ОСНОВЫ РЕГРЕССИОННОГО АНАЛИЗА. ПОЛНЫЙ ФАКТОРНЫЙ ЭКСПЕРИМЕНТ

Процесс определения явного вида уравнения регрессии получил название *регрессионного анализа*. Для различных математических планов эксперимента уравнение регрессии содержит различные составляющие:

а) для планов первого порядка уравнение регрессии включает линейные эффекты и парные взаимодействия:

$$y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + b_{12}X_1X_2 + \ldots + b_{n-1,n}X_{n-1}X_n \quad (1)$$

б) для планов второго порядка уравнение регрессии включает линейные эффекты, парные взаимодействия и квадратичные эффекты:

$$y = b_0 + b_1X_1 + \ldots + b_nX_n + b_{12}X_1X_2 + \ldots + b_{n-1,n}X_{n-1}X_n + b_{11}X_1^2 + \ldots b_{nn}X_n^2, \quad (2)$$

где b_0 – свободный член уравнения регрессии; $b_n, b_{12} \ldots b_{n-1,n}, b_{11} \ldots b_{nn}$ – коэффициенты регрессии; X_n – условное значение фактора x_n.

Предположим, что изучается влияние ряда факторов $z_i (i = 1, \ldots, k)$ на некоторую величину y. Для этого проводятся эксперименты по определенному плану, который позволяет реализовать все возможные комбинации факторов. Причем каждый фактор рассматривается лишь на двух фиксированных уровнях (верхнем и нижнем). Число всех экспериментов (опытов) в этом случае будет равно $n = 2^k$, где k – количество изучаемых факторов. Постановка опытов по такому плану называется полным факторным экспериментом типа 2^k (ПФЭ 2^k). План проведения экспериментов записывается в виде матрицы планирования, в которой в определенном порядке перечисляются различные комбинации факторов на двух уровнях. Например, в табл. 1 приведена матрица планирования ПФЭ 2^3 для трех факторов: z_1, z_2, z_3. Знак «+» говорит о том, что во время опыта значение фактора устанавливают на верхнем уровне, а знак «-» показывает, что значение фактора устанавливают на нижнем уровне.

Таблица 1 – Матрица планирования ПФЭ 2^3

Значение фактора	z_1	z_2	z_3
1	+	+	+
2	-	+	+

Окончание таблицы 2

Значение фактора	z_1	z_2	z_3
3	+	-	+
4	-	-	+
5	+	+	-
6	-	+	-
7	+	-	-
8	-	-	-

При проведении экспериментов получают значения исследуемой величины y для каждого опыта (или серии опытов). Затем переходят к построению математической модели. Под моделью понимается вид функции $y = f\left(z_1, z_2, \ldots, z_k\right)$, которая связывает изучаемый параметр со значениями факторов, лежащих в интервале между верхним и нижним уровнями. Эту функцию называют *уравнением регрессии*. По накопленному разными исследователями опыту работы с различными моделями можно считать, что самыми простыми моделями являются алгебраические полиномы. Для обработки результатов проведенных экспериментов и дальнейшего определения коэффициентов уравнения регрессии факторы приводят к одному масштабу. Это достигается путем кодирования переменных. Обозначим нижний уровень фактора z_i через z_i^-, а верхний уровень через z_i^+. Тогда новые кодированные переменные x_i будут определяться через z_i по формуле

$$X_i = \frac{z_i - z_i^0}{\varepsilon_i} \tag{3}$$

где z_i – натуральное значение i-го фактора; z_i^0 – натуральное значение i-го фактора на основном уровне; ε_i – интервал варьирования i-го фактора. При таком кодировании все новые переменные будут принимать значения от –1 до +1. Линейное уравнение регрессии относительно новых переменных имеет вид:

$$y = b_0 + b_1 X_1 + b_2 X_2 + \ \ldots \ + b_k X_k \tag{4}$$

Если требуется изучить влияние парных взаимодействий различных факторов на исследуемый параметр, то уравнение регрессии записывают в виде

(1). Прежде чем определять коэффициенты выбранной модели, матрицу планирования записывают относительно новых переменных. Далее матрицу дополняют (если это требует вид выбранного уравнения регрессии) столбцами знаков «+» и «–», соответствующих уровням, на которых будут находиться взаимодействия факторов. Знаки этих столбцов получают с помощью исходной матрицы планирования (табл. 2).

Таблица 2 – Матрица планирования для обработки результатов ПФЭ

Значение фактора	Факторы			Взаимодействия			Результаты опытов			Среднее результатов
	x_1	x_2	x_3	x_{12}	x_{13}	x_{23}	y_1	y_2	y_3	$y_{среднее}$
1	+	+	+	+	+	+				
2	–	+	+	–	–	+				
3	+	–	+	–	+	–				
4	–	–	+	+	–	–				
5	+	+	–	+	–	–				
6	–	+	–	–	+	–				
7	+	–	–	–	–	+				
8	–	–	–	+	+	+				

Обычно проводят несколько серий опытов для каждого эксперимента. Это необходимо для проверки уравнения на адекватность.

Адекватность – это способность модели предсказывать результаты эксперимента в некоторой области с требуемой точностью. Результаты опытов в каждом *j*-м эксперименте ($j = 1,\ldots,n$) записываются в правые столбцы матрицы планирования. В последнем столбце записываются средние выборочные значения полученных результатов для каждой серии опытов. Если каждый эксперимент повторяли *m* раз, то в матрице будет записано m столбцов $y_1, y_2, \ldots, y_m$.

Например, в табл. 2 видно, что каждый эксперимент повторялся три раза, т. е. $m = 3$. Коэффициенты уравнения регрессии находятся с помощью метода наименьших квадратов. Так как матрица планирования ПФЭ 2^k должна удовлетворять определенным требованиям, то формулы, определяющие коэффициенты уравнения регрессии, достаточно просты:

$$b_0 = \frac{1}{N}\sum_{j=1}^{N}\overline{y}_j \tag{5}$$

$$b_i = \frac{1}{N}\sum_{j=1}^{N}x_{ij}\overline{y}_j \tag{6}$$

$$b_{im} = \frac{1}{N}\sum_{j=1}^{N}x_{ij}x_{jm}\overline{y}_j \tag{7}$$

Некоторые из коэффициентов регрессии могут оказаться пренебрежимо малыми – незначимыми. Чтобы установить, значим коэффициент или нет, необходимо прежде всего вычислить оценку дисперсии, с которой он находится:

$$S^2_{\{y\}} = \frac{1}{N}\sum_{j=1}^{N}S_j^2 \tag{8}$$

Для проверки воспроизводимости опытов находится отношение наибольшей из оценок дисперсий к сумме всех оценок дисперсий (расчетное значение критерия Кохрена):

$$G_p = \frac{\max S_j^2}{\sum_{j=1}^{N}S_j^2} \tag{9}$$

Табулированные значения критерия Кохрена $G_т$ приведены в приложении. Для нахождения $G_т$ необходимо знать уровень значимости p, общее количество оценок дисперсий N и число степеней свободы f, связанных с каждой из них, причем $f = k - 1$. При выполнении условия $G_p \le G_т$ опыты считаются воспроизводимыми, а оценки дисперсий – однородными. Если опыты невоспроизводимы, то можно попытаться достигнуть воспроизводимости выявлением и устранением источников нестабильности эксперимента, а также использованием более точных методов и средств измерений. Наконец, если никакими способами невозможно достигнуть воспроизводимости, то математические методы планирования к такому эксперименту применять нельзя.

Следует отметить, что с помощью ПФЭ все коэффициенты определяются с одинаковой погрешностью. Значимость каждого коэффициента уравнения

регрессии устанавливается с помощью критерия Стьюдента (прил. 5), вычисляя его расчетное значение:

$$t_p = \frac{|b|}{\sqrt{S^2_{\{y\}}}} \tag{10}$$

где b – коэффициент уравнения регрессии, для которого устанавливается значимость. Каждое рассчитанное значение t_p сравнивают с табличным значением критерия Стьюдента $t_т$, которое выбирают для заданного уровня значимости p при числе степеней свободы $f = N(k-1)$.

Если выполняется условие $t_p \geq t_т$, то коэффициент считается значимым. В противном случае коэффициент регрессии незначим, и соответствующий член можно исключить из уравнения регрессии. Получив уравнение регрессии, следует проверить его адекватность с помощью критерия Фишера (прил. 1–3), который представляет собой отношение:

$$F_p = \frac{\max(S^2_{ад}; S^2_y)}{\min(S^2_{ад}; S^2_y)} \tag{11}$$

где – $S^2_{ад}$ оценка дисперсии адекватности, которая вычисляется как

$$S^2_{ад} = \frac{1}{N-B}\sum_{j=1}^{N}(y^э_j - y^p_j)^2 \tag{12}$$

где $y^э$, $y^р$ – экспериментальное и расчетное значения функции отклика, полученные в j-м опыте; B – количество коэффициентов в уравнении регрессии. При вычислении расчетного значения критерия Фишера по формуле (18) в числителе указывается большая, а в знаменателе – меньшая из оценок дисперсий. Уравнение регрессии адекватно описывает результаты эксперимента, если выполняется условие $F_р < F_т$, где $F_т$ – табличное значение критерия Фишера для принятого уровня значимости p и числа степеней свободы f_1 числителя и f_2 знаменателя. Если гипотеза об адекватности отвергается, необходимо перейти к более сложной форме или провести эксперимент с меньшим интервалом варьирования факторов.

Анализ результатов предполагает интерпретацию полученной модели. Интерпретацию модели можно производить только тогда, когда она записана в кодированных переменных. Только в этом случае на коэффициенты не влияет масштаб факторов, и мы можем по величине коэффициентов судить о степени влияния того или иного фактора. Чем больше абсолютная величина коэффициента, тем больше фактор влияет на отклик (изучаемый параметр). Следовательно, можно расположить факторы по величине их влияния. Знак «+» у коэффициента свидетельствует о том, что с увеличением значения фактора растет величина отклика, а при знаке «–» – убывает.

Для получения математической модели в натуральных переменных z_i в уравнение регрессии вместо x_i необходимо подставить их выражения. При переходе к натуральным переменным коэффициенты уравнения изменяются, и в этом случае пропадает возможность интерпретации влияния факторов по величинам и знакам коэффициентов. Однако если уравнение адекватно, то с его помощью можно определять значения исследуемой величины, не проводя эксперимента и придавая факторам значения, которые должны лежать между нижним и верхним уровнем.

Пример задания

Для исследования влияния некоторых технологических факторов на прочность приклеивания низа обуви полиуретановым клеем были поставлены эксперименты по плану ПФЭ 2^3, причем каждый эксперимент повторялся по три раза (табл. 3). В качестве факторов, влияющих на прочность y (кг/см^2), были выбраны следующие: z_1 – количество наносимого клея (г/см^2); z_2 – время активации клеевой пленки (с); z_3 – давление прессования при склеивании (кгс/см^2).

Требуется построить уравнение регрессии, учитывая все взаимодействия факторов, проверить полученную модель на адекватность и произвести ее интерпретацию.

Таблица 3 – Исходная матрица планирования ПФЭ 2^3

Значение фактора	Факторы			Результаты		
	z_1	z_2	z_3	y_1	y_2	y_3
1	+	+	+	7,4	8,4	6,4
2	–	+	+	8,6	7,0	7,8
3	+	–	+	12,3	9,0	9,3
4	–	–	+	5,8	5,8	5,7
5	+	+	–	18,8	17,0	15,2
6	–	+	–	8,4	8,4	6
7	+	–	–	11,8	7,0	9,4
8	–	–	–	10,5	7,8	8,1

Работу следует выполнять в следующем порядке:

1) кодируются переменные;

2) достраиваются матрицы планирования в кодированных переменных с учетом парных взаимодействий и дополняются столбцом средних значений отклика;

3) вычисляются коэффициенты уравнения регрессии;

4) проверяются вычисленные коэффициенты на значимость, предварительно определяя дисперсию воспроизводимости, и получается уравнение регрессии в кодированных переменных;

5) проверяется полученное уравнение на адекватность;

6) проводится интерпретация полученной модели;

7) выписывается уравнение регрессии в натуральных переменных.

Решение

1. Для каждого фактора принимаем основной уровень, интервал варьирования и зависимость кодированной переменной x_i от натуральной z_i. Оформляем результаты в таблицу (табл. 4).

Таблица 4 – Кодирование факторов

Факторы	Нижний уровень	Верхний уровень	Основной уровень	Интервал варьирования
z_1	0,06	0,02	0,04	0,02
z_2	300	60	180	120
z_3	8	2	5	3
Зависимость кодированной величины от натуральной $x_1 = \frac{z_1 - 0,04}{0,02} = 50z_1 - 2$ $x_2 = \frac{z_2 - 180}{120}$ $x_3 = \frac{z_3 - 5}{3}$				

2. Считаем средние выборочные результатов для каждого эксперимента:

$$\overline{y}_1 = \frac{1}{3}(7,4 + 8,4 + 6,4) = 7,4$$

$$\overline{y}_2 = \frac{1}{3}(8,6 + 7,0 + 7,8) = 7,8$$

$$\overline{y}_3 = \frac{1}{3}(12,3 + 9,0 + 9,3) = 10,2$$

$$\overline{y}_4 = \frac{1}{3}(5,8 + 5,8 + 5,7) = 5,77$$

$$\overline{y}_5 = \frac{1}{3}(18,8 + 17,0 + 15,2) = 17$$

$$\overline{y}_6 = \frac{1}{3}(8,4 + 8,4 + 6,0) = 7,6$$

$$\overline{y}_7 = \frac{1}{3}(11,8 + 7,0 + 6,0) = 9,4$$

$$\overline{y}_7 = \frac{1}{3}(10,5 + 7,8 + 8,1) = 8,8$$

Строим матрицу планирования с учетом всех взаимодействий и средних значений отклика (табл. 5).

Таблица 5 – Матрица планирования для обработки результатов ПФЭ

№ п/п	Факторы			Взаимодействия				Результаты опытов			Среднее результатов
	x_1	x_2	x_3	x_{12}	x_{13}	x_{23}	x_{123}	y_1	y_2	y_3	$y_{среднее}$
1	+	+	+	+	+	+	+	7,4	8,4	6,4	7,4
2	–	+	+	–	–	+	–	8,6	7,0	7,8	7,8
3	+	–	+	–	+	–	–	12,3	9,0	9,3	10,2
4	–	–	+	+	–	–	+	5,8	5,8	5,7	5,77
5	+	+	–	+	–	–	–	18,8	17,0	15,2	17,0
6	–	+	–	–	+	–	+	8,4	8,4	6,0	7,6
7	+	–	–	–	–	+	+	11,8	7,0	9,4	9,4
8	–	–	–	+	+	+	–	10,5	7,8	8,1	8,8

3. Вычисляем коэффициенты уравнения регрессии:

$$b_0 = \frac{1}{8}\sum_{j=1}^{8} \overline{y}_j = \frac{1}{8}(7,4+7,8+10,2+5,77+17+7,6+9,4+8,8) = 9,25$$

$$b_1 = \frac{1}{8}\sum_{j=1}^{8} x_{i1}\overline{y}_j = \frac{1}{8}(7,4-7,8+10,2-5,77+17-7,6+9,4-8,8) = 1,75$$

$$b_2 = \frac{1}{8}\sum_{j=1}^{8} x_{i2}\overline{y}_j = \frac{1}{8}(7,4+7,8-10,2-5,77+17+7,6-9,4-8,8) = 0,7$$

$$b_3 = \frac{1}{8}\sum_{j=1}^{8} x_{i3}\overline{y}_j = \frac{1}{8}(7,4+7,8+10,2+5,77-17-7,6-9,4-8,8) = -1,45$$

$$b_{12} = \frac{1}{8}\sum_{j=1}^{8} x_{i1}x_{j2}\overline{y}_j = \frac{1}{8}(7,4-7,8-10,2+5,77+17-7,6-9,4+8,8) = 0,5$$

$$b_{13} = \frac{1}{8}\sum_{j=1}^{8} x_{i1}x_{j3}\overline{y}_j = \frac{1}{8}(7,4-7,8+10,2-5,77-17+7,6-9,4+8,8) = -0,75$$

$$b_{23} = \frac{1}{8}\sum_{j=1}^{8} x_{i2}x_{j3}\overline{y}_j = \frac{1}{8}(7,4+7,8-10,2-5,77-17-7,6+9,4+8,8) = -0,9$$

$$b_{123} = \frac{1}{8}\sum_{j=1}^{8} x_{i1}x_{j2}x_{j3}\overline{y}_j = \frac{1}{8}(7,4-7,8-10,2+5,77-17+7,6+9,4-8,8) = -1,7$$

4. Находим дисперсию воспроизводимости. Столбец 6 вычисляем по формуле:

$$S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(y_{j1} - \overline{y_j}\right)^2\right]$$

Данные заносим в таблицу 6.

Таблица 6 – Расчет дисперсий

j	1			2	3	4	5	6
	y_1	y_2	y_3	$\bar{y}_j$	$(y_{j1} - \bar{y}_j)^2$	$(y_{j2} - \bar{y}_j)^2$	$(y_{j3} - \bar{y}_j)^2$	S^2_j
1	7,4	8,4	6,4	7,4	0	1	1	1
2	8,6	7,0	7,8	7,8	0,64	0,64	0	0,64
3	12,3	9,0	9,3	10,2	4,41	1,44	0,81	3,33
4	5,8	5,8	5,7	5,77	0,0009	0,0009	0,0049	0,0034
5	18,8	17,0	15,2	17,0	3,24	0	3,24	3,24
6	8,4	8,4	6,0	7,6	0,64	0,64	2,56	1,92
7	11,8	7,0	9,4	9,4	5,76	5,76	0	5,76
8	10,5	7,8	8,1	8,8	2,89	1	0,49	2,19

Суммируя элементы столбца 6 таблицы 6, получаем:

$$\sum_{j=1}^{8} S_j^2 = 18,0834$$

Отсюда получаем дисперсию воспроизводимости:

$$S_{\{y\}}^2 = \frac{1}{8}\sum_{j=1}^{8} S_j^2 = \frac{18,0834}{8} = 2,26$$

5. Определяем среднее квадратическое отклонение коэффициентов:

$$S_{\{y\}} = \sqrt{\frac{S_{\{y\}}^2}{n \cdot m}} = \sqrt{\frac{2,26}{8 \cdot 3}} = 0,293$$

Из таблиц распределения Стьюдента (прил. 5) по числу степеней свободы $n(m-1) = 8 \cdot 2 = 16$ при уровне значимости $\alpha = 0,05$ находим $t_{кр} = 2,12$. Следовательно, $t_{кр} \cdot S_{\{y\}} = 2,12 \cdot 0,293 = 0,52$. Сравнивая полученное значение с коэффициентами уравнения регрессии, видим, что все коэффициенты кроме $b_{1,2}$ больше по абсолютной величине 0,52. Следовательно, все коэффициенты кроме

$b_{1,2}$ значимы. Полагая $b_{1,2} = 0$, получаем уравнение регрессии в кодированных переменных:

$$y = 9,25 + 1,75x_1 + 0,7x_2 - 1,45x_3 - 0,75x_1x_3 - 0,9x_2x_3 - 1,7x_1x_2x_3$$

6. Проверим полученное уравнение на адекватность по критерию Фишера. Так как дисперсия воспроизводимости найдена в предыдущем пункте, то для определения расчетного значения критерия $F_{\text{расч}}$ необходимо вычислить остаточную дисперсию $S^2_{\text{ост}}$. Для этого найдем значения изучаемого параметра по полученному уравнению регрессии $\tilde{y}_j (j = 1,\ldots,8)$, подставляя +1 или –1 вместо x_i в соответствии с номером j эксперимента из табл. 5:

$$\tilde{y}_1 = 9,25 + 1,75 + 0,7 - 1,45 - 0,75 - 0,9 - 1,7 = 6,9$$
$$\tilde{y}_2 = 9,25 + 1,75(-1) + 0,7 - 1,45 - 0,75(-1) - 0,9 - 1,7(-1) = 8,3$$
$$\tilde{y}_3 = 9,25 + 1,75 + 0,7(-1) - 1,45 - 0,75 - 0,9(-1) - 1,7(-1) = 10,7$$
$$\tilde{y}_4 = 9,25 + 1,75(-1) + 0,7(-1) - 1,45 - 0,75(-1) - 0,9(-1) - 1,7 = 5,3$$
$$\tilde{y}_5 = 9,25 + 1,75 + 0,7 - 1,45(-1) - 0,75(-1) - 0,9(-1) - 1,7(-1) = 16,5$$
$$\tilde{y}_6 = 9,25 + 1,75(-1) + 0,7 - 1,45(-1) - 0,75 - 0,9(-1) - 1,7 = 8,1$$
$$\tilde{y}_7 = 9,25 + 1,75 + 0,7(-1) - 1,45(-1) - 0,75(-1) - 0,9 - 1,7 = 9,85$$
$$\tilde{y}_8 = 9,25 + 1,75(-1) + 0,7(-1) - 1,45(-1) - 0,75 - 0,9 - 1,7(-1) = 8,3$$

Находим остаточную дисперсию:

$$S^2_{ост} = \frac{3}{8-7}\sum_{j=1}^{8}(\tilde{y}_j - \overline{y}_j)^2 = 3[(6,9-7,4)^2 + (8,3-7,8)^2 + (10,7-10,2)^2 +$$
$$+(5,3-5,77)^2 + (16,5-17)^2 + (8,1-7,6)^2 + (9,85-9,4)^2 + (8,3-8,8)^2] =$$
$$= 3 \cdot 1,9234 = 5,77$$

Расчетное значение критерия Фишера $F_{\text{расч}}$:

$$F_{расч} = \frac{S^2_{ост}}{S^2_{\{y\}}} = \frac{5,77}{2,26} = 2,55$$

Табличное значение критерия $F_{\text{табл}}$ находим из таблицы критических точек распределения Фишера (прил. 1) при уровне значимости α = 0,05 по соответствующим степеням свободы $k_1 = n - r = 8 - 7 = 1$ и

$k_2 = n(m-1) = 8 \cdot 2 = 16$. $F_{табл} = 4,49$. Так как $F_{расч} = 2,8 < F_{табл} = 4,49$, то уравнение регрессии адекватно.

7. Проведем интерпретацию полученной модели:

$$y = 9,25 + 1,75x_1 + 0,7x_2 - 1,45x_3 - 0,75x_1x_3 - 0,9x_2x_3 - 1,7x_1x_2x_3$$

По уравнению видно, что наиболее сильное влияние оказывает фактор x_1 – количество наносимого клея, так как он имеет наибольший по абсолютной величине коэффициент. После него по силе влияния на отклик (прочность приклеивания низа обуви) идут: тройное взаимодействие всех факторов $x_1x_2x_3$; фактор x_3 – давление пресса при склеивании; парное взаимодействие x_2x_3 – сочетание времени активации клеевой пленки и уровня давления при склеивании; парное взаимодействие x_1x_3 – сочетание количества наносимого клея и уровня давления при склеивании; фактор x_2 – время активации клеевой пленки. Так как коэффициенты при x_1 и x_2 положительные, то с увеличением этих факторов увеличивается отклик, т.е. увеличивается прочность. Коэффициенты при x_3, x_1x_3, x_2x_3, $x_1x_2x_3$ отрицательные, это означает, что с уменьшением фактора x_3 и перечисленных взаимодействий значение отклика будет возрастать, а с увеличением – убывать.

8. Выписываем уравнение регрессии в натуральных переменных, подставляя вместо x_i их выражения через z_i, которые берем из табл. 4:

$$y = 9,25 + 17,5(50z_1 - 2) + 0,7\frac{z_2 - 180}{120} - 1,45\frac{z_3 - 5}{3} - 0,75(50z_1 - 2)\frac{z_3 - 5}{3} -$$
$$-0,9\frac{z_2 - 180}{120} \cdot \frac{z_3 - 5}{3} - 1,75(50z_1 - 2)\frac{z_2 - 180}{120} \cdot \frac{z_3 - 5}{3}$$

Преобразовав это уравнение, окончательно получаем его вид в натуральных переменных:

$$y = 10,87 - 62,5z_1 - 0,029z_2 - 1,23z_3 + 1,18z_1z_2 + 30z_1z_3 + 0,007z_2z_3 -$$
$$-0,236z_1z_2z_3$$