Calder Glass, Kotaro Ito, Zifeng Zhan

4/30/25

PSTAT175 Survival Analysis Project Proposal - Colon Cancer Dataset from the R Survival Package

Description of Dataset:

The dataset in this project is from a clinical trial evaluating the efficacy of adjuvant chemotherapy in patients with stage B/C colon cancer. There are two chemotherapy treatments: Levamisole and Levamisole+5-FU (Fluorouracil). Levamisole is a low-toxicity compound that was originally used to treat worm infestations in animals, while 5-FU is a moderately toxic chemotherapy agent used to treat cancer. It included 929 patients, each of whom was observed for a period of time until an event occurred: cancer recurrence or death, so each patient has two rows.

Scientific Questions:

First, it's imperative to investigate whether any of the treatments (Levamisole and Levamisole plus Fluorouracil) are significantly effective at keeping colon cancer patients alive longer than those without treatment. Furthermore, we are interested in comparing which treatment is more effective in elongating the survival times of patients with colon cancer. Another interesting question for instance, is whether the obstruction of the patient's colon can predict the length of survival time versus patients who experienced perforation instead.

Description of the Covariates:

Age - The age in years of the patient.
Sex - The sex of the patient encoded in binary: "1" being a man and "0" being a woman.
ID - Identification number for the given patient.
Study - The given study that the patients are in: this is just 1 because there is only 1 study.

rx - Whether or not the patient received a treatment: “Obs” indicates they did not receive a treatment, while “Lev(amisole)” and “Lev(amisole)+5-FU” are treatments that some patients received.
obstruct - Whether or not the colon of the patient is obstructed by a tumor: “1” indicates an obstruction while “0” indicates there is no obstruction.
perfor - Whether or not the colon of the patient is perforated: “1” indicates perforation while “0” indicates no perforation.
adhere - Whether or not the cancer is adhering to other organs: “1” indicates adherence while “0” indicates no adherence.
nodes - The total number of lymph nodes that were detected with colon cancer.
time - Time until the patient died or was censored.
status - Whether or not the patient was censored: 0 indicates “censor” while 1 indicates an “event”.
differ - The description of the removed cancer cells from the colon. A higher grade number (3 = Poorly differentiated) indicates that the cancer is growing quickly and more likely to spread, while a lower grade number (1 = Well differentiated) indicates that the cancer is slowly growing and less likely to spread. Other grades: (2 = Moderately differentiated)
extent - Description of the local spread of the tumor (1 is “submucosa”, 2 is “muscle”, 3 is “serosa”, and 4 is “contiguous structures”)
surg - Time from initial surgery to registering to be part of the study. “0” indicates a short time, while “1” indicates a long time.
node4 - Whether or not the patient has more than 4 positive lymph nodes: “1” indicates they have more than 4 positive lymph nodes, while “0” indicates they have 4 or fewer positive lymph nodes.
etype - Whether or not the event was a recurrence (“1”), where the cancer returned, or a death (“2”), where the person died.

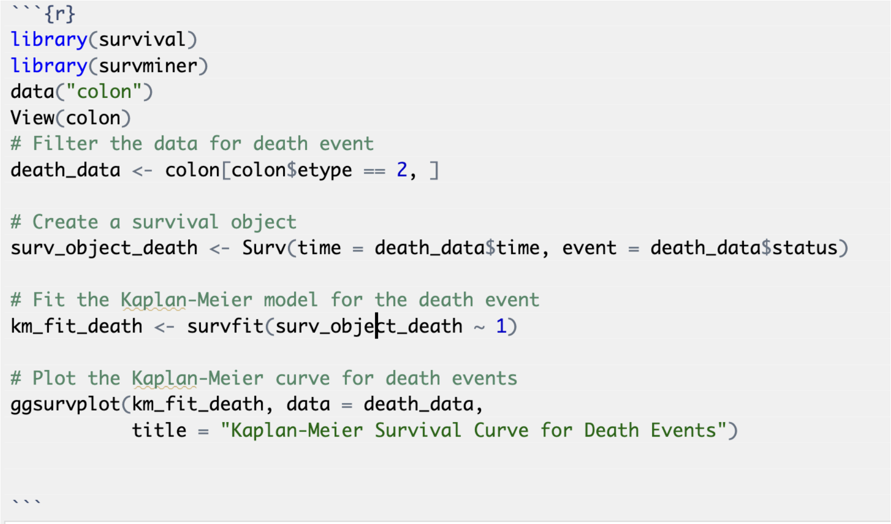
Bibliography:

1. *colon: Chemotherapy for Stage B/C colon cancer.* Colon Cancer. Retrieved May 1st, 2025, from https://rdrr.io/cran/survival/man/colon.html
2. CG Moertel, TR Fleming, JS MacDonald, DG Haller, JA Laurie, CM Tangen, JS Ungerleider, WA Emerson, DC Tormey, JH Glick, MH Veeder and JA Maillard, Fluorouracil plus Levamisole as an effective adjuvant therapy after resection of stage II colon carcinoma: a final report. Annals of Internal Med, 122:321-326, 1991.

Graph of the Data

For this analysis, a Kaplan-Meier plot will be created to visualize the survival probability of colon cancer patients over time in days. The event can be death or recurrence, and the survival time is until a recurrence happens or death.

- Kaplan-Meier survival curve for Death Event



~~~
```{r}
library(survival)
library(survminer)
data("colon")
View(colon)
# Filter the data for death event
death_data <- colon[colon$etype == 2, ]

# Create a survival object
surv_object_death <- Surv(time = death_data$time, event = death_data$status)

# Fit the Kaplan-Meier model for the death event
km_fit_death <- survfit(surv_object_death ~ 1)

# Plot the Kaplan-Meier curve for death events
ggsurvplot(km_fit_death, data = death_data,
           title = "Kaplan-Meier Survival Curve for Death Events")


```
~~~

Figure 1: Code of KM for Death Event

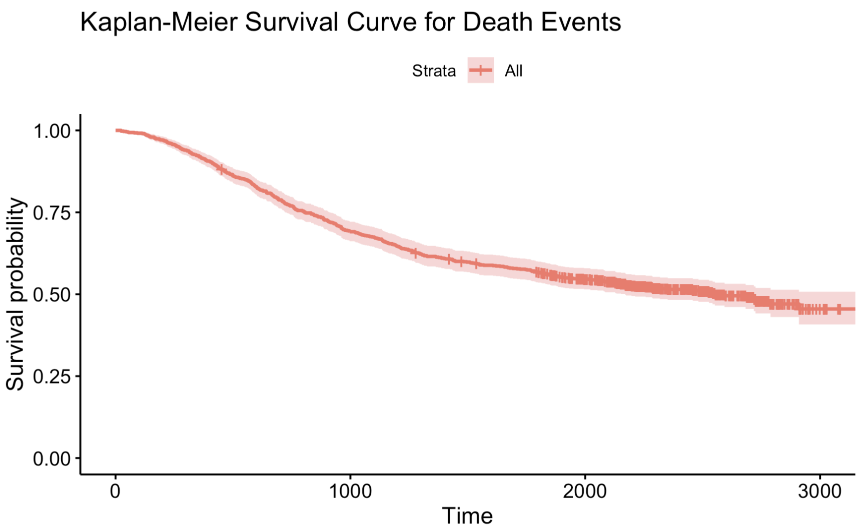
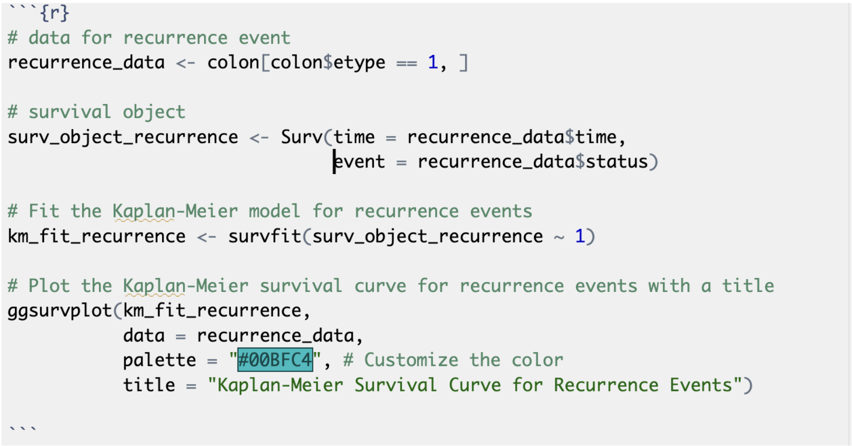


Figure 2: KM for Death Event

- Kaplan-Meier survival curve for Recurrence Event



~~~
```{r}
# data for recurrence event
recurrence_data <- colon[colon$etype == 1, ]

# survival object
surv_object_recurrence <- Surv(time = recurrence_data$time,
                               event = recurrence_data$status)

# Fit the Kaplan-Meier model for recurrence events
km_fit_recurrence <- survfit(surv_object_recurrence ~ 1)

# Plot the Kaplan-Meier survival curve for recurrence events with a title
ggsurvplot(km_fit_recurrence,
           data = recurrence_data,
           palette = "#00BFC4", # Customize the color
           title = "Kaplan-Meier Survival Curve for Recurrence Events")

```
~~~

Figure 3: Code of KM for Recurrence Event

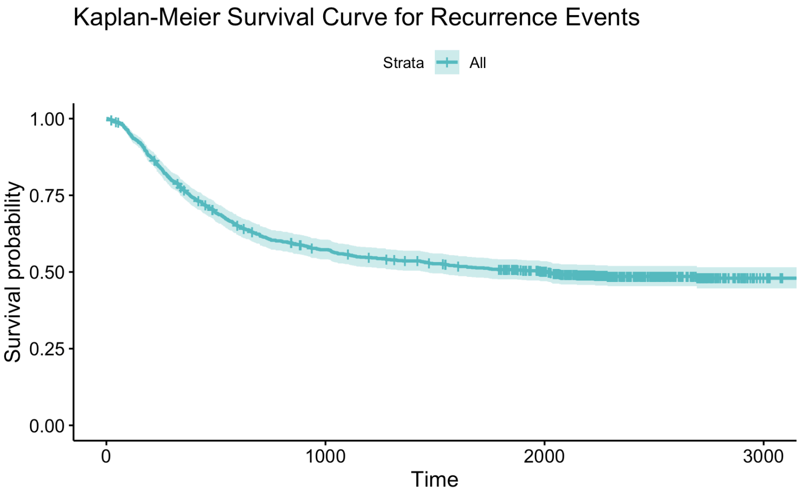


Figure 4: KM for Recurrence Event

Model fitting and proportional hazards assumptions check:

- **Kaplan-Meier:**
- **Cox proportional:**
- **AIC (or BIC) to find the best model**
- **Forward stepwise selection**
- **Log-log plot**
- **Cox zph**

Statistical Tools and Complications

One complication is the *etype* covariate for whether or not the patient experienced a recurrence of colon cancer or passed away from colon cancer. The issue with this is we don't we know if the patient died during the study or after the study.

Another complication is the *surg* covariate. It simplifies the time between the surgery and entering the study too much. It is unknown if there are certain time ranges that are more specific than "long vs short" that could indicate a positive or negative effect on the survival times of patients.

A statistical tool that may come in handy is principal components analysis. PCA is typically used to check for correlated features that can be reduced/eliminated in order to reduce the dimensionality of the dataset. With fewer dimensions, our model for the colon cancer will be more flexible, which made be advantageous.