Course: Computer statistics

Lecture 10: Logistic Regression. ROC - analysis.

Lecturer: Oleksandr Dykhovychnyi

Лекція 10. Логістична регресія. ROC - аналіз

I. Модель.
II. Оцінки
III. Аналіз рівняння регресії
IV. Прогноз
V. ROC аналіз

x2

I. Модель.

Регресійні моделі, які вже було розглянуто, передбачають, що залежна змінна Y була **кількісною**, Однак, у багатьох випадках змінна Y може бути нечисловою, назвімо її категоріальною. Приміром: чоловік/жінка; хворий/здоровий; або червоний/ зелений/ жовтий. При цьому незалежні змінні (предиктори) також можуть бути як кількісні, так і **категоріальні**.

Як правило, залежна змінна може набувати двох значень, тому мова йтиме про ймовірність набуття першого або другого значень. Нехай ця ймовірність $P(X)$ залежить від одного кількісного предиктора X. Якщо Y прийняв перше значення, то $P(X)=1,$ ні - $P(X)=0$.

Але виникає наступна проблема. Якщо ймовірність $P(X)$ записати у найпростішому вигляді лінійної залежності:

$$P(X)=a_0+a_1X,$$

то виникає протиріччя. Імовірність належить відрізку $[0,1]$, а права частина змінюється від $-\infty$ до $+\infty$.

Тому впроваджують наступне відношення: $\frac{P(X)}{1-P(X)}$, яке називають **шансом** або **ризиком** (**odds**) та його логарифм прирівнюють до лінійної функції:

$$\ln\frac{P(X)}{1-P(X)} = a_0 + a_1 X \ .$$

Звідки

$$P(X) = \frac{e^{a_0+a_1X}}{1+e^{a_0+a_1X}}$$

Таку залежність називають логістичною. Її графік має у випадку залежності ймовірності дефолту від фінансового балансу вигляд (рис.1):

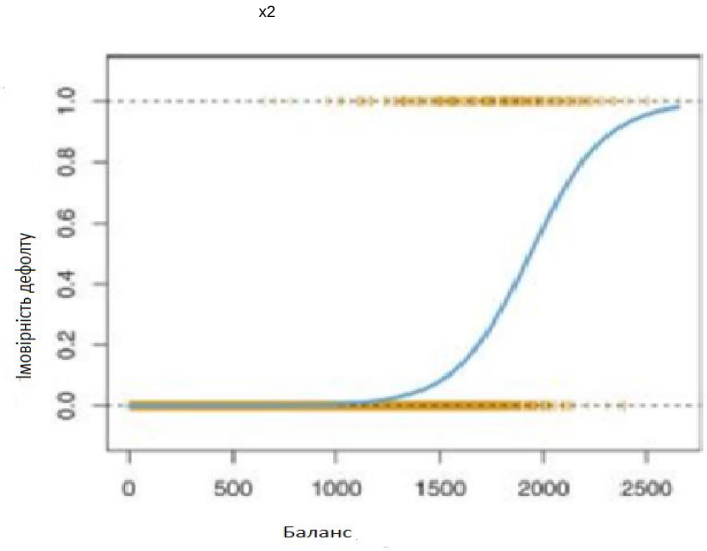

Рис1**.** (Рисунок згенеровано **RStudio**).

II. Оцінки

Оцінки коефіцієнтів отримують, як правило, не методом найменших квадратів, або методом максимальної вірогідності за вибіркою значень $(x_i, y_i), i = \overline{1,n}$.У прикладі розглядаються результати вступу до вишу у залежності від статі, балів з української, англійської, математики.

Приклади:

```
> library(ggplot2)
> library(Hmisc)
```

```
> my_df <- read.csv('C:/Users/adykhovychnyi/Desktop/R курс/Нова статистика/Логистическая регрессия/train.csv',sep=";")
 > my_df
    gender ukr engl math univer
1      boy  57   52   41      N
2      boy  44   33   54      N
3      boy  63   44   47      N
.....................................................................
65     boy  52   62   66      Y
66    girl  68   62   65      Y
.....................................................................
150   girl  63   65   65      Y
```

```
> ggplot(my_df, aes(ukr, math, col = gender))+
+   geom_point(size = 1)+
+       facet_grid(.~hon)+
+       theme(axis.text=element_text(size=15),
+       axis.title=element_text(size=15,face="bold"))
```

Будуємо діаграму залежності вступу від статі, математики та української (рис.2):

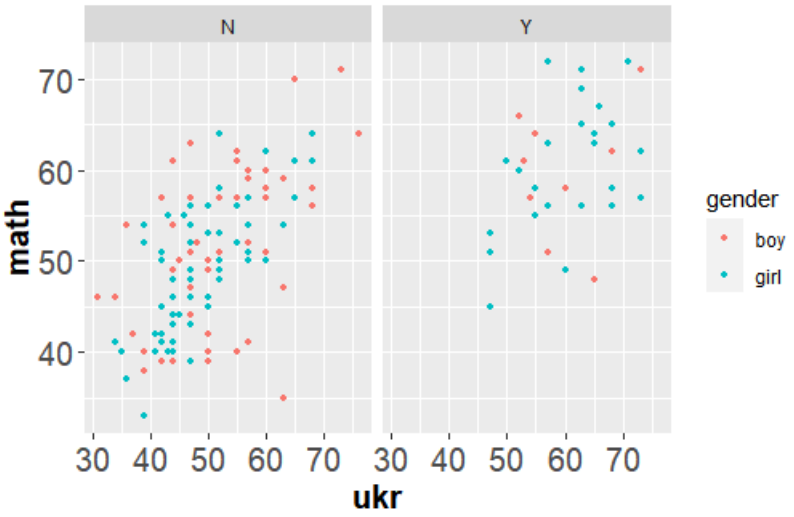

Рис. 2. (Рисунок згенеровано **RStudio**).

З діаграми видно, що більш успішно вступають дівчата.

III. Аналіз рівняння регресії

Переведемо символьні змінні у факторні:

Приклади:

```
> my_df$univer <- as.factor(my_df$univer)
> my_df$gender <- as.factor(my_df$gender)
```

Будуємо модель логістичної регресії:

```
> fit  <- glm(univer ~ ukr + math + gender, my_df, fami
ly = "binomial")
> summary(fit)


Call:
glm(formula = univer ~ ukr + math + gender, family = "binomial",
    data = my_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8173  -0.5989  -0.3086  -0.1087   2.3626

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.31013    2.25755  -5.896 3.73e-09 ***
ukr           0.06677    0.03291   2.029  0.04247 *
math          0.13907    0.04243   3.277  0.00105 **
gendergirl    1.18606    0.51326   2.311  0.02084 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 162.98  on 149  degrees of freedom
Residual deviance: 109.75  on 146  degrees of freedom
AIC: 117.75

Number of Fisher Scoring iterations: 5
```

Отримана модель має наступний вигляд:

Якщо учень хлопець (G=0), то

$$\ln\frac{P(U,M,G)}{1-P(U,M,G)} = -13.31013 + 0.06677U + 0.13907M + 1.18606G\,.$$

Тоді $P(U,M,G) = \dfrac{e^{-13.31013+0.06677U+0.13907M+1.18606G}}{1+e^{-13.31013+0.06677U+0.13907M+1.18606G}}$.

IV. Прогноз

Згідно отриманої моделі можна отримати прогноз.

Приклади:

```
> head(predict(object = fit))
        1         2         3         4         5         6
-3.802171 -2.862355 -2.567112 -2.244825 -4.130526 -1.593634
```

Тут ми маємо прогноз перших шести рядків таблиці **fit**, а саме, логарифмів ві д **odds.**
Якщо викликати функцію **predict** з опцією **type = "response":**

Приклади:

```
> head(predict(object = fit, type = "response"))
         1          2          3          4          5          6
0.02183486 0.05404619 0.07128525 0.09579675 0.01582013 0.16887324 ,
```

то отримаємо вже відповідні зпрогнозовані ймовірності вступу.

Створимо нову змінну, у яку впишемо розраховані ймовірності вступу:

Приклади:

```
> my_df$prob  <- predict(object = fit, type = "response")
```

Тоді у таблицю додається стовпець:

Приклади:

```
> my_df
    gender ukr engl math univer    prob
1      boy  57   52   41      N 0.021834862
2      boy  44   33   54      N 0.054046191
3      boy  63   44   47      N 0.071285255
4      boy  47   52   57      N 0.095796750
5      boy  50   59   42      N 0.015820126
.....................................................................................
```

За побудованою моделлю можна зробити прогноз

```
> test_df  <- read.csv('C:/Users/adykhovychnyi/Desktop/R курс/Нова статист
ика/Логистическая регрессия/test1.csv',sep=";")
> test_df$univer  <- NA #видаляю
> test_df$univer  <- predict(fit, newdata = test_df, type = "response")# прогн
озую
> test_df
   gender ukr engl math      univer
1    girl  68   59   53 0.447007543
2     boy  34   46   45 0.008311927
3     boy  73   62   73 0.847666060
```

```
4    girl  50   67   66 0.597046352
5     boy  42   49   43 0.010711632
6    girl  47   44   42 0.041297036
7     boy  63   49   49 0.092040129
8     boy  47   47   41 0.011318671
```

V. ROC-аналіз

Крива **ROC** (**Receiver Operator Characteristic**) є графіком, який часто використовується для відображення результатів бінарної класифікації. У бінарній класифікації вводять два класи: перший - з позитивними наслідками, другий - з негативними. ROC- крива демонструє, як змінюється кількість правильно класифікованих позитивних наслідків у порівнянні з кількістю неправильно класифікованих негативних наслідків.

Зазвичай у ROC-аналізі перший клас називають правдиво позитивним, а другий - хибно негативними. Це означає, що у класифікатора є параметр, який можна змінювати, і він визначає, як розділяються наслідки на два таких класи. Параметр, який розділяє на класи, зазвичай називають точкою відсікання (**cut-off value**). Це значення спричиняє різні види помилок, такі як помилка першого та другого роду. Щоб зрозуміти сутність цих помилок введемо таблицю, яка базується на результатах розділення на класи згідно моделі та фактичною належністю наслідків до класів.

Модель	Правдиво позитивно	Правдиво негативно
Позитивно	TP	FP
Негативно	FN	TN

- **TP** (*True Positives*) — правдиво класифіковані позитивні наслідки (істинно позитивні наслідки).
- **TN** (*True Negatives*) — правдиво класифіковані негативні наслідки (істинно негативні наслідки).

- **FN** (*False Negatives*) — позитивні наслідки, які було класифіковано як негативні (помилка I роду).

- **FP** (*False Positives*) — негативні наслідки, які було класифіковано як позитивні (помилка II роду).

Позитивний і негативний наслідки визначаються залежно від поставленої цілі. Приміром, при прогнозуванні ймовірності захворювання, "хворий пацієнт" може бути позитивним результатом, а "здоровий пацієнт" - негативним. У іншому контексті, коли ми визначаємо ймовірність того, що людина здорова, "здоровий пацієнт" стає позитивним наслідком. Коли проводиться аналіз, то застосовують або абсолютні показники, або відноснимі - частки (**rates**), які обчислюють у відсотках:
Частка правдиво позитивних наслідків (**True Positives Rate**):

$$TPR = \frac{TP}{TP + FN} \times 100\%$$

Частка хибно позитивних наслідків (**False Positives Rate**):

$$FPR = \frac{FP}{TN + FP} \times 100\%$$

Запровадимо два нових поняття - **чутливості і специфічності**, за допомого ю яких визначають якість нашого класифікатора.

Чутливість (Sensitivity) – називають частку правдиво позитивних наслідків:

$$S_\varepsilon = TPR = \frac{TP}{TP + FN} \times 100\%.$$

Специфічність (Specificity) - називають частку правдиво негативних наслідків, які правильно визначив класифікатор:

$$S_p = \frac{FP}{TN + FP} \times 100\%.$$

Відмітимо, що $FPR = 100 - Sp$

Модель, яка має високу чутливістю, правильно ідентифікує позитивні результати (виявляє хворих) у випадку наявності захворювання. З іншого боку, модель, яка має високу специфічнісь, правильно визначає негативні результати (виявляє здорових) у випадку відсутності захворювання. У медичному контексті, чутливий діагностичний тест максимально запобігає пропуску хворих, а специфічний тест визначає лише достеменно хворих.

Побудова ROC-кривої проводиться так:

1. Для кожного значення порога відсікання в інтервалі [0,1] кроком *dx*, приміром, 0,01 обчислюють чутливость S_ε і специфічність S_p .
2. Далі будують графік: вісь Y - S_ε , вісь X - FPR = 100-Sp, - частка хибно позитивних випадків.

Отримуємо криву (рис. 3).

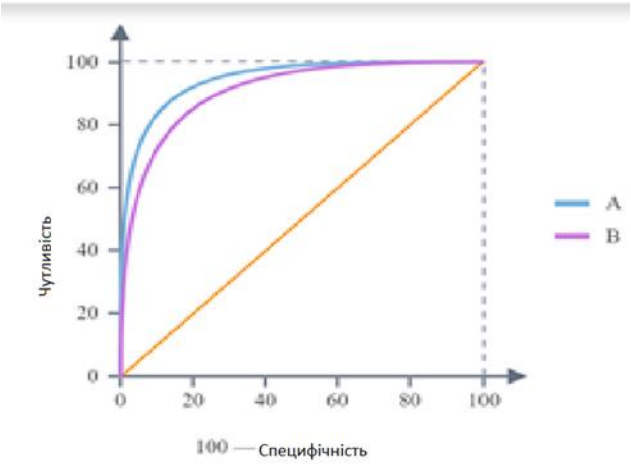

Рис. 3 (Рисунок згенеровано **RStudio**).

У випадку правильно побудованого класифікатора ROC-крива проходить через точки у верхньому лівому куті, де всі хворі будуть визначені правильно, а всі здорові - вірно віднесені до свого класу. Така крива відповідає ідеальній чутливості та специфічності. Якщо крива ближча до верхнього лівого кута, то модель забезпечує краще передбачення. На відміну, чим менше вигинів у кривій і чим вона ближче до діагоналі, тим меншу ефективність має модель. Діагональна лінія відповідає порожньому класифікатору, який не розрізняє класи.

Розташування ROC-кривих відносно одна до одної показує їх порівняльну ефективність. Крива, яка знаходиться вище і лівіше, свідчить про більшу здатність моделі до передбачення. Приміром, на рисунку 1 дві ROC-криві зображені на одному графіку. Очевидно, що модель «А» краще.

Порівняння ROC-кривих за допомогою візуального аналізу не завжди є ефективним способом визначення найкращої моделі. Одним з методів порівняння є оцінка площі під кривими. Теоретично ця площа може

змінюватися від 0 до 1, але, оскільки моделі зазвичай розташовані вище діагоналі, звичайно розглядаються значення від 0,5 («непотрібний» класифікатор) до 1,0 («ідеальна» модель). Це значення можна отримати, обчисливши площу під кривою, обмеженою осями координат, і зліва вгорі – емпірично визначеними точками. Числовий показник цієї площі називають AUC (Area Under Curve).

Можна припускати, що високі значення параметру AUC, забезпечує сильнішу прогностичну здатність моделі. Однак слід зазначити, що показник AUC призначений переважно для порівняльного аналізу кількох моделей, а не для оцінки чутливості і специфічності моделі. У деяких джерелах наводиться експертна шкала значень AUC, за якою можна оцінювати якість моделі:

Значення AUC	Модель
0,9-1,0	Відмінно
0,8-0,9	Дуже добре
0,7-0,8	Добре
0,6-0,7	Середнє
0,5-0,6	Незадовільне

Для ідеальної моделі чутливість та специфічність дорівнюють 100%. Але це недосяжно, до того ж, неможливо одночасно підвищити чутливість та специфічність. Компроміс досягається зміною порога відсікання.Тому треба визначити оптимальний поріг відсікання (**optimal cut-off value**).

Щоб визначити оптимальний поріг задають відповідний критерій одним з наступних варіантів:

1. Обмеження на мінімальні чутливість або специфічність даної моделі. Тоді, якщо потрібно забезпечити чутливість класифікатору не менше 75%, то за оптимальний поріг можна прийняти максимальну специфічність або чутливість, яка саме і досягається при 75%.

2. Вимога одночасного максимального сумарного значення чутливості і специфічності моделі, тобто $Cutt\_offo = \max_k (S_{\varepsilon k} + S_{pk})$.
3. Або вимога збалансованості чутливісті і специфічністі, тобто $Se \approx Sp$: $Cutt\_offo = \min_k \left| S_{\varepsilon k} - S_{pk} \right|$.

Порогом може бути точка перетину таких кривих, на осі X лежить поріг відсікання, а на осі Y відкладено чутливість та специфічність (рис. 4).

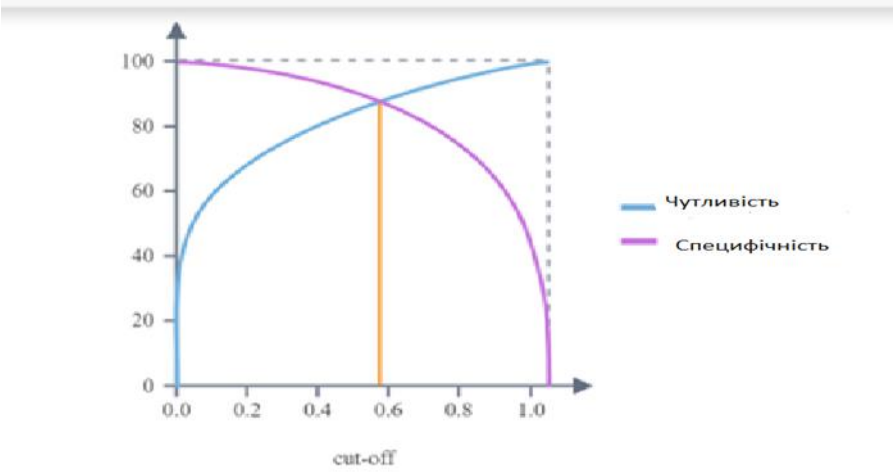

Рис. 4 (Рисунок згенеровано **RStudio**).

Розгляньмо, як це спрацьовує у нашому випадку. Для побудови ROC – кри вих використовують пакет **ROCR.** ROC – криву зображено на рисунку 5.

Приклади:

```
> library(ROCR)
> pred_fit <- prediction(my_df$prob, my_df$univer)
> perf_fit <- performance(pred_fit,"tpr","fpr")#Розрахунок ,"tpr","fpr"
>         plot(perf_fit, colorize=T , print.cutoffs.at = seq(0,1,by=0.1))
>         auc  <- performance(pred_fit, measure = "auc")
> str(auc)
```

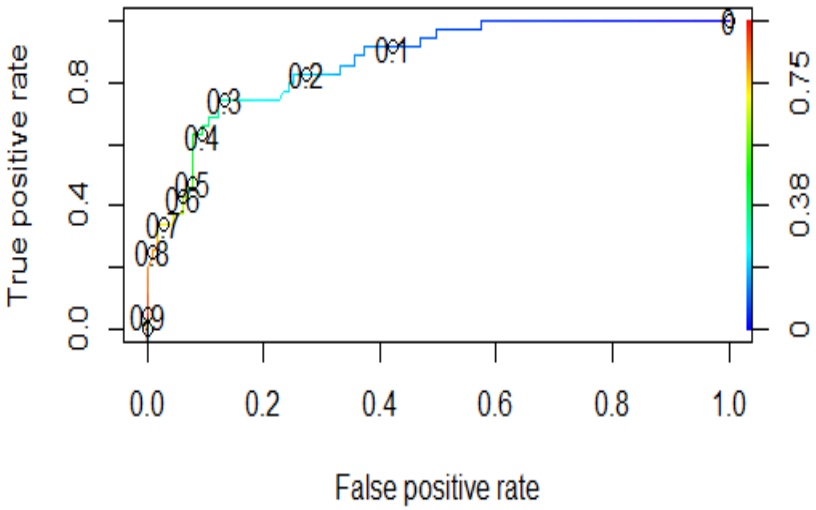

Рис. 5 (Рисунок згенеровано **RStudio**).

На кривій позначено значеннями (0.1, 0.2,…) порогові значення (*Cutt_offo*).

Приклади:

```
> auc <- performance(pred_fit, measure = "auc")
> str(auc)
Formal class 'performance' [package "ROCR"] with 6 slots
  ..@ x.name      : chr "None"
  ..@ y.name      : chr "Area under the ROC curve"
  ..@ alpha.name  : chr "none"
  ..@ x.values    : list()
  ..@ y.values    :List of 1
  .. ..$ : num 0.87
```

Отримуємо площу під ROC кривою **AUC =0.87**
Шукаємо порогове значення (рис.6)

Приклади:

```
> perf3 <- performance(pred_fit, x.measure = "cutoff", measure = "spec")
> perf4 <- performance(pred_fit, x.measure = "cutoff", measure = "sens")
> plot(perf3, col = "red", lwd =2)
> plot(add=T, perf4 , col = "green", lwd =2)
> legend(x = 0.6,y = 0.3, c("spec", "sens"),
+       lty = 1, col =c('red', 'green'), bty = 'n', cex = 1, lwd = 2)
> abline(v= 0.225, lwd = 2)
```

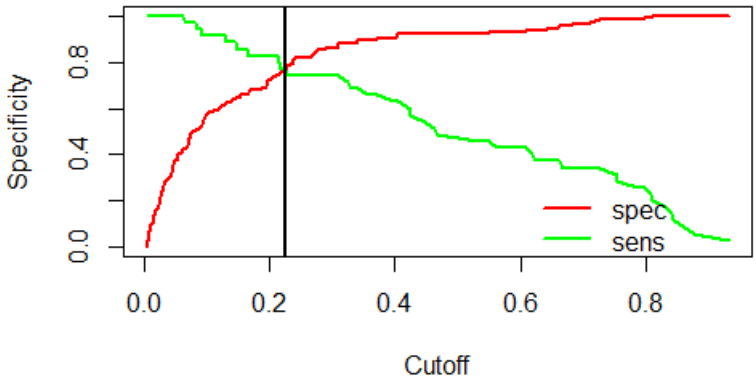

Рис. 6 (Рисунок згенеровано **RStudio**).

x2

Перевіряємо прогноз - змінна **correct**

Приклади:

```
> my_df$pred_resp  <- factor(ifelse(my_df$prob > 0.225, 1, 0), labels = c("N", "Y"))
> my_df$correct  <- ifelse(my_df$pred_resp == my_df$univer, 1, 0)
> my_df
    gender ukr engl math univer        prob pred_resp correct
1      boy  57   52   41      N 0.021834862         N       1
2      boy  44   33   54      N 0.054046191         N       1
3      boy  63   44   47      N 0.071285255         N       1
4      boy  47   52   57      N 0.095796750         N       1
5      boy  50   59   42      N 0.015820126         N       1
```

Визначаємо відсоток збігів:

Приклади:

```
> mean(my_df$correct)
[1] 0.7666667
```

Як бачимо, він достатньо високий.