Assignment 2

Jingwen GAO

Use the kidney data (survival package). Demonstrating some of the expository plotting methods and ggplot (and probably some dplyr) tools, generate a plot to describe and summarize the relationship between frailty (x) and time to infection by sex of patient.

–Note that sex is an integer variable so you'll need to use as.factor() or convert it to a factor, with labels for the levels.

–The x-y relationships should be optimally smoothed

–The plot should be labelled and annotated to "tell the story" of the relationship in each group.

–You'll need to consult the original Biometrics paper (or some authoritative source) about the frailty measure (its units, scale, etc) to provide a good axis label.

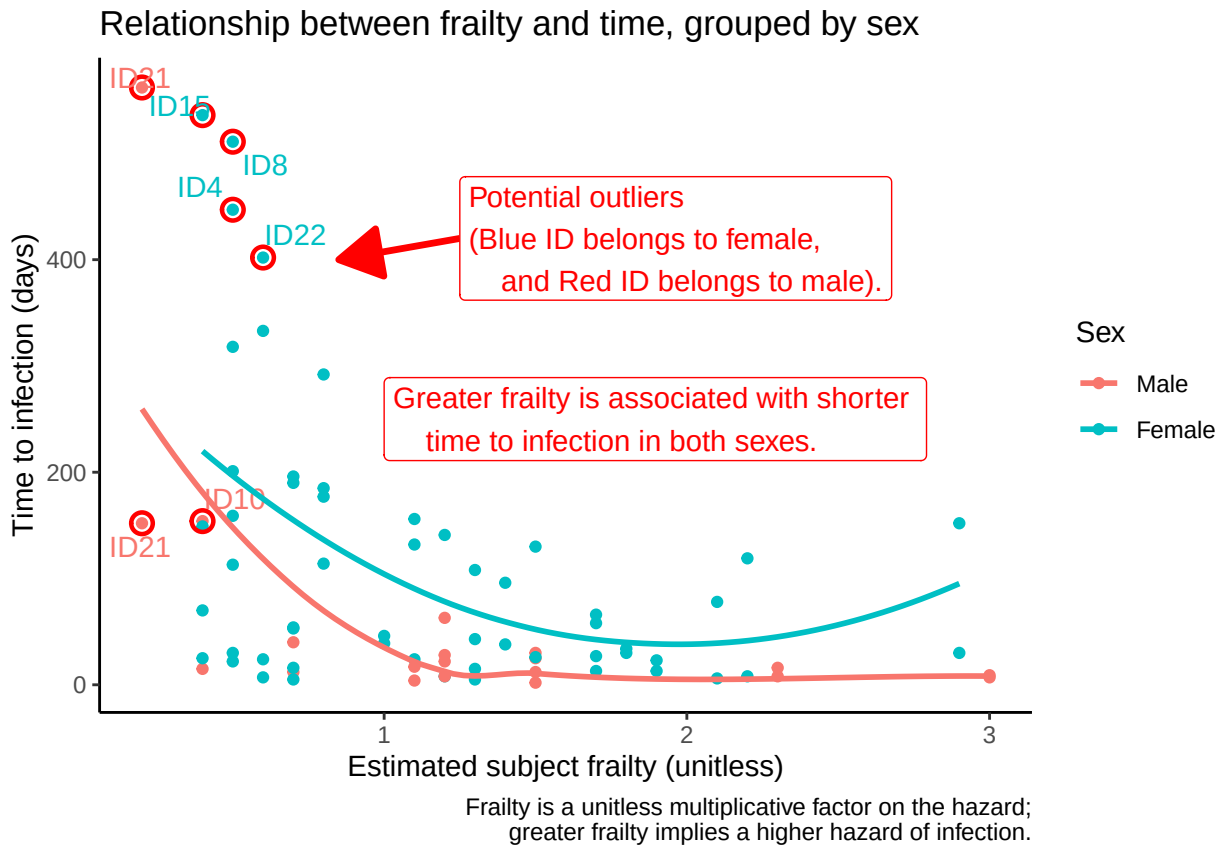
–Put brief summary text in a box on the plot.

–Unusual data point(s) can be annotated as you think is needed for a reader to appreciate the overall relationship (or that point's influence).

```
library(tidyr)
library(dplyr)
library(scales)
library(ggplot2)
library(survival)
library(ggrepel)
```

```
# find outliers by boxplot
g <- ggplot(kidney, aes(x = factor(sex), y = time)) + geom_boxplot()
built <- ggplot_build(g)$data[[1]]

#locate the outlier points
v_time <- unlist(built$outliers, use.names = FALSE)
v_sex  <- rep(built$x, sapply(built$outliers, length))
outs_tbl <- data.frame(sex = v_sex, time = v_time)

kidney_key <- paste(kidney$sex, kidney$time)
outs_key   <- paste(outs_tbl$sex, outs_tbl$time)
outliers   <- kidney[kidney_key %in% outs_key, ]

ggplot(kidney,
       aes(x = frail, y = time,
           colour = factor(sex, levels = c(1, 2), labels = c("Male", "Female")))) +
  geom_point() +
  geom_point(data = outliers,
             colour = "red", size = 3, shape = 1, stroke = 1.2,
             show.legend = FALSE) +
  geom_text_repel(data = outliers,
```

```
                aes(label = paste0("ID", id)),label.size=0.2,show.legend = F) +
geom_smooth(method = "loess", se = FALSE, span = 1) +
labs(title  = "Relationship between frailty and time, grouped by sex",
     x      = "Estimated subject frailty (unitless)",
     y      = "Time to infection (days)",
     color  = "Sex",
     caption= "Frailty is a unitless multiplicative factor on the hazard;
     greater frailty implies a higher hazard of infection.") +
theme_classic()+
annotate(
  geom = "label", x = 1, y = 250,

  label = "Greater frailty is associated with shorter
  time to infection in both sexes.",size=4,
  hjust = "left", color = "red"
)+
annotate(
  geom = "segment",
  x = 1.25, y = 420, xend = 0.85, yend = 400, color = "red",linewidth=1.2,
  arrow = arrow(type = "closed")
)+
annotate(
  geom = "label", x = 1.25, y = 420,

  label = "Potential outliers\n(Blue ID belongs to female,
  and Red ID belongs to male).",
  hjust = "left", color = "red"
)
```


Relationship between frailty and time, grouped by sex
ID31
ID15
ID8
ID4
ID22
ID10
ID21
Potential outliers
(Blue ID belongs to female,
and Red ID belongs to male).
Greater frailty is associated with shorter
time to infection in both sexes.
Time to infection (days)
400
200
0
1
2
3
Estimated subject frailty (unitless)
Frailty is a unitless multiplicative factor on the hazard;
greater frailty implies a higher hazard of infection.
Sex
Male
Female