SURVIVAL ANALYSIS - NCDB

Kelvin Ofori-Minta University of Texas at El Paso (UTEP)

July 05, 2022

Contents

1 Loading Data and Preparations

```
lung$Chemo<-factor(lung$Chemo,
                     levels=c("No Chemo", "Chemo"),
                     labels=c("No Chemo", "Chemo"))

lung$AGE_cat<-factor(lung$AGE_cat,
                     levels = c("50-60", "60-70", "70-80", "Above 80", "Below 50"),
                     labels = c("50-60", "60-70", "70-80", "Above 80", "Below 50"))

lung$SEX<-factor(lung$SEX,
                   levels=c("Female", "Male"),
                   labels=c("Female", "Male"))

lung$CDCC_TOTAL_BEST<-factor(lung$CDCC_TOTAL_BEST,
                             levels = c("0","1","2","3"),
                             labels = c("0","1","2","3"))

lung$TUMOR_SIZE_cat<-factor(lung$TUMOR_SIZE_cat,
                            levels=c("<=1cm","1cm-2cm","2cm-3cm","3cm-4cm","4cm-5cm"),
                            labels=c("<=1cm","1cm-2cm","2cm-3cm","3cm-4cm","4cm-5cm"))

lung$GRADE<-factor(lung$GRADE,
                   levels=c("Moderately differentiated","Poorly differentiated", "Undifferentia
                   labels=c("Moderately differentiated","Poorly differentiated", "Undifferentia

lung$Pathology<-factor(lung$Pathology,
                       levels = c("Adenocarcinoma","Other", "Squamous"),
                       labels = c("Adenocarcinoma","Other", "Squamous"))


lung$Visceral_Pleural_Invasion<-factor(lung$Visceral_Pleural_Invasion,
                                       levels = c("Other", "Present"),
                                       labels = c("Other", "Present"))

lung$LYMPH_VASCULAR_INVASION2<-factor(lung$LYMPH_VASCULAR_INVASION2,
                                      levels=c("Absent", "Present", "Unknown"),
                                      labels=c("Absent", "Present", "Unknown"))

lung$Margins<-factor(lung$Margins,
                     levels = c("Other","Positive","Zero"),
                     labels = c("Other","Positive","Zero"))

lung$Lymph_Nodes_Sampled<-factor(lung$Lymph_Nodes_Sampled,
                                 levels = c("<10",">=10", "Unknown"),
                                 labels = c("<10",">=10", "Unknown"))
```

```r
lung$Excision_less_than1<-factor(lung$Excision_less_than1,
                                 levels = c("FALSE","TRUE"),
                                 labels = c("FALSE", "TRUE"))
```

1.1 Partition Data

```r
require(caTools)
set.seed(1)
split = sample.split(lung$DX_LASTCONTACT_DEATH_MONTHS,SplitRatio = 0.85)
train=subset(lung, split==T)
test=subset(lung, split==F)
```

2 COXPH model for predictors of mortality - ALL DATA

```
library("survival")
# library("survminer")

lung$Chemo=relevel(as.factor(lung$Chemo), ref="No Chemo")
cox_fit1 <- coxph(Surv(DX_LASTCONTACT_DEATH_MONTHS,PUF_VITAL_STATUS) ~ Chemo +
                      AGE_cat+
                      SEX +
                      CDCC_TOTAL_BEST  +
                      TUMOR_SIZE_cat  +
                      GRADE  +
                      Visceral_Pleural_Invasion+
                      LYMPH_VASCULAR_INVASION2+
                      Margins  +
                      Lymph_Nodes_Sampled +
                      Excision_less_than1,
                    data = lung)

cox_fit1$coefficients #odds
```

```
##                       ChemoChemo                     AGE_cat60-70
##                    -0.0328323599                    -0.0427297280
##                     AGE_cat70-80                  AGE_catAbove 80
##                    -0.0325108982                    -0.0430062924
##                  AGE_catBelow 50                          SEXMale
##                    -0.0909712523                    -0.0510721635
##                 CDCC_TOTAL_BEST1                 CDCC_TOTAL_BEST2
##                    -0.0021252500                    -0.0085577321
##                 CDCC_TOTAL_BEST3            TUMOR_SIZE_cat1cm-2cm
##                     0.0285513604                    -0.0839515534
##            TUMOR_SIZE_cat2cm-3cm            TUMOR_SIZE_cat3cm-4cm
##                    -0.1181348822                    -0.0801962360
##            TUMOR_SIZE_cat4cm-5cm       GRADEPoorly differentiated
##                    -0.0791002404                    -0.0002581338
##            GRADEUndifferentiated                     GRADEUnknown
##                     0.0907022956                     0.0039056293
##         GRADEWell differentiated Visceral_Pleural_InvasionPresent
##                    -0.0500901193                     0.0572583599
##  LYMPH_VASCULAR_INVASION2Present  LYMPH_VASCULAR_INVASION2Unknown
##                    -0.0234197701                    -0.1408796161
##                  MarginsPositive                      MarginsZero
##                     0.0570926793                     0.1810093997
##          Lymph_Nodes_Sampled>=10       Lymph_Nodes_SampledUnknown
##                     0.0522284067                    -0.2228574697
##          Excision_less_than1TRUE
##                    -0.0199985495
```

```
exp(cox_fit1$coefficients) #HR
```

```
##                       ChemoChemo                     AGE_cat60-70 
##                        0.9677008                        0.9581703 
##                     AGE_cat70-80                  AGE_catAbove 80 
##                        0.9680119                        0.9579054 
##                  AGE_catBelow 50                          SEXMale 
##                        0.9130440                        0.9502101 
##                 CDCC_TOTAL_BEST1                 CDCC_TOTAL_BEST2 
##                        0.9978770                        0.9914788 
##                 CDCC_TOTAL_BEST3            TUMOR_SIZE_cat1cm-2cm 
##                        1.0289629                        0.9194758 
##            TUMOR_SIZE_cat2cm-3cm            TUMOR_SIZE_cat3cm-4cm 
##                        0.8885762                        0.9229352 
##            TUMOR_SIZE_cat4cm-5cm       GRADEPoorly differentiated 
##                        0.9239473                        0.9997419 
##            GRADEUndifferentiated                     GRADEUnknown 
##                        1.0949430                        1.0039133 
##         GRADEWell differentiated Visceral_Pleural_InvasionPresent 
##                        0.9511437                        1.0589294 
##  LYMPH_VASCULAR_INVASION2Present  LYMPH_VASCULAR_INVASION2Unknown 
##                        0.9768523                        0.8685939 
##                  MarginsPositive                      MarginsZero 
##                        1.0587539                        1.1984264 
##          Lymph_Nodes_Sampled>=10       Lymph_Nodes_SampledUnknown 
##                        1.0536164                        0.8002289 
##          Excision_less_than1TRUE 
##                        0.9802001
```

```
# lung$Chemo=relevel(as.factor(lung$Chemo), ref="No Chemo")
# cox_fit11 <- coxph(Surv(DX_LASTCONTACT_DEATH_MONTHS,PUF_VITAL_STATUS) ~
#                       AGE_cat+
#                       SEX +
#                       CDCC_TOTAL_BEST  +
#                       TUMOR_SIZE_cat  +
#                       GRADE  +
#                       Visceral_Pleural_Invasion+
#                       LYMPH_VASCULAR_INVASION2+
#                       Margins  +
#                       Lymph_Nodes_Sampled +
#                       Excision_less_than1,
#                     data = lung)
#
# cox_fit11$coefficients #odds
#
# exp(cox_fit11$coefficients) #HR
```

3 format results of cox model

```
require(kableExtra)
broom::tidy(cox_fit1 ,
            exp=TRUE) %>%
  kable()
```

term	estimate	std.error	statistic	p.value
ChemoChemo	0.9677008	0.0296857	-1.1059996	0.2687267
AGE_cat60-70	0.9581703	0.0312769	-1.3661761	0.1718837
AGE_cat70-80	0.9680119	0.0316406	-1.0275044	0.3041830
AGE_catAbove 80	0.9579054	0.0464372	-0.9261165	0.3543854
AGE_catBelow 50	0.9130440	0.0736842	-1.2346094	0.2169759
SEXMale	0.9502101	0.0218628	-2.3360333	0.0194895
CDCC_TOTAL_BEST1	0.9978770	0.0255017	-0.0833377	0.9335830
CDCC_TOTAL_BEST2	0.9914788	0.0361173	-0.2369425	0.8127014
CDCC_TOTAL_BEST3	1.0289629	0.0424419	0.6727165	0.5011277
TUMOR_SIZE_cat1cm-2cm	0.9194758	0.0911452	-0.9210746	0.3570115
TUMOR_SIZE_cat2cm-3cm	0.8885762	0.0919570	-1.2846749	0.1989059
TUMOR_SIZE_cat3cm-4cm	0.9229352	0.0937552	-0.8553788	0.3923415
TUMOR_SIZE_cat4cm-5cm	0.9239473	0.0952241	-0.8306742	0.4061577
GRADEPoorly differentiated	0.9997419	0.0252995	-0.0102031	0.9918592
GRADEUndifferentiated	1.0949430	0.1025368	0.8845825	0.3763818
GRADEUnknown	1.0039133	0.0381754	0.1023075	0.9185126
GRADEWell differentiated	0.9511437	0.0339343	-1.4760914	0.1399194
Visceral_Pleural_InvasionPresent	1.0589294	0.0362535	1.5793890	0.1142468
LYMPH_VASCULAR_INVASION2Present	0.9768523	0.0300965	-0.7781551	0.4364776
LYMPH_VASCULAR_INVASION2Unknown	0.8685939	0.0465820	-3.0243343	0.0024918
MarginsPositive	1.0587539	0.1968533	0.2900265	0.7717959
MarginsZero	1.1984264	0.1782101	1.0157079	0.3097685
Lymph_Nodes_Sampled>=10	1.0536164	0.0224537	2.3260485	0.0200160
Lymph_Nodes_SampledUnknown	0.8002289	0.0590297	-3.7753459	0.0001598
Excision_less_than1TRUE	0.9802001	0.0344942	-0.5797649	0.5620732

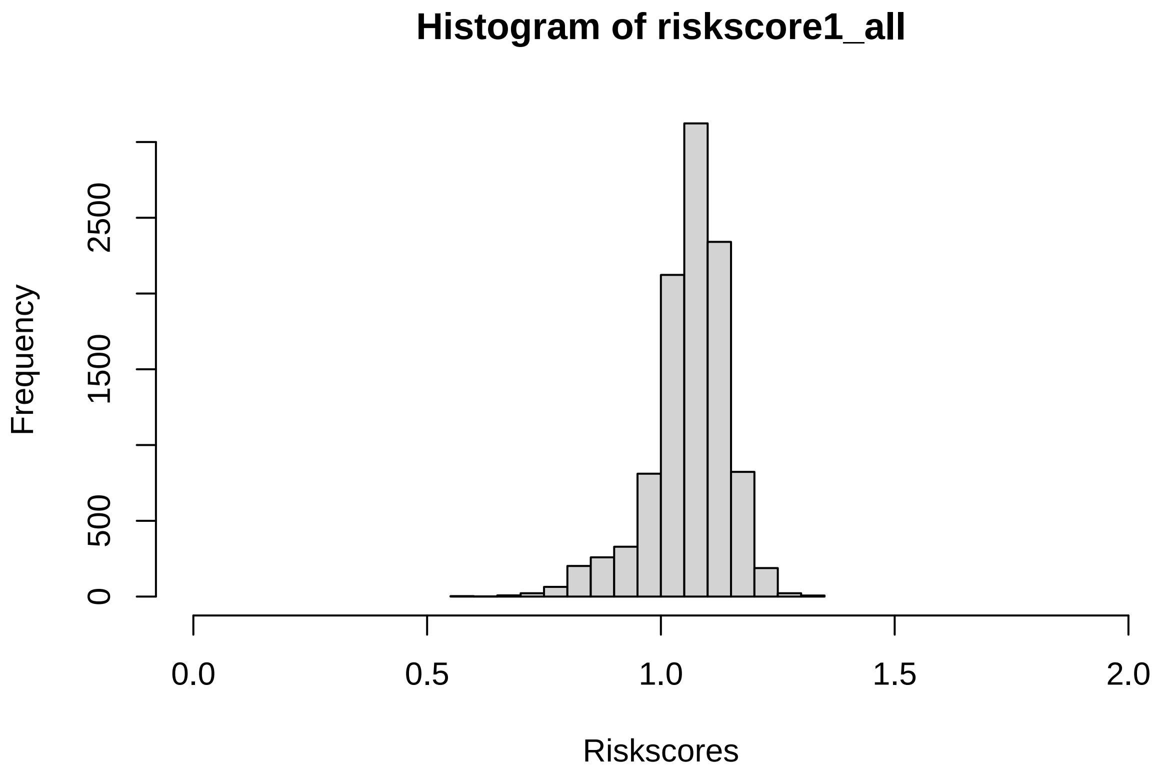
```
cox_fit1 %>%
  gtsummary::tbl_regression(exp=TRUE)
```

Characteristic	**HR**	**95% CI**	**p-value**
Chemo			
No Chemo			
Chemo	0.97	0.91, 1.03	0.3
AGE_cat			
50-60			
60-70	0.96	0.90, 1.02	0.2
70-80	0.97	0.91, 1.03	0.3
Above 80	0.96	0.87, 1.05	0.4
Below 50	0.91	0.79, 1.05	0.2
SEX			
Female			
Male	0.95	0.91, 0.99	0.019
CDCC_TOTAL_BEST			
0			
1	1.00	0.95, 1.05	>0.9
2	0.99	0.92, 1.06	0.8
3	1.03	0.95, 1.12	0.5
TUMOR_SIZE_cat			
<=1cm			
1cm-2cm	0.92	0.77, 1.10	0.4
2cm-3cm	0.89	0.74, 1.06	0.2
3cm-4cm	0.92	0.77, 1.11	0.4
4cm-5cm	0.92	0.77, 1.11	0.4
GRADE			
Moderately differentiated			
Poorly differentiated	1.00	0.95, 1.05	>0.9
Undifferentiated	1.09	0.90, 1.34	0.4
Unknown	1.00	0.93, 1.08	>0.9
Well differentiated	0.95	0.89, 1.02	0.14
Visceral_Pleural_Invasion			
Other			
Present	1.06	0.99, 1.14	0.11
LYMPH_VASCULAR_INVASION2			
Absent			
Present	0.98	0.92, 1.04	0.4
Unknown	0.87	0.79, 0.95	0.002
Margins			
Other			
Positive	1.06	0.72, 1.56	0.8
Zero	1.20	0.85, 1.70	0.3
Lymph_Nodes_Sampled			
<10			
>=10	1.05	1.01, 1.10	0.020
Unknown	0.80	0.71, 0.90	<0.001
Excision_less_than1			
FALSE			
TRUE	0.98	0.92, 1.05	0.6

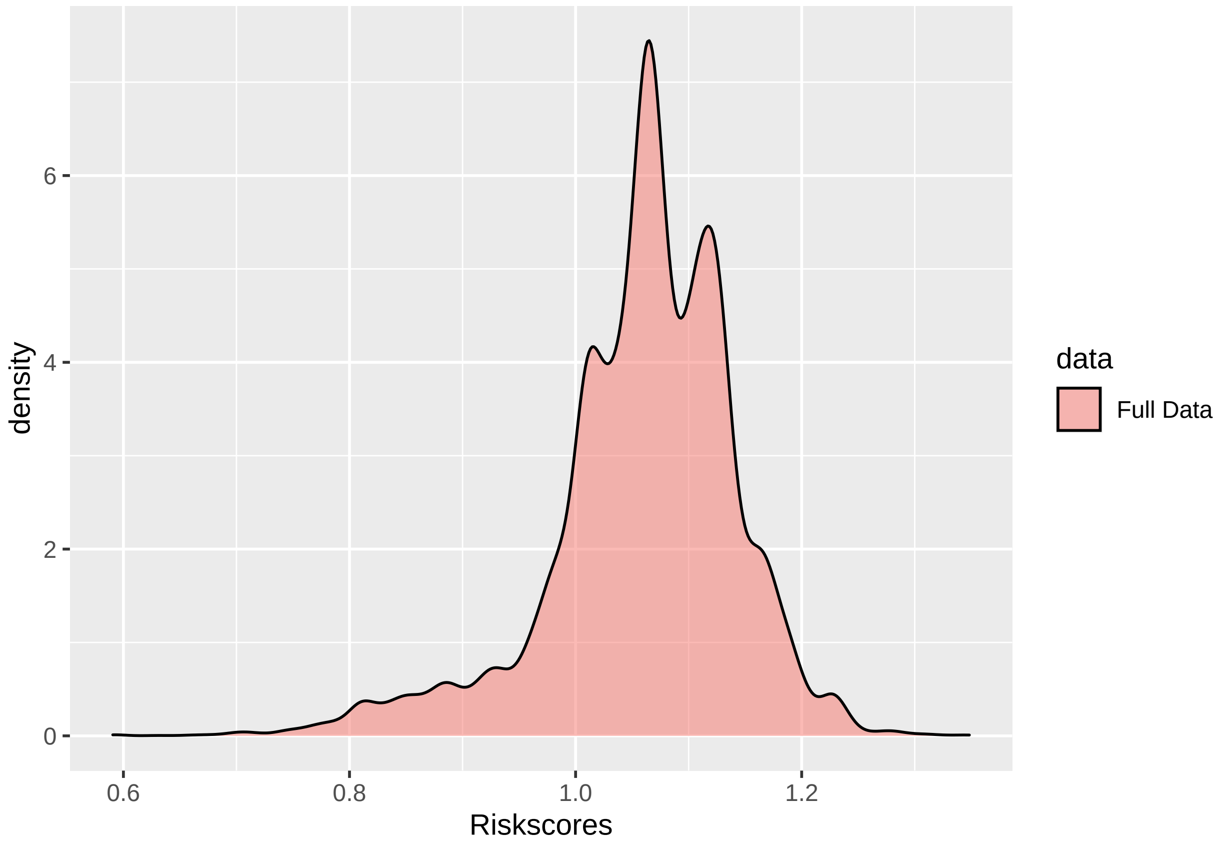
4 Predicting Risks scores and Hazard Ratio from COX PH Model

4.1 Distribution of Riskscores

```
require(ggplot2)
riskscore1_all=predict(cox_fit1,  type="risk") #the risk score exp(lp)
hist(riskscore1_all, xlim = c(0,2), xlab = "Riskscores")
```



```
#Density plot of riskscores
# TRAIN_RISK <- data.frame(rs=riskscore1_train)
# TEST_RISK <- data.frame(rs=riskscore1_test)
ALL_DATA <- data.frame(Riskscores=riskscore1_all)

# TRAIN_RISK$type<-'train'
# TEST_RISK$type<-'test'
ALL_DATA$data<-'Full Data'
ggplot(ALL_DATA, aes(Riskscores, fill=data)) + geom_density(alpha = 0.5)
```



```
# ggplot(TEST_RISK, aes(rs, fill=type)) + geom_density(alpha = 0.2)
# ggplot(TRAIN_RISK, aes(rs, fill=type)) + geom_density(alpha = 0.2)
#
# datlen<-rbind(TRAIN_RISK,TEST_RISK,ALL_DATA)
# ggplot(datlen, aes(rs, fill=type)) + geom_density(alpha = 0.2)
```

5 Hazard Ratios

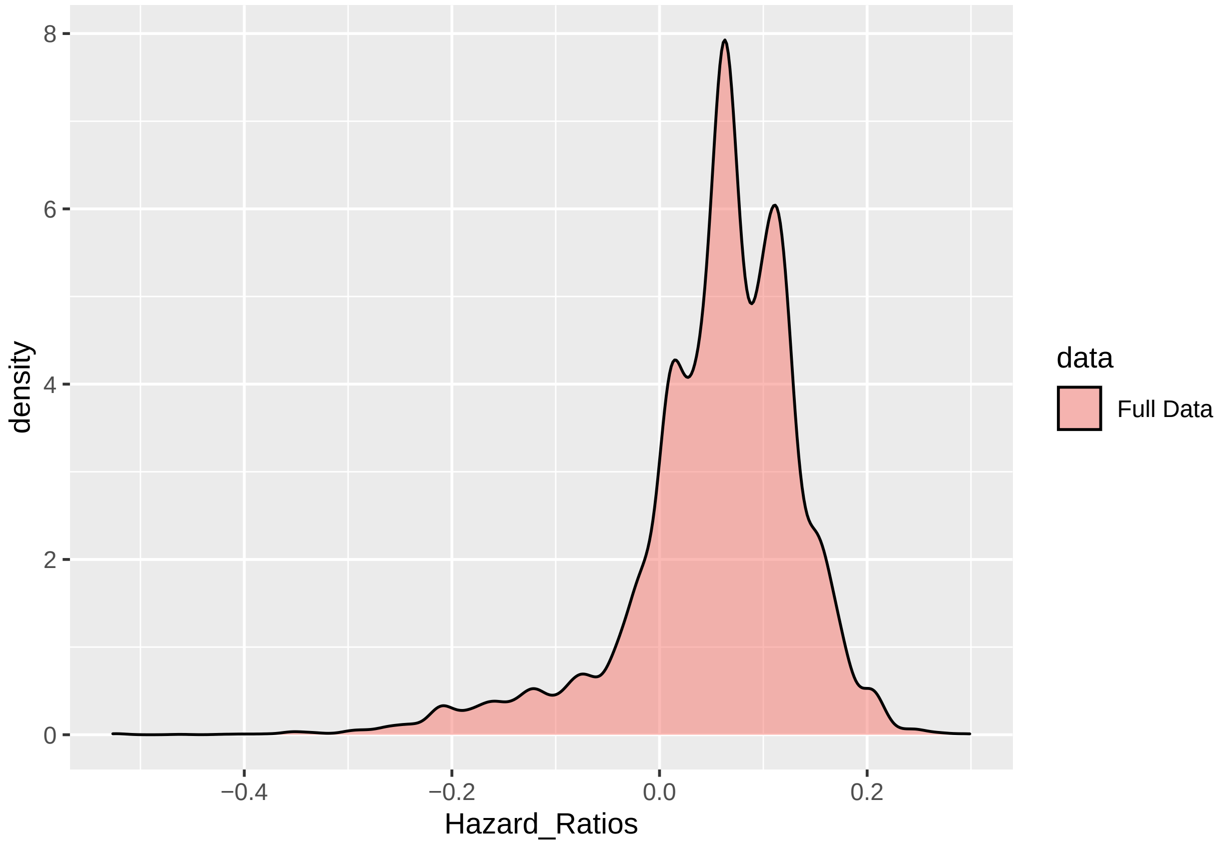
```
lphr3=predict(cox_fit1, type="lp") #predicted hazard ratio

# hist(lphr3, xlim = c(0,2), xlab = "HR")
# hist(1-lphr3, xlim = c(0,2), xlab = "HR")
range(lphr3)
```

```
## [1] -0.526645  0.298877
```

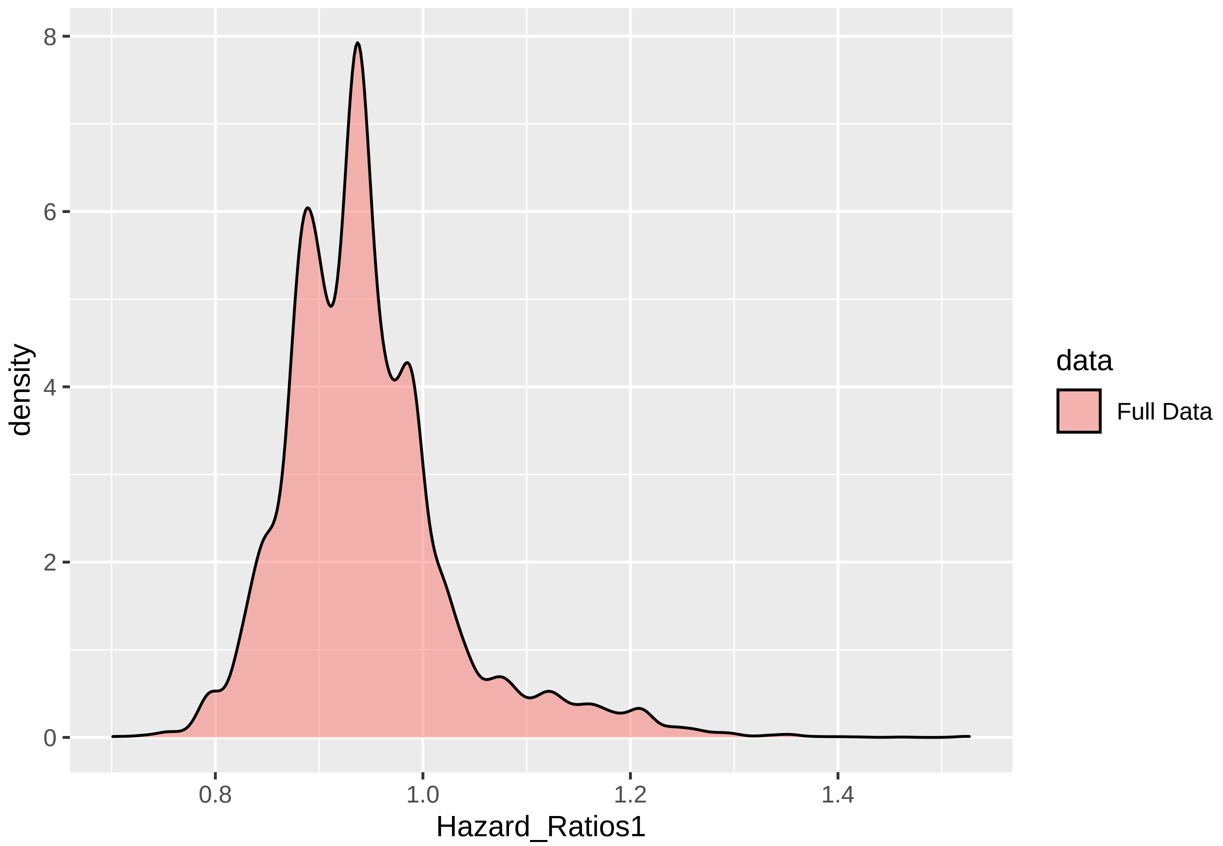
```
range(1-lphr3)
```

```
## [1] 0.701123 1.526645
```

```
ALLDATA <- data.frame(Hazard_Ratios=lphr3)
ALLDATA$data<-'Full Data'
ggplot(ALLDATA, aes(Hazard_Ratios, fill=data)) + geom_density(alpha = 0.5)
```

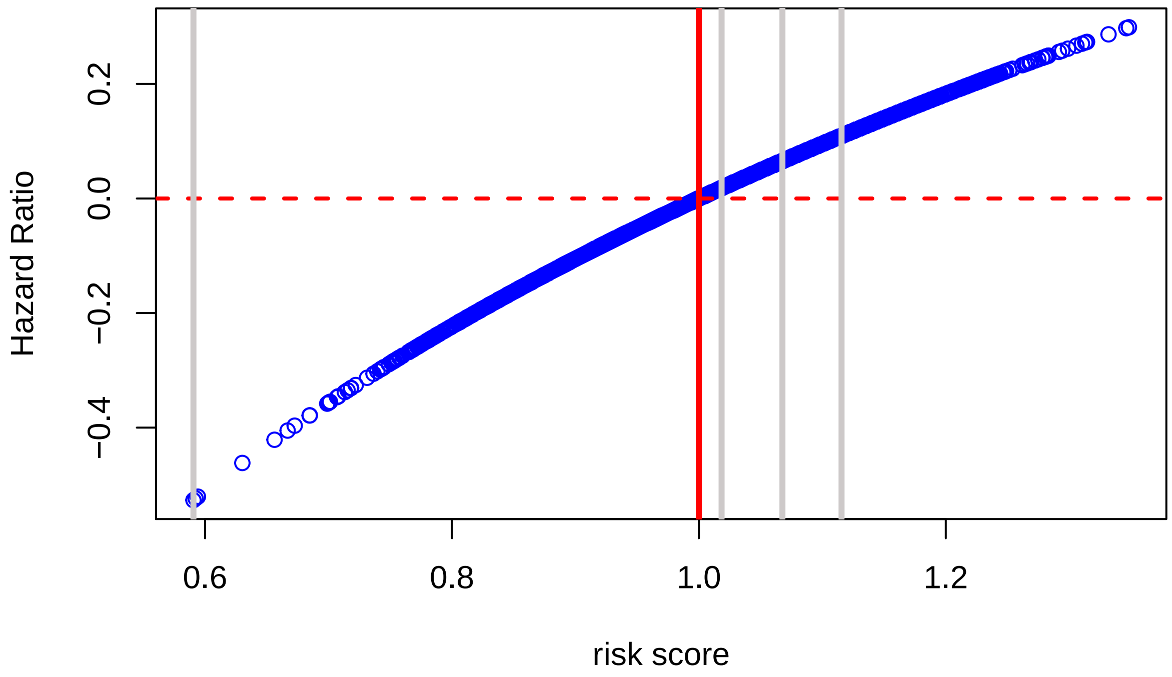


```
ALLDATA <- data.frame(Hazard_Ratios1=1-lphr3)
ALLDATA$data<-'Full Data'
ggplot(ALLDATA, aes(Hazard_Ratios1, fill=data)) + geom_density(alpha = 0.5)
```


8
6
4
2
0
density
0.8
1.0
1.2
1.4
Hazard_Ratios1
data
Full Data

6 Scatterplots of Riskscores vs Hazard Ratios

```
plot(riskscore1_all, lphr3, ylab = "Hazard Ratio" ,
     xlab="risk score", col="blue")
v=quantile(riskscore1_all)
abline(h=0, lty=2, lwd=2, col="red")
abline(v=1, lwd=3, col="red")
abline(v=v[2], lwd=3, col="snow3")
abline(v=v[1], lwd=3, col="snow3")
abline(v=v[3], lwd=3, col="snow3")
abline(v=v[4], lwd=3, col="snow3")
```



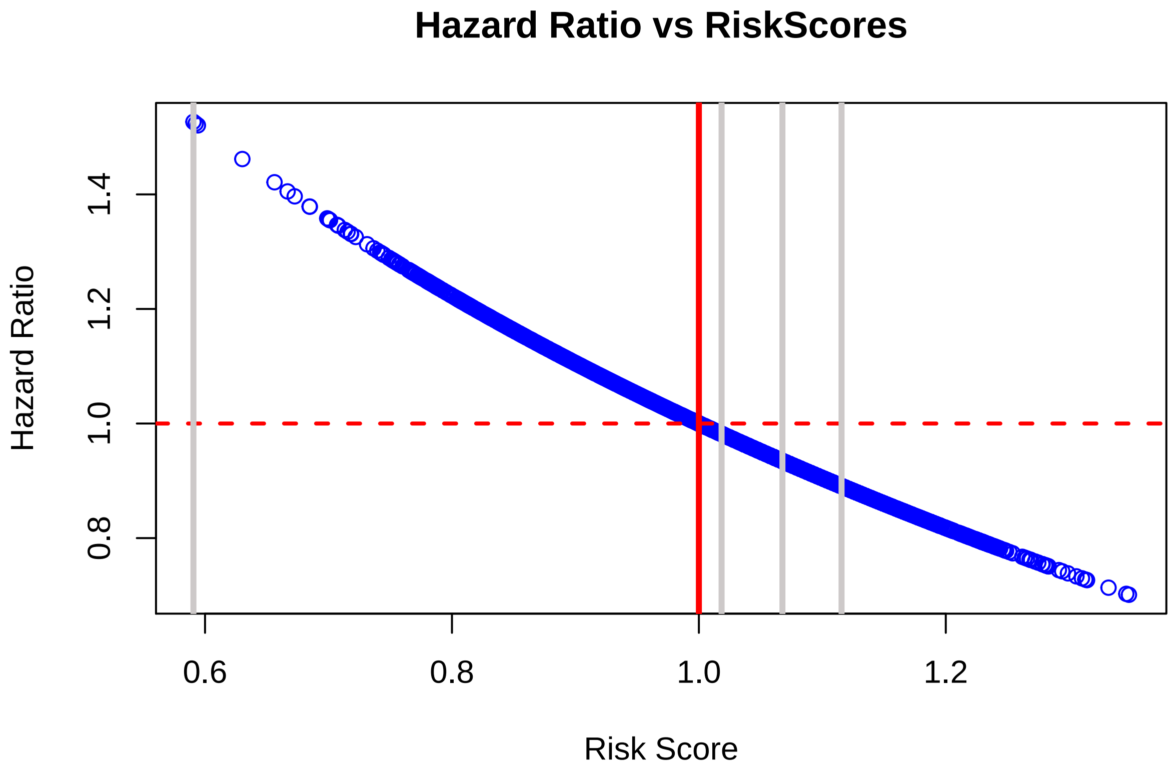
Comment

A positive HR indicates worse conditions/prognosis, while a negative coefficient indicates a better condition/prognosis.

Riskscores > 1 corresponds to increased hazards of mortality with multiple HRF. Riskscores < 1 corresponds to decreased hazards of mortality with multiple HRF, thus a survival benefit from chemotherapy.

The threshold for survival benefit is experienced when the risk score is < 1 at which point the hazards of mortality is decreased.

```
plot(riskscore1_all, 1-lphr3, ylab = "Hazard Ratio", xlab="Risk Score",
     main= "Hazard Ratio vs RiskScores",col="blue")# ylim=c(0.6,1.6), xlim=c(0.6,1.6)

abline(h=1, lty=2, lwd=2, col="red")
abline(v=1, lwd=3, col="red")
abline(v=v[2], lwd=3, col="snow3")
abline(v=v[1], lwd=3, col="snow3")
abline(v=v[3], lwd=3, col="snow3")
abline(v=v[4], lwd=3, col="snow3")
```



Subtracting HR from 1 gives reverse scale of HR reads, where lower HR indicates worse conditions, higher HR indicates better prognosis

As risk score increases, the hazard ratio gets worse (bad prognosis). A survival benefit from adjuvant chemotherapy is realised when the risk score is < 1 at which point the hazard ratio of mortality appears to be better.

7 Riskscores from Est.PH

8 Est.PH {survC1} - Derivation of a risk score by a Cox proportioal hazard model

8.1 Obtain Risk scores from the best predictors of mortality

```
#Provides risk score by fitting data to a Cox's proportional hazards model with a given set of
# Input data. The 1st column should be time-to-event, and the 2nd column is event indicator (1
#OUTPUT
# beta = Estimates for regression coefficient in the Cox model
# var = Variance-Covariance matrix for the beta above
# rs    = Risk score of each individual
# ft    = coxph object with the fitted model

library(survC1)
```

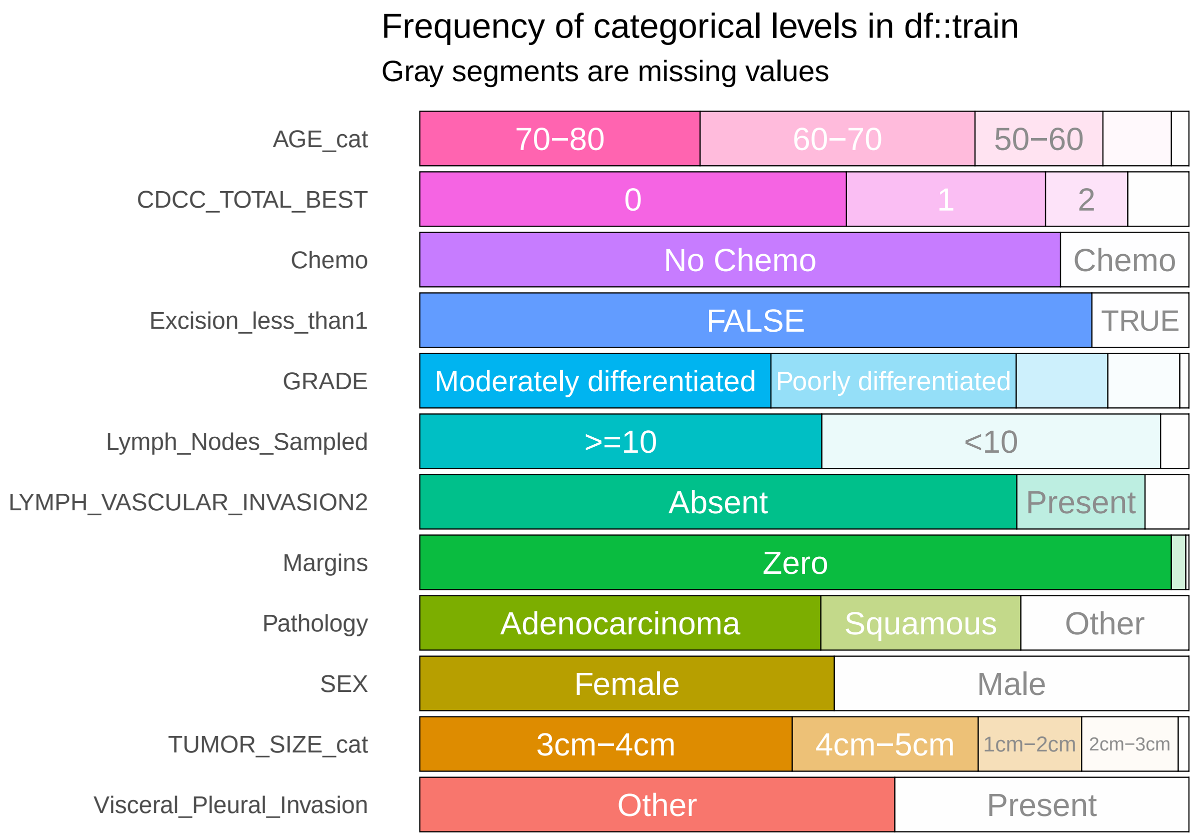
```
## Warning: package 'survC1' was built under R version 4.0.5
```

```
train1=lung[,c(1:2)] #time & status

train2 =lung[, c(3:15)] # other covariates

#convert other sub levels in all categorical covariates to integer
p = data.frame(lapply(train2, as.integer))

#combine numeric time & status with the numeric covariates
train_data = data.frame(cbind(train1,p))

#Make sure distribution of variables are not distorted after conversion
require(inspectdf)
```

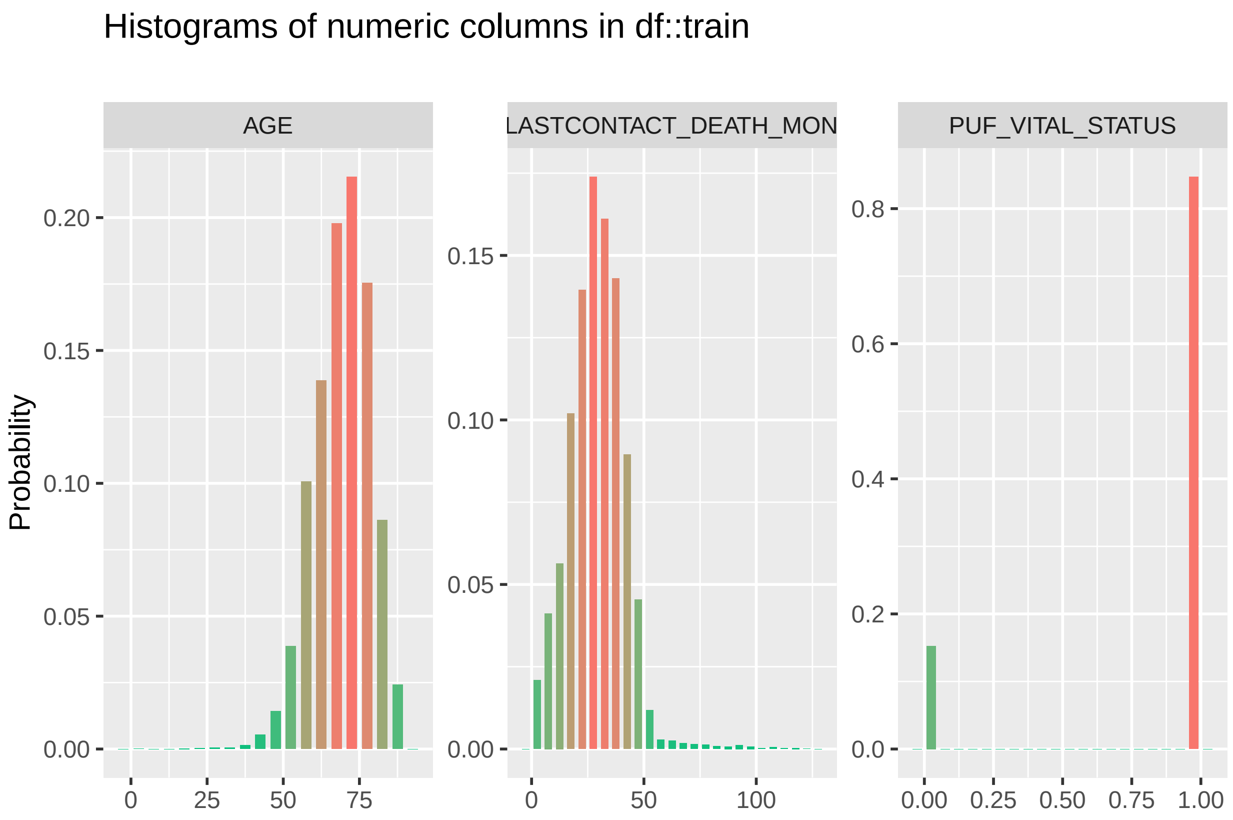
```
## Loading required package: inspectdf
```

```
## Warning: package 'inspectdf' was built under R version 4.0.5
```

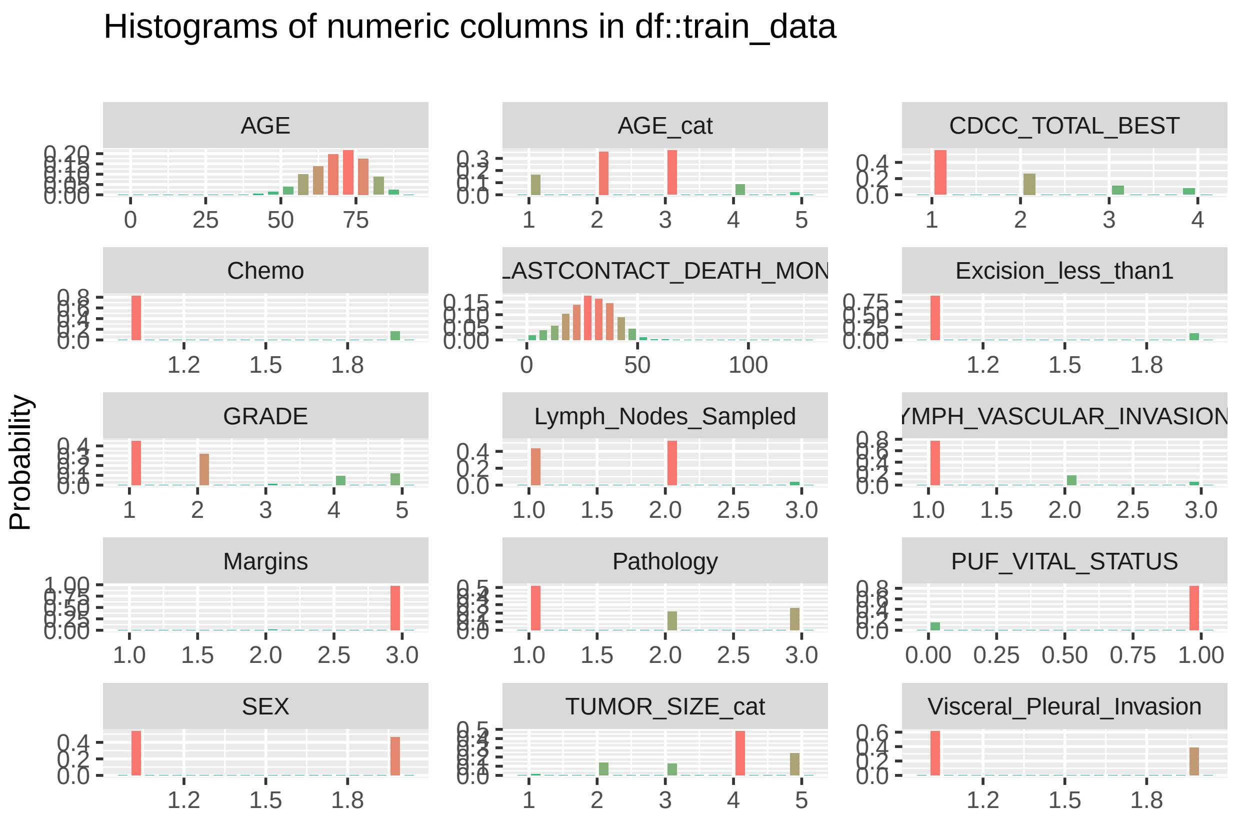
```
show_plot(inspect_cat(train)) # inspect categorical columns
```



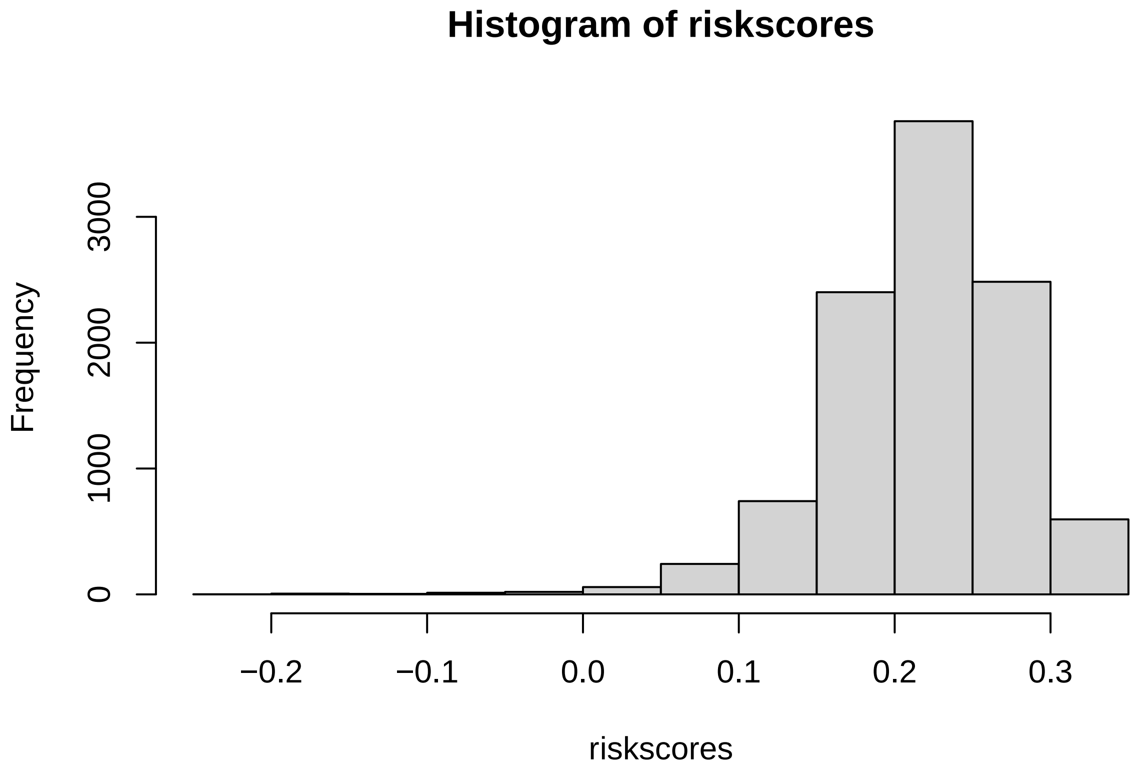
```
show_plot(inspect_num(train)) #inspect numeric columns
```



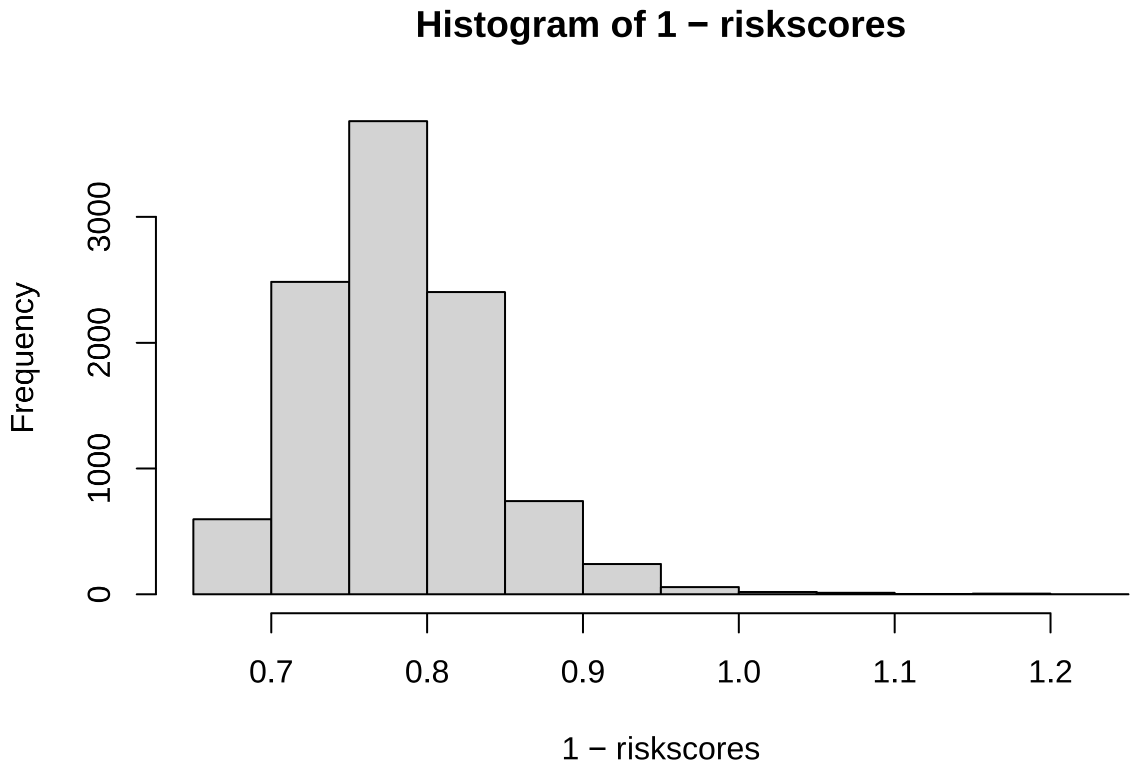
```
show_plot(inspect_num(train_data)) #inspect numeric columns
```



```
#obtain risk scores for each individual
rsmodel=Est.PH(train_data)
riskscores=rsmodel$rs
hist(riskscores)
```



```
hist(1-riskscores)
```



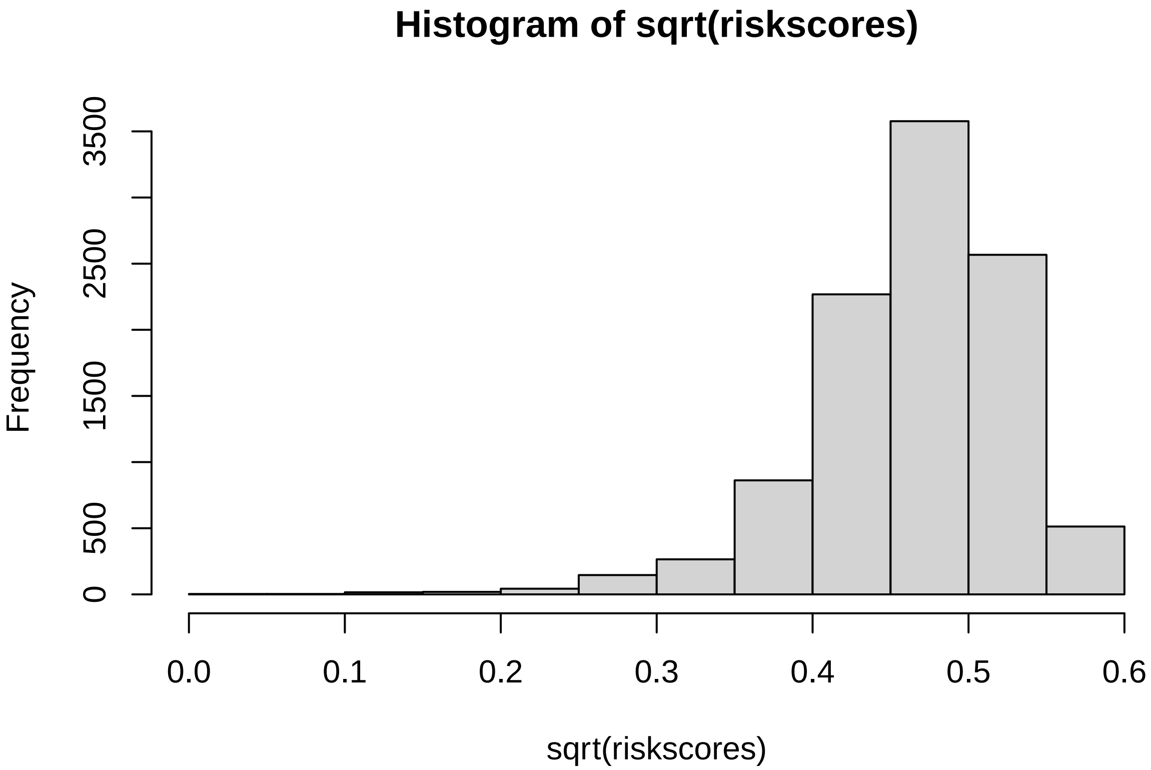
```
#riskscores1 = 1 - riskscores
```

```
#obtain hazard rates .... didnt work with this approach
coef=rsmodel$beta
exp(coef)
```

```
##                     covsChemo                       covsAGE
##                     0.9772929                     1.0010946
##                   covsAGE_cat                       covsSEX
##                     0.9791978                     0.9486315
##            covsCDCC_TOTAL_BEST           covsTUMOR_SIZE_cat
##                     1.0025016                     1.0005138
##                     covsGRADE                covsPathology
##                     0.9915676                     1.0041478
## covsVisceral_Pleural_Invasion  covsLYMPH_VASCULAR_INVASION2
##                     1.0490644                     0.9495022
##                   covsMargins       covsLymph_Nodes_Sampled
##                     1.1256563                     0.9935250
##       covsExcision_less_than1
##                     0.9654898
```

```
hist(sqrt(riskscores))
```

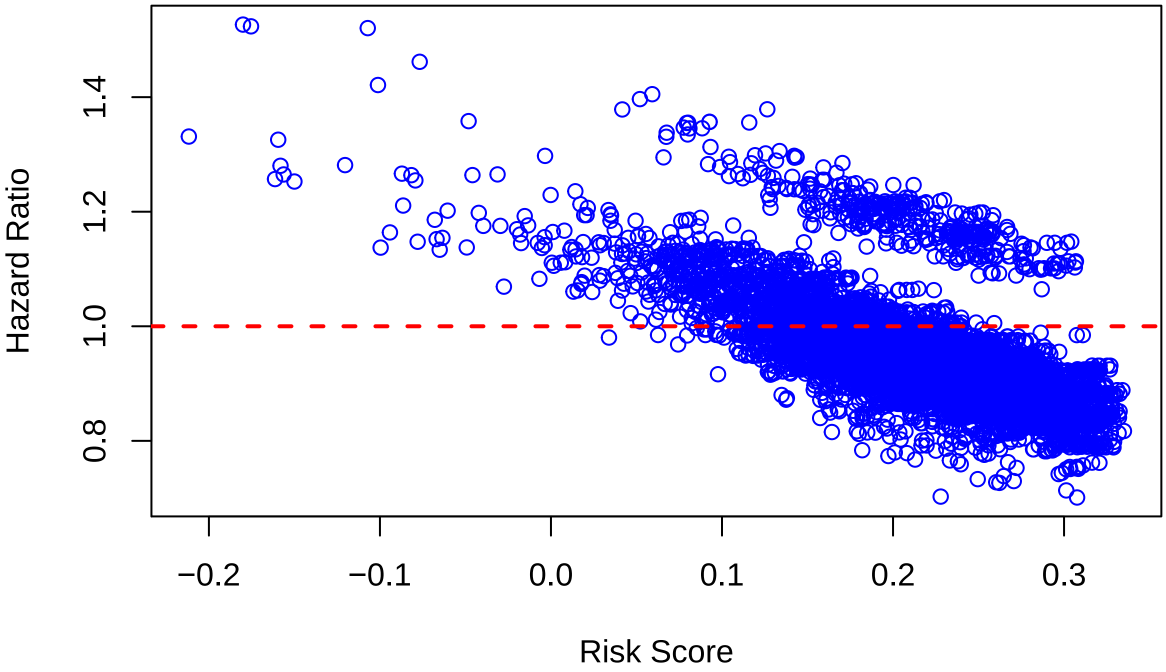
```
## Warning in sqrt(riskscores): NaNs produced
```



```
#riskscores1 = 1 - riskscores
```

```
# lphr3=predict(cox_fit3, lung, type="lp") #predicted hazard ratio
# hrrr=1-lphr3
```

```
plot(riskscores, 1-lphr3, ylab = "Hazard Ratio" ,
     xlab="Risk Score", col="blue")# ylim=c(0.6,1.6), xlim=c(0.6,1.6)

abline(h=1, lty=2, lwd=2, col="red")
abline(v=1, lwd=3, col="red")
abline(v=v[2], lwd=3, col="snow3")
abline(v=v[1], lwd=3, col="snow3")
abline(v=v[3], lwd=3, col="snow3")
abline(v=v[4], lwd=3, col="snow3")
```


Hazard Ratio
0.8
1.0
1.2
1.4
−0.2
−0.1
0.0
0.1
0.2
0.3
Risk Score