http://www.moderntechno.de/index.php/meit/article/view/meit31-00-076
DOI: 10.30890/2567-5273.2024-31-00-076

APPLICATION OF STATISTICAL ANALYSIS FOR MEDICAL DATA
ЗАСТОСУВАННЯ СТАТИСТИЧНОГО АНАЛІЗУ ДЛЯ МЕДИЧНИХ ДАНИХ

Doroshenko I.V. / Дорошенко І.В.
c. p.-m.s., as.prof. / к. ф.-м.н., доц.
ORCID: 0000-0001-8729-1768
Knihnitska T.V. / Кнігніцька Т.В.
Doctor of Philosophy in Mathematics and Statistics /
доктор філософії у галузі математики та статистики
ORCID: 0000-0003-4614-5945
Chernivtsi National University, Chernivtsi, Kotsyubynskoho 2, 58012
Чернівецький національний університет, Чернівці, вул.Коцюбинського 2, 58012

Анотація. *В статті розглянуто використання різноманітних статистичних підходів і методів машинного навчання в медицині. Проведено модельний аналіз на прикладі пацієнтів, де визначено фактори, що впливають на ймовірність виникнення інсульту. Здійснено аналіз даних, щоб встановити взаємозв'язок між фізичними характеристиками пацієнта, його шкідливими звичками, способом життя та ймовірністю виникнення інсульту. Для оцінки взаємозв'язку числових даних були побудовані моделі лінійної регресії та модель логістичної регресії для прогнозування ймовірності інсульту.*

Ключові слова:*. статистичний аналіз, машинне навчання, лінійна регресія, логістична регресія.*

Вступ.

Збільшення об’єму інформації в медицині та біології показало, що статистика є потужним інструментом концентрації знань, оскільки медицина є модельперш за все наукою експериментальною. Сучасні медичні дослідження є міждисциплінарними і тому вимагають обов’язкової участі спеціаліста-біостатистика.

1. Постановка задачі

Протягом останнього чвертьстоліття відбувся значний прогрес в галузі науки та техніки. Людство досягло успіхів у створенні роботів, які можуть виконувати різноманітні завдання надаючи допомогу у різних сферах людської діяльності. Сучасні досягнення вже не вражають нікого туристичним польотом у космос чи подорожжю міжконтинентальною ракетою навколо Землі. З неабиякими технологічними досягненнями люди, на жаль, продовжують стикатися з різними захворюваннями. Другою за частотою причиною смерті після раку є інсульт. За статистикою Всесвітньої організації охорони здоров'я, 11% всіх смертей пов'язані з крововиливом в мозок – інсультом. Відомо, що клітинний рівень організму людини залишається малодослідженим. Таким чином, застосування статистичного аналізу даних може допомогти виявити зв'язок між фізичними характеристиками пацієнтів, шкідливими звичками, віком та іншими факторами і випадками інсульту.

Основна мета цього дослідження полягає в пошуку відповідей на наступні питання:

- Чи впливає куріння на ймовірність інсульту?
- Чи впливає гіпертонія на ймовірність інсульту?
- Чи впливає вік на ймовірність інсульту?

- Чи існує лінійна залежність між індексом маси тіла та середнім рівнем глюкози в організмі, віком пацієнта?

Для вирішення цих завдань розглянемо моделі звичайної лінійної регресії та модель логістичної регресії. Основна різниця полягає в тому, що залежна змінна для лінійної регресії повинна бути числового типу, тоді як для логістичної регресії - факторного. Логістична регресія дозволяє класифікувати пацієнтів за ймовірністю на дві групи (0 - НІ, 1 - ТАК). Поза роботою над основними чотирма питаннями дослідження розглянемо описову статистику, обробку відсутніх значень та використання статистичних тестів. Інтелектуальний аналіз факторів, які впливають на ймовірність інсульту, може допомогти кожному робити висновки та розуміти важливі аспекти цього дослідження.

Дані були отримані з платформи Kaggle за наступним посиланням https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset. Дані складаються з 12 стовпців з 5110 записами. Інформація про атрибути містить:

1) id: унікальний ідентифікатор;
2) gender: «Чоловік», «Жінка»;
3) age: вік пацієнта;
4) hypertension: 0 - немає гіпертонії, 1 - є гіпертонія;
5) heart disease: 0 - немає захворювань серця, 1 - є захворювання серця;
6) ever married: "Ні" або "Так";
7) work type: "діти", " державна ", " ніколи не працював ", "приватна";
8) Residence type : "Сільський" або "Міський";
9) avg_glucose_level;
10) bmi (індекс маси тіла);
11) smoking_status : "раніше курив", "ніколи не курив", "курить " або "невідомо";
12) stroke (інсульт): 1 - був інсульт, або 0 - якщо ні.

Для візуалізації світової карти рівня смертності від інсульту на 100 000 населення ми використовуємо ще один набір даних із https://www.worldlifeexpectancy.com/cause-of-death/stroke/by-country/.

Побудуємо карту світу, на якій відображено рівні смертності від інсульту для країн (183 країни), які вказані в цьому наборі даних (рис. 1).

Як видно з рис. 1, країни з найнижчим рівнем смертності від інсульту включають Австралію, США, Канаду та країни Європи.

2. Результати

Набір даних містить числові та факторні змінні. Наприклад, змінна gender є фактором типу (чоловічий, жіночий, інший).

На рис. 2 представлено індекс маси тіла в залежності від віку. Після досягнення 40 років людина будь-якого індексу маси тіла може стикнутися з ризиком інсульту. Таким чином, ймовірність виникнення інсульту зростає після 40 років для всіх груп населення. Цікаво відзначити, що на тому ж рисунку індекс маси тіла не має чіткої взаємозв'язку з інсультом. Пацієнти з різними значеннями індексу маси тіла можуть уникнути ризику інсульту.

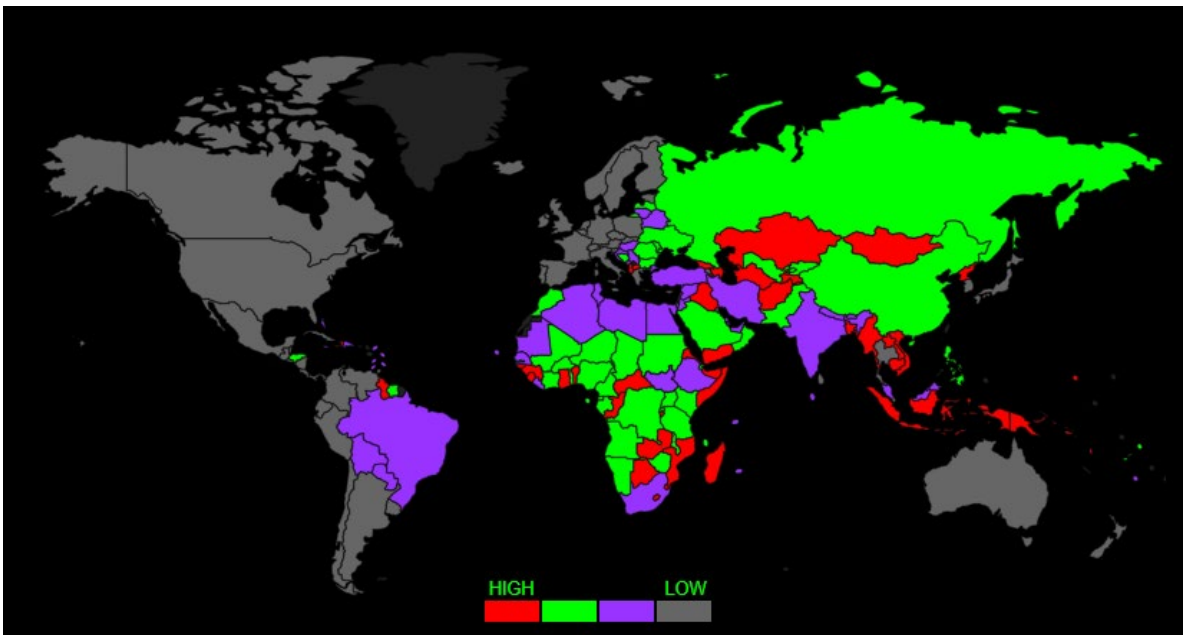

Рис. 1 «Смертність від інсульту на 100 000 населення»

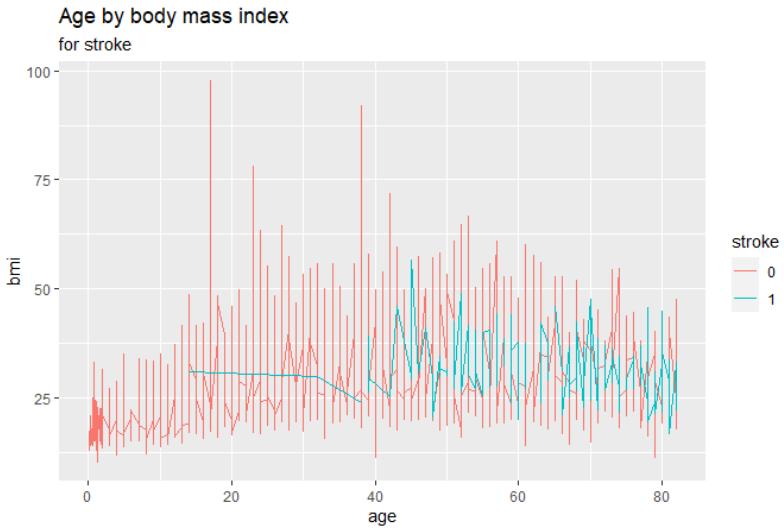

Рис. 2 «Індекс маси тіла за віком»

На рис. 3 представлена корелограма даних. Схематично видно, що атрибути або стовпці в початковому наборі даних не виявляють значущої кореляції між собою. Кореляція розглядається як слабка, якщо коефіцієнт кореляції знаходиться в межах від -0,3 до 0,3.

Тепер створимо модель лінійної регресії, щоб визначити, чи існує лінійна залежність між індексом маси тіла та середнім рівнем глюкози. На рис. 4 відображено цю взаємозалежність, де індекс маси тіла виступає залежною змінною, а середній рівень глюкози – незалежною змінною.

Рівняння лінійної регресії має такий вигляд:

$$bmi = 25.63 + 0.03*avg\_glucose\_level.$$

Тепер розглянемо, як вік впливає на середній рівень глюкози. З цією метою ми побудуємо іншу модель лінійної регресії та представимо її на графіку (рис. 5). З рисунку 5 видно, що зі збільшенням кількості років пацієнта середній рівень глюкози також зростає.

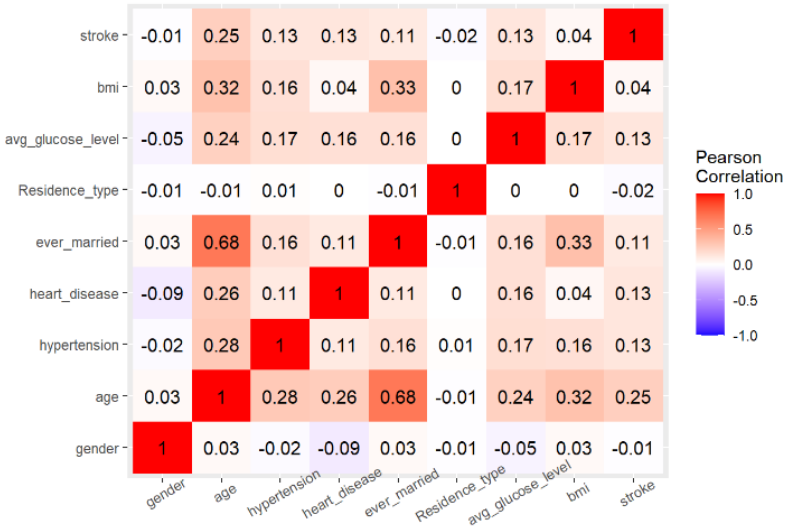

Рис.3 «Корелограма»

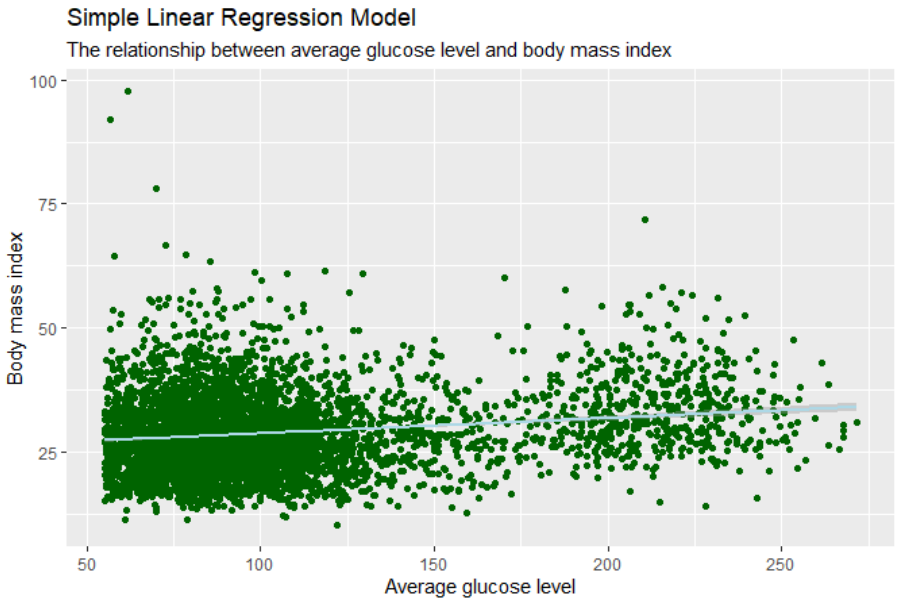

Рис. 4 «Модель лінійної регресії для BMI»

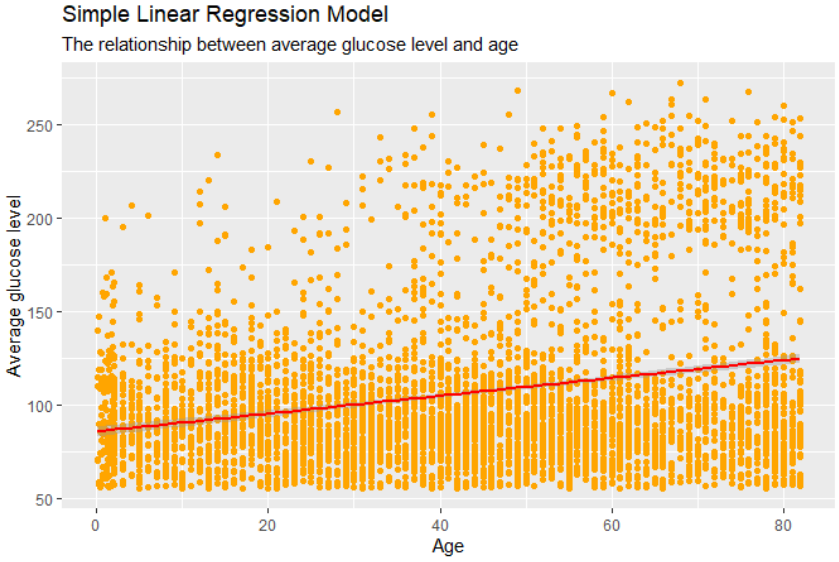

Рис. 5 «Модель лінійної регресії для змінної *average_glucose_level*»

Рівняння лінійної регресії для цієї залежності має наступний вигляд:

$$average\ \ glucose\ \ level = 85.53+0.47*age.$$

Важливо відзначити, що обидві побудовані моделі є нерепрезентативними. Іншими словами, існують інші фактори, які впливають на змінну відповіді. Цей висновок став очевидним після аналізу значень коефіцієнтів детермінації. Звичайно прийнято вважати модель прийнятною, якщо коефіцієнт детермінації перевищує 80%. У нашому випадку цей показник становить приблизно 10% в обох випадках.

Лінійна регресія не може служити інструментом для визначення ймовірності того, чи стане пацієнт жертвою інсульту. У нашому випадку відповідна змінна *stroke* може приймати лише два можливі значення. Тут на допомогу приходить логістична регресія, яка дозволяє отримати відповідь у вигляді ймовірності від 0 до 1.

Отже, логістична регресія - це метод, який використовується для прогнозування залежної змінної (інсульт), заданої незалежними змінними (вік, індекс маси тіла тощо), так, що залежна змінна є категоріальною.

Формула моделі логістичної регресії:

$$P(X) = \frac{e^{(\beta_0 + \beta_1 x)}}{e^{(\beta_0 + \beta_1 x)} + 1}.$$

У логістичній регресії, коли значення незалежної змінної збільшується на одиницю вимірювання, показник змінюється з логарифмом коефіцієнта β_0.

Таблиця 1. показує результати моделі логістичної регресії. Ми не перераховуємо всі незалежні змінні в таблиці, щоб не обтяжувати звіт. Змінні, не зазначені в таблиці, є статистично незначущими. Тобто вони не впливають на змінну *stroke*.

Таблиця 1. - Результати регресії

	Оцінка	Pr(>|t|)
(Intercept)	-5,922908	2.14e-13 ***
age	0,072741	< 2e-16 ***
hypertension1	0,563302	0,00348 **
heart_disease1	0,342615	0,13270
ever_marriedYes	-0,233803	0,37770
Residence_typeUrban	-0,102680	0,53266
avg_glucose_level	0,002709	0,06287.
` work_type_Self -employed`	-1,438607	0,09906
`smoking_status_never smoked	-0,244233	0,20289
smoking_status_smokes	-0,074173	0,77644

Тому змінні *age* та *hypertension* є статистично значимим для ймовірності отримати інсульт.

Висновки.

У даному дослідженні проведено модельний аналіз пацієнтів, де визначені фактори, що впливають на ймовірність виникнення інсульту у пацієнтів.

Проведений аналіз даних стосується встановлення зв'язку між фізичними характеристиками пацієнта, його шкідливими звичками, способом життя та ймовірністю інсульту. Для числових даних використовувались моделі лінійної регресії. Оскільки змінна відповіді "інсульт" має факторний характер, була розроблена модель логістичної регресії для прогнозування ймовірності виникнення інсульту. Всі аналізи та обробка даних виконані за допомогою середовища R.

Література:

[1] . Neil C. Jones, Pavel A. Pevzner An Introduction to Bioinformatics Algorithms. Cambridge, Massachusetts: London.- 2004. – 436 p.

Abstract. *The article explores the application of diverse statistical approaches and machine learning methods in medicine. A model analysis was conducted using patient data, where factors influencing the likelihood of stroke occurrence were identified. Data analysis was carried out to establish the correlation between the patient's physical characteristics, harmful habits, lifestyle, and the probability of stroke occurrence. Linear regression models and a logistic regression model were constructed to assess the correlation of numerical data and predict the probability of stroke occurrence.*

Keywords: *statistical analysis, machine learning, linear regression, logistic regression.*

Стаття відправлена: 20.02.2024 р.