RAkin\_ specifics\_of\_medical\_data

RAkin ALex

2024-10-31

## Загрузка и проверка датасета

trauma\_data <- readxl::read\_excel("data/trauma.xlsx")  
# Изменим единицы измерения роста и масы тела пациента на м и кг.  
### Почему-то версия кода с пайпом не работает  
# trauma\_data %>%  
# mutate(Height\_cm = Height %>% gsub("\"", "") %>% as.numeric())  
   
trauma\_data %>%   
 mutate(Height = as.numeric(gsub("\"", "", Height) ) \* 2.54 / 100,  
 Weight = Weight / 2.2,  
 Sex = as.factor(Sex),  
 Death = as.factor(Death),  
 ) -> trauma\_data  
  
  
# Переменная FOUR может иметь 0 в значении. Однако переменная Hb - не может иметь 0(получается что у пациента совсем не было гемоглобина). Подставим в наблюдения где Hb == 0 значение NA  
trauma\_data$Hb <- na\_if(trauma\_data$Hb, 0)

## Задание 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).

# описательные статистики  
  
skimr::skim(trauma\_data)

Data summary

|  |  |
| --- | --- |
| Name | trauma\_data |
| Number of rows | 1024 |
| Number of columns | 12 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| factor | 2 |
| numeric | 9 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Name | 0 | 1 | 9 | 25 | 0 | 1023 | 0 |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| Sex | 0 | 1 | FALSE | 2 | Mal: 640, Fem: 384 |
| Death | 0 | 1 | FALSE | 2 | 0: 615, 1: 409 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | 0 | 1.00 | 40.94 | 13.66 | 18.00 | 29.00 | 41.00 | 53.00 | 64.00 |
| Height | 0 | 1.00 | 1.84 | 0.08 | 1.62 | 1.77 | 1.85 | 1.91 | 2.03 |
| Weight | 0 | 1.00 | 87.97 | 8.29 | 60.59 | 82.58 | 88.14 | 93.38 | 115.05 |
| SBP | 0 | 1.00 | 110.78 | 7.45 | 90.00 | 106.00 | 110.00 | 116.00 | 134.00 |
| DBP | 0 | 1.00 | 84.78 | 8.42 | 58.00 | 78.00 | 84.00 | 90.00 | 110.00 |
| FOUR | 0 | 1.00 | 8.85 | 3.01 | 0.00 | 7.00 | 9.00 | 11.00 | 16.00 |
| GSC | 0 | 1.00 | 7.79 | 2.69 | 3.00 | 6.00 | 8.00 | 10.00 | 14.00 |
| Hb | 16 | 0.98 | 13.01 | 1.28 | 9.40 | 12.10 | 13.10 | 14.00 | 16.20 |

# Дополнительно рассчитайте у какого количества пациентов и в каком проценте случаев у пациентов был снижен уровень гемоглобина?  
# Для этого введем новую переменную low\_HB и применим функцию case\_when, так как будем задачать несколько условий одновременно.  
  
trauma\_data %>%  
 filter(!is.na(Hb)) %>%  
 mutate(Low\_HB = case\_when(  
 Sex == "Male" & Hb <= 13.5 ~ TRUE,  
 Sex == "Female" & Hb <= 12 ~ TRUE,  
 .default = FALSE  
 )) %>%  
 group\_by(Sex) %>%  
 # Расчитаем общее количество пациентов и процент пациетов с низким гемоглобином  
 summarise(  
 n = n(),  
 count\_lowHb\_patients = sum(Low\_HB),  
 percent\_lowHb\_patients = round(sum(Low\_HB) / n \* 100, 2),  
   
 ) %>%  
 ungroup() %>%  
 # Теперь добавим сторчку "всего"  
 bind\_rows(  
 summarise(  
 . ,  
 Sex = "Total\*",  
 count\_lowHb\_patients = sum(count\_lowHb\_patients),  
 percent\_lowHb\_patients = round(sum(count\_lowHb\_patients) / sum(n) \* 100, 2),  
 n = sum(n)  
 )  
 ) %>%  
 flextable() %>%  
 theme\_box() %>%  
 color(color = "black") %>%  
 align(align = "center", part = "all") %>%   
 add\_footer\_lines("\*16 missing values")

| **Sex** | **n** | **count\_lowHb\_patients** | **percent\_lowHb\_patients** |
| --- | --- | --- | --- |
| Female | 376 | 240 | 63.83 |
| Male | 632 | 259 | 40.98 |
| Total\* | 1,008 | 499 | 49.50 |
| \*16 missing values | | | |

## Задание 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)

# для ответа нужно сделать новую переменную body mass index, BMI  
trauma\_data %>%   
 mutate(BMI = Weight / (Height^2))-> trauma\_data   
tibble(  
 "Mean BMI" = round(mean(trauma\_data$BMI, na.rm = TRUE),2),  
 "Sd BMI" = round(sd(trauma\_data$BMI, na.rm = TRUE),2),  
 "% obesity" = round(sum(trauma\_data$BMI > 30, na.rm = TRUE)/NROW(trauma\_data$BMI)\*100, 2)  
)%>%  
 flextable() %>%  
 theme\_box() %>%  
 color(color = "black") %>%  
 align(align = "center", part = "all") %>%   
 add\_footer\_lines("n patients = 1024")

| **Mean BMI** | **Sd BMI** | **% obesity** |
| --- | --- | --- |
| 26.06 | 2.57 | 6.93 |
| n patients = 1024 | | |

## Задание 3

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?

roc\_curve\_Hb <- roc(Death ~ Hb,   
 data = trauma\_data,  
 ci = TRUE)

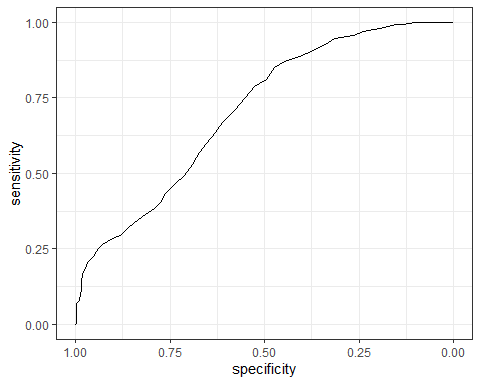
## Setting levels: control = 0, case = 1

## Setting direction: controls > cases

roc\_curve\_Hb

##   
## Call:  
## roc.formula(formula = Death ~ Hb, data = trauma\_data, ci = TRUE)  
##   
## Data: Hb in 605 controls (Death 0) > 403 cases (Death 1).  
## Area under the curve: 0.7078  
## 95% CI: 0.6763-0.7392 (DeLong)

roc\_curve\_Hb %>%   
 ggroc() +   
 theme\_bw()



### ROC- кривая близка к диагонали, а AUC 0.71, что свидетельствует о низком предсказании данной модели.

## Задание 4

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?

tibble(  
 AUC = roc\_curve\_Hb$auc,   
 CI\_min = roc\_curve\_Hb$ci[1],   
 CI\_max = roc\_curve\_Hb$ci[3]  
)

## # A tibble: 1 x 3  
## AUC CI\_min CI\_max  
## <auc> <dbl> <dbl>  
## 1 0.7077702 0.676 0.739

## Задание 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?

roc\_curve\_GSC <- roc(Death ~ GSC,   
 data = trauma\_data,  
 ci = TRUE)

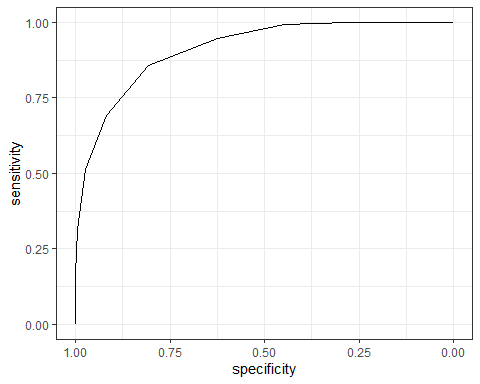
## Setting levels: control = 0, case = 1

## Setting direction: controls > cases

roc\_curve\_GSC

##   
## Call:  
## roc.formula(formula = Death ~ GSC, data = trauma\_data, ci = TRUE)  
##   
## Data: GSC in 615 controls (Death 0) > 409 cases (Death 1).  
## Area under the curve: 0.9124  
## 95% CI: 0.8959-0.9289 (DeLong)

roc\_curve\_GSC %>%   
 ggroc() +   
 theme\_bw()



# Найдем "лучшие" координаты   
  
roc\_curve\_GSC %>% coords(x = "best", best.method = "closest.topleft")

## threshold specificity sensitivity  
## 1 7.5 0.8081301 0.8557457

7.5 оптимальное значение для предсказания летального исхода в течение 24 часов по шкале комы Глазго. (При данном значении чувствительность и специфичной модели будет максимальными)

## Задание 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?

###### Дополнение. На занятии мы ввели новую переменную - среднее АД, давайте введем ее и решение  
  
trauma\_data <- trauma\_data %>%   
 mutate(MeanBP = (DBP + 1/3 \* (SBP - DBP)) %>% round(2))  
  
# Изменим датасет и найдем AUC у всех количественных переменных  
roc\_curve\_all <- trauma\_data %>%   
 select(Death, where(is.numeric) & -c(id)) %>%   
 pivot\_longer(cols = !Death) %>%   
 group\_by(name) %>%   
 # ci[2] - это сам AUC, ci[1] и ci[3] - 95 % ДИ  
 summarise(AUC = roc(Death, value, ci = T)$ci[2] %>% round(3),  
 AUC\_min = roc(Death, value, ci = T)$ci[1] %>% round(3),  
 AUC\_max = roc(Death, value, ci = T)$ci[3] %>% round(3))

roc\_curve\_all

## # A tibble: 10 x 4  
## name AUC AUC\_min AUC\_max  
## <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 0.708 0.676 0.739  
## 7 Height 0.482 0.445 0.518  
## 8 MeanBP 0.765 0.736 0.794  
## 9 SBP 0.786 0.759 0.814  
## 10 Weight 0.518 0.482 0.554

#### теперь найдем переменную с максимальным и миниммаьным AUC  
  
cat("Наибольшей площадью под ROC-кривой обладает переменная:",   
 roc\_curve\_all %>%  
 filter(AUC == max(AUC)) %>%  
 pull(name),   
 ". AUC = ",   
 roc\_curve\_all %>%  
 filter(AUC == max(AUC)) %>%   
 pull(AUC), "\n\n")

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

#### Можно автоматизировать, но надо писть функцию.  
 roc\_curve\_FOUR <- roc(Death ~ FOUR,  
 data = trauma\_data,  
 ci = TRUE)

## Setting levels: control = 0, case = 1  
## Setting direction: controls > cases

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

## threshold specificity sensitivity  
## 1 8.5 0.8308943 0.8948655

cat("Наименьшей площадью под ROC-кривой обладает переменная:",   
 roc\_curve\_all %>%  
 filter(AUC == min(AUC)) %>%  
 pull(name),   
 ". AUC = ",   
 roc\_curve\_all %>%  
 filter(AUC == min(AUC)) %>%   
 pull(AUC))

## Наименьшей площадью под ROC-кривой обладает переменная: Height . AUC = 0.482

Наибольшей площадью под ROC-кривой обладает переменная: FOUR . AUC = 0.934 Возможно это связано с тем , что данная шкала “имеет преимущества перед шкалой комы Глазго, а именно: точнее детализирует неврологический статус, распознает синдром запертого человека, дает оценку рефлексам ствола мозга, дает оценку дыхательному паттерну, выявляет различные стадии дислокации (вклинения) мозга”

Наименьшей площадью под ROC-кривой обладает переменная: Height . AUC = 0.482 (Вряд-ли рост как-либо будет смертность)