Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования
«Московский государственный технический университет им. Н. Э. Баумана»

Факультет «Машиностроительные технологии»

Кафедра «Литейные технологии»

Отчет по лабораторной работе № 2

по курсу «Основы научных исследований»

Планирование полного двухфакторного эксперимента. Регрессионный анализ

Вариант 34

Выполнил: студент гр. МТ5–17
Карпенко Д. Н.

Проверил: асс. каф. МТ5
Карпенко Д. Н.

Москва – 2016

Оглавление

Введение

Цель работы – ознакомление с методами математического планирования эксперимента и овладение методом регрессионного анализа экспериментальных данных.

Содержание работы:

1. Ознакомиться с теоретической частью работы по методическому указанию.
2. На основании экспериментальных данных провести проверку опытов на воспроизводимость.
3. Методом наименьших квадратов определить параметры линейной математической модели.
4. Провести проверку значимости полученных коэффициентов регрессии.
5. Проверить адекватность математической модели экспериментальным данным и сделать вывод о возможности ее применения.
6. Оформить отчет.

1 Теоретическая часть

1.1 Постановка задачи

Цель большинства экспериментальных исследований – изучение влияния ряда параметров на ход процесса или на характеристики изделия (материала) с тем, чтобы выбрать параметры, при которых изделие (материал) обладает требуемыми свойствами. В обоих случаях задача сводится к отысканию функциональной зависимости выходного параметра y от параметров, определяющих ход и результаты процесса, то есть:

$$y = f\left(x_1, x_2, ..., x_n\right), \qquad (1)$$

где y – выходной параметр, называемый также функцией отклика;

$x_1, \ldots, x_n$ – факторы эксперимента;

n – число факторов.

В технологических исследованиях обычно принимают, что функция отклика – случайная величина, имеющая нормальный закон распределения. На этом допущении основывается использование аппарата регрессионного анализа при проверке статистических гипотез о функции отклика. При этом каждой точке факторного пространства соответствуют μ и σ – математическое ожидание и дисперсия функции отклика. Из-за ограниченности выборки математические ожидания функции отклика оценивают как выборочные средние:

$$\mu_i = \frac{1}{m}\sum_{j=1}^{m} y_{ij}, \qquad (2)$$

где m – число параллельных измерений.

Геометрическое место точек математических ожиданий μ функции отклика $y = f(x_1, x_2, \ldots, x_n)$ в факторном пространстве называется регрессией y по x ($x = (x_1, x_2, \ldots, x_n)$), а уравнение, описывающее это геометрическое место точек в том же факторном пространстве, называется уравнением регрессии. Задача экспериментатора – найти такую функцию (уравнение), которая достаточно хорошо приближалась бы к уравнению регрессии, т.е. аппроксимировала его.

В качестве аппроксимирующих функций при планировании экспериментов чаще всего выбирают полиномы первого порядка, то есть линейные функции вида:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_n \cdot x_n \qquad (3)$$

и главная задача состоит в нахождении значений коэффициентов $b_1, \ldots, b_n$.

1.2 Метод наименьших квадратов

Пусть имеется N точек, в которых проведены эксперименты с целью построения уравнения регрессии (3). Чтобы снизить влияние случайных факторов, нарушающих воспроизводимость экспериментов, в каждой точке, эксперименты дублировали m_k раз – ставились параллельные опыты. Число параллельных опытов от точки к точке в общем случае различно, но всегда больше одного.

Аппроксимирующая кривая должна как можно ближе проходить к исходным (экспериментальным) точкам $\{\bar{y}_k\}$. Чаще всего при обработке эмпирической информации, особенно если она отягчена случайными ошибками, применяется метод наименьших квадратов (МНК), в котором в качестве критерия близости двух функций используется сумма квадратов разностей между экспериментально наблюдаемыми и модельными значениями $\{\hat{y}_k\}$, которая должна быть минимальной. Другими словами, идея метода наименьших квадратов состоит в том, чтобы на основании экспериментальный данных построить уравнение регрессии (3) так, чтобы функция

$$S(b_0, b_1, ..., b_n) = \sum_{k=1}^{N} (\bar{y}_k - \hat{y}_k) \cdot m_k \qquad (4)$$

имела минимум.

Подставив (3) в (4), получим:

$$S\left(b_0,b_1,...,b_n\right)=\sum_{k=1}^{N}\left(\bar{y}_k-\left(b_0+b_1\cdot x_{1k}+b_2\cdot x_{2k}+...b_n\cdot x_{nk}\right)\right)\cdot m_k\,. \qquad (5)$$

Для того чтобы найти значения коэффициентов $b_1,\ldots,b_n$, при которых функция (5) минимальна, надо найти первые производные и приравнять их нулю:

$$\begin{cases}\dfrac{\partial S\left(b_0,b_1,...,b_n\right)}{\partial b_0}=0;\\ \dfrac{\partial S\left(b_0,b_1,...,b_n\right)}{\partial b_1}=0;\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ \dfrac{\partial S\left(b_0,b_1,...,b_n\right)}{\partial b_n}=0.\end{cases} \qquad (6)$$

После подстановки (5) в (6) и ряда преобразований получим следующую систему уравнений:

$$\begin{cases}b_0\cdot\sum\limits_{k=1}^{N}m_k+b_1\cdot\sum\limits_{k=1}^{N}x_{1k}\cdot m_k+...+b_n\cdot\sum\limits_{k=1}^{N}x_{nk}\cdot m_k=\sum\limits_{i=1}^{N}\bar{y}_k\cdot m_k;\\ b_0\cdot\sum\limits_{k=1}^{N}x_{1k}\cdot m_k+b_1\cdot\sum\limits_{k=1}^{N}x_{1k}^2\cdot m_k+...+b_n\cdot\sum\limits_{k=1}^{N}x_{nk}\cdot x_{1k}\cdot m_k=\sum\limits_{k=1}^{N}\bar{y}_k\cdot m_k\cdot x_{1k};\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ b_0\cdot\sum\limits_{k=1}^{N}x_{nk}\cdot m_k+b_1\cdot\sum\limits_{k=1}^{N}x_{1k}\cdot x_{nk}\cdot m_k+...+b_n\cdot\sum\limits_{k=1}^{N}x_{nk}^2\cdot m_k=\sum\limits_{k=1}^{N}\bar{y}_k\cdot m_k\cdot x_{nk}.\end{cases} \qquad (7)$$

Линейная система уравнений (7) определена и может быть решена обычными методами, так что формально задача нахождения коэффициентов регрессии решена. Однако при произвольном выборе точек, в которых проводятся эксперименты, могут возникнуть трудности с вычислениями. Малые ошибки, например, при округлении (как вручную, так и с помощью ЭВМ), приводят к недопустимо большим ошибкам. Обойти эти проблемы можно только специальным выбором точек, в которых проводятся эксперименты.

1.3 Полный факторный эксперимент

Вычисления можно существенно упростить, если для системы (7) выполняются следующие четыре условия:

1. Число параллельных измерений m_k в каждой точке одинаковое: $m_k = m$.
2. Условие симметричности:

$$\sum_{k=1}^{N} x_{ik} = 0; \quad i = 1,...,n. \tag{8}$$

3. Условие ортогональности:

$$\sum_{k=1}^{N} x_{ik} \cdot x_{jk} = 0; \quad \forall i \neq j. \tag{9}$$

4. Условие нормировки:

$$\sum_{k=1}^{N} x_{ik}^2 = N; \quad i = 1,...,n. \tag{10}$$

С учетом этих условий после ряда преобразований система (7) сводится к простому виду:

$$\begin{cases} b_0 = \dfrac{1}{N} \cdot \sum\limits_{k=1}^{N} \overline{y}_k; \\ b_1 = \dfrac{1}{N} \cdot \sum\limits_{k=1}^{N} \overline{y}_k \cdot x_{1k}; \\ \dots\dots\dots\dots\dots\dots \\ b_n = \dfrac{1}{N} \cdot \sum\limits_{k=1}^{N} \overline{y}_k \cdot x_{nk}. \end{cases} \tag{11}$$

Для придания полной симметрии уравнениям (14) вводят фиктивную переменную x_0 такую, что $x_{0k} = +1$ для всех $k = 1,\dots, N$. Тогда:

$$b_i = \frac{1}{N} \cdot \sum_{k=1}^{N} x_{ik} \cdot \overline{y}_k; \quad i = 0,...,n. \tag{12}$$

Для построения плана, удовлетворяющего условиям (8)–(10) обычно используют особое преобразование переменных. Обычно планируют менять значения каждого фактора x_i в некоторых пределах (от $x_{i\,\min}$ до $x_{i\,\max}$).

Середину этого интервала $(x_{i\,\min} + x_{i\,\max})/2 = x_i^0$ назовем основным уровнем, а половину длины интервала $(x_{i\,\min} - x_{imax})/2 = \Delta x_i$ – интервалом варьирования. Преобразуем факторы из натурального масштаба в безразмерный по следующей формуле:

$$X_i = \frac{x_i - x_i^0}{\Delta x_i}. \qquad (13)$$

Если в экспериментах факторы будут принимать значения только на нижнем и верхнем уровне ($x_i^- = x_i^0 - \Delta x_i$ и $x_i^+ = x_i^0 + \Delta x_i$), то кодированные переменные X_i будут принимать значения +1 и –1. Эксперимент, проводимый по такому плану, называется полнофакторным экспериментом (ПФЭ) – см. табл. 1.

Таблица 1 – План ПФЭ для двух факторов

Номер опыта	Фактор	
	X_1	X_2
1	+1	+1
2	–1	+1
3	+1	–1
4	–1	–1

В методе ПФЭ опыты проводятся во всех вершинах гиперкуба (квадрата для случая двух факторов), при этом число опытов $N = 2^n$.

1.4 Регрессионный анализ

Регрессионный анализ применим при следующих условиях:

1. Функция отклика y – случайная величина с нормальным законом распределения.
2. Дисперсия y не зависит от абсолютной величины y.
3. Значения факторов – суть неслучайные величины.

Если число параллельных опытов велико, то постулат о нормальном законе распределения легко можно проверить с помощью стандартных статистических тестов (например, χ^2-критерия). Чаще, однако, этот постулат принимается на веру. Нарушение третьего условия обычно легко обнаруживается при реализации матрицы планирования. Практически третье условие означает, что ошибка при установлении и поддержании каждого фактора на заданном уровне существенно меньше, чем ошибка воспроизводимости опытов.

Проверка воспроизводимости

Второе условие проверяется с помощью критерия Кохрена, расчетное значение которого определяется по формуле:

$$G_P = \frac{S_{\max}^2}{\sum_{k=1}^{N} S_k^2}, \tag{14}$$

где S_k^2 – оценка дисперсии k-ого опыта:

$$S_k^2 = \frac{1}{m-1} \cdot \sum_{j=1}^{m} \left(y_{kj} - \bar{y}_k \right)^2; \tag{15}$$

$S_{\max}^2$ – максимальная из рассчитанных дисперсий.

Расчетное значение критерия Кохрена сравнивается с табличным значением G_T, которое находят по доверительной вероятности P, числу степеней свободы числителя $f_1 = m - 1$ и знаменателя $f_2 = N$). Если $G_P \le G_T$, то гипотеза об однородности дисперсий (второй постулат) подтверждается. В противном случае необходимо выяснить причины нестабильности экспериментов, ликвидировать их и повторить все сначала.

Проверка значимости коэффициентов

При одинаковом числе параллельных опытов (m) во всех точках плана матрицы дисперсию среднего значения функции отклика $S_{\bar{y}}^2$ и среднеквадратическое отклонение $S_{\bar{y}}$ вычисляют следующим образом:

$$S_{\bar{y}}^2 = \frac{S_y^2}{m}; \tag{16}$$

$$S_{\bar{y}} = \sqrt{\frac{S_y^2}{m}}\,, \tag{17}$$

где S_y^2 – дисперсия единичного измерения выходного параметра:

$$S_y^2 = \frac{1}{N}\sum_{k=1}^{N} S_k^2\,. \tag{18}$$

Значимость коэффициентов регрессии определяют по t-критерию Стьюдента. Для каждого коэффициента определяют расчетное значение критерия Стьюдента:

$$t_{iP} = \frac{|b_i| \cdot \sqrt{N}}{S_{\bar{y}}}\,. \tag{19}$$

Расчетное значение критерия Стьюдента сравнивают с табличным t_T, которое определяют по доверительной вероятности P и числу степеней свободы $f = N \cdot (m - 1)$. Если $t_{iP} > t_T$, то i-й коэффициент уравнения регрессии значим (гипотеза о том, что его математическое ожидание равно нулю, отвергается). В противном случае коэффициент незначим, и данный фактор надо исключить из уравнения регрессии как не влияющий на изучаемый процесс.

Проверка на адекватность

Проверка пригодности модели (уравнения регрессии) для использования, то есть проверка на адекватность, является самой главной. Для ее проведения в первую очередь нужно рассчитать дисперсию адекватности, характеризующую разброс значений функции отклика относительно функции регрессии:

$$S_{ад}^2 = \frac{1}{f_{ад}} \cdot \left(\sum_{k=1}^{N} \bar{y}_k^2 - N \cdot \sum_{i=0}^{n} b_i^2 \right), \tag{20}$$

где $f_{ад}$ – число степеней свободы дисперсии адекватности: $f_{ад} = N - (n + 1)$.

Дисперсия адекватности сравнивается с дисперсией среднего значения функции отклика, рассчитанной по формуле (17). Отношением этих дисперсий является расчетное значение критерия Фишера:

$$F_P = \frac{S_{ад}^2}{S_{\bar{y}}^2}. \tag{21}$$

Расчетное значение критерия Фишера сравнивается с табличным, которое определяется на основании доверительной вероятности P и числа степеней свободы числителя $f_{ад}$ и знаменателя f. Полученная модель будет считаться адекватной при выполнении условия $F_P \le F_T$. В противном случае уравнение регрессии неадекватно и пользоваться им нельзя.

2 Практическая часть

2.1 Задание

На основании результатов ПФЭ, приведенных в табл. 2, найти линейную зависимость мощности смесителя N от массы смеси $m_{см}$ и прочности смеси $\sigma_{см}$:

$$N = f\left(m_{см}, \sigma_{см}\right). \quad (22)$$

Таблица 2 – Результаты эксперимента

Номер опыта	Исходные значения		Кодовые значения		y: N, Вт			$\bar{N}$, Вт
	$m_{см}$, кг	$\sigma_{см}$, кПа	X_1	X_2				
1	3	60	–	–	205	210	215	210,000
2	5	100	+	+	375	385	390	383,333
3	5	60	+	–	230	235	245	236,667
4	3	100	–	+	310	315	325	316,667

Факторы в эксперименте принимали следующие значения:

- x_1: $m_{см}$ = (4 ± 1) кг;
- x_2: $\sigma_{см}$ = (80 ± 20) кПа.

2.2 Определение эмпирической зависимости

Эмпирическую зависимость будем искать в следующем виде:

$$y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2. \quad (23)$$

Проверка воспроизводимости опытов

В соответствии с планом обработки результатов измерения, изложенном в подразд. 1.4, в первую очередь следует проверить гипотезу об однородности дисперсий. Для этого вычислим дисперсии каждого из опытов по формуле (15):

$$S_1^2 = \frac{1}{3-1} \cdot \left((205-210)^2 + (210-210)^2 + (215-210)^2 \right) = 25;$$

$$S_2^2 = \frac{1}{2} \cdot \left(\left(375 - 383,333\right)^2 + \left(385 - 383,333\right)^2 + \left(390 - 383,333\right)^2 \right) =$$
$$= 58,3333;$$

$$S_3^2 = \frac{1}{2} \cdot \left(\left(230 - 236,667\right)^2 + \left(235 - 236,667\right)^2 + \left(245 - 236,667\right)^2 \right) =$$
$$= 58,3333;$$

$$S_4^2 = \frac{1}{2} \cdot \left(\left(310 - 316,667\right)^2 + \left(315 - 316,667\right)^2 + \left(325 - 316,667\right)^2 \right) =$$
$$= 58,3333.$$

Видно, что максимальная дисперсия характерна сразу для трех опытов, поэтому можно принять $S_{\max}^2 = S_2^2 = 58,3333$.

На основании этого вычислим по формуле (14) значение критерия Кохрена:

$$G_P = \frac{58,3333}{25 + 58,3333 + 58,3333 + 58,3333} = 0,291667.$$

Найдем табличное значение критерия Кохрена. Для этого зададимся уровнем доверительной вероятности P_G = 0,95. С учетом этого, а также числа степеней свободы числителя f_1 = 2 и знаменателя f_2 = 4 по таблице критических значений критерия Кохрена определяем: G_T = 0,77.

Видно, что расчетное значение меньше табличного, следовательно, дисперсии опытов однородны, а сами опыты воспроизводимы.

Расчет коэффициентов модели

Коэффициенты модели (23) рассчитаем с использованием данных табл. 2 по формуле (12):

$$b_0 = \frac{1}{4} \cdot \left(210 + 383,333 + 236,667 + 316,667\right) = 286,667;$$

$$b_1 = \frac{1}{4} \cdot \left(-210 + 383,333 + 236,667 - 316,667\right) = 23,3333;$$

$$b_2 = \frac{1}{4} \cdot \left(-210 + 383,333 - 236,667 + 316,667\right) = 63,3333.$$

Проверка значимости коэффициентов модели

Сначала по формулам (18), (16) и (17) вычислим дисперсию единичного измерения, дисперсию среднего значения функции отклика и соответствующее ей среднеквадратическое отклонение:

$$S_y^2 = \frac{25 + 58,3333 + 58,3333 + 58,3333}{4} = 50;$$

$$S_{\bar{y}}^2 = \frac{50}{3} = 16,6667;$$

$$S_{\bar{y}} = \sqrt{16,6667} = 4,08248.$$

Для проверки на значимость зададимся доверительной вероятностью P_t = 0,975 и рассчитаем число степеней свободы f = 4·(3 – 1) = 8. На основании этого по таблицам определим критическое значение критерия Стьюдента: t_T = 2,31.

Рассчитаем теперь значение критерия Стьюдента для каждого из факторов по формуле (19) и сравним с табличным значением:

$$t_{p0} = \frac{286,667 \cdot \sqrt{4}}{4,08248} = 140,437 > 2,31;$$

$$t_{p1} = \frac{23,3333 \cdot 2}{4,08248} = 11,431 > 2,31;$$

$$t_{p2} = \frac{63,3333 \cdot 2}{4,08248} = 31,0269 > 2,31.$$

Видно, что все коэффициенты модели значимы, поэтому окончательно уравнение регрессии (23) в кодовый переменных приобретает следующий вид:

$$y = 286,667 + 23,3333 \cdot X_1 + 63,3333 \cdot X_2. \qquad (24)$$

Применив формулу (13) к уравнению (24) можно получить уравнение регрессии в исходных переменных:

$$N = 286,667 + 23,3333 \cdot \frac{m_{см} - 4}{1} + 63,3333 \cdot \frac{\sigma_{см} - 80}{20},$$

что после несложных арифметических преобразований даст окончательное эмпирическое уравнение:

$$N = 23,3333 \cdot m_{см} + 3,16667 \cdot \sigma_{см} - 60\,. \qquad (25)$$

Проверка модели на адекватность

По формуле (20) вычислим дисперсию адекватности с учетом того, что $f_{ад} = 4 - (2 + 1) = 1$:

$$S^2_{ад} = \left(210^2 + 383,333^2 + 236,667^2 + 316,667^2\right) - \\ - 4 \cdot \left(286,667^2 + 23,3333^2 + 63,3333^2\right) = 400.$$

Вычислим значение критерия Фишера по формуле (21):

$$F_P = \frac{400}{16,6667} = 24\,.$$

Для определения табличного значения критерия Фишера зададимся допустимой вероятностью: $P_F = 0,95$. С учетом этого, а также числа степеней свободы числителя $f_{ад} = 1$ и знаменателя $f = 8$ по таблице определяем критическое значение критерия Фишера: $F_T = 5,32$.

Видно, что расчетное значение значительно больше табличного. Это говорит о том, что модель (24) и, следовательно, (25) не адекватно описывает данные эксперимента.

2.3 Результаты расчетов

Результаты расчетов представлены в табл. 3, сформированной учебной программой для лабораторной работы № 2 автоматически.

Таблица 5 – Результаты расчетов

Параметры	Результаты Программы	Результаты Пользователя
1. Расчетное значение критерия Кохрена	0.2917	0.2917
2. Табличное значение критерия Кохрена	0.77	0.77
Опыты воспроизводимы		
3. Коэффициент уравнения регрессии b_0	286.6667	286.667

Коэффициент уравнения регрессии b_1	23.3333	23.3333
Коэффициент уравнения регрессии b_2	63.3333	63.3333
4. Расчетное значение критерия Стьюдента t_{p0}	140.4374	140.437
Расчетное значение критерия Стьюдента t_{p1}	11.431	11.431
Расчетное значение критерия Стьюдента t_{p2}	31.02687	31.0269
5. Табличное значение критерия Стьюдента $t_т$	2.31	2.31
Все коэффициенты значимы		
6. Расчетное значение критерия Фишера F_p	24	24
7. Табличное значение критерия Фишера $F_т$	5.32	5.32
Уравнение неадекватно		

3 Выводы

1. В результате обработки экспериментальных данных была получена следующая эмпирическая модель:

$$N = 23,3333 \cdot m_{см} + 3,16667 \cdot \sigma_{см} - 60,$$

где N – мощность смесителя, Вт;

$m_{см}$ – масса смеси, кг;

$\sigma_{см}$ – прочность смеси, кПа.

2. Статистическая проверка показала, что полученная модель не адекватна экспериментальным данным, следовательно, ее нельзя использовать для описания исследуемого процесса. Учитывая, что расчетное значение критерия Фишера значительно больше табличного, функция отклика, вероятно, является существенно нелинейной. Поэтому для получения адекватной модели процесса необходимо использовать один из центральных композиционных планов второго порядка и провести дополнительные опыты в соответствующих точках. При этом будет получена модель второго порядка от тех же переменных, которая с высокой вероятностью окажется адекватной.