BIOST 537: Homework 1

Department of Biostatistics @ University of Washington

Alejandro Hernandez

Winter Quarter 2025

1. (Adapted from Klein and Moeschberger, 2003) A large number of healthy individuals were enrolled in a study beginning January 1, 1970, and were followed for 31 years to assess the age at which they developed breast cancer. Individuals had clinical exams every 3 years after enrollment. For four selected individuals described below in (b) - (e), discuss the types of censoring present.

(a) Discuss the type of truncation present in the study.

The decision to sample healthy individuals excludes individuals who have already developed breast cancer. These truncated patients never make it into the study are never known to exist, which induces a kind of bias. This is an example of **left truncation** of individuals who generally have shorter time-to-cancer measures. Without adjusting for truncation, we risk overestimating time to breast cancer development.

(b) Individual A enrolled in the study at age 32 and never developed breast cancer during the study.

Individual A is **right-censored** because their event presumably occurred after their observation period. Their time-to-event is at least 63 years, denoted 63+.

(c) Individual B enrolled in the study at age 40 and was diagnosed with breast cancer at the fifth exam after enrollment.

Individual B is **interval censored**, their time-to-event is within 52-55 years.

(d) Individual C enrolled in the study at age 56 and died from heart failure at age 61, without ever being diagnosed with breast cancer. Post-mortem examinations confirmed that this patient never developed breast cancer.

Individual C is **right-censored** because their time-to-event is not fully known, due to loss from a competing event. Their measured time-to-event is at least 61 years.

(e) Individual D enrolled in the study at age 43 and moved away from the community at age 57, without ever being diagnosed with breast cancer by investigators.

Individual D is also **right-censored** because their time-to-event is not fully known due to loss of follow-up. Their measured time-to-event is at least 57 years.

(f) If instead of age at onset of breast cancer we were interested in studying the time from enrollment into the study until onset of breast cancer, would your answers above be any different? If so, how?

Changing the initiating event from birth to study enrollment would change how we measure time-to-event and shorten the values we analyze.

2. You wish to study the distribution of time from diagnosis of Crohn's disease, a chronic inflammatory condition of the intestinal tract, until first bowel resection surgery or death. Suppose that you have access to a national registry of all current inflammatory bowel disease patients. On a particular date (i.e., the day of study sampling), you enroll a random collection of Crohn's disease patients currently in the registry and who have not yet had bowel resection surgery. The date of diagnosis is available for all patients, and you prospectively follow all recruited patients for 10 years. Suppose that A represents age at diagnosis, and that T is the time from diagnosis until either bowel resection surgery or death, whichever occurs first.

(a) Describe the types of censoring and truncation affecting the data from this study.

Censored individuals have incomplete observation.

- Patients enrolled in the study who stop following-up with researchers before their surgery or death has been known to occur are right-censored, because their event presumably occurs after their observation period.
- A patient may enroll in the study and never follow-up until some point afterward, when it is made known to the researchers that the patient's terminating event has occurred. If the exact time-to-event is uncertain, this patient's observation is interval censored.
- If a patient is observed for a period *before* the situation described above occur, such that the exact time-to-event is still unknown, their observation is interval censored.

Truncated individuals are never known to exist.

- Individuals who experienced the initiating event (diagnosis of Crohn's) and termintaing events (first bowel surgery or death) all before the study began are excluded from the study. These left-truncated individuals tend to have experienced shorter time-to-event, so their exclusion from study data may overrepresent longer survival times.
- Right truncation is prone to occur in retrospective studies and cannot occur here. Participants of this study cannot be excluded because their terminating event has not yet occurred.
- Individuals who have not yet experienced the initiating event or whose experience is unknown to the registry are excluded from the study. Particularly to this study, individuals who have Crohn's but are not on the national registry of all current inflammatory bowel disease patients will not be represented in the study. We cannot generalize their survival times or the impact of their exclusion on survival estimates, but if these individuals share demographics, those communities may not inform study conclusions.

(b) Consider the following statement: in this study, while T is subject to truncation, A is not and hence the observed ages at diagnosis are not affected by selection bias. Do you agree with this statement? Briefly explain why.

Are the ages at diagnosis affected by selection bias? Age of diagnosis is available for a random sample of Crohn's disease patients currently in a national registry of living inflammatory bowel disease patients. As mentioned in a previous response, people whose Crohn's diagnosis are unknown to the registry are not in the study, which may bias observed ages if there the purpose of their absence is systemic.

3. Read Dickson et al. (1989) to learn about the Mayo PBC dataset, and answer the following questions about that study:

(a) What is the study population?

The study population consisted of 312 patients with primary biliary cirrhosis (PBC) who participated in two randomized, placebo-controlled clinical trials at the Mayo Clinic to evaluate the use of D-penicillamine. Patients were enrolled between January 1974 and May 1984.

(b) What is the initiating event (or time zero)?

Time zero is time of entry into trial.

(c) What is the terminating event?

Mortality from any cause was treated as the terminating event.

(d) What is the time scale?

The time scale is months.

(e) What are the causes of censoring?

Patients were considered censored if they (1) failed to follow-up or (2) underwent liver transplantation. Transplant patients were censored at the date of transplantation.

(f) For each of the causes above, comment on whether you believe the underlying censoring mechanism may be related to the outcome of interest?

Liver transplantation often occurs in patients with more advanced disease, which implies a higher risk of mortality without intervention–violating an assumption of non-informative censoring. Loss of follow-up is often due to external factors, which may be related to the outcome (e.g., a patient suffering from advanced PBC may be inclined to leave the study).

(g) Figure 3 in Dickson et al. (1989) presents survival curves for 106 cross-validation patients. Based on the Kaplan-Meier curves you see on this plot, for each of the three risk groups, approximately what proportion of individuals die within five years?

The approximate proportion of individuals who died within five years among the (i) low-risk group was 15%, (ii) medium-risk group was 50%, and (iii) high-risk group was 80%.

4. Investigators followed a cohort of 238 individuals with heroin use disorder who entered methadone maintenance programs in either of two clinics in Sydney, Australia over an 18-month period during the late 1980s. The purpose of the study was to identify factors associated to retention in methadone maintenance. Time from entry into the study to exit from methadone maintenance is the duration of interest. Details can be found in Caplehorn and Bell (1991). In this problem, you will perform a parametric analysis of data from this study. The dataset in question, called methadone.csv, can be found on the canvas website and loaded using the read.csv() function.

The variables in this dataset are:

- `id`: patient identification number;
- `clinic`: clinic identifier;
- `event`: indicator of exit from maintenance;
- `time`: observed follow-up time (in days);
- `dose`: maintenance methadone dose (in mg/day);
- `prison`: indicator of previous incarceration.

(a) Compute the average follow-up time and the proportion of censored observations.

The average follow-up time is 402.57 days and the proportion of censored individuals is 0.63.

(b) Using the available data, fit exponential, Weibull and generalized gamma models to the distribution of time to exit from maintenance. For each model, report parameter estimates, associated 95% confidence intervals and maximum log likelihood value. (NOTE: For the exponential model, report on parameter λ. For the Weibull model, report on parameters λ and p. For the generalized gamma model, report on parameters μ, σ and Q.)

From the 238 individuals enrolled in the aforementioned study, we fit three parametric models of their time from entry into the study to exit from methadone maintenance.

- The exponential model estimates $\hat{\lambda} = 0.0016$ (95% CI: 0.0013-0.0018) and offers a maximum log-likelihood of -1118.93.
- The Weibull model estimates $\hat{\lambda} = 0.0016$ (95% CI: 0.0013-0.0018) and $\hat{p} = 1.2264$ (95% CI: 1.0603-1.3925); its maximum log-likelihood of our observed data is -1114.92.
- The generalized gamma model estimates $\hat{\mu} = 6.5502$ (95% CI: 6.2694-6.8311), $\hat{\sigma} = 0.6595$ (95% CI: 0.3260-0.9930), $\hat{Q} = 1.4682$ (95% CI: 0.3848-2.5516), and a maximum log-likelihood of -1114.36.

(c) Plot the survival function corresponding to each of the above parametric fits as well as a nonparametric estimator of the survival function on the same graph, and comment visually on the adequacy of the models considered.

See Figure 1. Figure 1 shows the three parametric survival estimators as compared to the nonparametric Kaplan-Meier estimator. Although the parametric models are roughly similar, the Weibull and generalized gamma are better at modeling the data than the exponential model. Considering these survival curves, we may find the Weibull model to be an adequate simplification of the generalized gamma.

(d) Is the Weibull model an appropriate simplification of the generalized gamma model in this example? Justify your answer by performing an appropriate statistical test.

A likelihood ratio test determined there is sufficient evidence in our sample the Weibull model is an adequate simplification of the generalized gamma model at $\alpha = 5\%$ (p=0.29).

(e) Using a Weibull model, provide an estimate and 95% confidence interval of:

i. the median time until exit from maintenance;

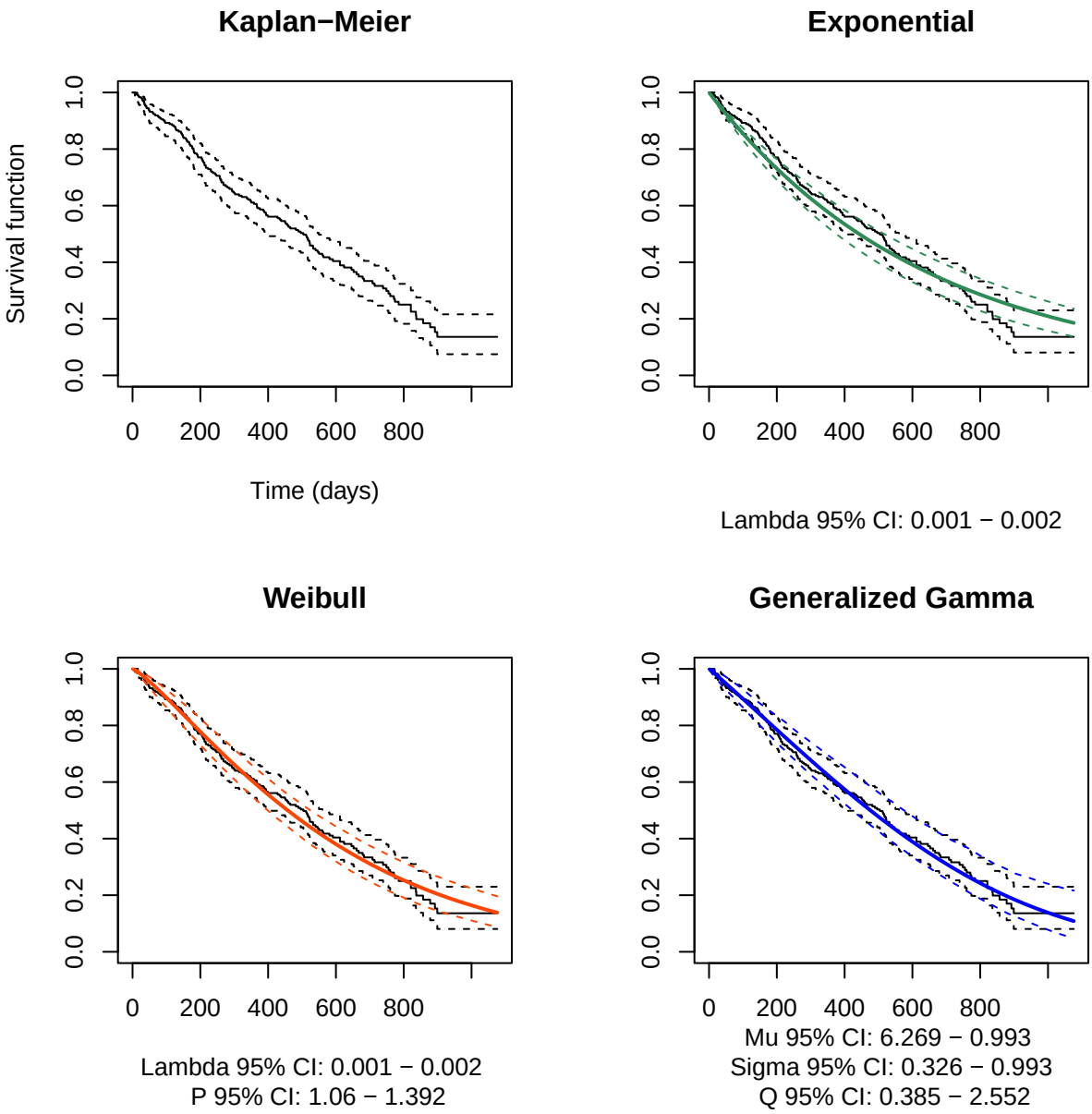

Figure 1: Survival estimators and corresponding parameter estimates

ii. the probability that no exit will occur by one year or, in other words, the probability an individual will not have exited by one year.

iii. the probability that no exit will occur by two years given that no exit has occurred by one year.

From the 238 individuals enrolled in the aforementioned study, we fit a parametric model of their time from entry into the study to exit from methadone maintenance, using the Weibull distribution. From this model, we estimate

- the median time until exit from maintenance is 458 days, approximately 1.3 years (95% CI: 397-518 days);
- the probability an individual will not have exited by one year is 0.59 (95% CI: 0.54-0.65);
- the probability an individual who has remained in the study at least a year will not have exited by two years is 0.49 (95% CI: 0.42-0.57).

(f) Is the exponential model an appropriate simplification of the Weibull model in this example? Justify your answer by performing an appropriate statistical test.

A likelihood ratio test determined there is insufficient evidence in our sample that the exponential model is an adequate simplification of the Weibull model at $\alpha = 5\%$ (p=0.0046).

(g) Separately fit an exponential model to the subset of individuals in clinic 1 and clinic 2. Report parameter estimates and corresponding 95% confidence intervals. Use the output of these two fits to determine whether the distribution of time to exit from maintenance differs significantly by clinic. Justify your answer by performing an appropriate statistical test.

From the 238 individuals enrolled in the aforementioned study, we fit two parametric models of time from entry into the study to exit from methadone maintenance: one for patients from clinic 1 and another for patients from clinic 2.

- The exponential survival model of clinic 1 patients estimates $\hat{\lambda} = 0.00205$ (95% CI: 0.00168-0.0024).
- The exponential survival model of clinic 2 patients estimates $\hat{\lambda} = 0.00077$ (95% CI: 0.00049-0.0011).

The λ parameter characterizes the survival function. We can tell from the non-overlapping confidence intervals above that the sample suggests these two groups have different survival rates, with clinic 1 patients having significantly longer times between study enrollment and exit. Figure 2 visualizes this difference.

A log-rank test determined there is sufficient evidence in our sample that time from entry into the study to exit from methadone maintenance differs significantly between clinics 1 and 2 at $\alpha = 5\%$ ($p < 0.001$).

(h) Repeat the last problem but substituting clinic by history of incarceration (i.e., prison)

From the 238 individuals enrolled in the aforementioned study, we fit two parametric models of time from entry into the study to exit from methadone maintenance: one for patients from with a history of incarceration and another for patients without a history of incarceration.

- The exponential model of patients with incarceration history estimates $\hat{\lambda} = 0.0015$ (95% CI: 0.0011-0.0018).
- The exponential model of patients without incarceration history estimates $\hat{\lambda} = 0.0017$ (95% CI: 0.0013-0.0021).

The overlapping confidence intervals above suggests these two groups may not have different survival rates; figure 3 shows they are visually similar.

A log-rank test determined there is insufficient evidence in our sample that time from entry into the study to exit from methadone maintenance differs significantly between patients with and without a history of incarceration at $\alpha = 5\%$ (p=0.26).

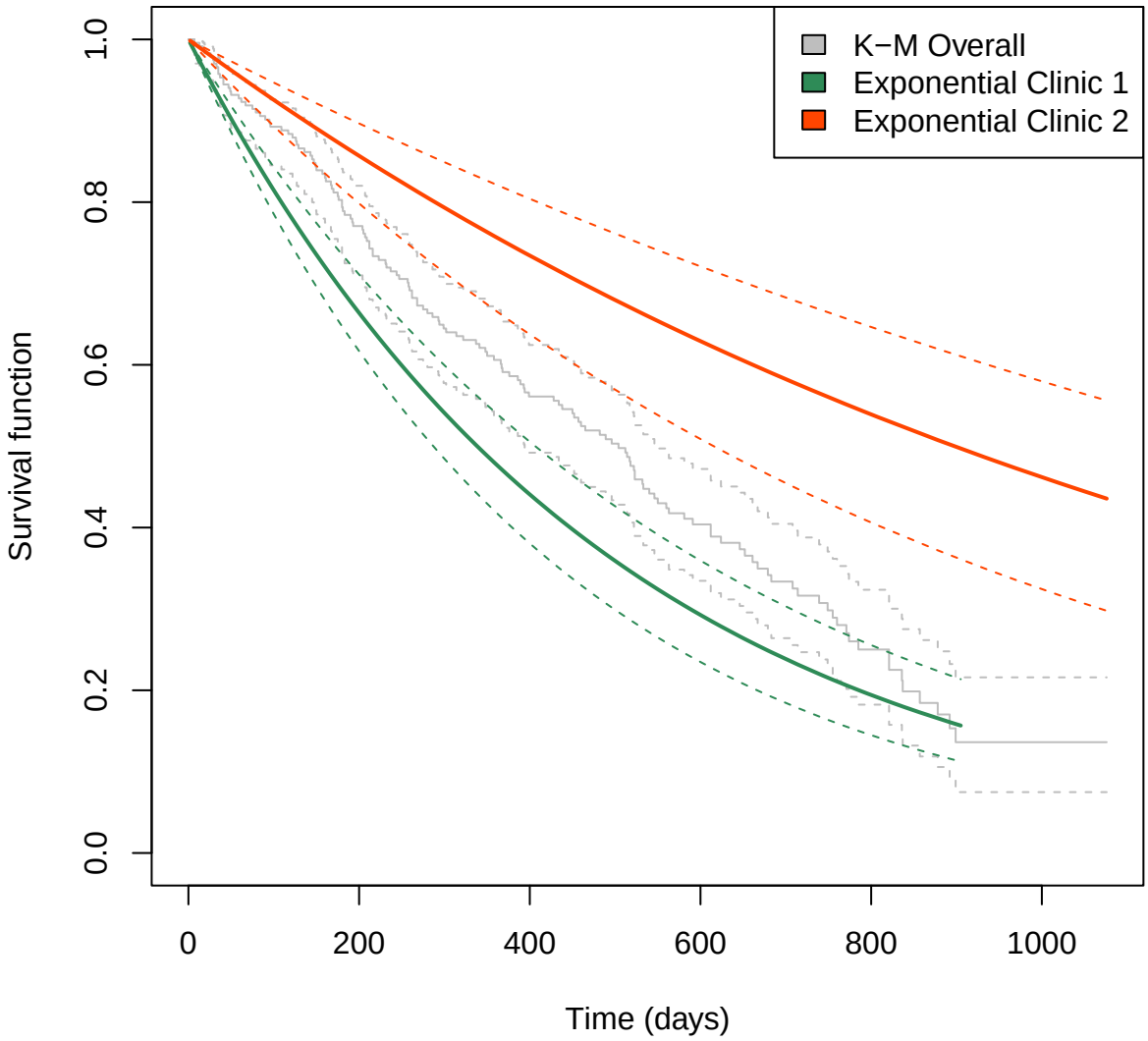

Figure 2: Time from entry to exit from methadone maintenence. Survival is modeled overall by a Kaplan-Meier estimator and by clinic with exponential models.

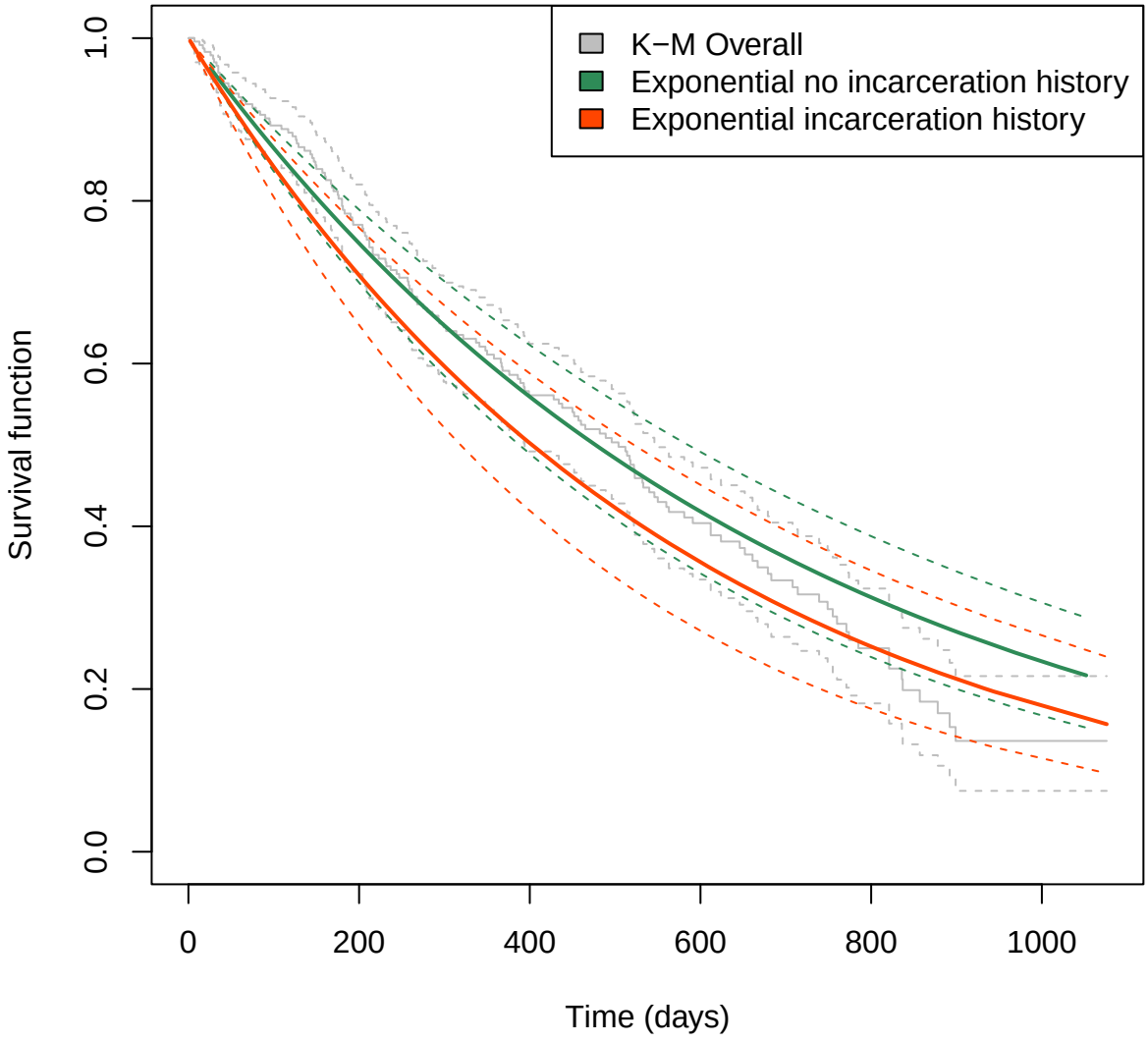

Figure 3: Time from entry to exit from methadone maintenence. Survival is modeled overall by a Kaplan-Meier estimator and by history of incarceration with exponential models.

Code Appendix

```
# Clear environment
rm(list=ls())

# Setup options
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE, results='hide')
options(knitr.kable.NA = '-', digits = 2)
labs = knitr::all_labels()
labs = labs[!labs %in% c("setup", "allcode")]
# Load relevant packages
library(survival)   # survival models
library(flexsurv)   # survival models
library(dplyr)      # data manipulation

# Load data
methadone <- read.csv("../data/methadone.csv")

anyNA(methadone)  # No missing data
head(methadone)

#### --- QUESTION 4 --- ####

## (a) Compute the average follow-up time and the proportion of censored observations.

# Summary statistics of follow-up time
with(methadone, summary(time))

# Summary of censored observations
with(methadone, mean(event))  # 63% event observed
with(methadone, sum(event))   # 150 observed events

## (b) Using the available data, fit exponential, Weibull, and generalized gamma models to the distribu

# Import function to fit models
source("../R/fitparametric.R")

# Encode survival times
surv_times <- with(methadone, survival::Surv(time, event))

# Fit parametric survival models
fitexp <- fitparametric(survobj = surv_times, dist = "exp")
fitweibull <- fitparametric(survobj = surv_times, dist = "weibull")
fitgengamma <- fitparametric(survobj = surv_times, dist = "gengamma")

# Table of parameter estimates
rbind(fitexp$coeff, fitweibull$coeff, fitgengamma$coeff)[,-4]

# Maximum log-likelihood
rbind(fitexp$loglik, fitweibull$loglik, fitgengamma$loglik)

## (c) Plot nonparametric and parametric estimators of the survival function
```

```r
# Table of coefficients (for easy access of parameter estimates in later plotting)
coef_tab <- rbind(fitexp$coeff,
                  fitweibull$coeff,
                  fitgengamma$coeff)[,-4] %>% round(digits = 3)

par(mfrow = c(2,2))
# Plot Kaplan-Meier nonparametric survival function
survival::survfit(surv_times ~ 1, conf.type = "log-log") %>%
  plot(main = "Kaplan-Meier",
       xlab = "Time (days)",
       ylab = "Survival function")
# Plot exponential survival function
plot(fitexp$fit,
     col = "seagreen",
     main = "Exponential",
     sub = paste("Lambda 95% CI:", coef_tab[1,2], "-", coef_tab[1,3]))
# Plot Weibull survival function
plot(fitweibull$fit,
     col = "orangered",
     main = "Weibull",
     sub = paste("Lambda 95% CI:", coef_tab[2,2], "-", coef_tab[2,3],
                 "\n P 95% CI:", coef_tab[3,2], "-", coef_tab[3,3]))
# Plot generalized gamma survival function
plot(fitgengamma$fit,
     col = "blue",
     main = "Generalized Gamma",
     sub = paste("Mu 95% CI:", coef_tab[4,2], "-", coef_tab[5,3],
                 "\n Sigma 95% CI:", coef_tab[5,2], "-", coef_tab[5,3],
                 "\n Q 95% CI:", coef_tab[6,2], "-", coef_tab[6,3]))

## (d) Test if the Weibull model is an appropriate simplification of the generalized gamma model

## Conduct a likelihood ratio test of the two models
# Calculate the test statistic
lrtstat <- -2*(fitweibull$loglik - fitgengamma$loglik)
lrtpval <- 1 - pchisq(q = lrtstat, df = 1)  # 1 df between Weibull and gen. gamma
# "If the p is low, the null must go!"
lrtpval < 0.05  # Sufficient evidence Weibull is adequate simplification of generalized gamma

## (e) Estimate quantities using the Weibull model

# restricted mean survival time
fitparametric(surv_times, dist = "weibull", feature = "mean",
              t = max(methadone$time))

# median survival time
fitparametric(surv_times, dist = "weibull", feature = "quantile")

# probability of surviving to one year
fitparametric(surv_times, dist = "weibull", feature = "survival",
              t = 365)

# probability of surviving to two years, given surviving to one year
```

```r
fitparametric(surv_times, dist = "weibull", feature = "condsurvival",
              t0 = 365, t = 2*365)

## (f) Test if the exponential model is an appropriate simplification of the Weibull model

## Conduct a likelihood ratio test of the two models
# Calculate the test statistic
lrtstat2 <- -2*(fitexp$loglik - fitweibull$loglik)
lrtpval2 <- 1 - pchisq(q = lrtstat2, df = 1)  # 1 df between exponential and Weibull
# "If the p is low, the null must go!"
lrtpval2 < 0.05  # Insufficient evidence exponential is adequate simplification of Weibull

## (g) Fit a separate exponential model, stratified by clinic

fitexp_clinic1 <- methadone %>%
  subset(clinic == 1) %>%
  with(., Surv(time, event)) %>%
  fitparametric(., dist = "exp")

fitexp_clinic2 <- methadone %>%
  subset(clinic == 2) %>%
  with(., Surv(time, event)) %>%
  fitparametric(., dist = "exp")

# Output model coefficients
rbind(fitexp_clinic1$coeff, fitexp_clinic2$coeff)

## Conduct a log-rank test of the two groups
logrank_pval <- survival::survdiff(surv_times ~ clinic, methadone)$pvalue
# "If the p is low, the null must go!"
logrank_pval < 0.05  # Sufficient evidence survival for clinics 1 and 2 differ

## (h) Fit a separate exponential model, stratified by incarceration history

fitexp_inc_without <- methadone %>%
  subset(prison == 0) %>%
  with(., Surv(time, event)) %>%
  fitparametric(., dist = "exp")

fitexp_inc_with <- methadone %>%
  subset(prison == 1) %>%
  with(., Surv(time, event)) %>%
  fitparametric(., dist = "exp")

# Output model coefficients
rbind(fitexp_inc_without$coeff, fitexp_inc_with$coeff)

## Conduct a log-rank test of the two groups
logrank_pval2 <- survdiff(surv_times ~ prison, methadone)$pvalue
# "If the p is low, the null must go!"
logrank_pval2 < 0.05  # Insufficient evidence survival differs between
                      # those with and without incarceration
```

```r
## Plot survival curves from question (4g)
## Plot survival curves
# Plot Kaplan-Meier estimator for all clinics
survfit(surv_times ~ 1, conf.type = "log-log") %>%
  plot(col = "grey",
       xlab = "Time (days)",
       ylab = "Survival function")

# Plot exponential estimators for clinics 1 and 2
lines(fitexp_clinic1$fit, col = "seagreen", main = "Clinic 1")
lines(fitexp_clinic2$fit, col = "orangered", main = "Clinic 2")
legend(x = "topright",
      legend = c("K-M Overall", "Exponential Clinic 1", "Exponential Clinic 2"),
       fill = c("grey", "seagreen","orangered"))
## Plot survival curves from question (4h)
# Plot Kaplan-Meier estimator for everyone
survfit(surv_times ~ 1, conf.type = "log-log") %>%
  plot(col = "grey",
       xlab = "Time (days)",
       ylab = "Survival function")

# Plot exponential estimators for patients with and without incarceration history
lines(fitexp_inc_without$fit, col = "seagreen", main = "No incarceration history")
lines(fitexp_inc_with$fit, col = "orangered", main = "Incarceration history")
legend(x = "topright",
      legend = c("K-M Overall",
                 "Exponential no incarceration history",
                 "Exponential incarceration history"),
       fill = c("grey", "seagreen","orangered"))
```

End of document.