УДК 61:[519.237.3+519.237.5]

О.Т. Девіняк[1,2], О.В. Ігнат[1], М.І. Кусій[2]

РЕГРЕСІЯ ПУАССОНА – МЕТОД ВИБОРУ ДЛЯ АНАЛІЗУ ЛІЧИЛЬНИХ ДАНИХ (НА ПРИКЛАДІ ЗАХВОРЮВАНОСТІ НАСЕЛЕННЯ ЗАКАРПАТСЬКОЇ ОБЛАСТІ НА КОЛОРЕКТАЛЬНИЙ РАК)

Державний вищий навчальний заклад «Ужгородський національний університет»[1], м. Ужгород, Україна
Львівський національний медичний університет імені Данила Галицького[2], м. Львів, Україна
Львівський державний університет безпеки життєдіяльності[3], м. Львів, Україна

e-mail: *o.devinyak@gmail.com*

Резюме: Наведено ряд аргументів на користь застосування регресії Пуассона з метою аналізу медичних та епідеміологічних показників, які засновані на кількості подій. Теоретичні викладки експериментально підкріплені наочним прикладом – аналізом захворюваності населення Закарпатської області на колоректальний рак.

Ключові слова: регресія *Пуассона*, множинна лінійна регресія, лічильні дані, захворюваність, колоректальний рак.

Вступ. Серед медичних та епідеміологічних показників, що збираються, вивчаються та моделюються під час наукових досліджень, часто зустрічаються лічильні (отримані при лічбі) дані. Прикладами таких даних є кількості випадків захворювання, кількості загиблих, кількості звернень до лікувально-профілактичного закладу, кількості виконаних обстежень, аналізів чи операцій – як абсолютні величини, так і приведені (наприклад, на 100 тис. населення). На сучасному етапі розвитку біостатистики та доказової медицини для визначення впливу різноманітних факторів на захворюваність (смертність) та прогнозування їх рівня використовують регресію Пуассона та її варіанти. Основоположні роботи щодо застосування регресії Пуассона до медичних даних з'явились ще на початку 1980-х років і стосувались дослідження захворюваності та смертності від ракових захворювань[10,11]. *L.S. Zeger* використовував регресію Пуассона для моделювання захворюваності на поліомієліт у США[13]. Зв'язок між кількістю випадків інсульту напротязі 1955-1989 рр. у м. Рочестер (Міннесота, США) та різними соціодемографічними та епідеміологічними показниками також досліджувалось за допомогою регресії *Пуассона*[12]. Однак досі вітчизняні науковці для вивчення лічильних даних використовують кореляційний аналіз або множинну лінійну регресію. Так, *А.М. Гупал* та співавтори використовували покрокову регресію (варіант множинної лінійної регресії із вибором незалежних змінних покроковим методом) для вивчення зв'язку стоматологічної захворюваності з елементним складом емалі зубів[6]. Модель лінійної регресії також використовували *О.Л. Гром* та *Д.Т. Садова* для дослідження захворюваності на туберкульоз населення Львівської області[3]. Вплив окремих антропогенних факторів на захворюваність населення України на рак легені визначали також за допомогою множинної лінійної регресії, зокрема залежність захворюваності від рівня викидів шкідливих речовин описували поліномом третього степеня[2]. Залежність захворюваності вузловим зобом від вмісту сполук міді у їжі жителів АР Крим підтверджували коефіцієнтом кореляції *Спірмена*[1]. Кореляційний аналіз та лінійну регресійну модель також використовували для вивчення зв'язку між загальною захворюваністю населення Луганської області та метеоумовами[5]. У роботі *В.З. Свиридюка* зв'язок захворюваностей на різні нозологічні одиниці підтвер-

джували коефіцієнтом кореляції *Спірмена*[7]. Інший непараметричний коефіцієнт кореляції (*Кендалла*) використовували для епідеміологічних досліджень дитячого туберкульозу в Україні[8]. Варто зазначити, що використання кореляційного аналізу дозволяє досліджувати ознаки лише попарно, не враховуючи одночасного впливу інших ознак, що є істотним недоліком вказаного підходу. Крім того, формування висновків на основі багатьох обчислених коефіцієнтів кореляції призводить до проблеми множинного порівняння[9]. Вибір правильного методу для дослідження є важливим, оскільки дозволяє коректно оцінити вплив факторів та здійснювати прогнозування із кращою точністю.

Мета дослідження – здійснити порівняння регресії *Пуассона* та множинної лінійної регресії як можливих підходів до аналізу лічильних даних на прикладі захворюваності населення Закарпатської області на колоректальний рак (КРР).

Матеріали та методи дослідження. Модель множинної лінійної регресії описує зв'язок між залежною змінною та незалежними факторами, виражений формулою

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon , \qquad (1)$$

де β є регресійними коефіцієнтами, а ε – похибка. При цьому регресійні коефіцієнти встановлюються методом найменших квадратів і, згідно теореми *Гаусса-Маркова*, якщо похибки є некорельованими та мають однакові дисперсії, то отримані коефіцієнти є незміщеними. Додатково, тестування статистичної значимості отриманих коефіцієнтів вимагає, щоб розподіл похибок був подібний до нормального розподілу. З іншого боку, якщо кількісні дані відображають число подій, що трапились у визначеному інтервалі часу, ймовірність яких не залежить від часу, що пройшов після останньої події, то вони описуються розподілом *Пуассона*. Порівняння функцій густини ймовірності розподілу *Пуассона* та нормального розподілу наведене на рис. 1.

Перш за все, варто зауважити, що розподіл Пуассона є дискретною функцією і визначений лише для цілочисельних аргументів. На графіку ці аргументи представлені точками, а лінії між точками насправді не мають смислу і наведені лише для зручності наочного порівняння. Якщо ж згадати, що досліджуваною величиною є кількість подій, то дійсно, кількість захворювань чи смертей не може бути дробовою. В той же час, нормальний розподіл є неперервним, а отже апріорі здатен лише приблизно відобразити розподіл похибки досліджуваної величини.

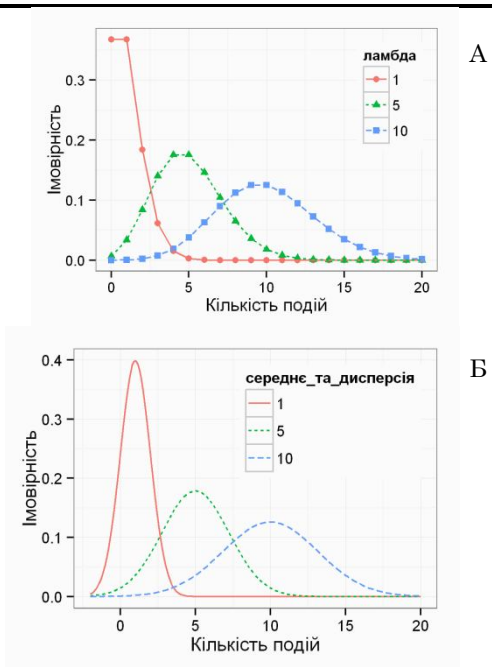

Рис. 1. **Функції густини імовірності розподілу *Пуассона* (А) та нормального розподілу (Б) при різних параметрах**

По-друге, нормальний розподіл задається двома параметрами: середнім та дисперсією. Причому побудова лінійної регресійної моделі методом найменших квадратів передбачає, що дисперсії похибок однакові. З іншого боку, дисперсії похибок кількості подій зростають разом із власне кількістю подій. Пояснимо це на прикладі. Нехай в деякій лікарні напротязі місяця зафіксували два випадки захворювання на хворобу А. Якщо в наступному місяці трапляться чотири випадки хвороби А, то є сенс говорити про зростання захворюваності. І нехай в деякій лікарні за той же час інша хвороба (позначимо її Б) була виявлена 100 раз. Якщо в наступному місяці захворюваність сягне числа 102, то реально цей результат не є значущим, і про зростання захворюваності говорити зарано. В той же час, оскільки в обох випадках відмінність між місяцями склала дві події, при використання моделі із константною дисперсією обом зростанням надається рівна вага. Це призведе до певного нехтування похибкою для малих чисел захворюваності та надмірне пристосування моделі до великих чисел. Цю проблему виключає застосування регресії *Пуассона*. Адже розподіл *Пуассона* задається одним параметром лямбда, який виступає одночасно і середнім, і дисперсією. Таким чином, для більшої кількості подій автоматично задається і більша дисперсія (рис. 1.). У деяких випадках статистичного аналізу потрібно задання іншої (зазвичай більшої) дисперсії розподілу, ніж при розподілі *Пуассона*. Це трапляється,

якщо кількість подій відома не точно, а з істотним наближенням, що є значно менше поширеним у медицині. Для цього існують від'ємні біноміальні регресійні моделі. Потретє, використання множинної лінійної регресії може призвести до прогнозування від'ємної кількості подій, що є нісенітницею. Тобто, регресія *Пуассона* має цілий ряд теоретичних переваг над множинною лінійною регресією для аналізу лічильних даних.

Регресія *Пуассона* використовує логарифм як канонічний зв'язок між залежною і незалежними змінними, тому модель регресії має вигляд:

$$\log(y) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon, \qquad (2)$$

А отже, значення залежної змінної можна отримати після експоненціювання:

$$y = e^{\beta_0} \times e^{\beta_1 x_1} \times ... \times e^{\beta_p x_p} \times e^{\varepsilon}, \qquad (3)$$

Для інтерпретації коефіцієнтів регресії β розглянемо випадок регресії *Пуассона* з однією незалежною змінною:

$$y = e^{\beta_0} \times e^{\beta_1 x_1} \qquad (4)$$

Збільшення незалежної змінної на одиницю дасть:

$$y = e^{\beta_0} \times e^{\beta_1 (x_1 + 1)} = e^{\beta_0} \times e^{\beta_1 x_1} \times e^{\beta_1} \qquad (5)$$

Таким чином, зміна в будь-якій незалежній змінній x має мультиплікативний вплив на залежну змінну y. Причому розмір цього впливу визначається експонентою регресійного коефіцієнта β.

Також варто знати, що, оскільки при зростанні кількості подій розподіл *Пуассона* наближається до нормального розподілу (нормальний розподіл є асимптотичним для розподілу *Пуассона*, при нескінченній кількості подій два розподіли є тотожними), то переваги регресії *Пуассона* поступово нівелюються. І якщо усі спостережувані кількості перевищують 30, регресію *Пуассона* можна безпечно замінити на множинну лінійну регресію (навіть у цьому випадку, краще використовувати логарифмічний зв'язок між незалежною та залежними змінними замість зв'язку тотожності).

Зазначивши теоретичні переваги регресії Пуассона над множинною лінійною регресією для аналізу лічильних даних, підтвердимо це фактичними результатами. Прикладом для статистичного аналізу в даному дослідженні було обрано інформацію щодо захворюваності населення Закарпатської області на КРР протягом 1995-2010 рр., наведену у праці *О.В. Ігната* та ін.[4]. Демографічна структура населення Закарпатської області отримана за допомогою офіційного сайту Державної служби статистики України http://ukrstat.gov.ua. Аналіз даних проводився у середовищі для статистичних обчислень R 3.0.1.

Результати дослідження та їх обговорення. Досліджувався зв'язок між кількістю зареєстрованих випадків КРР у Закарпатській області та роком спостереження, статтю і віком. Отримана модель множинної лінійної регресії характеризується коефіцієнтом детермінації R^2=0,940 та регресійними коефіцієнтами, наведеними у табл. 1.

Таблиця 1. **Коефіцієнти лінійної регресійної моделі для кількості випадків захворювання на КРР**

Незалежна змінна	Коефіцієнт	Стандартна похибка	p-величина
Зсув	-8,56	1,88	$1{,}10\times10^{-5}$
Рік	0,827	0,139	$1{,}73\times10^{-8}$
Стать (чол.)	5,02	1,28	0,000131
Вік (20-39)	4,47	2,02	0,0288
Вік (40-59)	42,0	2,02	$< 2\times10^{-16}$ *
Вік (60-79)	81,2	2,02	$< 2\times10^{-16}$ *
Вік (80+)	5,94	2,02	0,00387

Примітка: * 2×10^{-16} – найменше додатнє число, доступне для розміщення в оперативній пам'яті програмою R.

Зміст коефіцієнтів лінійної регресії – збільшення зсуву на потрібну кількість випадків захворюваності. При цьому сам зсув являє собою прогнозовану кількість випадків захворюваності на КРР для базової групи. У нашому дослідженні базовою є кількість захворювань жінок (базове значення статі) до 20 років (базове значення віку) у 1995 р. (базовий рік). Результати регресійного аналізу вивели від'ємне значення для зсуву, що, насправді, є неможливим, а справжня кількість пацієнтів із базової групи рівна нулю. Більше того, від'ємне значення зсуву підкріплене статистичною значимістю ($p=1{,}10\times10^{-5}$). Регресійний коефіцієнт для віку близький до одиниці. Це означає, що з кожним роком реєструють з діагнозом КРР в середньому на одну людину більше, ніж у попередньому році. Таким чином, можна говорити про досить слабке, однак статистично значиме ($p=1{,}73\times10^{-8}$) зростання. Коефіцієнти біля різних вікових груп вказують, що, порівняно із

пацієнтами до 20 років, кількість пацієнтів з КРР у групі 20-39 років в середньому більша на 4,6, у групі 40-59 – на 42, у групі 60-79 – на 81, а у групі пацієнтів старших 80 років – лише на 6. Слід зазначити, що отримані за допомогою множинної лінійної регресії коефіцієнти є малоінформативними для цілей епідеміологічних досліджень. Так, візьмемо до прикладу переважання на 5 випадків кількості захворювань чоловіків над жінками. Не знаючи усього контексту дослідження, визначити – багато це, чи мало – неможливо. Так, якщо кількість жінок з КРР складає всього 5-10, то спостерігаємо значну відмінність між групами, а якщо 500-1000, то відмінність несуттєва. Для виведення на основі результатів лінійної регресії звичних для епідеміології ризиків та співвідношень ризиків потрібно проводити додаткові обчислення.

Розглянемо тепер результати регресії *Пуассона* для тих же даних. Коефіцієнт детермінації для моделі регресії *Пуассона* вищий, ніж у випадку множинної лінійної регресії, і становить R^2=0,967. Варто зазначити, що коефіцієнт детермінації заснований на сумі квадратичних похибок, яка, в свою чергу, мінімізується лише при множинній лінійній регресії. А при регресії *Пуассона* мінімізується від'ємний логарифм правдоподібності. Тобто коефіцієнт детермінації не є оптимальним показником точності регресії *Пуассона*, і в даному дослідженні використаний для цілей порівняння.

Регресійні коефіцієнти після експоненціювання є співвідношеннями ризиків (табл. 2.).

Таблиця 2. **Коефіцієнти моделі регресії *Пуассона* для кількості випадків захворювання на КРР**

Незалежна змінна	Коефіцієнт	Стандартна похибка	Експонента коефіцієнта	p-величина
Зсув	-2,20	0,448	0,111	$9{,}68\times10^{-7}$
Рік	0,0308	0,00333	1,031	$< 2\times10^{-16}$
Стать (чол.)	0,187	0,0306	1,206	$9{,}35\times10^{-10}$
Вік (20-39)	3,39	0,455	29,60	$9{,}30\times10^{-14}$
Вік (40-59)	5,60	0,448	270,0	$< 2\times10^{-16}$
Вік (60-79)	6,26	0,448	520,8	$< 2\times10^{-16}$
Вік (80+)	3,66	0,453	39,0	$6{,}02\times10^{-16}$

Експонента зсуву є прогнозованою кількістю випадків захворювання на КРР в базових умовах (рік спостереження 1995, пацієнти до 20 років та жіночої статі). І справді, ця кількість становила нуль осіб, що і описується моделлю. Значення експоненти коефіцієнта для року, що становить 1,031 означає, що кожен наступний рік кількість захворювань підвищується у 1,031 раз, тобто на 3,1% порівняно з попереднім. Гіпотеза про зростання захворюваності підкріплена p-величиною, яка є меншою, ніж 2×10^{-16}.

Для порівняння, у випадку множинної лінійної регресії p-величина для аналогічної гіпотези склала $1{,}73\times10^{-8}$. Менші значення p-величини дозволяють з більшою впевненістю стверджувати отримані висновки.

Щодо зв'язку між кількістю захворювань і статтю, то обчислена експонента коефіцієнта свідчить про те, що ризик розвитку КРР у чоловіків у 1,206 раз (тобто на 20%) більший, ніж у жінок. Примітно, що цей висновок є більш адекватним, ніж різниця у 5 осіб, що була отримана за допомогою множинної лінійної регресії.

Випадки захворювання людей до 20 років рідкісні та поодиничні, тож відмінність між обома статями не те що досягає 5 осіб, а не спостерігається взагалі. У рамках же регресії Пуассона 0,111×1,206=0,134, що дає такий самий прогноз, що і для жінок – 0. Захворюваність у когорті людей від 20 до 39 років є вищою (в середньому за рік 4,0 випадки захворювання жінок та 5,2 – чоловіків). Для цього випадку прогнози множинної лінійної регресії такі:

для жінок: -8,56[зсув] + 4,47[вік] + (0+10*0,827)/2[середній доданок для року спостереження] = 0,045

для чоловіків: -8,56[зсув] + 4,47[вік] + (0+10*0,827)/2[середній доданок для року спостереження] + 5,02[стать]=5,065

Отже, спостерігаємо істотну помилку для жіночої статі і точне попадання для чоловічої.

Здійснивши прогноз моделлю регресії *Пуассона*, маємо:

для жінок: 0,111[зсув] × 29,60[вік] × (1-1,031^{11})/((1-1,031)×11)[середній множник для року спостереження] = 3,85

для чоловіків: 0,111[зсув] × 29,60[вік] × (1-1,031^{11})/((1-1,031)×11)[середній множник для року спостереження] × 1.206[стать] = 4,64,

що значно точніше, ніж у випадку множинної лінійної регресії.

Крім того, згідно моделі множинної лінійної регресії, відмінність між групами «до 20 років» та «20-39 років» за кількістю захворювань заледве досягла статистичної значимості (p=0,0288), тоді як для аналогічної гіпотези регресія *Пуассона* дає p=9,30×10^{-14}.

Візуальне порівняння точності двох моделей дозволяє оцінити істотну перевагу регресії *Пуассона* (рис. 2.)

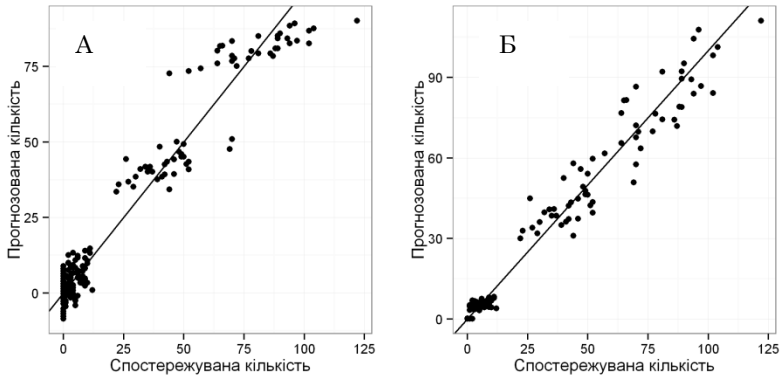

Рис. 2. **Точність прогнозування моделі множинної лінійної регресії (А) та регресії *Пуассона* (Б)**

Регресію *Пуассона* також можна застосовувати до вивчення захворюваності, тобто кількості випадків хвороби на визначену кількість (зазвичай на 100 тис.) населення. Так, увівши у формулу замість абсолютного числа відносну кількість випадків, маємо:

$$\log(y/N) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon, \qquad (6)$$

що після перенесення кількості населення N у праву частину дає

$$\log(y) = \log(N) + \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon. \qquad (7)$$

Так, порівняємо результати застосування множинної лінійної регресії та регресії *Пуассона* до аналізу захворюваності на КРР по відношенню до 100 тис. населення (табл. 3).

Таблиця 3. **Результати дослідження захворюваності за допомогою множинної лінійної регресії та регресії *Пуассона***

Незалежна змінна	Множинна лінійна регресія, R^2=0,767		Регресія *Пуассона*, R^2=0,959	
	Коефіцієнт	p-величина	Експонента коефіцієнта	p-величина
Зсув	-24,16	3,12×10^{-5}	0,0510	3,20×10^{-11}
Рік	1,575	0,000214	1,027	1,17×10^{-15}
Стать (чол.)	24,85	1,12×10^{-9}	1,692	< 2×10^{-16}
Вік (20-39)	2,342	0,699	28,52	1,71×10^{-13}
Вік (40-59)	27,09	1,48×10^{-5}	322,6	< 2×10^{-16}
Вік (60-79)	101,8	< 2×10^{-16}	1190	< 2×10^{-16}
Вік (80+)	74,88	< 2×10^{-16}	847,8	< 2×10^{-16}

По-перше, з огляду на коефіцієнти детермінації, точність моделі регресії *Пуассона* значно краща, а по-друге, p-величини у випадку регресії *Пуассона* істотно менші, що дає більшу впевненість у сформульованих на її основі висновках. Що ж до порівняння моделей, отриманих при вивченні кількості випадків КРР з моделями захворюваності на КРР, то найбільш суттєвою відмінністю є значне зростання коефіцієнта для статі. Врахувавши, що частка чоловіків у демографічному складі населення є нижчою (особливо у старших вікових групах), ризик захворювання на КРР у чоловіків насправді є значно вищий (на 69%), ніж у жінок.

Щоб забезпечити відтворюваність проведеного дослідження, вихідні дані захворюваності населення Закарпатської області на КРР розміщено в мережі інтернет за адресою http://pharm.uz.ua/files/krr-data.csv, а код на мові R, що був використаний при аналізі – за адресою http://pharm.uz.ua/files/krr-code.r.

Висновки:

Регресія *Пуассона* має значні переваги над множинною лінійною регресією при дослідженні лічильних даних. Крім коректного моделювання природи епідеміологічних показників (вірний розподіл похибки, зростання дисперсії разом зі зростанням значень показників, прогнозування лише додатніх чисел, зручне трактування експоненти коефіцієнтів), застосування регресії *Пуассона* призводить до вищих коефіцієнтів детермінації (а, отже, і нижчих похибок) та нижчих значень статистичної значимості, що було підтверджено на прикладі аналізу захворюваності населення Закарпатської області на колоректальний рак. Вказані аргументи роблять регресію *Пуассона* методом вибору при аналізі лічильних даних у медичних дослідженнях.

Література:

1. *Безруков О.Ф.* Роль концентрации медьсодержащих соединений в пище в возникновении патологии щитовидной железы / *О.Ф. Безруков, П.Е. Григорьев* // Таврический медико-биологический вестник. – 2010. – Т. 13, № 4 (52). – С. 11-13.
2. *Войтко О.В.* Вплив окремих антропогенних факторів на захворюваність населення України на рак легені / *О.В. Войтко, С.Т. Омельчук, Ю.М. Остапчук* // Онкологія. – 2009. – Т. 11, № 4. – С. 257-262.
3. *Гром О.Л.* До питання прогнозування фармацевтичної допомоги хворим на активний туберкульоз із врахуванням впливу факторів захворюваності / *О.Л. Гром, Д.Т. Садова* // Клінічна фармація, фармакотерпія та медична стандартизація. – 2010. – №1-2. – С. 35-40.
4. Захворюваність і виживаність хворих на колоректальний рак у Закарпатській області / *О.В. Ігнат, А.В. Русин, В.І. Русин* [та інш.] // Науковий вісник Ужгородського університету, серія «Медицина». – 2013. – №2 (47). – С. 42-52.
5. *Зубов О.Р.* Аналіз впливу метеоелементів на загальну захворюваність населення Луганської області / *О.Р. Зубов, Л.Г. Зубова, А.А. Кошурнікова* // Прикладна екологія. – 2009. – №1. – С. 2-9.
6. Комп'ютерні засоби в моделюванні процесів стоматологічної захворюваності / *А.М. Гупал, О.І. Остапко, Т.Я. Грачова, О.С. Воробйов* // Комп'ютерні засоби, мережі та системи. – 2009. – №8. – С. 52-57.
7. *Свиридюк В.З.* Використання коефіцієнта поєднання для характеристики етіологічних чинників хронічного панкреатиту за допомогою комп'ютерних технологій аналізу електронних баз даних (реєстрів) захворюваності / *В.З. Свиридюк* // Сучасна гастрентерологія. – 2008. – №1 (39). – С. 7-15.
8. *Речкіна О.О.* Епідеміологічні аспекти дитячого туберкульозу в Україні / *О.О. Речкіна, В.В. Куц* // Український пульмонологічний журнал. – 2007. – №2. – С. 53-56.
9. *Benjamini Y.* Simultaneous and selective inference: Current successes and future challenges / *Y. Benjamini* // Biometrical Journal. – 2010. – №52 (6). – P. 708-721.
10. *Frome E.L.* The analysis of rates using Poisson regression models / *E.L. Frome* // Biometrics. – 1983. – Vol.39. – P. 665-674.
11. *Frome E.L.* Use of Poisson regression models in estimating incidence rates and ratios / *E.L. Frome, H. Checkoway* // American Journal of Epidemiology. – 1985. –№2. – P. 309-323.
12. Stroke incidence, prevalence, and survival: secular trends in Rochester, Minnesota, through 1989 / *R.D. Brown, J.P. Whisnant, J.D. Sicks* [et al.] // Stroke. – 1996. – №3. – P. 373-380.
13. *Zeger L.S.* A regression model for time series of counts / *L.S. Zeger* // Biometrika. – 1988. – Vol.75. – P. 621–629.

УДК 61:[519.237.3+519.237.5]

РЕГРЕССИЯ ПУАССОНА – МЕТОД ВЫБОРА ДЛЯ АНАЛИЗА СЧЕТНЫХ ДАННЫХ (НА ПРИМЕРЕ ЗАБОЛЕВАЕМОСТИ НАСЕЛЕНИЯ ЗАКАРПАТСКОЙ ОБЛАСТИ НА КОЛОРЕКТАЛЬНЫЙ РАК)

О.Т. Девиняк[1,2], О.В. Игнат[1], М.И. Кусий[3]
ГВУЗ «Ужгородский национальный университет»[1], г. Ужгород, Украина
Львовский национальный медицинский университет имени Данила Галицкого[2], г. Львов, Украина
Львовский государственный университет безопасности жизнедеятельности[3], г. Львов, Украина

Резюме: Приведен ряд аргументов в пользу применения регрессии Пуассона с целью анализа медицинских и эпидемиологических показателей, которые основаны на количестве событий. Теоретические вы-

кладки экспериментально подкреплены наглядным примером – анализом заболеваемости населения Закарпатской области на колоректальный рак.

Ключевые слова: регрессия Пуассона, множественная линейная регрессия, счетные данные, заболеваемость, колоректальный рак.

UDC: 61:[519.237.3+519.237.5]

POISSON REGRESSION IS THE METHOD OF CHOICE FOR COUNT DATA ANALYSIS (ON THE EXAMPLE OF COLORECTAL CANCER MORBIDITY IN TRANSCARPATHIAN REGION)

O.T. Devinyak[1,2], O.V. Ihnat[1], M.I. Kusiy[3]
State University «Uzhgorod National University»[1], Uzhgorod, Ukraine
Danylo Halytsky Lviv National Medical University[2], Lviv, Ukraine
Lviv State University of Life Safety[3], Lviv, Ukraine

Summary: The number of arguments for using Poisson regression during based on count data medical and epidemiological indicators analysis are given. Theoretical statements are proved experimentally using the analysis of colorectal cancer morbidity in Transcarpathian region as a case study.

Keywords: Poisson regression, multiple linear regression, count data, morbidity, colorectal cancer.

Надійшла до редакції 18.12.2013 р.