Step1 Team Project Multivariate Analysis

Adrian White, Cesar Conejo, Xavier Bryant

12/6/2020

Introduction data set

We have selected the CRASH-2 data set provided by Vanderbilt School of Biostatistics for our project. It describes the outcome of a randomized controlled trial and economic valuation of the effects of tranexamic acid on death, vascular occlusive events, and transfusion requirement in bleeding trauma patients. Tranexamic acid reduces bleeding in trauma patients undergoing surgery but is an expensive treatment option. The trial's objective was to assess the effects and cost-effectiveness of an early administration of this medication.

Participants of the study were adults with, or at risk of, significant bleeding within 8 hours of injury. Sample randomization was determined by the allocation of an eight-digit sequence randomly generated by a computer. Patients and staff were masked to the treatment allocation of the tranexamic acid.

We have adjusted the original data set to remove some variables that were not relevant to our investigation. We have removed variables regarding the exact surgical procedures administered to patients, various IDs, and details on the patient outcome. We removed the health outcome columns because of complications regarding missing data, where the boolean structure of the columns relating to specific outcomes, like stroke or pulmonary embolism, left a large number of cases with missing values. Instead, we added a boolean variable for a general outcome of survival to assess the efficacy of the procedure, rather than looking at particular health outcomes in post-surgery for living patients.

We will be using variables regarding the sex, age, and injury of the patient as well as certain biometrics, like blood pressure, respiratory and heart rates, details on surgical blood transfusion, and a boolean variable on the survival of the patient. Our selection provides us with a balance of continuous and categorical variables, many of which are boolean, with minimal complications due to missing data. In summary, the data set consist of $n = 9497$ observations, with 11 columns, which $p = 8$ are quantitative and 3 are qualitative.

Moreover, the normal ranges of the biometric measurements are also added, in order to have a point of comparison with the observations present in the data sey and in this way determine if they are abnormal with respect to the normal metrics.

Summary variables in the data set

The variables in this dataset are the following:

1. sex: (Boolean) The sex of the patient (Male/Female)
2. age : (Numerical) Age of the patient(Years)
3. injurytime: (Numerical) Hours since injury (Hours)
4. injurytype: (Categorical) Type of injury {Blunt, Penetrating, Blunt and Penetrating}
5. sbp: (Numerical) Systolic Blood Pressure (mmHg). Normal range for adults at rest: less 120 mmHg.
6. rr: (Numerical) Respiratory Rate (breaths per minute). Normal range for adults at rest: 12 - 20 breaths per minute.
7. cc: (Numerical) Central Capillary Refille Time (seconds). Normal range for adults at rest. Less 3 seconds
8. hr: (Numerical) Heart Rate (beats per minute). Normal range for adults at rest: 60 - 100 bpm.
9. ndaysicu: (Numerical) Number of days in ICU (days)
10. ncell: (Numerical) Number of Units of Red Call Products Transfused.
11. Death: (Boolean) Indicator if the patient survived after the procedure

A summary of the data type is the following:

variable	type_variable	sub_type_variable
sex	Qualitative	Nominal
age	Quantitative	Continuous
injurytime	Quantitative	Continuous
injurytype	Qualitative	Nominal
sbp	Quantitative	Continuous
rr	Quantitative	Continuous
cc	Quantitative	Continuous
hr	Quantitative	Continuous
ndaysicu	Quantitative	Discrete
ncell	Quantitative	Continuous
death	Qualitative	Nominal

A review of the structure of the dataset is the following:

```
## 'data.frame':    9497 obs. of  11 variables:
##  $ sex       : Factor w/ 2 levels "male","female": 1 1 1 1 1 1 1 1 1 2 ...
##  $ age       : int  50 30 40 19 27 16 29 41 56 37 ...
##  $ injurytime: num  1 1 2 3 0.5 1 1 0.5 0.5 8 ...
##  $ injurytype: Factor w/ 3 levels "blunt","penetrating",..: 1 1 2 2 2 2 1 2 1 2 ...
##  $ sbp       : int  75 70 60 90 90 90 116 120 60 104 ...
##  $ rr        : int  28 26 20 30 26 28 15 15 9 23 ...
##  $ cc        : int  5 6 5 5 5 2 3 3 3 5 ...
##  $ hr        : int  120 130 120 90 96 118 118 70 100 92 ...
##  $ ndaysicu  : num  0 6 2 9 7 0 7 7 23 2 ...
##  $ ncell     : num  1 2 4 2 1 1 16 8 4 4 ...
##  $ death     : Factor w/ 2 levels "0","1": 2 1 2 2 1 1 1 1 1 1 ...
```

A summary of the values in the data set are:

```
##      sex            age          injurytime                      injurytype
##  male  :7906   Min.   :14.00   Min.   : 0.100   blunt                  :5211
##  female:1591   1st Qu.:24.00   1st Qu.: 1.000   penetrating            :2937
##                Median :31.00   Median : 3.000   blunt and penetrating:1349
##                Mean   :34.66   Mean   : 3.094
##                3rd Qu.:43.00   3rd Qu.: 4.500
##                Max.   :96.00   Max.   :48.000
##       sbp               rr              cc               hr
##  Min.   :  4.00   Min.   : 2.00   Min.   : 1.000   Min.   :  3.0
##  1st Qu.: 80.00   1st Qu.:20.00   1st Qu.: 2.000   1st Qu.: 96.0
##  Median : 90.00   Median :22.00   Median : 3.000   Median :110.0
##  Mean   : 93.13   Mean   :23.46   Mean   : 3.438   Mean   :108.1
##  3rd Qu.:104.00   3rd Qu.:28.00   3rd Qu.: 4.000   3rd Qu.:120.0
##  Max.   :225.00   Max.   :91.00   Max.   :20.000   Max.   :220.0
##     ndaysicu          ncell        death
##  Min.   : 0.000   Min.   : 0.000   0:7672
##  1st Qu.: 0.000   1st Qu.: 2.000   1:1825
##  Median : 1.000   Median : 3.000
##  Mean   : 4.137   Mean   : 3.912
##  3rd Qu.: 5.000   3rd Qu.: 5.000
##  Max.   :58.000   Max.   :60.000
```

Finally, the list of different values by column is the following:

Table 2: Count of distinct values of each variable

sex	age	injurytime	injurytype	sbp	rr	cc	hr	ndaysicu	ncell	death
2	81	78	3	153	58	16	154	47	47	2

Visual Analysis

Univariate Analysis

First, we will review the distribution of the variables involve in the dataset

In the case of age, the figure 1 reflects how this variable appears to be largely weighted to the left, with lower ages featuring more frequently than those that are greater, possibly reflecting that younger people often take more risk and work higher at-risk occupations, raising their chance of experiencing trauma involving bleeding.

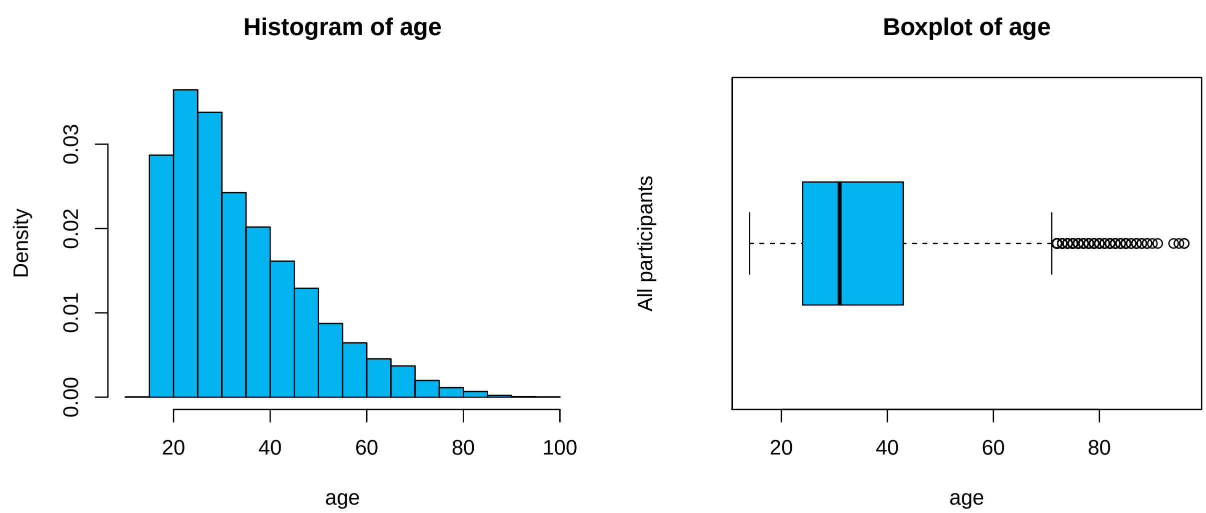


Figure 1: Distribution of age variable

The figure 2 show the distribution of the variable Injury time. We can see how this variable is highly positive skewed with almost all values falling below ten minutes since the injury was experienced. This is likely due to the fact that in cases of serious injury victims are brought to the hospital quite quickly. Then, we apply a *log* transformation `log_injurytime = log(injurytime + 1)` to this variables as we can see in the below of figure 2.

For *sbp* (Systolic Blood Pressure), the distribution is a fairly centrally balanced distribution around 90 mmHg. This is logical as a sample of biological characteristics observed in a population is likely to have most people around the mean and then a reasonably tight distribution of those who differ, similar to that of other biological features like height. Furthermore, most people are fairly young in the sample and therefore would have rates that deviant less from the norm, at a healthy level. The distribution is given by figure 3.

In the case of *rr* (Respiratory Rate) appears, similar to sbp, resembling a moderately balanced distribution around 22 respirations per minute, although is weighted more to the right. The distribution of this variable is showed in figure 4. Taking a *log* transformation `log_rr = log(rr)`, we have the new distribution in below part of figure 4.

In the case of *hr* (Heart rate), figure 5 show that this distribution seems fairly balanced at around 110, similar to the variables above, like *sbp* and *rr*.

For *cc* (Central capillary) refill has 75% of the observations below of 4 as we can appreciate in figure 6. However, the distribution is right-skewed. As a result, we apply a *log* transformation `log_cc = log(cc)` that is given in the below part figure 6.

The figure 7 shows the distribution of the *ndaysicu:* variable. In this case, the distribution is heavily weighted to the left and right-skewed. Most patients it seems, with injuries at high risk of bleeding, do not often need

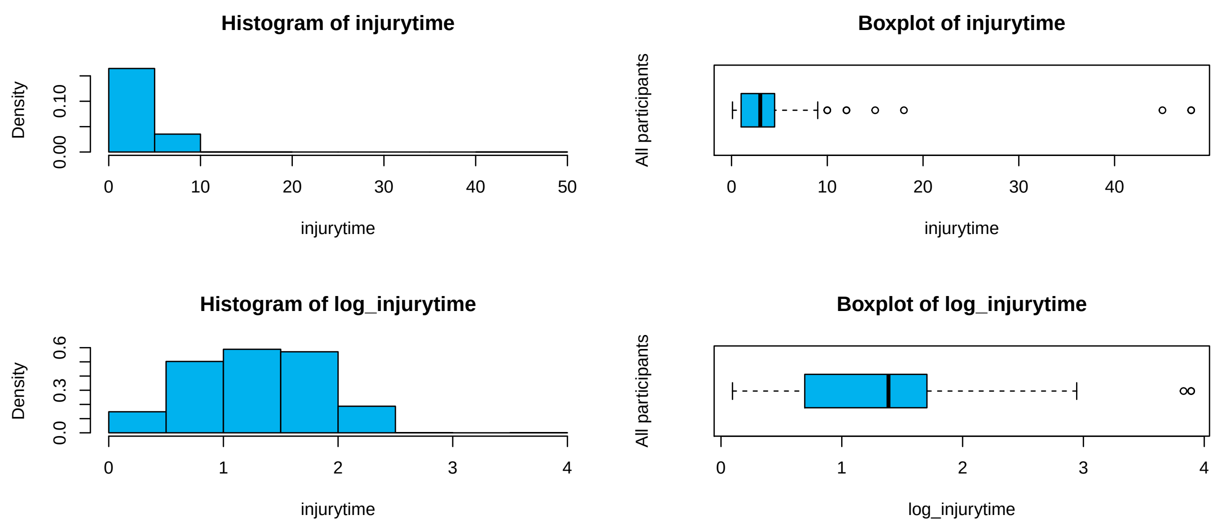


Figure 2: Distribution of injurytime variable and log of injurytime

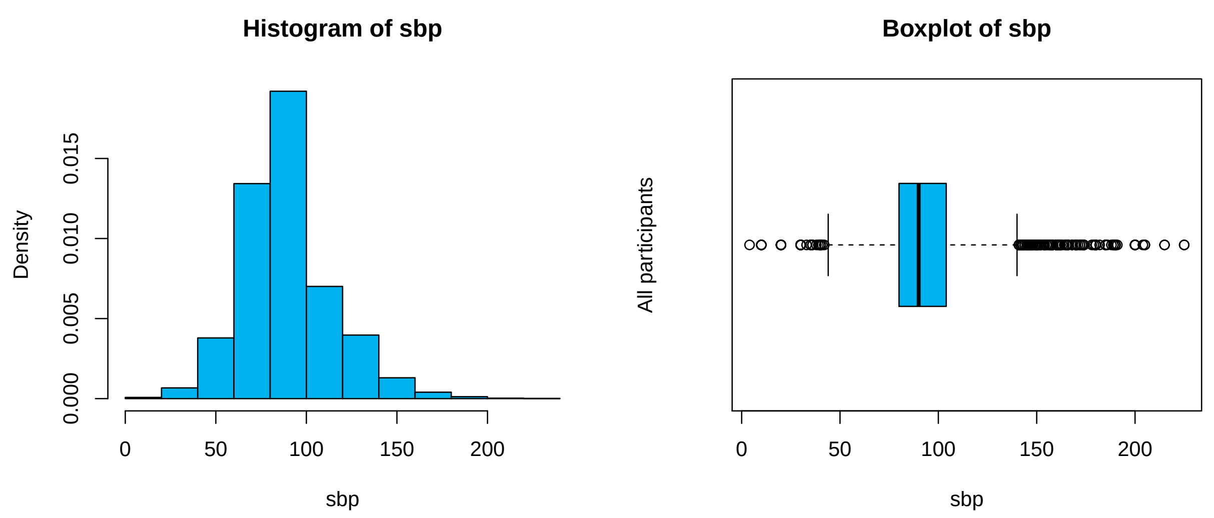


Figure 3: Distribution of sbp (Systolic Blood Pressure)

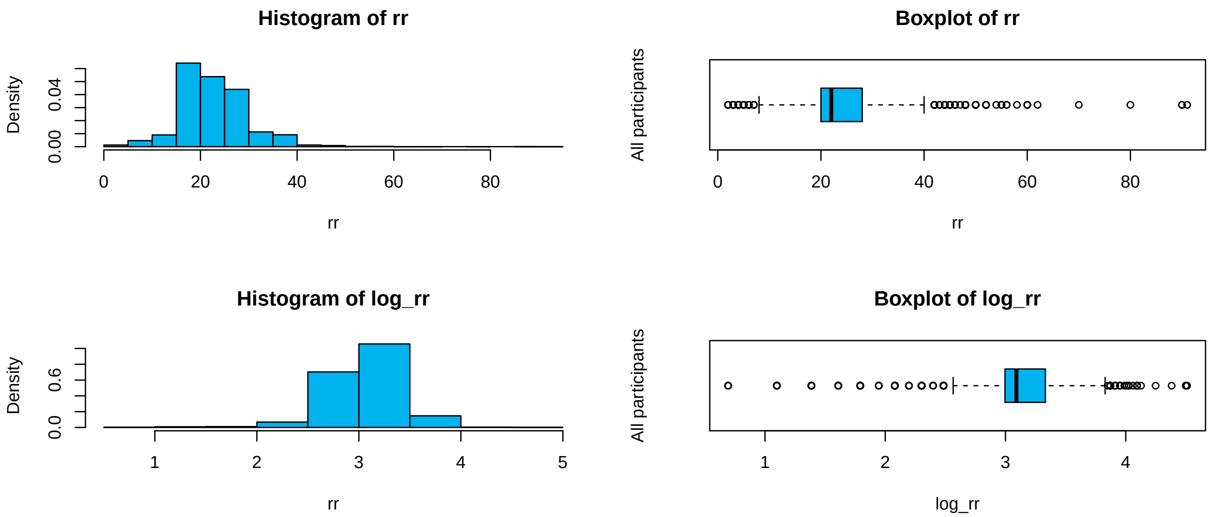


Figure 4: Distribution of rr (Respiratory Rate) and logrr

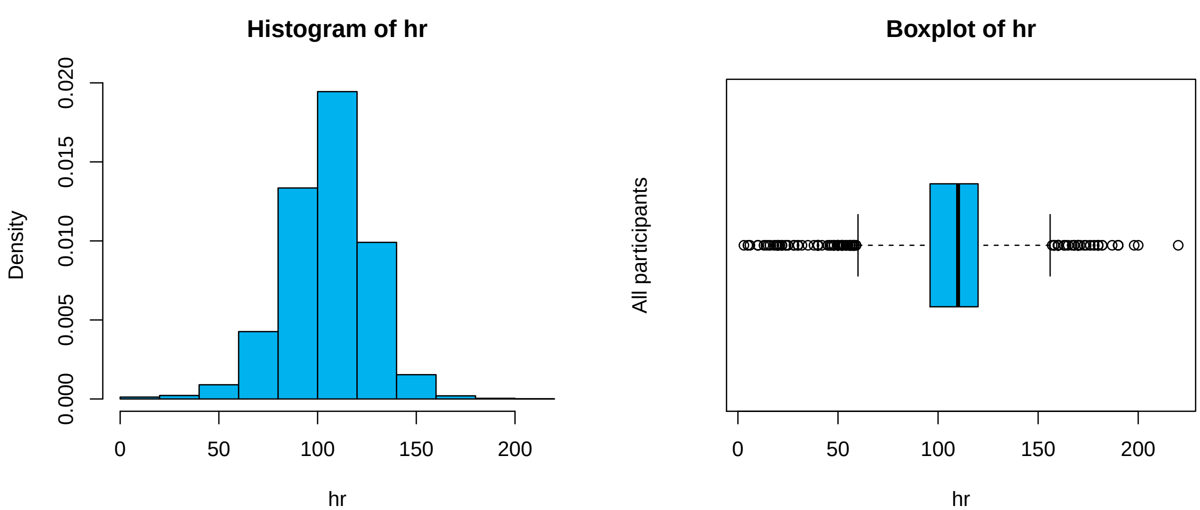


Figure 5: Distribution of hh (hearth Rate)

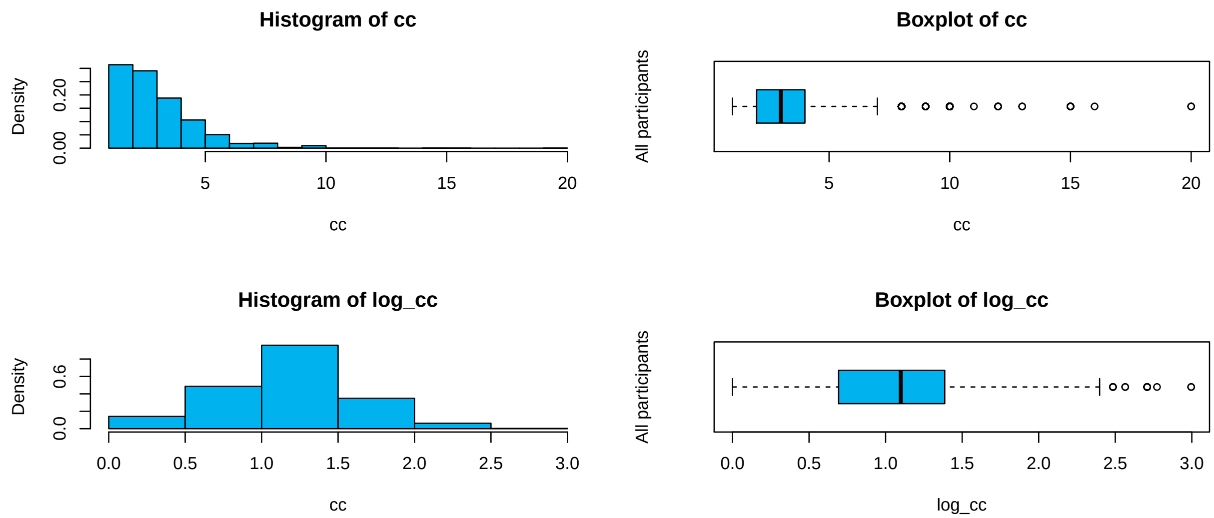


Figure 6: Distribution of cc (Central capillary) and log transformation

to remain in the hospital for long. The transformed distribution `log_ndaysicu = log(ndaysicu + 1)` is given in the below part of figure 7.

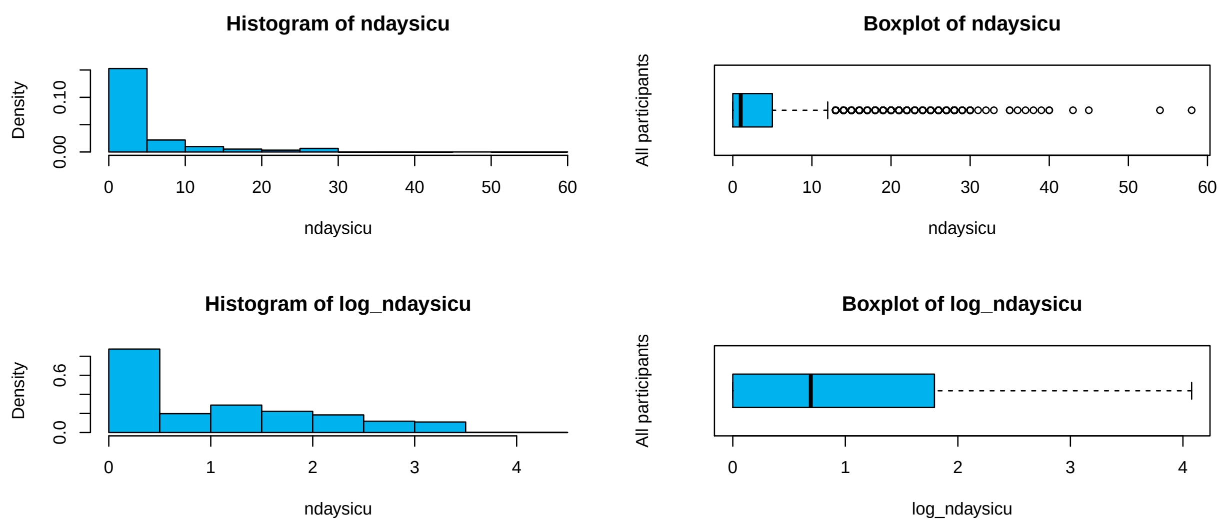


Figure 7: Distribution of ndaysicu and log transformation

Finally, for the *ncell* distribution is weighted to the left with a median of 3 as we can appreciate in the figure 8. The conclusion of this it that with many patients, only needing a small number of or zero units of red cell products transfused. Due to this variable is highly right skewed, we apply the *log* transformation `log_ncell = log(ncell + 1)` of the figure 8 in their below part of the figure.

For the categorical variables, we will focus on the distributions of deaths. The figure 9, shows that approximately for each death, 4 people survive. In the context of this problem, if we have an unbalanced proportion of people that survive can be considered as a sign that the drugs works.

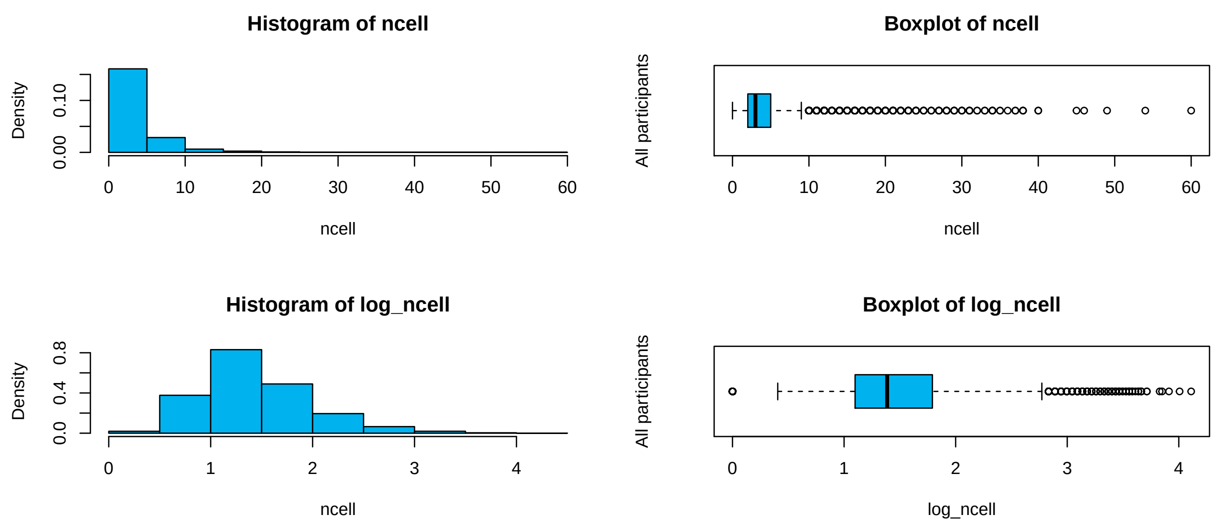


Figure 8: Distribution of ncell

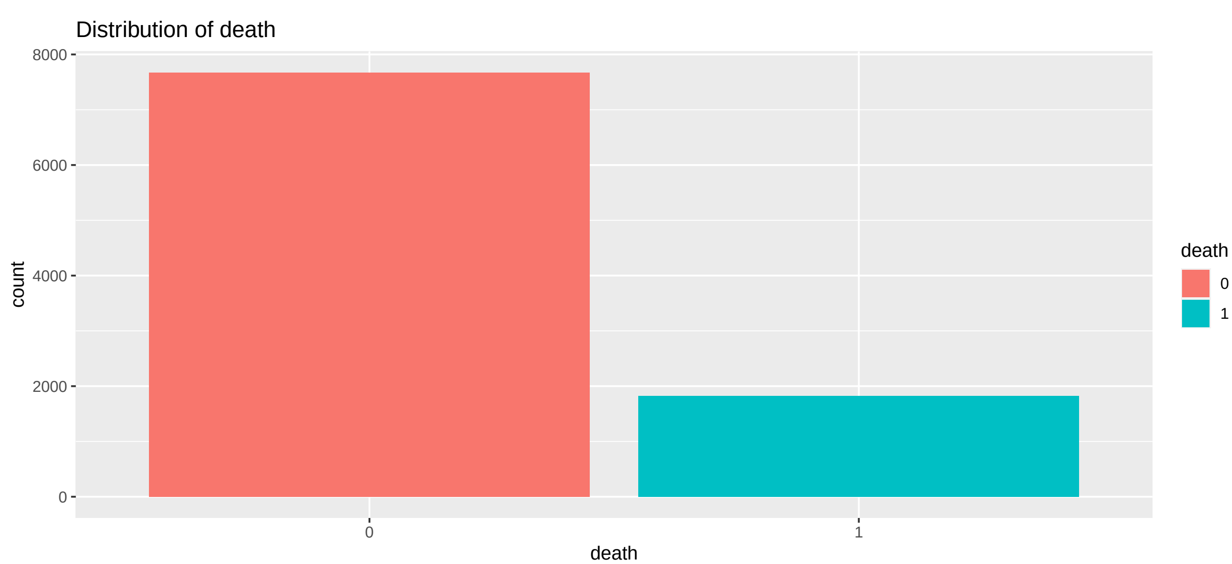


Figure 9: Distribution of deaths

Univariate Analysis by death - survival patients

On the other hand, we can study some relations of the quantitative variables in terms of the categorical variable death. Figure 10 . . .

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

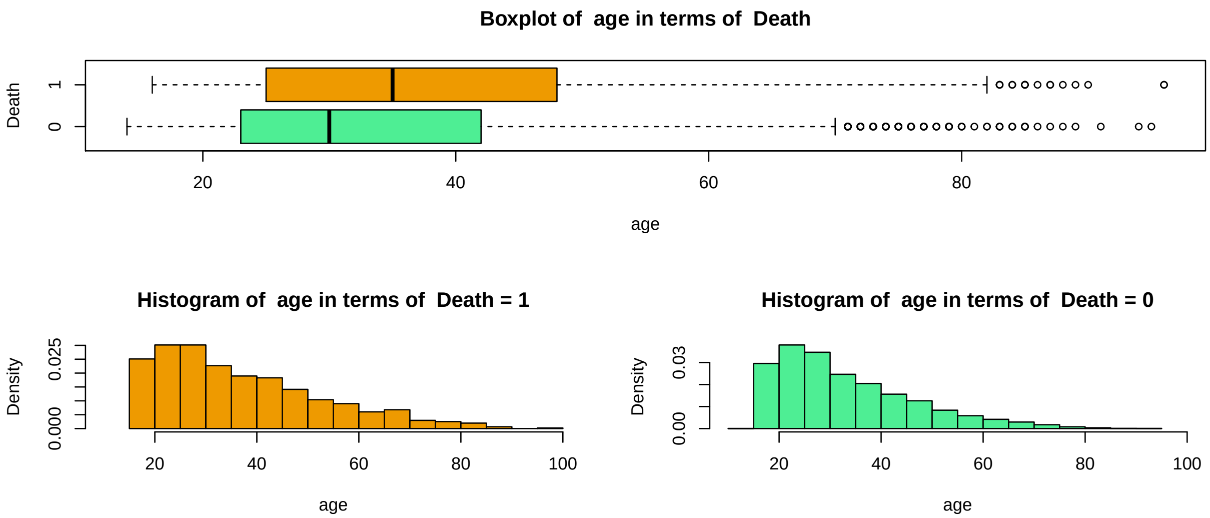


Figure 10: Distribution of age in terms of death

Figure 11 . . . .

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Figure 12 . . . .

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Figure 11 . . .

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Figure 14 . . .

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Figure 15 . . .

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

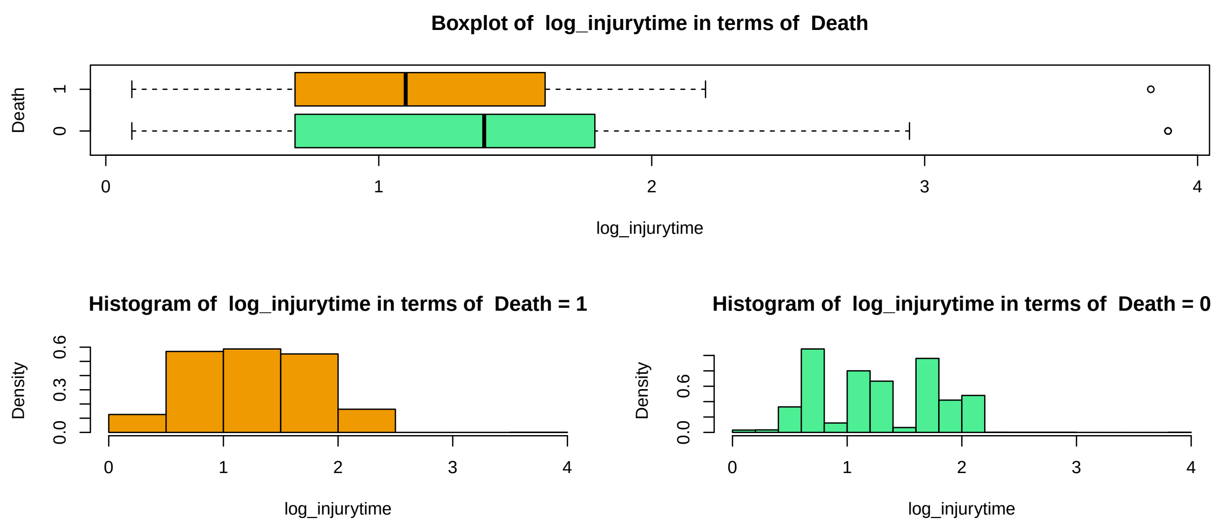


Figure 11: Distribution of the log injurytime in terms of death

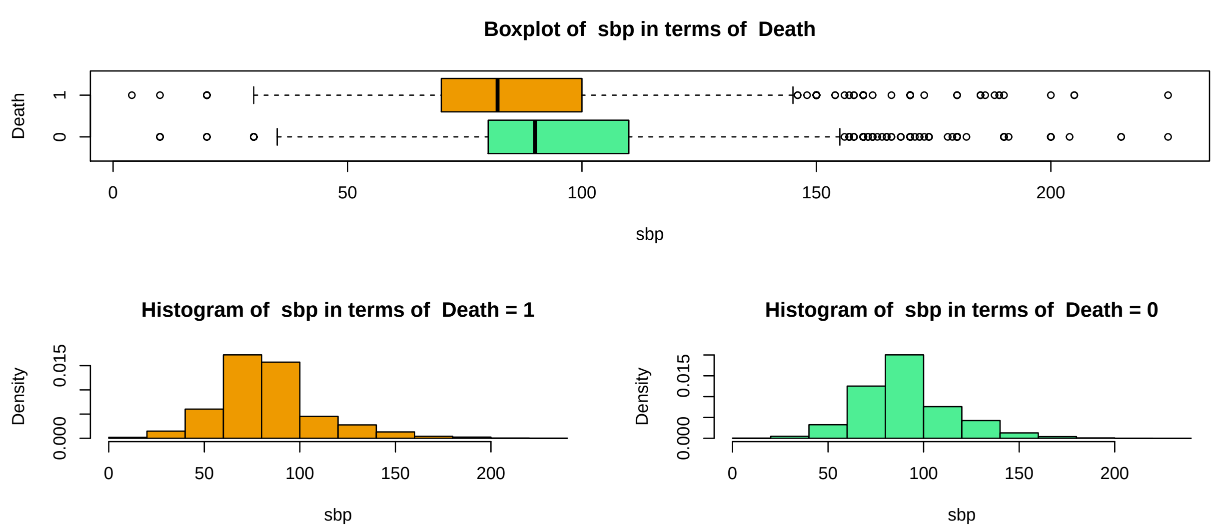


Figure 12: Distribution of sbp (Systolic Blood Pressure) in terms of death

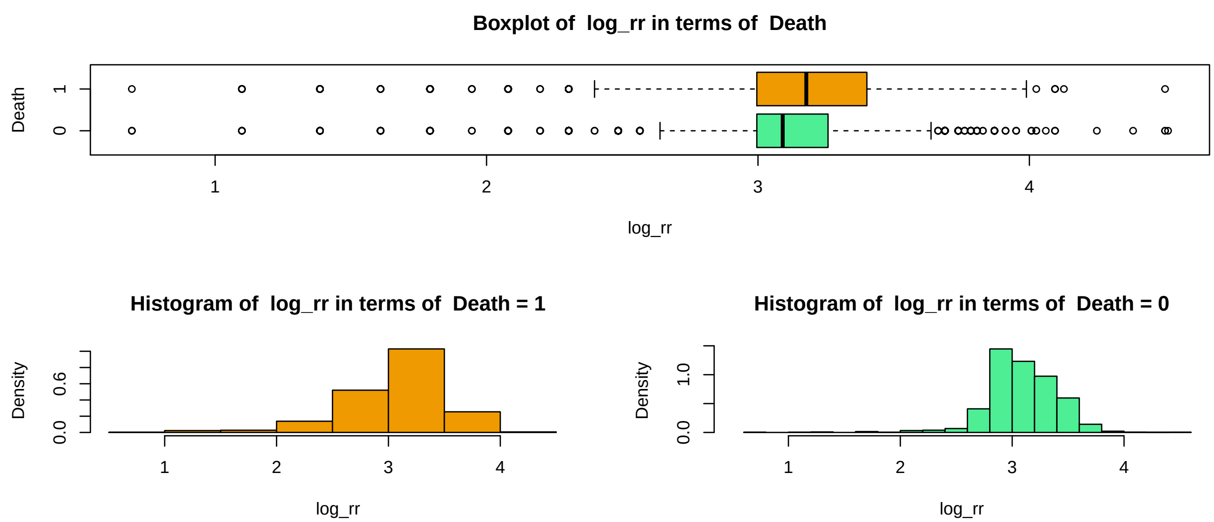


Figure 13: Distribution of log transformation of rr (Respiratory Rate) in terms of death

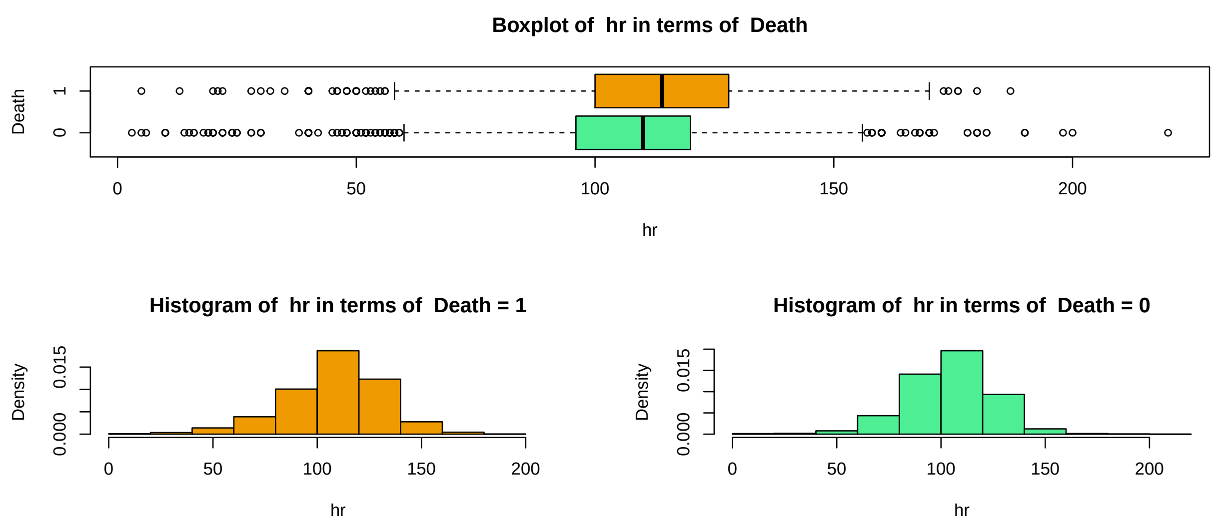


Figure 14: Distribution of hh (hearth Rate) in terms of death

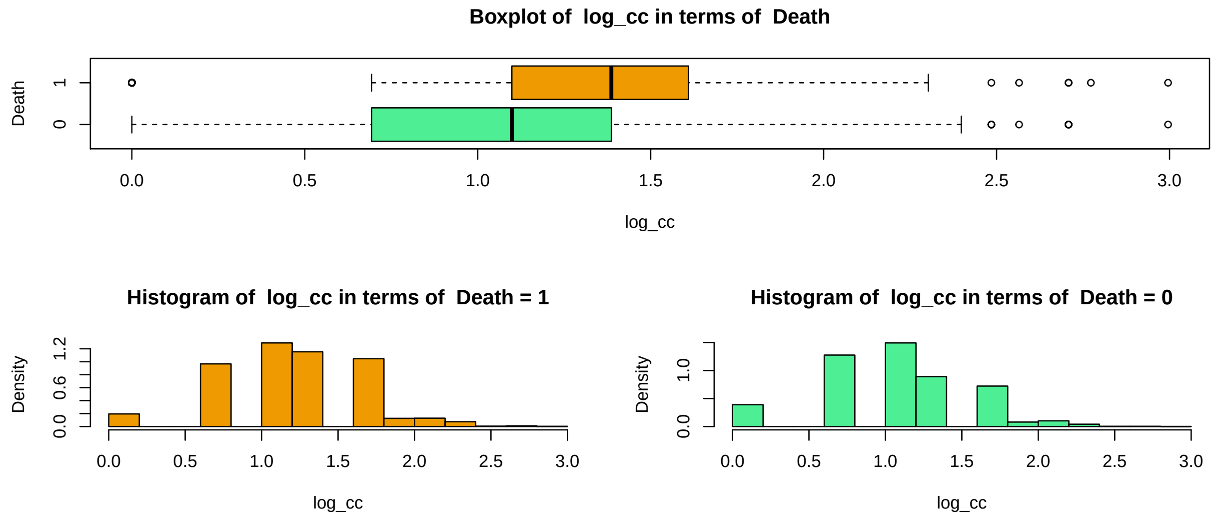


Figure 15: Distribution of log transformation of cc (Central capillary) in terms of death

Figure 16

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

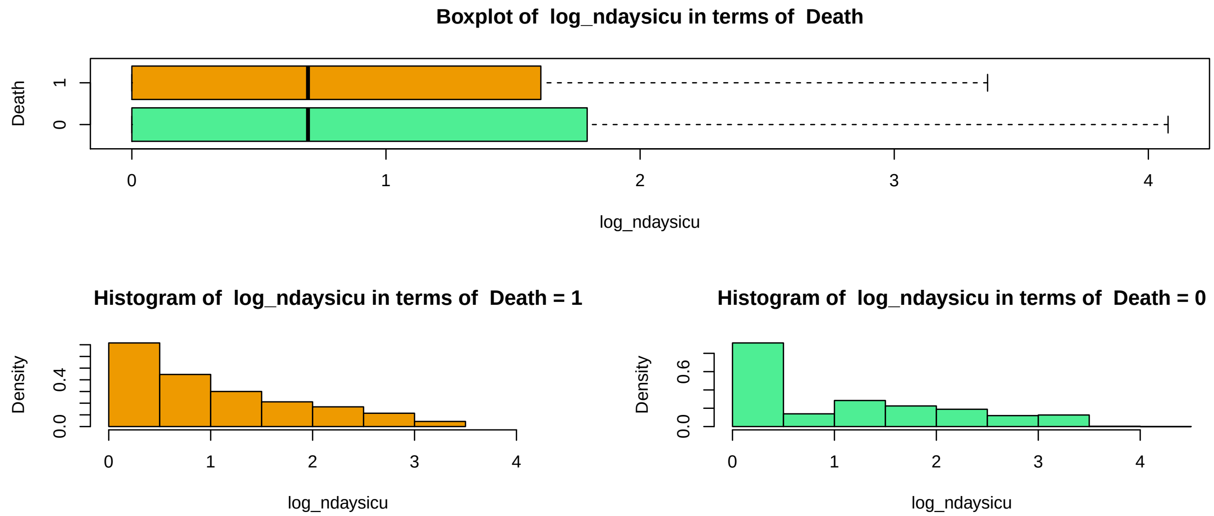


Figure 16: Distribution of log ndaysicu in terms of death

Figure 17 . . .

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Finally, the scatter plot in figure 23 . . .

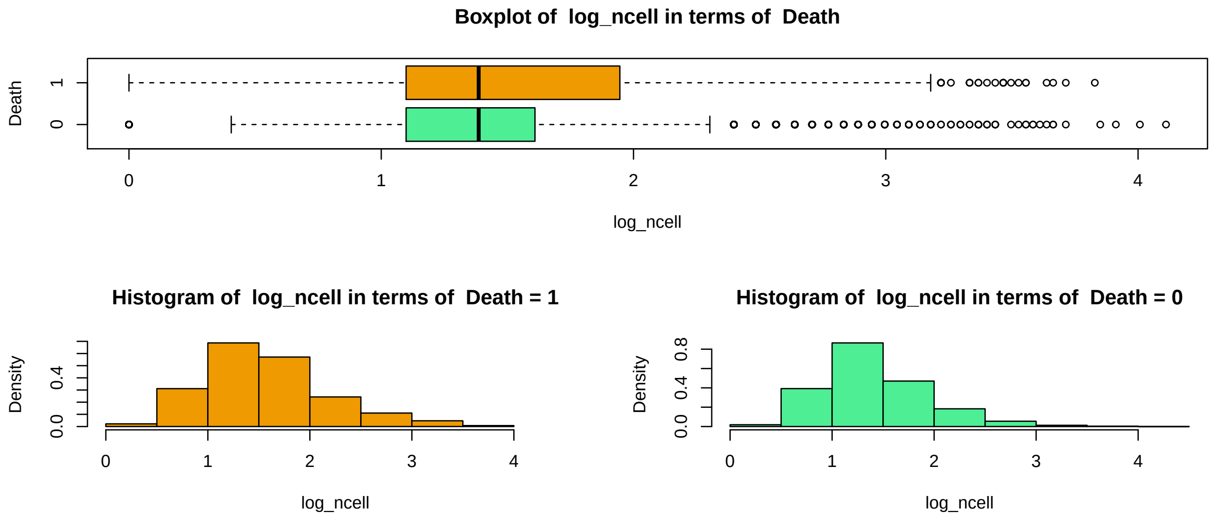


Figure 17: Distribution of log ncell in terms of death

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

The PCP plot in figure 24...

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

The Andrews' plot in figure 25...

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

In conclusion ....

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

In the next section we will make some inference using sample estimators of mean, covariance, correlation

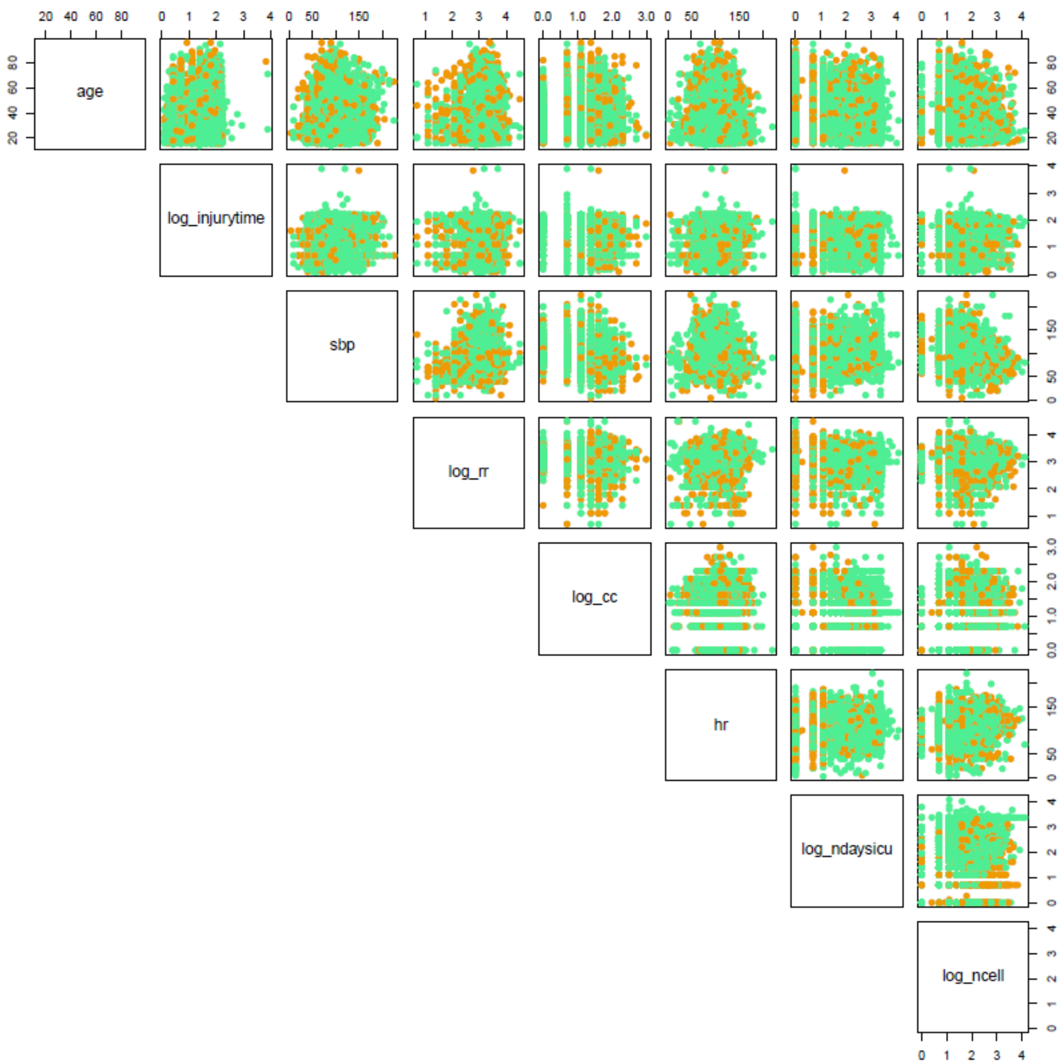


Figure 18: Scatter plot of all quantitiative variables

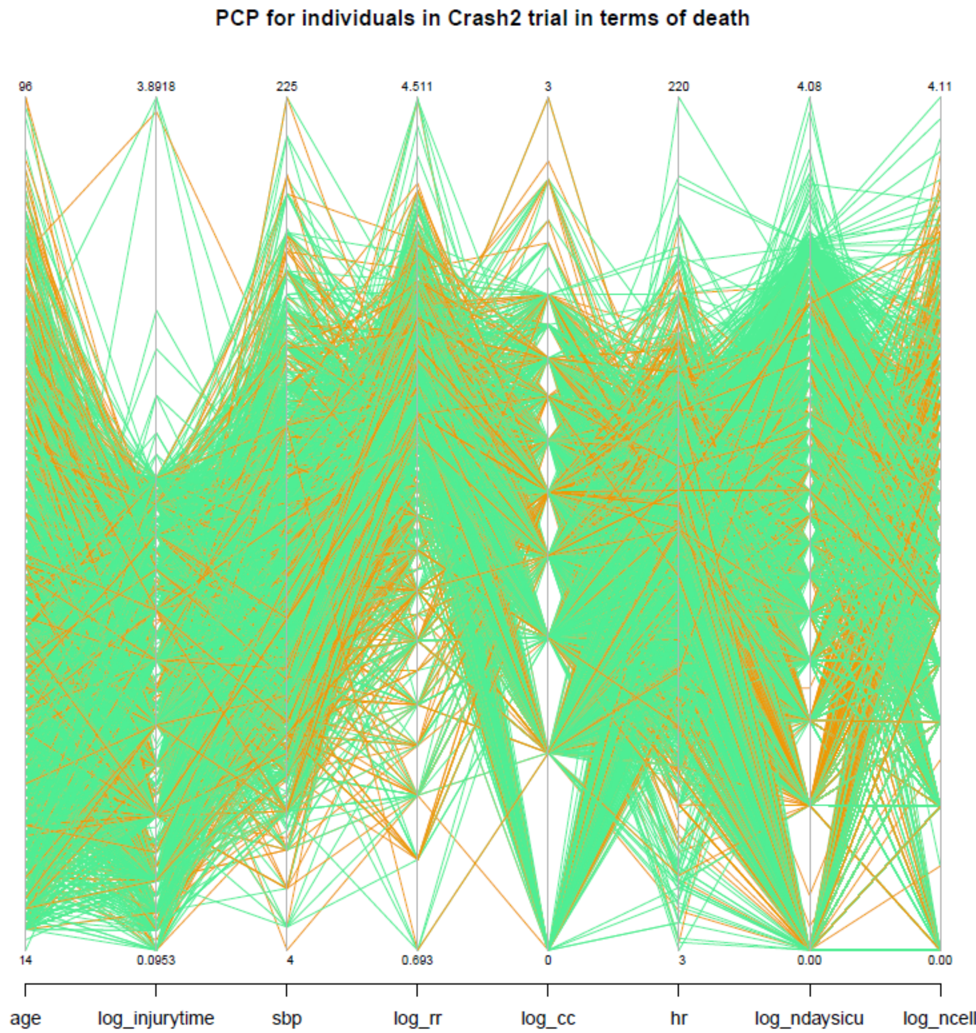


Figure 19: PCP plot of all quantitiative variables

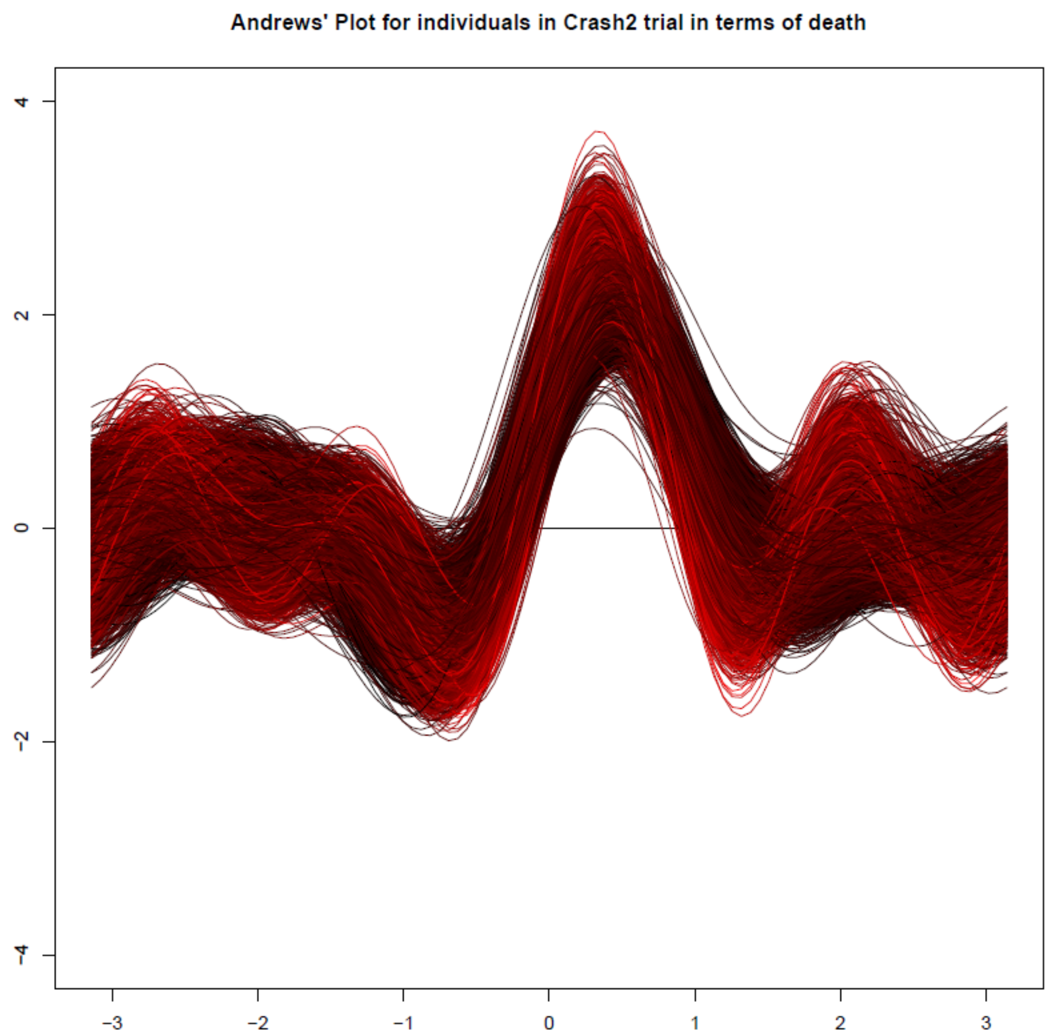


Figure 20: Andrews' plot of all quantitiative variables

Sample Estimators

Sample mean

All population...

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Table 3: Sample mean

age	sbp	hr	log_injurytime	log_rr	log_cc	log_ndaysicu	log_ncell
34.66474	93.12909	108.0621	1.26759	3.106722	1.116876	0.9976935	1.394665

Population with death = 1 ...

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Table 4: Sample mean death

age	sbp	hr	log_injurytime	log_rr	log_cc	log_ndaysicu	log_ncell
37.82849	87.33808	111.4559	1.240578	3.110272	1.232432	0.9390286	1.524647

Survival population

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Table 5: Sample mean survive

age	sbp	hr	log_injurytime	log_rr	log_cc	log_ndaysicu	log_ncell
33.91215	94.50665	107.2548	1.274015	3.105878	1.089387	1.011649	1.363745

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Sample covariance

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Table 6: Sample covariance matrix

age	sbp	hr	log_injurytime	log_rr	log_cc	log_ndaysicu	log_ncell
201.8020767	5.6687061	-26.9775900	0.5843356	0.0754442	0.3511496	1.1379316	0.1173186
5.6687061	604.1019094	-117.2846636	1.8258493	-0.3695028	-2.8828312	1.6117323	-1.9082923
-26.9775900	-117.2846636	459.8457610	-0.0753408	1.3004116	0.8685547	0.9063157	1.4739129
0.5843356	1.8258493	-0.0753408	0.2877629	-0.0037522	0.0126630	0.0863252	0.0034858
0.0754442	-0.3695028	1.3004116	-0.0037522	0.1092305	0.0119420	-0.0260511	-0.0025718
0.3511496	-2.8828312	0.8685547	0.0126630	0.0119420	0.2448313	0.0636306	0.0344845
1.1379316	1.6117323	0.9063157	0.0863252	-0.0260511	0.0636306	1.1508972	0.1949716
0.1173186	-1.9082923	1.4739129	0.0034858	-0.0025718	0.0344845	0.1949716	0.3404679

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Table 7: Sample covariance matrix death patients

age	sbp	hr	log_injurytime	log_rr	log_cc	log_ndaysicu	log_ncell
261.7057666	17.272379	-51.3011548	0.8025150	0.0945272	0.3294187	1.1929405	-0.3849756
17.2723792	705.035309	-116.7715375	2.4101714	0.4360450	-3.2088292	4.1185430	-2.2891556
-51.3011548	-116.771537	564.8052079	0.0045026	1.7090123	1.1424702	0.7975679	1.9352979
0.8025150	2.410171	0.0045026	0.2799950	0.0020301	0.0093817	0.0999836	-0.0265201
0.0945272	0.436045	1.7090123	0.0020301	0.1899808	0.0014615	-0.0137661	-0.0143161
0.3294187	-3.208829	1.1424702	0.0093817	0.0014615	0.2228162	-0.0160445	0.0149818
1.1929405	4.118543	0.7975679	0.0999836	-0.0137661	-0.0160445	0.8604265	0.1140010
-0.3849756	-2.289156	1.9352979	-0.0265201	-0.0143161	0.0149818	0.1140010	0.4261349

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Table 8: Sample covariance matrix survival patients

age	sbp	hr	log_injurytime	log_rr	log_cc	log_ndaysicu	log_ncell
184.6367863	8.3060201	-24.3595684	0.5577010	0.0676092	0.2486948	1.1796601	0.1156599
8.3060201	570.3044793	-111.6340107	1.6410806	-0.5550393	-2.6086143	0.9158253	-1.5962997
-24.3595684	-111.6340107	431.5566186	-0.0673383	1.1998768	0.6880413	0.9909257	1.2344841
0.5577010	1.6410806	-0.0673383	0.2894326	-0.0050994	0.0143642	0.0826221	0.0116551
0.0676092	-0.5550393	1.1998768	-0.0050994	0.0900403	0.0143148	-0.0289142	0.0000845
0.2486948	-2.6086143	0.6880413	0.0143642	0.0143148	0.2461653	0.0845804	0.0347028
1.1796601	0.9158253	0.9909257	0.0826221	-0.0289142	0.0845804	1.2191014	0.2164958
0.1156599	-1.5962997	1.2344841	0.0116551	0.0000845	0.0347028	0.2164958	0.3151667

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Sample correlation

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Table 9: Sample covariance matrix

age	sbp	hr	log_injurytime	log_rr	log_cc	log_ndaysicu	log_ncell
1.0000000	0.0162355	-0.0885593	0.0766800	0.0160691	0.0499569	0.0746681	0.0141536
0.0162355	1.0000000	-0.2225256	0.1384817	-0.0454874	-0.2370449	0.0611251	-0.1330613
-0.0885593	-0.2225256	1.0000000	-0.0065495	0.1834860	0.0818573	0.0393963	0.1177952
0.0766800	0.1384817	-0.0065495	1.0000000	-0.0211642	0.0477075	0.1500037	0.0111366
0.0160691	-0.0454874	0.1834860	-0.0211642	1.0000000	0.0730252	-0.0734743	-0.0133361
0.0499569	-0.2370449	0.0818573	0.0477075	0.0730252	1.0000000	0.1198711	0.1194407
0.0746681	0.0611251	0.0393963	0.1500037	-0.0734743	0.1198711	1.0000000	0.3114691
0.0141536	-0.1330613	0.1177952	0.0111366	-0.0133361	0.1194407	0.3114691	1.0000000

Figure 21 ...

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

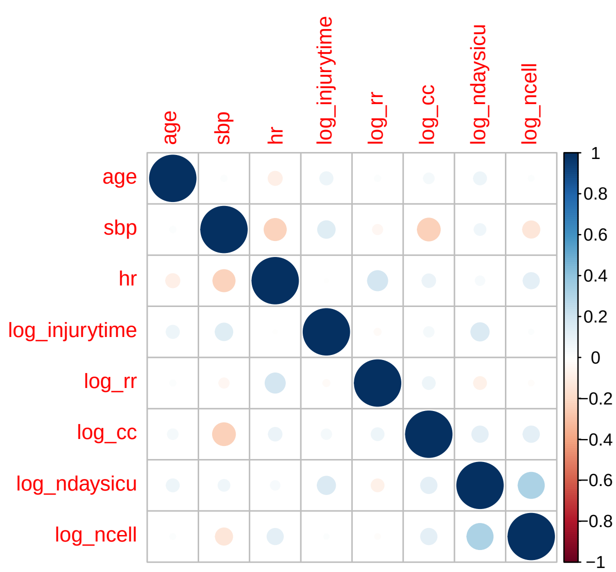


Figure 21: Sample Correlation Crash2 Dataset

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Table 10: Sample covariance matrix death patients

age	sbp	hr	log_injurytime	log_rr	log_cc	log_ndaysicu	log_ncell
1.0000000	0.0402106	-0.1334354	0.0937500	0.0134059	0.0431388	0.0794978	-0.0364547
0.0402106	1.0000000	-0.1850471	0.1715408	0.0376765	-0.2560164	0.1672173	-0.1320676
-0.1334354	-0.1850471	1.0000000	0.0003580	0.1649837	0.1018409	0.0361794	0.1247455
0.0937500	0.1715408	0.0003580	1.0000000	0.0088021	0.0375608	0.2037027	-0.0767761
0.0134059	0.0376765	0.1649837	0.0088021	1.0000000	0.0071034	-0.0340485	-0.0503150
0.0431388	-0.2560164	0.1018409	0.0375608	0.0071034	1.0000000	-0.0366435	0.0486202
0.0794978	0.1672173	0.0361794	0.2037027	-0.0340485	-0.0366435	1.0000000	0.1882687
-0.0364547	-0.1320676	0.1247455	-0.0767761	-0.0503150	0.0486202	0.1882687	1.0000000

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Table 11: Sample covariance matrix survival patients

age	sbp	hr	log_injurytime	log_rr	log_cc	log_ndaysicu	log_ncell
1.0000000	0.0255965	-0.0862962	0.0762902	0.0165817	0.0368888	0.0786281	0.0151619
0.0255965	1.0000000	-0.2250216	0.1277329	-0.0774554	-0.2201623	0.0347328	-0.1190669
-0.0862962	-0.2250216	1.0000000	-0.0060252	0.1924861	0.0667548	0.0432018	0.1058514
0.0762902	0.1277329	-0.0060252	1.0000000	-0.0315884	0.0538137	0.1390921	0.0385896
0.0165817	-0.0774554	0.1924861	-0.0315884	1.0000000	0.0961511	-0.0872717	0.0005018
0.0368888	-0.2201623	0.0667548	0.0538137	0.0961511	1.0000000	0.1543962	0.1245894
0.0786281	0.0347328	0.0432018	0.1390921	-0.0872717	0.1543962	1.0000000	0.3492684
0.0151619	-0.1190669	0.1058514	0.0385896	0.0005018	0.1245894	0.3492684	1.0000000

Figure 22…

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

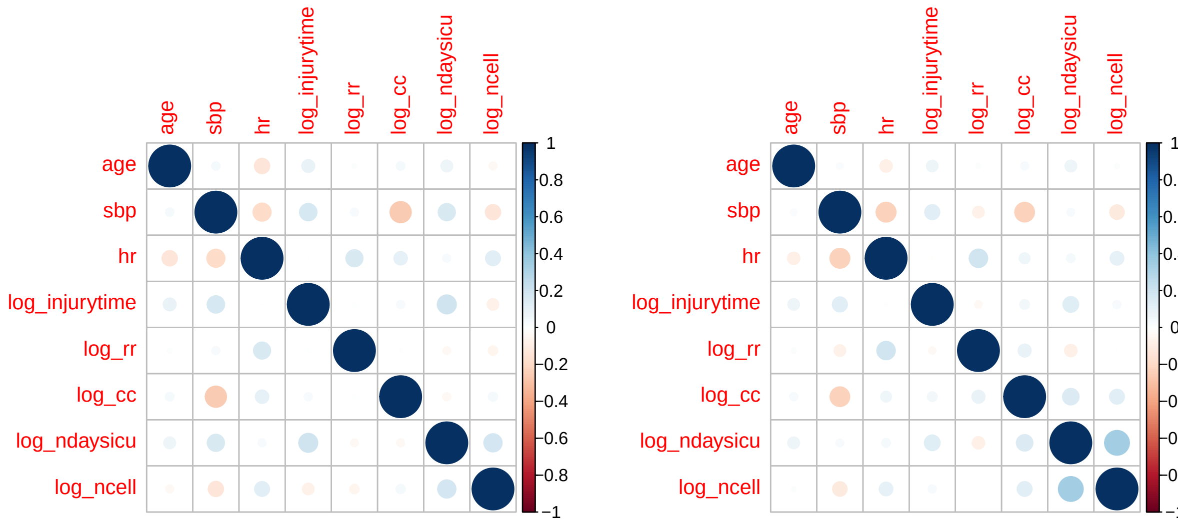


Figure 22: Sample Correlation for death and survival patients from Crash2 Dataset

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Outliers

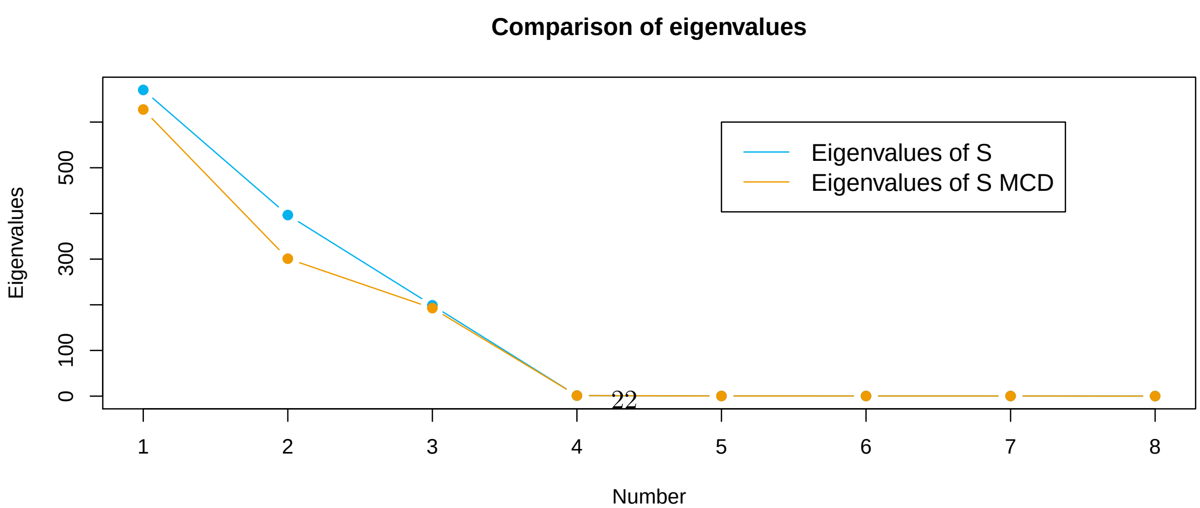
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Table 12: Sample robust media all population

age	sbp	hr	log_injurytime	log_rr	log_cc	log_ndaysicu	log_ncell
33.94433	92.32304	108.848	1.271321	3.137716	1.106643	0.9422688	1.357334

Comparison of the eigenvalues of the covariance matrix for sample and robust is given by figure 23 ...

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.



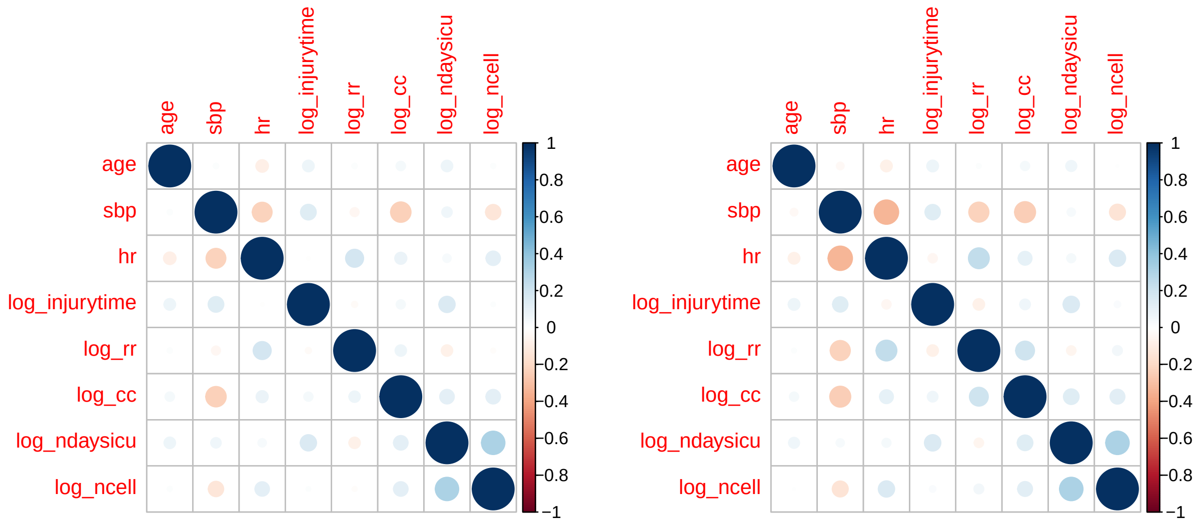


Figure 24: Correlation all population and MCD correlation

Principal Component Analysis

We can divide our quantitative variables in three categories:

1. Individual factors

  i) Age
  ii) injurytime

2. Biometrics

  i) sbp
  ii) rr
  iii) cc
  iv) hr

3. Medical attention

  i) ndaysicu
  ii) ncell

Then, we will discriminate how the principal components uses this variables.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

PCA Complete dataset

In figure 25, we plot the first 2 PC. In this case, there is approximately 40% of the variance explained with these two principal components

In green color, there is the survival patients and orange is the death patients. Is is difficult to see a clear sepatations between both populutions. In the same, there is no linear relationship between these two groups.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

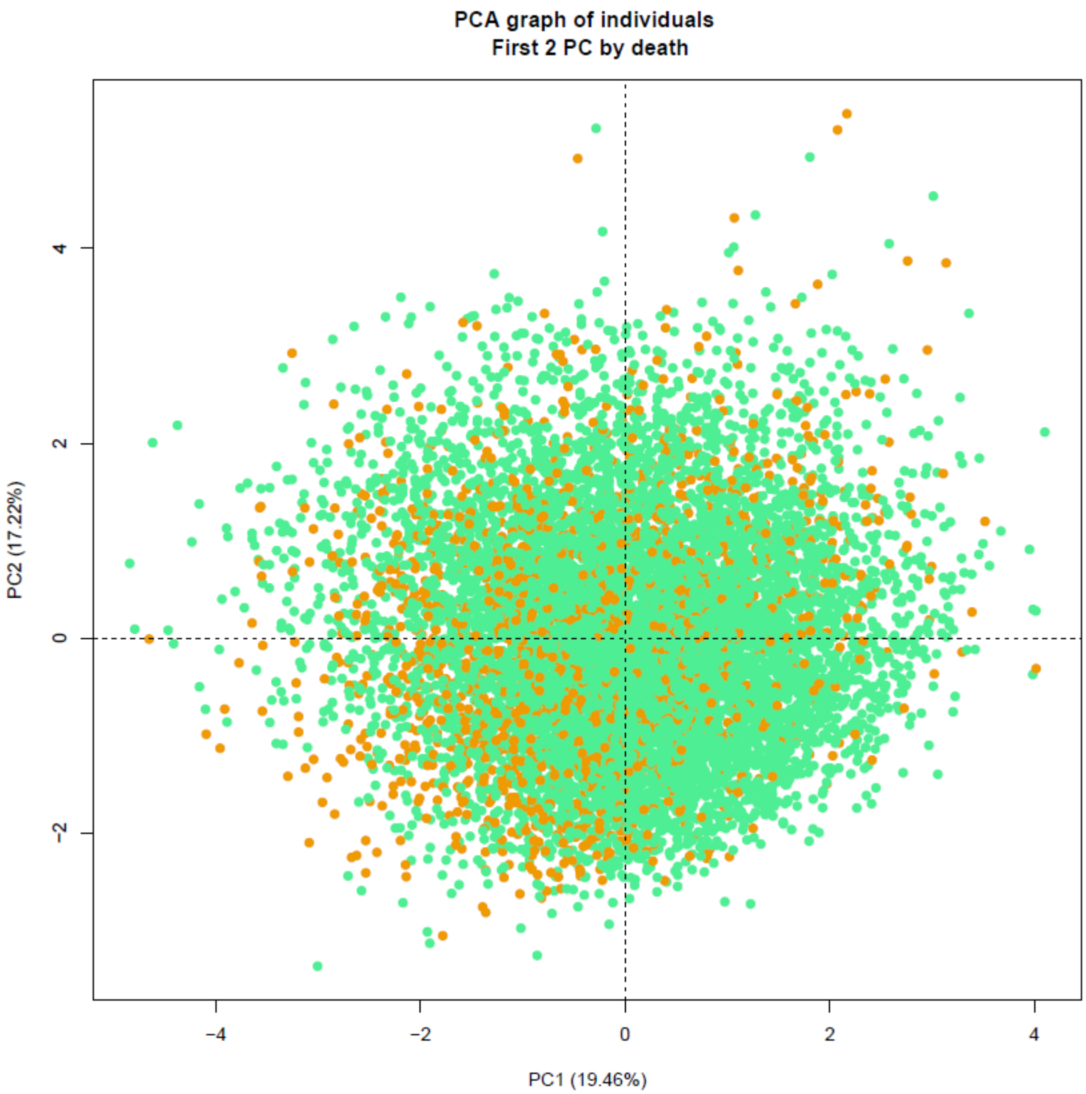


Figure 25: First two PCs based on the sample covariance all population

On the other hand, we will analize the weights of the fist two principal components. This can be seen in the figure 26. For the case of the first PC (Left), the largest positive value are associated with the specfic variable of *sbp*. Some studies, for example Banegas et all and Kurl et all suggest that Systolic blood pressure is a more frequent cardiovascular risk factor than other blood related measures and can be used in order to detect future diseases. So, it appears that this principal components reflects these effect for the patients.

In the case of the second Principal component (Right) the largest positive value with the medical attention and injury time. It suggets that previous you receive medical attention, it is possible that you can have more chance for survival.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

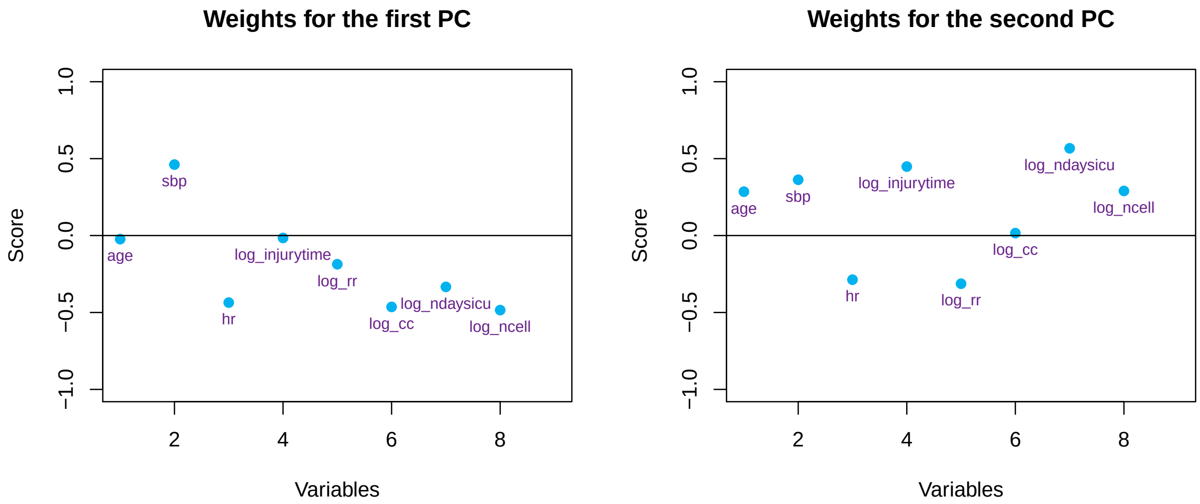


Figure 26: Loadings for the first and sencond PCs individually

On the other hand, when we consider the loadings for the first two PCs (figure 27) that larges value in magnitude are associated with the sbp pressure and medical attention.

It is important to notice the inverse relation between sbp and other biometric measures. For example

- Herakova et all report that in general deep breathing could reduce blood pressures (BP). So we
- Harvard Heart Letter reports that an *isolated increase in blood pressure can drop the heart rate a little*

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

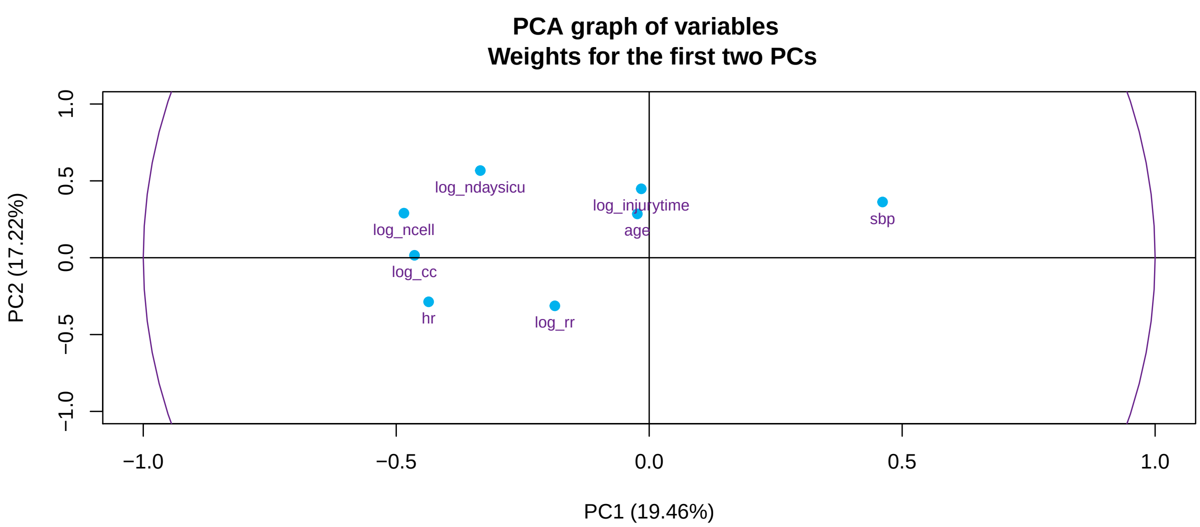


Figure 27: Loadings for the first two PC

Finally figure 28 reflects the PC scores and PC loadings. In this case, it is difficult to make conclusions due to the numbers of observed individuals, but we can make the following assevarations:

- There is a strong positive relation between age and injury time.
- There is no relation between the ndaysicu with the biometrical measures as sbp and hr. It makes sense, because this is a variable more related with the medical attention, as can be appreciated with the variables *ncell* and *injurytime*

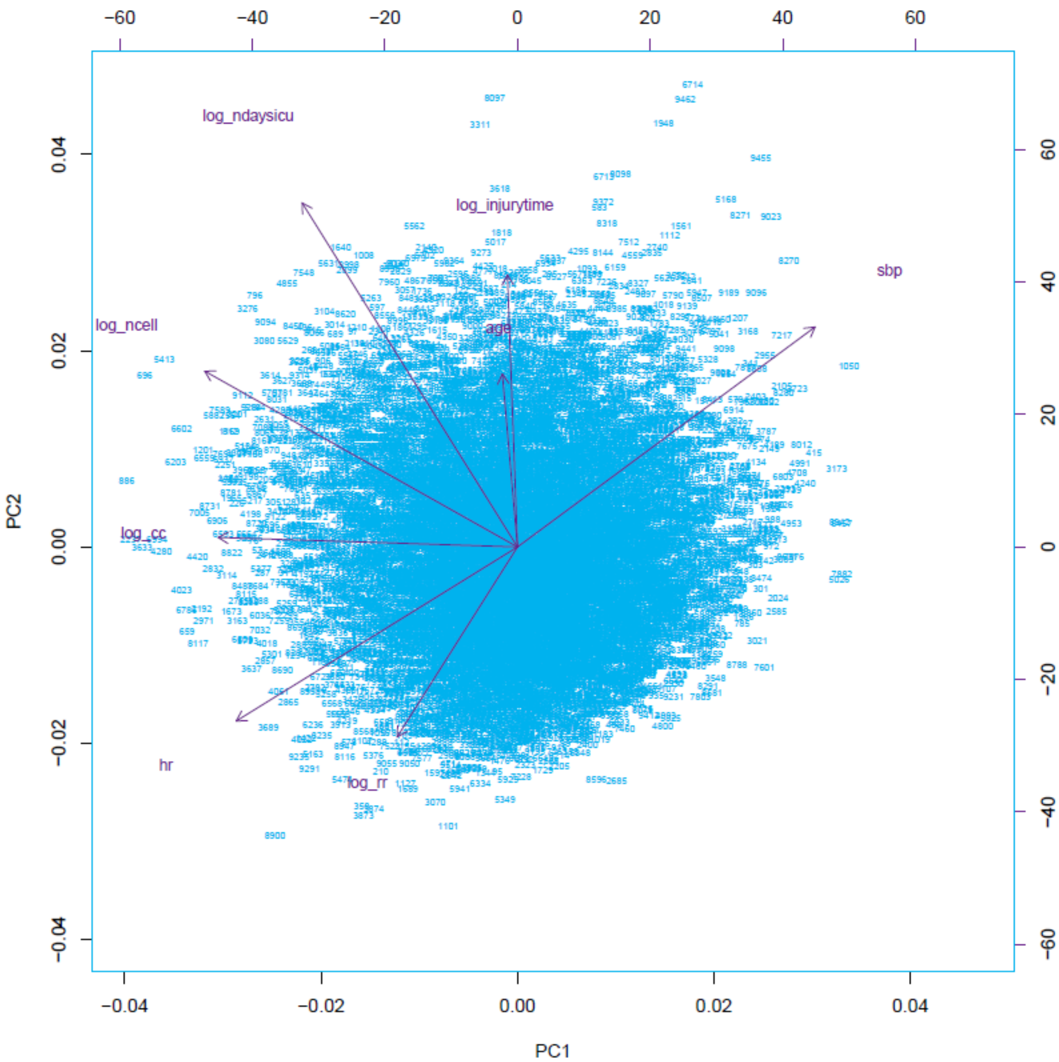


Figure 28: PC Scores and PC loading all population

The figure 29 shows the explained variance of the eigenvalue. We see that the first 4 principal components explains approximately 63% of the model.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

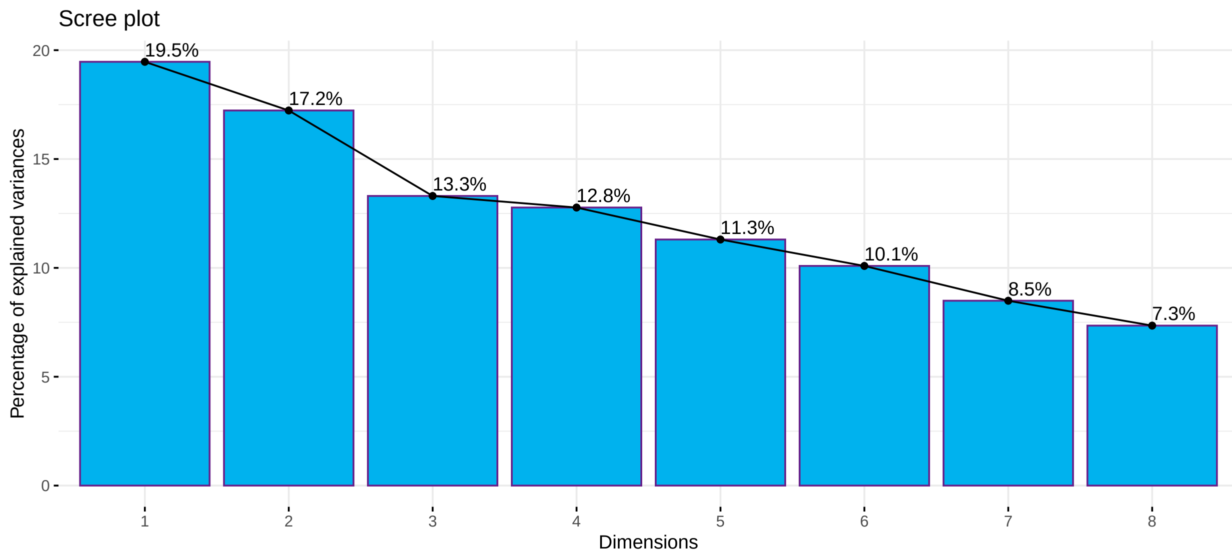


Figure 29: Eigenvalues of the sample correlation matrix

Figure 30 reflects the relations in the first four principal componets. We do not appreciate a clear distintion bewteen the groups, and the ausence of linear relation between them.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Finally figure 31 explain the correlation between the variables and the principal components. IN the case of tghe first principal components, in magnitude the biometrical measures and medical attention variables are the more related with them. In the case of the second principal component, the individual factors such age and injury time adquire more relevance in the analisis.

In case of age this i more influential in PC4, and it can explain the right orange points in figure 35.

IS NECESSARY TO DO THE SAME BY CLASS?

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

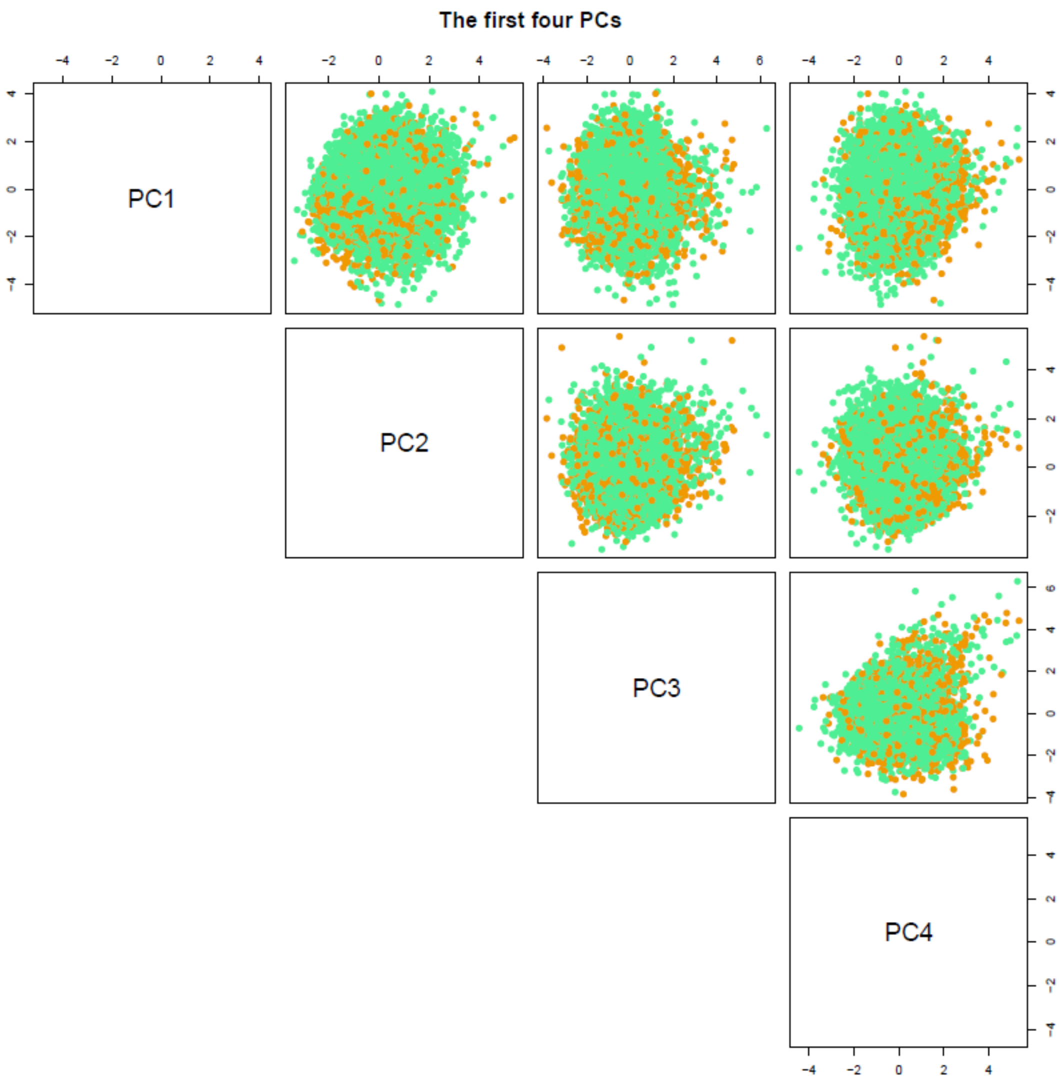


Figure 30: PC Scores and PC loading all population

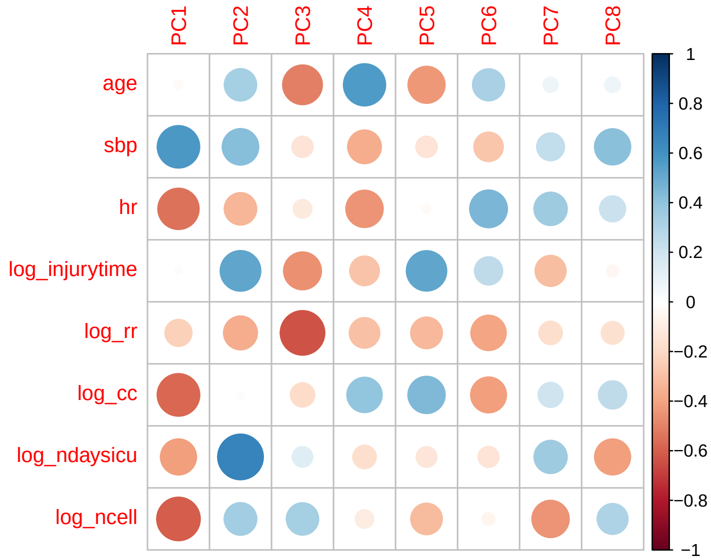


Figure 31: Correlation between dataset and all PC

PCA by Category

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Important. For these plot the groups are now female (blue) and male (gray)

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

figure 33. . .

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Figures 35 and 36 . . .

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

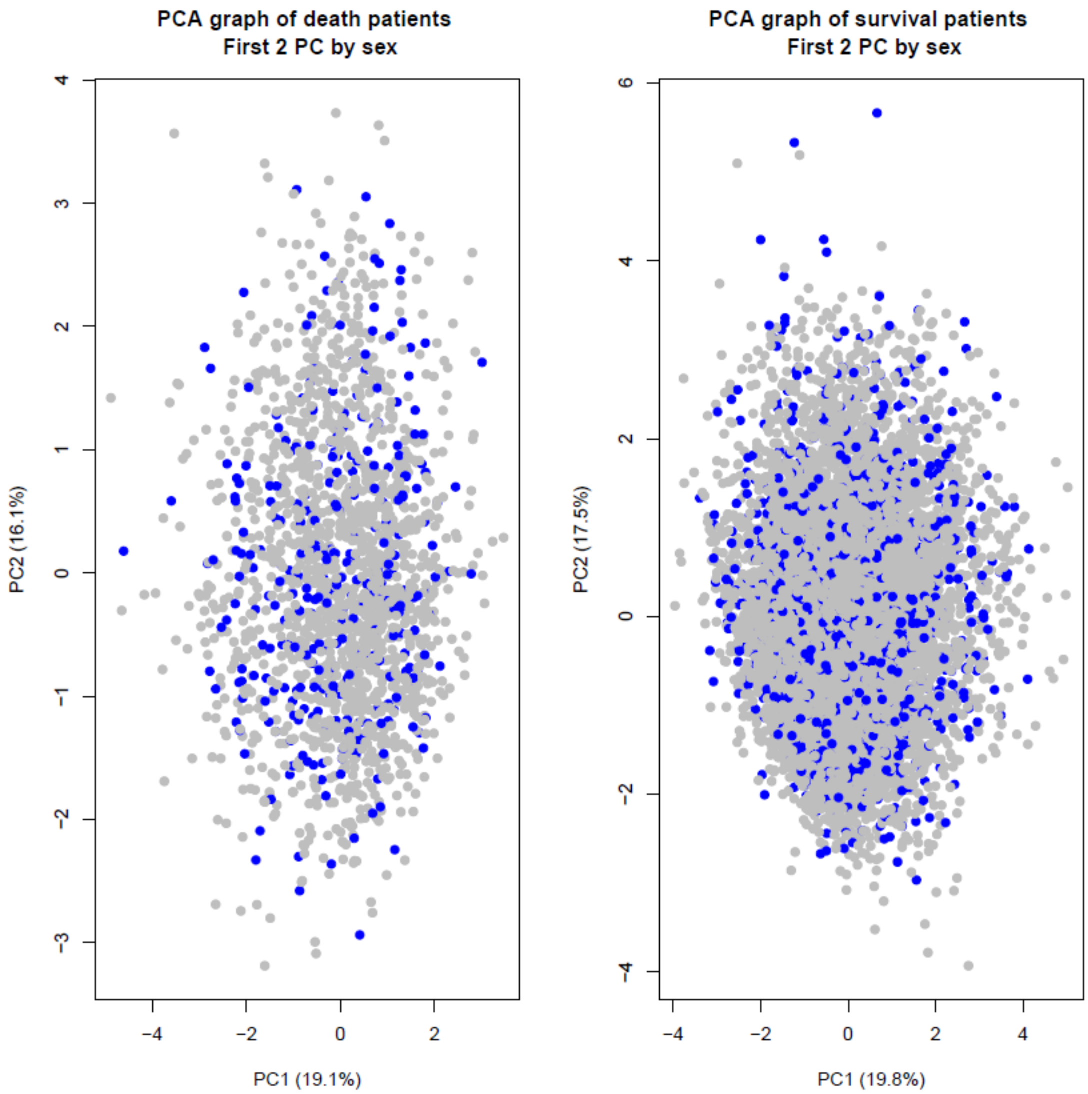


Figure 32: PC Scores and PC loading all population

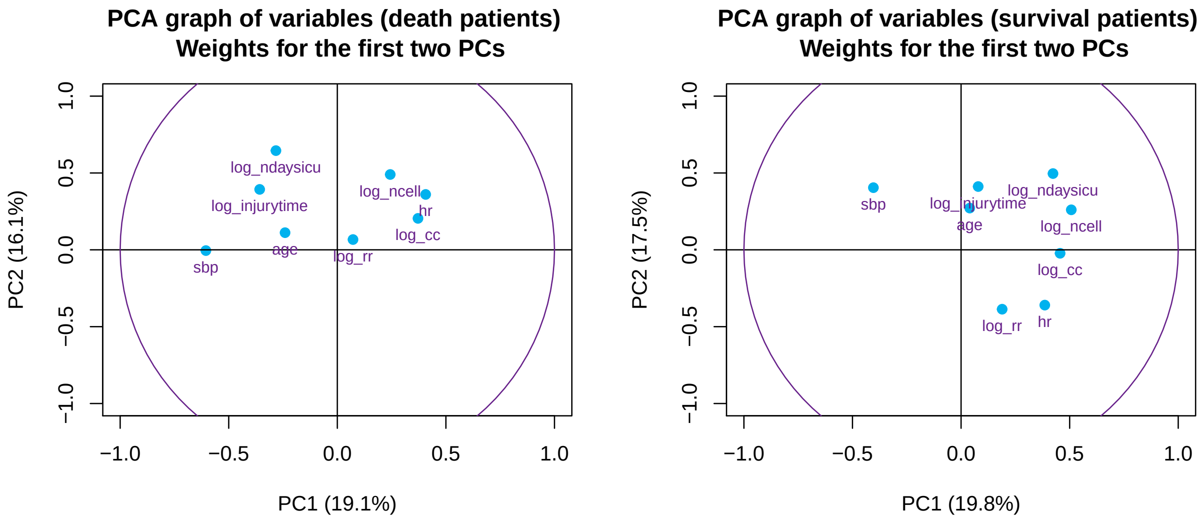


Figure 33: Loadings for the first two PC by death and survival patients

Conclusions

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

References

1. Banegas JR, de la Cruz JJ, Rodríguez-Artalejo F, Graciani A, Guallar-Castillón P, Herruzo R. Systolic vs diastolic blood pressure: community burden and impact on blood pressure staging. J Hum Hypertens. 2002 Mar;16(3):163-7. doi: 10.1038/sj.jhh.1001310. PMID: 11896505.
2. S. Kurl, J.A. Laukkanen, R. Rauramaa, T.A. Lakka, J. Sivenius, and J.T. Salonen. Systolic Blood Pressure Response to Exercise Stress Test and Risk of Stroke. Sep 2001 https://doi.org/10.1161/hs0901.095395Stroke. 2001;32:2036–2041

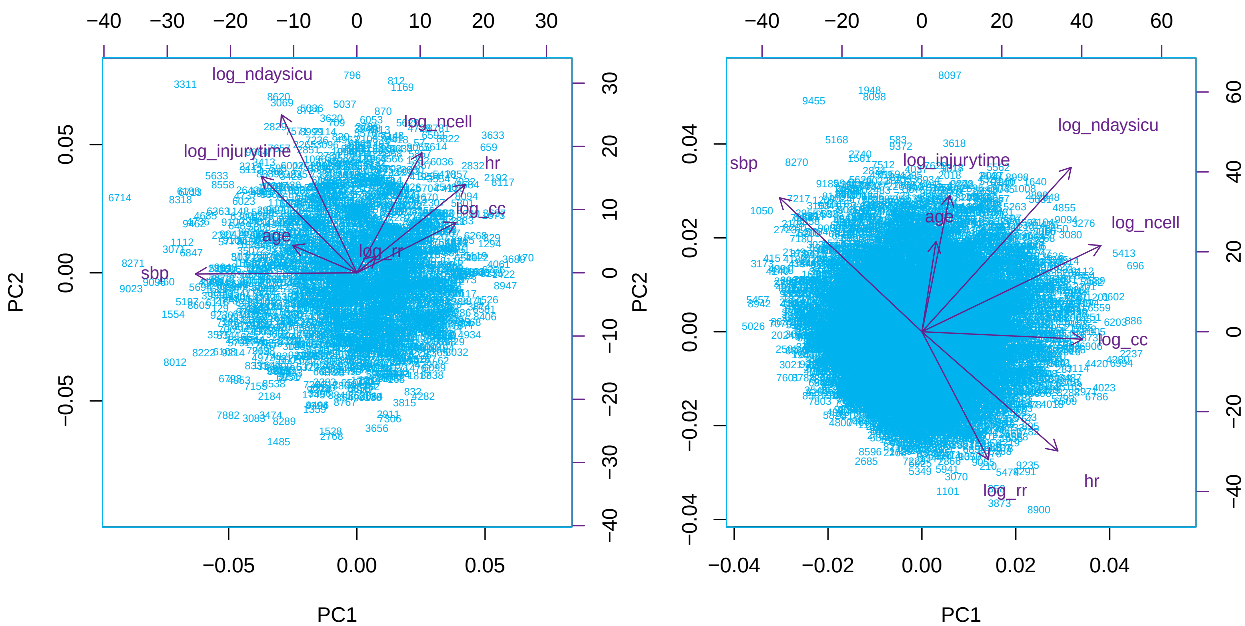


Figure 34: PC Scores and PC loading all population for death and survival patiens

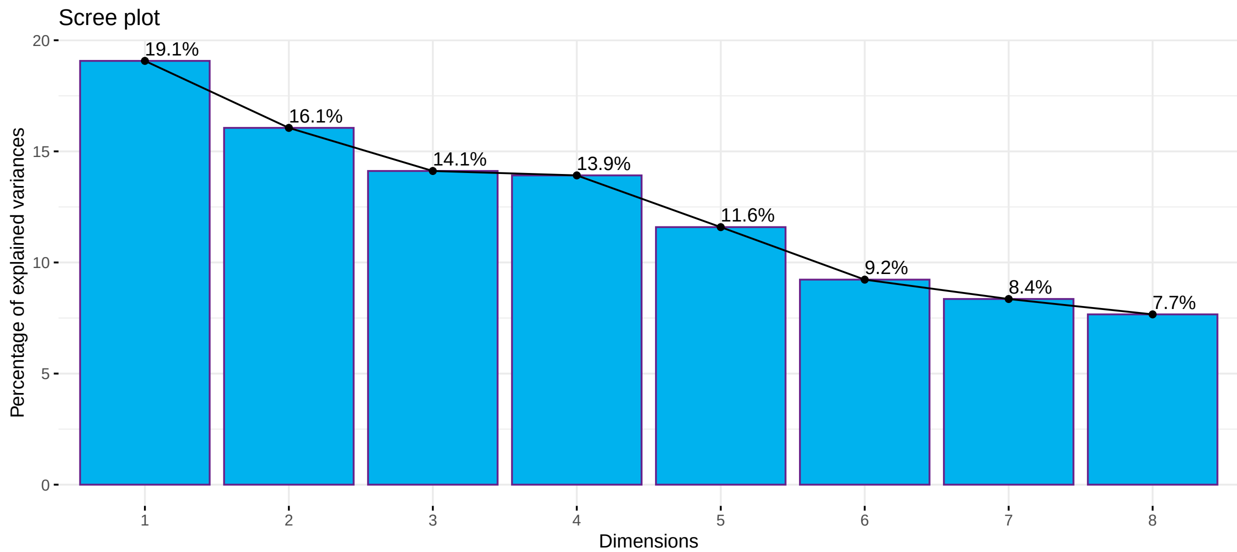


Figure 35: Eigenvalues of the sample correlation matrix for death patients

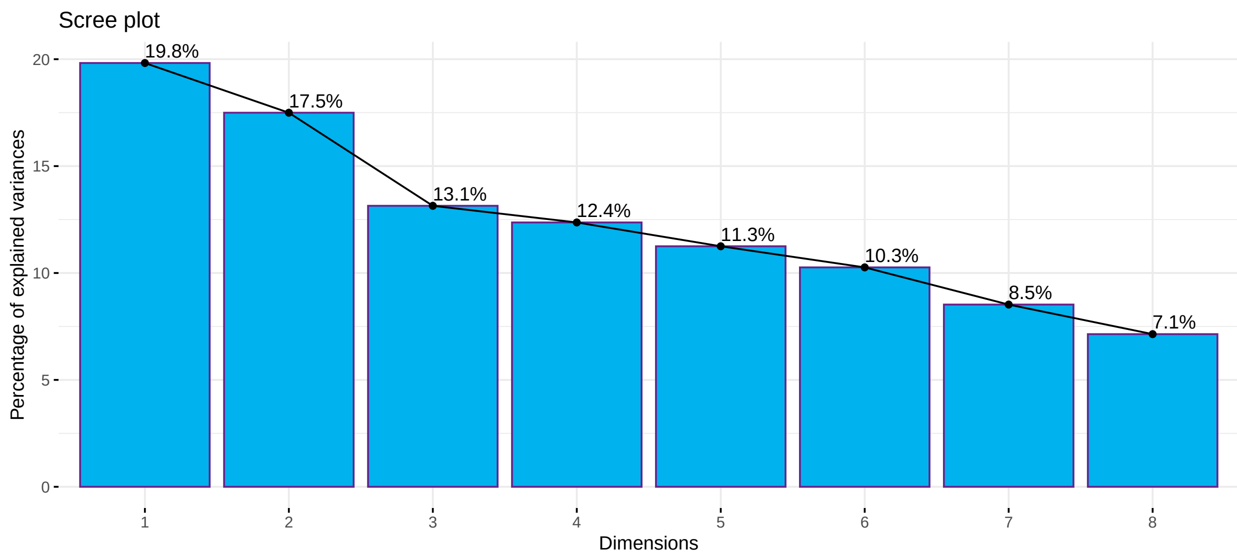


Figure 36: Eigenvalues of the sample correlation matrix for survival patients