Национальный Исследовательский Томский Политехнический Университет

Институт природных ресурсов

Кафедра химической технологии топлива и химической кибернетики

СТАТИСТИЧЕСКИЕ МОДЕЛИ ОБЪЕКТОВ НА ОСНОВЕ ПАССИВНОГО ЭКСПЕРИМЕНТА.

Активный эксперимент

- ***Активный*** эксперимент ставится по заранее составленному плану и обрабатывается по некоторому оптимальному алгоритму с целью составления математической модели или нелинейного полинома.

Пассивный эксперимент

- исследователь собирает некоторый объем экспериментальной информации, т.е. значений факторов x_i и выходного параметра y_i. Причем происходит это в режиме нормальной эксплуатации объекта. Данные (выборка) берутся из каких–либо журналов (например, оператора установки, регламента).
- Для получения статистических моделей в виде полиномов на основе данных используют методы корреляционного и регрессионного анализа.

МЕТОДЫ КОРРЕЛЯЦИОННОГО И РЕГРЕССИОННОГО АНАЛИЗА

.

Корреляционный анализ основывается на предпосылке о том, что переменные величины y (выходной параметр) и x_i (факторы) являются случайными величинами и между ними может существовать так называемая корреляционная связь, при которой с изменением одной величины изменяется распределение другой. Для колличественной оценки тесноты связи служит выборочный ***коэффициент корреляции.***

ВИДЫ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ:

- ***Простой*** коэффициент корреляции или коэффициент парной корреляции определяет величину (тесноту) зависимости между двумя переменными x или y.
- Коэффициент ***частной*** ***к***орреляции измеряет линейную зависимость между двумя переменными после устранения части зависимости, обусловленный зависимостью этих переменных с другими переменными.
- **Множественный** коэффициент корреляции определяет величину зависимости одной переменной от нескольких.

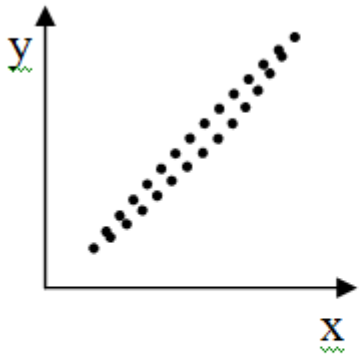
Коэффициент парной корреляции:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1) \cdot S_x \cdot S_y}$$

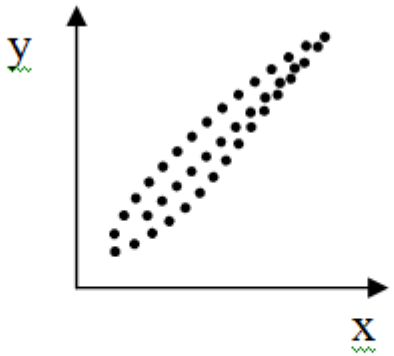
$$S_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}};$$

$$S_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}};$$

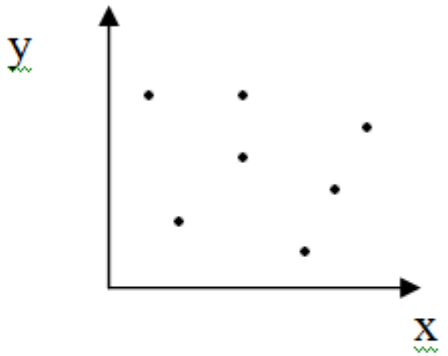
$0<r_{xy}<1$



Сильная корреляция
$r_{xy}\approx 0.97$



Слабая корреляция
$r_{xy}\approx 0.6\text{-}0.7$



Нет корреляции
$r_{xy}\approx 0.1\text{-}0.2$

- Если $r_{xy}=0$, то корреляции нет.

Коэффициент частной корреляции - оценивает степень влияния фактора x_1 на y при условии, что влияние x_2 на y исключено.

При исследовании зависимости **y** от $\mathbf{x_1}$ и $\mathbf{x_2}$ наличие корреляции между $\mathbf{x_1}$ и $\mathbf{x_2}$ и между **y** и $\mathbf{x_2}$ будет влиять на корреляцию между **y** и $\mathbf{x_1}$. Для того чтобы устранить влияние x_2 необходимо измерить корреляцию между y и x_1, при x_2=const.

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_2x_1}}{\left(1 - r^2_{yx_2}\right)^{\frac{1}{2}} \left(1 - r^2_{x_1x_2}\right)^{\frac{1}{2}}} \qquad r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_2x_1}}{\left(1 - r^2_{yx_1}\right)^{\frac{1}{2}} \left(1 - r^2_{x_1x_2}\right)^{\frac{1}{2}}}$$

Частный коэффициент оценивает степень влияния фактора x_1 на y при условии, что влияние x_2 на y исключено.

РЕГРЕССИОННЫЙ АНАЛИЗ

Постановка задачи:

По данной выборке объема ***n*** *найти уравнение приближенной регрессии и* ***оценить допускаемую при этом ошибку****. Эта задача решается методами корреляционного и регрессионного анализа.*

Т.е. нужно найти $\hat{y} = f(x)$

По сгущениям точек можно найти определенную зависимость, т.е. получить вид уравнения регрессии. При значительном разбросе точек регрессии не будет

Вид уравнения регрессии зависит от выбираемого метода приближения.

Обычно используется **метод наименьших квадратов**.

$$F = \sum_{i-1}^{n} \left(y_i - f(x_i)\right)^2 = \min \quad \text{или} \quad F = \sum_{i-1}^{n} \left(y_i - \hat{y}_i\right)^2 = \min$$

ЛИНЕЙНАЯ РЕГРЕССИОННАЯ МОДЕЛЬ

При моделировании ХТП во многих случаях связь между *X* и *Y* можно описать линейной зависимостью $\hat{y} = b_0 + b_1 x;$

Связь между входными (*x*) и выходными (*y*) параметрами:

Для нахождения коэффициентов уравнения регрессии b_0 и b_1 применим метод наименьших квадратов

$$F = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i\right)^2 = \mathbf{min}$$

Необходимым условием min функции является равенство нулю частных производных функции по искомым величинам (коэффициентам).

$$\begin{cases} \dfrac{\partial F}{\partial b_0} = -2 \sum\limits_{i=1}^{n} \left(y_i - b_0 - b_1 x_i\right) \cdot 1 = 0; \\ \dfrac{\partial F}{\partial b_1} = -2 \sum\limits_{i=1}^{n} \left(y_i - b_0 - b_1 x_i\right) \cdot x_i = 0 \end{cases}$$

$$\begin{cases} \sum \left(y_i - b_0 - b_1 x_i\right) = 0; \\ \sum \left(y_i - b_0 - b_1 x_i\right) \cdot x_i = 0; \end{cases} \qquad (2)$$

$$\begin{cases} n b_0 + b_1 \sum x_i = \sum y_i \\ b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i y_i \end{cases}$$

формулы для вычисления коэффициентов b_0 и b_1

$$b_0 = \frac{\begin{vmatrix} \sum y_i & \sum x_i \\ \sum x_i y_i & \sum x_i^2 \end{vmatrix}}{\begin{vmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix}} \qquad b_1 = \frac{\begin{vmatrix} n & \sum y_i \\ \sum x_i & \sum x_i y_i \end{vmatrix}}{\begin{vmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix}}$$

$$b_0 = \frac{\sum y_i \sum x_i^2 - \sum x_i y_i \sum x_i}{N\sum x_i^2 - \left(\sum x_i\right)^2}; \qquad b_1 = \frac{N\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2};$$

СТАТИСТИЧЕСКИЙ АНАЛИЗ РЕЗУЛЬТАТОВ

1. Для ***оценки тесноты*** линейной зависимости между факторами рассчитывают коэффициенты парной корреляции r по формуле:

$$r_{yx} = \frac{\sum_{i=1}^{N} \left(x_i - \bar{x}_1\right)\left(y_i - \bar{y}\right)}{(N-1)\cdot S_x \cdot S_y};$$

$$-1 \le r \le 1;$$

2. **Проверка однородности дисперсий.**

1) Определяется среднее по результатам параллельных опытов (если есть параллельные опыты):

$$\bar{y}_i = \frac{\sum_{u=1}^{m} y_{iu}}{m}; \quad i = 1,...N$$

m – число параллельных опытов

N – количество опытов в выборке

2) Определяются выборочные дисперсии:

$$S_i^2 = \frac{\sum_{u=1}^{m} \left(y_{iu} - \bar{y}_i\right)^2}{m-1}; i = \overline{1,N}$$

3) Суммируются дисперсии $\sum_{i=1}^{N} S_i^2;$

4) Выбирается максимальная дисперсия, составляется отношение:

$$G = \frac{S^2_{\max}}{\sum_{i=1}^{N} S_i^2};$$

$S^2_{\max}$ - максимальное значение выборочной дисперсии.

Проверяется однородность дисперсий по критерию Кохрена (при одинаковом количестве параллельных опытов).

Если $G < G_{табл.}(q, f_1, f_2)$, то дисперсии однородны.

число степеней свободы $f_1=m-1$; $f_2=N$;

5) Определяется дисперсия воспроизводимости $S^2_{воспр.} = \frac{\sum_{i=1}^{N} S_i^2}{N(m-1)}$

-для одинакового числа опытов: f= (N(m-1)).

3. **Оценивается значимость коэффициентов полинома по критерию Стьюдента** (предпосылка – отсутствие корреляции между факторами)

$$t_{b_i} = \frac{|b_i|}{S_{b_i}},$$

где b_i – i-ый коэффициент уравнения регрессии;

S_{b_i} - среднеквадратичное отклонение i-го коэффициента

Для случая линейного полинома $y=b_0+b_1x_1$ и вычисляются по следующим формулам $S^2_{b_0}$ $S^2_{b_1}$

$$S_{b_0} = \sqrt{\frac{S^2_{воспр.} \sum_{i=1}^{N} x_i}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i-1}^{N}\right)^2}}$$

$$S_{b_1} = \sqrt{\frac{S^2_{воспр.} \cdot N}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}}$$

Если $t_{b_1} > t_{табл.}(q, f)$, то коэффициент b_1 значим (значимо отличается от 0). В противном случае – незначим (≈0) и из уравнения может быть исключен.

4. **Проверка модели на адекватность осуществляется по критерию Фишера**.

Если $F = \frac{S^2_{ост}}{S^2_{воспр.}} < F_T(q, f_1, f_2)$, то модель адекватна (т.е. линейное уравнение регрессии адекватно описывает исследуемый объект).

для одинакового числа параллельных опытов $m_1=m_2=...m_n$.

$$S^2_{ост} = \frac{m\sum_{i=1}^{N}\left(\bar{y}_i - \hat{y}_i\right)^2}{N-l}$$

если опыты проведены без параллельных.

f_1 и f_2 – число степеней свободы (f_1 – для числителя, f_2 – для знаменателя).

$f_1=N-l$; (числ.);

l=n+1 – число членов аппроксимирующего полинома (число коэффициентов регрессии, включая свободный член).

$$S^2_{ост} = \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{N-l}$$

$f_2=N(m-1)$, (знамен.).

N – общее количество опытов.

n – количество факторов (x_1,x_2...)

Если не было параллельных опытов, то вместо проверки модели на адекватность выполняется ***оценка качества аппроксимации*** достигается сравнением остаточной дисперсии $S^2_{ост.}$ с дисперсией относительно среднего S^2_y

y_i – экспериментальное значение выходного параметра.

$$S^2_y = \frac{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}{N-l} \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

- среднее значение выходного параметра.

Уравнение регрессии будет иметь смысл, если дисперсия относительно среднего существенно больше, чем т.е. эти дисперсии должны отличаться значимо. Критерий Фишера в этом случае будет иметь вид:

$$F = \frac{S^2_y}{S^2_{ост}} > 1;$$

и в этом случае, чем $F>F_{табл}(q,f_1,f_2)$, тем уравнение регрессии эффективнее.

$f_1=N-1$; $f_2=N-1$; для выбранного q.

СТАТИСТИЧЕСКИЕ МОДЕЛИ В ВИДЕ НЕЛИНЕЙНЫХ ПОЛИНОМОВ

метод регрессионного анализа для составления статистической модели в виде полинома второй (или более высокой) степени:

$$\hat{y} = b_0 + \sum b_i x_i + \sum_{\substack{i=1 \\ i \neq j}}^{n} b_{ij} x_i x_j + \sum b_{ij} x_i^2 + ...+,...$$

Коэффициенты регрессии определяют также по МНК $F = \sum (y - \hat{y})^2 \to \min$

Пусть дано уравнение $\hat{y} = b_0 + b_1 x + b_2 x^2$ требуется определить b_0, b_1, b_2.

$$F = \sum_{i=1}^{N} \left(y_i - b_0 - b_1 x_i - b_2 x_i^2 \right)^2 \to \min$$

$$\begin{cases} \dfrac{\partial F}{\partial b_0} = -2\sum \left(y_i - b_0 - b_1 x_i - b_2 x_i^2 \right) \cdot 1 = 0; \\ \dfrac{\partial F}{\partial b_0} = -2\sum \left(y_i - b_0 - b_1 x_i - b_2 x_i^2 \right) \cdot x_i = 0; \\ \dfrac{\partial F}{\partial b_0} = -2\sum \left(y_i - b_0 - b_1 x_i - b_2 x_i^2 \right) \cdot x_i^2 = 0; \end{cases}$$

$$\begin{cases} Nb_0 + b_1\sum x_i + b_2\sum x_i^2 = \sum y_i \\ b_0\sum x_1 + b_1\sum x_i + b_2\sum x_i^3 = \sum x_i y_i \\ b_0\sum x_i^2 + b_1\sum x_i^3 + b_2\sum x_i^4 = \sum y_i x_i^2 \end{cases}$$

$$\begin{cases} Nb_0 + b_1 S_1 + b_2 S_2 = S_5 \\ b_0 S_1 + b_1 S_2 + b_2 S_3 = S_6 \\ b_0 S_2 + b_1 S_3 + b_2 S_4 = S_7 \end{cases};$$

Решая систему уравнений, вычисляем коэффициенты b_0, b_1, b_2.

$$b_0 = \frac{\begin{vmatrix} S_5 & S_1 & S_2 \\ S_6 & S_2 & S_3 \\ S_7 & S_3 & S_4 \end{vmatrix}}{\begin{vmatrix} N & S_1 & S_2 \\ S_1 & S_2 & S_3 \\ S_2 & S_3 & S_4 \end{vmatrix}} = \frac{S_5S_2S_4 + S_6S_3S_2 + S_7S_1S_3 - S_7S_2S_2 - S_6S_1S_4 - S_5S_3S_3}{NS_2S_4 + S_1S_3S_2 + S_2S_1S_3 - S_2^3 - S_1^2S_4 - NS_3^2};$$

- Аналогичным путем определяются коэффициенты параболы любого порядка. Исследования уравнений проводятся по статистическим критериям, также как в случае линейной регрессии. Однако, коэффициент корреляции r_{xy} рассчитывать не надо.

СТАТИСТИЧЕСКИЕ МОДЕЛИ НА ОСНОВЕ АКТИВНОГО ЭКСПЕРИМЕНТА

Активный эксперимент ставится по заранее составленному плану и обрабатывается по некоторому оптимальному алгоритму с целью составления математической модели. Одним из основных методов теории активного эксперимента является статистическое планирование эксперимента.

План эксперимента показывает расположение опытных точек в n-мерном факторном пространстве.

ПЛАНЫ ПЕРВОГО ПОРЯДКА
Полный факторный эксперимент

При планировании по схеме полного факторного эксперимента (ПФЭ) реализуются все возможные комбинации факторов на всех выбранных для исследования уровнях.

Необходимое количество опытов N при ПФЭ определяется по формуле:

$$N = l^n$$

N – число факторов;

l – число уровней, на которых варьируются факторы.

Уровни факторов – это границы исследуемой области по данному технологическому параметру.

В основном (обычно) применяется планирование на двух уровнях, т.е. l=2, тогда при n=2, $N=2^2=4$.

Нулевой (основной) уровень (центр плана эксперимента) – это некоторое начальное значение фактора при составлении математической модели.

Это точка с координатами

Интервал варьирования – часть области определения фактора, симметричная относительно его нулевого уровня.

Пример. Объект исследования – реактор, в котором выход продукта y зависит от двух факторов: температуры в реакторе (x_1) и давления (x_2). Известно априори, что T=100-200; P=10-20а, тогда 100 и 200, 10 и 20 – это два уровня, на которых варьируются факторы.

Верхний – 200^o и 20а

Нижний – 100^o и 10а

Основной нулевой уровень:150 15

Основной уровень:

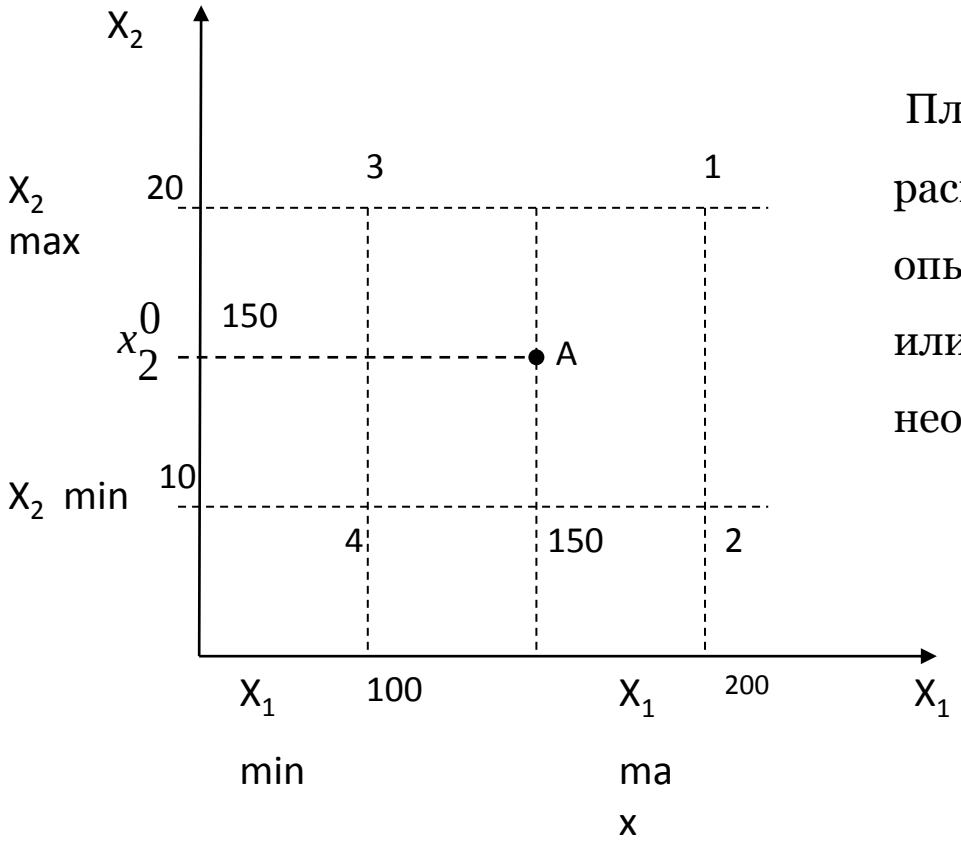
$$x_1^0 = \frac{x_1^{\max} + x_1^{\max}}{2} \quad x_2^0 = \frac{x_2^{\min} + x_{21}^{\max}}{2};$$

Интервалы варьирования:

$$\Delta X_1 = \frac{x_1^{\max} - x_1^{\min}}{2} = \frac{200 - 100}{2} = 50;$$

$$\Delta X_2 = \frac{x_2^{\max} - x_2^{\min}}{2} = \frac{20 - 10}{2} = 5;$$

В координатах на плоскости это можно представить следующим образом:



План эксперимента указывает расположение n – мерном пространстве опытных точек независимых переменных или условия всех опытов, которые необходимо провести

При ПФЭ эксперимент ставится только на границе области, т.А – центр области. В большинстве случаев эксперимент задается в виде матрицы планирования – это план (таблица), каждая строчка который представляет собой условия опыта, а каждый столбец матрицы соответствует значениям переменных в различных опытах.

Составим матрицу планирования для предыдущего примера.

X_1-T=100-200°C имеем два фактора,

X_2-P=10-20а, следовательно $N=2^n=4$.

Это ПФЭ типа 2^2:

N	X_1	X_2	y
1	100 min	10 min	Y_1
2	100 min	20 max	Y_2
3	200 max	10 min	Y_3
4	200 max	20 max	Y_4

Матрица планирования для ПФЭ 2^2 – все возможные комбинации факторов на двух уровнях. Это матрица планирования в натуральном масштабе.

Матрица планирования составляется для того, чтобы эксперимент провести по определенному плану, определить значения выходного параметра в каждом опыте и построить статистическую модель.

При планировании первого порядка получают математическую модель вида:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$ - Линейное уравнение

КОДИРОВАНИЕ ПЕРЕМЕННЫХ

Для удобства расчетов, перейдем от натуральных координат (натуральных единиц измерения) к безразмерным. Формула перехода или кодирования имеет вид:

$$X_i = \frac{x_i - x_i^0}{\Delta x_i},$$

x_i – значения (верхний или нижний уровень) натуральной переменной.

x_i^0 - основной уровень натуральной переменной.

Δx_i - интервал варьирования натуральной переменной.

X_i – кодированное значение i-го фактора (на верхнем или на нижнем уровне).

T=100-200°C P=10-20а

Перейдем от натуральных переменных к кодированным:

Для температуры

$$X_1^в = \frac{200-150}{50} = 1;$$
$$X_1^н = \frac{100-150}{50} = -1;$$

Для давления

$$X_2^в = \frac{20-15}{5} = 1;$$
$$X_2^н = \frac{10-15}{5} = -1;$$

Фактически мы обозначили значения факторов на верхнем уровне +1, (200,20), а на нижнем (100, 10) - -1;

Это матрица планирования в безразмерном масштабе.

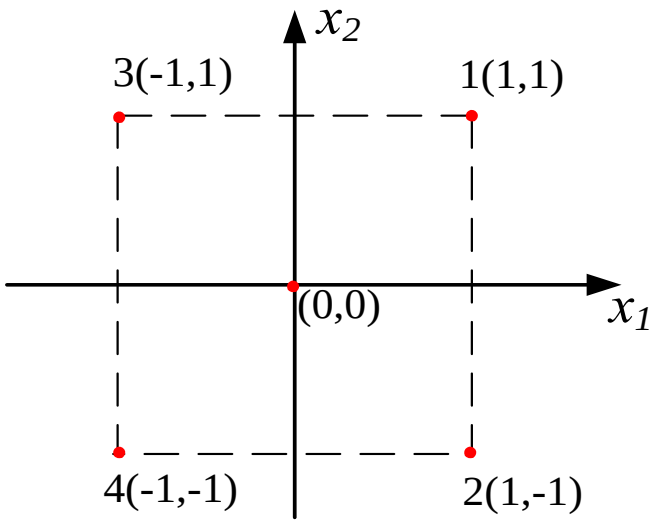
x_o – фиктивная переменная (+1), необходимое для вычисления свободного члена полинома.

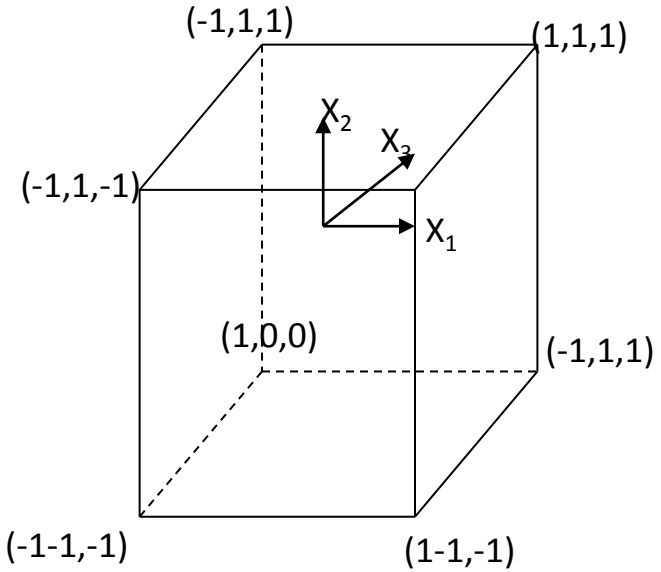
N	x_0	x_1	x_2
1	+1	+1	+1
2	+1	+1	-1
3	+1	-1	+1
4	+1	-1	-1

или

N	x_0	x_1	x_2
1	+	+	+
2	+	+	-
3	+	-	+
4	+	-	-

Расположение опытных точек в факторном пространстве будет следующим:





СВОЙСТВА МАТРИЦЫ ПЛАНИРОВАНИЯ

1. *ортогональность:*

скалярное произведение двух любых столбцов матрицы равно нулю:

$$\sum_{i=1}^{N} x_{ui} x_{ji} = 0; \qquad u \neq j;\ u,\ i = 1,\ldots,n$$

СВОЙСТВА МАТРИЦЫ ПЛАНИРОВАНИЯ

- *симметричность:*

сумма элементов всех столбцов матрицы, кроме первого, равна нулю:

$$\sum_{i=1}^{N} x_{iu} = 0, \qquad u = 1,\ldots,n;$$

СВОЙСТВА МАТРИЦЫ ПЛАНИРОВАНИЯ

- *нормировка:*

сумма квадратов элементов каждого столбца равна числу опытов

$$\sum_{i=1}^{N} x_{iu}^2 = N, \qquad u = 1,\ldots,n;$$

N	x_1	x_2	x_1x_2
1	+	+	+
2	-	+	-
3	+	-	-
4	-	-	+

СВОЙСТВА МАТРИЦЫ ПЛАНИРОВАНИЯ

4. Свойство ***ротатабельности:*** дисперсия предсказанного значения выходного параметра в любой точке факторного пространства при ПФЭ минимальна. Это означает, что ошибка определения коэффициентов регрессии в любой точке от центра плана одинакова и минимальна.

РАСЧЕТ КОЭФФИЦИЕНТОВ РЕГРЕССИИ.

После того, как составлен план, проводят эксперименты и на основании результатов рассчитываю коэффициенты в уравнении регрессии по формулам:

$$b_0 = \frac{1}{N}\sum_{i=1}^{N} x_{0i} \cdot y_i; \qquad b_i = \frac{1}{N}\sum_{i=1}^{N} x_{iu} \cdot y_i; \qquad b_{ij} = \frac{1}{N}\sum_{i=1}^{N} x_{iu} x_{ij} y_i;$$

u=1,..., n (факторы)

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2$$

Эти простые формулы получены благодаря **свойствам матрицы планирования** также на основании метода наименьших квадратов.

$$b_0 = \frac{\sum y_i \cdot \sum x_{1i}^2 - \sum x_{1i} y_i \cdot \sum x_i}{N\sum x_{1i}^2 - \left(\sum x_{ii}\right)^2} = \frac{\sum y_i \cdot N - 0 \cdot \sum x_i \cdot y_i}{N^2 - 0} = \frac{\sum y_i}{N};$$

$$b_1 = \frac{N\sum x_{1i} y_i - \sum x_{1i} \sum y_i}{N\sum x_{ii}^2 - \left(\sum x_{1i}\right)^2} = \frac{N \cdot \sum x_i y_i - 0 \cdot \sum y_i}{N^2 - 0} = \frac{\sum x_i y_i}{N};$$

b_{ij} - коэффициенты регрессии, характеризующие взаимодействие факторов.

Пример:

N	x_0	x_1	x_2	x_1x_2	Y
1	+1	1	1	1	85
2	+1	1	-1	-1	66
3	+1	-1	-1	-1	56
4	+1	-1	-1	1	50

$$b_0 = \frac{85+66+56+50}{4} = 64.25$$

$$b_1 = \frac{85+66-56-50}{4} = 11.25$$

$$b_2 = \frac{85-66+56-50}{4} = 6.25$$

$$b_{12} = \frac{85-66-56+50}{4} = 3.25$$

$$\hat{y} = 64.25 + 11.25x_1 + 6.25x_2 + 3.25x_1x_2$$

После вычисления коэффициентов регрессии приступают к статистическому анализу уравнения регрессии