Survival Analysis

Welsneil

2025-03-17

```r
library(survival)
```

```
## Warning: package 'survival' was built under R version 4.4.3
```

```r
library(survminer)
```

```
## Loading required package: ggplot2

## Loading required package: ggpubr

##
## Attaching package: 'survminer'

## The following object is masked from 'package:survival':
##
##     myeloma
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
train_df <- read.csv("D:/Documents/IAI-UET/AI - Churn Prediction/dataset/train_survival.csv")
test_df <- read.csv("D:/Documents/IAI-UET/AI - Churn Prediction/dataset/test_survival.csv")
```

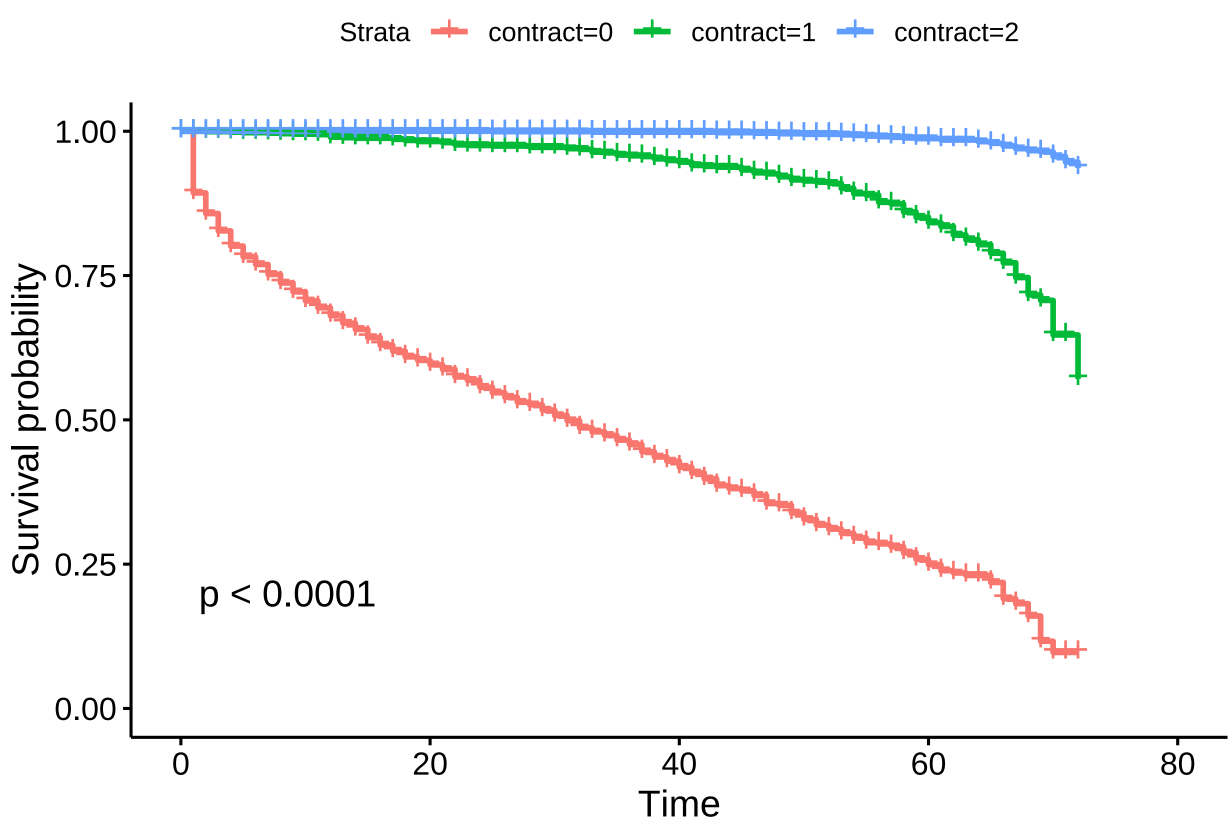
Including Plots

You can also embed plots, for example:

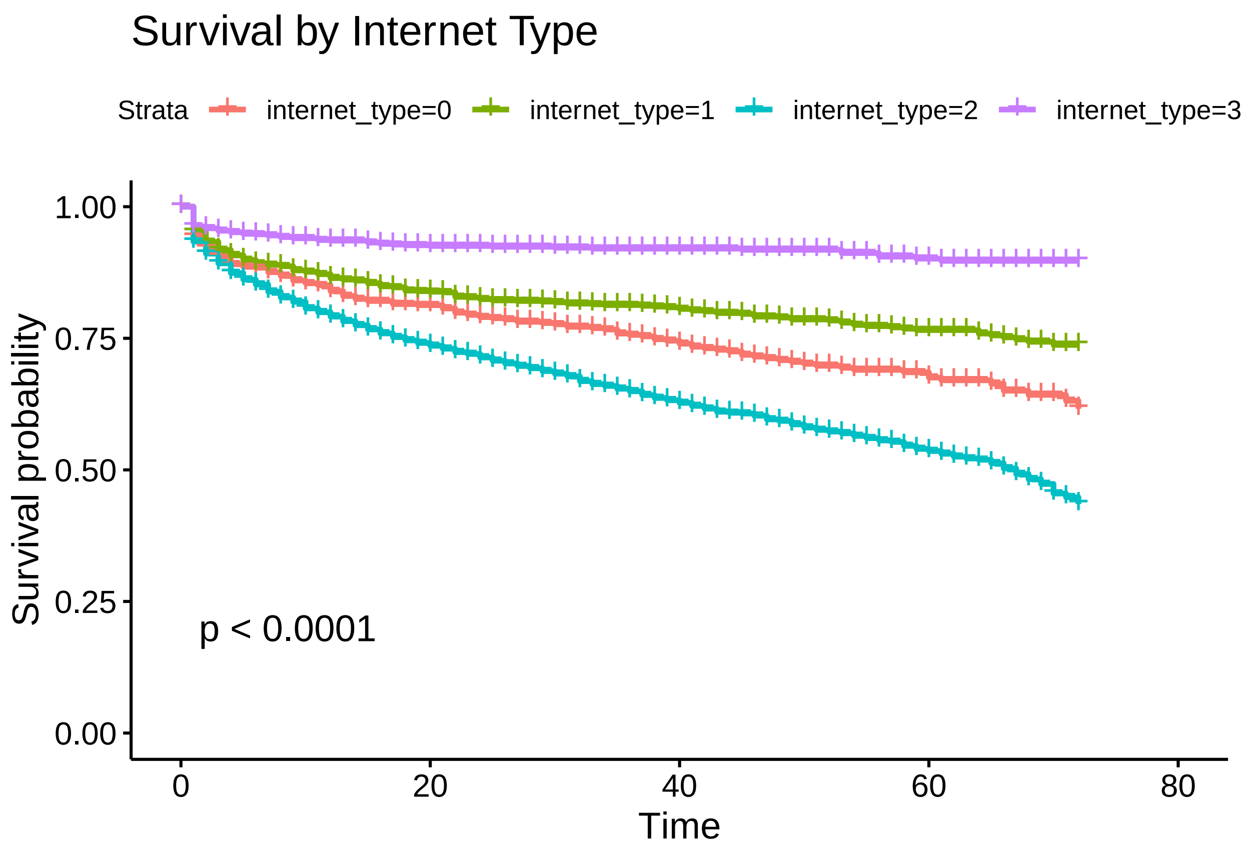
```r
train_df <- train_df %>%
  mutate(
    churn_value = as.numeric(churn_value),
    contract = as.factor(contract),
    internet_service = as.factor(internet_service),
    internet_type = as.factor(internet_type)
  ) %>% na.omit()

test_df <- test_df %>%
  mutate(
    churn_value = as.numeric(churn_value),
    contract = as.factor(contract),
    internet_service = as.factor(internet_service),
    internet_type = as.factor(internet_type)
  ) %>% na.omit()
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

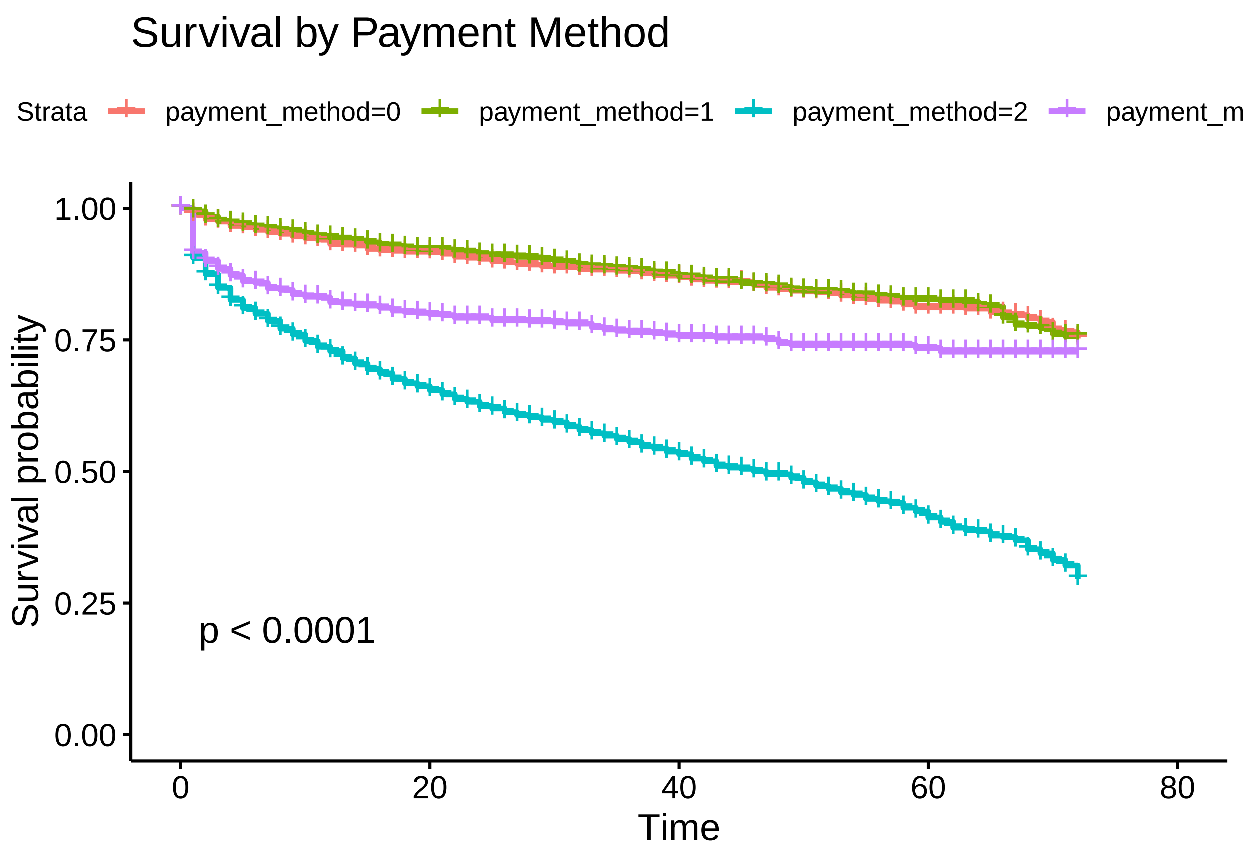
```r
surv_obj_train <- Surv(train_df$tenure, train_df$churn_value)


km_fit_contract <- survfit(surv_obj_train ~ contract, data=train_df)
ggsurvplot(km_fit_contract, pval=TRUE)
```



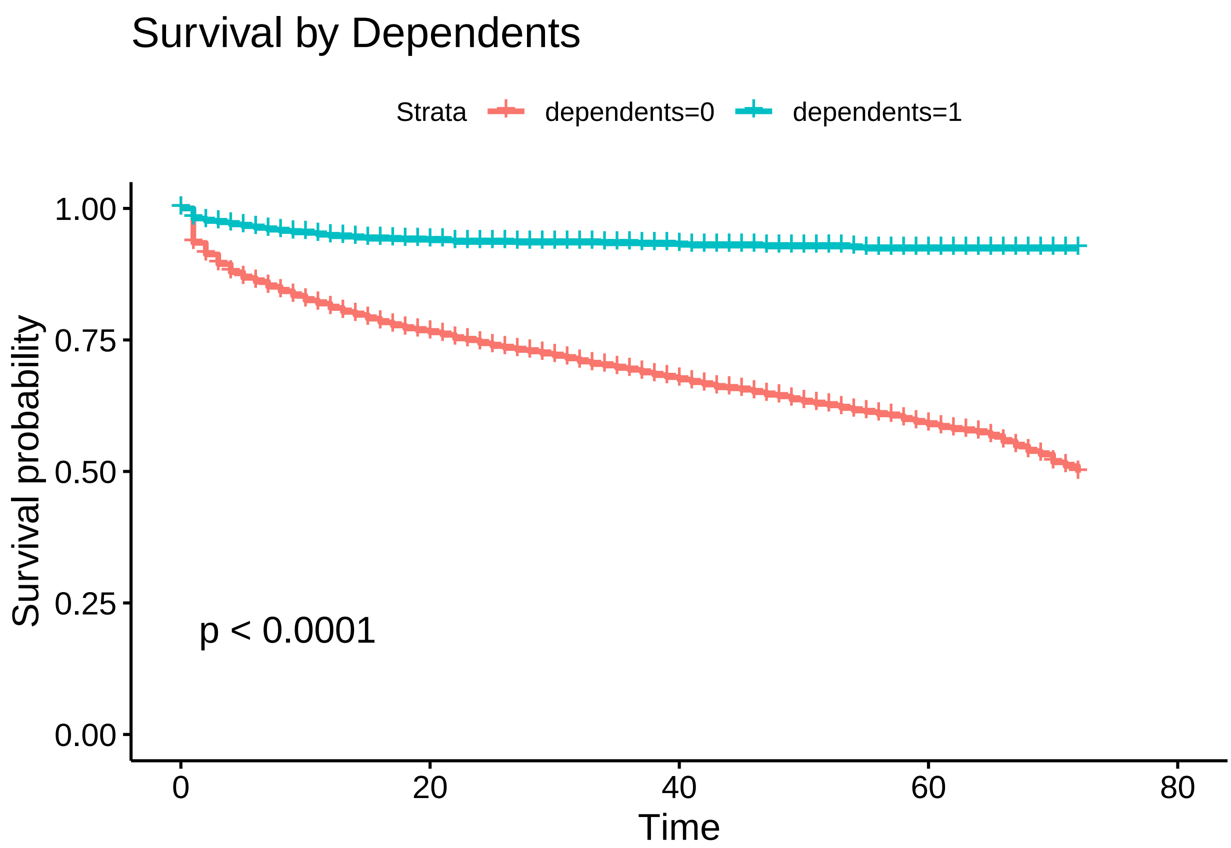
```r
km_fit_internet_type <- survfit(surv_obj_train ~ internet_type, data=train_df)
ggsurvplot(km_fit_internet_type, pval=TRUE, title="Survival by Internet Type")
```




```r
km_fit_payment_method <- survfit(surv_obj_train ~ payment_method, data=train_df)
ggsurvplot(km_fit_payment_method, pval=TRUE, title="Survival by Payment Method")
```




```r
km_fit_dependents <- survfit(surv_obj_train ~ dependents, data=train_df)
ggsurvplot(km_fit_dependents, pval=TRUE, title="Survival by Dependents")
```




```r
surv_obj_train <- Surv(time = train_df$tenure, event = train_df$churn_value)


cox_model <- coxph(Surv(tenure, churn_value) ~ contract + number_of_referrals + number_of_dependents +
                     monthly_charges + New_avg_service_fee + dependents + age + latitude + city +
                     internet_type + New_family_size_2 + total_charges + total_population +
                     payment_method + longitude + zip_code + New_family_size_3 +
                     New_contract_type_2 + avg_monthly_gb_download + senior_citizen,
                     # total_long_distance_charges + avg_monthly_long_distance_charges + offer + paperl
                   data = train_df)

summary(cox_model)
```

```
## Call:
## coxph(formula = Surv(tenure, churn_value) ~ contract + number_of_referrals +
##     number_of_dependents + monthly_charges + New_avg_service_fee +
##     dependents + age + latitude + city + internet_type + New_family_size_2 +
##     total_charges + total_population + payment_method + longitude +
##     zip_code + New_family_size_3 + New_contract_type_2 + avg_monthly_gb_download +
##     senior_citizen, data = train_df)
##
##   n= 5634, number of events= 1495
##
##                                 coef  exp(coef)   se(coef)       z Pr(>|z|)
## contract1                 -1.598e+00  2.023e-01  1.092e-01 -14.633  < 2e-16 ***
```

```
## contract2              -4.037e+00  1.764e-02  2.248e-01 -17.956  < 2e-16 ***
## number_of_referrals    -1.113e+00  3.287e-01  9.002e-02 -12.360  < 2e-16 ***
## number_of_dependents    2.105e-01  1.234e+00  1.062e-01   1.983  0.04741 *
## monthly_charges         1.557e+00  4.747e+00  1.013e-01  15.379  < 2e-16 ***
## New_avg_service_fee    -3.458e-01  7.076e-01  7.122e-02  -4.856 1.20e-06 ***
## dependents             -1.652e+00  1.916e-01  2.727e-01  -6.059 1.37e-09 ***
## age                     2.490e-02  1.025e+00  4.582e-02   0.544  0.58677
## latitude                1.283e-02  1.013e+00  8.597e-02   0.149  0.88140
## city                    2.184e-04  1.000e+00  8.699e-05   2.510  0.01206 *
## internet_type1         -1.109e-01  8.950e-01  1.017e-01  -1.091  0.27538
## internet_type2          1.462e-01  1.157e+00  1.034e-01   1.413  0.15756
## internet_type3         -1.352e+00  2.588e-01  1.898e-01  -7.120 1.08e-12 ***
## New_family_size_2True   7.389e-01  2.094e+00  7.208e-02  10.250  < 2e-16 ***
## total_charges          -3.738e+00  2.380e-02  1.042e-01 -35.883  < 2e-16 ***
## total_population        7.864e-02  1.082e+00  4.781e-02   1.645  0.10000 .
## payment_method          2.017e-01  1.223e+00  3.050e-02   6.614 3.75e-11 ***
## longitude               1.891e-02  1.019e+00  6.178e-02   0.306  0.75951
## zip_code               -5.760e-03  9.943e-01  5.928e-02  -0.097  0.92260
## New_family_size_3True   6.612e-01  1.937e+00  2.494e-01   2.651  0.00804 **
## New_contract_type_2True        NA         NA  0.000e+00      NA       NA
## avg_monthly_gb_download 1.895e-02  1.019e+00  4.616e-02   0.411  0.68138
## senior_citizen          2.656e-01  1.304e+00  9.587e-02   2.770  0.00561 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                         exp(coef) exp(-coef) lower .95 upper .95
## contract1                 0.20235     4.9420   0.16336   0.25063
## contract2                 0.01764    56.6778   0.01136   0.02741
## number_of_referrals       0.32869     3.0424   0.27552   0.39211
## number_of_dependents      1.23432     0.8102   1.00241   1.51989
## monthly_charges           4.74651     0.2107   3.89201   5.78862
## New_avg_service_fee       0.70765     1.4131   0.61546   0.81365
## dependents                0.19160     5.2192   0.11227   0.32697
## age                       1.02522     0.9754   0.93716   1.12155
## latitude                  1.01291     0.9873   0.85584   1.19880
## city                      1.00022     0.9998   1.00005   1.00039
## internet_type1            0.89503     1.1173   0.73332   1.09240
## internet_type2            1.15740     0.8640   0.94503   1.41749
## internet_type3            0.25881     3.8639   0.17839   0.37547
## New_family_size_2True     2.09354     0.4777   1.81772   2.41122
## total_charges             0.02380    42.0149   0.01941   0.02919
## total_population          1.08181     0.9244   0.98505   1.18808
## payment_method            1.22347     0.8173   1.15248   1.29883
## longitude                 1.01909     0.9813   0.90288   1.15027
## zip_code                  0.99426     1.0058   0.88519   1.11676
## New_family_size_3True     1.93705     0.5162   1.18798   3.15843
## New_contract_type_2True        NA         NA        NA        NA
## avg_monthly_gb_download   1.01913     0.9812   0.93098   1.11563
## senior_citizen            1.30416     0.7668   1.08075   1.57374
##
## Concordance= 0.94  (se = 0.002 )
## Likelihood ratio test= 5428  on 22 df,   p=<2e-16
## Wald test            = 2218  on 22 df,   p=<2e-16
## Score (logrank) test = 3962  on 22 df,   p=<2e-16
```

```r
base_surv <- basehaz(cox_model, centered = FALSE)


get_cumulative_hazard <- function(time_point){
  idx <- max(which(base_surv$time <= time_point))
  return(base_surv$hazard[idx])
}
```

```r
get_survival_probability <- function(tenure, hazard_score, base_surv, time_points) {

  cumulative_hazard <- sapply(time_points, function(t) get_cumulative_hazard(t + tenure))


  survival_probabilities <- exp(-cumulative_hazard)

  return(survival_probabilities)
}
```

```r
train_survival_features <- train_df %>%
  mutate(
    hazard_score = predict(cox_model, newdata = train_df, type = "lp"),
    baseline_hazard = sapply(tenure, get_cumulative_hazard),
    hazard_group = cut(hazard_score,
                       breaks=quantile(hazard_score, probs=seq(0, 1, 0.25)),
                       labels=c("Low", "Medium-Low", "Medium-High", "High"),
                       include.lowest=TRUE),

    survival_prob_3m = sapply(tenure, function(t) get_survival_probability(t, hazard_score, base_surv, 3
    survival_prob_6m = sapply(tenure, function(t) get_survival_probability(t, hazard_score, base_surv, 6
    survival_prob_12m = sapply(tenure, function(t) get_survival_probability(t, hazard_score, base_surv,
  ) %>%
  select(hazard_score, baseline_hazard, hazard_group, survival_prob_3m, survival_prob_6m, survival_prob_

# TEST survival features
test_survival_features <- test_df %>%
  mutate(
    hazard_score = predict(cox_model, newdata = test_df, type = "lp"),
    baseline_hazard = sapply(tenure, get_cumulative_hazard),
    hazard_group = cut(hazard_score,
                       breaks=quantile(train_survival_features$hazard_score, probs=seq(0, 1, 0.25)),
                       labels=c("Low", "Medium-Low", "Medium-High", "High"),
                       include.lowest=TRUE),
    survival_prob_3m = sapply(tenure, function(t) get_survival_probability(t, hazard_score, base_surv, 3
    survival_prob_6m = sapply(tenure, function(t) get_survival_probability(t, hazard_score, base_surv, 6
    survival_prob_12m = sapply(tenure, function(t) get_survival_probability(t, hazard_score, base_surv,
  ) %>%
  select(hazard_score, baseline_hazard, hazard_group, survival_prob_3m, survival_prob_6m, survival_prob_
```

```r
write.csv(train_survival_features, "survival_features_train.csv", row.names = FALSE)
write.csv(test_survival_features, "survival_features_test.csv", row.names = FALSE)
```

```r
write.csv(train_survival_features, "survival_features_train.csv", row.names = FALSE)
write.csv(test_survival_features, "survival_features_test.csv", row.names = FALSE)
```