Homework 3: Survival Analysis + CLV (R)

Yeva Stepanyan

2025-11-15

Libraries

```
library(survival)
library(flexsurv)
```

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

```
library(survminer)
```

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

## Loading required package: ggplot2

## Loading required package: ggpubr

##
## Attaching package: 'survminer'

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

```
library(ggplot2)
library(dplyr)
```

```
##
## 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
```

```
library(tidyr)
```

Loading the Dataset

```
df <- read.csv("telco.csv", stringsAsFactors = FALSE)

df$churn_flag <- ifelse(tolower(df$churn) == "yes", 1, 0)
df$tenure <- as.numeric(df$tenure)
```

Factoring the Categorical Variables

```
df$gender <- factor(df$gender)
df$ed <- factor(df$ed, levels = c("Did not complete high school","High school degree","Some college","C
df$custcat <- factor(df$custcat)
df$region <- factor(df$region)
```

Creating Survival Object

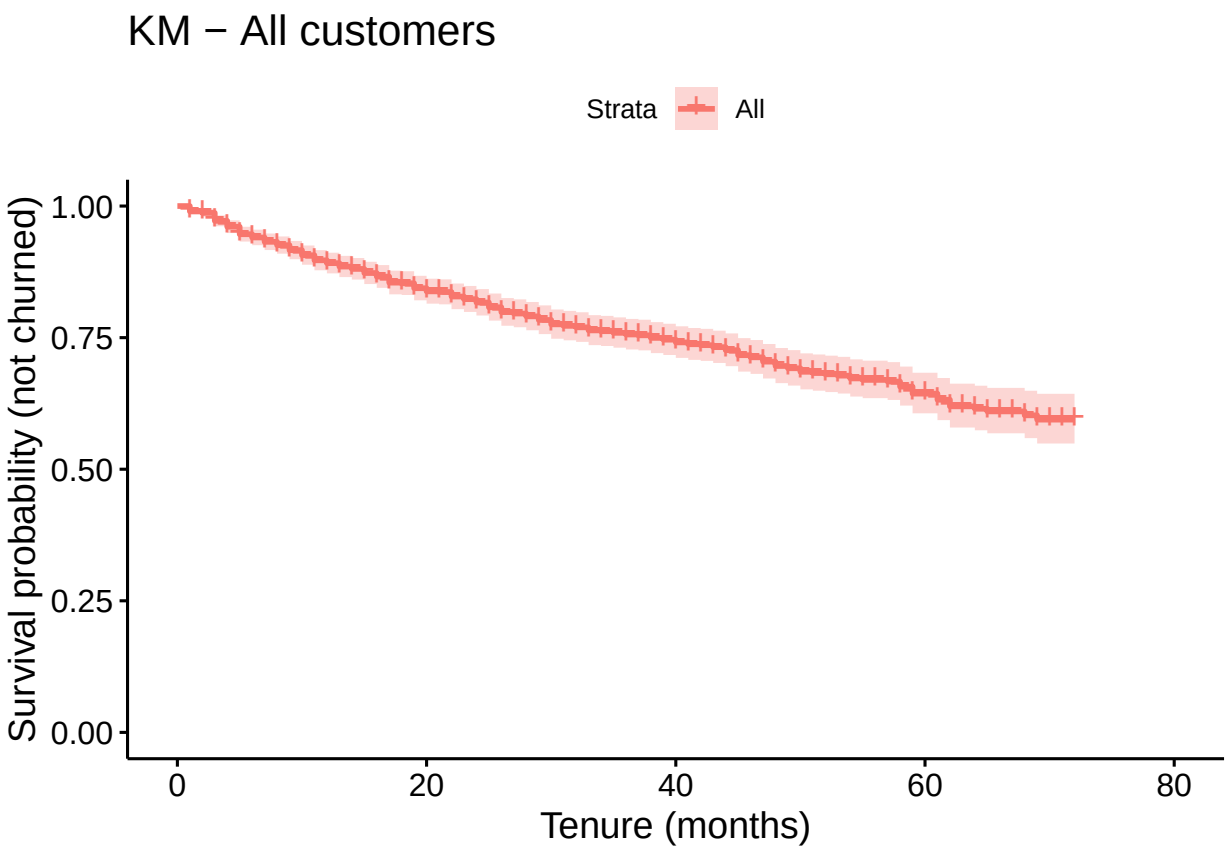
```
surv_obj <- Surv(time = df$tenure, event = df$churn_flag)
```

Kaplan-Meier

```
km <- survfit(surv_obj ~ 1, data = df)
km
```

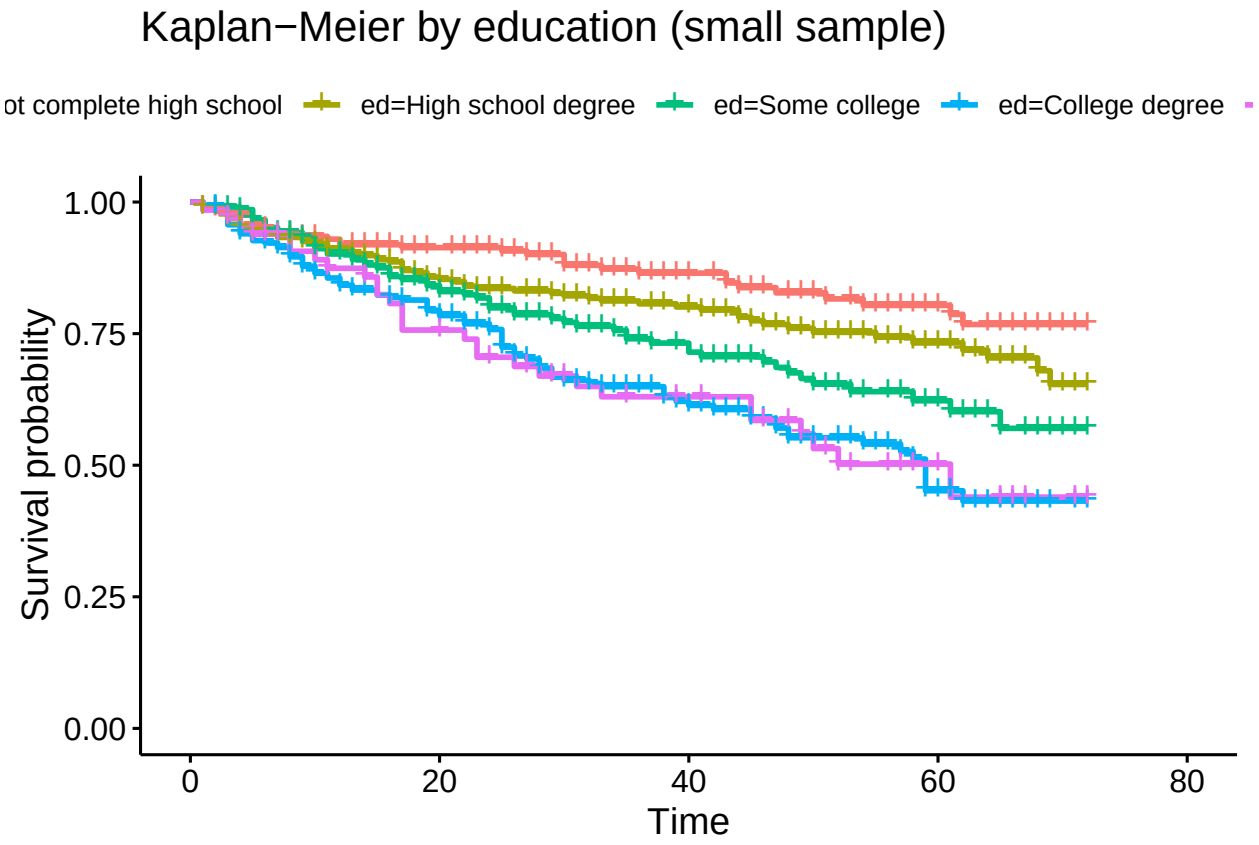
```
## Call: survfit(formula = surv_obj ~ 1, data = df)
##
##         n events median 0.95LCL 0.95UCL
## [1,] 1000    274     NA      NA      NA
```

```
ggsurvplot(km, data = df, xlab = "Tenure (months)", ylab = "Survival probability (not churned)", title =
```



KM by Education

```
if(nlevels(df$ed) > 1){
  ggsurvplot(survfit(surv_obj ~ ed, data = df), data = df,
             legend.title = "Education",
             title = "Kaplan-Meier by education (small sample)")
}
```



AFT Parametric Models

```
dists <- c("exponential","weibull","lognormal","loglogistic")  # survreg supports these names
models <- list()
for(dist in dists){
  formula <- as.formula("surv_obj ~ gender + ed + income")
  m <- try(survreg(formula, data = df, dist = dist), silent = TRUE)
  if(inherits(m, "try-error")){
    models[[dist]] <- NULL
  } else {
    models[[dist]] <- m
  }
}

flex_wb <- try(flexsurvreg(formula = surv_obj ~ gender + ed + income, data = df, dist = "weibull"), sil
if(!inherits(flex_wb, "try-error")) models[["flex_weibull"]] <- flex_wb
```

Model Comparison

```
aic_df <- data.frame(dist = character(0), AIC = numeric(0), stringsAsFactors = FALSE)
for(nm in names(models)){
  m <- models[[nm]]
```

```
  if(!is.null(m)){
    aic_df <- rbind(aic_df, data.frame(dist = nm, AIC = AIC(m)))
  }
}
aic_df <- aic_df %>% arrange(AIC)
aic_df
```

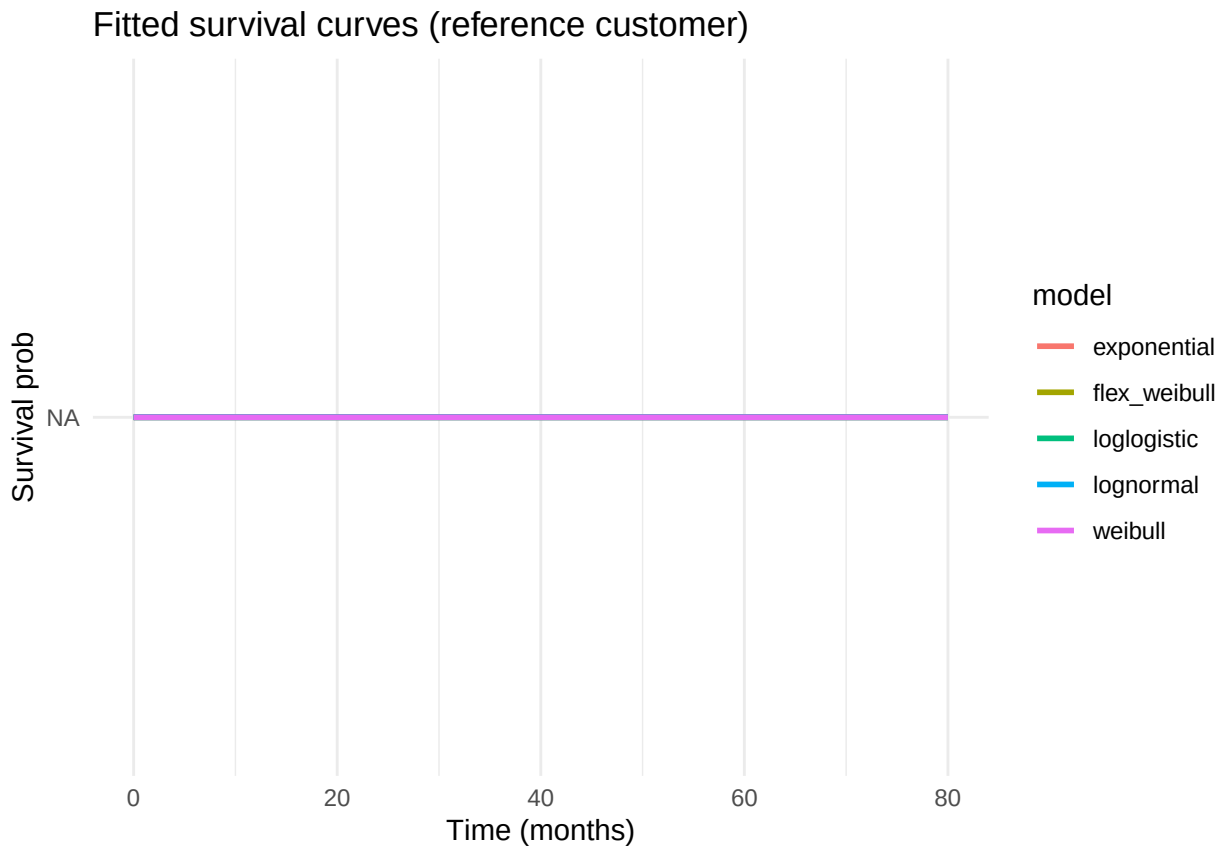
```
##           dist      AIC
## 1  exponential 3142.414
## 2 flex_weibull 3144.414
## 3      weibull 3144.414
## 4  loglogistic 3144.505
## 5    lognormal 3145.703
```

Visualizing Survival Curves

```
plot_df_all <- data.frame()
time_grid <- seq(0, 80, by = 1)

get_surv_probs <- function(fit_obj, newdata, times){
  if(inherits(fit_obj, "flexsurvflex")){
    s <- summary(fit_obj, newdata = newdata, t = times, type = "survival")
    probs <- s[[1]]$est
    return(probs)
  } else if(inherits(fit_obj, "survreg")){
    lp <- predict(fit_obj, newdata = newdata, type = "response")
    distname <- fit_obj$dist
    require(flexsurv)
    fs <- try(flexsurvreg(formula = update(fit_obj$call$formula, . ~ gender + ed + income),
                          data = eval(fit_obj$call$data),
                          dist = distname), silent = TRUE)
    if(!inherits(fs, "try-error")){
      s <- summary(fs, newdata = newdata, t = times, type = "survival")
      return(s[[1]]$est)
    } else {
      return(rep(NA, length(times)))
    }
  } else {
    return(rep(NA, length(times)))
  }
}

ref <- data.frame(gender = factor("Male", levels = levels(df$gender)),
                  ed = factor("High school degree", levels = levels(df$ed)),
                  income = median(df$income, na.rm = TRUE))

for(nm in names(models)){
  fit_obj <- models[[nm]]
  if(is.null(fit_obj)) next
  probs <- get_surv_probs(fit_obj, newdata = ref, times = time_grid)
  plot_df_all <- bind_rows(plot_df_all,
```

```
                          data.frame(time = time_grid, surv = probs, model = nm))
}

ggplot(plot_df_all, aes(x = time, y = surv, color = model)) +
  geom_line(size=1) +
  labs(title = "Fitted survival curves (reference customer)", x = "Time (months)", y = "Survival prob") +
  theme_minimal()
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```



Selecting the Final Model

```
if(nrow(aic_df) > 0){
  best <- aic_df$dist[1]
  cat("Selected final model by AIC:", best, "\n")
} else {
  best <- NA
  cat("No fitted models available.\n")
}
```

```
## Selected final model by AIC: exponential
```

```
final_model <- NULL
if(!is.na(best) && best %in% names(models)) final_model <- models[[best]]
final_model <- models[["flex_weibull"]]
```

Signifivant Features

```
if(!is.null(final_model) && inherits(final_model, "survreg")){
  s <- summary(final_model)
  coefs <- s$table
  sig_vars <- rownames(coefs)[which(coefs[,"p"] < 0.10)]
  cat("Significant parameter rows (p<0.10):\n")
  print(sig_vars)
}
```

CLV Using the Final Model

```
discount_annual <- 0.12
discount_month <- discount_annual / 12
T_max <- 60
acq_cost <- 0

get_S_for_customer <- function(customer_row, times){
  newd <- customer_row %>% select(gender, ed, income)
  newd$gender <- factor(newd$gender, levels = levels(df$gender))
  newd$ed <- factor(newd$ed, levels = levels(df$ed))
  if(inherits(final_model, "flexsurvflex")){
    s <- summary(final_model, newdata = newd, t = times, type = "survival")
    return(s[[1]]$est)
  } else if(inherits(final_model, "survreg")){
    distname <- final_model$dist
    fs <- try(flexsurvreg(formula = update(final_model$call$formula, . ~ gender + ed + income),
                          data = df,
                          dist = distname), silent = TRUE)
    if(!inherits(fs, "try-error")){
      s <- summary(fs, newdata = newd, t = times, type = "survival")
      return(s[[1]]$est)
    } else {
      return(rep(NA, length(times)))
    }
  } else {
    return(rep(NA, length(times)))
  }
}
```

CLV for Each Customer

```
clv_list <- vector("numeric", nrow(df))
times <- 1:T_max
for(i in seq_len(nrow(df))){
  cust <- df[i,]
  Svec <- get_S_for_customer(cust, times)
  monthly_margin <- 0.10 * (cust$income) / 12
  disc_f <- (1 + discount_month) ^ times
  clv_val <- sum(monthly_margin * Svec / disc_f, na.rm = TRUE) - acq_cost
  clv_list[i] <- clv_val
}
df$CLV_est <- clv_list

df %>% select(ID, region, tenure, income, ed, gender, churn_flag, CLV_est) %>%
  arrange(desc(CLV_est)) %>%
  head(10)
```

```
##    ID region tenure income                          ed gender churn_flag
## 1   1 Zone 2     13     64               College degree   Male          1
## 2   2 Zone 3     11    136    Post-undergraduate degree   Male          1
## 3   3 Zone 3     68    116 Did not complete high school Female          0
## 4   4 Zone 2     33     33           High school degree Female          1
## 5   5 Zone 2     23     30 Did not complete high school   Male          0
## 6   6 Zone 2     41     78           High school degree Female          0
## 7   7 Zone 3     45     19           High school degree Female          1
## 8   8 Zone 2     38     76           High school degree   Male          0
## 9   9 Zone 3     45    166               College degree   Male          0
## 10 10 Zone 1     68     72 Did not complete high school   Male          0
##    CLV_est
## 1        0
## 2        0
## 3        0
## 4        0
## 5        0
## 6        0
## 7        0
## 8        0
## 9        0
## 10       0
```

Outputs

Model AICs

```
aic_df
```

```
##           dist      AIC
## 1  exponential 3142.414
## 2 flex_weibull 3144.414
## 3      weibull 3144.414
## 4  loglogistic 3144.505
## 5    lognormal 3145.703
```

Sample CLV

```r
head(df %>% select(ID, CLV_est), 5)
```

```
##   ID CLV_est
## 1  1       0
## 2  2       0
## 3  3       0
## 4  4       0
## 5  5       0
```

```r
summary(df$tenure)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   17.00   34.00   35.53   54.00   72.00
```