Комментарии к ресурсу от 06.21 раздела «Мат. статистика»
проф. Ивановский Р.И.

Целевой аудиторией данного ресурса служат медработники, гуманитарии и все те, для кого математика не является доминантой профессии. Алгоритмическое содержание задач линейной регрессии приведено в теле ресурса, подробно описано в работе [1] и здесь не повторяется. Любая задача регрессии (ЗР) сопряжена с анализом связей двух или более переменных между собой и определением математических зависимостей между ними. В данном ресурсе рассматривается случай парных связей переменных y и x в рамках двух моделей.

Сопоставляются линейная модель (уравнение прямой линии) вида:

$$y = a + b \cdot x, \tag{1}$$

с простейшей нелинейной моделью второго порядка:

$$y = a + b \cdot x + c \cdot x^2, \tag{2}$$

где a, b, c – коэффициенты (множители при x^0, x^1, x^2).

В общем случае ЗР состоит в определении неизвестных коэффициентов (в нашем случае a, b или a, b, c) модели по данным наблюдений y и x так, чтобы найденное решение наилучшим образом соответствовало бы исходным данным. Это означает, что выражения типа (1) или (2), с подставленными в них значениями найденных коэффициентов, должны максимально (в силу потенциальных возможностей использованной модели) отражать свойства наблюдений y(x).

В ЗР переменные y и x называют откликом и фактором соответственно, а уравнения (1) или (2) – регрессионным соотношением.

Отметим также, что в ЗР важен тип зависимости отклика y от искомых коэффициентов регрессионного соотношения. Так, если эта зависимость – линейная, то ЗР также будет линейной независимо от вида выражения $y(x)$. Отсюда сразу следует, что для моделей (1) и (2) имеем линейные ЗР. Только в случаях, когда зависимость y от искомых коэффициентов – нелинейная, не линейной будет и ЗР.

Разделение ЗР на классы имеет определенный практический смысл, связанный с возможными методами их решения. Так, решение нелинейной ЗР достигается только путем применения алгоритма поиска экстремума функции многих переменных. Решение же любой линейной ЗР можно получить как непосредственным поиском экстремума, так и с помощью выражений, заранее полученных путем теоретического решения экстремальной задачи. Такой метод (алгоритм) и использован в рассматриваемом ресурсе. Он носит название алгоритма метода наименьших квадратов (МНК). Этот алгоритм удобен своей простотой, позволяет решить любую линейную ЗР. Поскольку алгоритм МНК представлен в ресурсе, остановимся здесь лишь на ряде его особенностей.

Пусть результаты n наблюдений отклика y_i и фактора x_i (i = 1, 2, .. n) находятся в ($n \times 1$)-столбцах **Y** и **x** соответственно. Тогда, для каждой i-ой пары имеем:

для модели (1): $y_i = a + b \cdot x_i$; (3)

ля модели (2): $y_i = a + b \cdot x_i + c \cdot (x_i)^2$. (4)

Совокупности n уравнений (3) или (4) образуют системы линейных алгебраических уравнений, в матричной форме имеющих вид $\mathbf{Y} = \mathbf{A} \cdot \beta$, которые и решаются для каждой модели в ЗР с применением МНК. В этих уравнениях:

- **Y** – вектор-столбец значений откликов с элементами y_i;
- **A** – матрица с n строками и m столбцами (m – число коэффициентов модели). Из (3) следует, что для модели (1) матрица **A** имеет два столбца: первый состоит из единиц, элементами второго служат x_i (т. е. второй столбец – это вектор **x**). Для модели (2) к этим двум столбцам добавляется третий, с элементами $(x_i)^2$;
- β — вектор-столбец коэффициентов модели. Так, для (1) $\beta = [a\ b]^T$; для модели (2) $\beta = [a\ b\ c]^T$. Здесь T — символ транспонирования матриц.

Центральной операцией МНК является вычисление вектора $\beta 0$ оптимальных оценок коэффициентов используемой модели. Определение $\beta 0$ позволяет получить множество дополнительных результатов, например, вектор $\mathbf{Y0} = \mathbf{A} \cdot \beta 0$. Этот вектор содержит оптимальные оценки отклика; его i-ый элемент соответствует выражениям (3) или (4), коэффициенты в которых взяты из $\beta 0$. Другим важным результатом ЗР является получение корреляционного отношения R, подробно рассмотренного в [1].

По найденным **Y0** и R можно судить о качестве модели в ЗР.

Элементы **Y0** (сплошные линии на графиках ресурса) должны наилучшим образом (в силу потенциальных возможностей используемой модели) отображать свойства откликов, например, быть близкими к значениям элементов вектора **Y**. Но сближение **Y0** и **Y** будет всегда сопровождаться ростом значений R, т. е. выбор лучшей (из нескольких) модели можно проводить по наибольшему значению R.

Существует несколько формул для R. В данном ресурсе R определяется двумя путями: отношением средних квадратических отклонений (СКО – корень квадратный из дисперсии) векторов **Y0** и **Y**, а также вычислением коэффициента корреляции этих векторов.

Ознакомительный характер ресурса позволил имитировать в нем исходные данные с введением повышающего и понижающего трендов в зависимости $y(x)$. Помеха имитирована датчиком нормально распределенных случайных чисел (встроенная функция rnorm) с нулевым средним и СКО, равным 0.4.

Из результатов решения задач регрессии для моделей (1) и (2) следует:

- Модели (2) имеют очевидное преимущество. Этот вывод не свидетельствует об уникальности именно полиномиальной модели (2). На ее месте могла бы быть модель третьего и более высокого порядка с тем же положительным эффектом. Резкая разница в качестве оценок будет наблюдаться только при переходе от линейной модели (1) к любой нелинейной.
- Линейные модели (1) непригодны при решении большинства прикладных задач. Модель этого вида – одна единственная среди бесконечного множества других – нелинейных моделей, каждая из которых описывает какую-либо кривую. Вероятность встретить реальную связь, удовлетворяющую уравнению прямой линии (1) практически равна нулю. Модель (1) дает оценки связей y и x лишь приблизительно, в общих чертах, поскольку прямая линия не в состоянии охарактеризовать свойства кривой.
- Попытка оценить связь y и x коэффициентом корреляции r автоматически сводит задачу к линейной постановке в рамках модели (1) со всеми негативными

последствиями. Это обстоятельство часто остается незаметным для исследователя, поскольку *r* обычно вычисляют без решения ЗР.

- Неоправданное, механическое использование *r* во всех ситуациях чревато появлением неверных выводов и ошибочных решений и рекомендаций.
- Анализ реальных связей должен предполагать решение ЗР и использование параметров *R*. Это обеспечит объективность оценок при моделях любого типа, уровня и направления связи, а также качества используемой модели в каждом конкретном случае.

1. Ивановский Р.И., Новожилов М.А. Анализ межканальных связей электроэнцефалограмм на основе корреляционных отношений //Математическая биология и биоинформатика. 2016. Т. 11. №. 2. С. 214-224. Статья размещена в разделе «Новости» (http://mas.exponenta.ru/news/), ресурс от 03.2020.