HDAT9600 Final Team 2 Assignment

Please see course outline / 'Announcements' for submission deadline

Team 2: Annie, Jason, Yosuke, Tiffany

Aug 19, 2022

Task 1

```
# Subsetting the ICU dataset into variables of interest and removing rows where SOFA is negative
icu_sub <- icu_patients_df1 %>% dplyr::select(in_hospital_death, ICUType, Age, SOFA, FiO2_max, RespRate_min) %>% filter(SOFA
!= -1)

# Getting the summary statistics of the dataset
sumtable(icu_sub)
```

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
in_hospital_death	1996	0.147	0.354	0	0	0	1
ICUType	1996						
… Coronary Care Unit	287	14.4%					
… Cardiac Surgery Recovery Unit	446	22.3%					
… Medical ICU	757	37.9%					
… Surgical ICU	506	25.4%					
Age	1996	64.52	17.397	16	53	78	90
SOFA	1996	6.683	4.042	0	3	9	22
FiO2_max	1996	0.789	0.248	0.28	0.5	1	1
RespRate_min	1996	14.251	3.792	4	12	17	24

We had a look at all the variables in the dataset before making a selection of 6 variables which we will investigate.

We selected the predictor variables (age,ICU type, SOFA score, maximum fractional inspired oxygen (FiO2) and minimum respiration rate) to predict the risk of in-hospital death.

Background Research supporting the selection of predictor variables

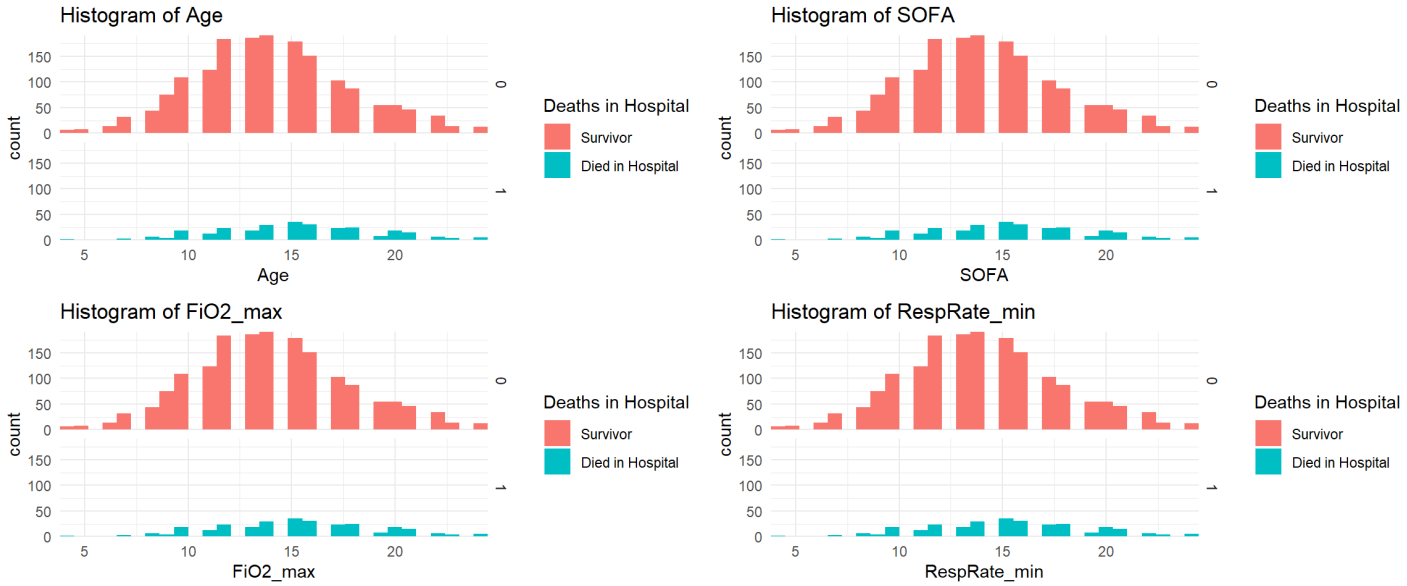
Based on the article "Relationship between age and in-hospital mortality during 15,345,025 non-surgical hospitalizations" from the Archives of Medical Science, the findings support in hospital death to be associated with the age of the patients. In this study, older patients have a greater odds of dying in hospital than younger patients.
In this article, "Infection as an independent risk factor for mortality in the surgical intensive care unit" from National Library of Medicine which evaluated mortality in hospital from surgical and medical ICU, points out that certain types of ICU are associated with high in hospital mortality.
Additionally, "Prognostic Accuracy of the SOFA Score, SIRS Criteria, and SOFA Score for In-Hospital Mortality Among Adults With Suspected Infection Admitted to the Intensive Care Unit" from PubMed propose that high SOFA scores are associated with higher in hospital mortality.
According to "Severity of respiratory failure at admission and in-hospital mortality in patients with COVID-19: a prospective observational multicentre study" from BMJ reports high FiO2 is independently associated with in-hospital mortality.
Lastly, "Mean nocturnal respiratory rate predicts cardiovascular and all-cause mortality in community-dwelling older men and women" from ERS reports that the association between low respiratory rate and in hospital and short term mortality.

An interesting finding with the summary statistics is that there might be issues with the data quality. As seen in the summary statistics of SOFA scores, the minimum value is negative which is not possible as the SOFA scores range from 0-24 and we removed it as they are missing values.

```
# Turn in_hospital_death variable to factor variables
icu_sub$in_hospital_death <- as.integer(as.factor(icu_sub$in_hospital_death)) - 1
```

The subset of ICU dataset has no missing data.

```
# for loop to produce all the plots to compare the distribution
 names <- colnames(icu_sub[-c(1:2)])
plots <- list()
  for (column in names) {
 plots[[column]] <- ggplot(data = icu_sub, aes(x = icu_sub[,column], fill = factor(in_hospital_death)))+ geom_histogram(bins
= 30) + theme_minimal() + scale_x_continuous(expand = c(0,0)) + scale_y_continuous(expand = c(0,0)) + ggtitle(paste("Histogr
am of", column)) + labs(x = column, fill = "Deaths in Hospital") + facet_grid(in_hospital_death~., scales = "fixed")+ scale_
fill_discrete(name = "Deaths in Hospital", labels = c("Survivor", "Died in Hospital"))
}
do.call(grid.arrange, plots)
```



Looking at the histograms, it is evident that majority of the patients did not die in the hospital and the distribution for all predictor variables except type of ICU for survivor and died in hospital are very similar. An interesting observation is that for the variable minimum respiration rate, the mean respiration rate is higher in those that died in hospital than those who survived. Another observation which aligns with the research is the predictor variable age, the mean age of patients is higher in the cohort that died in the hospital than those who survived.

```
# Fitting univariate logistic regression models
age_glm <- glm(in_hospital_death ~ Age, family = binomial, data = icu_sub)
Icu_type_glm <- glm(in_hospital_death ~ ICUType, family = binomial, data = icu_sub)
SOFA_glm <- glm(in_hospital_death ~ SOFA, family = binomial, data = icu_sub)
FiO2_glm <- glm(in_hospital_death ~ FiO2_max, family = binomial, data = icu_sub)
RespRate_glm <- glm(in_hospital_death ~ RespRate_min, family = binomial, data = icu_sub)
AIC(age_glm, Icu_type_glm, SOFA_glm, FiO2_glm, RespRate_glm)
```

	df <dbl>	AIC <dbl>
age_glm	2	1612.995
Icu_type_glm	4	1626.341
SOFA_glm	2	1581.240
FiO2_glm	2	1668.890
RespRate_glm	2	1643.241

5 rows

```
# Using AIC for model selection
full_mod <- glm(in_hospital_death~., family = binomial, data = icu_sub)
aic_suggest_mod <- stepAIC(full_mod, direction="both", scope=list(lower=~1, upper=~.^2), trace=F, data=icu_sub)
tab_model(aic_suggest_mod, show.intercept = TRUE, show.est = TRUE, show.ci = FALSE, show.se = TRUE, show.p = TRUE, show.stat
= TRUE, show.aic = TRUE, show.dev = TRUE, show.r2 = FALSE, show.fstat = TRUE, show.obs = FALSE)
```

	in hospital death			
Predictors	*Odds Ratios*	*std. Error*	*Statistic*	*p*
(Intercept)	0.00	0.00	-5.01	**<0.001**
ICUType [Cardiac Surgery Recovery Unit]	1.66	2.92	0.29	0.774
ICUType [Medical ICU]	1.20	1.52	0.15	0.884

ICUType [Surgical ICU]	0.13	0.18	-1.45	0.147
Age	1.04	0.02	2.45	**0.014**
SOFA	1.56	0.15	4.73	**<0.001**
RespRate min	1.03	0.02	1.72	0.085
ICUType [Cardiac Surgery Recovery Unit] * SOFA	0.78	0.06	-3.22	**0.001**
ICUType [Medical ICU] * SOFA	0.91	0.05	-1.82	0.068
ICUType [Surgical ICU] * SOFA	0.92	0.05	-1.55	0.122
Age * SOFA	1.00	0.00	-2.46	**0.014**
ICUType [Cardiac Surgery Recovery Unit] * Age	1.00	0.02	0.07	0.946
ICUType [Medical ICU] * Age	1.01	0.02	0.58	0.559
ICUType [Surgical ICU] * Age	1.04	0.02	2.12	**0.034**
Deviance	1409.259			
AIC	1437.259			

The AIC suggested models only contain four predictor variables which are type of ICU, age, SOFA score and minimum respiration rate with 3 interactions which are ICUTypeCardiac Surgery Recovery Unit:SOFA, Age:SOFA and ICUTypeSurgical ICU:Age.

```
# Model suggested by AIC
reduced_interactions1 <- glm(in_hospital_death ~ ICUType + Age + SOFA + RespRate_min +
    ICUType:SOFA + Age:SOFA + ICUType:Age, family = binomial, data = icu_sub)

# Tried to see if it dropping the minimum respiration rate would make any difference as the p-value is above 0.05
reduced_interactions2 <- glm(in_hospital_death ~ ICUType + Age + SOFA +
    ICUType:SOFA + Age:SOFA + ICUType:Age, family = binomial, data = icu_sub)

# display the results of both models
tab_model(reduced_interactions1, reduced_interactions2, show.intercept = TRUE, show.est = TRUE, show.ci = FALSE, show.se = T
RUE, show.p = TRUE, show.stat = TRUE, show.aic = TRUE, show.dev = TRUE, show.r2 = FALSE, show.fstat = TRUE, show.obs = FALSE
, dv.labels = c("reduced_interactions1", "reduced_interactions2"))
```

	reduced_interactions1				**reduced_interactions2**			
Predictors	*Odds Ratios*	*std. Error*	*Statistic*	*p*	*Odds Ratios*	*std. Error*	*Statistic*	*p*
(Intercept)	0.00	0.00	-5.01	**<0.001**	0.00	0.00	-4.76	**<0.001**
ICUType [Cardiac Surgery Recovery Unit]	1.66	2.92	0.29	0.774	1.38	2.44	0.18	0.854
ICUType [Medical ICU]	1.20	1.52	0.15	0.884	1.19	1.51	0.14	0.888
ICUType [Surgical ICU]	0.13	0.18	-1.45	0.147	0.12	0.17	-1.47	0.141
Age	1.04	0.02	2.45	**0.014**	1.04	0.02	2.42	**0.016**
SOFA	1.56	0.15	4.73	**<0.001**	1.58	0.15	4.87	**<0.001**
RespRate min	1.03	0.02	1.72	0.085				
ICUType [Cardiac Surgery Recovery Unit] * SOFA	0.78	0.06	-3.22	**0.001**	0.78	0.06	-3.23	**0.001**
ICUType [Medical ICU] * SOFA	0.91	0.05	-1.82	0.068	0.91	0.05	-1.92	0.054
ICUType [Surgical ICU] * SOFA	0.92	0.05	-1.55	0.122	0.91	0.05	-1.66	0.098
Age * SOFA	1.00	0.00	-2.46	**0.014**	1.00	0.00	-2.49	**0.013**
ICUType [Cardiac Surgery Recovery Unit] * Age	1.00	0.02	0.07	0.946	1.00	0.02	0.15	0.877

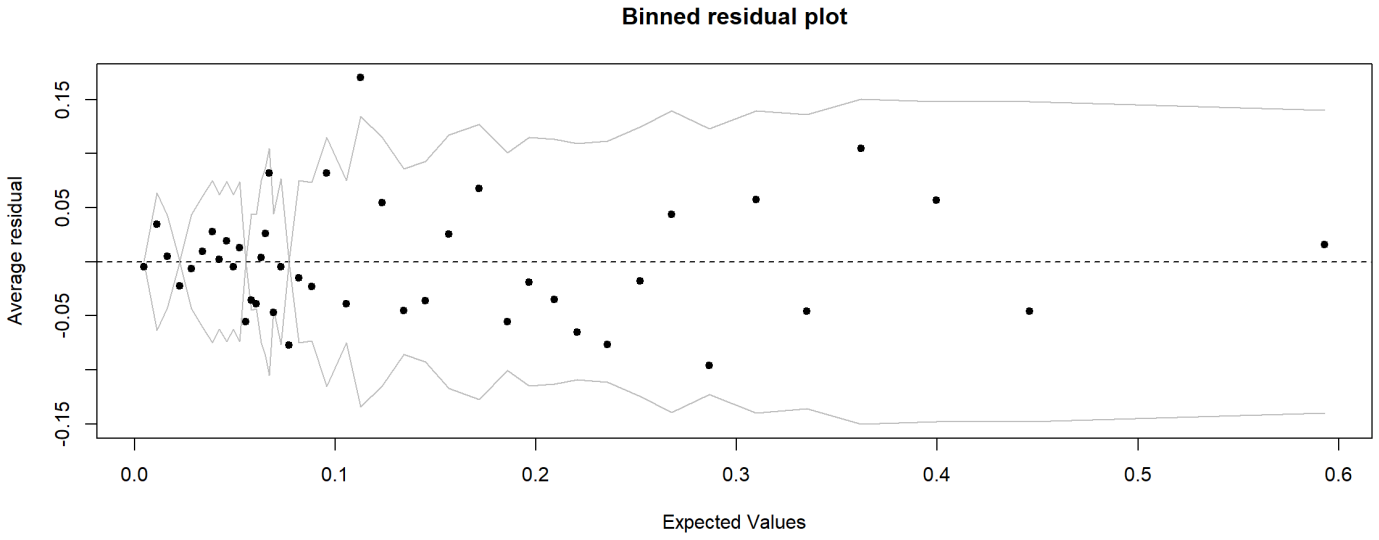
ICUType [Medical ICU] * Age	1.01	0.02	0.58	0.559	1.01	0.02	0.62	0.535
ICUType [Surgical ICU] * Age	1.04	0.02	2.12	**0.034**	1.04	0.02	2.15	**0.031**
Deviance	1409.259				1412.223			
AIC	1437.259				1438.223			

```
# Running an analysis of deviance
print(anova(reduced_interactions2, reduced_interactions1, test = "Chi"))
```

```
## Analysis of Deviance Table
##
## Model 1: in_hospital_death ~ ICUType + Age + SOFA + ICUType:SOFA + Age:SOFA +
##     ICUType:Age
## Model 2: in_hospital_death ~ ICUType + Age + SOFA + RespRate_min + ICUType:SOFA +
##     Age:SOFA + ICUType:Age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1983     1412.2
## 2      1982     1409.3  1   2.9635  0.08516 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the analysis of deviance, the p-value is above 0.05, showing that minimum respiration rate is not a significant predictor of in-hospital deaths. Even though the AIC for the second model is 1438.2 which is 0.9 higher then the model fitted by earlier, the simpler model is preferred based on the analysis of deviance.

```
# Checking the model fit using binned residual plot
binnedplot(predict(reduced_interactions2, type="response"), residuals(reduced_interactions2, type="response"))
```



The variance of the residuals in the binned residual plot seems to be constant and evenly distributed.

```
# Checking the goodness of fit of reduced model using Brier Score test
pred_prob_reduced <- predict(reduced_interactions1, type = "response")
brier_score_reduced <- mean((pred_prob_reduced - icu_sub$in_hospital_death)^2)
brier_score_reduced
```

```
## [1] 0.1072906
```

```
# Checking the goodness of fit of reduced model with interaction using Brier Score test
pred_prob_reduced_interactions2 <- predict(reduced_interactions2, type = "response")
brier_score_reduced <- mean((pred_prob_reduced_interactions2 - icu_sub$in_hospital_death)^2)
brier_score_reduced
```

```
## [1] 0.1074341
```

The difference in Brier scores between the two is very small, therefore the second model without minimum respiration rate is preferred based on the analysis of deviance conducted earlier.

Unsurprisingly, according to our model `reduced_interactions2` , an increase in `Age` and `SOFA` (Sequential Organ Failure Assessment) score increased mortality. For every year increase of `Age` , there is ~4% increase in mortality. For every `SOFA` point increase, there is a ~58% increase in mortality.

```
unimputed <- glm(in_hospital_death ~ ICUType + Age + SOFA +
    ICUType:SOFA + Age:SOFA + ICUType:Age, family = binomial, data = icu_patients_df0)
tab_model(unimputed, show.intercept = TRUE, show.est = TRUE, show.ci = FALSE, show.se = TRUE, show.p = TRUE, show.stat = TRU
E, show.aic = TRUE, show.dev = TRUE, show.r2 = FALSE, show.fstat = TRUE, show.obs = FALSE)
```

	in hospital death			
Predictors	*Odds Ratios*	*std. Error*	*Statistic*	*p*
(Intercept)	0.00	0.00	-4.52	**<0.001**
ICUType [Cardiac Surgery Recovery Unit]	1.19	2.04	0.10	0.921
ICUType [Medical ICU]	1.02	1.25	0.02	0.985
ICUType [Surgical ICU]	0.20	0.27	-1.18	0.238
Age	1.03	0.02	2.12	**0.034**
SOFA	1.47	0.13	4.31	**<0.001**
ICUType [Cardiac Surgery Recovery Unit] * SOFA	0.79	0.06	-3.06	**0.002**
ICUType [Medical ICU] * SOFA	0.93	0.05	-1.55	0.120
ICUType [Surgical ICU] * SOFA	0.91	0.05	-1.79	0.074
Age * SOFA	1.00	0.00	-1.84	0.065
ICUType [Cardiac Surgery Recovery Unit] * Age	1.00	0.02	0.15	0.883
ICUType [Medical ICU] * Age	1.01	0.02	0.60	0.547
ICUType [Surgical ICU] * Age	1.03	0.02	1.85	0.064
Deviance	1450.893			
AIC	1476.893			

```
# calculate p-value
(p_unimputed <- 1 - pchisq(unimputed$null.deviance - unimputed$deviance, unimputed$df.null - unimputed$df.residual))
```

```
## [1] 0
```

```
(p_imputed <- 1 - pchisq(reduced_interactions2$null.deviance - reduced_interactions2$deviance, reduced_interactions2$df.null
- reduced_interactions2$df.residual))
```

```
## [1] 0
```

```
# null model
reduced_null <-glm(in_hospital_death ~ 1, family = binomial, data = icu_sub)

# calculate p-value
print(anova(reduced_null, reduced_interactions2, test = "Chi"))
```

```
## Analysis of Deviance Table
##
## Model 1: in_hospital_death ~ 1
## Model 2: in_hospital_death ~ ICUType + Age + SOFA + ICUType:SOFA + Age:SOFA +
##     ICUType:Age
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      1995     1665.1
## 2      1983     1412.2 12   252.87 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking through the summary statistics and p-values of the imputed and the unimputed datasets, the ICUTypeSurgical ICU:Age interaction and Age:SOFA interaction became insignificant and in the unimputed dataset. The AIC for the unimputed dataset is 1476.9 which is much higher than the AIC for the imputed dataset of 1438.2. This means that the model is a better fit for the imputed dataset.

Summary Findings

After comparing different models, the final model we have fitted contains the predictor variables (type of ICU, age and SOFA scores) and with 3 interactions between type of ICU and age, type of ICU and SOFA scores, and age and SOFA scores. The p-value is 0 which suggests that the model is better than the null model.
Through the analysis of deviance, it is apparent that the model without the minimum respiration rate variable is the preferred model over the one with the variable as shown by the p-value above 0.05. Although the Brier Scores show the model with minimum respiration rate has a slightly better fit, but the differences is negligible if rounded to 3 decimal places where both values would be 1.107. The binned residual plot also did not raise any concerns as the variance of the residuals in the binned residual plot seems to be constant and evenly distributed. In conclusion, the findings of this study suggests that the in-hospital mortality rate is associated with the type of ICU, age of the patient and SOFA scores out of the initial subset of the predictor variables that were selected.

=====

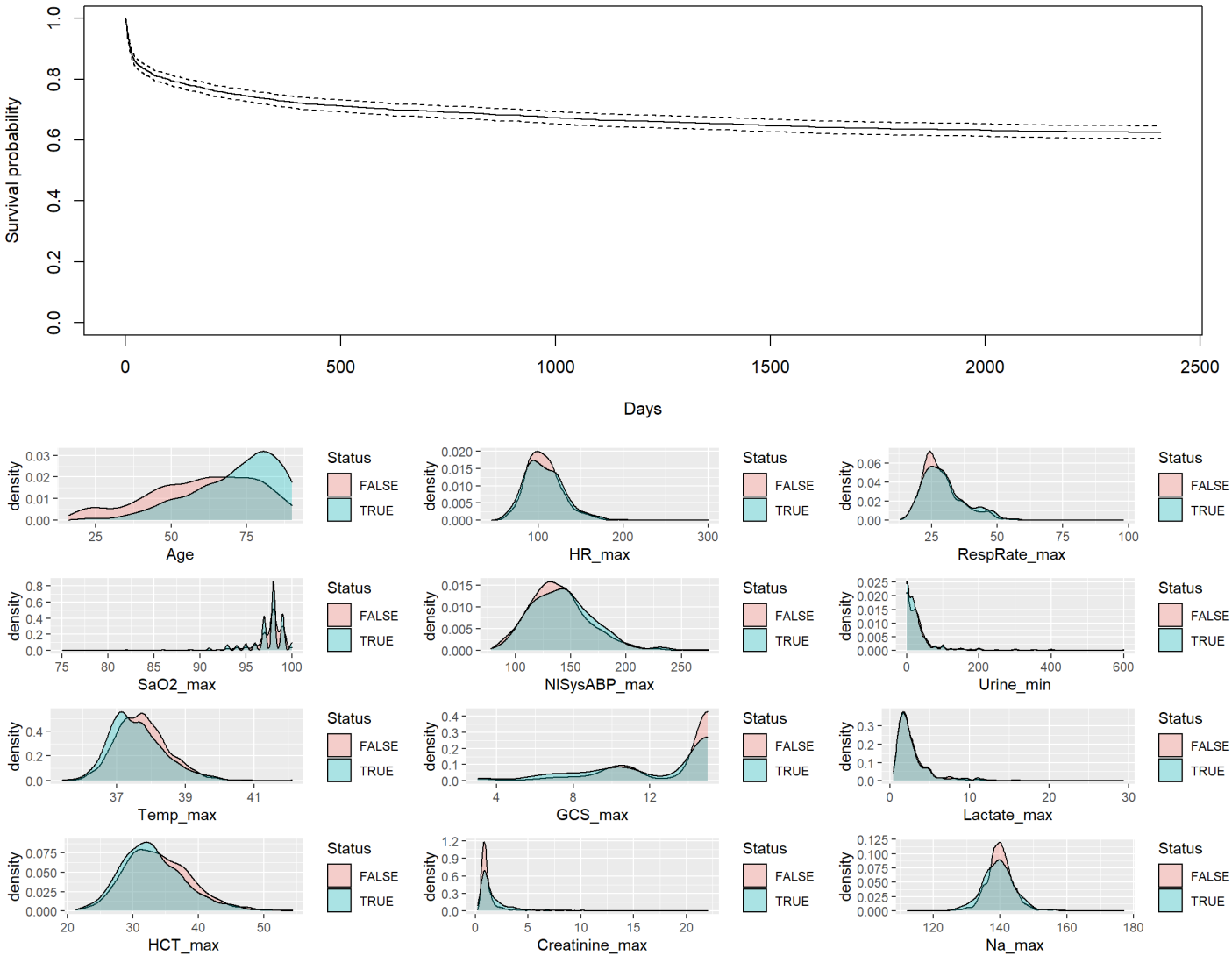
The explanatory variables I will use to predict survival will include:

- the `Age` demographic
- important vital signs: `HR_max` , `RespRate_max` , `SaO2_max` , `NISysABP_max` , `Temp_max` , `GCS_max` , `Urine_max`
- important biochemical markers chosen based on clinical experience: `Lactate_max` , `HCT_max` , `Creatinine_max` , `Na_max`

I will create a subset of the `icu_patients_df1` with the above predictor variables, and the outcome variables `Days` (survival) and `Status` (censoring).

```
icu_sub2 <- icu_patients_df1 %>% dplyr::select(Days, Status, Age, HR_max, RespRate_max, SaO2_max, NISysABP_max, Urine_min, T
emp_max, GCS_max, Lactate_max, HCT_max, Creatinine_max, Na_max)
icu_sub2c <- na.omit(icu_sub2)
attach(icu_sub2)
```

A brief exploratory data analysis is performed on the variables of interest. We show a survival curve for the overall data and density maps for each variable stratified by status.



We will now fit some univariate models for these predictors.

```
surv_object <- Surv(Days, Status)
coxmod_age <- coxph(surv_object ~ Age, data=icu_sub2)
coxmod_HR <- coxph(surv_object ~ HR_max, data=icu_sub2)
coxmod_RR <- coxph(surv_object ~ RespRate_max, data=icu_sub2)
coxmod_O2 <- coxph(surv_object ~ SaO2_max, data=icu_sub2)
coxmod_SBP <- coxph(surv_object ~ NISysABP_max, data=icu_sub2)
coxmod_U <- coxph(surv_object ~ Urine_min, data=icu_sub2)
coxmod_temp <- coxph(surv_object ~ Temp_max, data=icu_sub2)
coxmod_GCS <- coxph(surv_object ~ GCS_max, data=icu_sub2)
coxmod_lactate <- coxph(surv_object ~ Lactate_max, data=icu_sub2)
coxmod_HCT <- coxph(surv_object ~ HCT_max, data=icu_sub2)
coxmod_Cr <- coxph(surv_object ~ Creatinine_max, data=icu_sub2)
coxmod_Na <- coxph(surv_object ~ Na_max, data=icu_sub2)
# display the results
data.frame(rbind(summary(coxmod_age)$coefficients, summary(coxmod_HR)$coefficients,
summary(coxmod_RR)$coefficients, summary(coxmod_O2)$coefficients,
summary(coxmod_SBP)$coefficients, summary(coxmod_U)$coefficients,
summary(coxmod_temp)$coefficients, summary(coxmod_GCS)$coefficients,
summary(coxmod_lactate)$coefficients, summary(coxmod_HCT)$coefficients,
summary(coxmod_Cr)$coefficients, summary(coxmod_Na)$coefficients))
```

	coef <dbl>	exp.coef. <dbl>	se.coef. <dbl>	z <dbl>	Pr...z.. <dbl>
Age	0.033549811	1.0341190	0.002500182	13.418949	4.683064e-41
HR_max	0.002779443	1.0027833	0.001647579	1.686987	9.160587e-02
RespRate_max	0.014719540	1.0148284	0.004320335	3.407037	6.567227e-04
SaO2_max	-0.020685206	0.9795273	0.016313967	-1.267945	2.048178e-01
NISysABP_max	0.003502598	1.0035087	0.001402309	2.497735	1.249895e-02
Urine_min	-0.001925183	0.9980767	0.000717925	-2.681594	7.327233e-03
Temp_max	-0.196578503	0.8215368	0.049298379	-3.987525	6.676626e-05
GCS_max	-0.082041364	0.9212339	0.010612384	-7.730720	1.069401e-14
Lactate_max	0.057782574	1.0594846	0.016664719	3.467360	5.255980e-04
HCT_max	-0.025206607	0.9751084	0.007590586	-3.320772	8.976878e-04
Creatinine_max	0.101515397	1.1068470	0.014669061	6.920375	4.504509e-12
Na_max	-0.015698330	0.9844242	0.008287182	-1.894290	5.818646e-02

12 rows

From the univariate models, it appears `Age`, `RespRate_max`, `NISysABP_max`, `Temp_max`, `GCS_max`, `Urine_min`, `Lactate_max`, `Creatinine_max` are statistically significant predictors for survival.

We now fit some multivariate models.

1. Full subset of predictors `cox_mv_all`
2. Only those significant in the univariate models `cox_mv_uni`

```
surv_object_c <- Surv(icu_sub2c$Days, icu_sub2c$Status)
# use all variables
coxmod_mv_all <- coxph(surv_object_c ~ Age + HR_max + RespRate_max + SaO2_max + NISysABP_max + Urine_min + Temp_max + GCS_ma
x + Lactate_max + HCT_max + Creatinine_max + Na_max, data=icu_sub2c)
show_simple_sum(coxmod_mv_all)
```

```
##                          coef exp.coef.     se.coef.          z     Pr...z..
## Age              0.0328686234 1.0334148 0.0027486505 11.9580948 5.889773e-33
## HR_max           0.0040203059 1.0040284 0.0017030708  2.3606217 1.824433e-02
## RespRate_max     0.0036678101 1.0036745 0.0054167281  0.6771265 4.983257e-01
## SaO2_max        -0.0372253832 0.9634590 0.0173712712 -2.1429280 3.211888e-02
## NISysABP_max     0.0025187355 1.0025219 0.0014070571  1.7900734 7.344211e-02
## Urine_min       -0.0005791792 0.9994210 0.0007725334 -0.7497141 4.534269e-01
## Temp_max        -0.1979935467 0.8203751 0.0544572473 -3.6357612 2.771610e-04
## GCS_max         -0.1043350072 0.9009234 0.0138724549 -7.5210197 5.435064e-14
## Lactate_max      0.0503769324 1.0516674 0.0189938584  2.6522748 7.995144e-03
## HCT_max         -0.0271662860 0.9731994 0.0087220284 -3.1146752 1.841475e-03
## Creatinine_max   0.0839992315 1.0876281 0.0171472854  4.8986898 9.647784e-07
## Na_max          -0.0244340391 0.9758621 0.0084493770 -2.8918155 3.830229e-03
##                             test df       pvalue
## Wald test              278.4300 12 1.562337e-52
## Likelihood ratio test 300.9833 12 2.919117e-57
## Score (logrank) test  291.4059 12 2.987010e-55
```

```
# use only siginificant variables, according to the univariate models
coxmod_mv_uni <- coxph(surv_object_c ~ Age + RespRate_max + NISysABP_max + Urine_min + Temp_max + GCS_max + Lactate_max + HC
T_max + Creatinine_max, data=icu_sub2c)
show_simple_sum(coxmod_mv_uni)
```

```
##                          coef exp.coef.     se.coef.          z     Pr...z..
## Age              0.0321832393 1.0327067 0.0027457963 11.7209131 9.959275e-32
## RespRate_max     0.0042227577 1.0042317 0.0052014995  0.8118347 4.168865e-01
## NISysABP_max     0.0025517176 1.0025550 0.0014083617  1.8118340 7.001185e-02
## Urine_min       -0.0007802114 0.9992201 0.0007712887 -1.0115685 3.117444e-01
## Temp_max        -0.1720547601 0.8419331 0.0541889117 -3.1750916 1.497892e-03
## GCS_max         -0.0954797270 0.9089368 0.0137029143 -6.9678409 3.218418e-12
## Lactate_max      0.0567948295 1.0584386 0.0187449070  3.0298806 2.446505e-03
## HCT_max         -0.0263305627 0.9740131 0.0085009738 -3.0973584 1.952536e-03
## Creatinine_max   0.0852523093 1.0889918 0.0173088368  4.9253633 8.420377e-07
##                             test df       pvalue
## Wald test              260.5100  9 6.005030e-51
## Likelihood ratio test 283.8232  9 7.010649e-56
## Score (logrank) test  272.6358  9 1.638299e-53
```

```
# Accoding to the result of coxmod_mv_all, just use only significant vars
coxmod_mv_all_redc <- coxph(surv_object_c ~ Age + HR_max + SaO2_max + Temp_max + GCS_max + Lactate_max + HCT_max + Creatinin
e_max + Na_max, data=icu_sub2c)
show_test(coxmod_mv_all_redc) # show only test results because of page restriction
```

```
##                             test df       pvalue
## Wald test              275.5600  9 3.935870e-54
## Likelihood ratio test 296.8813  9 1.197089e-58
## Score (logrank) test  287.8221  9 9.965961e-57
```

```
# Accoding to the result of coxmod_mv_uni, just use only significant vars
coxmod_mv_uni_redc <- coxph(surv_object_c ~ Age + Temp_max + GCS_max + Lactate_max + HCT_max + Creatinine_max, data=icu_sub2
c)
show_test(coxmod_mv_uni_redc) # show only test results because of page restriction
```

```
##                             test df       pvalue
## Wald test              256.6400  6 1.565689e-52
## Likelihood ratio test 278.9843  6 2.591894e-57
## Score (logrank) test  267.9849  6 5.853762e-55
```

```
# display AICs of models
AIC(coxmod_mv_all, coxmod_mv_uni, coxmod_mv_all_redc, coxmod_mv_uni_redc)
```

	df <dbl>	AIC <dbl>
coxmod_mv_all	12	9027.128
coxmod_mv_uni	9	9038.289
coxmod_mv_all_redc	9	9025.230
coxmod_mv_uni_redc	6	9037.127

4 rows

```
# display results of anova, just show p values to reduce page
print(anova(coxmod_mv_all_redc, coxmod_mv_all)[[4]][2])
```

```
## [1] 0.250654
```

We compared Akaike's Information Criterion (AIC) of four models: `coxmod_mv_all` : a model with all variables , `coxmod_mv_uni` : a model with only significant variables in univariate models , `coxmod_mv_all_redc` : a model with non-significant variables removed from `coxmod_mv_all` , and `coxmod_mv_uni_redc` : a model with non-significant variables removed from `coxmod_mv_uni` . We found that the `coxmod_mv_all_redc` was the best model and chose it as our final model. Also, according to the p-value of anova, it is acceptable to choose a model with fewer variables. It was shown that the evaluations of the model by the AIC and anova function was improved by removing the variables `RespRate_max` , `NISysABP_max` , `Urine_min` .

```
# confirm PH assumption
prop <- cox.zph(coxmod_mv_all_redc)
prop
```

```
##                 chisq df       p
## Age             1.028  1 0.31058
## HR_max         14.867  1 0.00012
## SaO2_max        0.324  1 0.56901
## Temp_max        0.269  1 0.60397
## GCS_max        18.285  1 1.9e-05
## Lactate_max    31.174  1 2.4e-08
## HCT_max         2.244  1 0.13417
## Creatinine_max  0.607  1 0.43604
## Na_max          0.329  1 0.56644
## GLOBAL         57.746  9 3.6e-09
```

Confirming the proportional hazard assumption of each variable by coxzph, the p-values are small for three variables, and the global p-value are also low. Therefore, we see that there are violations that must be addressed in this model.

```
om.split <- survSplit( Surv(Days, Status) ~ ., data = icu_sub2c, cut=c(365, 730), episode= "tgroup")
coxmod_strata <- coxph(Surv(Days, Status) ~ Age + HR_max:strata(tgroup) + SaO2_max + Temp_max + HCT_max + Creatinine_max + N
a_max, data=om.split)
cox.zph(coxmod_strata)
```

```
##                       chisq df    p
## Age                   0.137  1 0.71
## SaO2_max              0.506  1 0.48
## Temp_max              0.145  1 0.70
## HCT_max               1.526  1 0.22
## Creatinine_max        0.702  1 0.40
## Na_max                1.720  1 0.19
## HR_max:strata(tgroup) 4.177  3 0.24
## GLOBAL                9.538  9 0.39
```

```
show_simple_sum(coxmod_strata)
```

```
##                                      coef exp.coef.    se.coef.         z
## Age                                0.032832700 1.0333776 0.002716817 12.084987
## SaO2_max                          -0.026830149 0.9735266 0.017899510 -1.498932
## Temp_max                          -0.076550029 0.9263066 0.052081650 -1.469808
## HCT_max                           -0.021537375 0.9786929 0.008722353 -2.469216
## Creatinine_max                     0.100314009 1.1055180 0.017064117  5.878652
## Na_max                            -0.014805388 0.9853037 0.008644662 -1.712662
## HR_max:strata(tgroup)tgroup=1      0.009072530 1.0091138 0.001796728  5.049473
## HR_max:strata(tgroup)tgroup=2     -0.007468377 0.9925594 0.006177306 -1.209002
## HR_max:strata(tgroup)tgroup=3     -0.010628687 0.9894276 0.004659188 -2.281232
##                                       Pr...z..
## Age                               1.267906e-33
## SaO2_max                          1.338913e-01
## Temp_max                          1.416137e-01
## HCT_max                           1.354094e-02
## Creatinine_max                    4.136203e-09
## Na_max                            8.677466e-02
## HR_max:strata(tgroup)tgroup=1     4.430294e-07
## HR_max:strata(tgroup)tgroup=2     2.266619e-01
## HR_max:strata(tgroup)tgroup=3     2.253473e-02
##                              test df       pvalue
## Wald test                 223.9900  9 3.036802e-43
## Likelihood ratio test     247.7759  9 2.940680e-48
## Score (logrank) test      235.3070  9 1.253770e-45
```

First, Days were grouped by <365, 365-730, >730 to three groups, and only HR_max was used to switch coefficients at the delimitations. Since GCS and Lactate are difficult to hold the PH assumptions even after this process, we decided to remove them as explanatory variables. We then confirmed that the PH assumption was valid for all variables and global. This process limited the significant variables to Age, HCT_max, and Creatine_max. For HR with stratification, the variables are significant for periods of less than 365 days and more than 730 days, but are no longer significant for the periods in between.

Similar to our `reduced_interactions2` GLM model, `coxmod_strata` showed for every year increase of `Age`, there is ~3% increase in mortality. Unsurprisingly, increases in maximum oxygen saturation (OR 0.97) and haematocrit were protective (OR 0.98). Also unsurprisingly, increases in creatinine (suggestive of poorer kidney function) increased mortality (OR 1.1). Interestingly, increase in maximum temperature was protective (OR 0.92); so was increase in maximum heart rate for `Days` > 365.

```
icu_sub2u <- icu_patients_df0 %>% dplyr::select(Days, Status, Age, HR_max, RespRate_max, SaO2_max, NISysABP_max, Urine_min,
Temp_max, GCS_max, Lactate_max, HCT_max, Creatinine_max, Na_max)
icu_sub2uc <- na.omit(icu_sub2u)

om.split.u <- survSplit( Surv(Days, Status) ~ ., data = icu_sub2uc, cut=c(365, 730), episode= "tgroup")
coxmod_strata_u <- coxph(Surv(Days, Status) ~ Age + HR_max:strata(tgroup) + SaO2_max + Temp_max + HCT_max + Creatinine_max +
Na_max, data=om.split.u)
#cox.zph(coxmod_strata_u)
show_simple_sum(coxmod_strata_u)
```

```
##                                       coef exp.coef.   se.coef.          z
## Age                                0.05081687 1.0521302 0.02997057  1.6955592
## SaO2_max                           0.12243876 1.1302499 0.14756565  0.8297240
## Temp_max                          -0.29334717 0.7457632 0.49917708 -0.5876615
## HCT_max                           -0.03766198 0.9630384 0.08755087 -0.4301725
## Creatinine_max                     0.46475991 1.5916320 0.35010058  1.3275040
## Na_max                            -0.05847782 0.9431992 0.07245750 -0.8070637
## HR_max:strata(tgroup)tgroup=1     -0.01117582 0.9888864 0.02596866 -0.4303581
## HR_max:strata(tgroup)tgroup=2     -0.10094003 0.9039872 0.10063922 -1.0029891
## HR_max:strata(tgroup)tgroup=3      0.09736730 1.1022652 0.04703329  2.0701785
##                                       Pr...z..
## Age                               0.08996939
## SaO2_max                          0.40669484
## Temp_max                          0.55675950
## HCT_max                           0.66707017
## Creatinine_max                    0.18434202
## Na_max                            0.41962978
## HR_max:strata(tgroup)tgroup=1     0.66693515
## HR_max:strata(tgroup)tgroup=2     0.31586614
## HR_max:strata(tgroup)tgroup=3     0.03843563
##                              test df    pvalue
## Wald test                 8.14000  9 0.5203770
## Likelihood ratio test    12.59906  9 0.1816031
## Score (logrank) test     10.17669  9 0.3363714
```

With only $n = 30$ observations without missing values, this model is unreliable. Indeed, most of the model coefficients have lost significance with the exception of `HR_max` for `Days` >730. However, it is worth noting the direction of influence has flipped compared to the original model (OR 1.1 vs. 0.99) with $n = 1608$. We would recommend not making any inferences from this model.