравнение регрессии потеряет свой смысл.

Прогноз, полученный подстановкой в уравнение регрессии ожидаемого значения параметра, является точечным. Вероятность реализации такого прогноза ничтожна мала. Целесообразно определить доверительный интервал прогноза. Для индивидуальных значений показателя интервал должен учитывать ошибки в положении линии регрессии и отклонения индивидуальных значений от этой линии.