ROC analysis

Victoria Zaitceva

2024-10-22

library(readxl)  
trauma <- read\_excel("trauma.xlsx")

• id – номер пациента в списке

• Name – имя фамилия пациента

• Sex – пол пациента

• Age – возраст пациента (лет)

• Height – рост пациента (дюймы1)

• Weight – масса тела пациента (фунты2)

• SBP – систолическое артериальное давление при поступлении (мм рт.ст.)

• DBP – диастолическое артериальное давление при поступлении (мм рт.ст.)

• FOUR – балл по шкале комы FOUR при поступлении

• GSC – балл по шкале комы Глазго при поступлении

• Hb – уровень гемоглобина при поступлении (г/дл)

• Death – летальный исход в течение 24 часов (1 – наступил, 0 – нет)

# 1

Дайте описательную статистику для переменных, включённых в датасет. Дополнительно рассчитайте, у какого количества пациентов и в каком проценте случаев у пациентов был снижен уровень гемоглобина? Используйте следующие референтные значения (Мужчины: 13.5–16 г/дл, Женщины: 12–14 г/дл).

trauma <- trauma %>%  
 mutate(Height = str\_replace(Height, "\"", "") %>% as.numeric(),  
 Height = Height \* 2.54, # из дюймов в см  
 Weight = Weight \* 0.453592) # из фунтов в кг  
  
# факторы  
trauma <- trauma %>%  
 mutate(across(c(Sex, Death), ~ as.factor(.x)))  
  
trauma %>%  
 select(-id, -Name) %>%  
 summary()

## Sex Age Height Weight SBP   
## Female:384 Min. :18.00 Min. :162.1 Min. : 60.46 Min. : 90.0   
## Male :640 1st Qu.:29.00 1st Qu.:177.0 1st Qu.: 82.41 1st Qu.:106.0   
## Median :41.00 Median :184.9 Median : 87.95 Median :110.0   
## Mean :40.94 Mean :184.0 Mean : 87.78 Mean :110.8   
## 3rd Qu.:53.00 3rd Qu.:190.8 3rd Qu.: 93.18 3rd Qu.:116.0   
## Max. :64.00 Max. :202.9 Max. :114.80 Max. :134.0   
## DBP FOUR GSC Hb Death   
## Min. : 58.00 Min. : 0.000 Min. : 3.000 Min. : 0.0 0:615   
## 1st Qu.: 78.00 1st Qu.: 7.000 1st Qu.: 6.000 1st Qu.:12.0 1:409   
## Median : 84.00 Median : 9.000 Median : 8.000 Median :13.1   
## Mean : 84.78 Mean : 8.853 Mean : 7.785 Mean :12.8   
## 3rd Qu.: 90.00 3rd Qu.:11.000 3rd Qu.:10.000 3rd Qu.:14.0   
## Max. :110.00 Max. :16.000 Max. :14.000 Max. :16.2

trauma <- trauma %>%  
 mutate(Hb\_low = case\_when(  
 Sex == "Мale" & Hb < 13.5 ~ 1,  
 Sex == "Female" & Hb < 12 ~ 1,  
 TRUE ~ 0  
 ),  
 Hb\_low = as.factor(Hb\_low)  
 )  
  
table(trauma$Hb\_low)

##   
## 0 1   
## 790 234

prop.table(table(trauma$Hb\_low)) %>% round(2)

##   
## 0 1   
## 0.77 0.23

Гемоглобин снижен у 23% пациентов.

# 2

Рассчитайте индекс массы тела у пациентов (кг / м2). Каков был средний (M (SD)) уровень ИМТ у пациентов, включённых в исследование? Какая доля пациентов имела ожирение (ИМТ > 30)?

trauma <- trauma %>%  
 mutate(BMI = Weight / ((Height/100)^2)) # из см в м  
  
  
mean\_sd <- trauma %>%   
 summarise(Mean = mean(BMI, na.rm = TRUE) %>% round(2),  
 SD = sd(BMI, na.rm = TRUE) %>% round(2))   
  
  
cat("Средний ИМТ:", mean\_sd$Mean,  
 "\nCо стандартным отклонением:", mean\_sd$SD,  
 "\nДоля пациентов с ожирением:",   
 prop.table(table(trauma$BMI > 30))[2] %>% round(2) \* 100, "%")

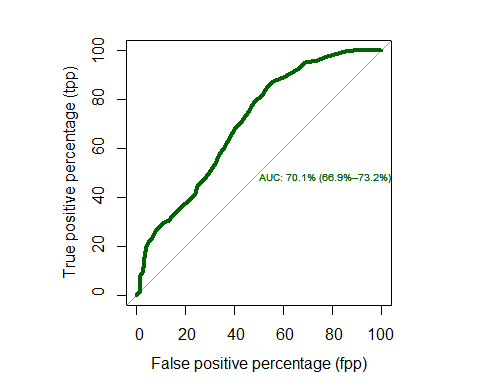
## Средний ИМТ: 26   
## Cо стандартным отклонением: 2.57   
## Доля пациентов с ожирением: 7 %

# 3,4

Как выглядит ROC-кривая для предсказания летального исхода в течение 24 часов по переменной, характеризующей уровень гемоглобина? Постройте график. Чем может быть обусловлена такая форма кривой?

Чему равна площадь под ROC-кривой, которую вы построили в вопросе 3? Чему равен 95% двусторонний ДИ для площади под ROC-кривой, которую вы построили в вопросе 3?

par(pty = "s")  
roc\_curve1 <- roc(Death ~ Hb, data = trauma, plot = TRUE,   
 print.auc = TRUE,  
 print.auc.cex = 0.65,  
 legacy.axes = TRUE, # to make it 1 - specificity  
 percent = TRUE,   
 ci = TRUE,  
 xlab = "False positive percentage (fpp)",  
 ylab = "True positive percentage (tpp)",  
 col = "darkgreen",  
 lwd = 4 # line width  
 )



roc\_curve1

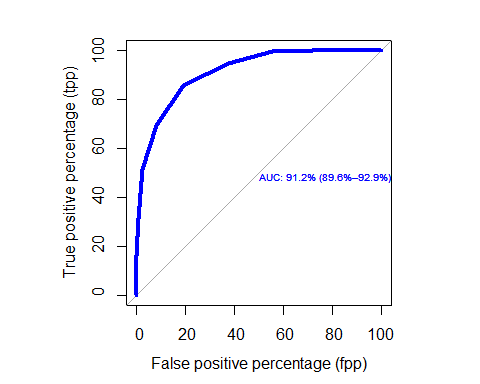
##   
## Call:  
## roc.formula(formula = Death ~ Hb, data = trauma, plot = TRUE, print.auc = TRUE, print.auc.cex = 0.65, legacy.axes = TRUE, percent = TRUE, ci = TRUE, xlab = "False positive percentage (fpp)", ylab = "True positive percentage (tpp)", col = "darkgreen", lwd = 4)  
##   
## Data: Hb in 615 controls (Death 0) > 409 cases (Death 1).  
## Area under the curve: 70.06%  
## 95% CI: 66.91%-73.21% (DeLong)

AUC = 0.7 Наша модель предсказывание летального исхода в течение 24 часов по уровню гемоглобина имеют среднюю точность - 70% случаев смерти могли быть предсказаны по уровню гемоглобина (95% CI: 66.91%-73.21%.)

# 5

Проведите ROC-анализ и определите, какое пороговое значение является оптимальным для предсказания летального исхода в течение 24 часов по шкале комы Глазго. Какой чувствительностью и специфичностью обладает данный порог?

par(pty = "s")  
roc\_curve2 <- roc(Death ~ GSC, data = trauma, plot = TRUE,   
 print.auc = TRUE,  
 print.auc.cex = 0.65,  
 legacy.axes = TRUE, # to make it 1 - specificity  
 percent = TRUE,   
 ci = TRUE,  
 xlab = "False positive percentage (fpp)",  
 ylab = "True positive percentage (tpp)",  
 col = "blue",  
 lwd = 4 # line width  
 )



roc\_curve2

##   
## Call:  
## roc.formula(formula = Death ~ GSC, data = trauma, plot = TRUE, print.auc = TRUE, print.auc.cex = 0.65, legacy.axes = TRUE, percent = TRUE, ci = TRUE, xlab = "False positive percentage (fpp)", ylab = "True positive percentage (tpp)", col = "blue", lwd = 4)  
##   
## Data: GSC in 615 controls (Death 0) > 409 cases (Death 1).  
## Area under the curve: 91.24%  
## 95% CI: 89.59%-92.89% (DeLong)

roc.df <- data.frame(  
 threshold = roc\_curve2$thresholds,  
 tpp = roc\_curve2$sensitivities,  
 fpp = 1 - roc\_curve2$specificities,  
 specificity = roc\_curve2$specificities  
)  
  
head(roc.df, 3)

## threshold tpp fpp specificity  
## 1 Inf 100 1.000000 0.000000  
## 2 13.5 100 -1.926829 2.926829  
## 3 12.5 100 -5.666667 6.666667

tail(roc.df, 3)

## threshold tpp fpp specificity  
## 11 4.5 31.54034 -98.34959 99.34959  
## 12 3.5 16.38142 -99.00000 100.00000  
## 13 -Inf 0.00000 -99.00000 100.00000

Верхний правый угол: когда оба Sensitivity (TPP) и 1 - specificity (FPP) равны 100% (т.е. специфичность равна 100%). Все пациенты с летальным исходом предсказаны правильно, но все пациенты без летального исхода предсказаны неправильно. Порог +бесконечность (бесконечно высокий балл по шкале комы Глазго)

Нижний левый угол: когда оба Sensitivity (TPP) и 1 - specificity (FPP) равны 0% (т.е. специфичность равна 0%). Никакой классификации не происходит.  
Порог -бесконечность (бесконечно низкий балл по шкале комы Глазго)

Если мы хотим найти оптимальный порог, то нам нужно найти точку, где TPP максимально, а FPP минимально. Например, мы хотим, чтобы TPP было больше 60%, а FPP меньше 80%:

roc.df[roc.df$tpp > 60 & roc.df$fpp < 80, ]

## threshold tpp fpp specificity  
## 1 Inf 100.00000 1.000000 0.000000  
## 2 13.5 100.00000 -1.926829 2.926829  
## 3 12.5 100.00000 -5.666667 6.666667  
## 4 11.5 100.00000 -13.796748 14.796748  
## 5 10.5 100.00000 -27.130081 28.130081  
## 6 9.5 99.26650 -43.715447 44.715447  
## 7 8.5 94.62103 -61.439024 62.439024  
## 8 7.5 85.57457 -79.813008 80.813008  
## 9 6.5 69.19315 -90.869919 91.869919

roc\_curve2 %>% coords(x = "best", best.method = "closest.topleft")

## threshold specificity sensitivity  
## 1 7.5 80.81301 85.57457

Таким образом, оптимальный порог для предсказания летального исхода в течение 24 часов по шкале комы Глазго равен 14.5 балла. При данном пороге чувствительность составляет 85.5%, а специфичность - 80.8%.

# 6

Какая из количественных переменных в датасете (включая рассчитанный вами ранее ИМТ) обладает наибольшей площадью под ROC-кривой? Как вы можете интерпретировать это знание? Какая количественная переменная имеет наименьшую площадь?

all\_rocs <- trauma %>%   
 select(-id) %>%  
 select(Death, where(is.numeric)) %>%  
 pivot\_longer(cols = !Death) %>%   
   
 group\_by(name) %>%   
   
 summarise(AUC = roc(Death, value, ci = T)$ci[2] %>% round(3),  
 AUC\_LCL = roc(Death, value, ci = T)$ci[1] %>% round(3), #lower confidence limit  
 AUC\_UCL = roc(Death, value, ci = T)$ci[3] %>% round(3)) #upper confidence limit  
  
all\_rocs

## # A tibble: 9 × 4  
## name AUC AUC\_LCL AUC\_UCL  
## <chr> <dbl> <dbl> <dbl>  
## 1 Age 0.527 0.491 0.563  
## 2 BMI 0.502 0.465 0.538  
## 3 DBP 0.744 0.714 0.774  
## 4 FOUR 0.934 0.92 0.948  
## 5 GSC 0.912 0.896 0.929  
## 6 Hb 0.701 0.669 0.732  
## 7 Height 0.482 0.445 0.518  
## 8 SBP 0.786 0.759 0.814  
## 9 Weight 0.518 0.482 0.554

cat("Наибольшей площадью под ROC-кривой обладает переменная:",   
 all\_rocs %>%  
 filter(AUC == max(AUC)) %>%  
 pull(name),  
 "\nAUC =",  
 all\_rocs %>%   
 filter(AUC == max(AUC)) %>% pull(AUC)  
)

## Наибольшей площадью под ROC-кривой обладает переменная: FOUR   
## AUC = 0.934

Таким образом, балл по шкале комы FOUR при поступлении имеет наибольшей предиктивной способностью (93% случаев будут предсказаны верно).

cat("Нулевой предсказательной мощностью обладают переменные:",   
 all\_rocs %>%  
 filter(AUC\_LCL <= 0.5 & AUC\_UCL >= 0.5) %>%  
 pull(name)  
)

## Нулевой предсказательной мощностью обладают переменные: Age BMI Height Weight