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. The treatment variable is unfortunately unavailable in our data set since it is proprietary to those who performed the original research. Our data is therefore contains the observations of the control subset and the treatment. We have sent an email to request if sharing this variable is possible. If we receive the variable, we will include it in Step 2 of this assignment.

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 consists 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 set 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 Refill Time (seconds). Normal range for adults at rest. Less than 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 involved in the data set.

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.

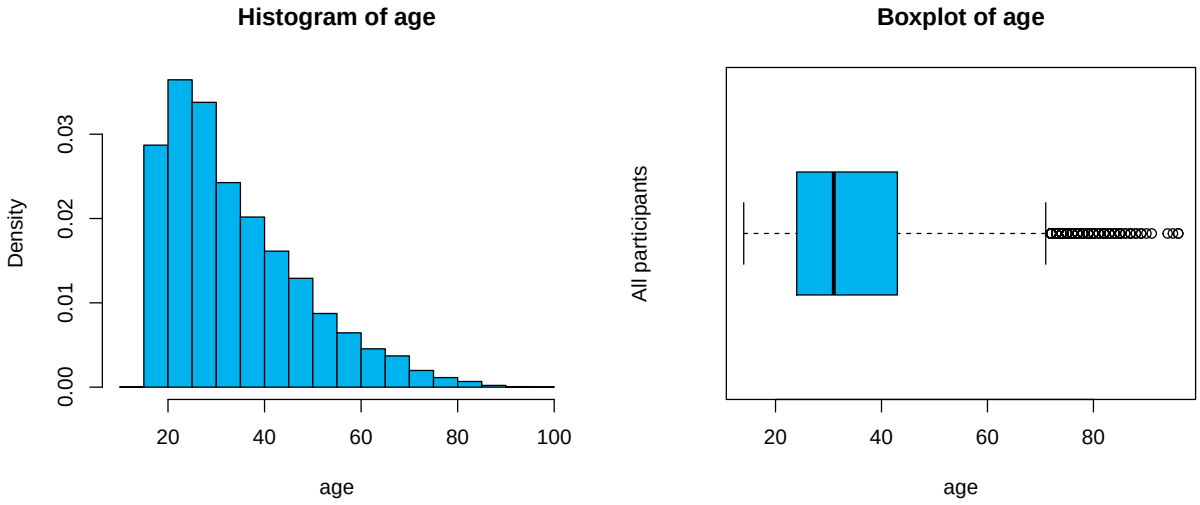

Figure 1: Distribution of age variable

The Figure 2 below shows the distribution of the variable Injury (*injurytime*) We can see how this variable is highly skewed to the left 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 as is the case for this data set on trauma. We apply a *log* transformation to this variable `log_injurytime = log(injurytime + 1)`, similarly to many other variables we have, because of the weight of the variables to the left of the distribution, in order to make a more normalized distribution. The *log* transformation as we can see in the below of Figure 2. After the *log* transformation, we can see that two observations remain quite separate , with a value of four, that appears as potential outlier to the distribution, as they have almost 4 times the median value.

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.

rr (Respiratory Rate) appears, similar to sbp, resembling a moderately balanced distribution around 22 respirations per minute, although *rr* is weighted more to the left. The distribution of this variable is showed in Figure 4. Taking a *log* transformation `log_rr = log(rr)`, we have the new distribution below as part of Figure 4. After the log transformation, the distribution remains quite spread out with a number fo values beyond the whiskers of the box plot. The *log* transformation in this case has not significantly changed the shape of the distribution, it does not shift much more towards a normal distribution after the *log* transformation.

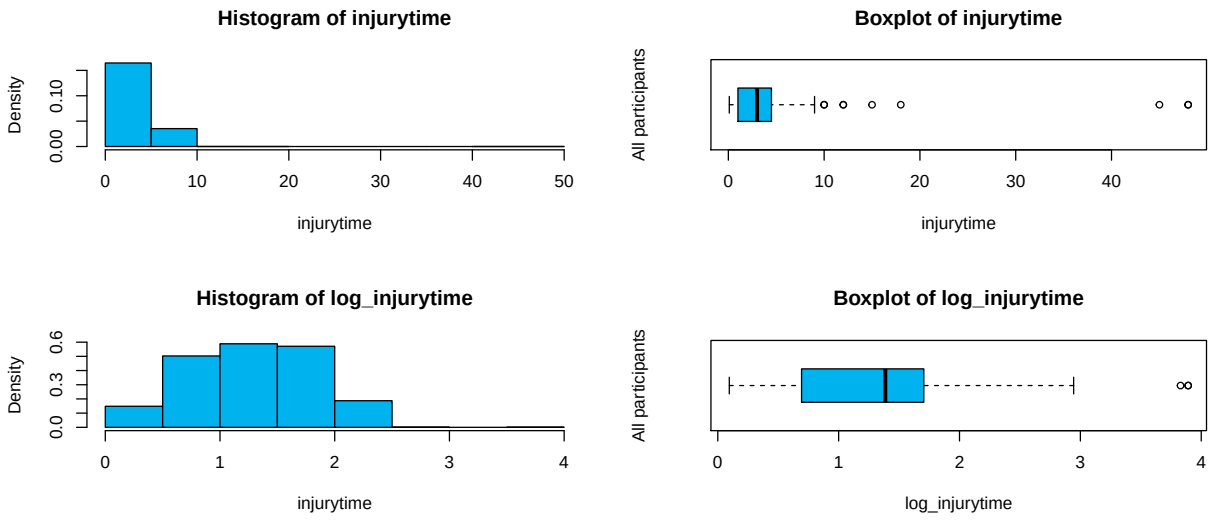

Figure 2: Distribution of injurytime variable and log of injurytime

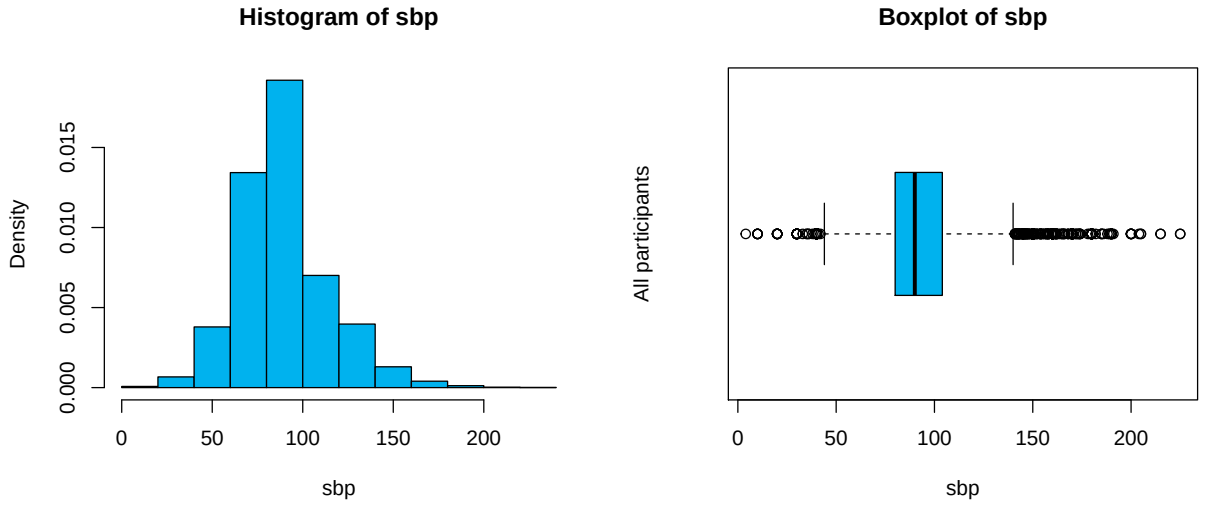

Figure 3: Distribution of sbp (Systolic Blood Pressure)

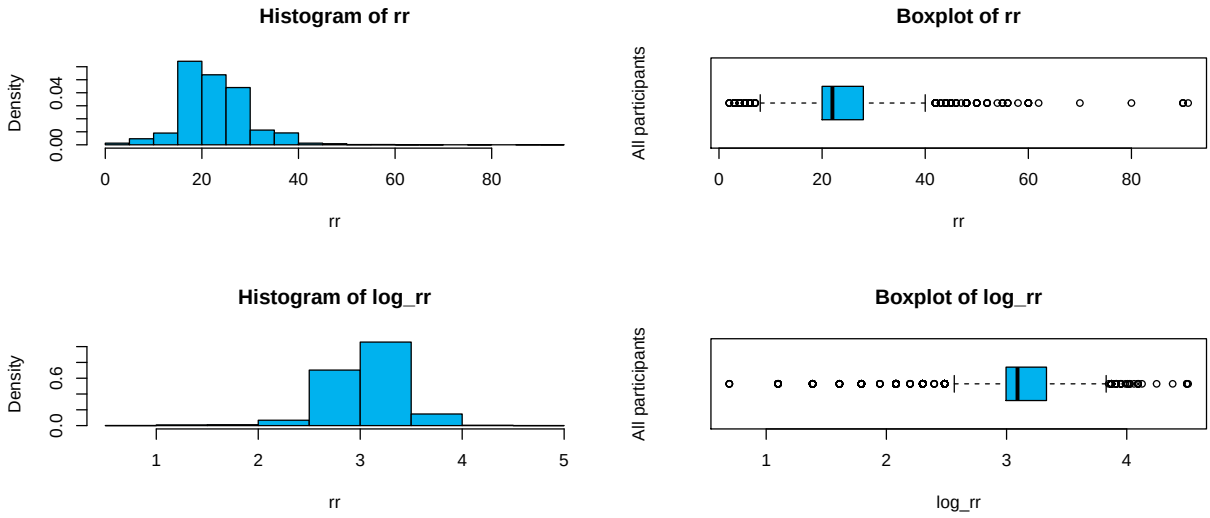

Figure 4: Distribution of rr (Respiratory Rate) and logrr

In the case of *hr* (Heart rate), Figure 5 has a distribution that seems fairly balanced at around 110, similar to the variables above, like *sbp* and *rr*. However, many values remain outside the whiskers and the IQR is quite tight, showing most people fall within a tight range but there is a number of people who have irregular rates on either side.

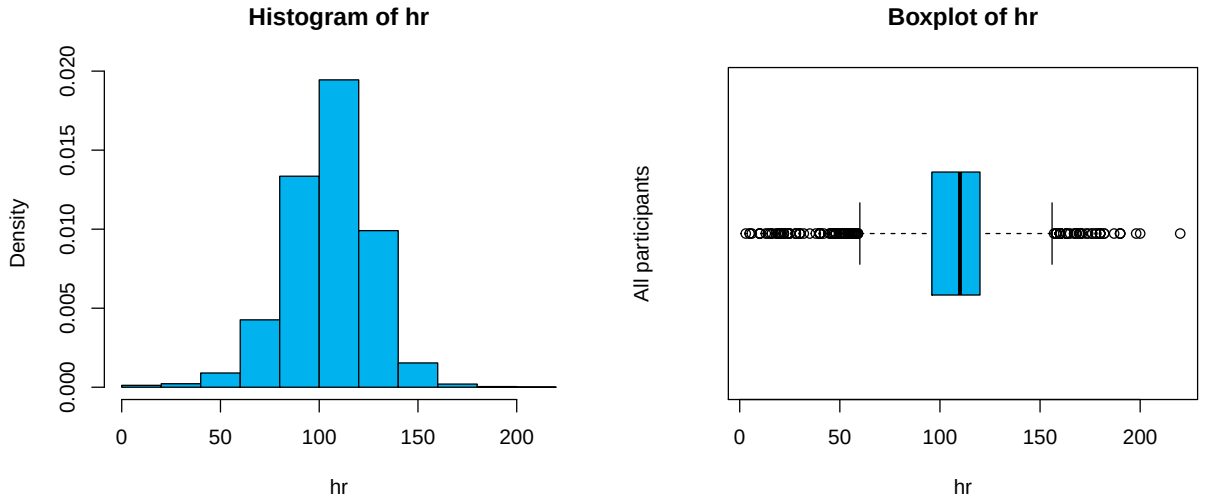

Figure 5: Distribution of hh (hearth Rate)

For *cc* (Central capillary) refill has 75% of the observations below roughly 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 log has made the distribution more normal, and was quite successful in this case. There remains a number of observations to the far right of the distribution, which could be possible outliers as the represent values approximately three times the median.

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

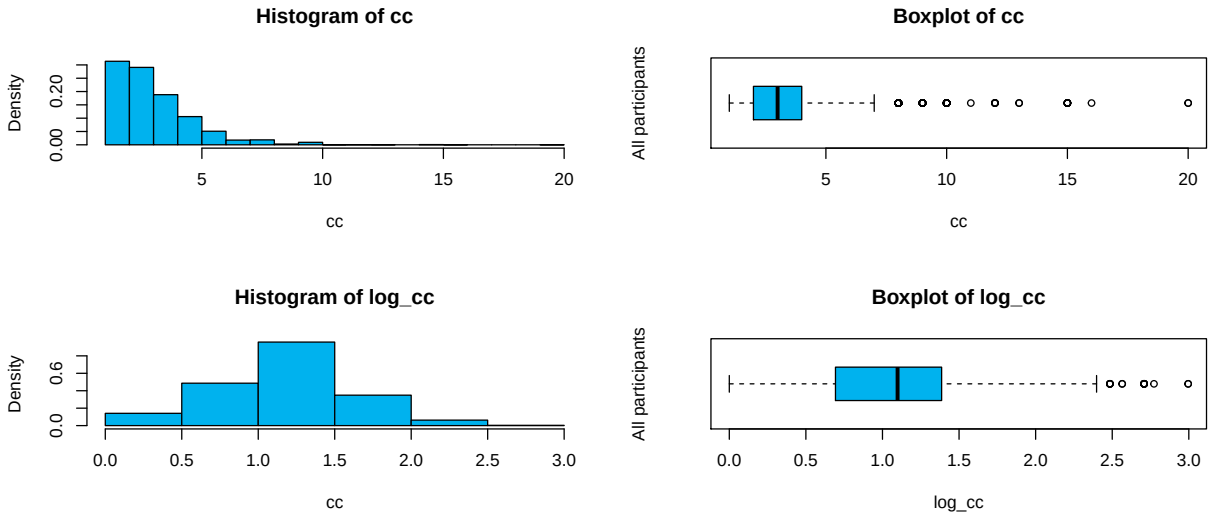

Figure 6: Distribution of cc (Central capillary) and log transformation

to remain in the hospital for long. The transformed distribution `log_ndaysicu = log(ndaysicu + 1)` is given in the below part of figure 7. After the *log* transformation, this distribution doesn't appear much more gaussian in nature. The values are still heavily weighted the left. No values appear outside the whiskers, however, therefore this leads us to believe that there is not much likelihood of outliers in this distribution.

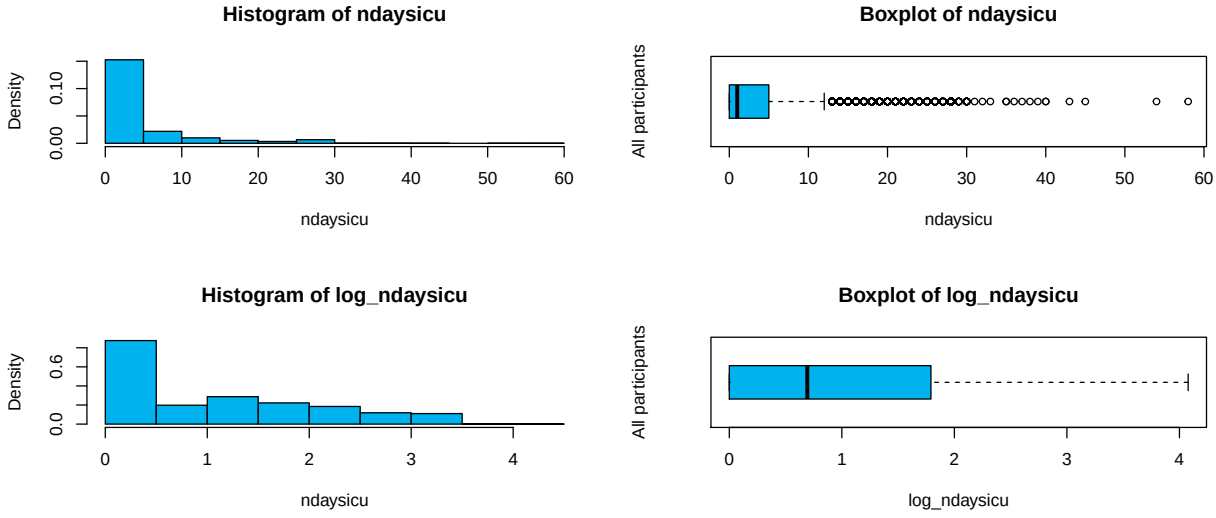

Figure 7: Distribution of ndaysicu and log transformation

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, is that many patients only need a small number of or zero units of red cell products transfused. Due to this variable being highly weighted to the right, we apply the *log* transformation `log_ncell = log(ncell + 1)` of the figure 8 in their below part of the Figure. After transformation, the variables visually appears to be more Gaussian, but a large number of values remain outside the whiskers on the right of the distribution, though none appear distinct enough for the others to be outliers.

For the categorical variables, we will focus on the distributions of deaths, where 0 represents survival. The

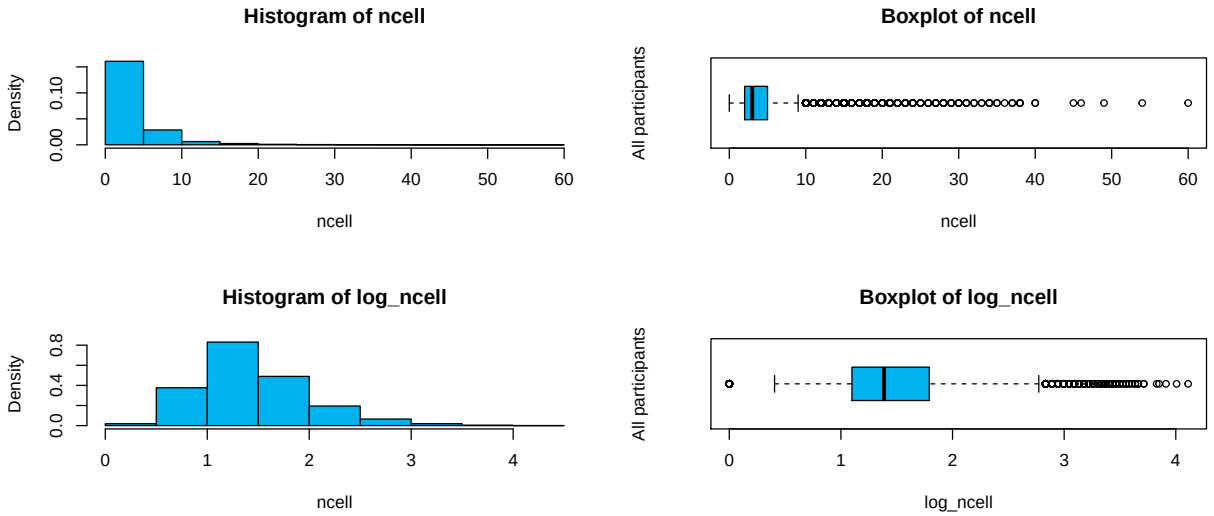

Figure 8: Distribution of ncell

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, it can be considered as a sign that most people survive trauma injuries in the data set and that with the administration of the drug, most people survive . Although, this is not necessarily due to the administration of the drug as both the control and treatment are contained in this data set.

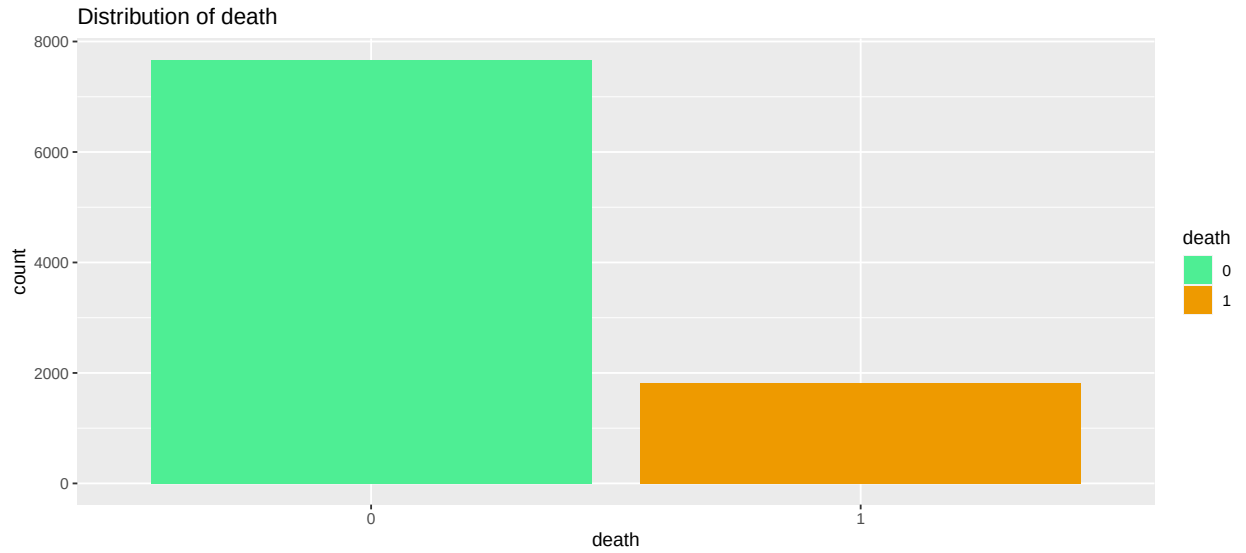

Figure 9: Distribution of deaths

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, shows us that those who survive (0) are on average younger than those who do not survive, which is logical, as younger people likely fair better in trauma accidents (at risk of significant bleeding, which this trail assesses). Aside from the median being slightly smaller than for those who survive, the distributions are fairly similar with a scew to the right. Participants in this trial are on average are quite young, roughly 3-, which could be due to sampling, but also because younger people are more likely to suffer from trauma accidents due to the jobs that they do and a increased propensity for risk taking behaviour.

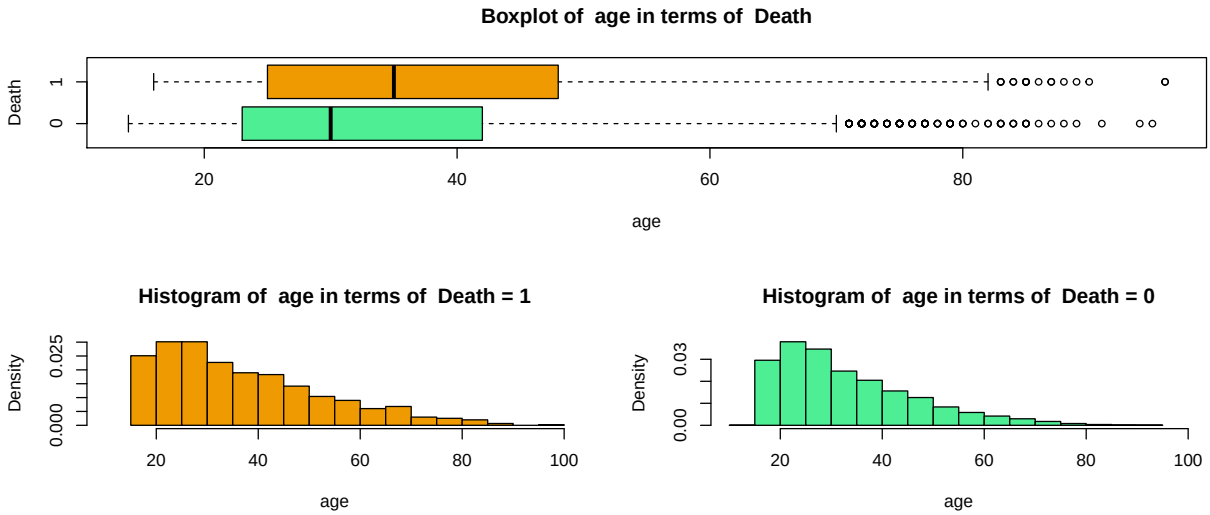

Figure 10: Distribution of Age in terms of death

Figure 11 showing *log* injury time in comparison, those who do not survive have longer relative injury time on average, and the IQR is also wider as well as the whiskers. However, there are still values of those with longer injury times who did survive, however they appare less frequently.

Figure 12 showing sbp (Systolic Blood Pressure) against death, shows that those who survive (O), have distribution lightly to the left of that of the those who do not survive. This could be logical as those who survive have a higher blood pressure, initially, then those who eventually do not.

Figure 11 compares *log* Respiratory rate (rr) with those who do not survive (1) have a slightly higher mean those who survive (1). This seems to lead us to believe, that those who do not survive, in some cases more often have a higher than average respiratory rate.

Figure 14, similar to *log* Respiratory Rate, the distribution and median of *log* Heart rate for those who do not survive is slightly higher than who survive.

Figure 15, also resembling hr and rr, cc (Central Capillary Refill Time in seconds) has a distribution further to the right, although more extreme than hr and rr, which could show that cc increases like hr and rr on average for people who will not survive. We also remain to see a limited number of of observations to the the extreme right, which could be outliers.

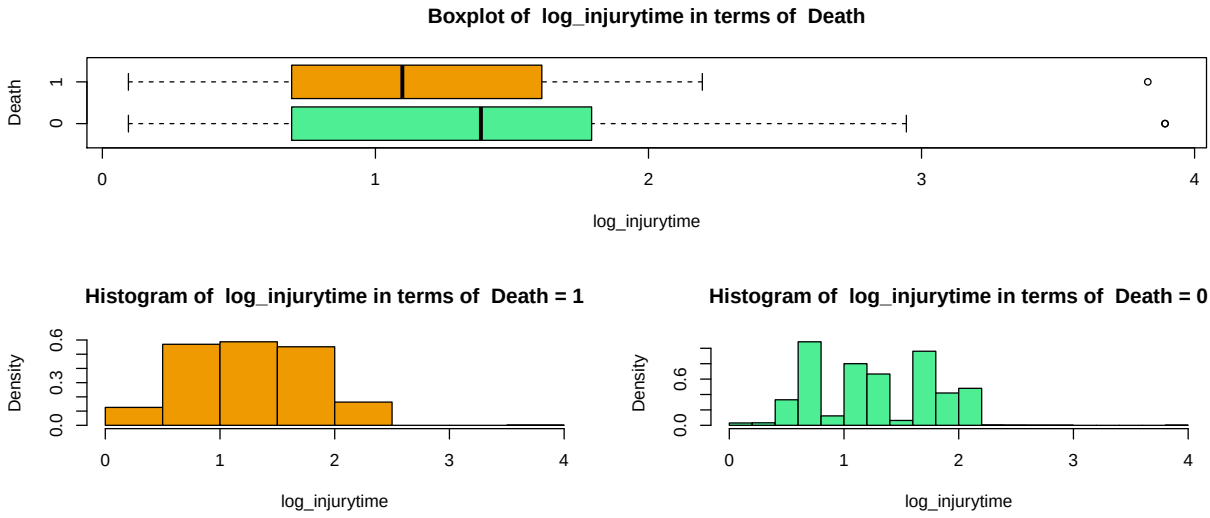

Figure 11: Distribution of the log injurytime in terms of death

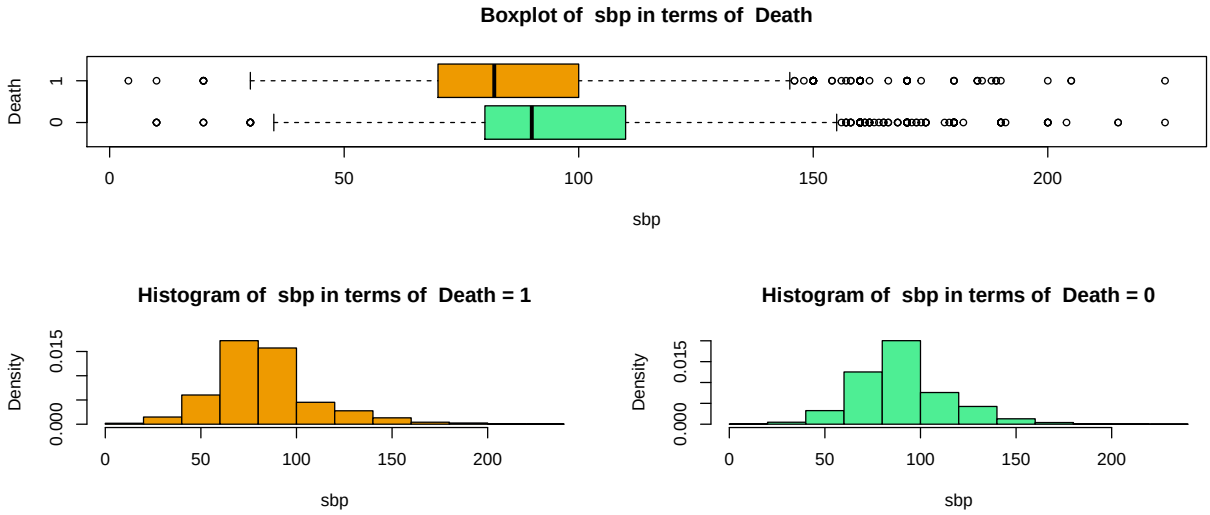

Figure 12: Distribution of sbp (Systolic Blood Pressure) in terms of death

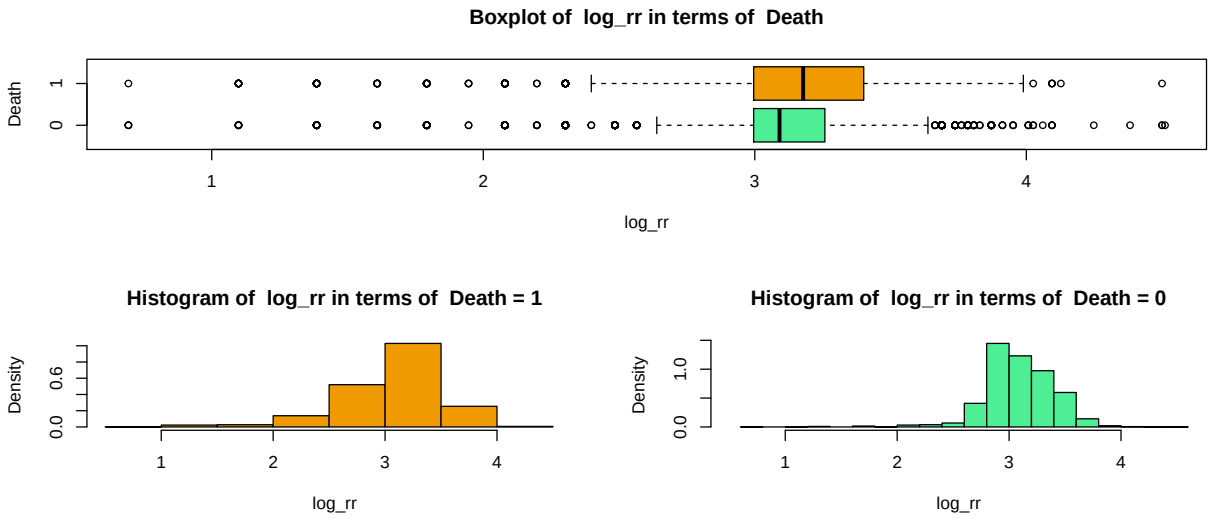

Figure 13: Distribution of log transformation of rr (Respiratory Rate) in terms of death

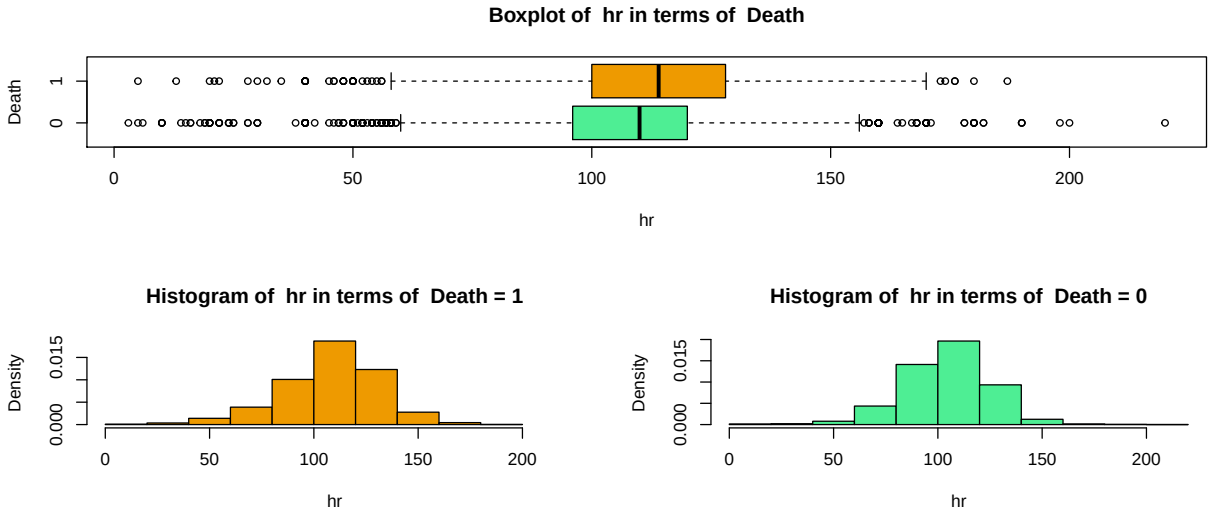

Figure 14: Distribution of hh (hearth Rate) in terms of death

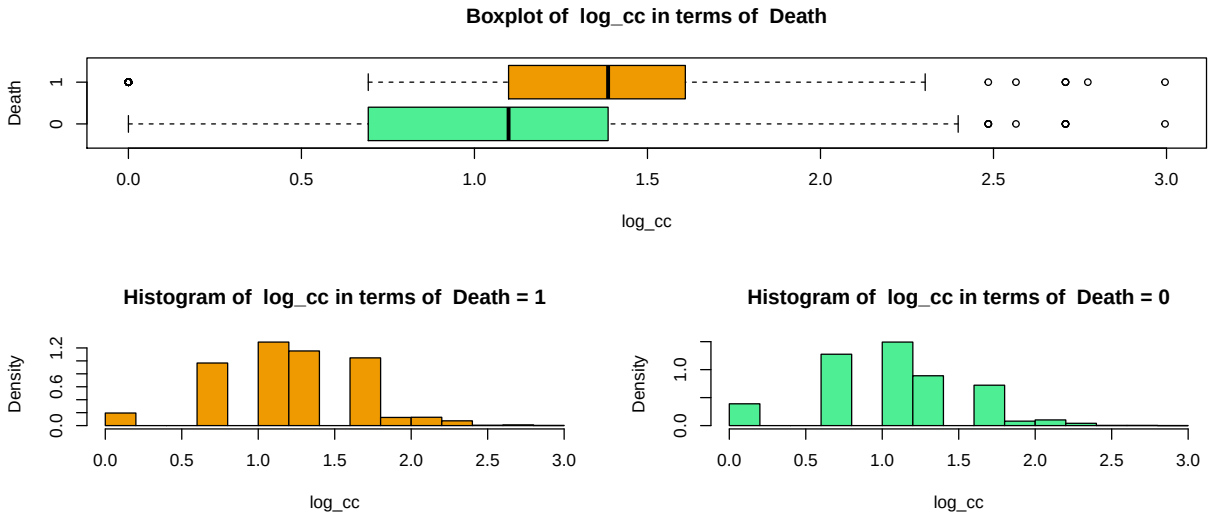

Figure 15: Distribution of log transformation of cc (Central capillary) in terms of death

Figure 16, shows that days in the ICU (ndaysicu), has a very similar distribution with an identical median, likely due to the fact that most people who suffer from trauma injuries only spend a very limited relative amount of time in ICU. We see that, for those who survive (0), the 75% quartile is slightly larger, and when reviewing the histogram, more observations are on the further right end of the spectrum (more relative days in hospital). This possibly indicates to us that those who are more likely to survive are kept in the ICU for longer to heal some cases.

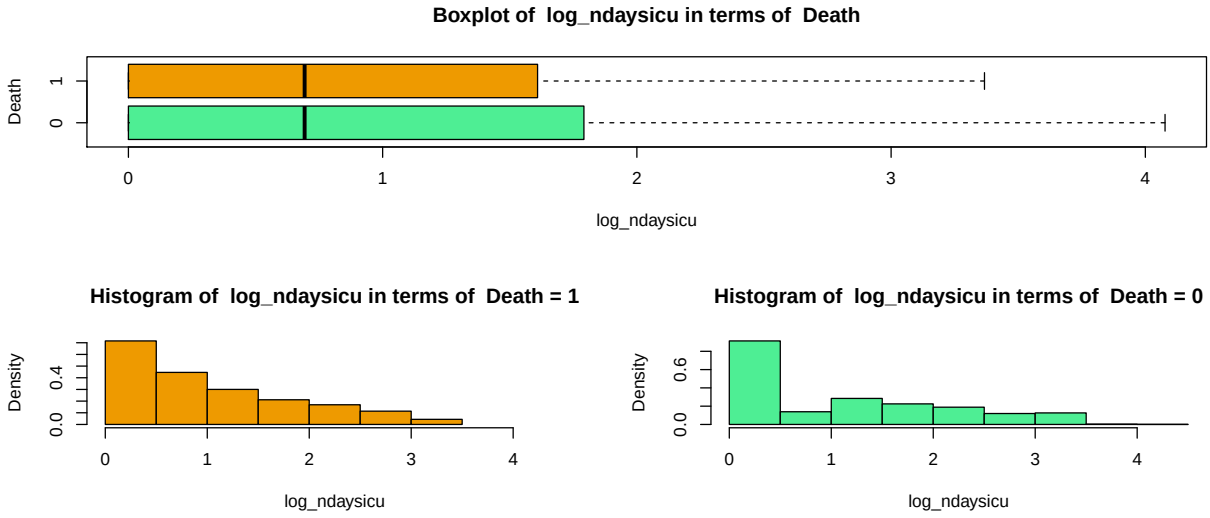

Figure 16: Distribution of log ndaysicu in terms of death

Figure 17 shows the *log* value of red blood cells transfused (ncell) accounting for death. Those who do not survive (1), have a broader distribution with the 75% quartile being larger, possibly indicating that for those who will not surive, more blood cells are transfused for more grave injuries in some cases.

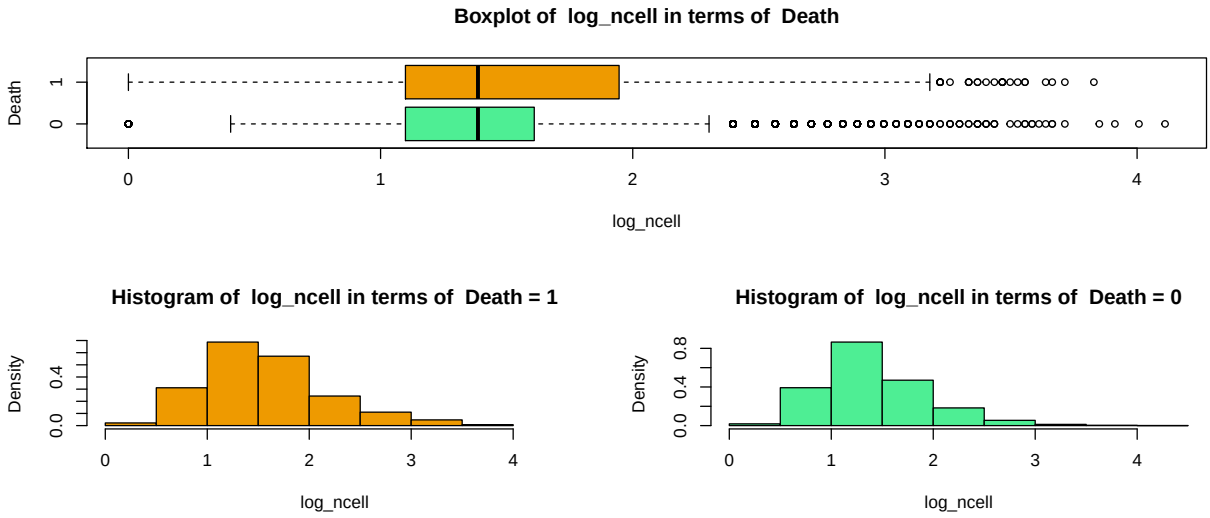

Figure 17: Distribution of log ncell in terms of death

Finally, the scatter plot in Figure 18 we cannot see any significant relationships but the variables, while accounting for Death. There does not seem be any visual relationship between any of the variables that are outstanding.

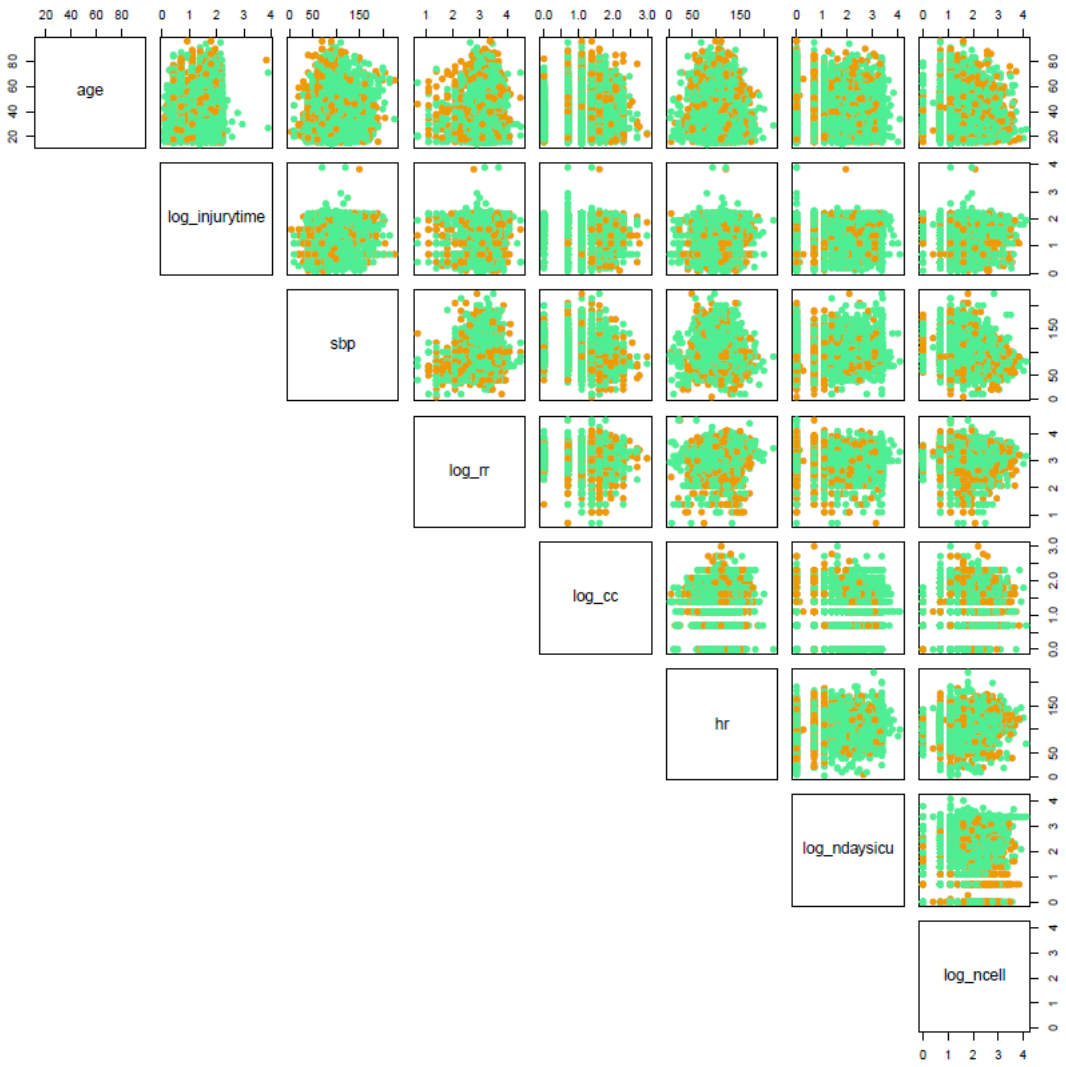

Figure 18: Scatter plot of all quantitiative variables

The PCP plot in Figure 19, generally matches with our observations about the scatterplot, that comparing those who survive and do not, follow the same visual pattern. For *log_cc*, *hr*, and *log_ndaysicu*, those who survive appear to have values that are more broadly dispersed, but this could be due to the fact those who survive represent a majority of the sample size. Although, comparing with the histograms and box plots, *log_cc*, *hr*, and *log_ndaysicu* do appear to have more dispersed distributions and wider IQRs.

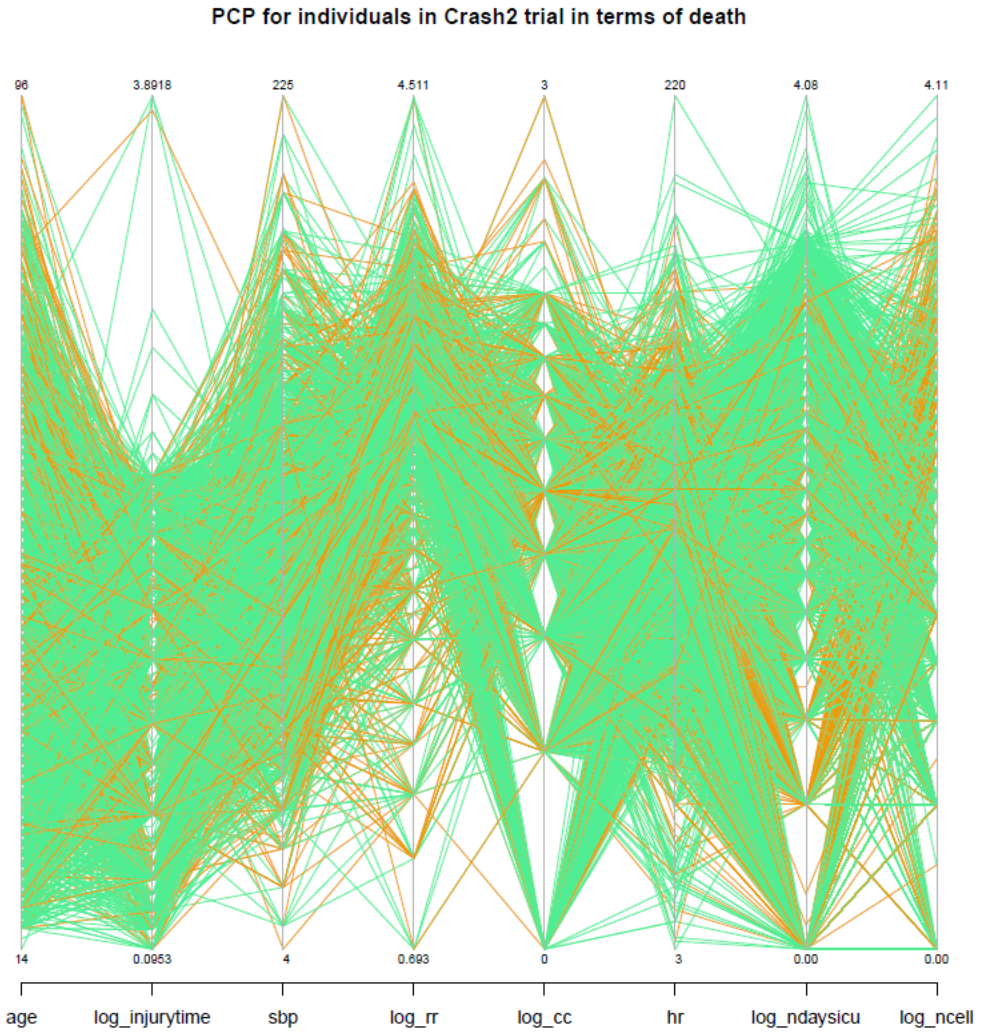

Figure 19: PCP plot of all quantitiative variables

The Andrews' plot in Figure 20, matches our previous analysis in Figure 18 and 19, that variables share similar distributions and relationships when accounting for Death, as the Andrew Plot does not show any dramatic differences between the two groups.

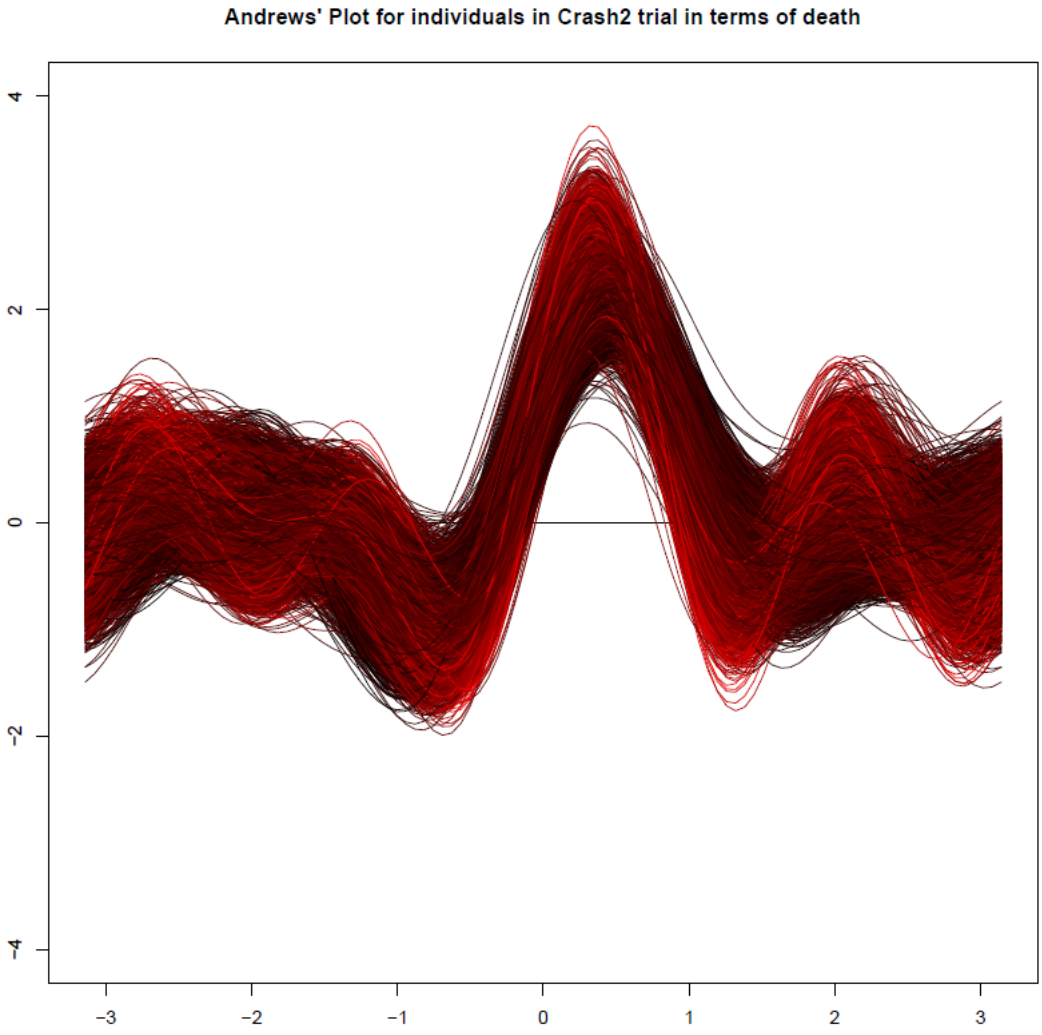

Figure 20: Andrews' plot of all quantitiative variables

In conclusion, we see that when accounting for difference categorical variables and Death, the distributions for our quantitative variables do not change significantly. We can see some minor changes in distributions and means. COMPLETE LATER.

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

Sample Estimators

Sample mean

Below is the sample mean vector for all the variables in our analysis, without controlling for death.

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

This table shows the mean while setting death to 1, or for those who do not survive.

Table 4: Sample mean: Do not survive

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

Now provided is the Sample mean for the variables when the patient survives, or when death = 0.

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

Although touched upon above, comparing the means when controlling for *death*, overall, we do see some differences in the variables. The means of those who survive are more similar to the overall sample mean as those who survive represent more of the sample population. In terms of difference between means we see that age is slightly higher for those who do not survive, which is logical, as those who are younger will likely be of better overall health and survive a trauma. *Sbp*, *hr*, *log_rr*, *log_cc* and *log_ncell* are all higher for those who do not survive possibly showing that those who have a higher Systolic Blood Pressure, Heart Rate, Respiratory Rate, continue...

Sample covariance

Below are the sample covariance matrices for the general sample, and then for those who do not survive (1) and those who survive (0). The diagonal is are the variances of the variables and other values outside of the diagonals identify the covariances for the corresponding variables on the in the row and column. First, below, is the Sample covariance matrix for the entire data set without controlling for death.

Table 6: Sample covariance matrix

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

Second, here, we provide the Sample covariance matrix for patients who did not survive ($death = 1$).

Table 7: Sample covariance matrix: Did not survive

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

The following below is the Sample covariance matrix for patients who survived ($death = 0$).

Table 8: Sample covariance matrix: Survive

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

We do not see in many of the categories, large ...

Ohter additional measure of dispersion is called the *Generalized variance*, which is defined as the determinant of the sample convariance. In the case of the three previous covaraince table, the generalized variance is given by the following table:

Table 9: Generalized variance

All population	Death	Survive
114978.7	313091.3	74406.57

Sample correlation

Below are the sample correlation matrices and the correlation plots for the general sample, and then for those who do not survive (1) and those who survive (0). The values outside of the diagonals identify the correlations for the corresponding variables on the in the row and column.

Below is the sample correlation matrix for the entire sample, without considering *death*.

Table 10: Sample correlation matrix

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

Figure 21, shown here is the correlation plot for the entire sample.

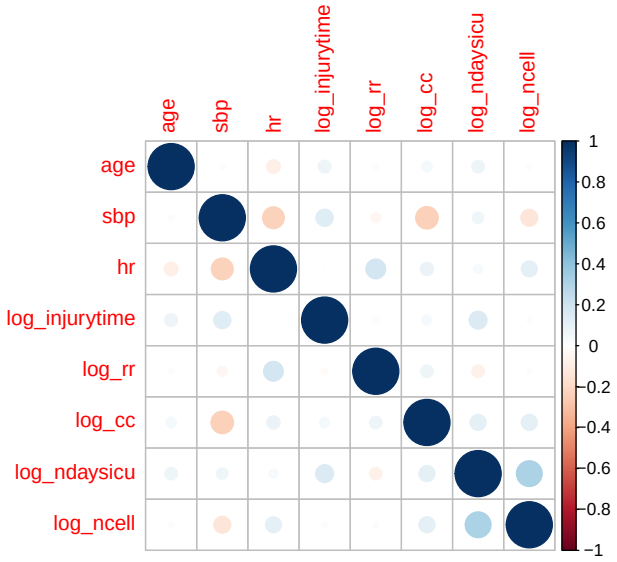

Figure 21: Sample Correlation Crash2 Dataset

Reviewing Figure 21 and the general sample covariance matrix we see that most values are fairly close to zero meaning most variables do not have a linear correlation.This is expected as our scatter plots did not visually reveal any linear relationships. The maximum absolute value we see is the correlation between log_ncell and

log_ndaysicu at 0.3113. Other correlation coefficients that exceed the absolute value of 0.1 are log_ndaysicu and log_injurytime, sbp and log_injurytime, log_rr and hr as well as hr and sbp at -0.2225. According to Ratner (2009), correlation coefficients with values below 0.3 are considered weak or minimal relationships. All of our correlation coefficients are within this category. Log_ncell and log_ndaysicu are the only ones to reach the threshold of a "moderate relationship". The fact that log_ncell and log_ndaysicu have a limited moderate relationship could be possibly be due to that fact those who stay in hospital for longer need more blood and those who have serious injuries need more blood and need to stay in hospital more often. We very small positive correlations between most variables but appears that Blood Pressue (sbp) has minor negative correlations with most variables, particularly with Heart Rate (hr), Central Capillary Refill Time (log_cc) and even more minorly, Number of Blood Cells transfused (log_ncell) (SPOKEN ABOUT LATER).

Below is the sample correlation matrix for those who did not survive (*death*=1)

Table 11: Sample correlation matrix: Did not survive

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

The following is the sample correlation matrix for those who survived (*death*=0)

Table 12: Sample correlation matrix survival patients

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

	age	sbp	hr	log_injurytime	log_rr	log_cc	log_ndaysicu	log_ncell
log_ncell	0.0151619	-0.1190669	0.1058514	0.0385896	0.0005018	0.1245894	0.3492684	1.0000000

Heatmap 22 ... and Figure 23, as provided are correlation plots while controlling for death

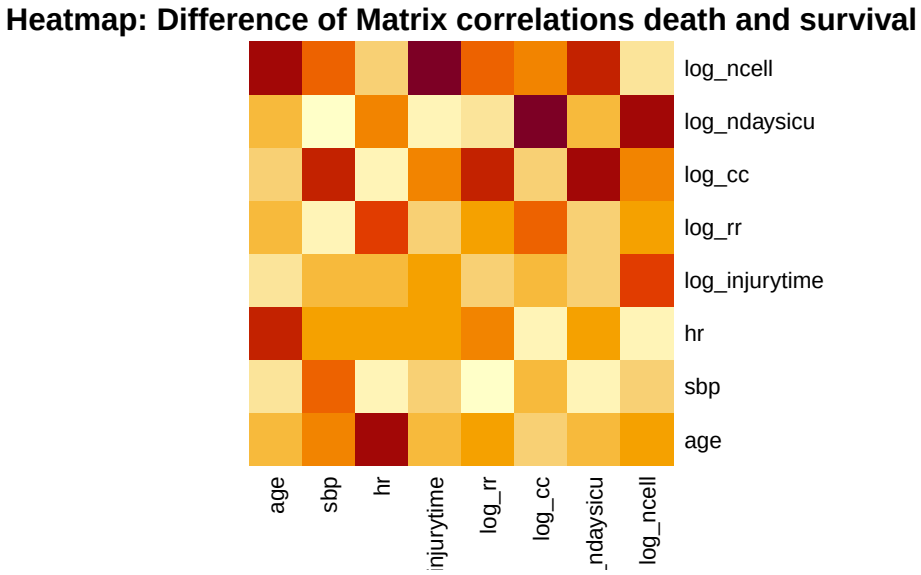

Figure 22: Heatmap difference between correlation atrices

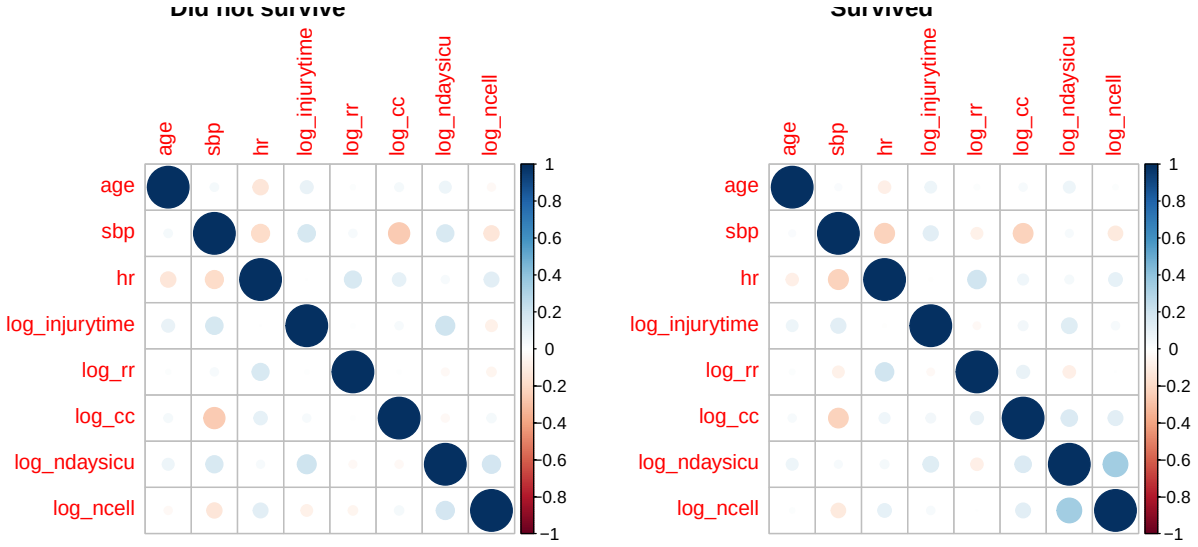

Figure 23: Sample Correlation for death and survival patients from Crash2 Dataset

We do not see any changes in relationships for those who do not survive and people that do. We do see that some correlations seem to get more minor for those who survive but this is likely due to the larger sample size of those who survive as their variances will be less. Some variances do increase for those who survive such as *log_ndaysicu* and *log_ncell*, almost doubles, and rises to a level of correlation at 0.3493 that is consider moderate (Ratner, 2009). This may be because, for those that do not survive, they receive a lot of blood to

attempt to save their lives, but do not survive long enough to stay in the ICU due to the severity of their trauma. This therefore would limit the covariance. For those who survive, they possibly could remain in hospital longer therefore receive more blood over those days, incresing the covariance.

TO INSERT: In conclusion, that since there is not much colinearity in these variables there is good potential for regression, however, the issue is that the two categories for death are not easily seperated, which means the analysis would likely not be accurate.

Outliers

INSET SOMEHWERE: We don't use shrinkage analysis because we have many more observations ($n = 9497$) than 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.

Table 13: Sample robust mean

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 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.

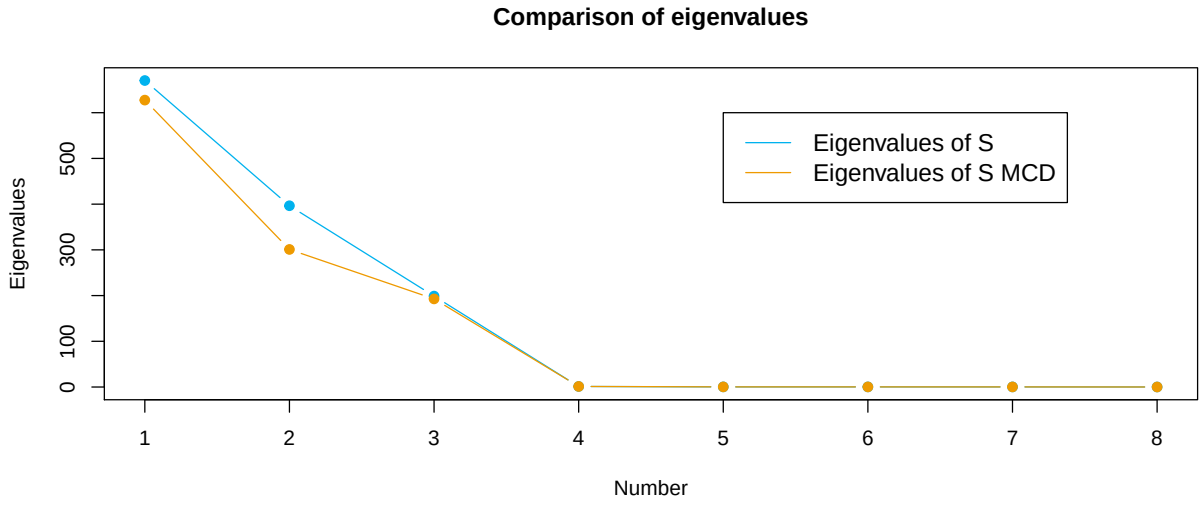

Figure 24: Comparison eigenvalues sample and robust covariance

The eigenvalues are used

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.

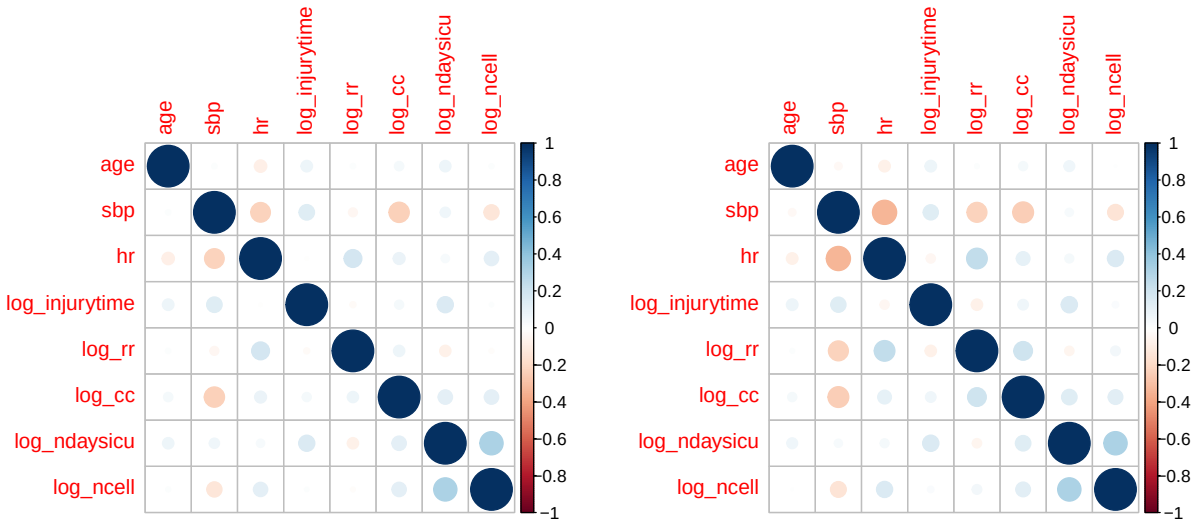

Figure 25: Correlation all population and MCD correlation

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.

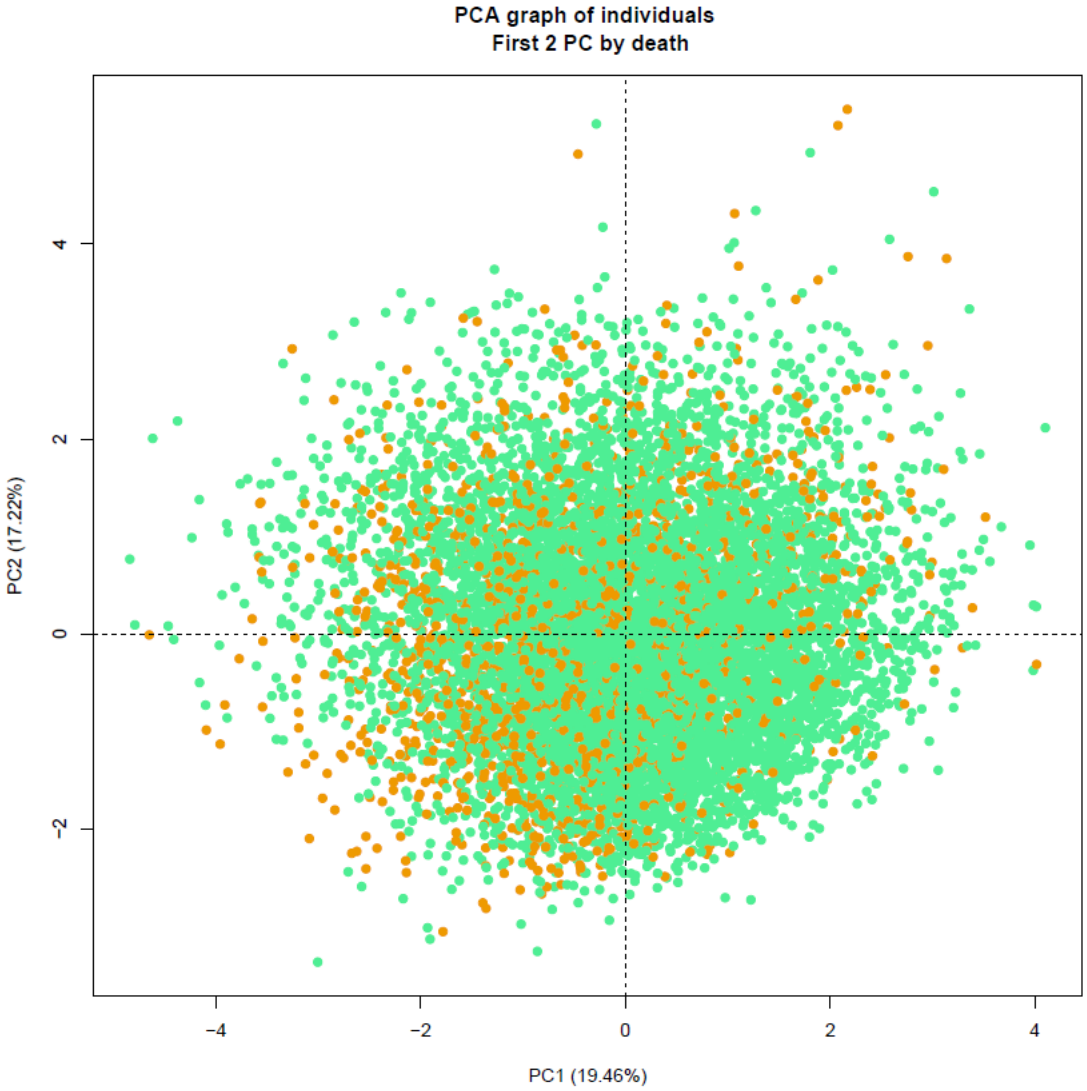

Figure 26: First two PCs based on the sample covariance all population

On the other hand, we will analize the weights of the fist two principal components. This can be seen in the figure 27. 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.

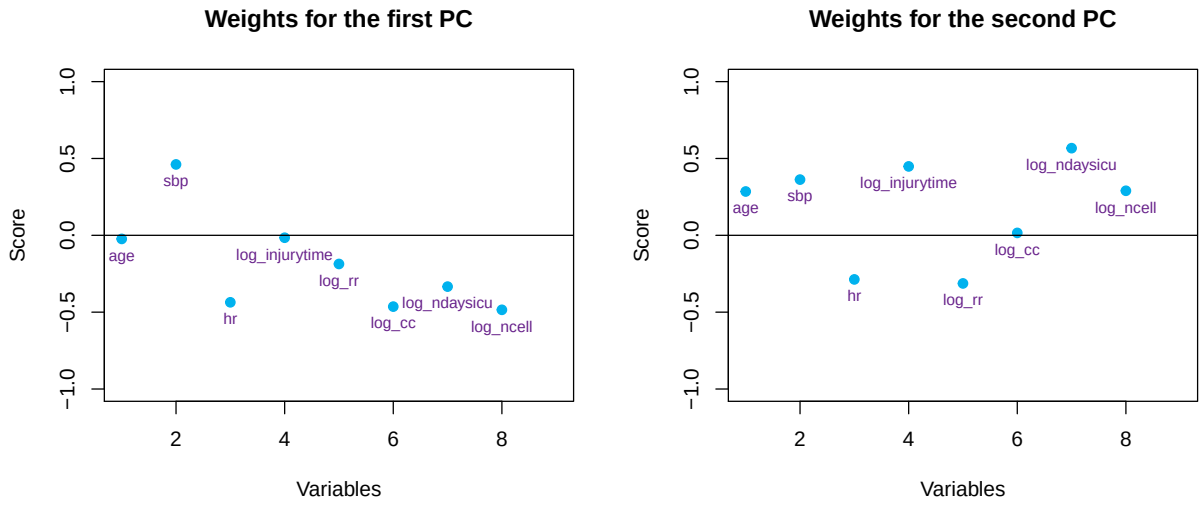

Figure 27: Loadings for the first and sencond PCs individually

On the other hand, when we consider the loadings for the first two PCs (figure 28) 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.

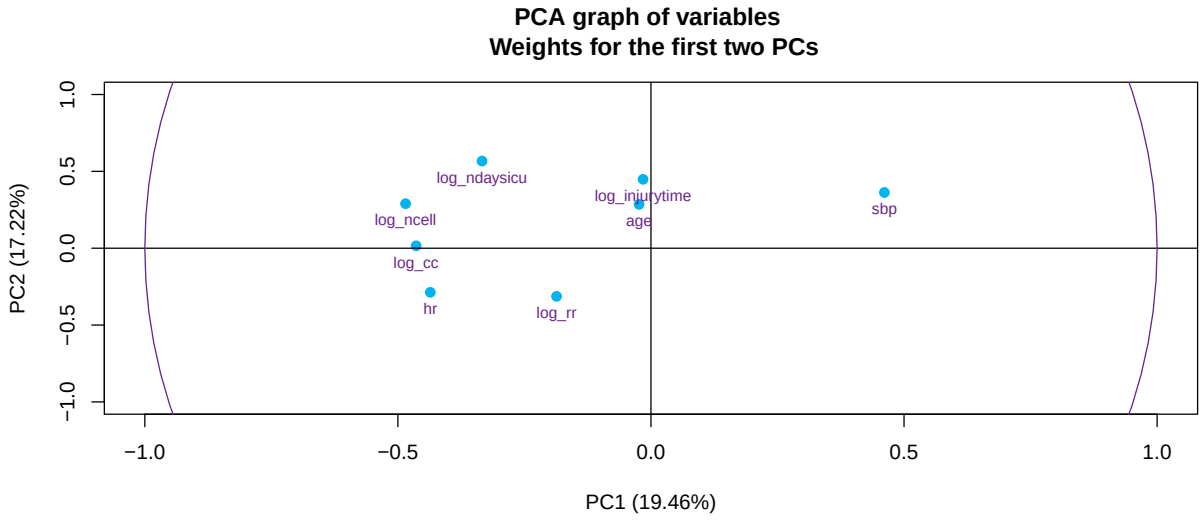

Figure 28: Loadings for the first two PC

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*

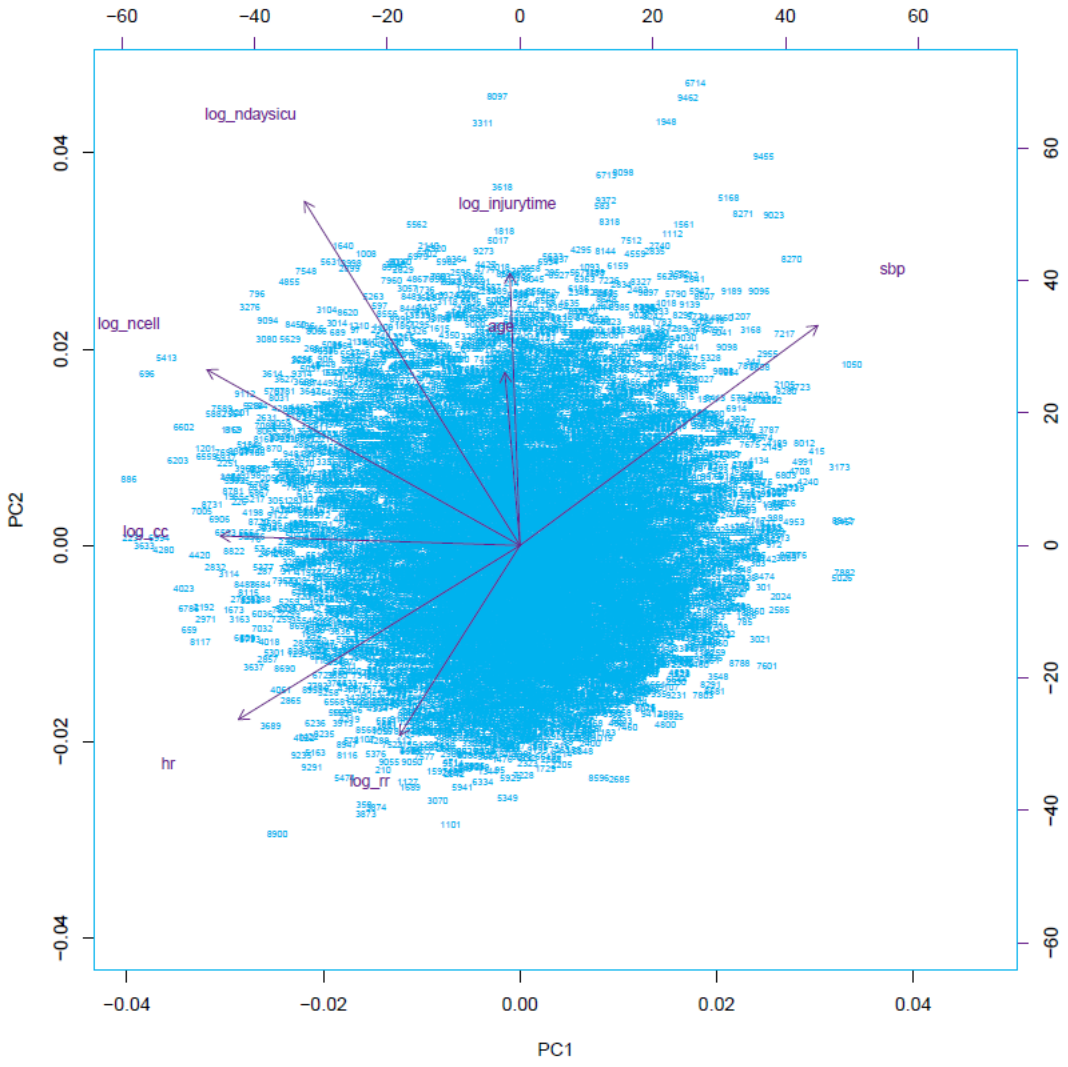

Figure 29: PC Scores and PC loading all population

The figure 30 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.

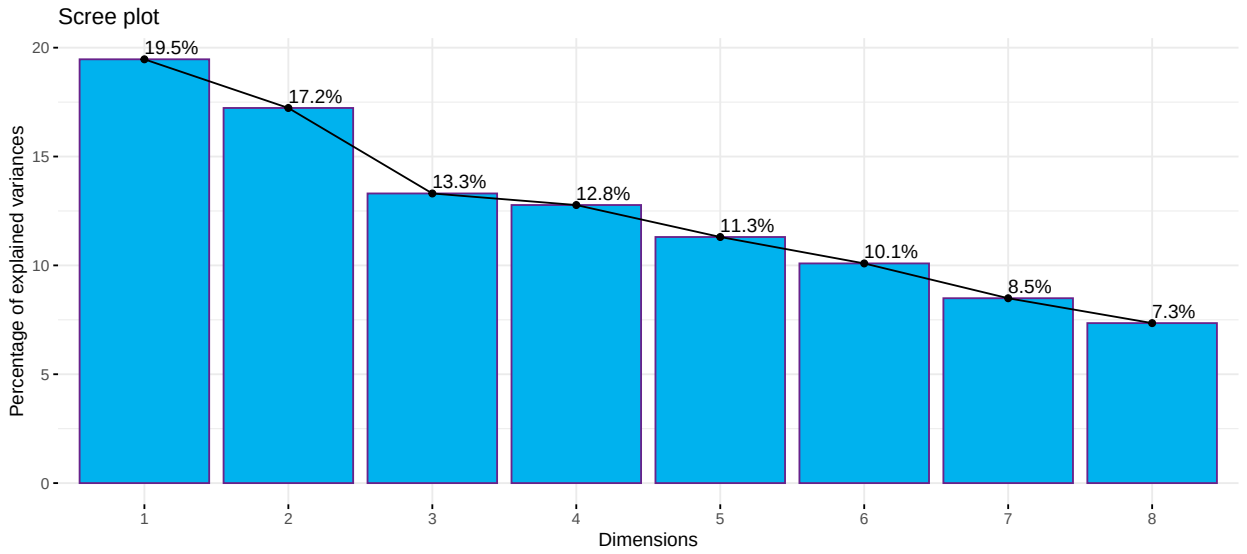

Figure 30: Eigenvalues of the sample correlation matrix

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 32 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.

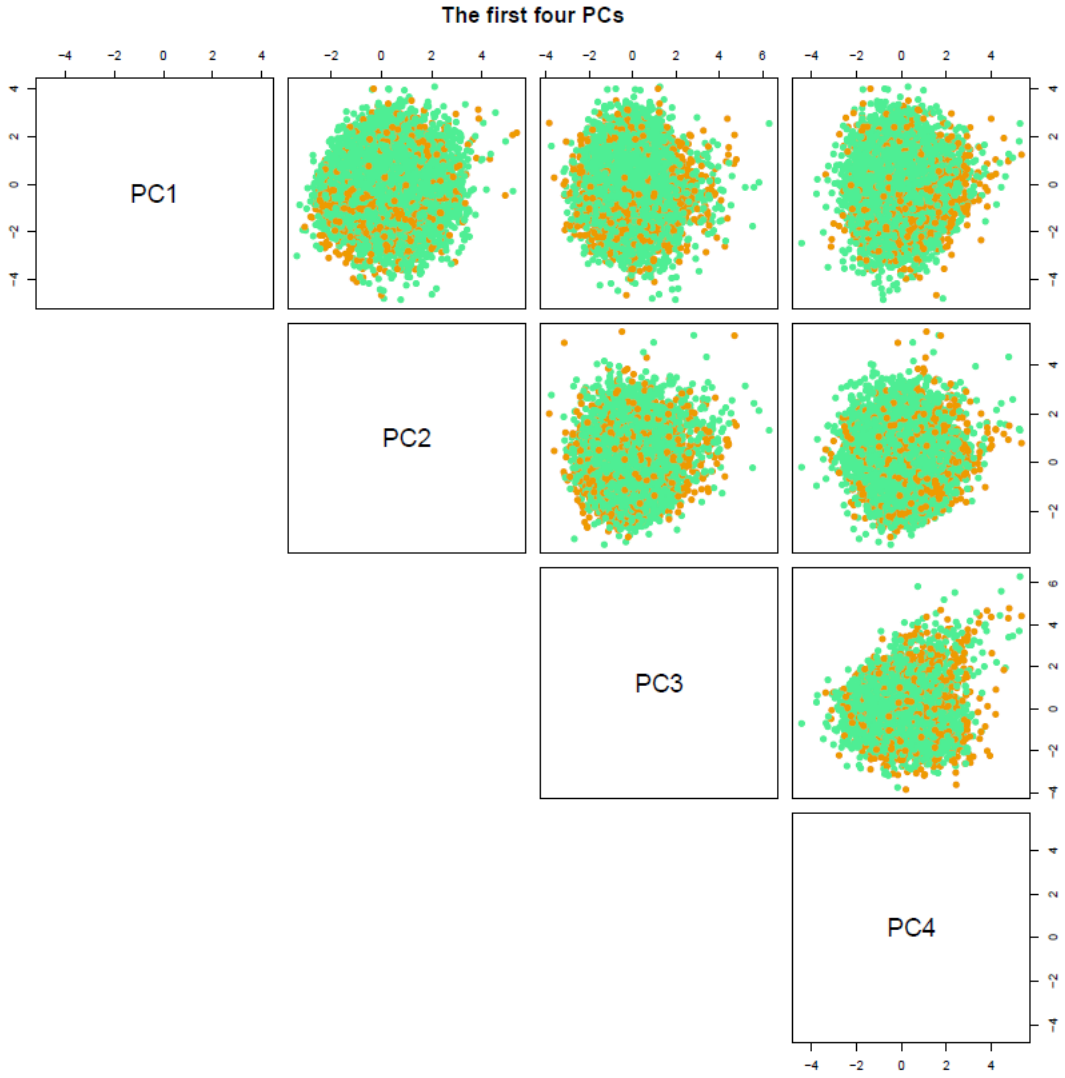

Figure 31: PC Scores and PC loading all population

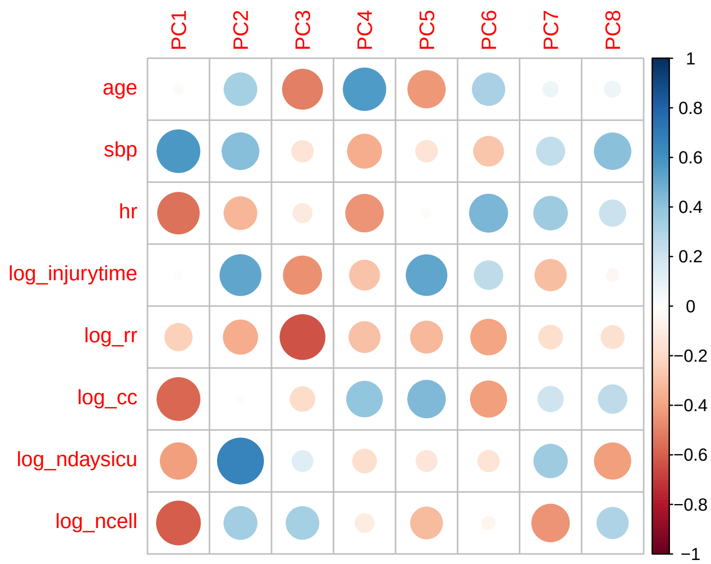

Figure 32: Correlation between dataset and all PC

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 34...

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 36 and 37 ...

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.

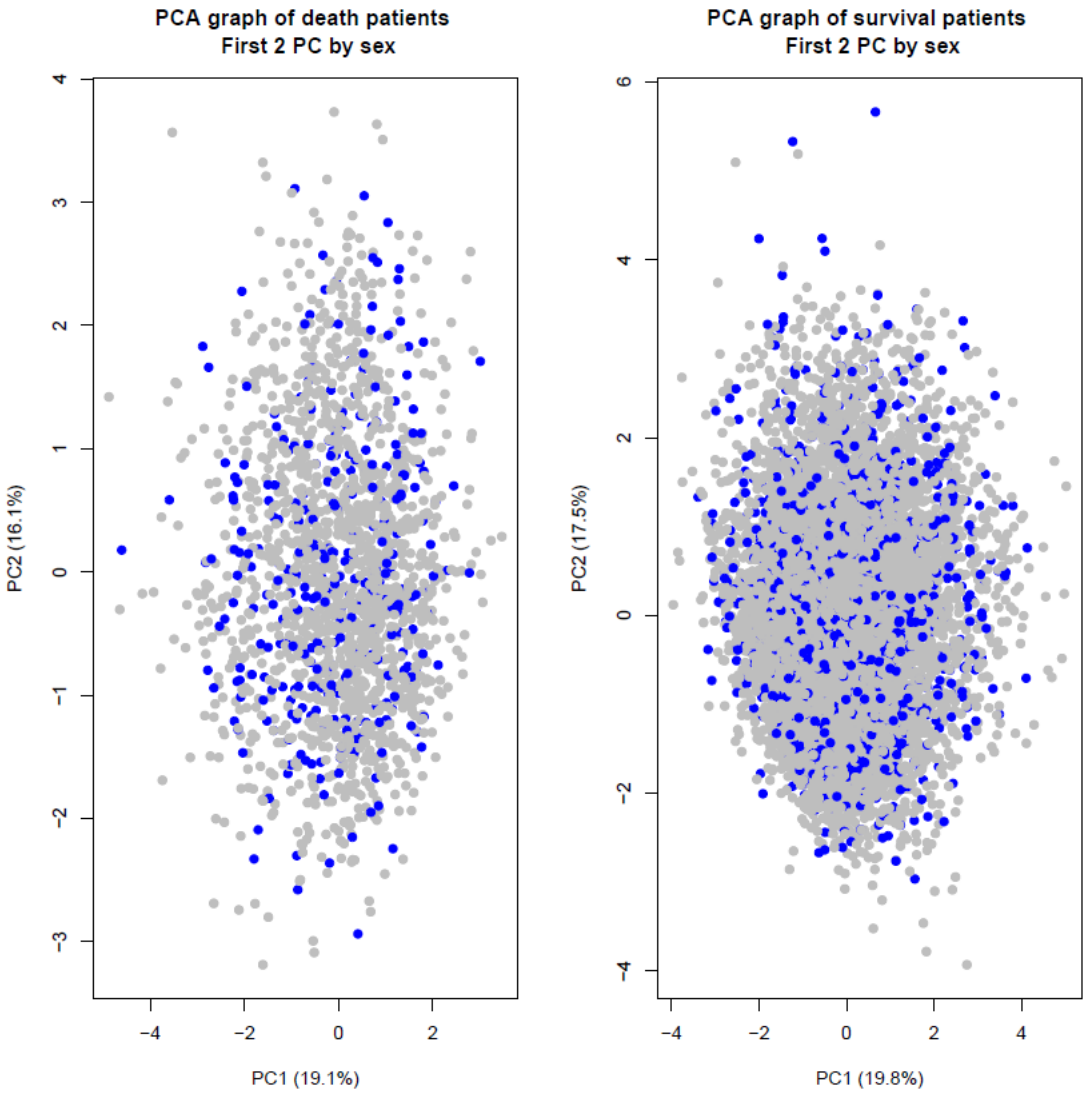

Figure 33: PC Scores and PC loading all population

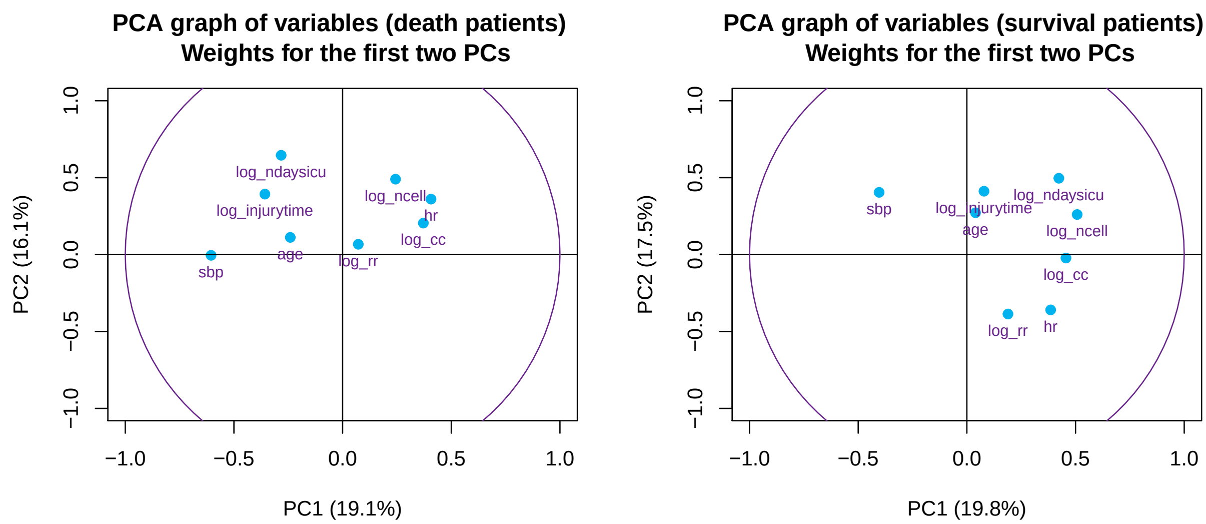

Figure 34: Loadings for the first two PC by death and survival patients

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, Rodriguez-Artalejo F, Graciani A, Guallar-Castillon 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

3. Ratner, B. The correlation coefficient: Its values range between $+1/-1$, or do they? J Target Meas Anal Mark 17, $139-142$ (2009). https://doi.org/10.1057/jt.2009.5

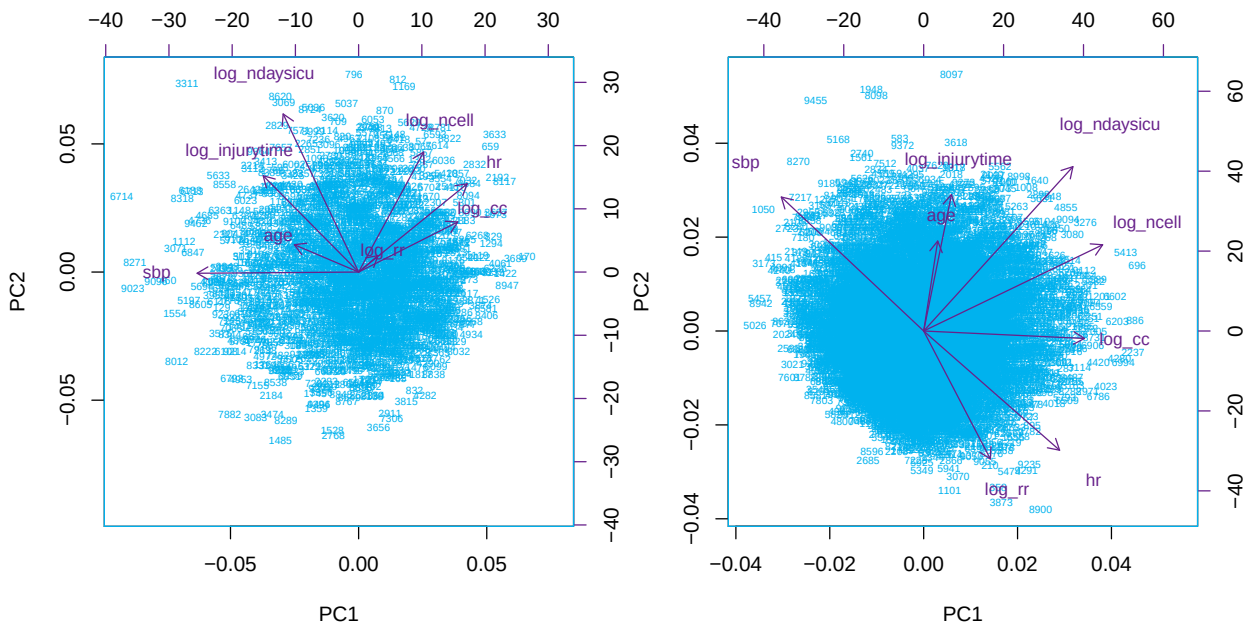

Figure 35: PC Scores and PC loading all population for death and survival patiens

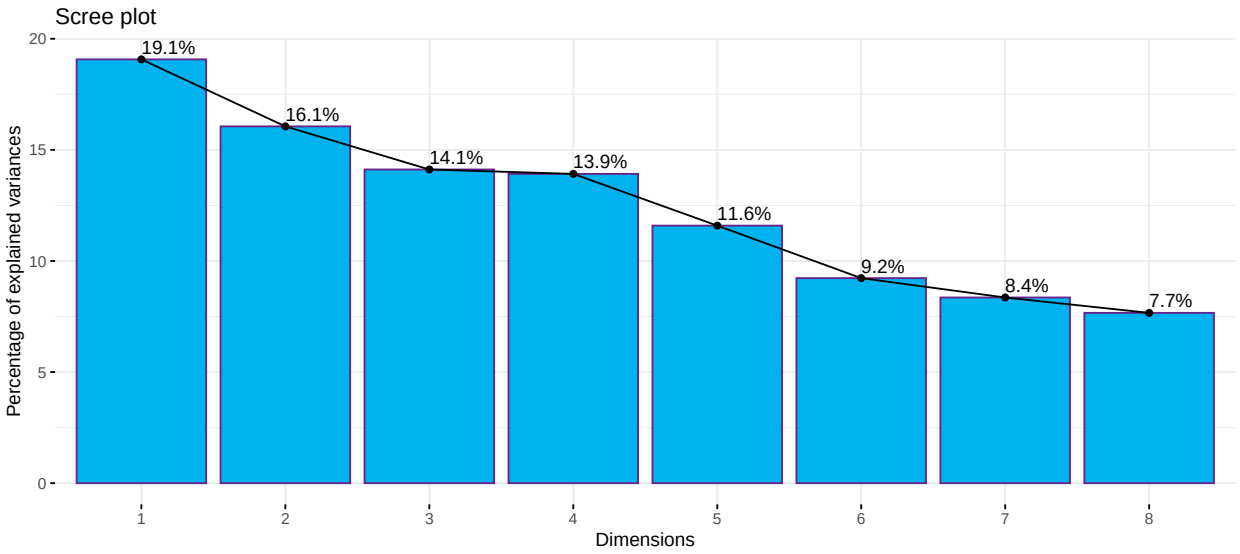

Figure 36: Eigenvalues of the sample correlation matrix for death patients

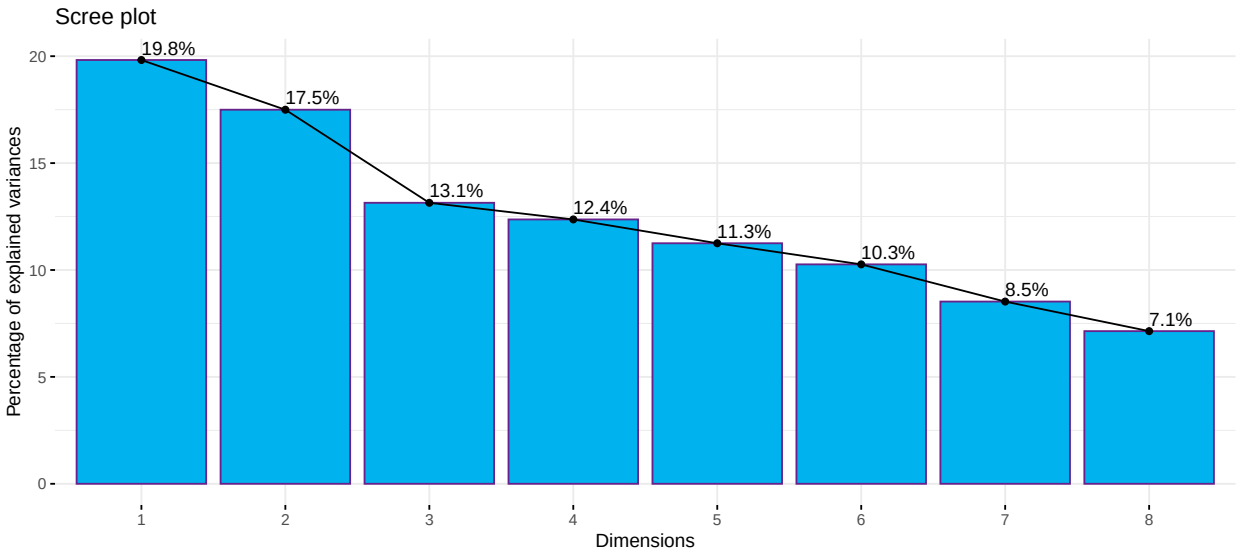

Figure 37: Eigenvalues of the sample correlation matrix for survival patients