ИЗУЧЕНИЕ МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ СТУДЕНТАМИ ФИЗИЧЕСКИХ СПЕЦИАЛЬНОСТЕЙ

Гончаренко Янина Владимировна, Парчук Мария Ивановна

ABSTRACT: *The article analyzes the methodological features of studying the method of least squares in the course of probability theory and mathematical statistics for students of physical specialties.*

KEYWORDS: *teaching methodology, the method of least squares, probability theory and mathematical statistics, students of physical specialties.*

В Физико-математическом институте Национального педагогического университета имени М.П. Драгоманова (Украина, Киев) осуществляется подготовка бакалавров специальности «Физика». Система подготовки учителя физики включает в себя циклы дисциплин гуманитарной и социально-экономической подготовки; фундаментальной и естественно-математической подготовки; профессиональной и практической подготовки. Они образуют единую систему и связаны между собой межпредметными и междисциплинарными связями. Полноценная профессиональная подготовка будущих учителей физики невозможна без изучения математических дисциплин, которые входят как во второй цикл подготовки, так и в третий, а также являются фундаментом или составляющими изучения других профессионально направленных дисциплин. Одной из дисциплин, которая обеспечивает формирование основных компетенций будущих учителей физики, является «Теория вероятностей и математическая статистика».

Курс «Теория вероятностей и математическая статистика» входит в цикл дисциплин фундаментальной и естественно-математической подготовки специалистов образовательно-квалификационного уровня «Бакалавр» направления подготовки 6.040203 Физика. На изучение отводится 126 часов (3,5 кредита ECTS).

Во время курса изучаются такие содержательные модули:

1. Теория вероятностей.
   - 1.1. Случайные события. Вероятность случайных событий.
   - 1.2. Случайные величины.
2. Элементы математической статистики.
   - 2.1. Первичная обработка эмпирических данных. Статистические оценки параметров распределения.
   - 2.2. Статистическая проверка статистических гипотез.
   - 2.3. Элементы регрессионного и корреляционного анализа.
   - 2.4. Элементы дисперсионного анализа. Элементы теории случайных процессов и теории массового обслуживания.

Во время изучения содержательного подмодуля «Элементы регрессионного и корреляционного анализа» рассматривается использование метода наименьших квадратов для нахождения оценок параметров функциональной зависимости между переменными, значения которых определяются из опыта. Вид искомой функциональной зависимости предполагается известным. Заметим, что математическую основу метода наименьших квадратов студенты рассматривают в курсе математического анализа при изучении функций многих переменных, однако при этом не рассматриваются вопросы, возникающие при практическом применении данного метода в условиях, когда исходные и сама модель данных носят стохастический характер. Применять метод наименьших квадратов «в полною силу» становится возможным только после изучения теории вероятностей.

В статье мы рассматриваем особенности изучения метода наименьших квадратов в курсе теории вероятностей и математической статистики студентами физических специальностей. Интерес к данной теме был вызван следующими методологическими проблемами: данный метод является одним из классических методов построения аппроксимирующих функций, однако, поскольку его математическая основа и область применимости рассматриваются в разных учебных дисциплинах, у студентов часто возникает определенный «разрыв» в знаниях, отсутствует целостное представление о методе и спектре его возможных применений; современные учебные пособия изобилуют прикладными задачами экономического содержания на применение данного метода, в то время как он не менее эффективен и при обработке результатов физических данных.

При изучении содержательного модуля «Элементы регрессионного и корреляционного анализа» в курсе «Теория вероятностей и математическая статистика» студентами физических специальностей мы предлагаем следующий объем теоретической информации.

Если в результате опыта получено $n+1$ пар значений $(x_i; y_i)$, где x_i - значения аргумента, а y_i - значение функции, то параметры аппроксимирующей функции $f(x)$ выбирают так, чтобы обратилась в минимум сумма

$S = \sum_{i=0}^{n}[y_i - f(x_i)]^2$.

Если в качестве аппроксимирующей функции взят многочлен, т. е.

$Q(x) = Q_m(x) = a_0 + a_1 x + \cdots + a_m x^m$, $(m \leq n)$,

то оценки его коэффициентов $\widetilde{a_k}$ определяются из системы $m+1$ нормальных уравнений

$\sum_{j=0}^{m} s_{k+j} a_j \equiv s_k a_0 + s_{k+1} a_1 + \cdots + s_{k+m} a_m = v_k$, $(k = 0, 1, 2, \ldots, m)$,

где $s_k = \sum_{i=0}^{n} x_i^k \quad (k = 0, 1, 2, \ldots, 2m)$, $v_l = \sum_{i=0}^{n} y_i x_i^l \quad (l = 0, 1, 2, \ldots, m)$.

Если значения x_i известны без ошибок, а значения y_i независимы и равноточны, то оценка дисперсии $\widetilde{\sigma^2}$ величины y_i определяется формулой

$$\tilde{\sigma}^2 = \frac{1}{n-m} S_{min};\ S_{min} = \sum_{i=0}^{n} \varepsilon_i^2\ ;\ \varepsilon_i = y_i - \tilde{f}(x_i),$$

где S_{min} – значение S, вычисленное в предложении, что коэффициенты полинома $Q(x) = Q_m(x)$ заменены их оценками, найденными из системы нормальных уравнений.

При нормальном законе распределения величин y_i изложенный метод является наилучшим способом нахождения аппроксимирующей функции $f(x)$ [6].

Несложно заметить, что квадратическая функция S непрерывная, выпуклая и ограниченная снизу ($S \geq 0$), а значит имеет минимум.

Необходимым условием существования минимума непрерывно дифференцируемой функции двух переменных является равенство нулю ее частных производных.

В случае линейной зависимости ($m = 1$) имеем: $y = a_0 + a_1 x$ и

$$\begin{cases} \frac{\partial S}{\partial \hat{a}_0} = -2\sum(y_i - \hat{a}_0 - \hat{a}_1 x_i) = 0 \\ \frac{\partial S}{\partial \hat{a}_1} = -2\sum(y_i - \hat{a}_0 - \hat{a}_1 x_i)x_i = 0 \end{cases},$$

выполнив преобразования системы, получим:

$\begin{cases} \widehat{a_1} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}; \\ \widehat{a_0} = \bar{y} - \widehat{a_1}\bar{x}. \end{cases}$ [2].

В случае неравноточных измерений, когда величины y_i имеют различные дисперсии σ_i^2, все предыдущие формулы остаются в силе, если величины S, s_k, v_i заменить соответственно на

$$S' = \sum_{i=0}^{n} p_i^2 (y_i - a_0 - a_i x_i - \cdots - a_m x_i^m)^2,$$

$$s'_k = \sum_{i=0}^{n} p_i^2 x_i^k \quad (k = 0, 1, 2, \ldots, 2m),$$

$$v'_l = \sum_{i-0}^{n} p_i^2 y_i x_i^l \quad (l = 0, 1, 2, \ldots, m),$$

где «веса» p_i^2 величин y_i равны

$$p_i^2 = \frac{A^2}{\sigma_i^2};$$

A^2 - произвольный коэффициент пропорциональности.

Если y является функцией нескольких аргументов z_k, то роль величин z_k могут играть любые линейно независимые функции $f_k(x)$ некоторого аргумента x. Например, если функция y, заданная в интервале $(0, 2\pi)$, аппроксимируется тригонометрическим многочленом

$$y = \lambda_0 + \sum_{k=1}^{m}{}' (\lambda_k \cos kx + \mu_k \sin kx),$$

то при равноотстоящих значениях аргумента x_i оценки коэффициентов λ_k и μ_k определяются формулами Бесселя:

$$\tilde{\lambda}_0 = \frac{1}{n}\sum_{i=0}^{n-1} y_i; \; \tilde{\lambda}_k = \frac{2}{n}\sum_{i=0}^{n-1} y_i \cos kx_i;$$

$$\tilde{\mu}_k = \frac{2}{n}\sum_{i=0}^{n-1} y_i \sin kx_i \quad (k = 1, 2, \ldots, m).$$

Для применимости формул Бесселя необходимо, чтобы значения y_i подчинялись закону нормального распределения с одинаковой дисперсией σ_2.

При сложной функциональной зависимости и достаточно малой области изменения аргументов z_k вычисления упрощаются, если разложить функцию y в ряд по степеням отклонений аргументов от их приближенного значения (например, от их среднего).

Приведем примеры задач на использование метода наименьших квадратов, которые можно рекомендовать студентам физических специальностей.

Задача 1. Исследовалась зависимость между барометрическим давлением и точкой кипения воды. Цель исследования – оценить высоту над уровнем моря по температуре кипения воды. В горных условиях для определения барометрического давления удобнее

измерять температуру кипения воды. В таблице 1 приведены данные об измерениях барометрического давления и точки кипения воды в 17 экспериментах.

Номер	Давления	Точка кипения	Номер	Давления	Точка кипения
1	20,79	194,5	10	24,01	201,3
2	20,79	194,3	11	25,14	203,6
3	22,4	197,9	12	26,57	204,6
4	22,67	198,4	13	28,49	209,5
5	23,15	199,4	14	27,76	208,6
6	23,35	199,9	15	29,04	210,7
7	23,89	200,9	16	29,88	211,9
8	23,99	201,1	17	30,66	212,2
9	24,02	201,4			

Таблица 1

Предполагая, что зависимость барометрического давления от точки кипения воды линейная: $y = b_0 + b_1 x$, найти оценки максимального правдоподобия параметров b_0, b_1 и дисперсии σ^2. Построить доверительные интервалы для b_0, b_1, σ^2. Проверить гипотезу о значимости регрессии.

Решение. Построим точечный график зависимости эмпирических значений y от x.

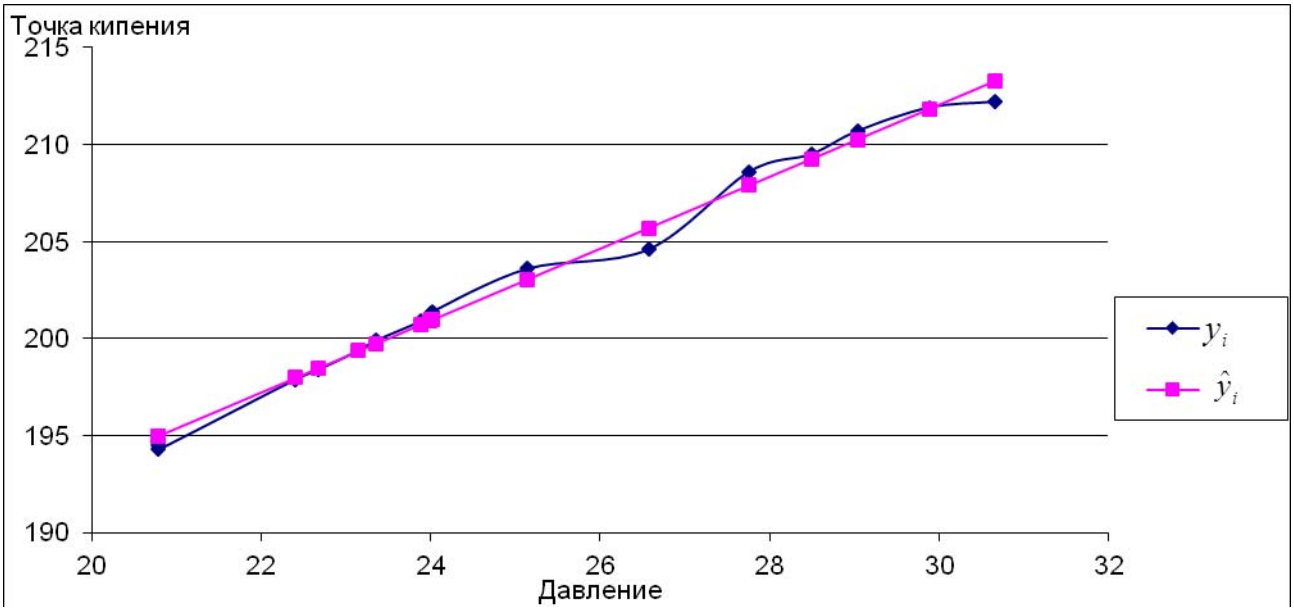

Рис. 1.

Форма графика позволяет сделать предположение о линейной форме зависимости: $y = b_0 + b_1 x$.

Вычислим выборочные средние значения, дисперсии и средние квадратические отклонения давления и точки кипения:

1) средние значения:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{426{,}6}{17} = 25{,}09\,;\ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{3450{,}2}{17} = 202{,}95\,.$$

2) дисперсии:

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{152,28}{17} = 8,98;$$

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} = \frac{530,78}{17} = 31,22;$$

3) середние квадратические отклонения:

$\sigma_x = \sqrt{8,98} = 3$; $\sigma_y = \sqrt{31,22} = 5,59$.

1. Вычислим коэффициент кореляции:

$$r = \sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y} = \frac{283,09}{17 \cdot 3 \cdot 5,59} = \frac{283,09}{285,09} = 0,99.$$

Поскольку значения коэффициента корреляции близко к 1, то между исследуемыми величинами существует прямая существенная связь.

2. Оценим параметры b_0 **и** b_1 линейного уравнения регрессии $y = b_0 + b_1 x$ с помощью метода наименьших квадратов.

$$b_1 = \sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x^2} = \frac{283,09}{17 \cdot 8,98} = 1,85;$$

$$b_0 = \bar{y} - b_1\bar{x} = 202,95 - 1,85 \cdot 25,09 = 156,53.$$

Уравнение регрессии выглядит так:

$\hat{y} = 156,53 + 1,85x$.

Анализируя параметры уравнения регрессии, можно сделать вывод: при увеличении факторного показателя (давления) на 1 (1 Па) результативный показатель (точка кипения) увеличится на 1,85 (1,85 □C); если значения факторного показателя будут равны 0, тогда значения результативного будут равны 156,53.

3. Проверим адекватность построенной модели.

Вычислим дисперсию погрешности:

$$\sigma_\varepsilon^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} = \frac{4,52}{17} = 0,27.$$

Дисперсия погрешности приближается к нулю. Это позволяет выдвинуть гипотезу о адекватности построенной модели.

Применим критерий Фишера, то есть проверим нулевую гипотезу $H_0 : \beta_1 = 0 (\hat{y}_i = \bar{y})$ при альтернативной $H_1 : \beta_1 \neq 0 (\hat{y}_i \neq \bar{y})$, где β_1 - наклон обобщенной регрессионной модели.

$\hat{y}_i$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
194,99	63,3845	0,2415
194,99	63,3845	0,4781
197,97	24,8297	0,0049
198,47	20,1012	0,0048
199,36	12,9272	0,0018
199,73	10,4034	0,0297
200,73	4,9570	0,0301

200,91	4,1674	0,0355
200,97	3,9439	0,1874
200,95	4,0177	0,1235
203,04	0,0074	0,3147
205,69	7,4614	1,1761
209,24	39,4831	0,0694
207,89	24,3350	0,5097
210,25	53,3054	0,1989
211,81	78,4120	0,0084
213,25	106,05	1,1046
Сума	521,1715	4,5197

Таблица 2.

Вычислим наблюдаемое значение ***F-критерия***

$$F_{спост.} = \frac{(n-2)\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \frac{15 \cdot 521{,}1715}{4{,}52} = 1729{,}551.$$

Согласно таблицам F-критерия и уровнем значимости $\alpha = 0{,}05$ (надежностью 0,95) и числами степеней свободы 1 и 15 находим $F_{1,15}(0{,}05) = 0{,}83$.

Поскольку $F_{ñïîñò.} \rangle F_{1,15}(0{,}05)$, то построенная нами регрессионная модель может считаться адекватной с вероятностью 0,95.

4. Проверим значимость параметров b_0 и b_1.

Вычислим среднюю квадратическую погрешность уравнения регрессии:

$$\hat{\sigma}_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \frac{4{,}5197}{15} = 0{,}3$$

и оценки дисперсий параметров b_0 и b_1:

$$\hat{\sigma}_{b_0} = \hat{\sigma}_\varepsilon \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 10{,}445;\ \hat{\sigma}_{b_1} = \frac{\hat{\sigma}_\varepsilon}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 0{,}02431\ .$$

Проверим нулевую гипотезу $H_0 : \beta i = 0$ при альтернативной $H_1 : \beta i \neq 0$ согласно критерию Стьюдента.

Определим наблюдаемые значения критерия Стьюдента для параметров регрессии:

$$t_{ñïîñò.}(b_0) = \frac{|b_0|}{\hat{\sigma}_{b_0}} = \frac{156{,}53}{10{,}445} = 14{,}986;\ t_{ñïîñò.}(b_1) = \frac{|b_1|}{\hat{\sigma}_{b_1}} = \frac{1{,}85}{0{,}02431} = 76{,}10.$$

Критическое значение $t_{15}(0{,}05) = 0{,}1315$.

Поскольку наблюдаемые значения больше, чем $t_{15}(0,05)$, то нулевая гипотеза ($H_0 : \beta i = 0$) для каждого параметра отбрасывается, а значит, вычисленные значения b_0 и b_1 являются статистически значимыми.

Построим доверительные интервалы для параметров уравнения регрессии β_0 и β_1:

$$\beta_1 = b_1 \pm t_{\frac{\alpha}{2}} \hat{\sigma}_{b_1} = 1{,}85 \pm 0{,}1315 \cdot 0{,}02431 = 1{,}85 \pm 0{,}0032;$$

$$\beta_0 = b_0 \pm t_{\frac{\alpha}{2}} \hat{\sigma}_{b_0} = 156{,}53 \pm 0{,}1315 \cdot 10{,}445 = 156{,}53 \pm 1{,}3735,$$

что означает:

$$P\{1{,}8468 < \beta_1 < 1{,}8532\} = 0{,}95; \; P\{155{,}1565 < \beta_0 < 157{,}9035\} = 0{,}95.$$

Задача 2. Высота h падения тела за время t определяется формулой

$$h = a_0 + a_1 t + a_2 t^2,$$

где a_0 – путь, пройденный телом к моменту начала отсчета времени, a_1 - скорость тела в момент начала отсчета времени, a_2 - половина ускорения силы тяжести g.

Определить оценки коэффициентов a_0, a_1, a_2 и оценить точность определения ускорения силы тяжести указанным методом на основании серии равноточных измерений, результаты которых приведены в таблице 3.

t, сек	h, см	t, сек	h, см	t, сек	h, см	t, сек	h, см	t, сек	h, см
$\frac{1}{30}$	11,86	$\frac{4}{30}$	26,69	$\frac{7}{30}$	51,13	$\frac{10}{30}$	85,44	$\frac{13}{30}$	129,54
$\frac{2}{30}$	15,67	$\frac{5}{30}$	33,71	$\frac{8}{30}$	61,49	$\frac{11}{30}$	99,08	$\frac{14}{30}$	146,48
$\frac{3}{30}$	20,60	$\frac{6}{30}$	41,93	$\frac{9}{30}$	72,90	$\frac{12}{30}$	113,77		

Таблица 3.

Задача 3. Конденсатор заряжен до напряжения U_0, отвечающего моменту начала отсчета времени, после чего он разряжается через некоторое сопротивление. Напряжение U измеряется с округлением до 5 В в различные моменты времени. Результаты измерений приведены в таблице 4.

i	t_i, *сек*	U_i, *В*	i	t_i, *сек*	U_i, *В*	i	t_i, сек	U_i, В
0	0	100	4	4	30	8	8	10
1	1	75	5	5	20	9	9	5
2	2	55	6	6	15	10	10	5
3	3	40	7	7	10			

Таблица 4.

Известно, что зависимость U от t имеет вид

$$U = U_0 e^{-at}.$$

Выбрать коэффициенты U_0 и a, составить доверительные интервалы для U_0 и a при доверительной вероятности $\alpha = 0{,}90$.

Задача 4. Величины сжатия x_i стального бруса под действием нагрузки y_i, а также значения дисперсий σ_i^2, характеризующих точность измерения y_i, приведены в таблице 5.

i	0	1	2	3	4

x_i, мк	5	10	20	40	60
y_i, кг	51,33	78	144,3	263,6	375,2
σ_i^2	82,3	25	49,3	51,3	46,7

Таблица 5.

Найти линейную зависимость $y = a_0 + a_1 x$, отвечающую закону Гука; пстроить доверительные интервалы для коэффициентов a_k $(k = 0, 1)$, а также доверительные границы для неизвестного истинного значения нагрузки при x от 5 *мк* при доверительной вероятности $\alpha = 0{,}9$.

«Веса» измерений, отвечающих каждому значению сжатия x_i, принять обратно пропорциональными величинам σ_i^2.

В заключение можем сделать следующие выводы:

- метод наименьших квадратов является эффективным методом обработки результатов физических экспериментов;
- изучение его основ в курсе математического анализа является недостаточным для дальнейшего применения при решении практических и профессионально ориентированных задач, поскольку не включает представлений о стохастической природе модели и пределах применимости метода;
- в курсе теории вероятностей и математической статистике у студентов физических специальностей необходимо сформировать представление о возможности применения данного метода в зависимости от природы исходных данных, умение находить наилучшие в определенном смысле аппроксимирующие функции разных типов, оценивать погрешность параметров, полученных методом наименьших квадратов, строить доверительные интервалы с заданным уровнем надежности;
- изучение данного метода студентами физиками открывает широкие возможности реализации прикладной и профессиональной направленности обучения математике, реализации межпредметных связей, в частности путем решения задач на обработку результатов физических экспериментов.

Список использованной литературы

1. Гончаренко Я.В. Економетрія. Методичні вказівки для виконання розрахункової робот из економетрії. – К.:НПУ імені М.П. Драгоманова, 2005. – 90 с.
2. Лещинський О.Л. Економетрія: Навч. посіб. для студ. вищ. навч. закл./ О.Л. Лещинський, В.В. Рязанцева, О.О. Юнькова. – К.: МАУП, 2003. – 208 с.: іл.
3. Лук'яненко І.Г., Краснікова Л.І Економетрика: Підручник. – К.: Товариство «Знання», КОО, 1998. – 494 с. (з табл., граф.)
4. Освітньо-кваліфікаційна характеристика бакалавра зі спеціальності 6.010100 Педагогіка і методика середньої освіти. Фізика напряму підготовки 0101 Педагогічна освіта. – К., 2003. – 63 с.
5. Освітньо-професійна програма підготовки бакалавра зі спеціальності 6.010100 Педагогіка і методика середньої освіти. Фізика напряму підготовки 0101 Педагогічна освіта. – К., 2003. –73 с.
6. Сборник задач по теории вероятностей, математической статистике и теории случайных функций, под. ред. А.А. Свешникова. М.: «Наука», - 1970. – 656 с.
7. Турчин В.Н. Теория вероятностей и математическая статистика. Основные понятия, примеры и задачи. Учебник для студентов высших учебных заведений. – Днепропетровск, ЧМА – ПРЕСС. – 2012. – 576 с.