ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ И НАУКИ РФ
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ДИЗАЙНА И ТЕХНОЛОГИИ

КАФЕДРА ВЫСШЕЙ МАТЕМАТИКИ

Ю.К. ОСТРОВСКИЙ

МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ МОДУЛЯ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

ЧАСТЬ I. Критерий согласия Пирсона. Метод наименьших квадратов.

УЧЕБНОЕ ПОСОБИЕ

Утверждено в качестве учебного пособия
Редакционно – издательским советом МГУДТ

МГУДТ 2007

УДК 519.2(076)
О78

Куратор РИС А.П.Жихарев

Работа рассмотрена на заседании кафедры высшей математики и рекомендована к печати.

Зав. кафедрой А.А.Горский

Автор:

Ю.К.Островский, к.ф.-м.н.

Рецензент: профессор Алексашенко А.А., д.т.н.

О78 **Островский Ю.К.** Методическое обеспечение модуля по математической статистике. Часть I. Критерий согласия Пирсона. Метод наименьших квадратов: Учебное пособие/ Островский Ю.К.- М.: ИИЦ МГУДТ, 2007 – 22 с.

Работа содержит 50 вариантов заданий по критерию согласия χ^2 и линейному регрессионному анализу. Такое пособие можно использовать для самостоятельных занятий и расчетно-графических работ на механическом, технологическом, экономическом и других факультетах при освоении модуля по теории вероятностей и математической статистике.

УДК 519.2(076)

© МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ДИЗАЙНА И ТЕХНОЛОГИИ, 2007

СОДЕРЖАНИЕ

с
Введение…… …… 4

Задание 1. Проверка статистической гипотезы о нормальном распределении генеральной совокупности по критерию согласия Пирсона 5

Литература ……...…….. 12

Задание 2. Метод наименьших квадратов …….. 13

Литература ……... 20

Введение.

Цель этого пособия состоит в том, чтобы развить у студентов способности к статистической обработке данных экспериментов, в том числе с помощью вычислительных устройств.

Каждое из двух заданий содержит необходимые краткие теоретические сведения, а также ссылки на литературу, где такие сведения изложены подробно и доказательно.

Все задания могут быть выполнены с помощью калькулятора, однако более продуктивным стал бы анализ статистических данных с помощью компьютера или, что тоже не бесполезно, овладение средами вроде MathCad, MathLab и т.д.

Задание № 1. Проверка статистической гипотезы о нормальном распределении генеральной совокупности по критерию согласия Пирсона.

При анализе статистических данных часто возникает проблема создания гистограммы этих данных с целью дальнейшей проверки гипотезы о том или другом распределении случайной величины по какому-нибудь критерию согласия. Прежде чем заняться такой проверкой, необходимо корректно выбрать число интервалов гистограммы. Решающим фактором при этом является, конечно, опыт, однако для начинающего пользу может оказать формула Штюргесса [1]:

$$k = 1 + 3.32 \lg N \quad , \qquad (1)$$

где k - округленное до целого число интервалов гистограммы, N - объем выборки.

При использовании критерия χ^2 обычно рассматривают выборки с объемом не менее 100 и числом элементов n_i выборки, попавших в i - ый интервал гистограммы, не менее 5 (в противном случае интервалы объединяют).

Кроме того, выборка должна быть репрезентативной, т.е. давать достаточное представление об особенностях генеральной совокупности (по крайней мере, быть случайной), а выборочные оценки должны быть:
- состоятельны, т.е. стремиться по вероятности с увеличением объема выборки к оцениваемому параметру;
- несмещенными, т.е. их математические ожидания должны совпадать с оцениваемыми параметрами;
- эффективными, т.е. иметь минимальную дисперсию.

Будем считать n_i случайной величиной (с.в.), подчиняющейся биноминальному закону распределения с математическим ожиданием

(м.о.) и средним квадратическим отклонением (с.к.о.) $\sqrt{Np_i(1-p_i)}$, где p_i - вероятность попадания с.в. в i - ый интервал выборки. При $N >> 1$ и $p_i << 1$ с.к.о $\cong \sqrt{Np_i}$, а с.в. $\xi_i = \frac{n_i - Np_i}{\sqrt{Np_i}}$ можно считать распределенной нормально. Тогда

$$\sum_{i=1}^{k} \xi_i \sqrt{p_i} = \frac{\sum_{i=1}^{k} \left(n_i - Np_i\right)}{\sqrt{N}} = 0 \quad , \qquad (2)$$

$$\chi^2 = \sum_{i=1}^{k} \xi_i^2 = \sum_{i=1}^{k} \frac{\left(n_i - Np_i\right)^2}{Np_i} \quad , \qquad (3)$$

где $N = \sum_{i=1}^{k} n_i$.

В случае проверки гипотезы о нормальном распределении с.в. существуют еще два (кроме (2)) уравнения, линейные по $\frac{n_i}{N}$, определяющие выборочное среднее $x_в$ и выборочное с.к.о. $\sigma_в$:

$$x_в = \sum_{i=1}^{k} \frac{n_i}{N} \cdot x_i \quad , \qquad (4)$$

$$\sigma_в^2 = \sum_{i=1}^{k} \frac{n_i}{N} \cdot \left(x_i - x_в\right)^2 \quad , \qquad (5)$$

где $x_i = \frac{x_i + x_{i+1}}{2}$ - середина i - го интервала выборки;

$i = 1,2,3,...,k; x_1, x_2,...,x_{k+1}$ - границы интервалов выборки.

Количество связей (2), (4) и (5), равное r, называются числом связей подгонки, а число $(k-r)$ - числом степеней свободы подгонки [3].

Подгонкой будем называть или подбор такого распределения $P(x)$ с.в. X с плотностью распределения $f_0(x)$, или коррекцию $f_0(x) \to f(x)$, для которых вероятности $p_i = \int_{x_i}^{x_{i+1}} f_0(x)dx$ или $p_i = \int_{x_i}^{x_{i+1}} f(x)dx$ оптимальны в смысле минимальности χ^2 (3).

Суть критерия согласия Пирсона состоит в сравнении числа χ^2 (3), полученного по выборочным данным, с числом χ_q^2, определяемым из уравнения:

$$\alpha = \int_{\chi_q^2}^{+\infty} k_n(x)dx \quad , \tag{6}$$

где $\alpha = 1 - P$ - уровень значимости гипотезы о нормальном распределении с.в.; $k_n(x)$ - плотность распределения χ^2(3); $n = k - r$; $P = P\{\chi^2 < \chi_q^2\}$ - доверительная вероятность; $k_n(x) = \dfrac{x^{\frac{n}{2}-1} \cdot e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)}$;

$\Gamma(p) = \int_0^{+\infty} e^{-z} z^{p-1} dz$ - гамма-функция Эйлера.

Вывод формулы для $k_n(x)$ можно посмотреть, например, в работе [2].

Если $\chi^2 > \chi_q^2$, то гипотезу о нормальном распределении генеральной совокупности отвергают (и принимают в противном случае).

В настоящей работе $k = 10, r = 3, \alpha = 0.05$. Интеграл (6) затабулирован в таблицах, имеющихся во всех учебниках и задачниках по теории

вероятностей и математической статистике и, конечно, в работах [1-5]. Из этих таблиц для $n = 10 - 3 = 7$ и указанному значению α находим $\chi_q^2 = 14.1$.

Примерная схема выполнения такой работы могла бы быть следующей.

1. По формулам $x_i = x_{\min} + h \cdot (i-1)$ и $x_i = 0.5(x_i + x_{i+1})$ вычисляются границы x_i интервалов $(i = 1,2,3,\ldots,k+1)$ и середины x_i интервалов $(i = 1,2,3,\ldots,k)$ выборки.
2. По формулам (4) и (5) вычисляем $x_в$ и $\sigma_в$ и полагаем $m_x \cong x_в, \sigma_x \cong \sigma_в$.
3. Строим график плотности распределения

$$f_0(x) = \frac{\exp\left\{-\frac{\left(x - m_x\right)^2}{2\sigma_x^{\,2}}\right\}}{\sqrt{2\pi} \cdot \sigma_x}$$ и гистограмму относительных частот

$$\frac{n_i}{h \cdot N}.$$

4. Вычисляем безразмерные границы $z_i = \frac{x_i - m_x}{\sigma_x}, i = 1,2,3,\ldots,k+1,$

интервалов выборки и находим «теоретические» частоты

$$p_i = \Phi\left(z_{i+1}\right) - \Phi\left(z_i\right), i = 2,3,4,\ldots,k-1, p_1 = \Phi(z_2) - \Phi(-\infty),$$
$$p_{10} = \Phi(+\infty) - \Phi(z_k),$$

где $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z \exp\left\{-\frac{t^2}{2}\right\} dt$ - интеграл Лапласа,

затабулированный в таблицах, содержащихся в учебниках и задачниках по математической статистике. Зная, что $\Phi(\pm\infty) = \pm 0.5$, следует соблюдать аккуратность при экстраполяции промежуточных значений $\Phi(z_i)$ при пользовании таблицами,

надежнее при вычислении частот p_i использовать численные квадратуры: $p_i = \frac{1}{\sqrt{2\pi}} \int_{z_i}^{z_{i+1}} \exp\left\{-\frac{t^2}{2}\right\} dt, i = 2,3,4,\ldots,k-1,$

$$p_1 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_2} \exp\left\{-\frac{t^2}{2}\right\} dt = 0.5 - \frac{1}{\sqrt{2\pi}} \int_{0}^{|z_2|} \exp\left\{-\frac{t^2}{2}\right\} dt,$$

$$p_{10} = \frac{1}{\sqrt{2\pi}} \int_{z_k}^{+\infty} \exp\left\{-\frac{t^2}{2}\right\} dt = 0.5 - \frac{1}{\sqrt{2\pi}} \int_{0}^{z_k} \exp\left\{-\frac{t^2}{2}\right\} dt.$$

5. По формуле (3) вычисляем χ^2 и сравниваем это число с χ_q^2. Если $\chi^2 < \chi_q^2$, то гипотезу о нормальном распределении генеральной совокупности принимаем, если $\chi^2 > \chi_q^2$, то гипотезу отвергаем и переходим к следующему пункту.
6. Вычисляем центральные (выборочные) моменты 3,4 и 5 порядков по формулам: $\mu_s^{(0)} = \sum_{i=1}^{k} \frac{n_i}{N} \cdot \left(x_i - x_в\right)^s, s = 3,4,5,$ и вводим поправки Шеппарда [4,5], вспоминая, что $\mu_2^{(0)} = \sigma_в^2$:

$$\mu_{2ш}^{(0)} = \mu_2^{(0)} - \frac{h^2}{12}, \mu_{3ш}^{(0)} = \mu_3^{(0)}, \mu_{4ш}^{(0)} = \mu_4^{(0)} - \frac{\mu_2^{(0)}}{2} h^2 + \frac{7}{240} h^4,$$

$$\mu_{5ш}^{(0)} = \mu_5^{(0)} - \frac{5\mu_3^{(0)}}{6} h^2.$$

Полагаем далее $\sigma_ш = \sqrt{\mu_{2ш}^{(0)}}$ и сравниваем исправленные моменты с исходными. Если отличие не превосходит 1-2%, то поправки Шеппарда не учитываем в дальнейших вычислениях.

7. Вычисляем коэффициенты асимметрии, эксцесса и ресимметрии:

$$A_ш=\frac{\mu_{3ш}^{(0)}}{\sigma_ш^3},E_ш=\frac{\mu_{4ш}^{(0)}}{\sigma_ш^4}-3,R_ш=\frac{\mu_{5ш}^{(0)}}{\sigma_ш^5}\ .$$

8. В разложении Эджворта [4,5] для плотности распределения с.в. учитываем только члены, линейные по $A_ш,E_ш,R_ш$:

$$f_э(z)=f_0(z)-\frac{A_ш}{6}f_0^{(3)}(z)+\frac{E_ш}{24}f_0^{(4)}(z)-\frac{R_ш-10A_ш}{120}f_0^{(5)}(z)$$

. В этой формуле

$$f_0(z)=\frac{1}{\sqrt{2\pi}\cdot\sigma_ш}\exp\left\{-\frac{z^2}{2}\right\},$$

$$f_0^{(2)}=\left(z^2-1\right)f_0(z),$$

$$f_0^{(3)}(z)=\left(-z^3+3z\right)f_0(z),$$

.

$$f_0^{(4)}(z)=\left(z^4-6z^2+3\right)f_0(z),$$

$$f_0^{(5)}=\left(-z^5+10z^3-15z\right)f_0(z),z=\frac{x-m_x}{\sigma_ш}.$$

9. Строим график функции $f_э(x)$.

10. Так как функция распределения $F(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp\left\{-\frac{t^2}{2}\right\}dt=\Phi(z)+\frac{1}{2}$,

то, учитывая, что $F^{(m)}(z)=f_*^{(m-1)}(z),f_*^{(0)}(z)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{z^2}{2}\right\},m$ - натуральное, получим разложение Эджворта для функции распределения: $F_э(z)=F(z)-\frac{A_ш}{6}f_*^{(2)}(z)+\frac{E_ш}{24}f_*^{(3)}(z)-\frac{R_ш-10A_ш}{120}f_*^{(4)}(z)$.

Учитывая, что $f_*^{(s)}(\pm\infty)=0,s=0,1,2,..$, находим вероятности

$p_{iэ} = F_э(z_{i+1}) - F_э(z_i)$. Необходимо отметить, что разности $F(z_{i+1}) - F(z_i)$ уже были вычислены для значений $z_i = \dfrac{x_i - m_x}{\sigma_x}$. В последнем разложении для функции распределения $F_э(z)$ нужно было бы положить $z_i = \dfrac{x_i - m_x}{\sigma_ш}$.

11. Вычисляем $\chi_э^2 = N\sum_{i=1}^{k}\dfrac{(n_i / N - p_{iэ})}{p_{iэ}}$ и сравниваем с χ_q^2. При $\chi_э^2 < \chi_q^2$ гипотезу о нормальном распределении генеральной совокупности принимаем (с учетом поправок на асимметрию, эксцесс и т.д.); в противном случае гипотеза отвергается.
12. В отчете о проделанной работе должны быть представлены: гистограмма относительных частот выборки, графики плотностей распределений $f(x)$ и $f_э(x)$, значения $N, m_x, \sigma_x, \sigma_ш, A_ш, E_ш, R_ш, \chi^2, \chi_э^2$.

В Табл.1 даны варианты заданий. Для уровня значимости 0.05 необходимо проверить гипотезу о нормальном распределении генеральной совокупности, используя критерий Пирсона, и применяя, если это необходимо, разложение Эджворта.

Табл.1.

№№	$x_{\min}$	h	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}
1	55.55	2.24	7	16	27	67	86	57	53	32	25	8
2	51.75	2.74	17	16	27	67	86	117	143	92	51	28
3	37.42	2.69	5	8	16	32	54	31	19	12	8	6
4	57.42	3.69	5	9	16	33	61	32	19	11	8	5
5	27.42	1.69	5	10	15	34	60	33	18	12	9	5
6	32.21	2.57	6	15	20	39	65	59	29	19	14	11
7	41.23	2.13	5	10	14	34	59	33	17	10	6	5
8	61.03	3.73	6	12	16	36	61	35	19	12	8	6
9	51.63	2.98	8	14	18	38	61	37	21	13	10	7
10	51.63	2.98	10	16	20	40	63	39	23	15	12	9
11	48.63	2.77	12	19	23	42	62	40	21	16	14	10
12	29.32	2.18	19	39	77	88	90	62	41	39	15	10

13	33.16	2.59	9	9	17	28	39	42	71	53	25	12
14	93.55	5.82	29	39	57	98	119	162	97	53	25	12
15	84.24	4.77	15	19	47	68	102	131	69	41	17	11
16	81.12	5.61	9	19	37	78	121	74	51	31	17	7
17	78.26	5.11	17	19	27	38	52	74	85	36	16	11
18	74.16	4.18	21	22	22	23	24	26	28	39	22	17
19	38.25	3.11	15	17	37	77	78	47	33	22	15	8
20	56.09	4.25	11	13	27	37	48	59	88	62	35	18
21	65.18	4.11	21	33	57	87	48	32	20	15	9	8
22	16.19	2.25	5	8	11	19	31	39	48	52	33	18
23	26.19	2.05	15	18	21	29	41	49	58	62	43	28
24	12.04	1.20	10	13	16	24	36	44	53	58	38	22
25	22.24	1.88	20	33	56	74	56	34	23	18	8	6

В заключение этого задания продемонстрируем на Рис.1графически выполнение одного из вариантов Табл.1.

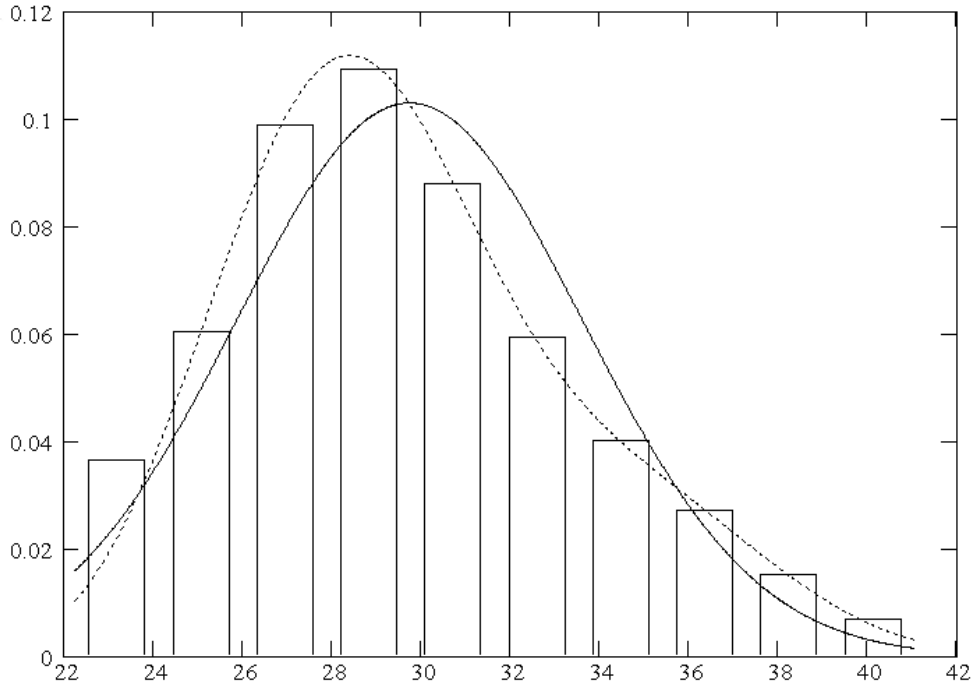

Рис.1

На Рис.1 изображены гистограмма относительных частот и графики плотностей распределений $f(x)$ и $f_э(x)$ (пунктирная кривая).

Литература

1. Ахназарова С.Л., Кафаров В.В. Методы оптимизации эксперимента в химической технологии.- М.: Высшая школа, 1985
2. Смирнов Н.В., Дунин-Барковский И.В. Курс теории вероятностей и математической статистики.- М.: Наука, 1969
3. Брандт З. Статистические методы анализа наблюдений.- М.: Мир, 1975
4. Крамер Г. Математические методы статистики.- М.: Мир, 1975
5. Уилкс С. Математическая статистика.- М.: Наука, 1967

Задание № 2. Метод наименьших квадратов.

Графическим отображением результатов какого-либо эксперимента часто является диаграмма рассеяния, т.е. геометрическое место точек, координаты которых представляют собой значения измеримых параметров опыта.

В этом задании рассматривается простейшая двухмерная диаграмма рассеяния и предполагается, что одна из координат такой диаграммы не подвержена влиянию случайных факторов.

Анализ таких диаграмм осуществляется обычно двумя способами.

Во-первых, если по каким-либо причинам известна форма зависимости условного математического ожидания с.в. Y при конкретном значении x, т.е.

$$M[Y \mid x] = \sum_{i=1}^{m} a_i f_i(x),$$

(2.1)

где $f_i(x)$ - известные линейно независимые функции, a_i - уточняемые коэффициенты регрессии (2.1), то при количестве измерений $n > m$ параметров опыта x_j, y_j не возникает никаких проблем при таком анализе.

Во-вторых, если для данной диаграммы рассеяния сложно указать систему линейно независимых функций $\{f_i(x)\}$, то приходится заниматься их возможным перебором, используя в качестве критерия отбора, например, наименьшее значение остаточной суммы квадратов $Q^* = \sum_{i=1}^{m} \left(y_i - M[Y \mid x_i]\right)^2$, где $y_i (i = 1,2,...,n)$ - экспериментальные значения с.в. Y, критерий адекватности Фишера [2.1-2.5] и т.п. Однако в этом случае необходимо соблюдать крайнюю осторожность в интерпретации результатов такого анализа [2.2,2.4].

Пусть с.в. Y принимает значения y_i и

$$y_i = M[Y \mid x_i] + \delta_i \quad ,$$

(2.2)

где $M[Y \mid x]$ имеет вид (2.1), а случайные ошибки δ_i $(i = 1,2,...,n)$ удовлетворяют следующим условиям:

- математические ожидания $M[\delta_i] = 0$,
- дисперсия $D[\delta_i] = \sigma^2$ и не зависит от x,
- δ_i независимы для различных x_i и плотность их распределения

$$p(\delta_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\delta_i^2}{2\sigma^2}\right\}.$$

Плотность распределения системы ошибок $\{\delta_i\}$ в этом случае будет иметь вид:

$$P_{\Sigma}\left(x_1, x_2, ..., x_n; a_1, a_2, ..., a_m, \sigma^2\right) = \prod_{i=1}^{n} p(\delta_i) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left\{ -\frac{\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} a_j f_j(x_i)\right)^2}{2\sigma^2} \right\}$$

(2.3)

и называется функцией максимального правдоподобия Фишера [2.5].

Можно доказать (см., например, [2.2]), что условия

$$\frac{\partial P_{\Sigma}}{\partial a_j} = \frac{\partial P_{\Sigma}}{\partial \sigma^2} = 0,\ j = 1,2,...,m$$

(2.4)

дают наилучшие оценки для коэффициентов a_j регрессии и с.к.о., причем оценка для σ^2 оказывается несколько смещенной и для устранения такого смещения будем умножать дисперсию σ_s^2, вычисленную из (2.4), на число $\frac{n}{n-m}$, т.е. полагать

$$\hat{\sigma} = \sqrt{\frac{n}{n-m}}\sigma_s \quad .$$

(2.5)

В целях компактности изложения перейдем в описании МНК к матричной форме. Введем матрицы:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ .. \\ y_n \end{pmatrix}, \Delta = \begin{pmatrix} \delta_1 \\ \delta_2 \\ .. \\ \delta_n \end{pmatrix}, A = \begin{pmatrix} a_1 \\ a_2 \\ .. \\ a_m \end{pmatrix}, F = \begin{pmatrix} f_1(x_1) & f_2(x_1) & ... & f_m(x_1) \\ f_1(x_2) & f_2(x_2) & ... & f_m(x_2) \\ ... & ... & ... & ... \\ f_1(x_n) & f_2(x_n) & ... & f_m(x_n) \end{pmatrix} ,$$

(2.6)

заметив, что размерность матрицы F равна $n \times m, n > m$ и ее ранг $R(F) = m$. Тогда $\Delta = Y - F \cdot A$ и функция P_{Σ} (2.3) будет иметь вид

$$P_{\Sigma} = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left\{ -\frac{\Delta^T \cdot \Delta}{2\sigma^2} \right\} \quad ,$$

(2.7)

где Δ^T - транспонированная к Δ матрица.

Дифференцируя (2.7) по Δ и приравнивая производную к 0, получим оценку для матрицы коэффициентов регрессии

$$\hat{A} = \left(F^T F\right)^{-1} F^T Y \quad ,$$

(2.8)

которая оказывается несмещенной $\left(M[\hat{A}] = M[A]\right)$.

В условиях нашей задачи $D[Y] = \sigma^2 E$, где E - единичная матрица размерностью $n \times n$, поэтому

$$D[\hat{A}] = \left(F^T F\right)^{-1} F^T D[Y] F \left(F^T F\right)^{-1} = \sigma^2 \left(F^T F\right)^{-1} .$$

(2.9)

Наконец, из (2.4-2.7) получим несмещенную оценку

$$\hat{\sigma}^2 = \left(Y - F\hat{A}\right)^T \left(Y - F\hat{A}\right)/(n-m) = \left(Y^T Y - Y^T F\hat{A}\right)/(n-m) \ .$$

(2.10)

Подчеркнем, что с.к.о. σ системы ошибок $\{\delta_i\}$ нам неизвестно. Можно показать, что оценки (2.8-2.10) являются эффективными [2.2,2.3].

Построим доверительные интервалы для оцениваемых параметров. Введем две с.в.

$$V = Q/\sigma^2 \qquad \text{и} \qquad U_j = \frac{\hat{a}_j - a_j}{\sqrt{D[\hat{a}_j - a_j]}} \quad ,$$

(2.11)

где $Q=\sum_{i=1}^{n}\left[y_i-\sum_{j=1}^{m}\hat{a}_j f_j(x_i)\right]^2=\left(Y-F\hat{A}\right)^T\left(Y-F\hat{A}\right)$; $\hat{a}_j, a_j, D[\hat{a}_j-a_j]=D[\hat{a}_j]$ - матричные элементы матриц $\hat{A}, A, D\left[\hat{A}\right]$.

С.в. V распределена по закону χ^2_{n-m} с $n-m$ степенями свободы (для оценок регрессионных коэффициентов уже использованы m уравнений (2.4)). Плотность такого распределения есть

$$k_{n-m}(x)=\frac{x^{(n-m-2)/2}\exp\{-x/2\}}{2^{(n-m)/2}\Gamma\left(\frac{n-m}{2}\right)}\ ,$$ где $\Gamma(p)=\int e^{-z}z^{p-1}dz$ - гамма-функция Эйлера.

Вывод формулы для плотности распределения $k_{n-m}(x)$ можно найти в работе [2.1].

Доверительная вероятность $P[\chi^2_{n-m,\alpha/2}<Q/\sigma^2<\chi^2_{n-m,1-\alpha/2}]=1-\alpha$, где α - уровень значимости, есть интеграл от $k_{n-m}(x)$ по dx в указанных пределах, который затабулирован в статистических таблицах. Следовательно, доверительный интервал для σ^2 имеет вид:

$$\frac{(n-m)\hat{\sigma}^2}{\chi^2_{n-m,1-\alpha/2}}<\sigma^2<\frac{(n-m)\hat{\sigma}^2}{\chi^2_{n-m,\alpha/2}}\ ,$$

(2.12)

где $\hat{\sigma}^2$ известно из (2.10), а квантили распределения Пирсона можно найти в статистических таблицах. Извлекая из обеих частей неравенства (2.12) квадратный корень, найдем доверительный интервал для с.к.о.:

$$\sqrt{\frac{(n-m)\hat{\sigma}^2}{\chi^2_{n-m,1-\alpha/2}}}<\sigma<\sqrt{\frac{(n-m)\hat{\sigma}^2}{\chi^2_{n-m,\alpha/2}}}\ .$$

(2.12’)

Рассмотрим с.в. $t^{(j)}=U_j\cdot\sqrt{\frac{n-m}{V}}$, распределенную по закону Стьюдента. Плотность такого распределения с $n-m$ степенями свободы имеет вид:

$$s_{n-m}(x) = \frac{\Gamma\left(\frac{n-m+1}{2}\right)}{\Gamma\left(\frac{n-m}{2}\right)\sqrt{\pi(n-m)}}\left(1+\frac{x^2}{n-m}\right)^{-(n-m+1)/2} .$$

(2.13)

Вывод формулы (2.13) можно найти в работе [2.1].

Учитывая (2.11) и требуя, чтобы $\left|t^{(j)}\right| < t_{n-m,1-\alpha/2}$, получим доверительные интервалы для коэффициентов регрессии a_j:

$$\hat{a}_j - \sqrt{\frac{Q \cdot D[\hat{a}_j]}{(n-m)\sigma^2}} \cdot t_{n-m,1-\alpha/2} < a_j < \hat{a}_j + \sqrt{\frac{Q \cdot D[\hat{a}_j]}{(n-m)\sigma^2}} \cdot t_{n-m,1-\alpha/2} .$$

(2.14)

Квантили распределения Стьюдента $t_{n-m,1-\alpha/2}$ определяются из формулы

$$1-\alpha/2 = \int_{-\infty}^{t_{n-m,1-\alpha/2}} s_{n-m}(x)dx ,$$

(2.15)

где α - уровень значимости, и печатаются в статистических таблицах.

Аналогично строится доверительная полоса Уоркинга-Хотеллинга:

$$\sum_{j=1}^{m} \hat{a}_j f_j(x) - \sqrt{\frac{Q \cdot D\left[\sum_{j=1}^{m}(\hat{a}_j - a_j) f_j(x)\right]}{(n-m)\sigma^2}} \cdot t_{n-m,1-\alpha/2} < \sum_{j=1}^{m} a_j f_j(x) <$$

$$< \sum_{j=1}^{m} \hat{a}_j f_j(x) + \sqrt{\frac{Q \cdot D\left[\sum_{j=1}^{m}(\hat{a}_j - a_j) f_j(x)\right]}{(n-m)\sigma^2}} \cdot t_{n-m,1-\alpha/2} ,$$

(2.16)

где

$$D\left[\sum_{j=1}^{m}(\hat{a}_j - a_j) f_j(x)\right] = \sum_{j=1}^{m} f_j^2(x) D[\hat{a}_j - a_j] + 2\sum_{r<}\sum_{s}^{m} f_r(x) f_s(x) M[(\hat{a}_r - a_r)(\hat{a}_s - a_s)] \ .$$

(2.17)

Схема выполнения задания 2, представленного в Табл.2, могла бы быть следующей. Из статистических таблиц при $n = 10, m = 2, \alpha = 0.1$ (двухсторонний критерий) определяем квантили распределений Пирсона и Стьюдента: $\chi^2_{8,0.05} = 2.73, \chi^2_{8,0.95} = 15.5, t_{8,0.95} = 1.86$. Далее:

1. Вычисляем матричные элементы $g_{ij} = \sum_{k=1}^{10} f_i(x_k) f_j(x_k)$ квадратной матрицы $F^T F$, $i, j = 1,2$, вычисляем $\det\left(F^T F\right)$ и строим симметрическую ковариационную матрицу $\left(F^T F\right)^{-1} = \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{12} & \Psi_{22} \end{pmatrix}$.

2. Находим матрицу коэффициентов регрессии $\hat{A} = \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = \left(F^T F\right)^{-1} F^T Y$.

3. По формуле (2.10) найдем оценку для дисперсии $\hat{\sigma}^2 = Q/8$, где $Q = \sum_{i=1}^{10}\left[y_i - \sum_{j=1}^{2}\hat{a}_j f_j(x_i)\right]^2 = \left(Y - F\hat{A}\right)^T\left(Y - F\hat{A}\right) = Y^T Y - Y^T F\hat{A}$, и построим доверительные интервалы для $\sigma^2, \sigma, a_1, a_2$ по формулам (2.12), (2.12’), (2.14), в частности:

$$\hat{a}_1 - \sqrt{\frac{Q \cdot \Psi_{11}}{8}} \cdot t_{8,0.95} < a_1 < \hat{a}_1 + \sqrt{\frac{Q \cdot \Psi_{11}}{8}} \cdot t_{8,0.95},$$

$$\hat{a}_2 - \sqrt{\frac{Q \cdot \Psi_{22}}{8}} \cdot t_{8,0.95} < a_2 < \hat{a}_2 + \sqrt{\frac{Q \cdot \Psi_{22}}{8}} \cdot t_{8,0.95}.$$

4. Определяем дисперсию отклонений оценочной регрессии $\hat{Y} = \hat{a}_1 f_1(x) + \hat{a}_2 f_2(x)$ от условного математического ожидания $M[Y \mid x]$ (2.1) по формуле (2.17) для $j = 1,2, r = 1, s = 2$ и для значений, например, $x = x_i$ находим границы доверительной полосы Уоркинга-Хотеллинга (2.16), принимая во внимание, что
$$\frac{D\left[\sum_{j=1}^{m}(\hat{a}_j - a_j) f_j(x)\right]}{\sigma^2} = f_1^2(x)\Psi_{11} + f_2^2(x)\Psi_{22} + 2 f_1(x) f_2(x)\Psi_{12}.$$

5. На масштабной бумаге изображаем в декартовой системе координат XOY диаграмму рассеяния (x_i, y_i), график функции $\hat{Y} = \hat{a}_1 f_1(x) + \hat{a}_2 f_2(x)$ и границу доверительной полосы Уоркинга-Хотеллинга.

На Рис.2 дана графическая демонстрация решения одного из вариантов задания 2, содержащегося в Табл.2. Кружками обозначена диаграмма рассеяния, регрессионная кривая $\hat{Y} = \hat{a}_1 f_1(x) + \hat{a}_2 f_2(x)$ - сплошная линия, полоса
Уоркинга-Хотеллинга – пунктирные линии.

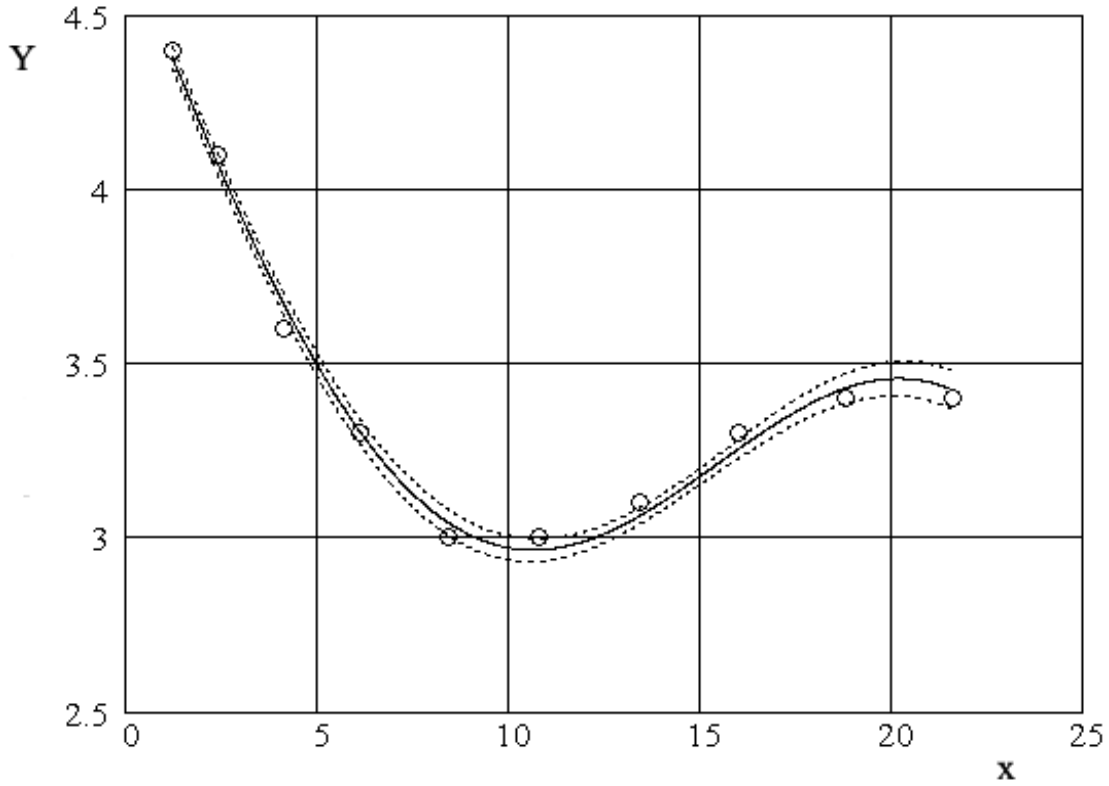

Рис.2.

Литература.

2.1. Смирнов Н.В., Дунин-Барковский И.В. Курс теории вероятностей и математической статистики.- М.: Наука, 1969

2.2. Себер Дж. Линейный регрессионный анализ.- М., Мир, 1980

2.3. Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке.-М., Мир, 1980

2.4. Худсон Д. Статистика для физиков.- М., Мир, 1967

2.5. Брандт З. Статистические методы анализа наблюдений.- М., Мир, 1975

$x_1, x_2, \ldots\ldots, x_{10} \Rightarrow$			1.2	2.4	4.1	6.2	8.4	10.8	13.4	16.0	18.8	21.6
№	$f_1(x)$	$f_2(x)$	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
1	$x^{0.13}$	$x^{-1.2}$	9.0	11.3	12.7	13.6	14.3	14.8	15.3	15.6	16.0	16.3
2	$e^{-0.11x}$	$e^{0.021x}$	-0.1	0.7	1.6	2.6	3.5	4.5	5.2	6.0	6.6	7.4
3	$\ln(1+x)$	$x^{0.3}$	10.6	14.0	17.0	20.1	22.0	24.3	25.6	27.6	28.6	30.3
4	$arctgx$	$x^{1.14}$	-1.6	-1.0	0.5	2.8	5.6	9.0	13.0	16.4	21.1	25.1
5	$\arcsin 0.01x$	$\cos 0.1x$	-1.85	-1.93	-1.69	-1.52	-1.02	-0.63	-0.06	0.55	1.16	1.79
6	$x^{-1.3}$	$x^{1.13}$	-3.3	5.0	13.5	22.0	33.0	43.0	56.0	68.0	83.3	96.0
7	$arctg0.3x$	$\sqrt{x}$	-0.41	-0.29	-0.26	-0.32	-0.53	-0.73	-1.06	-1.21	-1.56	-1.79
8	$\ln\dfrac{x+1}{x+3}$	$x^{0.32}$	-7.4	-6.0	-4.9	-4.4	-3.9	-3.7	-3.6	-3.5	-3.5	-3.5
9	$e^{-0.01x^2}$	$\sqrt{x}$	0.5	-1.4	-3.9	-6.4	-9.2	-11.4	-13.9	-15.3	-17.4	-18.8
10	$(x^2+x)^{1/4}$	$1/(1+x)$	3.7	5.9	7.9	9.9	11.2	13.5	14.3	16.3	17.2	18.8
11	$(x^2+1)^{-1}$	$\ln x$	-0.4	3.0	5.4	7.3	8.4	9.6	10.2	11.3	11.7	12.3
12	$e^{-\sqrt{x}}$	$1/x$	0.02	0.23	0.17	0.10	0.04	0.0	-0.02	-0.03	-0.04	-0.05
13	$\sin 0.03x$	$\cos 0.05x$	-4.7	-4.6	-4.0	-3.80	-3.01	-2.41	-1.59	-0.71	0.26	1.26
14	$\sqrt{x}$	$x^{-2.3}$	-0.1	-5.2	-7.8	-9.9	-11.4	-13.3	-14.6	-16.1	-17.2	-18.6
15	$\arccos 0.03x$	$\sin 0.2x$	4.4	4.1	3.6	3.4	3.0	3.0	3.1	3.2	3.5	3.4
16	$\ln(x^2+2)$	$\sqrt{x+1}$	-0.8	-2.5	-4.3	-5.8	-6.7	-7.5	-7.9	-8.4	-8.7	-9.0
17	$e^{-0.1x}$	$x^{-0.5}$	5.2	5.1	4.5	3.7	2.8	2.2	1.5	1.1	0.7	0.4
18	$xe^{-0.02x}$	1	-0.8	0.3	1.8	3.4	5.0	6.8	8.2	9.7	10.9	12.0
19	$e^{-(x-3)^2/100}$	$e^{-(x-5)^2/50}$	-3.2	-3.1	-3.0	-2.6	-2.2	-1.7	-1.1	-0.6	-0.3	-0.2
20	$x^{1.1}$	$x^{-1.1}$	4.0	6.0	10.0	15.0	21.0	27.5	35.0	42.0	51.0	59.0
21	$2^{0.1x}$	$e^{-0.4x}$	12.4	11.5	11.9	12.5	14.8	17.0	20.0	24.5	30.0	36.0
22	$x^{0.1}$	$\ln x$	9.8	12.5	14.4	16.5	17.4	18.7	19.2	20.5	21.0	21.8
23	3^{-x}	$x^{1.1}$	-0.4	-2.5	-4.6	-7.6	-10.2	-14.0	-17.0	-21.5	-24.3	-30.0
24	$\sin(1/x)$	$x^{-1.1}$	14.5	7.4	4.5	2.7	2.2	1.5	1.3	1.1	0.9	0.8
25	$x^{-1.3}$	$\ln x$	2.0	6.0	10.0	12.5	15.0	16.5	18.5	19.4	20.5	21.4

Табл.2

Учебное издание

Островский Юрий Константинович, к.ф.-м.н., доцент

Методическое обеспечение модуля по математической статистике.
Часть I. Критерий согласия Пирсона. Метод наименьших квадратов.

Учебное пособие

Компьютерная верстка Островский Ю.К.

Технический редактор Киреев Д.А.

Ответственный за выпуск Морозов Р.В.

Бумага офсетная. Печать на ризографе.

Усл. печ. л. 1.3 Тираж 50 экз. Заказ №_____

Информационно-издательский центр МГУДТ

115998, Москва, ул. Садовническая, 33
тел./факс: (095) 506-72-71
e-mail: rfrost @ yandex.ru

Отпечатано в ИИЦ МГУДТ

Табл.2

$x_1, x_2, \ldots\ldots, x_5 \Rightarrow$			1.2	2.4	4.1	6.2	8.4
№	$f_1(x)$	$f_2(x)$	y_1	y_2	y_3	y_4	y_5
1	$x^{0.13}$	$x^{-1.2}$	9.0	11.3	12.7	13.6	14.3
2	$e^{-0.11x}$	$e^{0.021x}$	-0.1	0.7	1.6	2.6	3.5
3	$\ln(1+x)$	$x^{0.3}$	10.6	14.0	17.0	20.1	22.0
4	$arctgx$	$x^{1.14}$	-1.6	-1.0	0.5	2.8	5.6
5	$\arcsin 0.01x$	$\cos 0.1x$	-1.85	-1.93	-1.69	-1.52	-1.02
6	$x^{-1.3}$	$x^{1.13}$	-3.3	5.0	13.5	22.0	33.0
7	$arctg 0.3x$	$\sqrt{x}$	-0.41	-0.29	-0.26	-0.32	-0.53
8	$\ln\frac{x+1}{x+3}$	$x^{0.32}$	-7.4	-6.0	-4.9	-4.4	-3.9
9	$e^{-0.01x^2}$	$\sqrt{x}$	0.5	-1.4	-3.9	-6.4	-9.2
10	$(x^2+x)^{1/4}$	$1/(1+x)$	3.7	5.9	7.9	9.9	11.2
11	$(x^2+1)^{-1}$	$\ln x$	-0.4	3.0	5.4	7.3	8.4
12	$e^{-\sqrt{x}}$	$1/x$	0.02	0.23	0.17	0.10	0.04
13	$\sin 0.03x$	$\cos 0.05x$	-4.7	-4.6	-4.0	-3.80	-3.01

Табл.2

$x_1, x_2, \ldots\ldots, x_5 \Rightarrow$			1.2	2.4	4.1	6.2	8.4
№	$f_1(x)$	$f_2(x)$	y_1	y_2	y_3	y_4	y_5
14	$\sqrt{x}$	$x^{-2.3}$	-0.1	-5.2	-7.8	-9.9	-11.4
15	$\arccos 0.03x$	$\sin 0.2x$	4.4	4.1	3.6	3.4	3.0
16	$\ln(x^2+2)$	$\sqrt{x+1}$	-0.8	-2.5	-4.3	-5.8	-6.7
17	$e^{-0.1x}$	$x^{-0.5}$	5.2	5.1	4.5	3.7	2.8
18	$xe^{-0.02x}$	1	-0.8	0.3	1.8	3.4	5.0
19	$e^{-(x-3)^2/100}$	$e^{-(x-5)^2/50}$	-3.2	-3.1	-3.0	-2.6	-2.2
20	$x^{1.1}$	$x^{-1.1}$	4.0	6.0	10.0	15.0	21.0
21	$2^{0.1x}$	$e^{-0.4x}$	12.4	11.5	11.9	12.5	14.8
22	$x^{0.1}$	$\ln x$	9.8	12.5	14.4	16.5	17.4
23	3^{-x}	$x^{1.1}$	-0.4	-2.5	-4.6	-7.6	-10.2
24	$\sin(1/x)$	$x^{-1.1}$	14.5	7.4	4.5	2.7	2.2
25	$x^{-1.3}$	$\ln x$	2.0	6.0	10.0	12.5	15.0

№№	x_6	x_7	x_8	x_9	x_{10}
	10.8	13.4	16.0	18.8	21.6
	y_6	y_7	y_8	y_9	y_{10}
1	14.8	15.3	15.6	16.0	16.3
2	4.5	5.2	6.0	6.6	7.4
3	24.3	25.6	27.6	28.6	30.3
1	9.0	13.0	16.4	21.1	25.1
5	-0.63	-0.06	0.55	1.16	1.79
6	43.0	56.0	68.0	83.3	96.0
7	-0.73	-1.06	-1.21	-1.56	-1.79
8	-3.7	-3.6	-3.5	-3.5	-3.5
9	-11.4	-13.9	-15.3	-17.4	-18.8
10	13.5	14.3	16.3	17.2	18.8
11	9.6	10.2	11.3	11.7	12.3
12	0.0	-0.02	-0.03	-0.04	-0.05
13	-2.41	-1.59	-0.71	0.26	1.26
14	-13.3	-14.6	-16.1	-17.2	-18.6
15	3.0	3.1	3.2	3.5	3.4
16	-7.5	-7.9	-8.4	-8.7	-9.0
17	2.2	1.5	1.1	0.7	0.4
18	6.8	8.2	9.7	10.9	12.0
19	-1.7	-1.1	-0.6	-0.3	-0.2
20	27.5	35.0	42.0	51.0	59.0
21	17.0	20.0	24.5	30.0	36.0
22	18.7	19.2	20.5	21.0	21.8
23	-14.0	-17.0	-21.5	-24.3	-30.0
24	1.5	1.3	1.1	0.9	0.8
25	16.5	18.5	19.4	20.5	21.4