УДК 37.013 (045)

МЕТОДИКА ПОСТРОЕНИЯ КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫХ МОДЕЛЕЙ КОЛЕБАНИЙ УРОЖАЙНОСТИ

METHODOLOGY FOR CONSTRUCTION OF CORRELATION-REGGRESSION MODELS OF VOLUME VIBRATION

И. Ю. Каневская
Саратовский государственный аграрный университет
имени Н.И. Вавилова, г. Саратов

I. Yu. Kanevskaya
Saratov State Agrarian University
Named after N.I. Vavilov, city of Saratov

Аннотация:
Методика построения корреляционно-регрессионных моделей колебаний урожайности. Цель установить связь между двумя случайными признаками или факторами (корреляционный анализ) и зависимость между исследуемыми параметрами (регрессионный анализ).

Ключевые слова:
теории вероятностей, математическая статистика, корреляционный анализ, регрессионный анализ.

Annotation:
Method of constructing correlation-regression models of crop yields. The goal is to establish a relationship between two random sign-factors or factors (correlation analysis) and the relationship between the parameters studied (regression analysis).

Keywords:
Probability theory, mathematical statistics, correlation analysis, regression analysis.

Задача статистика установить связь между явлениями, выявить причинно-следственные отношения между явлениями, выявить факторы, которые влияют на эти явления, на их вариацию. В статистике используют различные методы и приемы. При построении моделей урожайности в качестве совокупностей выступают отклонения фактических уровней урожайности от тренда. После получения информации об урожайности и факторов, влияющих на нее, ее систематизируют, т.е. составляют упорядоченную систему статистических показателей по урожайности. Несколько показателей урожайности: потенциальная урожайность, плановая урожайность, ожидаемая урожайность (виды на урожай), урожайность на корню (биологическая урожайность). Но их однородность зависит от того,

насколько точно и правильно подобран тренд урожайности и все ее изменения на данном отрезке времени. Качественная однородность отклонений урожайности от тренда обуславливается рядом причин, вызывающих эти отклонения [1].

Главной причиной колебаний урожайности являются колебания метеорологических факторов, в воздействия которых на урожайность не предвидится качественного изменения. Но есть и другие условия влияющие на уровень урожайности. Это географические, агротехнические, микробиологические, почвенные, биологические, экономические и другие.

Отбор факторов, включаемых в модель колебаний урожайности, осуществляется на основе качественного анализа колеблемой урожайности и ее связи с факторными признаками в конкретном районе, области или регионе по определенной культуре и непосредственной количественной оценке тесноты связи этих признаков с результативным.

Природных факторов очень много, которые влияют на колеблемость урожайности. Всех их перечислить и включить невозможно. Обычно включают в модель не более 5-6 переменных. Главной задачей моделирования колеблемости урожайности является выбор небольшого числа основных факторов, в достаточной мере объясняющих колеблемость урожайности. Задача корреляционного анализа установить – есть ли связь между нами выбранными признаками? Если есть, то определить какая? Прямая или обратная. Определить характер силы признака в воздействии на урожайность: тесная, средняя или слабая. Затем установить, как влияют одни признаки на другие и записать в виде уравнения регрессии, который описывает корреляционную зависимость между результативным признаком и факторами. Для отбора факторов, проверки реальности и существенности связи выбранного фактора с колебаниями урожайности используются методы аналитической группировки и дисперсионного анализа. Малое число наблюдений по отдельным областям не всегда позволяет доказать существенность влияния фактора на результативный признак на ряду с очевидностью последнего. Для этого надо взять данные по разным регионам за несколько лет, т.е. совокупность «область - лет» по метеорологическим и агрономическим статистическим данным.

Решение задачи опирается на разные приемы статистики [2]. Методику построения корреляционно-регрессивных моделей рассмотрим на примере урожайности кукурузы в Саратовской области. Саратовская область — это юго-восток Восточно-Европейской равнины, в Нижнем Поволжье. Площадью - 101200 км^2. Климат в области резко-континентальный.

В целом сельское хозяйство области имеет зерново-скотоводческую специализацию. В ТОП-20 регионов по валовым сборам кукурузы по состоянию на 02 ноября 2016 года также вошла Саратовская область (192,7 тыс. тонн, 2,0%) – 12 место. Кукурузу выращивают в основном: Краснокутском, Пугачёвском и Энгельсском районе и других.

Проведем статистическую обработку результатов хозяйственной деятельности по кукурузе шестидесяти хозяйств Саратовской области, по двум признакам X и У

Таблица 1

X – затраты на производство кукурузы (в тыс. руб.) У – произведено кукурузы (в т. ц.)			
X	Y	X	Y
20	52	70	41
51	39	28	29
38	74	66	34
35	57	52	66
80	12	39	58
12	81	12	56
48	41	50	21
62	59	66	21
74	52	72	17
22	22	28	53
90	76	92	49
47	50	21	66
11	67	51	51
77	60	61	64
65	15	69	67
42	50	43	34
40	72	14	26
70	42	50	62
62	20	67	16
33	51	90	56
49	53	30	55
49	68	54	82
63	42	84	48
24	61	27	25
53	32	86	19
46	61	51	44
87	46	53	39
63	58	60	51
58	31	44	63
78	22	71	36

X – затраты на производство кукурузы (в тыс. руб.),
У – произведено кукурузы (в т. ц.).

Проведем расчеты по X. Найдем минимальное и максимальное значения выборки: $x_{min} = 11$ $x_{max} = 92$

Вычислим размах варьирования $R_x = 92 - 11 = 81$. Найдем длину частичного интервала $h_x = 81 / 6{,}9 = 11{,}74 \approx 12$

Получим следующее разбиение выборки на интервалы:

11 – 23 – 35 – 47 – 59 – 71 – 83 – 95

Таблица 2

Интервалы	Середина интервала xi	Разноска	Частота mi
11 – 23	17		7
23 – 35	29		6
35 – 47	41		8
47 – 59	53		14
59 – 71	65		13
71 – 83	77		6
83 – 95	89		6
			60

Итак, получили дискретный вариационный ряд:

Таблица 3

xi	17	29	41	53	65	77	89
mi	7	6	8	14	13	6	6

Построим многоугольник распределения (полигон).

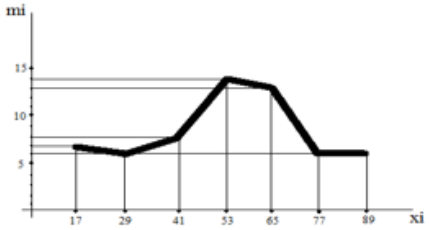

Рис.1

Вычислим выборочную среднюю, для этого составим следующую рабочую таблицу:

Таблица 4

xi	17	29	41	53	65	77	89	
mx	7	6	8	14	13	6	6	60
ximi	119	174	328	742	845	462	534	3204

$x = 3204 / 60 = 53{,}4$

$x_{низ} = (119 + 174 + 328 + 742) / (7 + 6 + 8 + 14) = 1363 / 35 = 38{,}9$

$x_{высп} = (845 + 462 + 534) / (13 + 6 + 6) = 1841 / 25 = 73{,}6$

Модой Мо называют варианту, которая имеет наибольшую частоту.

Мо(x) = 53

Медианой Ме называют варианту, которая делит вариационный ряд на две части, равные по числу вариант.

Me(x) = 53

Таблица 5

xi	Mi	xi – x	(xi – x)mi	(xi – x)²mi
17	7	-36,4	-254,8	9274,72
29	6	-24,4	-146,4	3572,16
41	8	-12,4	-99,2	1230,08
53	14	-0,4	-5,6	2,24
65	13	11,6	150,8	1749,28
77	6	23,6	141,6	3341,76
89	6	35,6	213,6	7604,16
			-506	
	60		+506	26774,4

Вычислим выборочное среднее квадратическое отклонение:

$Sx = \sqrt{26774,4} / 59 = \sqrt{453,8} = 21,3$

Коэффициент вариации V – это отношение выборочного среднего квадртического отклонения к выборочной средней, выраженное в процентах

$Vx = (Sx / x_{высш}) * 100\% = (21,3 / 73,6) * 100\% = 28,94\%$

Схема расчета $x^2_{фак}$:

1. Построить вариационный ряд для наблюдаемого признака и вычислить его параметры: выборочную среднюю x и среднее квадратическое отклонение Sx .

2. В предположении нормального распределения, вычислить теоретические частоты miT по формуле $miT = (n*hx) / Sx * \varphi(ui)$ где n – объем выборки, hx – шаг (разность между двумя соседними вариантами), $\varphi(ui)$ - дифференциальная функция Лапласа (приложение 2) , а аргумент функции ui вычисляется по формуле $ui = (xi – x) / Sx$.

3. Вычисляют эмпирическое значение $x^2_{фак.}$

4. По таблице (приложение 3) критических точек распределения x^2, по заданному уровню значимости α и числу степеней свободы k=ν- 3 (ν - число групп выборки) находят $x^2_{кр}(\alpha;k)$ правосторонней критической области.

5. Если $x^2_{фак.} \leq x^2_{кр.}$ - нет оснований отвергнуть выдвинутую гипотезу о нормальном распределении генеральной совокупности, т.е. эмпирические и теоретические частоты различаются незначимо (случайно). В противном случае, если $x^2_{фак.} > x^2_{кр.}$, гипотезу отвергают, и различия между эмпирическими и теоретическими частотами считают значимыми.

$(n * hx) / Sx = (60 * 12) / 21,3 = 33,8$

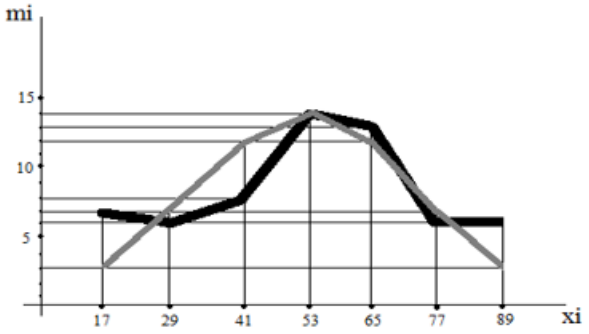

Таблица 6

xi	Mi	ui	φ (ui)	i T	mi – mi T	(mi – mi T)²	(mi – mi T)² / mi T
17	7	-1,70	0,0940	3	4	16	5,3
29	6	-1,14	0,2083	7	-1	1	0,14
41	8	-0,58	0,3372	2	-4	16	1,33
53	14	-0,01	0,3989	4	0	0	0
65	13	0,54	0,3448	2	1	1	0,08
77	6	1,10	0,2179	7	-1	1	0,14
89	6	1,67	0,0989	3	3	9	3
	60			58			9,99

$x^2_{кр.}(\alpha;k) = (0,05;3) = 7,8$ $x^2_{фак.} > x^2_{кр.}$ 9,99 > 7,8

Схема расчета ω^2: построить вариационный ряд для наблюдаемого признака и вычислить его параметры x и S_x, в предположении нормального распределения, вычислить теоретические частоты, вычислить частности по формуле Wi = mi / n. Для каждой группы ряда вычислить накопленную частость.

Вычислить эмпирическое значение $\omega^2_{фак.}$, сравнить вычисленное значение $\omega^2_{фак.}$ с критическим значением $\omega^2_{кр.}$ Если $\omega^2_{фак.} \leq \omega^2_{кр.}$ - нет оснований отвергнуть выдвинутую гипотезу о нормальном распределении генеральной совокупности. В противном случае, если $\omega^2_{фак.} > \omega^2_{кр.}$, гипотезу отвергают, и различие между эмпирическими и теоретическими частотами считают значимым.

Таблица 7

xi	mi	Wi	Fi	miT	WiT	FiT	$(Fi - FiT)^2$
17	7	0,12	0,12	3	0,05	0,05	0,0049
29	6	0,1	0,22	7	0,12	0,17	0,0025
41	8	0,13	0,35	12	0,21	0,38	0,0009
53	14	0,23	0,58	14	0,24	0,62	0,0016
65	13	0,22	0,80	12	0,21	0,83	0,0009
77	6	0,1	0,90	7	0,12	0,95	0,0025
89	6	0,1	1	3	0,05	1	0
	60	1		58	1		0,0133

$\omega^2_{фак.} = 0,0133$ $\omega^2_{кр.} = 0,461 / 60 = 0,0077$ $\omega^2_{фак.} > \omega^2_{кр.}$

Генеральной средней называют среднее арифметическое значений признака генеральной совокупности.

Доверительным называют интервал, который с заданной надежностью покрывает заданный параметр

$x - (Sx / \sqrt{n}) * t < a_x < x + (Sx / \sqrt{n}) * t$

$53,4 - (21,3 / \sqrt{60}) * 2,001 \leq a_x < 53,4 + (21,3 / \sqrt{60}) * 2,001$

$53,4 - 27,11 \leq a_x < 53,4 + 27.11$

$26,29 \leq a_x < 80,51$

Генеральной дисперсией называется среднее арифметическое квадратов отклонений значений признака генеральной совокупности от их среднего значения. Генеральным средним квадратическим отклонением называют квадратный корень из генеральной дисперсии

$Sx / (1 + q) < Q_x < Sx / (1 - q)$

$21,3 / (1 + 0,188) \leq Q_x < 21.3 / (1 - 0,188)$

$17,93 \leq Q_x < 26,23$

Аналогично проведем расчеты по Y, получим все результаты статистической обработки

Таблица 8

№	Параметр	Обозначение	Значения параметров	
			X	Y
1	Объем выборки	n	60	60
2	Размах варьирования	Rx , Ry	81	70
3	Выборочное среднее	$\bar{x}$, $\bar{y}$	12	10
4	Частные средние:			
	- низшая	$\bar{x}_{низ.}$, $\bar{y}_{низ.}$	38,9	31,4
	- высшая	$\bar{x}_{высш.}$, $\bar{y}_{высш.}$	73,6	62,2
5	Мода	Mo(x) , Mo(y)	53	55
6	Медиана	Me(x) , Me(y)	53	55
7	Среднее квадратическое отклонение (стандарта)	Sx , Sy	21,3	18,3
8	Соответствие нормальному распределению по критерию			
	x^2, y^2 – Пирсона	$x^2_{фак.}$, $y^2_{фак.}$	9,99	7,52
	ω^2 – Смирнова	$\omega^2_{фак.}$	0,0133	0,014
9	Коэффициент вариации	Vx , Vy	28,94%	29,42%

10	Доверительный интервал генеральной средней	a_x , a_y	26,29 – 80,51	43,08 – 52,52
11	Доверительный интервал генерального среднего квадртического отклонения	Q_x , Q_y	17,93 – 26,23	15,4 – 22,54

В работе проведена статистическая обработка данных 60 хозяйств по затратам на производство кукурузы X (в тыс. руб.) и произведенной кукурузы Y (в т. ц.), а также установлен вид регрессионной связи и наличие корреляционной зависимости между указанными признаками.

Статистическая обработка данных по признаку X. Полученные статистические характеристики дают сделать следующие выводы:

1. затраты, на производство кукурузы, в среднем по выбранным хозяйствам составляет $\bar{x} = 53,4$ тыс. рублей. В большинстве хозяйств она больше Mo(x) = 53 тыс. руб. При этом наиболее передовые хозяйства засеивают $\bar{x}_{высш.} = 73,6$ тыс. руб., а отстающие $\bar{x}_{низ.} = 38,9$тыс. руб.;
2. затраты отведенных на кукурузу относительно, наиболее характерной для данной выборки, $\bar{x} = 53,6$ тыс. руб. характеризуется выборочным средним квадратическим отклонением или стандартой Sx = 21,3 тыс. рублей. В процентах это отклонение выражает коэффициент вариации Vx = 28,94%;
3. проведенная проверка согласия эмпирического и теоретического нормального распределения по критериям x^2 – Пирсона и ω^2 – Смирнова подтвердила, что распределение данной выборки можно считать, с надежностью p = 0,95, подчиняющимся закону нормального распределения, что дает основание использовать формулы нормального распределения для вычисления интервальных оценок;
4. с надежностью 0,95 определены доверительные интервалы генеральной средней $26,29 \leq a_x < 80,51$ и генерального среднего квадратического отклонения $17,93 < Q_x < 26,23$. Следовательно, в среднем по области количество затрат, производства кукурузы, для всех хозяйств будет находиться в пределах от 26,29 тыс. рублей до 80,51 тыс. рублей, а среднее квадратическое отклонение от 17,93 тыс. рублей до 26,23 тыс. рублей.

Статистическая обработка данных по признаку У. Полученные статистические характеристики дают сделать следующие выводы:

1. произведенной кукурузы в среднем по выбранным хозяйствам составляет $\bar{y} = 47,8$ т. ц. В большинстве хозяйств она равна Mo(y) = 55 т. ц. и Me(y) = 55 т. ц. (в данном случае получили многомодальный ряд). При этом наиболее передовые хозяйства получают $\bar{y}_{высш.} = 62,2$ т. ц. кукурузы, а отстающие $\bar{y}_{низ.} = 31,4$ т. ц.;
2. производство кукурузы относительно, наиболее характерной для данной выборки, произведенного $\bar{y} = 47,8$ т. ц. характеризуется выборочным средним квадратическим отклонением или стандартой Sy = 18,3 т. ц. В процентах это отклонение выражает коэффициент вариации Vy = 29,42%;

3. проведенная проверка согласия эмпирического и теоретического нормального распределения по критериям y^2 – Пирсона и ω^2 – Смирнова подтвердила, что распределение данной выборки можно считать, с надежностью p = 0,95, подчиняющимся закону нормального распределения, что дает основание использовать формулы нормального распределения для вычисления интервальных оценок;

4. с надежностью 0,95 определены доверительные интервалы генеральной средней $43,08 < a_x < 52,52$ и генерального среднего квадратического отклонения $15,4 < Q_y < 22,54$. Следовательно, в среднем по области произведенной кукурузы, будет находиться в пределах от 43,08 т. ц. до 52,52 т. ц., а среднее квадратическое отклонение от 15,4 т. ц. до 22,54 т. ц. [3].

Статистическая оценка адекватности полученных уравнений регрессии проводилась по анализу отклонений и объясняет около 50 % урожайности. Такие характеристики говорят о применении полученных моделей колебаний урожайности для прогнозирования. Аппарат математической статистики широко используется во всех областях и является незаменимым средством достижения наибольшей эффективности сельского хозяйства и экономики в целом.

Список литературы:

1. Белько, И. В. Теория вероятностей, математическая статистика, математическое программирование. [Электронный ресурс]: учебное пособие / И. В. Белько, И. М. Морозова, Е. А. Криштапович. — Электрон. текстовые данные. — М.:НИЦ ИНФРА-М, Нов. знание, 2016. - 299 с.: 60х90 1/16. - (Высшее образование: Бакалавриат) (Переплёт 7БЦ) ISBN 978-5-16-011748-5. - Режим доступа: http://znanium.com/catalog.php?bookinfo=542521 – Загл. с экрана.

2. Сапожников, П. Н. Теория вероятностей, математическая статистика в примерах, задачах и тестах [Электронный ресурс]: учебное пособие / П. Н. Сапожников, А. А. Макаров, М. В. Радионова. — Электрон. текстовые данные. — М.:КУРС, НИЦ ИНФРА-М, 2016. - 496 с.: 60х90 1/16. - (Бакалавриат и магистратура) (Переплёт 7БЦ) ISBN 978-5-906818-47-8. - Режим доступа: http://znanium.com/catalog.php?bookinfo=548242 – Загл. с экрана.

3. Бунимович Е.А., Булычев В.А. Вероятность и статистика. 5-9 кл.: пособие для общеобразоват. учреждений. – 3-е изд., стереотип. – М.: Дрофа, 2009. – 159 с.