HDAT9600 Final Team Assignment

Submission deadline - 03-May-2024

Team 3: Michelle Kim z5161954, Georgina Jacko z5441162, Harvey Lee z , Zhenyu Zhang z

2024-05-03

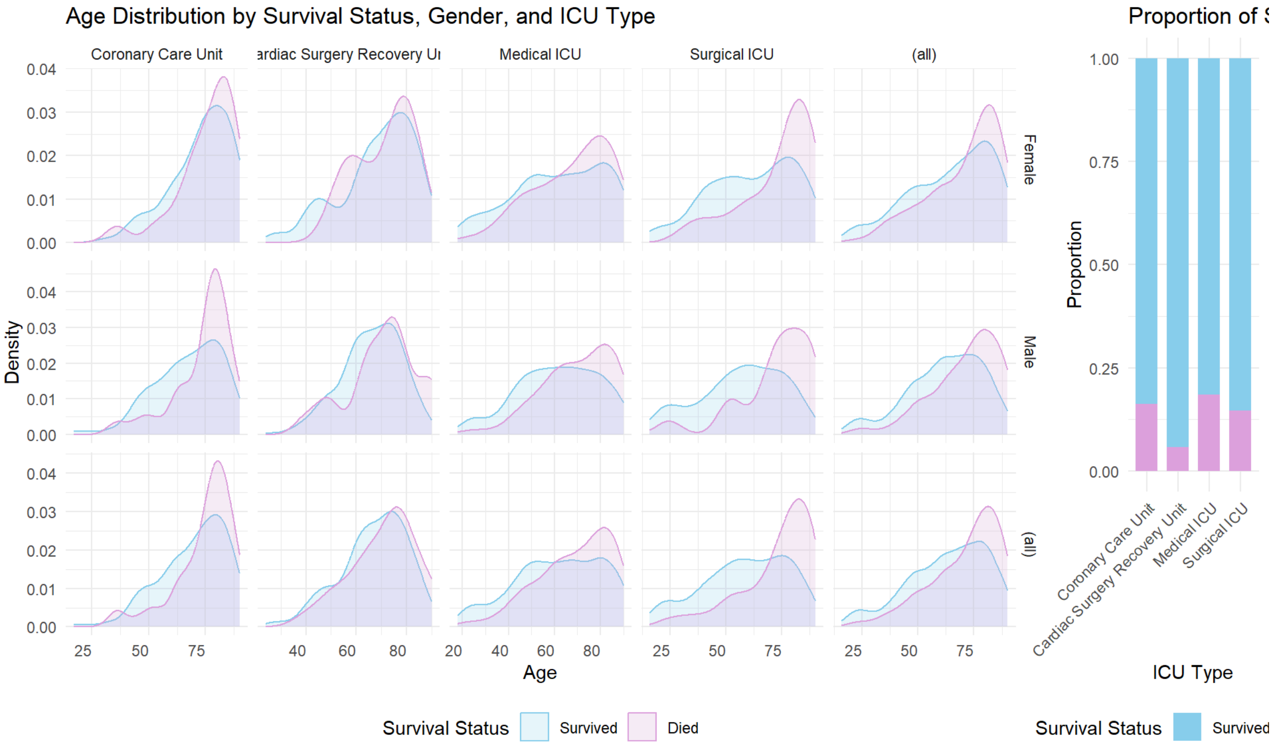
Task 1 Part A: Exploratory data analysis

Step 1: Data Clean

- **Unique Patient Count:** There were 2061 and 2061unique patients contained in the dataset df0 and df1, confirming dataset consistency in terms of patient entries across both versions.
- **NaN or Infinite Values:**
  - *DF0*: There are variables with NaN or infinite values such as `Height` , `DiasABP_diff` , `SysABP_diff` , and others mainly associated with blood pressure and weight (146 for easch of `Weight_diff` , `Weight_max` , `Weight_min`). All of them has been Converted to NA.
  - *DF1*: No NaN or infinite values reported, which indicates either preprocessing steps were different or the issues were resolved in this version of the dataset.
- **Missing Values in df1:**
  - We set a **30%** threshold for maximum acceptable proportion of missing data in df1, and impute missing values for acceptable variables using median.
  - *High Missingness in Both Datasets*: Both datasets show significant numbers of missing values across various variables, particularly clinical measurements like `Albumin_diff` , `ALP_diff` , `Bilirubin_max` , and physiological parameters like `DiasABP_max` , `NIDiasABP_max` , etc.
  - `Height` (992, 48.13% missing) given the high rate of missing values, and low probability of importance to clinical outcome it has been excluded.
  - Blood pressure-related variables (`DiasABP_diff` , `DiasABP_max` , `DiasABP_min` , `NIDiasABP_diff` , `NIDiasABP_max` , `NIDiasABP_min` , `NIMAP_diff` , `NIMAP_max` , `NIMAP_min` , `NISysABP_diff` , `NISysABP_max` , `NISysABP_min` , `SysABP_diff` , `SysABP_max` , `SysABP_min`): All these have a high degree of missingness. Since these are many related variables with substantial missing data, we therefore considered only MAP, given the complete information.
  - *Survival Variable*: A crucial variable like `Survival` has a notably high count of missing values (1288） , as it is the number of days between ICU admission and death for patients who died. Thus, no missing values impute for it.
  - *SAPS1*: `SAPS1` has 96 missing values (4.657933 % missing), which is a relatively small proportion of the dataset. Since SAPS-I is an important score for ICU prognosis, it might be worth imputing these values in df1 rather than excluding the variable.
  - *Weight-related variables* (`Weight_diff` , `Weight_max` , `Weight_min`): 146 missing values (7.08% missing) - Given this is a smaller proportion of missing data and weight is possibly an important variable for analysis it may be worth imputing values.
- **Variables to be excluded from df1 dataset due to high missing data:** DiasABP_diff, DiasABP_max, DiasABP_min, Height, SysABP_diff, SysABP_max, SysABP_min
- **Remaining total missing values in the cleaned df1 dataset:** There are SAPS1, SOFA, NIDiasABP_diff, NIDiasABP_max, NIDiasABP_min, NIMAP_diff, NIMAP_max, NIMAP_min, NISysABP_diff, NISysABP_max, NISysABP_min, Weight_diff, Weight_max, Weight_min with total 1288 missing values remaining. Given the complexity and breadth of clinical data, such a level of missingness can be considered manageable, particularly the missing values are spread across multiple variables rather than concentrated in a few.

Step 2: Exploratory data analysis

Demographics



There were 2061 patients in total, with a higher proportion of males (approximately 55.7%) compared to females. - The proportion of patients who died in the hospital was about 14.41%. Patients range in age from 16 to 90, with a median age of 67, suggesting a predominantly older patient population. Survivors have a slightly lower mean age (63.29) compared to the overall (64.41) and non-survivor (71.05) groups. The distribution of patients across ICU types shows the largest number in the Medical ICU (38.23%) and the least in the Coronary Care Unit (14.41%). As Height has been excluded from both datasets and weight excluded from df0 due to high missing data, no more consideration.

Survival in a hospital ICU is influenced by numerous factors. Older age is often associated with a higher risk of death in an ICU, and this is supported by the plots showing a notable difference in the distribution of age for survivor and in hospital died patients. The plots shows very similar proportions of survivors/in-hospital deaths for each gender, suggesting it is also not a useful predictor, the plots for ICU type showed notably different proportions of survivors/in-hospital deaths per ICU type, with a significantly higher chance of survival in the cardiac surgery recovery unit, suggesting it may have some value as a predictor.

Clinical measures
Clinical measures are critical in predicting ICU survival as they encapsulate various facets of a patient’s physiological state. In our dataset, we have segregated 36 clinical measures into different physiological systems for for fitting **Generalized Linear Models (GML)** to get the potential predictors for our model.
- **Respiratory Function** is captured by PaO2, FiO2, and RespRate. - **Coagulation & Blood** metrics include Platelets, pH, and Glucose. - **Liver Function** is indicated by Bilirubin, Albumin, ALP, ALT, and AST. - **Cardiovascular System** measures comprise HR, MAP, TroponinI, TroponinT, and Lactate. - **Central Nervous System** is assessed primarily through GCS. - **Renal Function** involves Creatinine and Urine output. - **Immune System** response is gauged by Temperature and WBC counts.

Each clinical measures represented by up to three types of variables: the minimum (_min), the maximum (_max), and the difference (_diff) observed during the first 24 hours in the ICU. Furthermore, for variables like arterial blood pressure, both invasive and non-invasive measurements are taken into account, reflecting systolic, diastolic, and mean arterial pressures. Given the logistic regression’s sensitivity to multicollinearity, it is imperative to select the most representative variable among related measures to avoid high correlations that could skew model accuracy. Therefore, our analysis aims to discern not just the significance of each clinical measure in predicting ICU outcomes, but also to identify which among the related variables (minimum, maximum, or difference) most effectively captures the impact on survival. This approach helps in honing in on the most critical predictors and reduces redundancy in the model.

Significant predictors identified include elevated levels of Albumin, AST, Bilirubin, BUN, Creatinine, GCS, Glucose, HCO3, Lactate, respiratory rate, and urine output, suggesting that their maximum values during the first 24 hours are pivotal. Conversely, substantial changes in heart rate, sodium levels, pH, temperature, and non-invasive systolic blood pressure (_diff variables) are strongly associated with ICU mortality, emphasizing the impact of acute fluctuations in physiological states.

Task 1 Part B: Predictive logistic model

Multiple logistic regression models were applied for predicting in-hospital mortality

By consider significance (`p-value`) and absolute `Estimate value` of predictors, we integrating following predictors and creating intact model strive to provide a holistic view of a patient’s health and potential survival odds in the ICU: 1. `Age` 2. `SAPS1` 3. `SOFA` 4. `Albumin_max` 5. `ALT_max` 6. `AST_max` 7. `Bilirubin_max` 8. `BUN_max` 9. `Cholesterol_max` 10. `Creatinine_max` 11. `GCS_max` 12. `GCS_min` 13. `Glucose_max` 14. `HR_diff` 15. `Lactate_max` 16. `Na_diff` 17. `NISysABP_diff` 18. `PaO2_max` 19. `pH_diff` 20. `pH_min` 21. `RespRate_max` 22. `Temp_diff` 23. `Urine_max` 24. `Urine_min` 25. `WBC_max` 26. `ICUType`

```
# Creating intact model with all significant predictors
intact_model <- glm(in_hospital_death ~ Age + SAPS1 + SOFA + Albumin_max + ALT_max + AST_max + Bilirubin_max + BUN_max + Cholesterol_max +
                        Creatinine_max + GCS_max + GCS_min + Glucose_max + HR_diff + Lactate_max + Na_diff +
                        NISysABP_diff + PaO2_max + pH_diff + pH_min + RespRate_max + Temp_diff + Urine_max +
                        Urine_min + WBC_max + ICUType,
                    family = binomial(link = "logit"), data = icu_patients_df1_cleaned)
#summary(intact_model)
# Using stepwise regression to find an optimal set of predictors for intact model
stepwise_result <- stepAIC(intact_model, direction = "both", trace = FALSE) #stepwise_result2 <- stats::step(intact_model,direction="both", trace=0)
# Variance Inflation Factor (VIF) check
vif_values <- vif(stepwise_result, type = 'terms')
#print(vif_values)
summary(stepwise_result)
```

```
##
## Call:
## glm(formula = in_hospital_death ~ Age + SAPS1 + SOFA + Albumin_max +
##     ALT_max + AST_max + Bilirubin_max + BUN_max + Creatinine_max +
##     GCS_max + GCS_min + HR_diff + NISysABP_diff + PaO2_max +
##     pH_diff + Urine_max + ICUType, family = binomial(link = "logit"),
##     data = icu_patients_df1_cleaned)
##
## Coefficients:
##                                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -2.9586700  0.7516403  -3.936 8.28e-05 ***
## Age                                   0.0261907  0.0052084   5.029 4.94e-07 ***
## SAPS1                                 0.0580310  0.0230064   2.522 0.011656 *
## SOFA                                  0.0583371  0.0280534   2.079 0.037572 *
## Albumin_max                          -0.2497119  0.1199924  -2.081 0.037428 *
## ALT_max                              -0.0010521  0.0004431  -2.375 0.017571 *
## AST_max                               0.0007670  0.0002632   2.914 0.003570 **
## Bilirubin_max                         0.0463521  0.0135900   3.411 0.000648 ***
## BUN_max                               0.0154032  0.0037981   4.055 5.00e-05 ***
## Creatinine_max                       -0.1037505  0.0541855  -1.915 0.055527 .
## GCS_max                              -0.1708445  0.0275421  -6.203 5.54e-10 ***
## GCS_min                               0.0546851  0.0282780   1.934 0.053133 .
## HR_diff                               0.0078886  0.0043060   1.832 0.066948 .
## NISysABP_diff                         0.0075894  0.0039740   1.910 0.056160 .
## PaO2_max                             -0.0018209  0.0007904  -2.304 0.021244 *
## pH_diff                               2.5952276  1.1409727   2.275 0.022931 *
## Urine_max                            -0.0007388  0.0002126  -3.475 0.000511 ***
## ICUTypeCardiac Surgery Recovery Unit -0.7094156  0.3114644  -2.278 0.022746 *
## ICUTypeMedical ICU                   -0.0874038  0.2103697  -0.415 0.677793
## ICUTypeSurgical ICU                   0.0258841  0.2331446   0.111 0.911599
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1699.7  on 2060  degrees of freedom
## Residual deviance: 1344.0  on 2041  degrees of freedom
## AIC: 1384
##
## Number of Fisher Scoring iterations: 6
```

Although it is known that SAPS1 and SOFA are calculated from some of the clinical measures, it can be seen from the correlation coefficients above that none of the predictors are highly correlated, so there is no major violation of the assumption of no multicollinearity.

Steps to Refine the Model:

- **stepAIC:**
  - method adds or removes predictors based on the AIC value in a stepwise manner to identify a model that balances model complexity with information loss. StepAIC() helped refine the model by retaining significant predictors and removing those less contributive variables.
  - The summary of the stepwise regression model presents similar significant variables and AIC as found in the previous model.
  - Predictors shows a relatively high p-value (>0.05),implies that it may not contribute meaningful explanatory power to the model in predicting in-hospital death.
- **VIF values:**
  - The VIF values common thresholds are 5 or 10, below that suggests that there isn’t a concerning level of multicollinearity among the main effects in the model.
  - High GVIF values for interaction terms are not uncommon, as these terms are products of their component variables and can inherit their collinearity.
  - High VIF/GVIF values do not imply that a model is invalid; rather, they suggest that the precision of the coefficient estimates for the related variables may be reduced.
- **Removing less important predictors:**
  - Drop Non-Significant (especially low abs of Estimate value) Predictors: `Glucose_max` , `Urine_min` , `Cholesterol_max` , `RespRate_max`

- Address Multicollinearity: `ALT_max` and `AST_max` both have high VIFs and are related liver enzymes, need redundancy.
- Clinical Relevance: `GCS_min` (representing the lowest Glasgow Coma Score), despite its higher p-value, are clinically relevant and known to be associated with patient outcomes, need keep.

```
# Then, our refined model is:
intact_model_rf <- glm(in_hospital_death ~ Age + SAPS1 + SOFA + Albumin_max  + ALT_max + AST_max + Bilirubin_max + BUN_max  + C
reatinine_max + GCS_max + GCS_min + HR_diff + Lactate_max + Na_diff + NISysABP_diff + PaO2_max + pH_diff  + Temp_diff + Urine_m
ax + WBC_max + ICUType,
                       family = binomial(link = "logit"), # for binary dependent variables
                       data = icu_patients_df1_cleaned)
```

Comparing Refined Model and Intact Model with ANOVA test:

```
# Comparing models with ANOVA
anova_results <- anova(intact_model, intact_model_rf, test = "Chisq")
print(anova_results)
```

```
## Analysis of Deviance Table
##
## Model 1: in_hospital_death ~ Age + SAPS1 + SOFA + Albumin_max + ALT_max +
##     AST_max + Bilirubin_max + BUN_max + Cholesterol_max + Creatinine_max +
##     GCS_max + GCS_min + Glucose_max + HR_diff + Lactate_max +
##     Na_diff + NISysABP_diff + PaO2_max + pH_diff + pH_min + RespRate_max +
##     Temp_diff + Urine_max + Urine_min + WBC_max + ICUType
## Model 2: in_hospital_death ~ Age + SAPS1 + SOFA + Albumin_max + ALT_max +
##     AST_max + Bilirubin_max + BUN_max + Creatinine_max + GCS_max +
##     GCS_min + HR_diff + Lactate_max + Na_diff + NISysABP_diff +
##     PaO2_max + pH_diff + Temp_diff + Urine_max + WBC_max + ICUType
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      2032     1338.2
## 2      2037     1341.7 -5    -3.43    0.634
```

Residual Degrees of Freedom: The first model (`intact_model`) has 1522 degrees of freedom, while the second model (`intact_model_rf`) has 1526. This suggests that the second model has four fewer predictors than the first.
Residual Deviance: The residual deviance for the first model is 1076.9 compared to 1081.5 for the second model. Lower residual deviance generally indicates a better fit to the data.
Difference in Deviance: The difference in deviance between the two models is -4.5503, which means that removing the four predictors (going from the first model to the second model) has increased the deviance slightly, indicating a slight loss in fit.
Pr(>Chi): The p-value of 0.3366 suggests that the increase in deviance (loss in fit) by removing these four predictors is statistically significant. It will be reasonable to prefer the simpler model (`intact_model_rf`)
Two-way interactions:

```
# Creating two-way interaction models from previous models
two_way_model <- update(intact_model_rf, . ~ .^2)
# Dropping each two-way interaction term to test significance
drop_interaction_result <- drop1(two_way_model, test = "Chisq")
drop_interaction_result
```

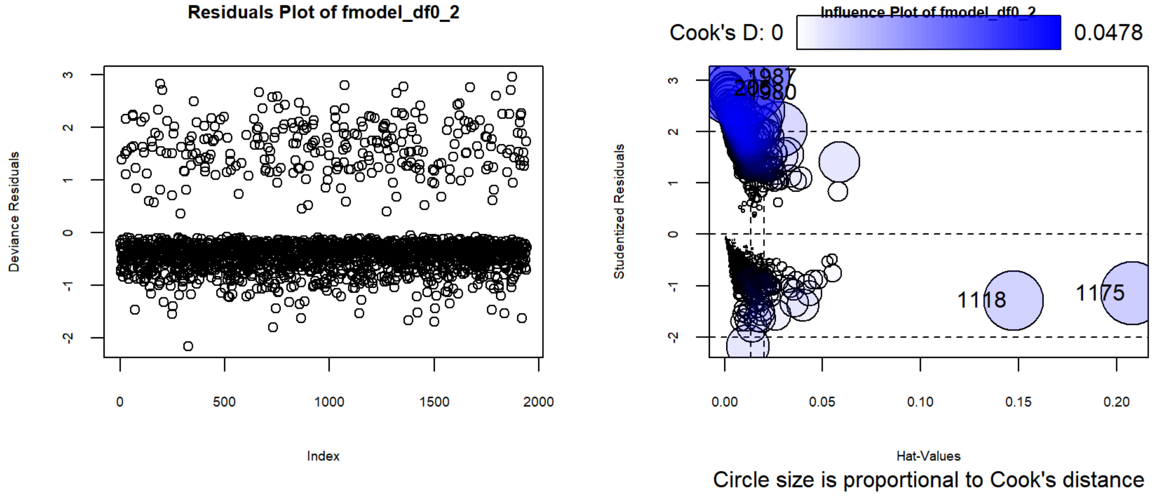
1. Most interaction terms do not significantly improve the model (indicated by high p-values), which implies that the main effects of the predictors are sufficient for the prediction.
2. A few interaction terms do show significance (indicated by low p-values), which might suggest that the relationship between predictors and the in_hospital_death is more complex than additive.
3. Notably: Considering the inclusion of interaction terms, it is essential to evaluate their clinical relevance and the potential for overfitting. Models with too many interactions, especially in smaller datasets, can fit the noise in the data rather than the underlying relationships, leading to models that may not generalize well to new data.
Inclusion of Main Effects: Typically, when including an interaction term in a regression model, should also include the main effects of the interacting variables. This approach avoids the model misattributing the effects that should be captured by the main effects alone to the interaction term.
`SOFA:pH_diff` (Pr(>Chi) = 2.912e-05) and `SOFA:WBC_max` (Pr(>Chi) = 0.000378) shown significant in two-way interaction model, they can be candidate predictors.

#### Re-fit Refined Model to the unimputed data frame df0 Before refitting the model, it’s important to acknowledge that several variables were excluded due to high levels of missing data. Proceeding with a direct refit results in a seemingly optimal AIC value of 229.87; however, this comes at the cost of losing 1819 observations. Such a reduction in the dataset size could impact the reliability and generalizability of the model.

Model Refit Adjustments:
- Exclude variables (`Albumin_max` , `ALT_max` , `AST_max` , `Bilirubin_max` , `Lactate_max` , `NISysABP_diff` , `PaO2_max` , `pH_diff`) with high levels of missing data.
- Can further exclude variables `Na_diff` (74 missing), `Urine_max` (70 missing) and `WBC_max` (76 missing).
- Exclude less important variables.
- Try include `Glucose_max` and `Urine_min` , which are viable predictors not significantly affected by missing data.
- Try introduce interaction term `SOFA:WBC_max` to explore additional predictive relationships.

```
## 
## Call:
## glm(formula = in_hospital_death ~ Age + SAPS1 + SOFA + BUN_max + 
##     Creatinine_max + GCS_max + GCS_min + HR_diff + Urine_max + 
##     ICUType, family = binomial(link = "logit"), data = icu_patients_df0_impute)
## 
## Coefficients:
##                                        Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                          -3.8803602  0.6378783  -6.083 1.18e-09 ***
## Age                                   0.0211763  0.0051627   4.102 4.10e-05 ***
## SAPS1                                 0.0720861  0.0225430   3.198 0.001385 ** 
## SOFA                                  0.0959063  0.0275033   3.487 0.000488 ***
## BUN_max                               0.0183532  0.0040510   4.531 5.88e-06 ***
## Creatinine_max                       -0.1290924  0.0613355  -2.105 0.035318 *  
## GCS_max                              -0.1690829  0.0276875  -6.107 1.02e-09 ***
## GCS_min                               0.0836136  0.0282872   2.956 0.003118 ** 
## HR_diff                               0.0071132  0.0043032   1.653 0.098335 .  
## Urine_max                            -0.0007173  0.0002140  -3.352 0.000803 ***
## ICUTypeCardiac Surgery Recovery Unit -0.9460586  0.3031276  -3.121 0.001802 ** 
## ICUTypeMedical ICU                    0.0014586  0.2115710   0.007 0.994499    
## ICUTypeSurgical ICU                   0.0233587  0.2342317   0.100 0.920563    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1609.2  on 1941  degrees of freedom
## Residual deviance: 1296.5  on 1929  degrees of freedom
##   （因为不存在，119个观察量被删除了）
## AIC: 1322.5
## 
## Number of Fisher Scoring iterations: 6
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## Age            1.246316  1        1.116385
## SAPS1          2.305918  1        1.518525
## SOFA           2.425421  1        1.557376
## BUN_max        2.309878  1        1.519828
## Creatinine_max 2.175403  1        1.474925
## GCS_max        1.801511  1        1.342204
## GCS_min        3.290971  1        1.814103
## HR_diff        1.155935  1        1.075144
## Urine_max      1.136340  1        1.065993
## ICUType        1.424330  3        1.060723
```



```
##        StudRes         Hat       CookD
## 206   2.842125 0.001170271 0.004871899
## 1118 -1.286021 0.147353090 0.016872382
## 1175 -1.142622 0.207781312 0.018912670
## 1980  2.758181 0.007815360 0.023011980
## 1987  3.061980 0.007768811 0.047825732
```

Key Observations:

Significant Predictors: Several variables are statistically significant predictors of the outcome: **Age** and **SAPS1** show a positive relationship with the likelihood of in-hospital death, indicating that higher values increase the risk. **SOFA** scores, reflecting organ failure severity, are also positively associated with mortality. **GCS_max** negatively impacts mortality, suggesting that higher GCS scores (indicating better neurological status) are

associated with lower mortality risk. **Urine_max** and **GCS_min** have notable influences with their respective positive and negative coefficients indicating their importance in the model. Among ICU types, only the **Cardiac Surgery Recovery Unit** shows a significant negative impact compared to the baseline ICU type, suggesting lower mortality risk in this unit.
Model Fit: The model reduces the residual deviance significantly from the null model, indicating a good fit to the data.
AIC and Model Complexity: The AIC of 1322.5 suggests the model is relatively efficient in balancing goodness of fit with model complexity.
Variance Inflation Factor (VIF): The VIF values common thresholds are 5 or 10, below that suggests that there isn't a concerning level of multicollinearity among the main effects in the model. Most VIFs are acceptable, though GCS_min shows a 3.29 VIF indicating potential multicollinearity issues (GCS_max exist), which might affect the stability of the regression coefficients.
The Residuals Plot of the adjusted refit logistic regression model shows a fairly random scatter of deviance residuals against the index, which is a good sign as it indicates that the residuals are well-behaved and do not exhibit any obvious patterns that might suggest non-linearity or other issues with the model specification. However, there is some noticeable spread in the residuals as they fluctuate around zero, especially evident in residuals beyond 1 and -1, suggesting that the model may not fully capture all the systematic variation in the data or may be experiencing issues such as outliers or influential observations.
The Influence Plot provides a visual assessment of the influence of individual data points on the regression estimates. This plot illustrates both the leverage (hat-values) and the impact of each point on the regression coefficients through Cook's distance. In this plot, we can see a few points with notably high Cook's distances (points 206, 1118 and 1175, notably), indicating they are influential points and potentially outliers. These points could be distorting the regression results and might warrant further investigation. The plot also highlights several other points with moderate influence on the model.

Influential point check

```
## Significant Values at or near Min/Max:
## Index 206, SOFA: 0.000000 near/at Min
## Index 206, GCS_max: 15.000000 near/at Max
## Index 206, GCS_min: 15.000000 near/at Max
## Index 206, Died in hospital
## Index 1118, GCS_min: 3.000000 near/at Min
## Index 1118, HR_diff: 212.922109 near/at Max
## Index 1118, Survivor
## Index 1175, BUN_max: 197.000000 near/at Max
## Index 1175, Creatinine_max: 22.000000 near/at Max
## Index 1175, GCS_min: 3.000000 near/at Min
## Index 1175, Survivor
```

For example, after check influential_point 1175, we find that:
- **BUN_max:**: Blood urea nitrogen measures the amount of nitrogen in the blood that comes from urea, a waste product processed by the kidneys, 197 is maximum in whole data.
- **Creatinine_max:** A maximum creatinine level of 22 is maximum in whole data.
- **GCS_min:** The minimum Glasgow Coma Scale (GCS) score of 3 is the lowest possible score, indicating severe brain injury or deep unconsciousness.
- **Urine_min:** no volume of urine output.
- **In-hospital death (0):** Despite severe kidney dysfunction, deep unconsciousness, and possible respiratory distress, the patient survived the hospital stay.This is kind of counterintuitive by consider the clinical parameters. It's might be essential to review the medical records for such patients to understand the context better, validate the data accuracy, and decide on the inclusion in the analysis.

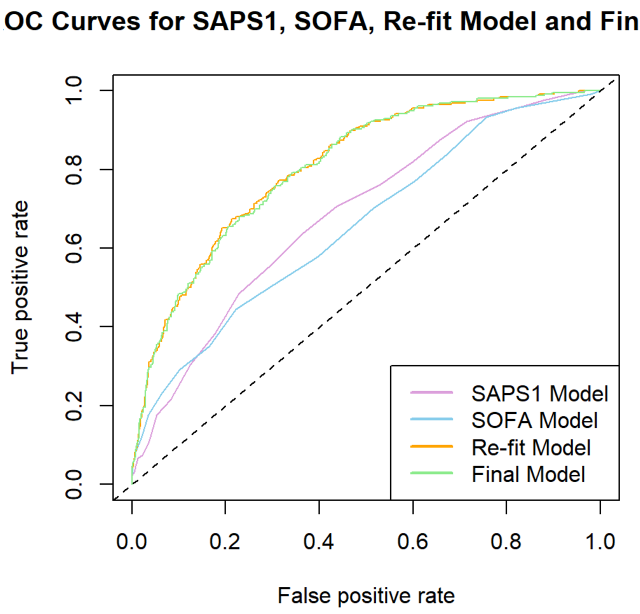
Checking and Aligning Data

```
## Brier score for SAPS1 Model:  0.1217737
```

```
## Brier score for SOFA Model:  0.1214202
```

```
## Brier score for Final Model:  0.1054758
```

```
## Brier score for Final Model with interaction terms:  0.1057297
```



```
## AUC for SAPS1 Model: 0.6864565
```

```
## AUC for SOFA Model: 0.6591956
```

```
## AUC for Re-fit Model: 0.8085496
```

```
## AUC for Final Model: 0.8068159
```

Analysis of ROC Curves and Model Performance:

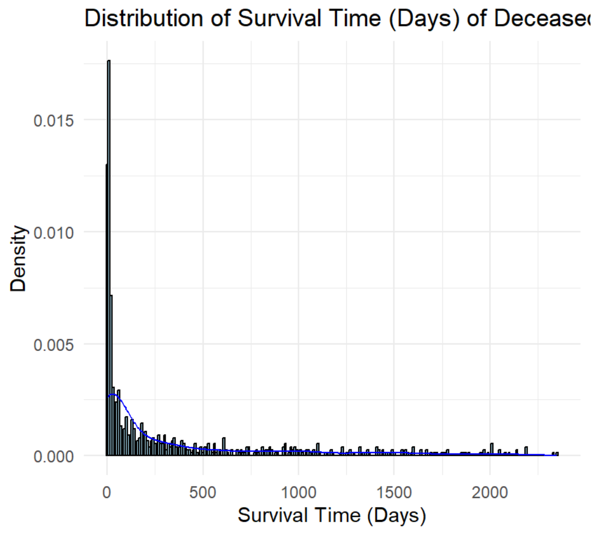
Brier Score: Lower Brier scores indicate better model calibration, where the predicted probabilities are closer to the actual outcomes. Both Re-fit Model and Final Model show better performance than the simpler SAPS1 and SOFA models, with the Re-fit model (have more) slightly outperforming Final Model.

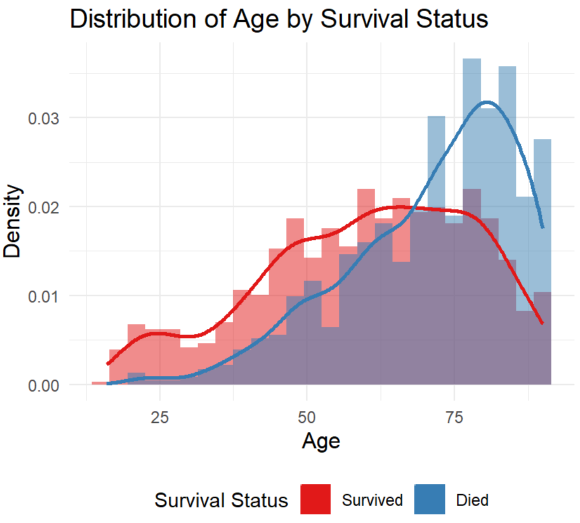
AUC: Higher AUC values indicate better discriminative ability, i.e., the model's ability to distinguish between patients who survived and those who did not. Similar to the Brier scores, Both Re-fit Model and Final Model show superior performance.

ROC Curves: The Receiver Operating Characteristic (ROC) curves show distinct performances in predicting in-hospital death. While SAPS1 and SOFA models are simpler and less computationally intensive, do not perform as well as the Re-fit Model and Final Model, which justifies the inclusion of additional predictors did improve model accuracy.

The introduction of extra predictors does not improve the Re-fit model Brier Score and AUC significantly, indicating that while these terms add complexity, they do not substantially enhance the model's ability to discriminate between outcomes.

Task 2 Part A: Exploratory data analysis



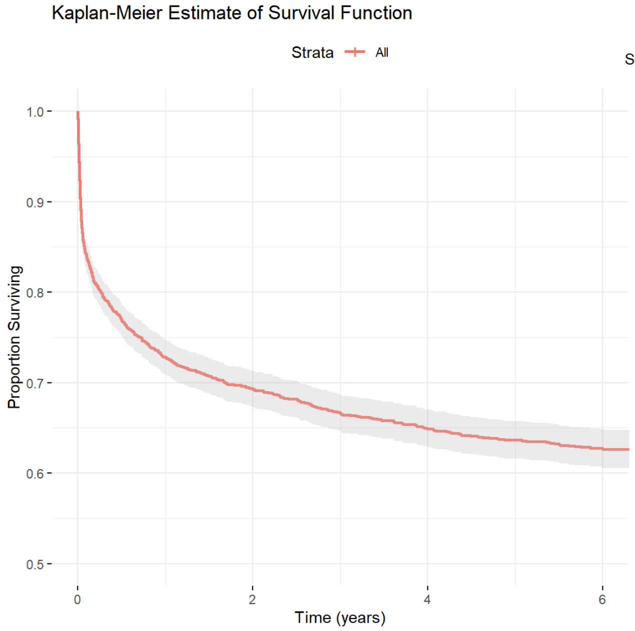


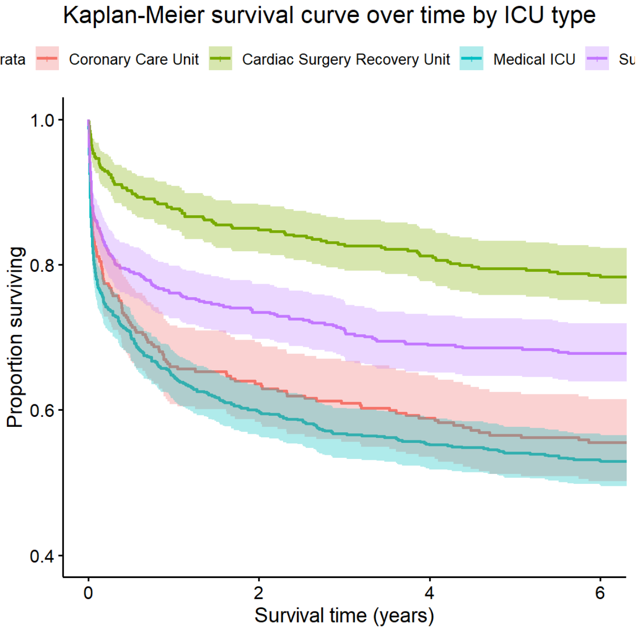
Survival Status of Patients

Status	Frequency	Proportion
Survived	1288	62.49%
Deceased	773	37.51%

Distribution of Survival Time (Days) of Deceased Patients:
- A heavy concentration of data at the lower end of the survival time spectrum, illustrateing that most deceased patients have a very short survival time post-admission. This indicates that many deaths occur relatively soon after hospital admission, which might reflect the severity of illness at the time of admission. It also underscores the urgency of effective early intervention in critical care settings, particularly for older patients.
- If including values at 2408 days, the presence of a sharp peak at the far right end could indicate right-censoring at a study endpoint or a standard follow-up period.

Distribution of Age of All Patients:
- **The distribution of ages among patients who survived** (represented by the red bars and density curve) spans a broad range, and there’s a notable peak around the middle age range, suggesting that younger patients have a higher survival rate.
- **The distribution of ages among patients who died** (depicted by the blue bars and density curve) also covers a wide age range but with a peak at an older age, indicating older patients possibly having a higher risk of mortality.
- **Elderly Vulnerability** - The right skew in the distribution of patients who died — extending further into older ages — emphasizes that elderly patients are particularly vulnerable and more likely to die. This trend might be due to the higher prevalence of comorbidities and lower physiological reserves that come with aging.
- **Overlap and Implications** - There’s a significant overlap in the age distributions around the middle age range (around 50 to 70 years), indicating that while age is a factor, other variables also play critical roles in survival outcomes. Medical interventions and health conditions could significantly impact outcomes in this age range.





Overall Survival Analysis (Left plot): The left MK curve shows the overall survival function for all patients. It depicts a sharp decline in survival probability within the initial period, followed by a gradual decrease over time. The “Number at risk” table below the curve indicates the number of individuals still being tracked at each time point, which decreases as time progresses due to events (deaths).
Survival Analysis by ICU Type (Right plot): The right MK curves stratify patients by ICU type, illustrating the survival probabilities across different units: Coronary Care Unit (CCU), Cardiac Surgery Recovery Unit (CSRU), Medical ICU, and Surgical ICU. Notably, the survival curves diverge significantly among the types, with CSRU showing higher survival probabilities but Medical ICU and CCU displaying lower probabilities throughout the observed period.

Task 2 Part B: Explanatory survival model

Univariable Cox Proportional Hazards Model

```
# In data instraction "ICU stays of less than 48 hours have been excluded.", But I still see the 1 and 2 days value in `Days` variable.
#filtered_2days_data <- icu_patients_df1_cleaned %>% filter(Days >= 2)
cox_model_age <- coxph(Surv(Days, Status) ~ Age, data = icu_patients_df1_cleaned) # not sure if we should exclude < 2 days data as "ICU stays of less than 48 hours have been excluded." data = filtered_2days_data
summary(cox_model_age)
```

```
## Call:
## coxph(formula = Surv(Days, Status) ~ Age, data = icu_patients_df1_cleaned)
##
##   n= 2061, number of events= 773
##
##        coef exp(coef) se(coef)     z Pr(>|z|)
## Age 0.03355   1.03412  0.00250 13.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## Age     1.034      0.967     1.029     1.039
##
## Concordance= 0.646  (se = 0.01 )
## Likelihood ratio test= 209.4  on 1 df,   p=<2e-16
## Wald test            = 180.1  on 1 df,   p=<2e-16
## Score (logrank) test = 187  on 1 df,   p=<2e-16
```

Coefficient (Age): The coefficients in model are positive indicating that with additional year of age, risk of the event (death) increased.
Hazard Ratio (exp(coef)): Each additional year of age increases the risk of death by approximately 3.5%.
Concordance: It measures the model’s predictive accuracy 0.646 suggest a moderate predictive ability.
Statistical Tests: All three statistical tests (Likelihood ratio test, Wald test, and Score (logrank) test) confirm the significant relationship between age and survival in both models.

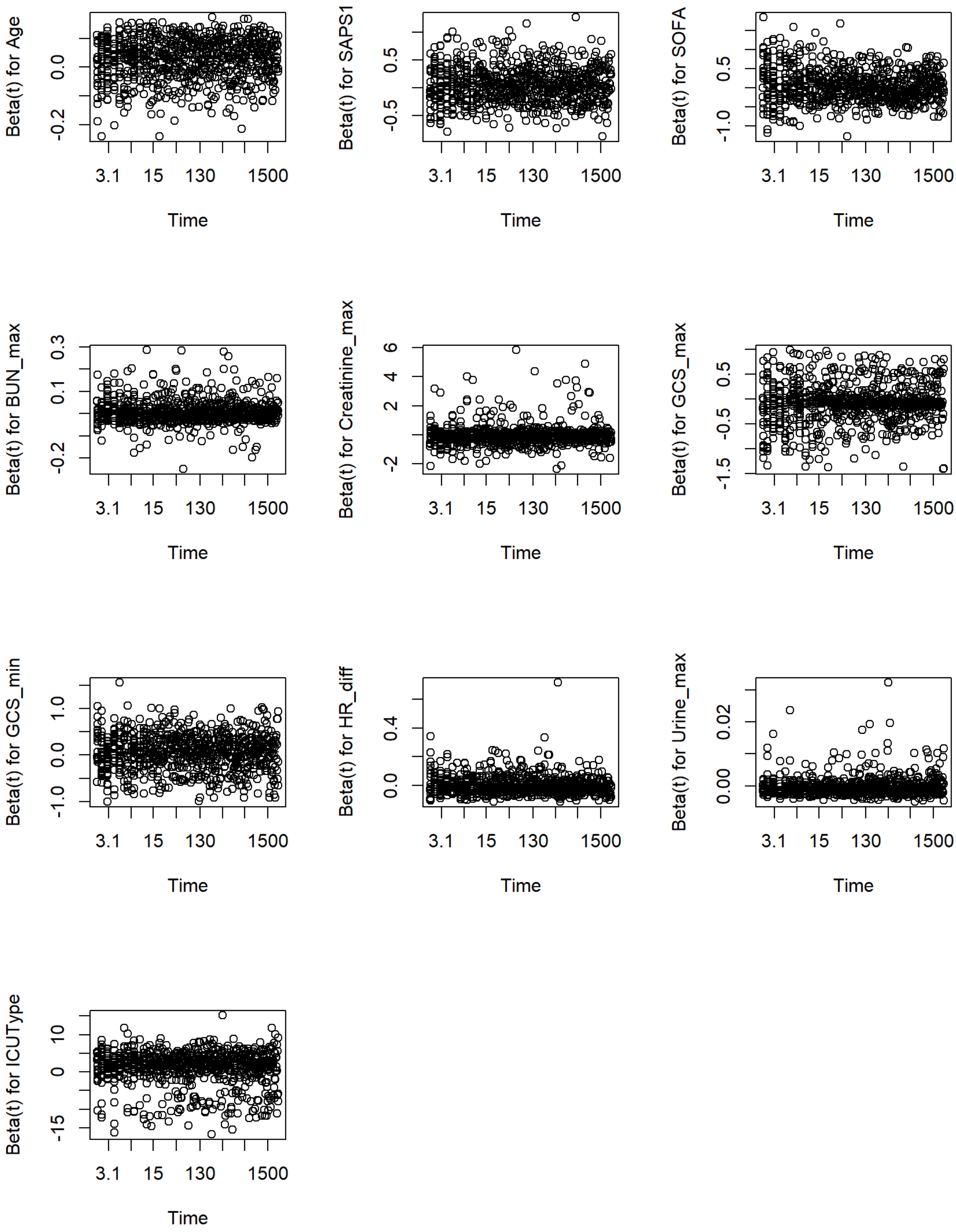
Diagnostic plots and Stepwise Cox Model

```
## Start:  AIC=10983.01
## Surv(Days, Status) ~ Age + SAPS1 + SOFA + BUN_max + Creatinine_max +
##     GCS_max + GCS_min + HR_diff + Urine_max + ICUType
##
##                  Df   AIC
## - HR_diff         1 10981
## <none>              10983
## - Creatinine_max  1 10983
## - Urine_max       1 10989
## - SOFA            1 10991
## - BUN_max         1 10999
## - GCS_min         1 11000
## - SAPS1           1 11001
## - GCS_max         1 11017
## - ICUType         3 11020
## - Age             1 11096
##
## Step:  AIC=10981.14
## Surv(Days, Status) ~ Age + SAPS1 + SOFA + BUN_max + Creatinine_max +
##     GCS_max + GCS_min + Urine_max + ICUType
##
##                  Df   AIC
## <none>              10981
## - Creatinine_max  1 10981
## + HR_diff         1 10983
## - Urine_max       1 10988
## - SOFA            1 10989
## - BUN_max         1 10997
## - GCS_min         1 10998
## - SAPS1           1 11002
## - GCS_max         1 11015
## - ICUType         3 11021
## - Age             1 11094
```

Summary of Stepwise Cox Model

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
Age	0.0274595	1.0278400	0.0026814	10.2406599	0.0000000
SAPS1	0.0527927	1.0542111	0.0110645	4.7713630	0.0000018
SOFA	0.0468344	1.0479485	0.0145792	3.2124229	0.0013162
BUN_max	0.0082433	1.0082774	0.0018997	4.3393830	0.0000143
Creatinine_max	-0.0385600	0.9621739	0.0269852	-1.4289326	0.1530236
GCS_max	-0.0882243	0.9155555	0.0146272	-6.0315338	0.0000000
GCS_min	0.0601750	1.0620224	0.0140899	4.2707758	0.0000195
Urine_max	-0.0002670	0.9997331	0.0000960	-2.7813743	0.0054129
ICUTypeCardiac Surgery Recovery Unit	-0.7530384	0.4709335	0.1477731	-5.0959100	0.0000003

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
ICUTypeMedical ICU	0.0792308	1.0824541	0.1047697	0.7562377	0.4495067
ICUTypeSurgical ICU	-0.1700844	0.8435936	0.1211985	-1.4033541	0.1605113



The `cox.zph` function in R tests the proportional hazards assumption of a Cox regression model.

1. **Plot Interpretation**:

- The x-axis represents the transformed time, and the y-axis shows the scaled Schoenfeld residuals.
- A horizontal line at zero (sometimes shown with confidence bands) is the reference line indicating no deviation from proportionality.
- If the points form a random pattern around the horizontal line, it suggests that the proportional hazards assumption holds for that variable.
- Systematic patterns, trends, or a non-random dispersion suggest a violation of the assumption.

2. **General Observation**:

- Age, GCS max, SOFAF and ICUType plots show some pattern or deviation from the zero line, particularly for Age and ICUType, which suggests potential violations of the proportional hazards assumption, indicating the effect of these variables on the hazard changes over time.

3. **The proportional hazards assumption test**: performed using the scaled Schoenfeld residuals, which should be approximately independent of time if the proportional hazards assumption holds.

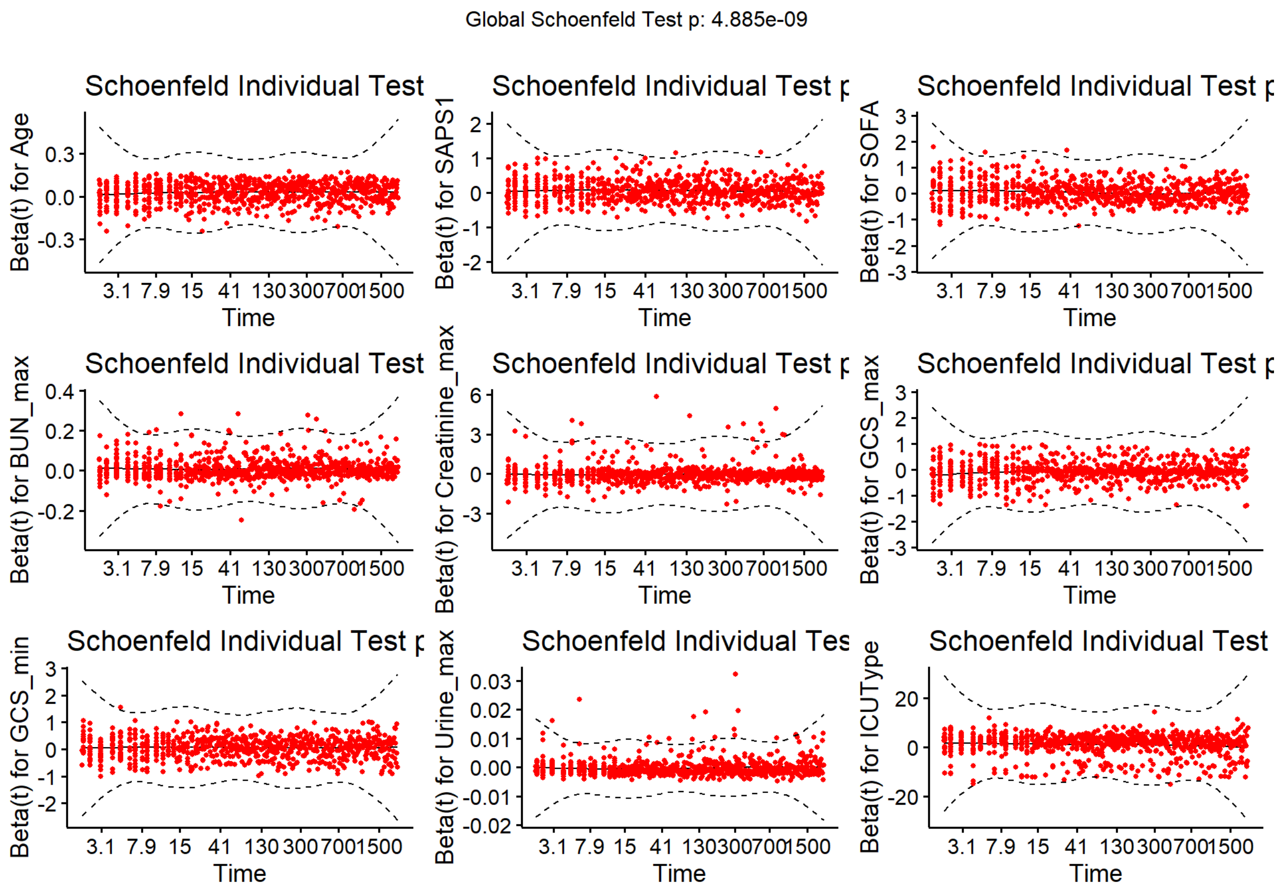
- chisq: This column shows the chi-square statistic for the test of the null hypothesis that there is no time dependence of the covariate’s effect. A higher value indicates more evidence against the null hypothesis.
- df: This column indicates the degrees of freedom associated with the chi-square test for each covariate. Most individual tests will have 1 degree of freedom unless a covariate is categorical with more than two levels (like ICUType).
- p: This column provides the p-value associated with the chi-square test. A small p-value (typically less than 0.05) suggests that the effect of the covariate is not proportional over time, indicating a violation of the proportional hazards assumption. For covariates with significant tests (low p-values), consider modeling them with time-varying coefficients or including interaction terms with time to account for their changing effects.
4. **Stepwise Cox model interpretation**:
- coef (Coefficient): This represents the estimated effect of each covariate on the hazard rate, assuming all other covariates are held constant. A positive coefficient increases the hazard rate, a negative coefficient reduces the hazard rate.
- exp(coef) (Hazard Ratio): The factor by which the hazard rate is multiplied for a one-unit increase in the covariate.Above 1 indicates increased risk; below 1 indicates decreased risk.
- se(coef) (Standard Error): The standard deviation of the estimated coefficient, indicating the precision of the estimate. Smaller values suggest more precise estimates.
- Z-score: The ratio of the coefficient to its standard error. It’s used to test the null hypothesis that the coefficient is zero (no effect).
- Pr(>|z|) (P-value): This indicates the probability of observing the given result, if the null hypothesis (that the coefficient is zero) is true. A small p-value less than 0.05 suggests that the effect is statistically significant.
5.**Model Coefficients and Hazard Ratios: (Stepwise) Cox model summary**: - **Age**: For each additional year, the hazard of death increases by about 2.78%. This is a common finding in clinical settings where older age is often associated with higher mortality risks. - **SAPS1 (Simplified Acute Physiology Score)**: Each point increase in SAPS1 is associated with a 5.42% increase in the hazard of death. This indicates that higher severity scores are linked to worse outcomes, as expected. - **SOFA (Sequential Organ Failure Assessment)**: Each point increase in the SOFA score increases the hazard by 4.79%. This score, which assesses the extent of a patient’s organ function or rate of failure, helps in predicting the likelihood of mortality, with higher scores indicating higher risk. - **BUN_max (Maximum Blood Urea Nitrogen)**: For each unit increase in BUN, the hazard of death increases by 0.83%. High BUN levels can indicate kidney dysfunction, which is a critical factor in patient outcomes. - **Creatinine_max**: Surprisingly, higher maximum creatinine levels are associated with a slightly reduced hazard of death (3.8% decrease per unit increase), although this predictor is not statistically significant (p = 0.153). - **GCS_max (Maximum Glasgow Coma Scale)**: Each point increase in the maximum GCS, which measures consciousness level, is associated with a 8.44% decrease in the hazard of death. Higher GCS scores, indicating better neurological function, are associated with better outcomes. - **GCS_min (Minimum Glasgow Coma Scale)**: Conversely, each point increase in the minimum GCS score reduces the hazard by 6.20%, emphasizing the importance of neurological status in survival outcomes. - **Urine_max**: Each unit increase in maximum urine output slightly reduces the hazard by 0.03%, reflecting better kidney function and fluid balance. - **ICUType**: The type of ICU also plays a significant role: - **Cardiac Surgery Recovery Unit**: Being in this unit reduces the hazard of death by 52.91% compared to the baseline ICU type, indicating potentially better treatment post-cardiac surgery. - **Medical ICU**: There is an increase in hazard by 8.25%, but this is not statistically significant. - **Surgical ICU**: There’s a slight reduction in hazard (15.64%), but again, it’s not significant.

Assess Model Assumptions

```
test_proportional_hazards <- cox.zph(stepwise_cox_model) # Assessing proportional hazards assumption
print(test_proportional_hazards) # Check the proportional hazards assumption
```

```
##                  chisq df       p
## Age              2.443  1  0.1180
## SAPS1           16.471  1 4.9e-05
## SOFA            27.093  1 1.9e-07
## BUN_max          2.483  1  0.1151
## Creatinine_max  0.845  1  0.3580
## GCS_max         30.817  1 2.8e-08
## GCS_min         15.997  1 6.3e-05
## Urine_max        6.170  1  0.0130
## ICUType         12.058  3  0.0072
## GLOBAL          61.495 11 4.9e-09
```

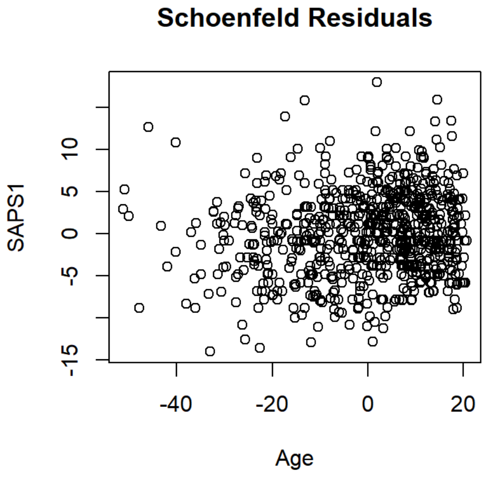
```
#plot(test_proportional_hazards) # Plotting the Schoenfeld residuals to visually inspect any trends over time
ggcoxzph(test_proportional_hazards) # the scaled Schoenfeld residuals along with a smooth curve
```

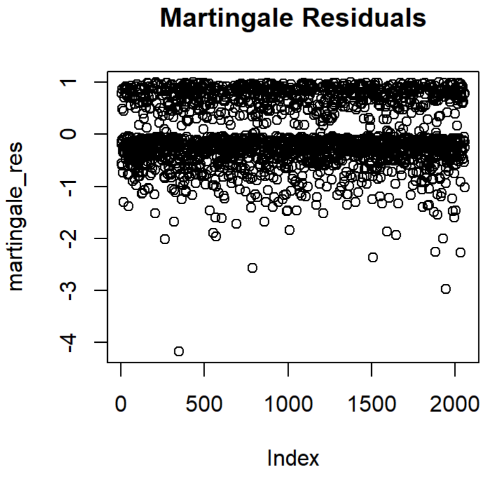


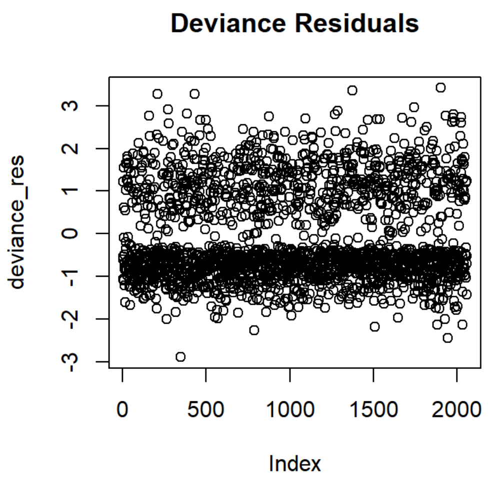
The results and the plots from the Schoenfeld residuals test indicate **significant concerns** regarding the proportional hazards assumption for several variables in your Cox regression model.

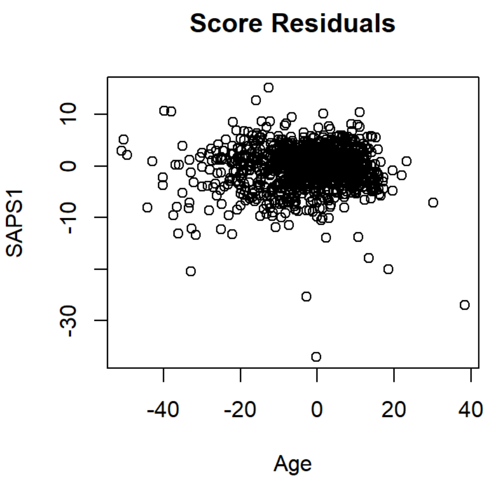
- `SAPS1` , `SOFA` , `GCS_max` , `GCS_min` , and `ICUType` show significant p-values (p < 0.05), suggesting that the effect of these covariates on the hazard changes over time. This violates the proportional hazards assumption.
- `Age` , `BUN_max` and `Creatinine_max` are the only three have p-values greater than 0.05, suggesting they meet the proportional hazards assumption, means effect on the hazard is consistent over time. However, `Age` and `BUN_max` have p-values close to 0.1, indicating a borderline case where reviewing stability over time still needs to be considered.
- In addition, the `global` test has a p-value of 4.9e-09, indicating strong evidence against the proportional hazards assumption for the model as a whole.
- For variables that violate the assumption, stratification or using time-dependent covariates in the model is the method to handle the non-proportionality.Alternatively, interactions between these variables and time could be modeled to better fit the data.
- In addition, based on summary of both multivariable Cox Model and stepwise cox model, variable `HR_diff` can be dropped.

Diagnostic and Goodness-of-Fit









1. Schoenfeld Residuals:
- Calculates Schoenfeld residuals, which are used to check the proportional hazards assumption of a Cox model. Each residual represents the contribution of an individual covariate to the hazard at each event time.
- Plots these residuals against time., if the residuals display any systematic pattern over time, it suggests a violation of the proportional hazards assumption.
- In this case, residuals lack of clear patterns would suggest that the proportional hazards assumption holds for age.

2. Martingale Residuals:
- Computes Martingale residuals, which represent the difference between the observed number of events and the expected number under the model, for each individual.
- Detecting non-linear effects of covariates and potential outliers. Ideally, residuals should be randomly scattered around zero without clear patterns.
- In deviance residuals plot, a random scatter indicates good model fit, while patterns or trends could suggest model inadequacies.

3. Deviance Residuals:
- Calculates deviance residuals, providing a measure of how well the model predicts each observation.
- Like Martingale residuals, these should ideally scatter randomly around zero, aiding in identifying poorly predicted observations.
- A random scatter indicates good model fit.

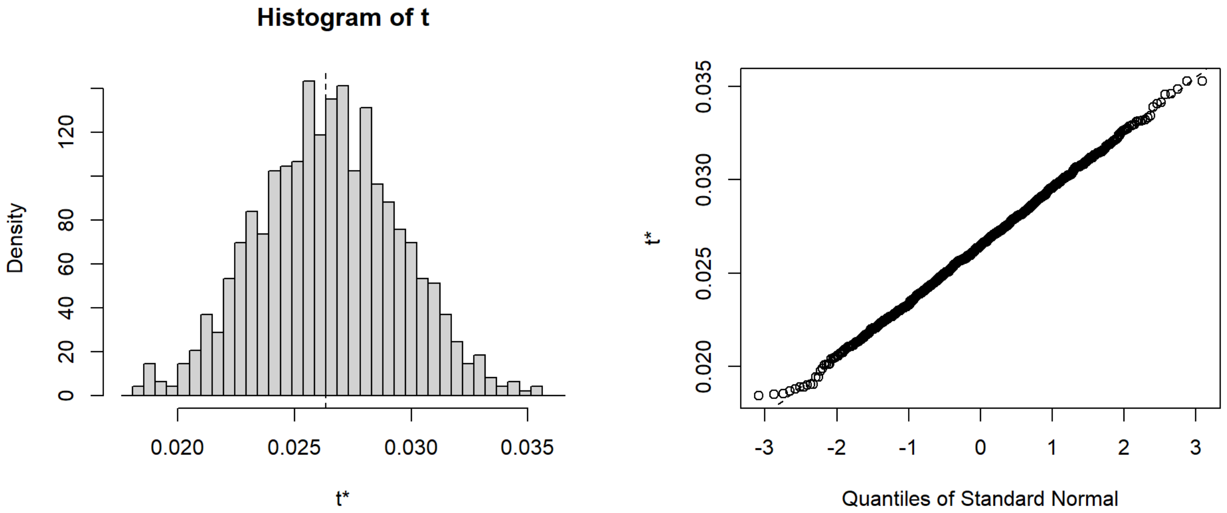
4. Score Residuals:
- Computes score residuals, assessing the contribution of each observation to the overall model fit.
- Identify influential cases where an individual observation markedly affects the coefficient estimates.
- Dense clustering without extreme outliers is ideal.

```
# Fit initial Cox model to df0 data
initial_cox_model_df0 <- coxph(Surv(Days, Status) ~ Age + SAPS1 + SOFA +  BUN_max  + Creatinine_max + GCS_max + GCS_min + Urine
_max + ICUType,
                          data = icu_patients_df0_cleaned)
summary(initial_cox_model_df0)
# Fit proportional hazards assumption violate Cox model to df0 data
ph_violate_cox_model_df0 <- coxph(Surv(Days, Status) ~ SAPS1 + SOFA + GCS_max + GCS_min + Urine_max + ICUType,
                          data = icu_patients_df0_cleaned)
summary(ph_violate_cox_model_df0)
# Fit proportional hazards assumption accept Cox model to df0 data
ph_accept_cox_model_df0 <- coxph(Surv(Days, Status) ~ Age + BUN_max + Creatinine_max,
                          data = icu_patients_df0_cleaned)
summary(ph_accept_cox_model_df0)
# Check proportional hazards assumption
cox.zph_res_df0 <- cox.zph(initial_cox_model_df0)
print(cox.zph_res_df0)
plot(cox.zph_res_df0) #diagnostic plots

#stepwise_cox_model_df0 <- stepAIC(initial_cox_model_df0, direction = "both") # Perform stepwise selection based on AIC
#kable(summary(stepwise_cox_model_df0)$coefficients, caption = "Summary of Stepwise Cox Model")
```

Please Note: When Perform stepwise selection on re-fit Cox model, the error “number of rows in use has changed: remove missing values?” occurred, as the data used for the Cox model fitting contains missing values.

Bootstrapping to validate the model:



- **Histogram of t**: represents the distribution of the bootstrapped estimates for one of the coefficients from the Cox proportional hazards model. It helps visualize the variability and bias of the estimate. The variable `t` on the x-axis represents the coefficient values obtained in each bootstrap iteration. In here, the distribution looks symmetric about the mean, suggesting that the bootstrap samples do not show much bias.
- **Quantile-Quantile (QQ) plot**: This plot compares the quantiles of the bootstrap estimates against the quantiles of a standard normal distribution. The points lie close to the diagonal line, which implies that the distribution of the bootstrap estimates is approximately normal.

####Analysis of Survival Time by Time Group:**

```
# Splitting the dataset into 'Early' and 'Late' groups based on a 1200-day threshold (close to median)
icu_patients_df0_impute$time_group <- ifelse(icu_patients_df0_impute$Days <= 1200, "Early", "Late")
# Fitting Cox proportional hazards models separately for each time group to investigate the effects of clinical variables on su
rvival
cox_model_early <- coxph(Surv(Days, Status) ~ Age + SAPS1 + SOFA + BUN_max + Creatinine_max + GCS_max + GCS_min + Urine_max + I
CUType + strata(time_group),
                         data = subset(icu_patients_df0_impute, time_group == "Early"), ties = "breslow")
cox_model_late <- coxph(Surv(Days, Status) ~ Age + SAPS1 + SOFA + BUN_max + Creatinine_max + GCS_max + GCS_min + Urine_max + IC
UType + strata(time_group),
                        data = subset(icu_patients_df0_impute, time_group == "Late"), ties = "breslow")
summary(cox_model_early)
summary(cox_model_late)
# Additionally, fitting a model with time interactions to analyze how predictor effects change over time
cox_model_interaction <- coxph(Surv(Days, Status) ~ Age + SAPS1 * time_group + SOFA * time_group + BUN_max + Creatinine_max + G
CS_max * time_group + GCS_min * time_group + Urine_max + ICUType * time_group,
                               data = icu_patients_df0_impute, ties = "breslow")
summary(cox_model_interaction) # Check the model summary
```

Early Phase:
- `SAPS1` , `SOFA` , `BUN_max` , `GCS_max` , `GCS_min` , and `ICUType Cardiac Surgery Recovery Unit` are significant predictors of the hazard, with `SAPS1` and `SOFA` showing a positive relationship, indicating higher scores are associated with a higher hazard of death or event.
- `Age` and `Creatinine_max` do not show significant effects during this early phase.
- The `ICUType Cardiac Surgery Recovery Unit` shows a notable decrease in hazard (HR < 1), suggesting a lower risk compared to the baseline ICU type.

Late Phase:
- `Age` is the only significant predictor with a positive coefficient, suggesting that as age increases, the risk of death or an event increases in the later phase of follow-up.
- Other covariates, including `SAPS1` , `SOFA` , and markers such as `BUN_max` and `Creatinine_max` , are not significant, which might suggest that their impact is more pronounced early in the treatment or their effects have stabilized in the late phase.

Combined Analysis with Time Interaction: Interactions between the covariates and the time variable (time_group) can also be added to a single model to assess if the effect of these covariates changes significantly over the different time periods.
- The interaction terms generally are not significant, indicating that the proportional hazards assumption might still hold when splitting the model into two time periods without needing to include interactions explicitly.
- The main effects of `SAPS1` , `SOFA` , `GCS_max` , and `GCS_min` remain significant, but their hazard ratios shift slightly, likely due to adjusting for the split in follow-up time.

References:

- Bagshaw, S. M., Webb, S. A., Delaney, A., George, C., Pilcher, D., Hart, G. K., & Bellomo, R. (2009). Very old patients admitted to intensive care in Australia and New Zealand: A multi-centre cohort analysis. Critical Care, 13(2), R45. https://ccforum.biomedcentral.com/articles/10.1186/cc7768 (https://ccforum.biomedcentral.com/articles/10.1186/cc7768)
- Boumendil, A., Aegerter, P., Guidet, B., & CUB-Réa Network. (2005). Treatment intensity and outcome of patients aged 80 and older in intensive care units: A multicenter matched-cohort study. Journal of the American Geriatrics Society, 53(1), 88-93.

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-5415.2005.53016.x (https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-5415.2005.53016.x)

- de Rooij, S. E., Govers, A. C., Korevaar, J. C., Giesbers, A. W., Levi, M., & de Jonge, E. (2008). Cognitive, functional, and quality of life outcomes of patients aged 80 and older who survived at least one year after planned or unplanned surgery or medical intensive care treatment. Journal of the American Geriatrics Society, 56(5), 816-822. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1532-5415.2008.01671.x (https://onlinelibrary.wiley.com/doi/full/10.1111/j.1532-5415.2008.01671.x)
- Lerolle, N., Trinquart, L., Bornstain, C., Tadié, J. M., Imbert, A., Diehl, J. L., Fagon, J. Y., & Guérot, E. (2010). Increased intensity of treatment and decreased mortality in elderly patients in an intensive care unit over a decade. Archives of Internal Medicine, 170(1), 59-65. https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/415749 (https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/415749)
- Valentin, A., Jordan, B., Lang, T., Hiesmayr, M., & Metnitz, P. G. (2003). Gender-related differences in intensive care: A multiple-center cohort study of therapeutic interventions and outcome in critically ill patients. Critical Care Medicine, 31(7), 1901-1907. https://journals.lww.com/ccmjournal/Abstract/2003/07000/Gender_related_differences_in_intensive_care__A.2.aspx (https://journals.lww.com/ccmjournal/Abstract/2003/07000/Gender_related_differences_in_intensive_care__A.2.aspx)
- Fuchs, L., Chronaki, C. E., Park, S., Novack, V., Baumfeld, Y., Scott, D., McLennan, S., Talmor, D., & Celi, L. A. (2012). ICU admission characteristics and mortality rates among elderly and very elderly patients. Intensive Care Medicine, 38(10), 1654-1661. https://link.springer.com/article/10.1007/s00134-012-2629-6 (https://link.springer.com/article/10.1007/s00134-012-2629-6)
- Minne, L., Abu-Hanna, A., & de Jonge, E. (2008). Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review. Critical Care, 12(6), R161. https://ccforum.biomedcentral.com/articles/10.1186/cc7160 (https://ccforum.biomedcentral.com/articles/10.1186/cc7160)
- Sacanella, E., Pérez-Castejón, J. M., Nicolás, J. M., Masanés, F., Navarro, M., Castro, P., & López-Soto, A. (2009). Functional status and quality of life 12 months after discharge from a medical ICU in healthy elderly patients: A prospective observational study. Critical Care, 13(2), R46. https://ccforum.biomedcentral.com/articles/10.1186/cc7769 (https://ccforum.biomedcentral.com/articles/10.1186/cc7769)
- Heyland, D. K., Cook, D. J., Rocker, G. M., Dodek, P. M., Kutsogiannis, D. J., & Peters, S. (2005). The attributable morbidity and mortality of ventilator-associated pneumonia in the critically ill patient. The American Journal of Respiratory and Critical Care Medicine, 171(10), 1173-1179. https://www.atsjournals.org/doi/full/10.1164/rccm.200405-644OC (https://www.atsjournals.org/doi/full/10.1164/rccm.200405-644OC)
- Iapichino, G., Morabito, A., Mistraletti, G., Ferla, L., & Radrizzani, D. (2010). Age, illness severity, and ICU organizational factors influence long-term survival