Assignment3

Jingwen GAO

2025-09-27

```
library(survival)
library(dplyr)
```

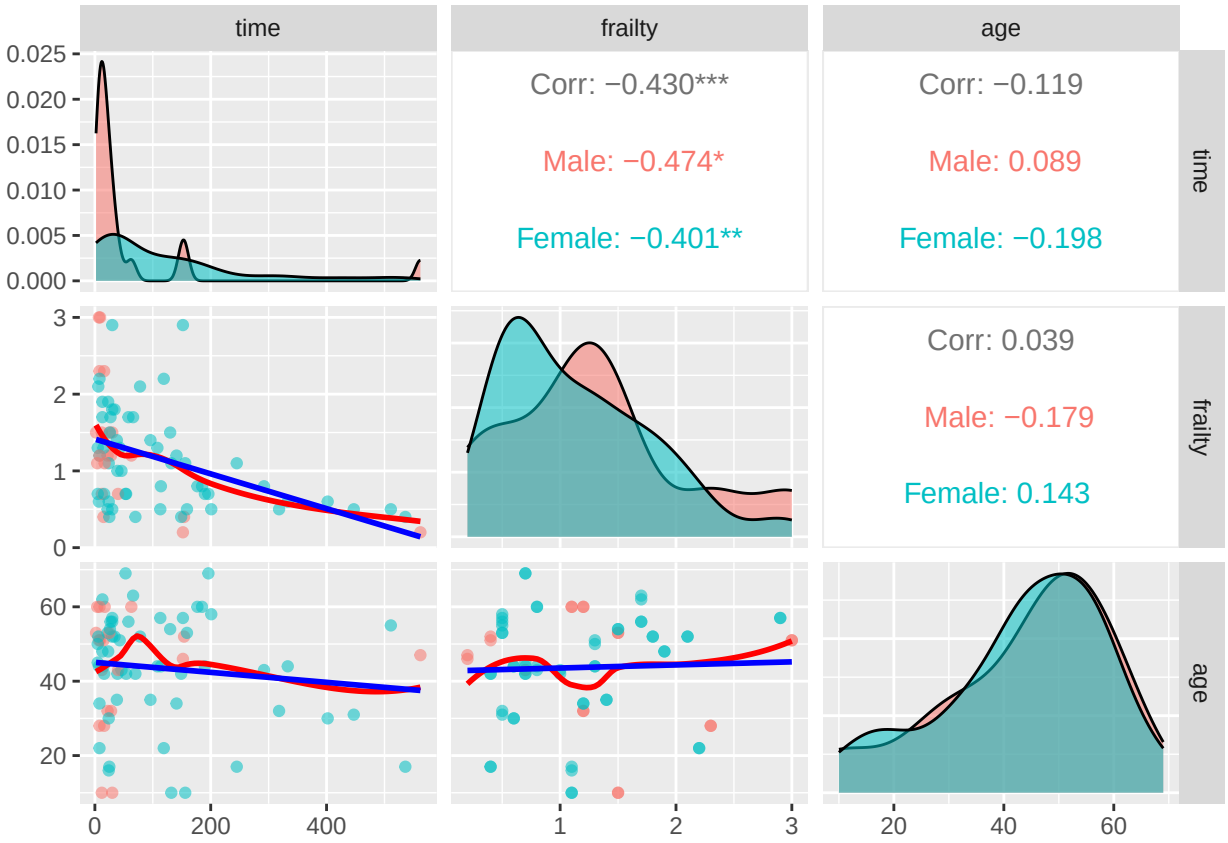
scatterplot matrix

To begin our analysis, we can generate a scatterplot matrix to have a general overview of the relationships of thes variables.

```
library(GGally)
new <- survival::kidney %>%
  transmute(
    time,
    frailty = frail,
    age,
    sex = factor(sex, levels = c(1,2), labels = c("Male","Female"))
  )

my_fn <- function(data, mapping, ...){
  p <- ggplot(data = data, mapping = mapping) +
    geom_point() +
    geom_smooth(method=loess, se=F, fill="red", color="red", ...) +
    geom_smooth(method=lm, se=F, fill="blue", color="blue", ...)
  p
}

GGally::ggpairs(
  new,
  columns = c("time", "frailty", "age"),
  aes(color = sex, alpha = 0.7),
  lower = list(continuous = my_fn)
)
```



By viewing the correlations, we observe that

- Time vs. Frailty: The correlations within male (-0.474) and female (-0.401) groups are similar, suggesting that sex does not strongly affect this relationship. The overall correlation (-0.430) is significantly negative, indicating that higher frailty is associated with shorter recurrence time, consistent with the definition of frailty as a multiplicative factor of infection risk.
- Time vs. Age: The correlation is close to zero (0.119), implying no significant linear association. Thus, age does not appear to be a primary determinant of infection time.
- Frailty vs. Age: The correlation is near zero (0.039), suggesting that frailty is largely independent of age and provides predictive information itself.
- Distribution of Time: The distribution is highly right-skewed with heavy tails, violating normality assumptions. A log-transformation may be required to stabilize variance and improve model fit.

time vs sex

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

kidney1 <- survival::kidney |>
  mutate(sex = factor(sex, levels = c(1, 2), labels = c("Male", "Female"))) |>
  filter(!is.na(time))
```

1. numeric summaries by sex

```
library(moments)
sum_tbl <- kidney1 |>
  group_by(sex) |>
  summarise(
    n = n(),
    mean = mean(time),
    median = median(time),
    sd = sd(time),
    mad = mad(time),
    IQR = IQR(time),
    p25 = quantile(time, 0.25),
    p75 = quantile(time, 0.75),
    min = min(time),
    max = max(time),
    skewness = skewness(time),
    kurtosis = kurtosis(time)
  )
sum_tbl
```

```
## # A tibble: 2 x 13
##   sex        n  mean median    sd   mad   IQR   p25   p75   min   max skewness
##   <fct>  <int> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>
## 1 Male      20  59.3   16.5  126.  13.3  23.8  8.75  32.5     2   562     3.43
## 2 Female    56 117.    62    130.  76.4 132   24.8  157.      5   536     1.65
## # i 1 more variable: kurtosis <dbl>
```

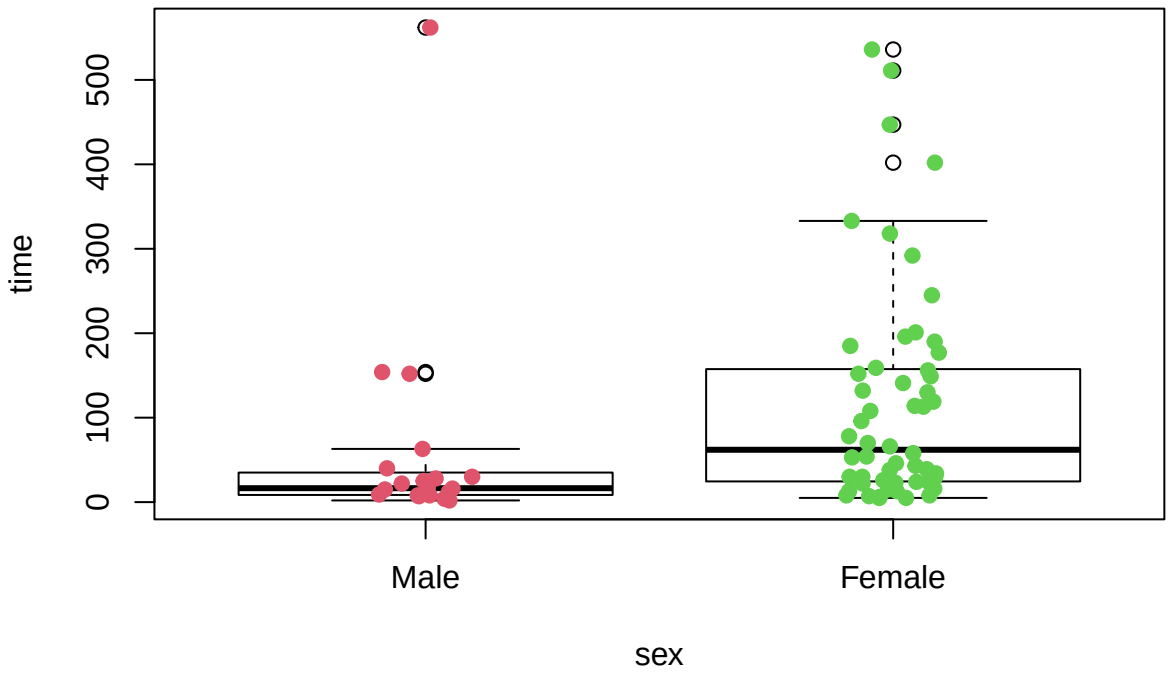
From the numeric summarise, we know that

- The male group is much smaller, so results are more sensitive to outliers;
- Both groups have means much larger than their medians, indicating that the distributions are right-skewed, with a few patients having very long infection times. The male group shows this pattern even more strongly.
- The standard deviation is large for both groups, meaning that both genders have high variability in infection time. The male group is more clustered than the female group, as seen from the MAD column.
- The skewness values confirm the right-skew (3.43 for males vs. 1.65 for females). The male group is therefore more asymmetric.
- The kurtosis values (14.05 for males vs. 5.19 for females) indicate that the male distribution is much more heavy-tailed, which is consistent with the presence of extreme long survivors and may also reflect the small sample size.

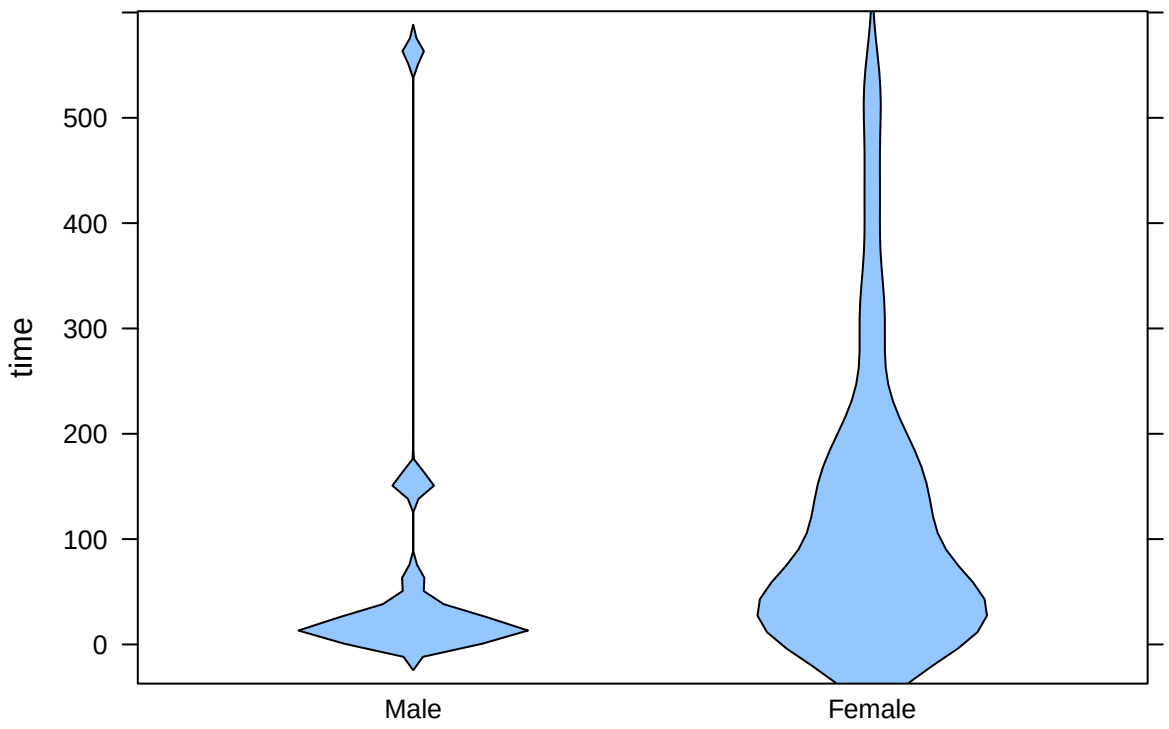
We can combine these statistics to boxplots and violin plots.

```
library(lattice)
boxplot(time ~ sex, data = kidney1, col = "white")
stripchart(time ~ sex,
           data = kidney1,
           method="jitter",
```

```
        pch = 19,
        col = 2:4,
        vertical = T,
        add = T)
```




```
bwplot(time~sex,data=kidney1,
       panel=panel.violin,
       add=T)
```



We can see that the distribution of recurrence time is highly skewed for both sexes.

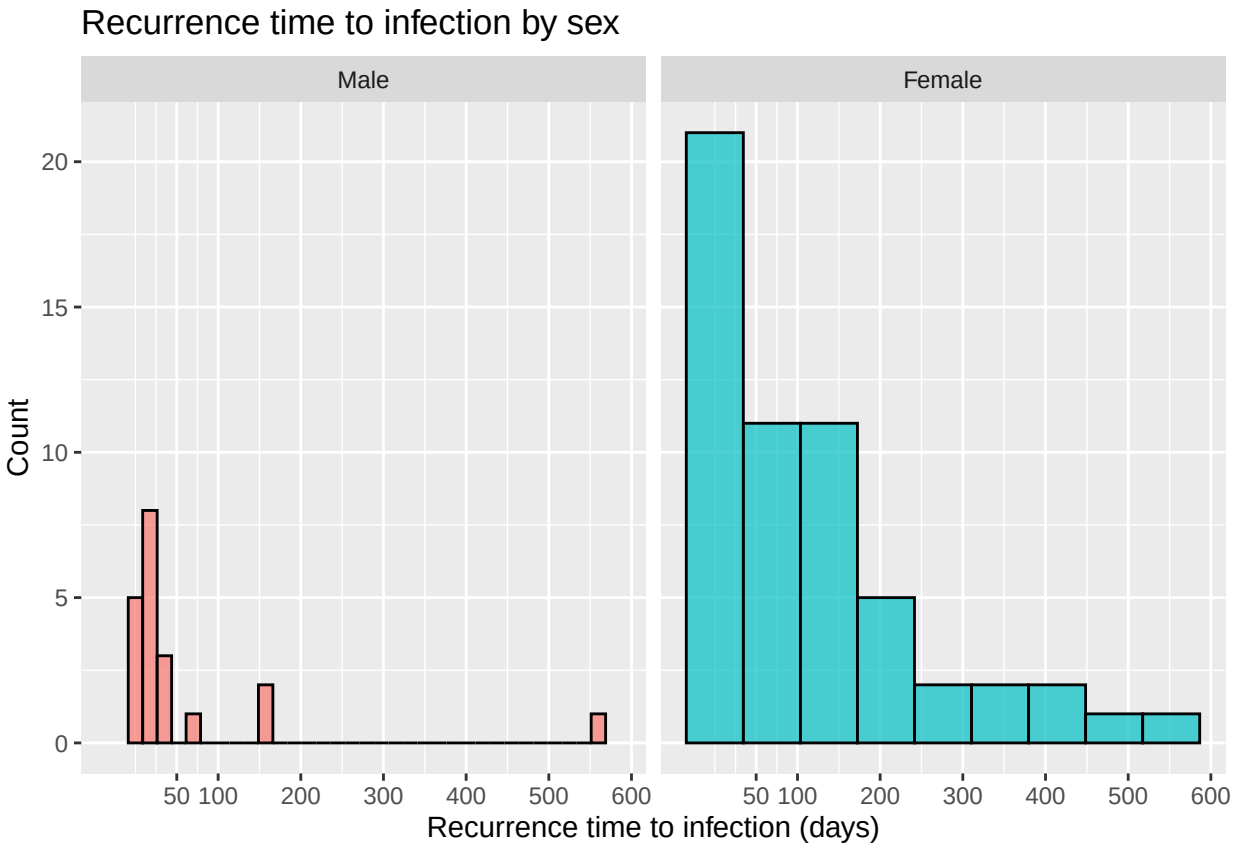
- Male group: Most cases are concentrated at very short times, with a few extreme outliers extending beyond 500 days. This makes the distribution very narrow at the bottom but with a long tail.
- Female group: The distribution is wider and more dispersed, with recurrence times spread across a much larger range. The central mass is higher than in the male group, consistent with the longer median.

2. histograms with group-specific FD binwidths

```
bw_tbl <- kidney1 |>
  group_by(sex) |>
  summarise(bw = 2 * IQR(time) / (n()^(1/3)), .groups = "drop")

ggplot(kidney1, aes(x = time, fill = sex)) +
  geom_histogram(data = subset(kidney1, sex == "Male"),
                 binwidth = bw_tbl$bw[bw_tbl$sex == "Male"],
                 color = "black", alpha = 0.7) +
  geom_histogram(data = subset(kidney1, sex == "Female"),
                 binwidth = bw_tbl$bw[bw_tbl$sex == "Female"],
                 color = "black", alpha = 0.7) +
  facet_grid(cols = vars(sex)) +
  scale_x_continuous(breaks = c(50, 100, 200, 300, 400, 500, 600, 700)) +
```

```
  labs(x = "Recurrence time to infection (days)", y = "Count",
       title = "Recurrence time to infection by sex") +
  theme(legend.position = "none")
```

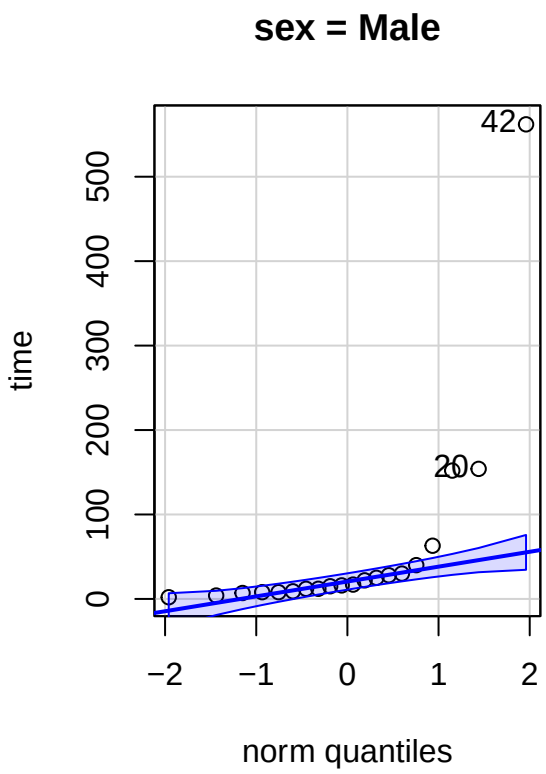


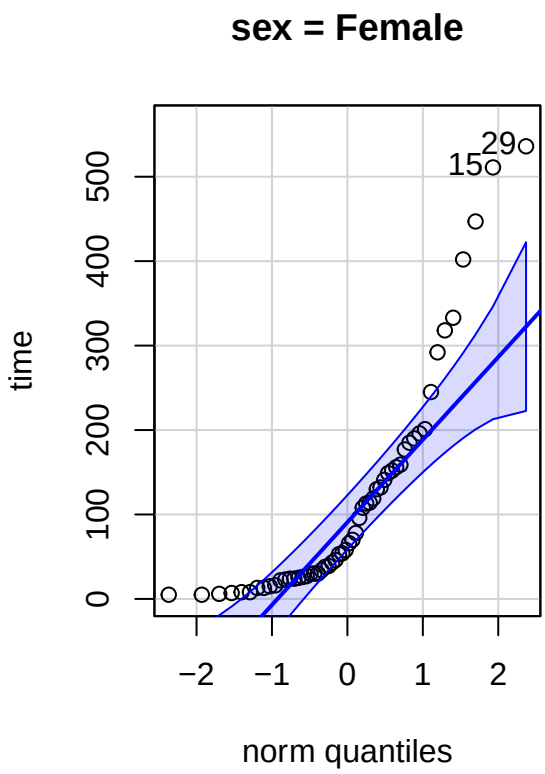
- Most observations for both sexes fall within ~50 days; longer infection-free times become increasingly rare.
- The male panel shows a mid-range gap (300–500 days), which is likely due to small n rather than a structural absence.

3. Normality analysis

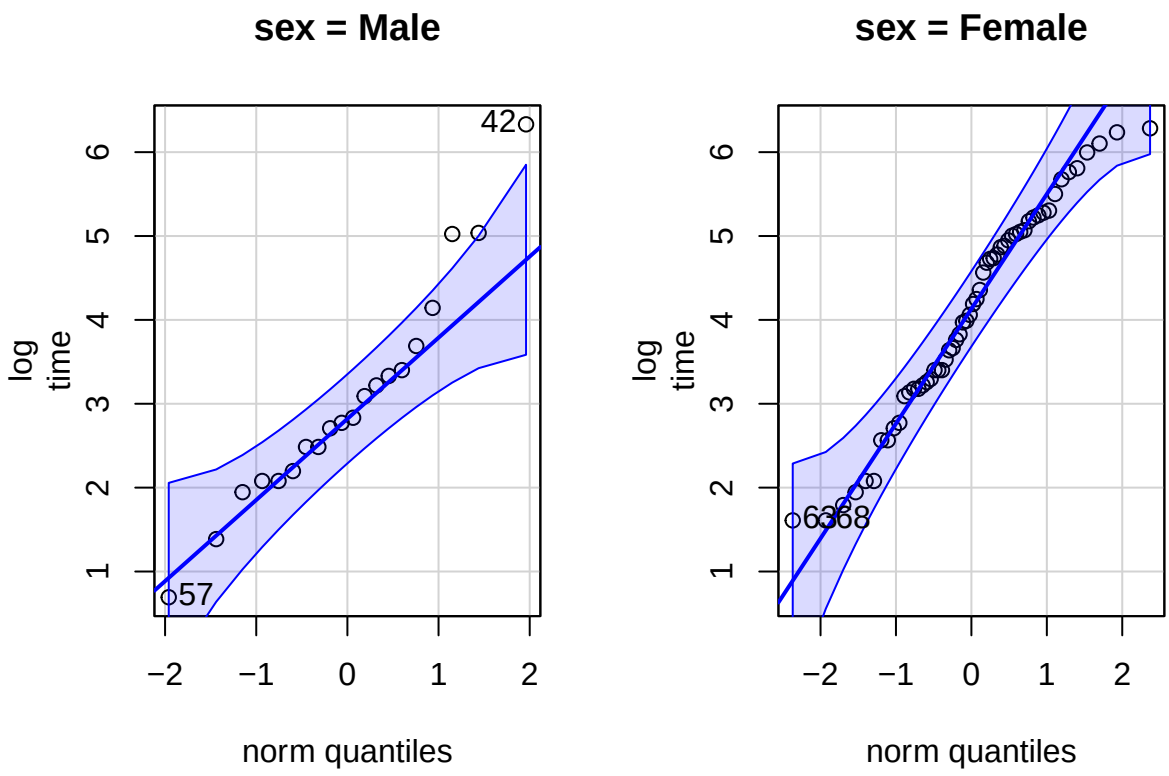
Based on our previous exploration, we found general skewness and kurtosis for both genders. Now we perhaps need qqplots to consider the validity to do OLS regression.

```
library(car)
qqPlot(time~sex,data=kidney1)
```






```
qqPlot(log(time)~sex,data=kidney1)
```



- On the raw time scale, residuals deviate strongly from the 45° line, confirming non-normality due to skewness and heavy tails.
- On the log-transformed scale, the points for both males and females lie closer to the straight line, especially in the central quantiles.
- Some deviations remain in the tails, particularly for the male group, but overall the log transformation greatly improves approximate normality.

4. Recommendation

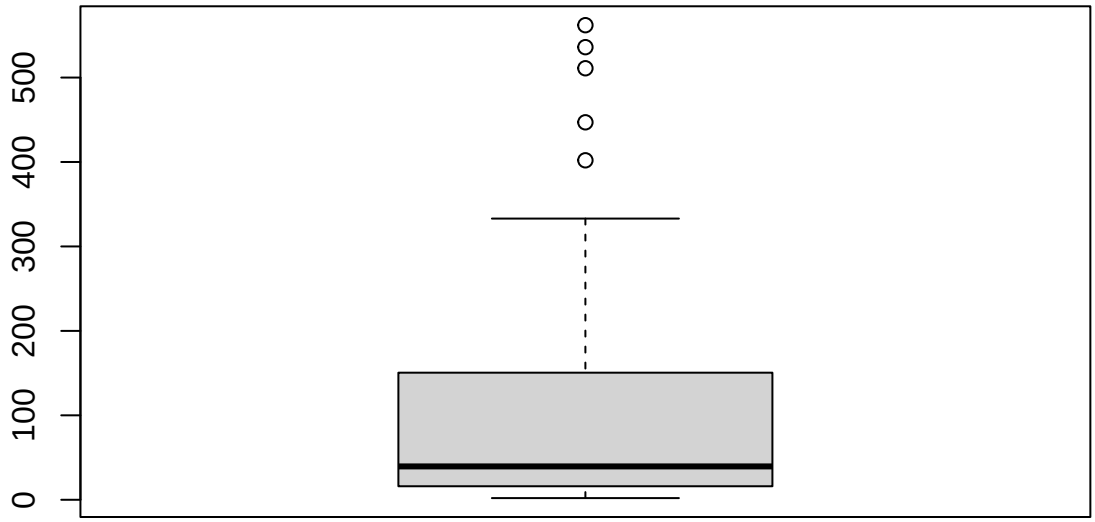
- Fit lm(log(time) ~ sex). Log(time) addresses strong right-skew, heavy tails, and non-constant variance.
- Male n is small and contains extreme long survivors. Perhaps more data is needed for better fitting.
- Provide medians and IQRs by sex to complement mean-based OLS results.

time vs. frailty

```
kidney2 <- survival::kidney |>
  filter(!is.na(time))|>
  filter(!is.na(frail))
```

First, we want to use boxplot method to identify some possible outiliers.

```
outliers<-boxplot(kidney2$time)$out
```



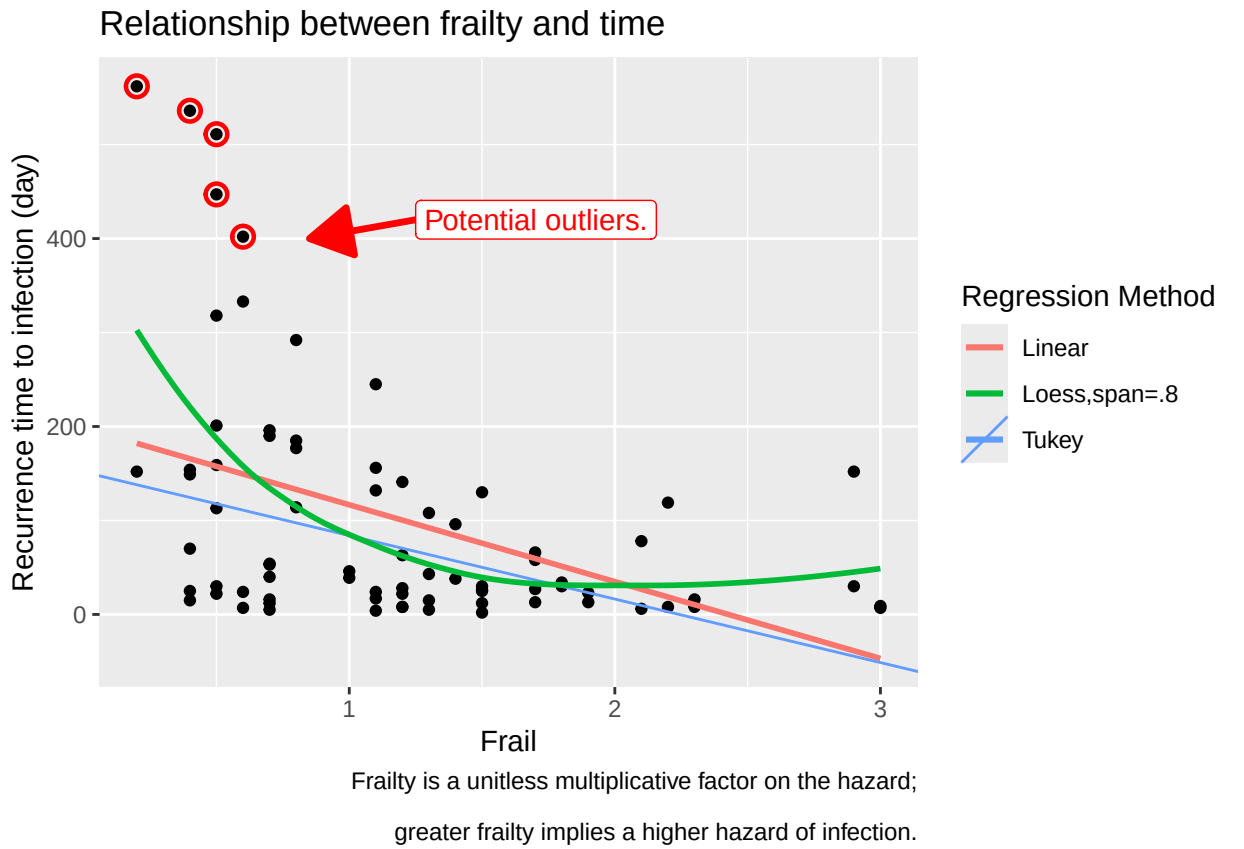

```
out<-kidney2|>
  filter(time%in%outliers)
```

```
tukey<- line(kidney2$frail,kidney2$time,iter=10)

kidney2 |>
  ggplot(aes(x=frail, y=time)) +
  geom_point() +
  geom_point(data=out, colour="red",size=3,shape=1,stroke=1.2,
             show.legend=FALSE)+
  geom_smooth(method="lm",se=F,show.legend = T,aes(colour = "Linear")) +
  geom_abline(aes(intercept = tukey$coefficients[1],
              slope = tukey$coefficients[2],colour = "Tukey"))+
  geom_smooth(method = "loess",se=F,aes(color="Loess,span=.8"),
              show.legend = T,span=.8)+
  labs(x="Frail",
       y="Recurrence time to infection (day)",
       colour="Regression Method",
       title ="Relationship between frailty and time",
       caption="Frailty is a unitless multiplicative factor on the hazard;

 greater frailty implies a higher hazard of infection.")+
   annotate(geom="label",x=1.25,y=420,
```

```
            label="Potential outliers.",
            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")
)
```



Frailty is a unitless multiplicative factor on the hazard;

greater frailty implies a higher hazard of infection.

According to the scatterplot, we can view

- a general decreasing trend of recurrence time as frail increases. This corresponds to the definition of Frailty, as frailty increases, hazard increases, so expected time decreases.
- Also, we discovered that when Frailty is very small, there are 5 outliers which indicates unusually long infection time. This might influence the linear regression effect since simply OLS regression is not resistant to outliers.
- Therefore, we might have to think about tukey regression or polynomial regression.

```
fit_lm   <- lm(time ~ frail, data = kidney2)
fit_loess <- loess(time ~ frail, data = kidney2, span = 0.8, degree = 2)
fit_tukey <- tukey

yhat_lm    <- predict(fit_lm)
yhat_loess <- predict(fit_loess)
yhat_tukey <- fit_tukey$coefficients[1] + fit_tukey$coefficients[2]*kidney2$frail
```

```
res_lm    <- kidney2$time - yhat_lm
res_loess <- kidney2$time - yhat_loess
res_tukey <- fit_tukey$residuals

rse <- data.frame(
  regressionMethod = c("Linear","Tukey","Loess"),
  RSE = c(summary(fit_lm)$sigma,
          sqrt(sum(res_tukey^2)/(nrow(kidney2)-2)),
          fit_loess$s),
  MAD = c(mad(res_lm),
          mad(res_tukey),
          mad(res_loess))
)

rse
```

```
##   regressionMethod      RSE      MAD
## 1           Linear 118.9639 97.19962
## 2            Tukey 123.2162 83.78934
## 3            Loess 113.9217 79.68662
```

By summarizing all the RSEs, we verify that

- Linear regression did get affected by some outliers, resulting in a relatively large MAD in residuals.
- Loess regression displays better regression performance in both RSE and MAD, therefore Loess regression in the analysis between infection time vs. frailty is recommended.

time vs age

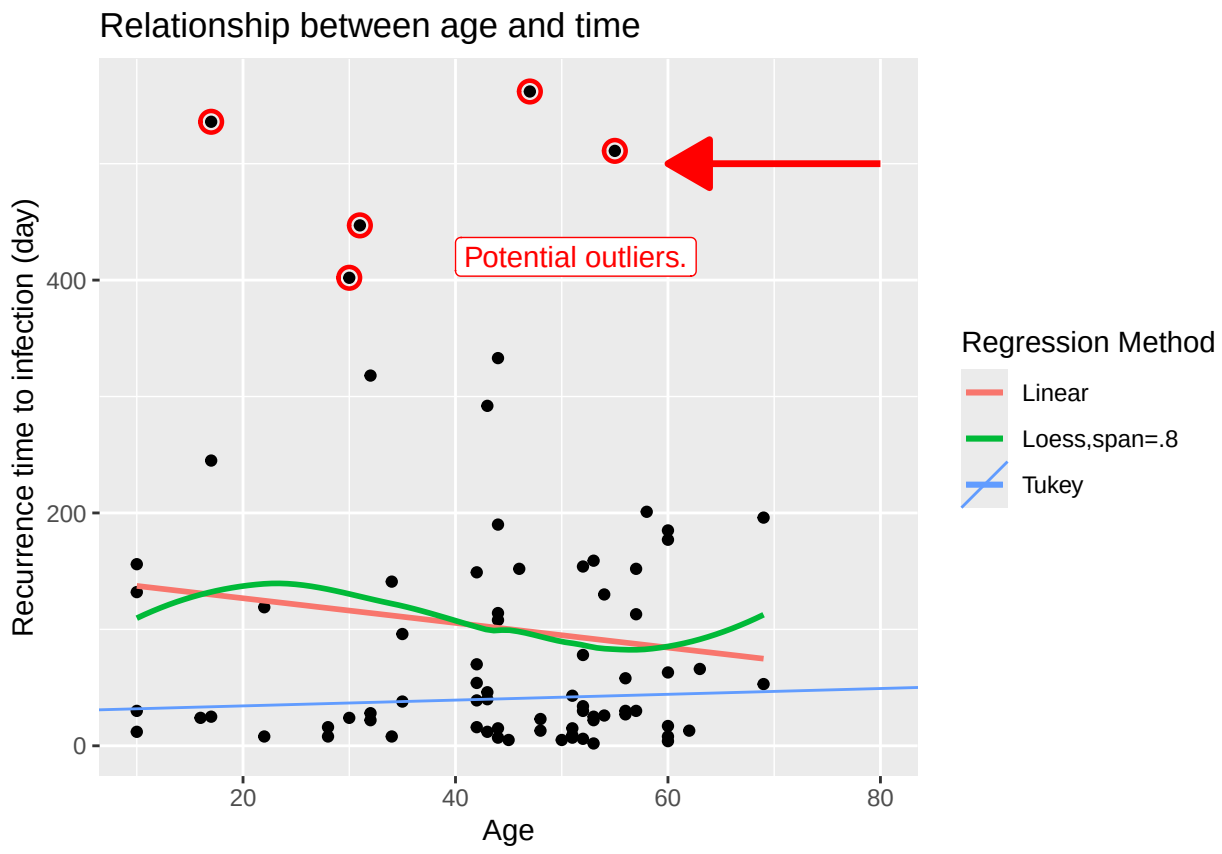
Followed by our analysis of time vs. frailty, we do the similar procedure to time vs. age.

```
kidney3 <- survival::kidney |>
  filter(!is.na(time))|>
  filter(!is.na(age))
```

```
tukey3<- line(kidney3$age,kidney3$time,iter=10)

kidney3 |>
  ggplot(aes(x=age, y=time)) +
  geom_point() +
  geom_point(data=out, colour="red",size=3,shape=1,stroke=1.2,
             show.legend=FALSE)+
  geom_smooth(method="lm",se=F,show.legend = T,aes(colour = "Linear")) +
  geom_abline(aes(intercept = tukey3$coefficients[1],
              slope = tukey3$coefficients[2],colour = "Tukey"))+
  geom_smooth(method = "loess",se=F,aes(color="Loess,span=.8"),
              show.legend = T,span=.8)+
  labs(x="Age",
       y="Recurrence time to infection (day)",
       colour="Regression Method",
```

```
      title ="Relationship between age and time")+
  annotate(geom="label",x=40,y=420,
           label="Potential outliers.",
           hjust="left",color="red")+
  annotate(geom="segment",x=80,y=500,
           xend=60,yend=500,color="red",linewidth=1.2,
           arrow=arrow(type="closed")
)
```




```
fit_lm2   <- lm(time ~ age, data = kidney3)
fit_loess2 <- loess(time ~ age, data = kidney3, span = 0.8, degree = 2)
fit_tukey2 <- tukey3

yhat_lm2    <- predict(fit_lm2)
yhat_loess2 <- predict(fit_loess2)
yhat_tukey2 <- fit_tukey2$coefficients[1] + fit_tukey2$coefficients[2]*kidney3$age

res_lm2    <- kidney3$time - yhat_lm2
res_loess2 <- kidney3$time - yhat_loess2
res_tukey2 <- fit_tukey2$residuals

rse2 <- data.frame(
  regressionMethod = c("Linear","Tukey","Loess"),
  RSE = c(summary(fit_lm2)$sigma,
          sqrt(sum(res_tukey2^2)/(nrow(kidney3)-2)),
```

```
          fit_loess2$s),
  MAD = c(mad(res_lm2),
          mad(res_tukey2),
          mad(res_loess2))
)

rse2
```

```
##   regressionMethod      RSE      MAD
## 1           Linear 130.8545 59.35648
## 2            Tukey 146.2177 47.96971
## 3            Loess 133.3251 59.60419
```

For time vs age,

- OLS regression achieves the lowest RSE, but it is strongly influenced by outliers, leading to relatively large variability.
- Tukey's robust regression yields the lowest MAD, showing stability against extreme long survivors.
- From our `scatterplotmatrix`, the linear relationship between time and age is weak, while loess captures possible nonlinear patterns across age groups, suggesting that a nonlinear fit may be more appropriate than simple linear regression.
- However, from the scatterplot, we think that general relationship (both linear and nonlinear) between these two factors is weak, meaning that age might not be a primary factor to time. We should emphasize on the analysis of other possible factors.