УДК 311.1:330.43:633.1

Дронь В. С., кандидат фізико-математичних наук
Головне управління статистики у Чернівецькій області

ЗАСТОСУВАННЯ МОДИФІКОВАНИХ ЛІНІЙНИХ РЕГРЕСІЙНИХ МОДЕЛЕЙ ДЛЯ ПРОГНОЗУВАННЯ ПОКАЗНИКІВ У РОСЛИННИЦТВІ

Рецензент – доктор економічних наук, професор І. М. Школа

Запропоновано використання модифікації лінійних регресійних моделей для дослідження і прогнозування соціально-економічних показників, зокрема у рослинництві, за наявності додаткової інформації щодо зв'язку між величинами. У випадку відомого значення залежної величини при певному значенні незалежної змінної фіксується точка, через яку проходить пряма регресії. Наведено приклад застосування методу в процесі дослідження урожайності пшениці озимої. Здійснено порівняння за основними критеріями якості регресійних моделей, отриманих класичним і модифікованим методами.

Ключові слова: *проста лінійна регресія, прогноз, регресійна пряма, критерії якості прогнозних моделей, урожайність зернових культур.*

Постановка проблеми. Лінійні регресійні моделі встановлюють лінійну залежність між двома і більше змінними. При цьому одна зі змінних вважається залежною змінною й розглядається як функція від інших незалежних змінних.

У традиційному випадку [7, с. 44] проста вибіркова лінійна регресійна модель записується у вигляді:

$$y = b_0 + b_1 x + e, \qquad (1)$$

де: $y = \{y_1, ..., y_n\}$ – вектор спостережень за залежною змінною, $x = \{x_1, ..., x_n\}$ – вектор спостережень за незалежною змінною, $e = \{e_1, ..., e_n\}$ – вектор випадкових величин (помилок або відхилень), b_0, b_1 – невідомі параметри регресійної моделі.

Найчастіше параметри моделі (1) – невідомі b_0 і b_1 – знаходять за методом найменших квадратів, мінімізуючи суму квадратів відхилень

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = f(b_0, b_1) \rightarrow \min. \qquad (2)$$

Із геометричної точки зору побудована за (1) модель задає пряму на площині, де b_0 – перетин із віссю ординат, а b_1 – нахил. Економічний зміст параметрів b_0 і b_1 найчастіше трактують так:

1) величина b_0 вказує на значення залежної змінної y, яке набувається за нульового значення незалежної змінної x (у випадку часового аргументу – у початковий момент часу);

2) величина b_1 вказує, на скільки одиниць у середньому змінюється значення залежної змінної y, якщо значення незалежної змінної x змінити на одну одиницю.

Проте величина b_0 не завжди матиме такий економічний зміст, адже нуль може бути за межами області визначення незалежної змінної x.

Отже, модель (1) задає пряму, що проходить через точку $(0, b_0)$, проте ця точка може не мати економічного змісту або взагалі суперечити економічним законам чи закономірностям.

Аналіз основних досліджень і публікацій, у яких започатковано розв'язання проблеми. Регресійним моделям надається неабияка увага як у працях вітчизняних [3, 5, 7, 8], так і зарубіжних [1, 2, 9] дослідників, проте на вказану недоречність звертається недостатньо уваги. Один із підходів розв'язання даної проблеми запропонували О. І. Кулинич та Р. О. Кулинич [6]. Перевагою їхнього методу статистичних рівнянь залежностей є послаблення вимог щодо наявності чисельної сукупності об'єктів спостереження.

Метою даної роботи є формулювання підходу, який би давав змогу дати однозначне трактування параметрам регресійної моделі та забезпечував високий рівень її адекватності.

Завданням роботи є формулювання методу регресійної моделі з фіксованою точкою, порівняння за основними критеріями якості регресійних моделей, отриманих класичним та модифікованим методами, наведення прикладів застосування методу у рослинництві.

Для проведення дослідження були використані дані державного статистичного спостереження за формою №29-сг «Підсумки збору врожаю сільськогосподарських культур, плодів, ягід та винограду», проведені Головним управлінням статистики у Чернівецькій області, методи кореляційно-регресійного аналізу.

Результати досліджень. Регресійну модель у вигляді (1) доцільно будувати, коли відомо, що змінні мають залежність – лінійну чи близьку до лінійної. Часто існують випадки, коли із загальних соціально-економічних законів чи законо-

мірностей (тобто з економічного змісту величин) відома додаткова інформація про значення залежної змінної $y = y_0$ за певного значення незалежної змінної $x = x_0$. Наприклад, якщо x – це кількість реалізованого товару, а y – виручка від його реалізації, то, очевидно, що завжди в разі відсутності реалізації ($x_0 = 0$) виручка також буде нульовою ($y_0 = 0$). Побудована за спостережуваними даними модель типу (1) у даному випадку не обов'язково матиме значення параметра $b_0 = 0$, що суперечитиме вищевказаному висновку.

У таких випадках доцільно регресійну модель будувати у модифікованому вигляді – з фіксованою точкою:

$$y = y_0 + b(x - x_0) + e, \qquad (3)$$

де: x, y, e – ті ж величини, що й у моделі (1), b – невідомий параметр, (x_0, y_0) – фіксована точка, яка береться з відомого апріорного факту (при значенні $x = x_0$ залежна змінна y набуває значення y_0).

У моделі (3) лише один параметр, економічний і геометричний зміст якого збігається, як для b_1 із моделі (1).

Шукатимемо значення параметра також методом найменших квадратів, тобто з виразу

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y_0 - b(x_i - x_0))^2 = F(b) \;\rightarrow\; \min \qquad (4)$$

Мінімум функції однієї змінної (4) досягається за необхідної умови, коли перша похідна дорівнює нулеві, тобто

$$F'(b) = -2\sum_{i=1}^{n} (y_i - y_0 - b(x_i - x_0))(x_i - x_0) = 0 .$$

Звідси отримуємо

$$\sum_{i=1}^{n} (y_i - y_0)(x_i - x_0) = b\sum_{i=1}^{n} (x_i - x_0)^2 ,$$

або

$$b = \frac{\sum_{i=1}^{n} (y_i - y_0)(x_i - x_0)}{\sum_{i=1}^{n} (x_i - x_0)^2} . \qquad (5)$$

Зауважимо, що однією з властивостей простої вибіркової лінійної регресії (1), в якій невідомі параметри b_0 і b_1 визначені за методом найменших квадратів, є те, що регресійна пряма проходить через середню точку (це рівнозначне тому, що сума помилок дорівнює нулю) [7, с. 51]. Звідси маємо, що моделі (1) і (3) стають тотожними, якщо взяти за фіксовану точку середні значення спостережуваних змінних $x_0 = \bar{x}$, $y_0 = \bar{y}$. У цьому випадку формула (5) для b набуде відомого вигляду:

$$b = b_1 = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)} , \qquad (6)$$

де: $\mathrm{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$ – вибірковий коефіцієнт коваріації між x і y; $\mathrm{var}(x) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$ – вибіркова дисперсія величини x.

Вказане вище зауваження можна узагальнити: якщо побудована лінія регресії

$y = b_0 + b_1 x$

за моделлю (1) і точка (x_0, y_0) належить цій прямій, то моделі (1) і (3) збігаються.

Також є очевидним, що метод побудови ліній регресії з фіксованою точкою можна легко поширити на інші регресійні криві, в тому числі й багатофакторної регресії, зокрема на лінії, які є лінійними відносно параметрів та які можуть бути зведені до лінійних відносно невідомих параметрів кривих.

Здійснимо порівняння традиційного методу побудови прямої регресійної моделі та лінії регресії з фіксованою точкою на прикладі статистичних даних щодо урожайності пшениці озимої у сільськогосподарських підприємствах Чернівецької області у 2011 році.

За даними Головного управління статистики у Чернівецькій області [4], у 2011 році сільськогосподарськими товаровиробниками всіх категорій було зібрано 5928,8 тис. ц зернових і зернобобових культур у вазі після доробки, у тому числі сільськогосподарськими підприємствами (включаючи фермерські господарства) – 2542,4 тис. центнерів.

Основною зерновою культурою сільськогосподарських формувань Буковини є пшениця озима. За даними річної державної статистичної звітності за формою №29-сг «Підсумки збору врожаю сільськогосподарських культур, плодів, ягід та винограду», її вирощуванням у 2011 році займалося 222 товаровиробники області. На площі 26512 га ними було зібрано 1032,5 тис. ц цієї цінної зернової культури (табл. 1).

Побудуємо регресійні моделі за статистичними даними по згаданих 222 підприємствах загалом та у розрізі районів області. Очевидно, що існує залежність, яка є близькою до лінійної між такими величинами: площа збирання культури (незалежна детермінована величина) та валовий збір культури (залежна випадкова величина). Валовий збір сільськогосподарської культури

1. Збір урожаю пшениці озимої сільськогосподарськими підприємствами Чернівецької області у 2011 році

Регіон	Кількість підприємств	Площа, га	Валовий збір, ц	Урожайність, ц/га
Чернівецька область	222	26511,9	1032495,2	38,9
м. Чернівці	3	130,2	3336,0	25,6
Вижницький район	20	1008,2	26491,5	26,3
Герцаївський район	4	1340,0	49858,0	37,2
Глибоцький район	11	1331,0	45357,0	34,1
Заставнівський район	24	6385,3	245439,0	38,4
Кельменецький район	29	4877,7	202406,0	41,5
Кіцманський район	60	4612,0	203965,7	44,2
Новоселицький район	35	3538,5	131887,0	37,3
Путильський район	–	–	–	–
Сокирянський район	16	2124,5	75242,0	35,4
Сторожинецький район	12	431,7	16393,0	38,0
Хотинський район	8	732,8	32120,0	43,8

Джерело: [4]

залежить від багатьох факторів: родючості земель, клімату, кількості опадів, якості насіння, кількості добрив і термінів підживлення, термінів оранки, висівання й збору врожаю та багатьох інших. Вплив усіх цих факторів вважатимемо результатом дії інтегрованого випадкового чинника. Проте найбільше валовий збір довільної культури залежить від площі її висівання (посадки), а точніше, площі збирання урожаю. Основною характеристикою такої залежності є показник «урожайність культури».

При побудові лінійної регресійної моделі у традиційному вигляді (1) методом найменших квадратів за формулою (2) одержимо результати, подані у графах 3–5 таблиці 2.

У всіх випадках, окрім м. Чернівців, критерій Фішера

$$F = \frac{MSR}{MSE} \qquad (7)$$

підтверджує адекватність моделі. У формулі (7) позначено: $MSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$ – середній квадрат, що пояснює регресію,

$MSE = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ – середній квадрат помилок з урахуванням числа ступенів вільності, $\hat{y}_i$ – значення регресійної моделі при *i*-му випробовуванні.

2. Побудова лінійних регресійних моделей залежності валового збору пшениці озимої від площі збирання

Регіон	Кількість спостережень	Значення перетину b_0 моделі (1)	Значення нахилу b_1 моделі (1)	Значення коефіцієнта детермінації R^2 моделі (1)	Значення нахилу b моделі (3)
Чернівецька область	222	-574,7	43,8	0,95	43,0
м. Чернівці	3	164,0	21,8	0,94	24,1
Вижницький район	20	210,3	22,1	0,77	22,8
Герцаївський район	4	-3310,2	47,1	0,99	42,1
Глибоцький район	11	-169,9	35,5	0,99	35,1
Заставнівський район	24	-417,8	40,0	0,98	39,7
Кельменецький район	29	-1791,9	52,1	0,97	50,0
Кіцманський район	60	-822,7	54,9	0,95	53,1
Новоселицький район	35	-765,0	44,8	0,96	43,3
Сокирянський район	16	-393,3	38,4	0,80	37,3
Сторожинецький район	12	-88,0	40,4	0,99	40,0
Хотинський район	8	-251,4	46,6	0,96	44,8

Джерело: авторські розрахунки на основі даних статистичного спостереження

Для м. Чернівців даних трьох підприємств надто мало для повноцінного застосування методу найменших квадратів.

Як бачимо з даних таблиці 2, у жодному випадку значення перетину b_0 не дорівнює нулеві. У двох випадках з одинадцяти за регресійною моделлю при нульовій площі отримуємо ненульові збори, у дев'яти випадках значення b_0 менше нуля, що взагалі не має жодного економічного трактування. Отже, можна зробити висновок, що при всіх формальних ознаках адекватності моделей, вони абсолютно непридатні для пояснення збору урожаю на невеликих площах, тим більше для прогнозування.

Наприклад, у Кельменецькому районі достатньо значна кількість сільськогосподарських підприємств займалася вирощуванням пшениці озимої. За великого рівня адекватності побудованої регресійної моделі (за коефіцієнтом детермінації вона на 97 % пояснює зміну валового збору) модель дає абсолютно неприйнятні (від'ємні) значення валового збору на площах менших, аніж 34 гектари. Водночас із 29 товаровиробників району 11 мали площу збирання, що менша за цю величину.

Значення невідомого параметра лінійних регресійних моделей із фіксованою точкою вигляду (3) за спостережуваними даними по кожному з регіонів області подано в останній графі таблиці 2. У виразі (3) фіксованою точкою було взято точку (0,0), що випливає з так званого правила «відсутності рогу достатку» – з нульової площі не можна зібрати ніякого урожаю.

Легко помітити, що значення коефіцієнта нахилу b моделей типу (3) знаходиться ближче до значень фактичних середніх урожайностей по кожному з регіонів Буковини (остання графа таблиці 1), ніж значення b_1 моделі (1). Тобто, регресійні моделі типу (3) не лише адекватно задають можливі валові збори пшениці озимої на невеликих площах, але й їх параметр точніше, ніж у моделях типу (1), відображає свій економічний зміст, який у лінійному випадку задає одночасно граничну та середню урожайності.

Зіставимо моделі типу (1) та (3) за іншими критеріями якості [7, с. 65–68]. За своїми властивостями регресійні моделі типу (1) мають нульову середню помилку (*mean error*) прогнозу ($ME=\frac{1}{n}\sum_{i=1}^{n} e_i$). Для моделей типу (3) *ME* не дорівнює нулю, тому за цим критерієм моделі типу (1) мають перевагу. Зауважимо, що критерій *ME* характеризує ступінь зміщення прогнозів і для правильних прогнозів повинен прямувати до нуля за умови значної кількості спостережень.

Значення інших критеріїв якості регресійних моделей подано у вигляді таблиці 3.

Середня абсолютна помилка (*mean absolute error*) прогнозу $MAE=\frac{1}{n}\sum_{i=1}^{n}|e_i|$ визначає середнє значення помилки без урахування знаку. За даними таблиці 3 бачимо, що у трьох випадках даний критерій кращий для моделей типу (1), у дев'яти випадках – кращий для моделі типу (3).

3. Критерії якості для регресійних моделей залежності валового збору пшениці озимої від площі збирання

Регіон	Модель (1)				Модель (3)			
	MAE	*MAD*	*MSE*	*MAPE*	*MAE*	*MAD*	*MSE*	*MAPE*
Чернівецька область	1270,6	1270,6	8569422,1	536,9	1167,0	1220,0	8848603,1	80,9
м. Чернівці	167,8	167,8	34028,6	45,3	166,2	187,6	44565,8	20,5
райони								
Вижницький	619,9	619,9	2035859,8	443,2	523,0	604,9	2073166,4	32,0
Герцаївський	1411,3	1411,3	2909515,6	64,9	2301,2	1827,8	8335796,0	103,6
Глибоцький	228,1	228,1	113301,5	211,9	203,7	223,6	135246,3	30,9
Заставнівський	1590,1	1590,1	8379348,7	284,8	1469,9	1572,6	8523666,3	36,7
Кельменецький	1863,1	1863,1	9280947,8	238,0	2054,9	1802,5	11831396,2	103,9
Кіцманський	1289,0	1289,0	4601982,0	1155,3	1115,8	1207,0	5165146,1	158,7
Новоселицький	1096,5	1096,5	2871597,3	367,6	931,5	999,0	3334971,6	58,9
Сокирянський	1748,4	1748,4	11749911,0	199,7	1712,9	1705,9	11848781,9	77,3
Сторожинецький	166,0	166,0	87040,3	159,9	133,4	152,6	93327,1	61,8
Хотинський	516,7	516,7	416506,9	28,3	527,8	541,0	438254,1	79,6

Джерело: авторські розрахунки

Абсолютне середнє відхилення (*mean absolute deviation*) $MAD=\frac{1}{n}\sum_{i=1}^{n}\left|e_i-\bar{e}\right|$ характеризує рівень відхилень похибок від їх середнього значення. Як і для попереднього критерію, у дев'яти з дванадцяти випадків ця величина менша для моделей типу (3).

Поширеним у виборі оптимальних моделей прогнозування є критерій – середній квадрат помилки (*mean square error*) $MSE=\frac{1}{n}\sum_{i=1}^{n}e_i^2$, який у випадку моделей (1) збігається з дисперсією помилок. У порівнянні з критерієм *MAE* він надає більшої негативної ваги великим відхиленням. Для моделей типу (1) він має менше значення. Поєднуючи два критерії – *MSE* і *MAE,* можна дійти висновку, що в середньому відхилення прогнозів від спостережуваних значень частіше кращі у моделях типу (3), проте саме за цими моделями частіше зустрічаються великі відхилення.

Останній розрахований критерій – абсолютна середня процентна помилка (*mean absolute persentage error*) $MAPE=\frac{1}{n}\sum_{i=1}^{n}\frac{\left|e_i\right|}{y_i}\cdot 100\%$ частіше використовується при порівнянні точності прогнозів різнорідних об'єктів, оскільки він характеризує відносну точність прогнозу. Проте його іноді використовують також для характеристики адекватності моделей: при значенні показника *MAPE* менше 10 % вважається висока якість моделі, 10–20 % – добра, 20–50 % – задовільна, понад 50 % – незадовільна. Такий критерій спонукає досить критично ставитися до окремих побудованих регресійних моделей, передусім, побудованих за формулою (1).

Висновки. На прикладі побудованих лінійних регресійних моделей для продуктивності зернової ниви було проілюстровано, що при використанні модифікованого підходу до побудови регресійних моделей частіше отримуємо не гірші прогнозні значення. Проте регресійні моделі з фіксованою точкою враховують додаткові теоретичні відомості про об'єкт дослідження, й усі параметри моделі завжди мають економічне трактування. Запропонований підхід доцільно використовувати при побудові регресійних моделей за наявності додаткової інформації щодо значення змінних.

БІБЛІОГРАФІЯ

1. *Айвазян С. А.* Прикладная статистика в задачах и упражнениях: [Учебник для вузов] / С. А. Айвазян, В. С. Мхитарян. – М. : ЮНИТИ-ДАНА, 2001. – 270 с.

2. Економічне прогнозування: вступ / К. Холден, Д. А. Піл, Дж. Л. Томпсон. – К. : Інформтехніка-ЕМЦ, 1996. – 216 с.

3. *Єріна А. М.* Теорія статистики: [Практикум] / А. М. Єріна, З. О. Пальян. – К. : Товариство «Знання», КОО, 1997. – 325 с.

4. Збір урожаю сільськогосподарських культур, плодів, ягід та винограду в області за 2011 р. : [статистичний бюлетень] / Головне управління статистики у Чернівецькій області. – Чернівці, 2012. – 57 с.

5. *Крамченко Л. І.* Економічна статистика : [Навч. посібник] / Л. І. Крамченко, Н. П. Лутчин, Б. С. Москаль. – Львів: «Новий Світ – 2000», 2004. – 364 с.

6. *Кулинич О. І.* Теорія методу статистичних рівнянь залежностей / О. І. Кулинич, Р. О. Кулинич // Прикладна статистика: проблеми теорії та практики : Зб. наук. пр. – Вип. 1 / Держкомстат України, Держ. акад. статистики, обліку та аудиту. – К. : ДП Інформаційно-аналітичне агентство. – 2007. – С. 62–76.

7. *Лук'яненко І. Г.* Економетрика : [Підручник] / І. Г. Лук'яненко, Л. І. Краснікова – К.: Товариство «Знання», КОО, 1998. – 494 с.

8. Методологічні положення зі статистики. – Вип.2 , Т. 1 : [за редакцією О. Г. Осауленка] / Державний комітет статистики України – К. : ІВЦ Держкомстату України, 2006. – 504 с.

9. *Ханк Д. Э.* Бизнес-прогнозирование, 7-е издание : [Пер. с англ.] / Д. Э. Ханк, Д. У. Уичерн, А. Дж. Райте. – М. : Издательский дом «Вильямс», 2003. – 656 с.