ВикипедиЯ

Регрессионный анализ

Материал из Википедии — свободной энциклопедии

Регрессио́нный анализ — набор статистических методов исследования влияния одной или нескольких независимых переменных $X_1, X_2, ..., X_p$ на зависимую переменную Y. Независимые переменные иначе называют регрессорами или предикторами, а зависимые переменные — критериальными. Терминология *зависимых* и *независимых* переменных отражает лишь математическую зависимость переменных (*см. Корреляция*), а не причинно-следственные отношения. Наиболее распространенный вид регрессионного анализа — линейная регрессия, когда находят линейную функцию, которая, согласно определенным математическим критериям, наиболее соответствует данным. Например, в методе наименьших квадратов вычисляется прямая(или гиперплоскость), сумма квадратов между которой и данными минимальна.

Содержание

Цели регрессионного анализа

1. Определение степени детерминированности вариации критериальной (зависимой) переменной предикторами (независимыми переменными)
2. Предсказание значения зависимой переменной с помощью независимой(-ых)
3. Определение вклада отдельных независимых переменных в вариацию зависимой

Математическое определение регрессии

Строго регрессионную зависимость можно определить следующим образом. Пусть $Y, X_1, X_2, \ldots, X_p$ — случайные величины с заданным совместным распределением вероятностей. Если для каждого набора значений $X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p$ определено условное математическое ожидание

$y(x_1, x_2, \dots, x_p) = \mathbb{E}(Y \mid X_1 = x_1, X_2 = x_2, \dots, X_p = x_p)$ (уравнение регрессии в общем виде),

то функция $y(x_1, x_2, \dots, x_p)$ называется **регрессией** величины Y по величинам $X_1, X_2, \dots, X_p$, а её график — **линией регрессии** Y по $X_1, X_2, \dots, X_p$, или **уравнением регрессии**.

Зависимость Y от $X_1, X_2, \dots, X_p$ проявляется в изменении средних значений Y при изменении $X_1, X_2, \dots, X_p$. Хотя при каждом фиксированном наборе значений $X_1 = x_1, X_2 = x_2, \dots, X_p = x_p$ величина Y остаётся случайной величиной с определённым распределением.

Для выяснения вопроса, насколько точно регрессионный анализ оценивает изменение Y при изменении $X_1, X_2, \dots, X_p$, используется средняя величина дисперсии Y при разных наборах значений $X_1, X_2, \dots, X_p$ (фактически речь идет о мере рассеяния зависимой переменной вокруг линии регрессии).

В матричной форме уравнение регрессии (УР) записывается в виде: $Y = BX + U$, где U — матрица ошибок. При обратимой матрице X◤X получается вектор-столбец коэффициентов B с учётом U◤U=min(B). В частном случае для X=(±1) матрица X◤X является рототабельной, и УР может быть использовано при анализе временны́х рядов и обработке технических данных.

Метод наименьших квадратов (расчёт коэффициентов)

На практике линия регрессии чаще всего ищется в виде линейной функции $Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_N X_N$ (линейная регрессия), наилучшим образом приближающей искомую кривую. Делается это с помощью метода наименьших квадратов, когда минимизируется сумма квадратов отклонений реально наблюдаемых Y от их оценок $\hat{Y}$ (имеются в виду оценки с помощью прямой линии, претендующей на то, чтобы представлять искомую регрессионную зависимость):

$$\sum_{k=1}^{M} (Y_k - \hat{Y}_k)^2 \to \min$$

(M — объём выборки). Этот подход основан на том известном факте, что фигурирующая в приведённом выражении сумма принимает минимальное значение именно для того случая, когда $Y = y(x_1, x_2, \dots x_N)$.

Для решения задачи регрессионного анализа методом наименьших квадратов вводится понятие **функции невязки**:

$$\sigma(\bar{b}) = \frac{1}{2} \sum_{k=1}^{M} (Y_k - \hat{Y}_k)^2$$

Условие минимума функции невязки:

$$\begin{cases} \frac{\partial \sigma(\bar{b})}{\partial b_i} = 0 \\ i = 0...N \end{cases} \Leftrightarrow \begin{cases} \sum\limits_{i=1}^{M} y_i = \sum\limits_{i=1}^{M} \sum\limits_{j=1}^{N} b_j x_{i,j} + b_0 M \\ \sum\limits_{i=1}^{M} y_i x_{i,k} = \sum\limits_{i=1}^{M} \sum\limits_{j=1}^{N} b_j x_{i,j} x_{i,k} + b_0 \sum\limits_{i=1}^{M} x_{i,k} \\ k = 1, \ldots, N \end{cases}$$

Полученная система является системой $N+1$ линейных уравнений с $N+1$ неизвестными $b_0, \ldots, b_N$.

Если представить свободные члены левой части уравнений матрицей

$$B = \begin{pmatrix} \sum\limits_{i=1}^{M} y_i \\ \sum\limits_{i=1}^{M} y_i x_{i,1} \\ \vdots \\ \sum\limits_{i=1}^{M} y_i x_{i,N} \end{pmatrix},$$

а коэффициенты при неизвестных в правой части — матрицей

$$A = \begin{pmatrix} M & \sum\limits_{i=1}^{M} x_{i,1} & \sum\limits_{i=1}^{M} x_{i,2} & \ldots & \sum\limits_{i=1}^{M} x_{i,N} \\ \sum\limits_{i=1}^{M} x_{i,1} & \sum\limits_{i=1}^{M} x_{i,1} x_{i,1} & \sum\limits_{i=1}^{M} x_{i,2} x_{i,1} & \ldots & \sum\limits_{i=1}^{M} x_{i,N} x_{i,1} \\ \sum\limits_{i=1}^{M} x_{i,2} & \sum\limits_{i=1}^{M} x_{i,1} x_{i,2} & \sum\limits_{i=1}^{M} x_{i,2} x_{i,2} & \ldots & \sum\limits_{i=1}^{M} x_{i,N} x_{i,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum\limits_{i=1}^{M} x_{i,N} & \sum\limits_{i=1}^{M} x_{i,1} x_{i,N} & \sum\limits_{i=1}^{M} x_{i,2} x_{i,N} & \ldots & \sum\limits_{i=1}^{M} x_{i,N} x_{i,N} \end{pmatrix},$$

то получаем матричное уравнение: $A \times X = B$, которое легко решается методом Гаусса. Полученная матрица будет матрицей, содержащей коэффициенты уравнения линии регрессии:

$$X = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_N \end{pmatrix}$$

Для получения наилучших оценок необходимо выполнение предпосылок МНК (условий Гаусса — Маркова). В англоязычной литературе такие оценки называются *BLUE* (*Best Linear*

Unbiased Estimators — «наилучшие линейные несмещенные оценки»). Большинство исследуемых зависимостей может быть представлено с помощью МНК нелинейными математическими функциями.

Интерпретация параметров регрессии

Параметры b_i являются частными коэффициентами корреляции; $(b_i)^2$ интерпретируется как доля дисперсии Y, объяснённая X_i, при закреплении влияния остальных предикторов, то есть измеряет индивидуальный вклад X_i в объяснение Y. В случае коррелирующих предикторов возникает проблема неопределённости в оценках, которые становятся зависимыми от порядка включения предикторов в модель. В таких случаях необходимо применение методов анализа корреляционного и пошагового регрессионного анализа.

Говоря о нелинейных моделях регрессионного анализа, важно обращать внимание на то, идет ли речь о нелинейности по независимым переменным (с формальной точки зрения легко сводящейся к линейной регрессии), или о нелинейности по оцениваемым параметрам (вызывающей серьёзные вычислительные трудности). При нелинейности первого вида с содержательной точки зрения важно выделять появление в модели членов вида X_1X_2, $X_1X_2X_3$, свидетельствующее о наличии взаимодействий между признаками X_1, X_2 и т. д. (см. Мультиколлинеарность).

См. также

- Корреляция
- Мультиколлинеарность
- Автокорреляция
- Перекрёстная проверка
- Линейная регрессия на корреляции

Литература

- *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. Множественная регрессия = Applied Regression Analysis. — 3-е изд. — М.: «Диалектика», 2007. — С. 912. — ISBN 0-471-17082-8.
- *Фёрстер Э., Рёнц Б.* Методы корреляционного и регрессионного анализа = Methoden der Korrelation - und Regressiolynsanalyse. — М.: Финансы и статистика, 1981. — 302 с.
- *Захаров С. И., Холмская А. Г.* Повышение эффективности обработки сигналов вибрации и шума при испытаниях механизмов // Вестник машиностроения : журнал. — М.: Машиностроение, 2001. — № 10. — С. 31—32. — ISSN 0042-4633 (https://www.worldcat.org/search?fq=x0:jrnl&q=n2:0042-4633).
- *Радченко С. Г.* Устойчивые методы оценивания статистических моделей: Монография. — К.: ПП «Санспарель», 2005. — С. 504. — ISBN 966-96574-0-7, УДК: 519.237.5:515.126.2, ББК 22.172+22.152.
- *Радченко С. Г.* Методология регрессионного анализа: Монография. — К.: «Корнийчук», 2011. — С. 376. — ISBN 978-966-7599-72-0.

Источник — https://ru.wikipedia.org/w/index.php?title=Регрессионный_анализ&oldid=103660903

Эта страница в последний раз была отредактирована 2 декабря 2019 в 16:03.

Текст доступен по лицензии Creative Commons Attribution-ShareAlike; в отдельных случаях могут действовать дополнительные условия.
Wikipedia® — зарегистрированный товарный знак некоммерческой организации Wikimedia Foundation, Inc.