# Самостійна робота №5
# з дисципліни "асимптотична статистика"
# Варіант №4

Горбунова Даніела Денисовича
4 курс бакалаврату
група "комп'ютерна статистика"

12 травня 2021 р.

## 1 Вступ.

У даній роботі побудована модель логістичної регресії для даних про якість вина. Для моделі отримані довірчі інтервали коефіцієнтів, що оцінюються, проаналізовано якість прогнозування на початкових даних. В кінці звіту покажемо застосування обраної моделі на підвибірці невеликого обсягу та відповідні показники якості прогнозу.

## 2 Хід роботи.

### 2.1 Початкові дані.

Як було частково зазначено у попередньому розділі, ми будемо працювати з даними про результатами аналізу зразків вина, обраного з трьох різних виноградників (змінна Site). У варіанті №4, ми обмежимося лише першими двома виноградниками. Змінні, на основі яких побудуємо модель, такі:

- $X_1$ = Alcogol - вміст спиртів;
- $X_2$ = Ash - кількість золи;
- $X_3$ = Malic acid - вміст яблучної кислоти;
- $X_4$ = Proline - вмість проліну.

їх можна узагальнити як кількість досліджуваних речовин та домішок у вині (для спрощення, надалі позначимо через $X^j := (X_1^j, X_2^j, X_3^j, X_4^j)^\top$).

### 2.2 Постановка задачі та розв'язання.

Ми розгляаємо задачу бінарної класифікації: за вказаною кількістю речовин $X$ у вині потрібно спрогнозувати номер виноградника $Y$, де продукт було вироблено. Для цього розглянемо модель логістичної регресії для опису апостеріорної ймовірності того, що вино вироблено на обраному винограднику (наприклад, на першому) за змінними (2.1):

$$\mathbb{P}\left(Y^j = 1|X_1^j, X_2^j, X_3^j, X_4^j\right) = \frac{1}{1+\exp(-(b_0 + \sum_{i=1}^4 b_i X_i^j))} = g(b_0, b, X^j),$$

де параметри $b_0,\ b = (b_1, b_2, b_3, b_4)^\top$ треба оцінити за спостереженнями. Для цього скористаємося методом найбільшої вірогідності. Будемо вважати, що спостереження $X^j$ мають невипадкову природу. В силу дихотомічності $Y$ та незалежності спостережень, можна записати функцію вірогідності вигляду:

$$L(Y, b_0, b) = \prod_{j=1}^n g(b_0, b, X^j)^{Y_j}(1 - g(b_0, b, X^j))^{1-Y_j},\ b_0 \in \mathbb{R},\ b \in \mathbb{R}^4$$

Тоді оцінювання $b_0, b$ зводиться до розв'язання оптимізаційної задачі:

$$L(Y, b_0, b) \to \max_{b_0, b}$$

або це еквівалентно, в силу монотонності та зростання натурального логарифму, такій умові:

$$\ln L(Y, b_0, b) \to \max_{b_0, b} \tag{1}$$

Наступні кроки спираються на використанні наближених методів обчислення екстремальних значень, які будуть задовольняти умові (1). Для цього застосуємо техніку нелінійної оптимізації, що реалізована в R:

```
g <- function(t) { 1/(1 + exp(-t)) } # Логістична функція
# Для спрощення викладок, ми розбили програмну реалізацію цільової функції на дві частини
f <- function(b.0, b, x) { g(b.0 + t(b)%*%x) } # Апостеріорна ймовірність
lik.function <- function(b.0, b, X, Y) # Логарифмічна функція вірогідності
{
  f.0 <- function(u) {g(b.0 + t(b)%*%u)}
  f.val <- apply(X, 1, f.0)
  lik.val <- prod(f.val^Y * (1 - f.val)^(1-Y))
  log(lik.val)
}
# Функція, яка наближено обчислює точку максимуму логарифмічної функції вірогідності
mle.estimation <- function(X, Y, z.0 = rep(0.01, ncol(X) + 1))
{
  to.minimize <- function(z) {-lik.function(z[1], z[-1], X, Y)}
  nlm.struct <- nlm(to.minimize, z.0)
  print(summary(nlm.struct))
  nlm.struct$estimate
}
```

Після виконання програмного коду на початкових даних, отримали наступні оцінки для невідомих параметрів $b_0$ та $b$:

$$\hat{b}_0 = 166.31342521,\ \hat{b} = (-0.02645128, -10.25041790, -2.84288246, -1.92734266)^\top$$

### 2.3 Вибір значущих змінних в моделі.

### 2.4 Прогнозування.

## 3 Висновки.