Survival Analysis Exam Arezki BALI:

2025-12-21

```
library(tidyverse)
```

```
## Warning: package 'tidyr' was built under R version 4.5.2
```

```
## Warning: package 'readr' was built under R version 4.5.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.6
## v forcats   1.0.1     v stringr   1.6.0
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.2
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(survival)
library(broom)
```

```
## Warning: package 'broom' was built under R version 4.5.2
```

```
library(survminer)   # For Kaplan-Meier plots
```

```
## Loading required package: ggpubr
##
## Attaching package: 'survminer'
##
## The following object is masked from 'package:survival':
##
##     myeloma
```

```
library(survivalROC) # For Time-Dependent ROC
```

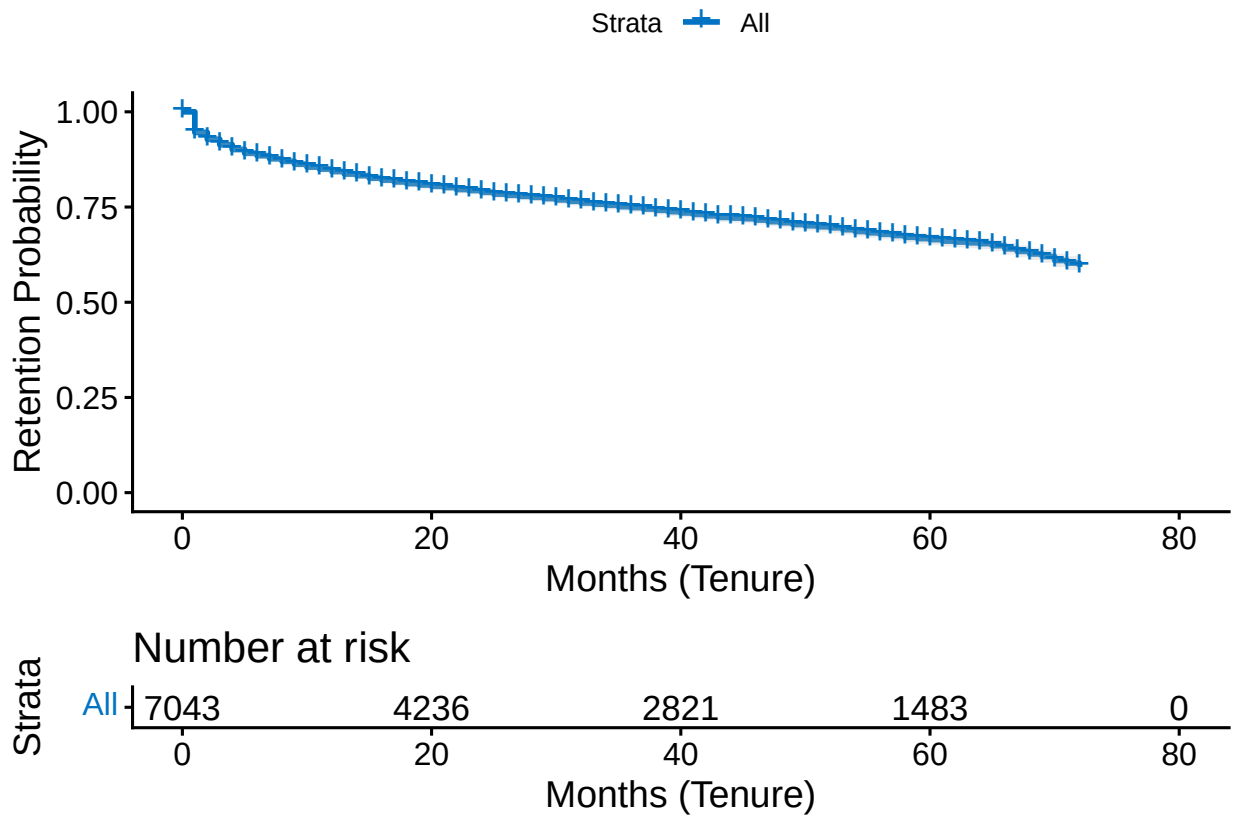
```
# Data Loading
dat <- read.csv("/Users/arezkibali/Downloads/WA_Fn-UseC_-Telco-Customer-Churn.csv")
```

```
# Convert 'TotalCharges' to numeric (this introduces NAs for blank strings)
dat$TotalCharges <- as.numeric(dat$TotalCharges)
```

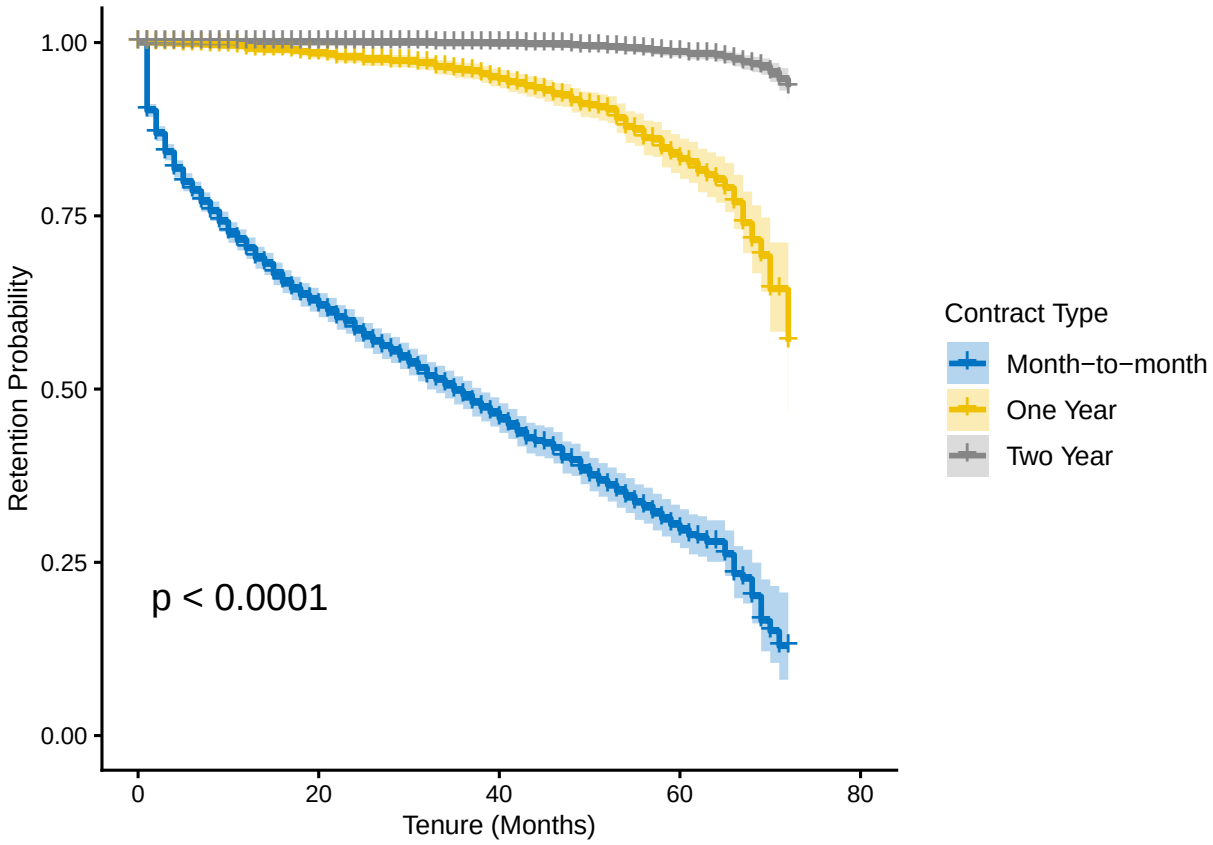
```r
# Handle Missing Values (The "Emicvaz" improvement: Impute 0 instead of dropping)
# Logic: If tenure is 0, TotalCharges should be 0.
dat$TotalCharges[is.na(dat$TotalCharges)] <- 0

# Convert Target 'Churn' to Binary (1 = Yes, 0 = No)
dat$Churn <- ifelse(dat$Churn == "Yes", 1, 0)

# Convert Categorical Predictors to Factors
dat$Contract <- as.factor(dat$Contract)
dat$PaymentMethod <- as.factor(dat$PaymentMethod)
dat$InternetService <- as.factor(dat$InternetService)
dat$gender <- as.factor(dat$gender)
```

```r
#  Kaplan-Meier Plots)

# 1. Fit the Kaplan-Meier overall
km_fit_contract <- survfit(Surv(tenure, Churn) ~ 1, data = dat)

# 2. Plot
ggsurvplot(km_fit_contract,
           data = dat,
           pval = TRUE,             # Show p-value
           risk.table = TRUE,       # Show table of people at risk
           conf.int = TRUE,         # Show confidence intervals
           palette = "jco",         # Nice color palette
           main = "Kaplan-Meier: Actual Retention by Contract Type",
           xlab = "Months (Tenure)",
           ylab = "Retention Probability")
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## i The deprecated feature was likely used in the ggpubr package.
##   Please report the issue at <https://github.com/kassambara/ggpubr/issues>.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## Warning in .pvalue(fit, data = data, method = method, pval = pval, pval.coord = pval.coord, : There a
##  This is a null model.

## Ignoring unknown labels:
## * fill : "Strata"
## Ignoring unknown labels:
## * fill : "Strata"
## Ignoring unknown labels:
## * fill : "Strata"
## Ignoring unknown labels:
## * fill : "Strata"
## Ignoring unknown labels:
## * colour : "Strata"
```




```
#  Kaplan-Meier by contract
km_fit_contract <- survfit(Surv(tenure, Churn) ~ Contract, data = dat)

ggsurvplot(km_fit_contract,
           data = dat,
           pval = TRUE,
           conf.int = TRUE,
           palette = "jco",

           # --- VISUAL FIXES ---
           legend = "right",
           legend.title = "Contract Type",
           legend.labs = c("Month-to-month", "One Year", "Two Year"),
           xlab = "Tenure (Months)",
           ylab = "Retention Probability",
           font.x = 10,
           font.y = 10,
           font.tickslab = 9
)
```




```
# Cox: MODEL BUILDING


# 1. Define the Full Model WITHOUT 'TotalCharges'
# Total Charges is mainly Monthly Charge* Tenure...sort of colinearity that will effect the model.

Mfull <- coxph(Surv(tenure, Churn) ~ MonthlyCharges + Contract +
                 InternetService + PaymentMethod + PaperlessBilling +
                 gender + Partner + Dependents,
               data = dat)

# 2. Run Standard Stepwise Selection
MAIC <- step(Mfull, direction = "both", trace = 0)

print("Best Model Selected")
```

```
## [1] "Best Model Selected"
```

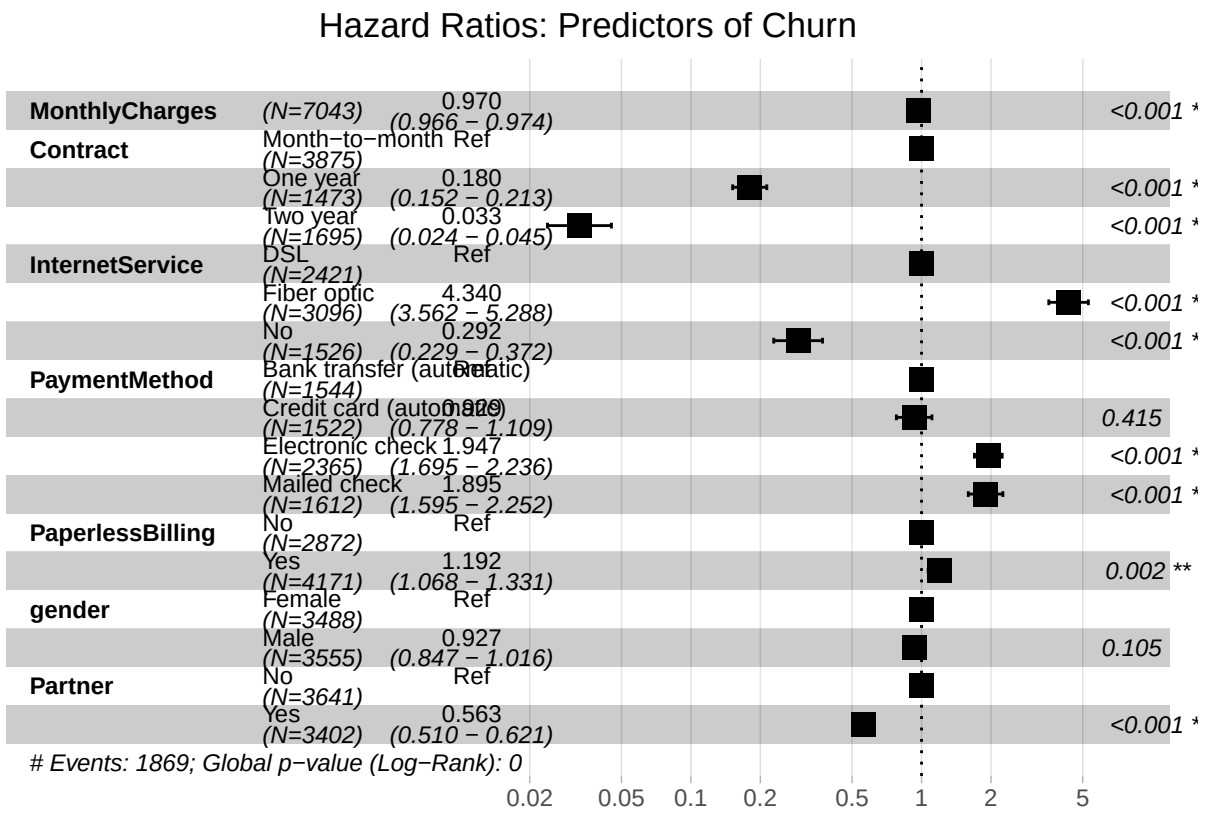
```
summary(MAIC)
```

```
## Call:
## coxph(formula = Surv(tenure, Churn) ~ MonthlyCharges + Contract +
##     InternetService + PaymentMethod + PaperlessBilling + gender +
##     Partner, data = dat)
##
```

```
##   n= 7043, number of events= 1869
##
##                                        coef exp(coef)  se(coef)       z
## MonthlyCharges                       -0.030634  0.969830  0.002138 -14.327
## ContractOne year                     -1.715405  0.179891  0.087037 -19.709
## ContractTwo year                     -3.415544  0.032859  0.162976 -20.957
## InternetServiceFiber optic            1.467957  4.340359  0.100783  14.566
## InternetServiceNo                    -1.232081  0.291685  0.124228  -9.918
## PaymentMethodCredit card (automatic) -0.073819  0.928840  0.090607  -0.815
## PaymentMethodElectronic check         0.666345  1.947107  0.070686   9.427
## PaymentMethodMailed check             0.639425  1.895391  0.088069   7.260
## PaperlessBillingYes                   0.175693  1.192073  0.056092   3.132
## genderMale                           -0.075324  0.927443  0.046401  -1.623
## PartnerYes                           -0.575206  0.562589  0.050268 -11.443
##                                      Pr(>|z|)
## MonthlyCharges                        < 2e-16 ***
## ContractOne year                      < 2e-16 ***
## ContractTwo year                      < 2e-16 ***
## InternetServiceFiber optic            < 2e-16 ***
## InternetServiceNo                     < 2e-16 ***
## PaymentMethodCredit card (automatic)  0.41523
## PaymentMethodElectronic check         < 2e-16 ***
## PaymentMethodMailed check            3.86e-13 ***
## PaperlessBillingYes                   0.00173 **
## genderMale                            0.10452
## PartnerYes                            < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                      exp(coef) exp(-coef) lower .95 upper .95
## MonthlyCharges                         0.96983     1.0311   0.96577   0.97390
## ContractOne year                       0.17989     5.5589   0.15168   0.21335
## ContractTwo year                       0.03286    30.4335   0.02387   0.04522
## InternetServiceFiber optic             4.34036     0.2304   3.56237   5.28825
## InternetServiceNo                      0.29168     3.4284   0.22865   0.37210
## PaymentMethodCredit card (automatic)   0.92884     1.0766   0.77771   1.10934
## PaymentMethodElectronic check          1.94711     0.5136   1.69520   2.23645
## PaymentMethodMailed check              1.89539     0.5276   1.59490   2.25249
## PaperlessBillingYes                    1.19207     0.8389   1.06797   1.33060
## genderMale                             0.92744     1.0782   0.84682   1.01574
## PartnerYes                             0.56259     1.7775   0.50980   0.62084
##
## Concordance= 0.852  (se = 0.004 )
## Likelihood ratio test= 3245  on 11 df,   p=<2e-16
## Wald test            = 1813  on 11 df,   p=<2e-16
## Score (logrank) test = 2929  on 11 df,   p=<2e-16
```

```
# VISUALIZATION: FOREST PLOT (Hazard Ratios)

library(survminer)

# Create the Forest Plot
p <-ggforest(MAIC,
         data = dat,
```

```
        main = "Hazard Ratios: Predictors of Churn",
        fontsize = 0.8,
        refLabel = "Ref",
        cpositions = c(0.02, 0.22, 0.4)
)

p
```



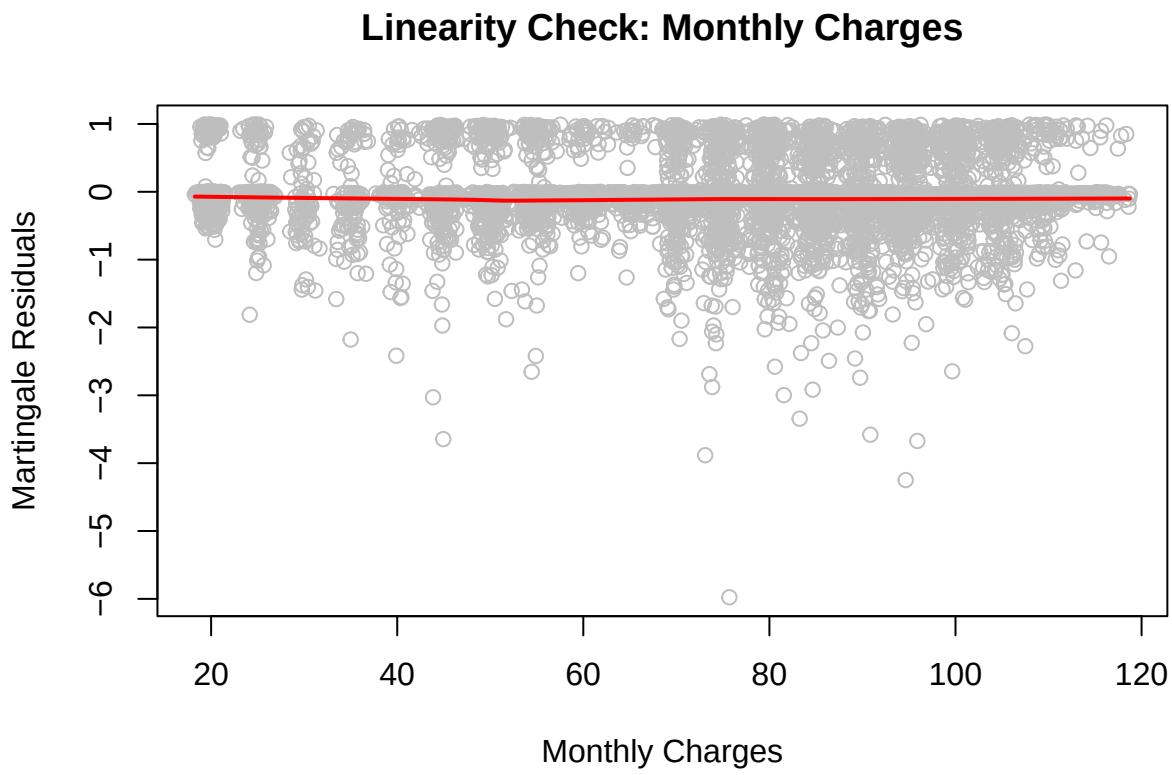

```
# Save  with specific dimensions (Width x Height)
ggsave("ForestPlot_Tall.png", plot = p, width = 10, height = 12)
```

```
#VALIDATION: TRAIN/TEST SPLIT

set.seed(123) # For reproducibility
train_index <- sample(1:nrow(dat), size = 0.5 * nrow(dat)) # 50% split
d_train <- dat[train_index, ]
d_test  <- dat[-train_index, ]

# Fit model on TRAINING data only
M_train <- coxph(Surv(tenure, Churn) ~ MonthlyCharges + Contract +
                   InternetService + PaymentMethod + PaperlessBilling,
                 data = d_train)

# Predict risk scores on TESTING data
d_test$risk_score <- predict(M_train, newdata = d_test, type = "lp")
```

```r
# Calculate C-Statistic (Concordance) on Test Data
# 0.5 = Random, 1.0 = Perfect
perf_test <- coxph(Surv(tenure, Churn) ~ risk_score, data = d_test)
c_index <- summary(perf_test)$concordance[1]

print(paste("C-Statistic (Test Set):", round(c_index, 3)))
```

```
## [1] "C-Statistic (Test Set): 0.847"
```

```r
# DIAGNOSTICS: CHECKING ASSUMPTIONS

# A. Martingale Residuals (Check Linearity of MonthlyCharges)
dat$residuals <- residuals(MAIC, type = "martingale")
plot(dat$MonthlyCharges, dat$residuals, col = "gray",
     main = "Linearity Check: Monthly Charges",
     xlab = "Monthly Charges", ylab = "Martingale Residuals")
lines(lowess(dat$MonthlyCharges, dat$residuals), col = "red", lwd = 2)
```



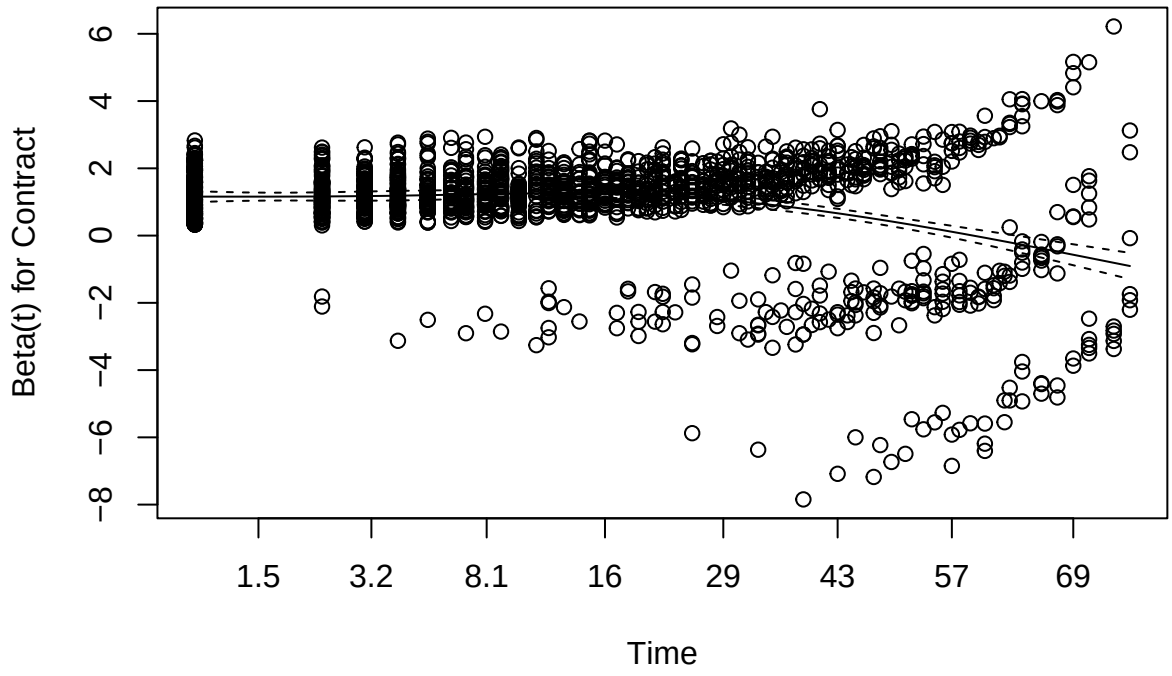

```r
# B. Proportional Hazards Test (Schoenfeld Residuals)
# Checks if risk ratios are constant over time
test_ph <- cox.zph(MAIC)
print(" Proportional Hazards Test ")
```

```
## [1] " Proportional Hazards Test "
```

```r
print(test_ph)
```

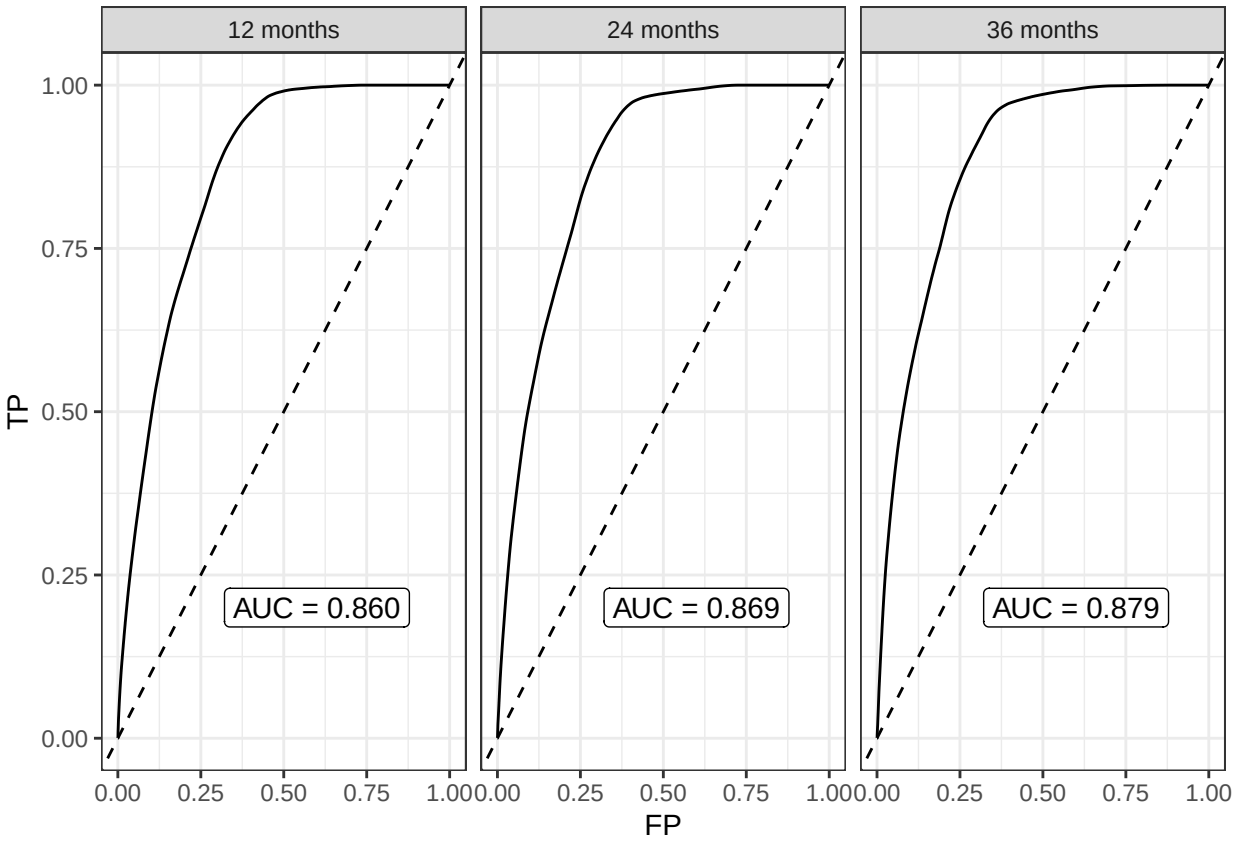
```
##                     chisq df       p
## MonthlyCharges    105.393  1 < 2e-16
## Contract           89.719  2 < 2e-16
## InternetService    41.496  2 9.8e-10
## PaymentMethod      48.453  3 1.7e-10
## PaperlessBilling    3.732  1   0.053
## gender              0.012  1   0.913
## Partner            20.289  1 6.7e-06
## GLOBAL            253.807 11 < 2e-16
```

```r
# Visual check for 'Contract' (Likely violator)
plot(test_ph, var = "Contract")
```




```r
library(survivalROC)
library(dplyr)
library(purrr)
library(tidyr)
library(ggplot2)

times_to_check <- c(12, 24, 36)

# If Churn and tenure are already correct, keep it simple:
```

```r
d_test <- dat %>%
  mutate(lp = predict(MAIC, newdata = ., type = "lp"))

roc_df <- map_df(times_to_check, function(t) {
  roc <- survivalROC(
    Stime        = d_test$tenure,
    status       = d_test$Churn,
    marker       = d_test$lp,
    predict.time = t,
    method       = "NNE",
    span         = 0.25 * nrow(d_test)^(-0.20)
  )

  tibble(
    t   = t,
    auc = roc$AUC,
    FP  = roc$FP,
    TP  = roc$TP
  )
})

ggplot(roc_df, aes(FP, TP)) +
  geom_line() +
  geom_abline(linetype = "dashed") +
  geom_label(
    data = roc_df %>% distinct(t, auc),
    aes(x = 0.6, y = 0.2, label = sprintf("AUC = %.3f", auc)),
    inherit.aes = FALSE
  ) +
  facet_wrap(~t, labeller = labeller(t = \(x) paste0(x, " months"))) +
  theme_bw()
```


12 months
24 months
36 months
TP
FP
1.00
0.75
0.50
0.25
0.00
AUC = 0.860
AUC = 0.869
AUC = 0.879