5.5 Построение и анализ линейной по параметрам модели

Вычислительная процедура построения линейной по параметрам регрессионной модели сводится к использованию соотношения $\mathbf{B} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$. Это соотношение позволяет найти параметры модели, но не решает вопроса соответствия построенной модели и объекта исследования.

Как уже было отмечено, в основе метода наименьших квадратов лежат три предположения:

1. Предположение о нормальном распределении ошибок.
2. Предположение о независимости измерений.
3. Предположение о равной точности измерений.

Если хотя бы одно из этих предположений нарушено, то применение метода наименьших квадратов недопустимо: полученная ЭС-модель не будет наилучшим (исходя из принципа максимального правдоподобия) описанием объекта.

Проверка предположений должна выполняться до построения модели. Проверка становится возможной только тогда, когда в каждом u-м из $u = \overline{1,N}$ экспериментов измерение отклика повторяется $m_u > 1$ раз. Эти m_u измерений, соответствующие одной точке факторного пространства, называют *параллельными* измерениями.

Проверка первых двух предположений требует существенного увеличения числа параллельных измерений, поэтому на практике ограничиваются только проверкой гипотезы о равной точности измерений. С целью снижения возможной взаимной зависимости обычно выполняют рандомизацию измерений (проводят измерения в случайном порядке).

Пусть в u-й точке выполнено m_u параллельных измерений. Оценки средних и дисперсий находятся по известным правилам:

$$\bar{y}_u = \frac{1}{m_u}\sum_{i=1}^{m_u} y_{ui},\ u = \overline{1,N},$$

$$s_u^2 = \frac{1}{m_u - 1}\sum_{i=1}^{m_u}\left(y_{ui} - \bar{y}_u\right)^2 = \frac{1}{f_u}\sum_{i=1}^{m_u}\left(y_{ui} - \bar{y}_u\right)^2,\ u = \overline{1,N},$$

где $f_u = m_u - 1$ – *число степеней свободы* выборочной дисперсии (число параллельных испытаний, уменьшенное на число найденных по выборке оценок: при вычислении выборочной дисперсии уже найдено выборочное среднее). Первый индекс u в обозначении отклика y_{ui} является номером эксперимента, второй индекс i – номером параллельного испытания в этом эксперименте.

В предположении о равной точности измерений найденные оценки дисперсий позволяют вычислить *дисперсию воспроизводимости* (*дисперсию эксперимента*):

$$s_e^2 = \frac{f_1 s_1^2 + f_2 s_2^2 + \ldots + f_N s_N^2}{f_1 + f_2 + \ldots + f_N} = \frac{1}{f_e}\sum_{u=1}^{N} f_u s_u^2 = \frac{1}{f_e}\sum_{u=1}^{N}\sum_{i=1}^{m_u}\left(y_{ui} - \bar{y}_u\right)^2, \tag{5.9}$$

где

$$f_e = \sum_{u=1}^{N} f_u = \sum_{u=1}^{N} (m_u - 1) = \sum_{u=1}^{N} m_u - N$$

– число степеней свободы дисперсии воспроизводимости (полное число измерений, включая параллельные, за вычетом числа экспериментов в различных точках).

Если в каждом эксперименте число параллельных измерений одинаково

$$m_1 = m_2 = \ldots = m_N = M,$$

то соотношение (5.9) упрощается:

$$s_e^2 = \frac{1}{MN - N} \sum_{u=1}^{N} (M-1) s_u^2 = \frac{1}{N} \sum_{u=1}^{N} s_u^2; \qquad (5.10)$$

дисперсия воспроизводимости вычисляется как среднее арифметическое всех выборочных дисперсий.

При $m_u > 3$ проверить гипотезу о равенстве генеральных дисперсий для всех N экспериментов можно по *критерию Бартлета*. Вычисляются величины

$$B = 2{,}303\left(f_e \lg s_e^2 - \sum_{u=1}^{N} (m_u - 1) \lg s_u^2 \right);$$

$$C = 1 + \frac{1}{3(N-1)} \left[\left(\sum_{u=1}^{N} \frac{1}{m_u - 1} \right) - \frac{1}{f_e} \right],$$

после чего вычисляется статистика B/C. Можно приближенно считать, что данная статистика подчинена χ^2-распределению с $N-1$ степенями свободы.

Пусть, например, требуется проверить гипотезу о равной точности измерений в серии из $N = 5$ экспериментов при числе параллельных испытаний $m_1 = m_2 = \ldots = m_5 = M = 4$ (табл. 5.1).

Таблица 5.1

Номер эксперимента	Номер параллельного измерения			
	1	2	3	4
1	0,955452	1,018464	1,011975	0,984515
2	1,969206	2,004065	2,000176	2,005457
3	2,991986	2,973529	3,003962	2,988799
4	4,035914	3,97457	4,0353	3,955538
5	4,953677	5,017916	5,030431	5,043914

Для каждого из пяти экспериментов найдем оценки математического ожидания и дисперсии по формулам

$$\bar{y}_u = \frac{1}{4} \sum_{i=1}^{4} y_{ui}, \; s_u^2 = \frac{1}{3} \sum_{i=1}^{4} (y_{ui} - \bar{y}_u)^2.$$

Найденные оценки сведены в табл. 5.2.

Далее:

$$f_e = N(M-1) = 5 \cdot 3 = 15,$$

$$C = 1 + \frac{1}{3(5-1)}\left(\frac{5}{3} - \frac{1}{15}\right) = \frac{17}{15},$$

$$s_e^2 = \frac{1}{5}\sum_{u=1}^{N} s_u^2 \approx 9{,}2 \cdot 10^{-4},$$

$$B = 2{,}303\left(15 \lg s_e^2 - 3\sum_{u=1}^{N} \lg s_u^2\right) \approx 5{,}50,$$

$$\frac{B}{C} \approx 4{,}85.$$

Таблица 5.2

Номер эксперимента	1	2	3	4	5
Выборочное среднее	0,993	1,995	2,990	4,000	5,011
Выборочная дисперсия	$8{,}30 \cdot 10^{-04}$	$2{,}94 \cdot 10^{-04}$	$1{,}57 \cdot 10^{-04}$	$1{,}72 \cdot 10^{-03}$	$1{,}60 \cdot 10^{-03}$

Вероятность критического события, состоящего в том, что истинное (неизвестное) значение статистики B/C окажется не меньшим наблюдаемого на опыте[1]:

$$P\left(\frac{B}{C} \geq 4{,}85\right) \approx 0{,}3.$$

Поэтому на уровне значимости $\alpha = 0{,}05 < 0{,}3$ нет оснований отвергать гипотезу о равной точности измерений.

При совпадающем числе параллельных испытаний наиболее удобным способом проверки предположения о равной точности измерений оказывается применение *C-критерия* (*критерий Кохрена*). При использовании этого критерия проверка всех $N(N-1)/2$ гипотез о равенстве дисперсий во всех парах серий параллельных испытаний не выполняется. Данный критерий является средством, позволяющим сделать вероятностное суждение о наличии серии измерений, точность в которой существенно ниже средней точности всего эксперимента. О задаче в такой постановке говорят как о задаче *проверки однородности дисперсий*.

Использование *C*-критерия начинается с вычисления статистики

$$C = \max\left\{s_u^2\right\} \Big/ \sum_{u=1}^{N} s_u^2, \tag{5.11}$$

равной отношению максимальной из оценок дисперсий к сумме всех оценок. Эмпирическое значение статистики сравнивается с критическим значением[2]:

$$C_\alpha = \left(1 + \frac{N+1}{F_{\alpha/N}(M-1, (N-1)(M-1))}\right)^{-1}, \tag{5.12}$$

---

[1] Для нахождения вероятности можно использовать: для рабочего листа табличного процессора MS Excel – функцию ХИ2РАСП(4,85;4); для рабочего листа Open/LibreOffice Calc: верное с точки зрения определения выражение: 1-CHISQDIST(4,85;4;1).

[2] http://en.wikipedia.org/wiki/Cochran's_C_test

где $F_\beta(a,b)$ – β-квантиль распределения Фишера для чисел степеней свободы α и β. При выполнении неравенства

$$C < C_\alpha$$

гипотеза о равной точности измерений не отвергается.

После проверки однородности дисперсий параллельных опытов и отыскания параметров регрессионной модели для каждого из найденных параметров необходимо проверить гипотезу о равенстве истинного значения параметра нулю. Если в условиях эксперимента отвергать данную гипотезу нет оснований, то говорят, что параметр *статистически незначим*.

Для проверки значимости параметров β_j, $j = \overline{1, L}$ находят значения статистик

$$t_j = \frac{\beta_j}{\sqrt{s_e^2 c_{jj}}}, \qquad (5.13)$$

где s_e^2 – дисперсия воспроизводимости, c_{jj} – диагональный элемент ковариационной матрицы. Статистика (5.13) подчинена распределению Стьюдента с $f = f_e = N(M-1)$ степенями свободы (в случае неортогональных планов это выполнено лишь приближенно).

Гипотеза о равенстве нулю неизвестного истинного значения *j*-го параметра должна быть отвергнута в пользу двусторонней альтернативы (т.е., параметр должен быть признан *статистически значимым*), если вероятность

$$p_j = \int_{-\infty}^{-|t_j|} f(x)dx + \int_{|t_j|}^{\infty} f(x)dx = 1 - 2\int_0^{|t_j|} f(x)dx \qquad (5.14)$$

критического события, состоящего в том, что при указанной гипотезе будет получено значение β_j, *не меньшее* найденного в эксперименте, оказывается *меньше* заданного уровня значимости α (в соотношении (5.14) подынтегральная функция является плотностью распределения Стьюдента).

Все незначимые параметры модели обнуляются; это соответствует отбрасыванию некоторых слагаемых модели (изменению ее вида). Если ковариационная матрица для исходной модели не являлась диагональной (план был не ортогональным), то параметры модели необходимо пересчитать заново.

Изложенное определяет *итерационный процесс регрессионного анализа*: наличие статистически незначимых оценок параметров модели требует изменения ее вида, повторного отыскания параметров и последующей проверки статистической значимости каждого из них.

Заключительным шагом анализа является проверка *адекватности* полученной ЭС-модели результатам эксперимента. Для ее выполнения вычисляется *остаточная дисперсия*, или *дисперсия адекватности* – величина, пропорциональная сумме квадратов разностей между предсказанными моделью и эмпирическими значениями отклика. Если в каждой точке факторного пространства выполняется M параллельных измерений, то дисперсия адекватности вычисляется по правилу:

$$s_{ad}^2 = \frac{M}{N-L}\sum_{u=1}^{N}\left(y_u - f(\mathbf{x}_u)\right)^2 , \tag{5.15}$$

где N – число экспериментов (число различных точек плана эксперимента), L – число искомых параметров модели (число слагаемых).

Затем вычисляется значение статистики

$$F = \frac{s_{ad}^2}{s_e^2} , \tag{5.16}$$

где s_e^2 – дисперсия воспроизводимости. Статистика (5.16) подчинена распределению Фишера с $f_{ad} = N - L$ и $f_e = N(M-1)$ степенями свободы. Гипотеза адекватности модели эксперименту отвергается, если вероятность

$$p = \int_F^{\infty} f(x)dx = 1 - \int_0^F f(x)dx$$

критического события (состоящего в том, что при адекватной модели значение F будет не меньшим, чем реально обнаруженное) окажется меньше выбранного уровня значимости (в последнем соотношении подынтегральная функция $f(x)$ является плотностью распределения Фишера).