ДОНСКОЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

УПРАВЛЕНИЕ ДИСТАНЦИОННОГО ОБУЧЕНИЯ И ПОВЫШЕНИЯ КВАЛИФИКАЦИИ

Кафедра «Технология вяжущих веществ, бетонов и строительной керамики»

Методические указания

к практической работе по теме

«Регрессионный анализ нелинейных моделей»

Автор

Серебряная И.А.

Ростов-на-Дону, 2017

Аннотация

Методические указания регламентируют порядок организации однофакторного эксперимента и математической обработки его результатов. Указания содержат правила выбора факторов и функции отклика, определение вида регрессионной связи, методики построения уравнений регрессии и статистической оценки результатов.

Предназначены для обучающихся по направлению подготовки 27.03.01 «Стандартизация и метрология», 27.04.02 «Управление качеством», 08.04.01 «Строительство».

Автор

к.т.н., доцент
кафедры «ТВВБиСК»
Серебряная И.А.

Оглавление

1. ПЛАНИРОВАНИЕ ОДНОФАКТОРНОГО ЭКСПЕРИМЕНТА.

1.1. В практике экспериментальных исследований довольно часто решаются задачи, в которых необходимо установить влияние одной переменной величины X на другую переменную Y. В таких задачах X является независимой детерминированной величиной, значения которой задаются исследователем, и называется фактором. Величина Y является зависимой переменной, представляет собой результаты опытов и называется откликом. В реальных условиях на результаты опытов оказывает влияние не только изменение значений фактора X, но и многие неконтролируемые и неуправляемые факторы. В следствие этого каждому значению фактора X = x соответствует не одно, а некоторая совокупность возможных значений отклика Y, т. е. переменная Y является случайной величиной, связь между переменными X и Y носит вероятностный (стохастический) характер.

В практических задачах экспериментатор получает не множество значений Y, соответствующих фиксированному значению X = x, а ограниченное число, иногда одно значение Y = y. Функция Y = f(x), описывающая изменение случайной переменной Y при изменении неслучайной переменной X, называется функцией регрессии и описывается уравнением регрессии. Эксперимент, в котором изучается влияние на функцию Y только одной независимой переменной X, называется однофакторным.

1.2. Планирование эксперимента должно начинаться с четкого определения цели исследований и выбора критерия ее достижения – отклика Y. В качестве цели исследований могут быть приняты, например, повышение прочности, однородности или морозостойкости материала, снижение себестоимости и т. п. Соответственно критериями достижения поставленной цели могут служить: прочностные показатели, количество циклов замораживания и оттаивания, удельный расход цемента, расход пара и др.

В эксперименте может быть один или несколько откликов. Каждый отклик должен отвечать следующим требованиям: иметь физический смысл, оцениваться количественно, быть устойчивым к малым случайным воздействиям.

При выборе фактора X необходимо использовать априорную, т. е. имеющуюся до проведения опыта информацию с тем, чтобы выбранный фактор оказывал существенное влияние на изучаемую функцию.

В любом случае факторы должны быть управляемыми, т. е. экспериментатор должен иметь возможность задавать и выдерживать в опытах требуемые значения фактора. Большое значение в эксперименте имеет правильный выбор интервалов изменения (варьирования) факторов. Слишком узкий диапазон изменений фактора может привести к неверному выводу о несущественном влиянии данного фактора на изучаемую функцию. Однако и неоправданное увеличение диапазона нежелательно, т. к. приводит к увеличению числа опытов или же к снижению точности результатов. Основанием для выбора диапазона варьирования могут служить литературные данные, а при отсутствии таковых задачу решают на интуитивном уровне.

Значения фактора X в эксперименте следует принимать таким образом, чтобы каждое последующее значение X_i отличалось от предыдущего на постоянную величину – шаг h. Это упрощает вычисление коэффициентов уравнения регрессии, построение графиков, повышает точность решения.

Исходя из удобства построения графика парной зависимости, в эксперименте должно быть не менее пяти опытных точек, а для облегчения обработки результатов количество опытов N принимают нечетным.

1.3. При проведении эксперимента необходимо оценить ошибку воспроизводимости $s_э$, которая вычисляется по результатам дублирующих (параллельных) опытов, представляющих собой полное повторное воспроизведение всех условий опыта в определенной точке. Количество дублирующих опытов m принимается не менее трех. Для расчета ошибки эксперимента вначале рассчитывают выборочное среднее y (1) и выборочную дисперсию $s^2_э$ (2):

$$\bar{y} = \frac{1}{m}\sum_{i=1}^{\bar{m}} y_i \quad , \qquad (1)$$

$$S_э^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2 \quad , \qquad (2)$$

затем $S_э = \sqrt{S_э^2}$, (3)

2. ОПРЕДЕЛЕНИЕ ВИДА РЕГРЕССИОННОЙ СВЯЗИ

2.1. В ряде случаев выбор типа уравнения, которым можно описать зависимость Y = f(x), можно сделать на основе теоретических представлений или работ других авторов, изучавших аналогичные зависимости. Чаще форма регрессионной связи неизвестна и ее надо установить по результатам эксперимента. Наиболее простым способом является построение графика зависимости по экспериментальным значениям "на глаз". Полученную эмпирическую линию сравнивают с типичными графиками распространенных уравнений. По виду выбранного уравнения регрессии следует проанализировать эмпирическую линию регрессии, определив вначале, является зависимость линейной или криволинейной. Линейная зависимость описывается линейным уравнением регрессии:

$$\hat{y} = a + bx, \qquad (4)$$

Если зависимость криволинейная, то следует выяснить, является ли она монотонно возрастающей (убывающей) или содержит характерные точки – максимум, минимум, точки перегиба. В последнем случае необходимо убедиться не являются ли эти точки следствием ошибок опыта. Проверкой может служить постановка дополнительных опытов в области характерных точек.

2.2. Часто эмпирические зависимости могут быть описаны уравнением параболы второго порядка вида:

$$\hat{y} = a + bc + cx^2, \qquad (5)$$

Графики этой зависимости могут иметь различную степень кривизны и обладают одной экстремальной точкой (максимум или минимум). В принятом диапазоне варьирования фактора экстремальной точки, то принимается уравнение третьего порядка, если три – то четвертого порядка и т.д.

2.3. Большая группа кривых может быть описана трансцендентными и степенной функциями, но такие уравнения регрессии мало удобны для пользования, поэтому при расчетах выполняются преобразования, позволяющие привести эти функции к линейному виду.

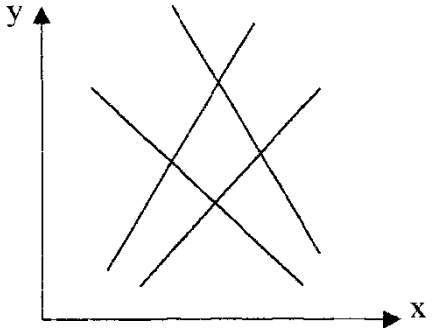
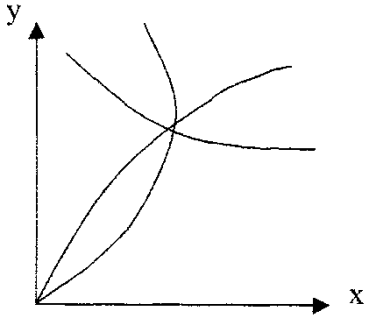
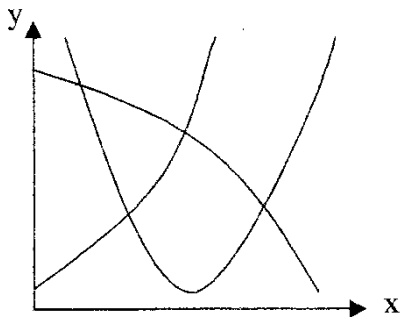
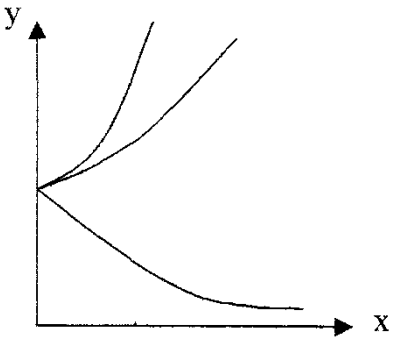
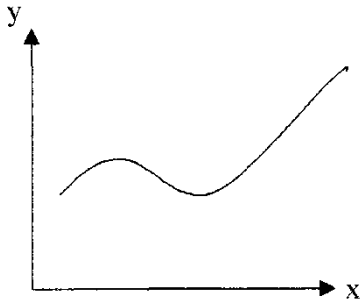
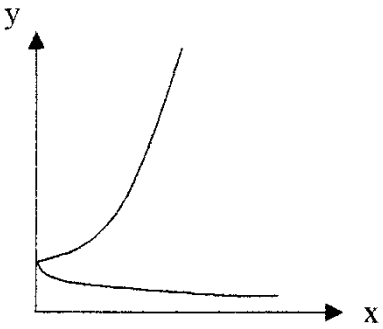
Иногда эмпирическая кривая похожа на несколько кривых, описываемых различными уравнениями. Возможны также случаи, когда уравнение достаточно точно описывает зависимость между

X и Y, но график этого уравнения не похож на эмпирическую кривую. Хорошие результаты дает метод, при котором экспериментальная зависимость описывается несколькими уравнениями, а окончательное решение принимается после сравнения расчетных значений y по каждому уравнению с экспериментальными величинами y. Можно считать, что уравнение регрессии удовлетворительно описывает экспериментальные данные, если:

- отклонения результатов эксперимента от расчетных величин минимальны;
- отклонения по абсолютной величине близки друг к другу;
- число отклонений со знаком "+" и знаком "-" примерно равно и отмечается тенденция их чередования.

Более надежную и объективную оценку точности принятого уравнения дает регрессионный анализ (раздел 5 настоящих МУ).

Однофакторные зависимости y = f(x)

График зависимости	Уравнение. Способ расчета	График зависимости	Уравнение. Способ расчета
	$y = a + bx$		$y = ax^{в}$
	Определение коэффициентов из системы уравнений (8,11)		Привидение к линейному виду логарифмированием: $Z=A+bx'$, где $Z = lgy$, $A = lga$, $x' = Igx$
	$y=a+bx+cx^2$		$y = xe^{bx}$
	Определение коэффициентов МНК по стандартной программе		Приведение к линейному виду логарифмированием: $Z = A + Bx$, где $Z = lgy$, $A = lga$, $B = b \cdot lge$
	$y=a+bx+ cx^2+dx^2$		$y = ab^x$
	Определение коэффициентов МНК по стандартной программе		Приведение к линейному виду логарифмированием: $Z = A + Bx$, где $Z = lgy$, $A = lga$, $B = lgl$

Однофакторные зависимости y = f(x)

3. ЛИНЕЙНАЯ РЕГРЕССИЯ

3.1. Наиболее простым видом зависимости является линейная, которая описывается уравнением вида:

$$\hat{y} = a + bx,$$

где $\hat{y}$ – расчетное значение функции;

a, b – коэффициенты уравнения, которые надо вычислить.

Если бы не случайный характер функции y, для определения коэффициентов уравнения (5) было бы достаточно поставить всего два опыта. Из-за разброса опытных данных приходиться ставить большее число опытов N. При этом число возможных уравнений значительно больше, чем неизвестных, и их решение дает различные значения коэффициентов. Задача состоит в том, чтобы найти значения коэффициентов, которые наилучшим образом будут описывать все опытные данные.

3.2. Наиболее корректным решением этой задачи является метод наименьших квадратов – МНК.

Из-за ошибок эксперимента и несовершенства гипотезы линейной связи всегда между опытными значениями y_1, и расчетными y; будет расхождение:

$$\Delta_i = y_i - \hat{y}_i, \qquad (6)$$

Наилучшее линейное приближение достигается в том случае если минимизируется не сумма абсолютных отклонений по всем N опытам $\sum_{i=1}^{N} |\Delta_i|$, а сумма квадратов отклонений:

$$\sum_{i=1}^{N} \Delta_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - a - bx_i)^2 \to \min, \qquad (7)$$

Для нахождения минимума функции (7) необходимо приравнять нулю частные производные по неизвестным a и b . После дифференцирования и алгебраических преобразований получаем систему нормальных уравнений:

$$\left.\begin{aligned} Na + b\sum_{}^{N} x_i &= \sum_{}^{N} y_i \\ a\sum_{}^{N} x_i + b\sum_{}^{N} x_i^2 &= \sum_{}^{N} y_i x_i \end{aligned}\right\}, \qquad (8)$$

Решение системы уравнений (8) производится любым известным способом.

3.3. Расчет коэффициентов регрессии по МНК является наиболее точным, но требует большого объема вычислений. Менее точным, но более простым является расчет по способу средней, который рекомендуется применять, например, при определении формы связи. При этом способе считается достаточным, чтобы сумма всех отклонений экспериментальных данных от теоретической линии уравновешивалась:

$$\sum_{i=1}^{N} \Delta_i = \sum_{i=1}^{N} (y_i - a - bx_i) = 0, \qquad (9)$$

Для нахождения коэффициентов прямой по этому способу все опытные данные разбиваются на две примерно равные части, и для каждой из них записывается

Условие (9):

$$\left.\begin{aligned} \sum_{i=1}^{n} (y_i - a - bx_i) &= 0 \\ \sum_{i=n+1}^{N} (y_i - a - bx_i) &= 0 \end{aligned}\right\}, \qquad (10)$$

где N – общее число опытов;

n – число опытов в первой группе.

После алгебраических преобразований системы (10) получим:

$$\left.\begin{aligned} an + b\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ a(N-n) + b\sum_{i=n+1}^{N} x_i &= \sum_{i=n+1}^{N} y_i \end{aligned}\right\}, \qquad (11)$$

Решив систему уравнений (11), получим значения коэффициентов a и b.

4. ТРАНСЦЕНДЕНТНАЯ И СТЕПЕННАЯ РЕГРЕССИИ

4.1. Теоретически любую зависимость можно описать полиномиальным уравнением. Но в ряде случаев для описания сложных кривых необходимо использовать полиномы высоких порядков, что усложняет расчеты и практическое использование полученных уравнений. Поэтому для описания сложных зависимостей нередко используют некоторые трансцендентные функции, такие как показательная или логарифмическая, которые содержат небольшое число коэффициентов.

Для упрощения вычисления коэффициентов трансцендентной регрессии путем алгебраических преобразований уравнение приводят к линейной функции, содержащей только два искомых коэффициента. Эта процедура называется линеаризацией. Ниже показаны приемы линеаризации отдельных видов трансцендентных и степенной функции.

4.2. Показательная функция в общем случае записывается в виде:

$$\hat{y} = ab^x \quad \text{или} \quad \lg \hat{y} = \lg a + x \lg b, \qquad (12)$$

Приняв: lgy = Z, lga = A и lgb = B, получаем линейное уравнение вида:

$$Z = A + Bx \qquad (13)$$

Разновидностями показательных уравнений являются:

$\hat{y} = ae^{bx}$, $\hat{y} = c + ae^{bx}$ и др.

Оба уравнения приводятся к линейному виду: Z = A + Bx,

где Z = lgy, A = lga, B = 1ge·b = 0,4343b,

Однако для второго уравнения после вычисления коэффициентов и обратного преобразования условных переменных должна быть введена постоянная "c".

4.3. Уравнение логарифмического вида $\hat{y} = a + b \lg x$, приводится к линейному виду путем введения новой переменной x/ = lgx :

$$\hat{y} = a + bx' \qquad (14)$$

4.4. Степенное уравнение вида: $y = ax^b$ преобразуется логарифмированием: lgy=lga+b-lgx.

Перейдя к условным переменам Z = lgy и $x^{/}$ = lgx, получим линейное уравнение:

$$Z = A + bx. \qquad (15)$$

Коэффициенты в линейных уравнениях (13) – (15) вычисляются теми же методами, что и для линейной регрессии. На заключительном этапе расчета выполняется обратное преобразование: переход от условных переменных к фактическим – x и y.

5. РЕГРЕССИОННЫЙ АНАЛИЗ УРАВНЕНИЙ

5.1. Уравнение регрессии позволяет определить форму связи и количественно описать изучаемую зависимость. Однако, учитывая ее вероятностный характер, необходимо объективно оценить "качество" полученного уравнения, т. е. точность описания поведения объекта и информационную ценность математического описания. С этой целью проводят регрессионный анализ уравнения.

Регрессионные уравнения представляют собой вероятностные математические модели поведения случайной величины Y. Суждения о свойствах таких моделей принимаются с некоторой вероятностью P, которая называется доверительной вероятностью. При этом допускается определенный риск α= 1-P, называемый уровнем значимости.

Для большинства технических задач поискового характера достаточна доверительная вероятность P = 0,90—0,95, что соответствует уровню значимости α = 0,10-0,05.

5.2. Регрессионный анализ включает проверку адекватности и информационной ценности математического уравнения (модели).

Для оценки адекватности, т. е. соответствия полученного уравнения опытным данным, проверяют гипотезу о равенстве дисперсии неадекватности $S^2_{на}$ и дисперсии воспроизводимости эксперимента

$$H_o: \sigma^2_{на} = \sigma^2_э, \qquad (16)$$

Эту гипотезу проверяют по критерию Фишера, фактическое значение которого рассчитывают по формуле:

$$F_{факт} = S^2_{на} / S^2_э \qquad (17).$$

Дисперсия неадекватности $\overline{S}^2$ характеризуется отклонение расчетных значений отклика $\hat{y}$ от опытных y:

$$\overline{S}^2_{на} = \frac{1}{N-B}\sum^{N}(y_i - \hat{y}_i)^2, \qquad (18)$$

где N — число опытов,

B – число коэффициентов регрессии в уравнении,

y_i и $\hat{y}_i$ – опытное и расчетное значения отклика в i-ой точке.

Дисперсия воспроизводимости $\overline{S}^2_э$ рассчитывается по формуле (2). Для проверки гипотезы (16) находят значение $F_{табл}$ при α = 0,05 (0,10), числе степеней свободы числителя $f_1 = N - B$ и знаменателя $f_2 = m - 1$ (m – число параллельных опытов) и сравнивают его с факт . Если $F_{факт} < F_{табл}$, то гипотезу H_o не отвергают и с риском α считают уравнение адекватным. Если же $F_{факт} > F_{табл}$, то уравнение считается неадекватным и требуется построение другого более точного.

5.3. Для оценки информационной способности уравнения проверяют гипотезу о равенстве общей дисперсии выхода и дисперсии неадекватности Но:

$$H_0: \sigma^2_{общ} = \sigma^2_{на} \qquad (19)$$

Эту гипотезу проверяют по критерию Фишера, значение которого рассчитывается по формуле:

$$F_u = S^2_{общ} / S^2_{на}. \qquad (20)$$

Общая дисперсия характеризует отклонение всех опытных значений y_i , от среднего $\overline{y} = \frac{1}{N}\sum^{N}(y_i)$ и рассчитывается по формуле:

$$\overline{S}^2_{общ} = \frac{1}{N-1}\sum^{N}(y_i - \overline{y})^2 \qquad (21)$$

По таблицам находят значение критерия $F_{табл}$ при α==0,05 (0,10), fi =N-1

и $f_2 = N - B$ и сравнивают с $F_и$. если $F_и < F_{табл}$, то гипотезу Но (19) не отвергают, т. е. можно считать, что полученное уравнение

регрессии описывает опытные данные так же, как простейшее уравнение $\hat{y} = \overline{y}$ и информационной ценности не имеет.

Если же $F_и > F_{табл}$ то гипотеза Но отвергается и можно считать, что анализируемое уравнение обладает информационной ценностью.

СПИСОК ЛИТЕРАТУРЫ

1. Таблицы планов эксперимента для факторных и полиномиальных моделей: Справочник / Под ред. В.В.Налимова. – М: Металлургия, 1982. – 752 с.

2. Адлер Ю.П., Маркова Е.В. Грановский Ю.В. Планирование эксперимента при поиске оптимальных условий. – М.: Наука, 1976. – 279 с.