5. РЕГРЕССИОННЫЙ АНАЛИЗ

5.1 Предмет регрессионного анализа. Предикторы и отклик. Математическая теория эксперимента

Регрессионным анализом называют область математической статистики, связанную с выявлением и аналитическим выражением зависимостей между одной несколькими неслучайными величинами x_1, x_2, ..., x_n и доступной для измерения случайной величиной y. Регрессионный анализ является основным средством концентрации, «свертки» эмпирической информации. Такую редукцию информации упрощенно называют «сглаживанием экспериментальных данных». Некорректность этой формулировки обусловлена тем, что задача аппроксимации решается исключительно средствами анализа, без привлечения методов математической статистики. Постановка задач регрессионного анализа обычно предполагает не только нахождение аналитической зависимости, но и последующий анализ результатов, включающий проверку статистических гипотез.

Исследуемая система может пониматься как *черный ящик* (рис. 5.1). Внутреннее содержание системы остается неизвестным для исследователя. Регистрации доступны лишь значения зависимой переменной y, называемой *откликом системы*.

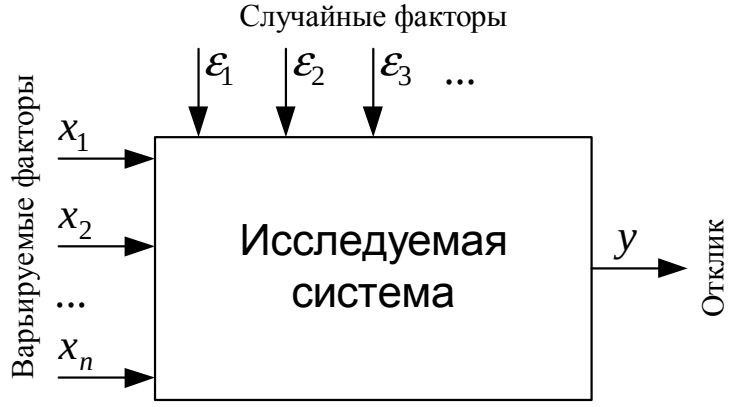

Рис. 5.1. Эксперимент над системой

Пусть производится эксперимент, в ходе которого имеется возможность произвольно выбирать (варьировать) значения n независимых переменных

$$\left(x_1, x_2, ..., x_n\right) = \mathbf{x},$$

называемых *входными переменными*, *варьируемыми факторами* или *предикторами*.

На значение отклика, вместе с объективными закономерностями функционирования системы, оказывают влияние случайные факторы. Последние выражают либо внутренне присущую отклику изменчивость, либо влияние на него обстоятельств, не учтенных в эксперименте (в частности, влияние несовершенства средств измерений). Путь при отсутствии случайных факторов связь меж-

ду откликом и входными переменными дается зависимостью $y = \psi(\mathbf{x})$. Тогда наблюдаемое случайное значение отклика можно представить в виде суммы

$$y = \psi(\mathbf{x}) + \varepsilon,$$

в которой первое слагаемое закономерно зависит от входных переменных **x**, а второе связано с влиянием случайных факторов. Это слагаемое можно считать «ошибкой» эксперимента.

Методы регрессионного анализа на основе экспериментальных данных позволяют получить аналитическую зависимость отклика от варьируемых факторов

$$y = f(x_1, x_2, ..., x_n) = f(\mathbf{x}).$$

Эту зависимость называют *экспериментально-статистической моделью* (ЭС-моделью), *статистической моделью* или *регрессионной моделью*. Корректное применение методов регрессионного анализа предполагает не только получение ЭС-модели, но и решение вопроса об адекватности полученного описания – соответствия его неизвестной истинной зависимости $y = \psi(\mathbf{x})$ отклика от входных переменных.

Регрессионный анализ тесто связан с математической теорией эксперимента – дисциплиной, в рамках которой разрабатываются методы постановки активного эксперимента, позволяющие сделать обоснованный выбор в ситуации альтернативы между недостаточным объемом объективной информации и затратами ресурсов на проведение эксперимента в условиях воздействия случайных факторов. Если цель эксперимента определена, то методы математической теории эксперимента (называемые также методами планирования эксперимента) позволяют ответить на вопросы о минимально необходимом объеме измерений и требуемом характере измерений.

5.2 Принцип максимального правдоподобия и метод наименьших квадратов

Интуитивно понятно, что предсказанное «хорошей» моделью значение $f(\mathbf{x})$ в точке **x** должно быть по возможности близко к наблюдаемому в эксперименте значению отклика. В то же время, формально определить понятие близости функции $f(\mathbf{x})$ к множеству точек $\{(\mathbf{x}, y)\}$ можно по-разному. Например, можно потребовать чтобы в минимум обращалась сумма абсолютных величин $|f(\mathbf{x}) - y|$ отклонений предсказанных и эмпирических значений, или же сумма длин перпендикуляров (отрезков нормалей) к графику, опущенных из точек $\{(\mathbf{x}, y)\}$.

Общий вид модели (класс аналитической зависимости) выбирается до начала аналитических процедур регрессионного анализа. Достаточно распространена практика, когда «выбор» вида регрессионной модели осуществляется только на основании «внешнего вида» данных (численных значений критериев близости регрессионной модели и экспериментальных точек). Такой выбор часто приводит к моделям, не отражающим характерные особенности исследуе-

мых явлений и содержащим «информационный шум» – посторонние члены, не связанные с объективными закономерностями.

Выбор модели должен осуществляться из неформальных соображений, на основе накопленного опыта (отражением которого является интуиция исследователя) и доступной информации об объекте исследования.

Параметрическая регрессионная модель включает L неизвестных параметров $b_1, b_2, \ldots, b_L$:

$$y = f(\mathbf{x}, b_1, b_2, ..., b_L), \tag{5.1}$$

подлежащих определению на основе экспериментальных данных $\{(\mathbf{x}, y)\}$.

Методологической основой построения регрессионных моделей (в частности – методологической основой нахождения параметров) является *принцип максимального правдоподобия*: *наилучшим описанием системы является такое, при котором для модели максимальна вероятность предсказания отклика*.

Пусть для (измеренных в эксперименте) значений отклика выполняются предположения:

1. Ошибки измерений отклика (и измеренные значения) подчинены т.н. *нормальному закону*; математическое ожидание отклика при этом оказывается равным неизвестному истинному значению $\psi(\mathbf{x}_u)$.

2. Измерения независимы и равноточны – стандартные отклонения отклика (и ошибок измерений) во всех опытах равны σ.

Вместе с принципом максимального правдоподобия сделанные предположения составляют основу большинства методов регрессионного анализа. При этих предположениях в каждом u-м из N опытов результат измерения y_u будет случайной величиной Y_u, плотность вероятности которой

$$g_u(y_u) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{[y_u - \psi(\mathbf{x}_u)]^2}{2\sigma^2} \right).$$

Эксперимент есть событие, состоящее в том, что случайные величины $Y_1, Y_2, \ldots, Y_N$ приняли значения $y_1, y_2, \ldots, y_N$. Так как случайные величины Y_u непрерывны, то вероятность события $Y_u = u_u$ равна нулю; можно лишь поставить вопрос о вероятности попадания величины Y_u на малый интервал $[y_u, y_u + dy_u)$. Эта вероятность

$$P(y_u \le Y_u < y_u + dy_u) = g_u(y_u)dy_u = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{[y_u - \psi(\mathbf{x}_u)]^2}{2\sigma^2} \right) dy_u .$$

Так как измерения выполняются независимо, то вероятность P произведения событий $Y_u \in [y_u, y_u + dy_u)$, $u = \overline{1, N}$ оказывается равной произведению вероятностей сомножителей:

$$P = \prod_{u=1}^{N} P(y_u \le Y_u < y_u + dy_u) = \prod_{u=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{[y_u - \psi(\mathbf{x}_u)]^2}{2\sigma^2} \right) dy_u =$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^N \exp\left(-\frac{1}{2\sigma^2} \sum_{u=1}^{N} [y_u - \psi(\mathbf{x}_u)]^2 \right) dy_1 dy_2 ... dy_n ,$$

Обозначим

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N dy_1 dy_2 ... dy_n = K$$

– величина, не зависящая ни от номера u опыта, ни от характера зависимости $\psi(\mathbf{x}_u)$ отклика от входных переменных. Тогда вероятность получить значения, близкие к наблюдаемым на опыте:

$$P = K \exp\left(-\frac{1}{2\sigma^2}\sum_{u=1}^{N}\left[y_u - \psi(\mathbf{x}_u)\right]^2\right). \quad (5.2)$$

Эта вероятность возрастает вместе с увеличением показателя степени и максимальна, если сумма в показателе достигает минимума:

$$\sum_{u=1}^{N}\left[y_u - \psi(\mathbf{x}_u)\right]^2 = \min. \quad (5.3)$$

Но функция $\psi(\mathbf{x})$ неизвестна, а задача как раз и состоит в нахождении модели $f(\mathbf{x})$, которая была бы близка к ней. Поэтому при сделанных предположениях из принципа максимального правдоподобия следует, что *наилучшей моделью будет такая, для которой сумма квадратов отклонений эмпирических значений y_u от значений, предсказанных моделью, обращается в минимум*:

$$S(y_u, f, \mathbf{x}_u) = \sum_{u=1}^{N}\left[y_u - f(\mathbf{x}_u)\right]^2 = \min. \quad (5.4)$$

Методом наименьших квадратов называют метод отыскания параметров модели, который обеспечивает обращение в минимум суммы квадратов отклонений (невязок) наблюдаемых и предсказанных значений.

Метод наименьших квадратов является важнейшим методом обработки эмпирической информации. Он привлекает минимум дополнительных предположений о природе опытных данных и в силу этого имеет высокую общность. Обычно этот метод принято рассматривать в курсе анализа, в разделе теории функций нескольких переменных. При этом метод излагается в терминах анализа применительно к узкому кругу приложений (как правило – применительно к отысканию параметров линейной регрессии). Такое изложение является недостаточным для практики, так как причины выбора критерия близости функции $f(\mathbf{x})$ к множеству экспериментальных точек $\{(\mathbf{x}, y)\}$ остаются нераскрытыми.

5.3 Однофакторная линейная регрессия

Пусть в процессе исследования варьировалась одна независимая переменная x и после проведения N экспериментов получены значения y_u, $u = \overline{1, N}$. Требуется методом наименьших квадратов подобрать параметры линейной экспериментально-статистической модели

$$y = ax + b.$$

Для модели указанного вида сумма квадратов отклонений будет равна: