Очистка базы данных

Задание 1-2. Анализ данных. Описательная статистика

Залание 3-4. ROC-анализ для предсказания летального исхода в течение 24 часов по переменной, характеризующей уровень гемоглобина

Задание 5. ROC-анализ предсказания летального исхода в течение 24 часов в зависимости от балла по шкале комы Глазго при поступлении

Задание . Анализ площади под ROC-кривой всех количественных данных

ДЗ. Специфика медицинских данных

Code

Oksana Plastinina

2025-11-18

Show

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.2     ✓ tibble    3.3.0
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ purrr     1.1.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

Show

```
##
## Attaching package: 'flextable'
##
## The following object is masked from 'package:purrr':
##
##     compose
```

Show

```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

Show

Show

Очистка базы данных

Анализ значений переменных исходной базы данны

Show

```
## Статистика по всем переменных исходной базы данных
```

Show

```
##        id             Name               Sex                 Age        
##  Min.   :   1.0   Length:1024        Length:1024        Min.   :18.00  
##  1st Qu.: 256.8   Class :character   Class :character   1st Qu.:29.00  
##  Median : 512.5   Mode  :character   Mode  :character   Median :41.00  
##  Mean   : 512.5                                         Mean   :40.94  
##  3rd Qu.: 768.2                                         3rd Qu.:53.00  
##  Max.   :1024.0                                         Max.   :64.00  
##     Height             Weight           SBP             DBP        
##  Length:1024        Min.   :133.3   Min.   : 90.0   Min.   : 58.00  
##  Class :character   1st Qu.:181.7   1st Qu.:106.0   1st Qu.: 78.00  
##  Mode  :character   Median :193.9   Median :110.0   Median : 84.00  
##                     Mean   :193.5   Mean   :110.8   Mean   : 84.78  
##                     3rd Qu.:205.4   3rd Qu.:116.0   3rd Qu.: 90.00  
##                     Max.   :253.1   Max.   :134.0   Max.   :110.00  
##       FOUR             GSC              Hb            Death        
##  Min.   : 0.000   Min.   : 3.000   Min.   : 0.0   Min.   :0.0000  
##  1st Qu.: 7.000   1st Qu.: 6.000   1st Qu.:12.0   1st Qu.:0.0000  
##  Median : 9.000   Median : 8.000   Median :13.1   Median :0.0000  
##  Mean   : 8.853   Mean   : 7.785   Mean   :12.8   Mean   :0.3994  
##  3rd Qu.:11.000   3rd Qu.:10.000   3rd Qu.:14.0   3rd Qu.:1.0000  
##  Max.   :16.000   Max.   :14.000   Max.   :16.2   Max.   :1.0000
```

Show

```
##
## Типы всех переменных исходной базы данных
```

Show

```
## tibble [1,024 × 12] (S3: tbl_df/tbl/data.frame)
##  $ id    : num [1:1024] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Name  : chr [1:1024] "Jecelle Oberly" "Halie Mccone" "Kole Cook" "Patricia Davis"
...
##  $ Sex   : chr [1:1024] "Female" "Female" "Male" "Female" ...
##  $ Age   : num [1:1024] 24 35 39 20 63 60 63 26 20 25 ...
##  $ Height: chr [1:1024] "68\"" "69.8\"" "72.8\"" "70.5\"" ...
##  $ Weight: num [1:1024] 176 184 210 195 169 ...
##  $ SBP   : num [1:1024] 102 116 122 120 112 118 122 114 104 106 ...
##  $ DBP   : num [1:1024] 78 86 102 100 84 88 100 94 84 78 ...
##  $ FOUR  : num [1:1024] 10 9 12 11 9 13 10 11 10 7 ...
##  $ GSC   : num [1:1024] 8 7 12 9 7 11 9 10 9 5 ...
##  $ Hb    : num [1:1024] 11.4 12.2 15 12.2 12.4 15 11.8 14.4 11.4 12.2 ...
##  $ Death : num [1:1024] 1 0 0 0 1 0 0 0 0 1 ...
```

Show

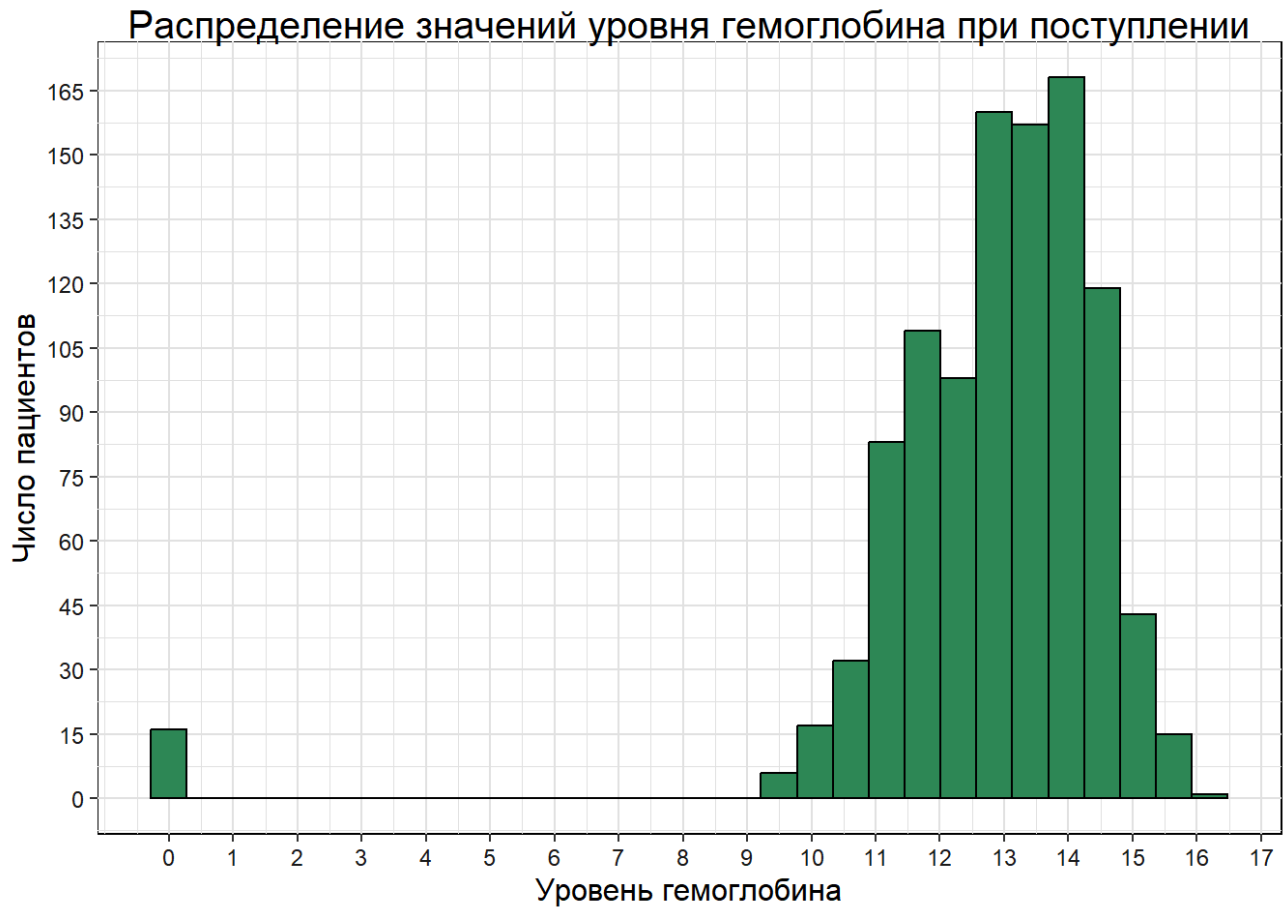
```
##
## Анализ разности САД и ДАД
```

Show

```
##       dif
##  Min.   :20.00
##  1st Qu.:22.00
##  Median :26.00
##  Mean   :25.99
##  3rd Qu.:30.00
##  Max.   :32.00
```

Show

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Значение уровня гемоглобина при поступлении иммет минимальный показатель 0, что не является возможным. Для того, чтобы определить, как распределены значения данного признака в исходных данных построим гистограмму. Из графика видно, что примерно 20 значений соответсвуют 0. Остальные значения больше 9 и меньше 17, что соответсвует возможным значениям.

В типах данных есть ошибки внесения переменной Height из-за чего она определяестя как текстовая переменная. Также Death лучше скорее рассматривать как фактор.

Значения остальных переменных рассматриваем как возможные.

Подготовка базы данных для работы

Show

```
## Статистика по всем переменных базы данных для дальнейшей работы
```

Show

```
##        id             Name               Sex                 Age        
##  Min.   :   1.0   Length:1008        Length:1008        Min.   :18.00  
##  1st Qu.: 257.8   Class :character   Class :character   1st Qu.:29.00  
##  Median : 512.5   Mode  :character   Mode  :character   Median :41.00  
##  Mean   : 513.6                                         Mean   :40.89  
##  3rd Qu.: 769.2                                         3rd Qu.:53.00  
##  Max.   :1024.0                                         Max.   :64.00  
##      Height          Weight           SBP             DBP        
##  Min.   :63.80   Min.   :133.3   Min.   : 90.0   Min.   : 58.00  
##  1st Qu.:69.70   1st Qu.:181.8   1st Qu.:106.0   1st Qu.: 78.00  
##  Median :72.85   Median :194.0   Median :110.0   Median : 84.00  
##  Mean   :72.43   Mean   :193.7   Mean   :110.8   Mean   : 84.78  
##  3rd Qu.:75.10   3rd Qu.:205.5   3rd Qu.:116.0   3rd Qu.: 90.00  
##  Max.   :79.90   Max.   :253.1   Max.   :134.0   Max.   :110.00  
##       FOUR             GSC              Hb         Death      Heightm      
##  Min.   : 0.000   Min.   : 3.00   Min.   : 9.40   0:605   Min.   :1.62  
##  1st Qu.: 7.000   1st Qu.: 6.00   1st Qu.:12.10   1:403   1st Qu.:1.77  
##  Median : 9.000   Median : 8.00   Median :13.10           Median :1.85  
##  Mean   : 8.845   Mean   : 7.78   Mean   :13.01           Mean   :1.84  
##  3rd Qu.:11.000   3rd Qu.:10.00   3rd Qu.:14.00           3rd Qu.:1.91  
##  Max.   :16.000   Max.   :14.00   Max.   :16.20           Max.   :2.03  
##     Weightkg    
##  Min.   : 60.60  
##  1st Qu.: 82.60  
##  Median : 88.20  
##  Mean   : 88.03  
##  3rd Qu.: 93.40  
##  Max.   :115.00
```

Из базы данных удалили пациентов с невозможным значения Hb (выборка в целом большая, а количество таких удаленных пациентов - 16, что не должно повлиять на репрезентативность). Сделали Death фактором. Исправили ошибки внесения переменной Height и сделелали ее числовой. Создали переменные Heightm и Weightkg, соответствующие росту в м и массе в кг.

Задание 1-2. Анализ данных. Описательная статистика

Анализ качественных переменных

Show

Таблица 1 - Описательные статистики качественных признаков

Пол	**Количество пациентов и их % от общего числа**
Исход: не наступил летальный исход в течение 24 часов	
Женщина	220 (21.8%)
Мужчина	385 (38.2%)
Всего с таким	605 (60%)

Всего пациентов в исследовании 1008.
Из них 376 (37.3%) пациентов - женщины, 632 (62.7%) - мужчины.

Пол	Количество пациентов и их % от общего числа
исходом	
Исход: наступил летальный исход в течение 24 часов	
Женщина	156 (15.5%)
Мужчина	247 (24.5%)
Всего с таким исходом	403 (40%)

Всего пациентов в исследовании 1008.
Из них 376 (37.3%) пациентов - женщины, 632 (62.7%) - мужчины.

Из таблицы видно, что большую часть пациентов (62.7%) составляют мужчины. У 60% всех пациентов исследования не наступил летальный исход в течени 24 часов.

Анализ количественных переменных

Сразу введем в базу данных переменную ИМТ для оценки параметров массы и роста

Show
Show
Show

Таблица 2 - Описательные статистики количественный признаков

Параметр	Возраст, год	САД, мм рт.ст.	ДАД, мм рт.ст.	FOUR	GSC	Hb, г/дл	Рост, м	Масса, кг	ИМТ
Исход: не наступил летальный исход в течение 24 часов									
Количество NA	0	0	0	0	0	0	0	0	0
Количество наблюдений	605	605	605	605	605	605	605	605	605
Среднее	40.3	113.7	87.6	10.6	9.3	13.4	1.8	88.3	26.1
Медиана	40	114	88	10	9	13.5	1.85	88.2	26
Стандатрное отклонение	13.61	6.76	7.83	2.21	2.07	1.19	0.08	8.46	2.6
25-75 квартили	29 - 52	108 - 118	82 - 94	9 - 12	8 - 11	12.5 - 14.3	1.78 - 1.91	83 - 93.6	24 - 28
Мин-макс значение	18 - 64	96 - 134	66 - 110	3 - 16	4 - 14	9.4 - 16.2	1.62 - 2.01	60.6 - 115	18 - 35
Исход: наступил летальный исход в течение 24 часов									
Количество	0	0	0	0	0	0	0	0	0

САД - систолическое артериальное давление при поступлении;
ДАД - диастолическое артериальное давление при поступлении;
FOUR – балл по шкале комы FOUR при поступлении;
GSC – балл по шкале комы Глазго при поступлении;
Hb – уровень гемоглобина при поступлении.

Параметр	Возраст, год	САД, мм рт.ст.	ДАД, мм рт.ст.	FOUR	GSC	Hb, г/дл	Рост, м	Масса, кг	ИМТ
NA									
Количество наблюдений	403	403	403	403	403	403	403	403	403
Среднее	41.7	106.4	80.5	6.2	5.5	12.4	1.8	87.7	26.1
Медиана	43	106	80	6	5	12.7	1.85	88.1	26
Стандатрное отклонение	13.68	6.21	7.45	1.99	1.77	1.2	0.08	7.93	2.59
25-75 квартили	30 - 53	102 - 110	76 - 86	5 - 8	4 - 7	11.45 - 13.3	1.76 - 1.9	82.05 - 93.1	24 - 28
Мин-макс значение	18 - 64	90 - 124	58 - 102	0 - 11	3 - 10	9.4 - 14.8	1.64 - 2.03	65.8 - 111.2	19 - 33

САД - систолическое артериальное давление при поступлении;
ДАД - диастолическое артериальное давление при поступлении;
FOUR – балл по шкале комы FOUR при поступлении;
GSC – балл по шкале комы Глазго при поступлении;
Hb – уровень гемоглобина при поступлении.

Из таблицы можно сделать следующие выводы: - база данных не содержит пропущенных значений; - количество пациентов исследования с летальным исходом в течение 24 часов меньше; - перменная возраст и ИМТ примерно совпадают в двух группах; - значения САД и ДАД у пациентов без летального исхода несколько выше; - баллы по шкале комы FOUR и Глазго у пациентов без летального исхода выше, что согласуется с концепцией этих тестов; - значения уровня гемоглобина у пациентов без летального исхода несколько выше;

Рассмотрим подробнее пациентов с анемией

Для определения состояния анемии использовалии слеюущие референтные значения: Мужчины - 13.5–16 г/дл, Женщины - 12–14 г/дл.

Show

Таблица 3 - Количество пациентов с анемией и их соотношения в разных группах

Пол	Количество пациентов	Количество пациентов с анемией	% числа пациентов с анемией в этой группе среди этого пола	% числа пациентов с анемией среди всех пациентов	% числа пациентов с анемией среди всех пациентов с анемией
Исход: не наступил летальный исход в течение 24 часов					
Женщина	220	95	25%	9.4%	20.7%
Мужчина	385	71	11%	7%	15.5%
Исход: наступил летальный исход в течение 24 часов					
Женщина	156	131	35%	13%	28.6%
Мужчина	247	161	25%	16%	35.2%

Пол	Количество пациентов	Количество пациентов с анемией	% числа пациентов с анемией в этой группе среди этого пола	% числа пациентов с анемией среди всех пациентов	% числа пациентов с анемией среди всех пациентов с анемией

Количество пациентов с анемией - 458, что составляет 45.4% от общего числа пациентов.
Референтные значения: Мужчины - 13.5–16 г/дл, Женщины - 12–14 г/дл.

45.4% пациентов страдают анемией, женщины и мужчины составляют примерно равные доли этой группы. У большей часть пациентов с анемией (63.8%) наступил летальный исход.

Анемия в нашем исследовании втречается чаще у женщин, чем у мужчин, 60% и 36% их них имеют сниженный гемоглобин соответственно. Большая половина женщин и мужчин с анемией имели летальный исход в результате ЧМТ в течение 24 ч, 58% и 69% соответственно.

Рассмотрим подробнее пациентов с высоким ИМТ (>30)

Show

Таблица 4 - Количество пациентов с высоким ИМТ (>30) и их соотношение в группах с разным исходом

Летальный исход в течение 24 ч	Количество пациентов	Количество пациентов с высоким ИМТ (% от всех пациентов с высоким ИМТ)	% пациентов с высоким ИМТ в этой группе
Не наступил	605	33 (64.7%)	5.5 %
Наступил	403	18 (35.3%)	4.5 %

Show

```
## Средний уровень ИМТ пациентов, включенных в исследование - 26 (+/-2.6). Из таблицы 1 видно, что это среднее совпадает в двух группах исхода.
## Количество пациентов с ожирением (ИМТ>30) - 51, что составляет 5.1% от общего числа пациентов, включенных в исследование. При этом в группе с летальным исходом таких пациентов меньше, 18 (35.3%), что составляет 4.5 % этой группы. В группе без летального исхода - 33 (64.7%) таких пациентов, что составляет 5.5 % этой группы. Т.к. доли пациентов среди групп исхода примерно равны, данный фактор скорее всего не влияет на исход.
```

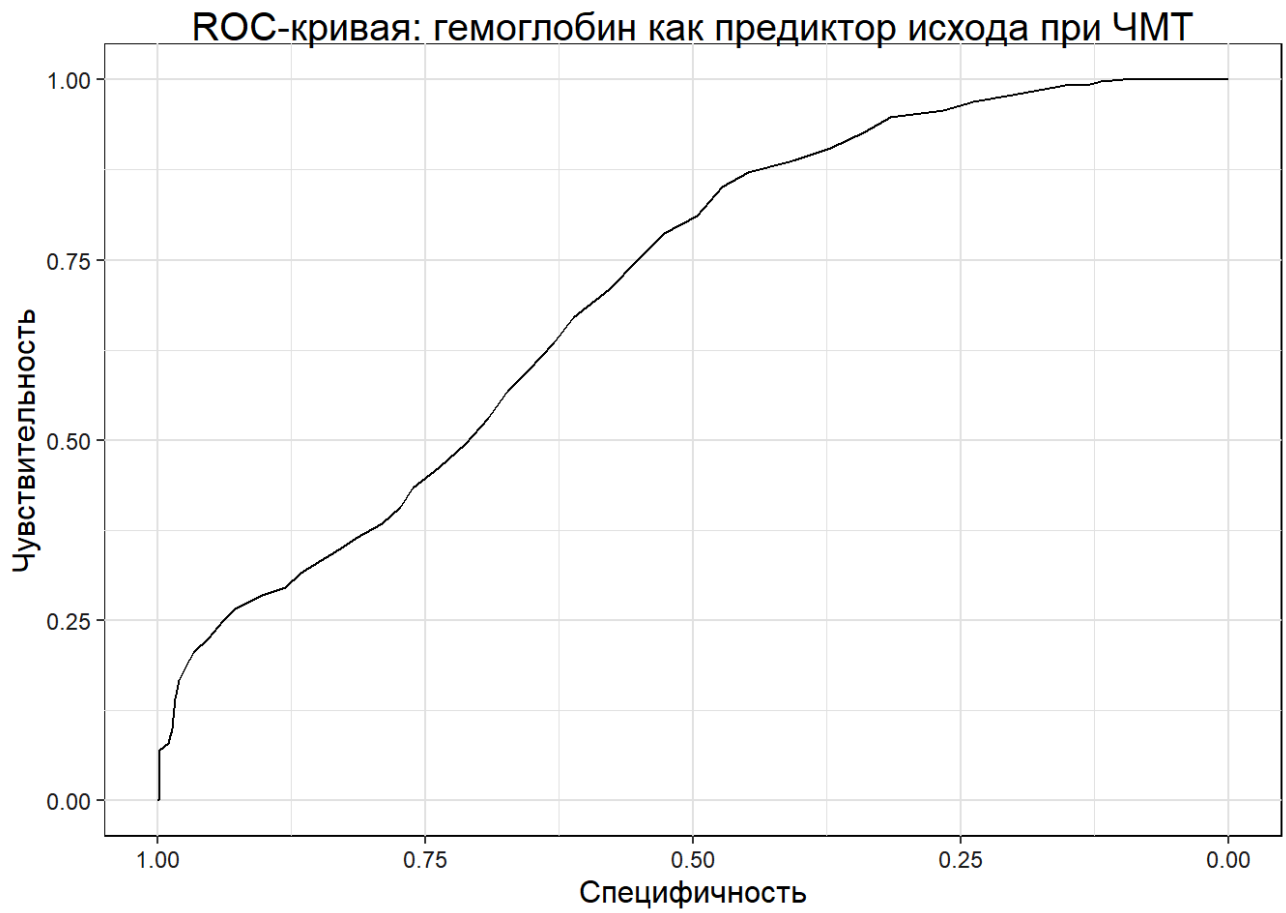
Залание 3-4. ROC-анализ для предсказания летального исхода в течение 24 часов по переменной, характеризующей уровень гемоглобина

Залание 3. Анализ ROC-кривой

Show

```
## Setting levels: control = 1, case = 2
```

```
## Setting direction: controls > cases
```



ROC-кривая выгнута в сторону верхнего левого угла, который соответствует 100% чувствительности и специфичности, что свидетельствует о возможности предсказания летального схода в течение 24 часов (после госпитализации) по причине получения черепно-мозговой травмы на основании уровня гемоглобина. Такая ROC-кривая согласуется с выводами, сделанными по таблице 3 с соотношением пациентов с анемией в разных группах исхода. Здесь также можно выделить два небольших пика, возможно соответствующих двум разных пороговым значениям уровня гемоглобина у мужчин и женщин, что скорее всего может влиять на точность теста при обобщении результатов на мужчин и женщин.

Задание 4. Анализ площади под ROC-кривой

Show

```
## Setting levels: control = 1, case = 2
```

```
## Setting direction: controls > cases
```

```
##
## Call:
## roc.default(response = as.numeric(trauma$Death), predictor = trauma$Hb,     ci = TRU
E)
##
## Data: trauma$Hb in 605 controls (as.numeric(trauma$Death) 1) > 403 cases (as.numeric
(trauma$Death) 2).
## Area under the curve: 0.7078
## 95% CI: 0.6763-0.7392 (DeLong)
```

Согласно результатам ROC-анализа, гемоглобин является статистически значимым предстказателем летального исхода в течение 24 часов (после госпитализации) по причине получения черепно-мозговой травмы. Площадь под ROC-кривой - значение AUC равно 0.71, 95% ДИ не включает значение 0.5 [0.68-0.74], соответствущее случайной диагностике исхода.

Направление связи controls (случаи НЕ наступления летального исхода) > cases (случаи наступления летального исхода), связывает состояние анемии и повышенного риска летального исхода, что мы предполагали из таблицы 3.

Задание 5. ROC-анализ предсказания летального исхода в течение 24 часов в зависимости от балла по шкале комы Глазго при поступлении

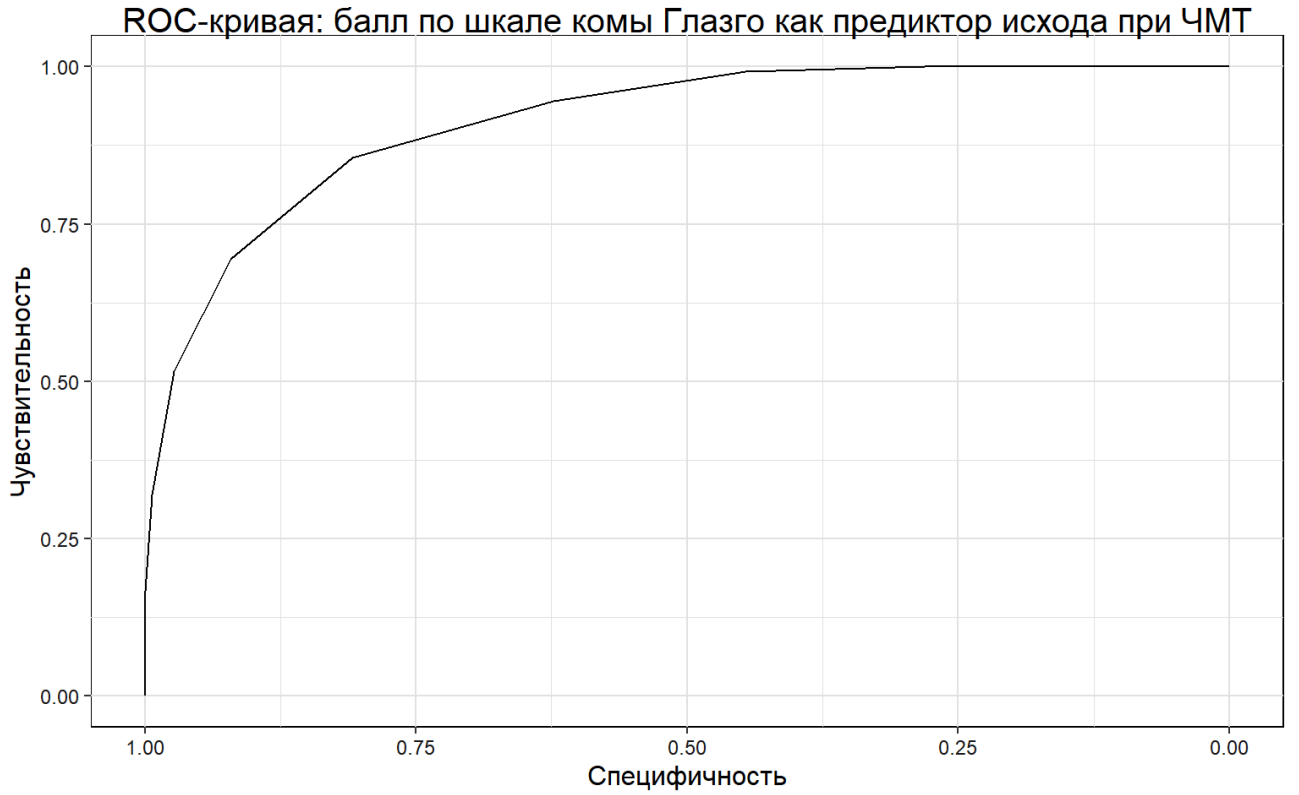
Анализ ROC-кривой и площади под ROC-кривой

Show

```
## Setting levels: control = 1, case = 2
```

```
## Setting direction: controls > cases
```

Show



Show

```
##
## Call:
## roc.default(response = as.numeric(trauma$Death), predictor = trauma$GSC,     ci = TR
UE)
##
## Data: trauma$GSC in 605 controls (as.numeric(trauma$Death) 1) > 403 cases (as.numeri
c(trauma$Death) 2).
## Area under the curve: 0.9129
## 95% CI: 0.8963-0.9295 (DeLong)
```

Исходя из визуализации ROC-кривой и результатов ROC-анализа, балл по шкале комы Глазго является статистически значимым предсказателем летального исхода в течение 24 часов (после госпитализации) по причине получения черепно-мозговой травмы. Площадь под ROC-кривой - значение AUC равно 0.91, 95% ДИ не включает значение 0.5 [0.896-0.929], соответствущюее случайной диагностике исхода.

Направление связи controls (случаи НЕ наступления летального исхода) > cases (случаи наступления летального исхода), связывает баллы по шкале комы Глазго и вероятности выживания. Данный результат сопоставим с концепцией шкалы комы Глазго.

Анализ оптимального порогового значения для предсказания летального исхода в течение 24 часов по шкале комы Глазго

Show

Таблица 5 - Оптимальное пороговое значение для предсказания летального исхода по шкале комы Глазго, определенное методом ближайшей точки к левому краю

Оптимальное порогове значение	Специфичность	Чувствительность
7.5	0.8082645	0.8560794

Оптимальным для предсказания летального исхода в течение 24 часов (после госпитализации) по причине получения черепно-мозговой травмы на основании балла по шкале комы Глазго является пороговое значение равное 7.5. При таком пороговом значении (<7.5) тест будет обладает:

86% чувствительностью (т.е. в 86% случаев верно определяет случаи наступления летального исхода в течение 24 часов, в 14% случаев - получим ложноотрицательный результат)

81% специфичностью (т.е. 81% случаев НЕ наступления летального исхода в течение 24 часов мы определим верно, в 19% случаев - получим ложноположительный результат).

Задание . Анализ площади под ROC-кривой всех количественных данных

Show

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

```
## Setting direction: controls < cases
```

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

```
## Setting direction: controls < cases
```

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

```
## Setting direction: controls > cases
```

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

```
## Setting direction: controls > cases
```

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

```
## Setting direction: controls > cases
```

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

```
## Setting direction: controls > cases
```

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

```
## Setting direction: controls > cases
```

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

```
## Setting direction: controls < cases
```

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

```
## Setting direction: controls < cases
```

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

```
## Setting direction: controls > cases
```

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

```
## Setting direction: controls > cases
```

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

```
## Setting direction: controls > cases
```

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

```
## Setting direction: controls > cases
```

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

```
## Setting direction: controls > cases
```

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

```
## Setting direction: controls < cases
```

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

```
## Setting direction: controls < cases
```

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

```
## Setting direction: controls > cases
```

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

```
## Setting direction: controls > cases
```

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

```
## Setting direction: controls > cases
```

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

```
## Setting direction: controls > cases
```

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

```
## Setting direction: controls > cases
```

Show

Таблица 6 - Значения AUC для всех количественных признаков

Переменная	AUC	AUC_LCL	AUC_UCL

Переменная	AUC	AUC_LCL	AUC_UCL
ИМТ	0.499	0.463	0.535
Возраст	0.528	0.492	0.565
Уроверь гемоглобина	0.708	0.676	0.739
ДАД	0.742	0.712	0.772
САД	0.784	0.756	0.812
Балл по шкале комы Глазго	0.913	0.896	0.929
Балл по шкале комы FOUR	0.934	0.920	0.948

Наименьшие площади под кривой (значения AUC), соответсвующие случайному оперделению признака 0.49 и 0.53, имеют переменные возраста и ИМТ, их 95% ДИ включают значение 0.5. Данные переменные имеют похожие статистики у двух групп исхода (то, что мы смотрели выше в таблице 2: среднее, медианы и др.) и не подходят для разделения групп и предсказания исхода.

Наилучшей предсказательной силой обладают баллы комы по шкале Глазго и FOUR, их значения AUC соответственно 0.91 и 0.93 с узкими 95% ДИ близкими к 1, что подтверждает их потенциальную состоятельность в оценке.

Остальные переменные (уровень гемоглобина, САД и ДАД) имеют промежуточные, но высокие значения AUC, 95% ДИ которых не включают 0.5, что показывает потенциально возможное использование данных парамнтров в оценке исхода.