# Exploratory Data Analysis

Yijia Jiang

2022-11-28

## 1. Data Overview

In this study, the dataset we used is called 'colon', which is found in the 'survival' R package. These are data from one of the first successful trials of adjuvant chemotherapy for colon cancer. Levamisole is a low-toxicity compound previously used to treat worm infestations in animals; 5-FU is a moderately toxic (as these things go) chemotherapy agent. There are two records per person, one for recurrence and one for death. There are 1858 observations and 16 variables.

- id: patient id
- study: 1 for all patients
- rx: treatment - Obs(ervation), Lev(amisole), Lev(amisole)+5-FU
- sex: sex (0 = Female, 1 = Male)
- age: observation age in years
- obstruct: obstruction of colon by tumour (0 = No, 1 = Yes)
- perfor: perforation of colon (0 = No, 1 = Yes)
- adhere: adherence to nearby organs (0 = No, 1 = Yes)
- nodes: number of lymph nodes with detectable cancer
- time: days until event or censoring
- status: censoring status (0 = Censored, 1 = Event)
- differ: differentiation of tumour (1 = Well, 2 = Moderate, 3 = Poor)
- extent: extent of local spread (1 = Submucosa, 2 = Muscle, 3 = Serosa, 4 = Contiguous structures)
- surg: time from surgery to registration (0 = Short, 1 = Long)
- node4: more than 4 positive lymph nodes (0 = No, 1 = Yes)
- etype: event type (1 = recurrence, 2 = death)

The primary endpoints are the death of patients and the recurrence of patients. The type of censoring is right censoring, which means patients left the study before their death.

```
data(cancer, package = "survival")
colon_tb <- as_tibble(colon)

colon_tb$sex <- factor(colon_tb$sex, levels = c(0,1), labels = c("Female", "Male"))
colon_tb$obstruct <- factor(colon_tb$obstruct, levels = c(0,1), labels = c("No", "Yes"))
colon_tb$perfor <- factor(colon_tb$perfor, levels = c(0,1), labels = c("No", "Yes"))
colon_tb$adhere <- factor(colon_tb$adhere, levels = c(0,1), labels = c("No", "Yes"))
colon_tb$status <- factor(colon_tb$status, levels = c(0,1), labels = c("Censored", "Dead"))
colon_tb$differ <- factor(colon_tb$differ, levels = c(1,2,3), labels = c("Well", "Moderate","Poor"))
colon_tb$extent <- factor(colon_tb$extent, levels = c(1,2,3,4), labels = c("Submucosa", "Muscle","Seros
colon_tb$surg <- factor(colon_tb$surg, levels = c(0,1), labels = c("Short", "Long"))
colon_tb$node4 <- factor(colon_tb$node4, levels = c(0,1), labels = c("No", "Yes"))
```

```r
label(colon_tb$rx) <- "Treatment"
label(colon_tb$sex) <- "Sex"
label(colon_tb$age) <- "Age"
label(colon_tb$obstruct) <- "Obstruction of colon by tumour"
label(colon_tb$perfor) <- "Perforation of colon"
label(colon_tb$adhere) <- "Adherence to nearby organs"
label(colon_tb$nodes) <- "Number of lymph nodes with detectable cancer"
label(colon_tb$time) <- "Time until event or censoring"
label(colon_tb$status) <- "Censoring status"
label(colon_tb$differ) <- "Differentiation of tumour"
label(colon_tb$extent) <- "Extent of local spread"
label(colon_tb$surg) <- "Time from surgery to registration"
label(colon_tb$node4) <- "More than 4 positive lymph nodes"

units(colon_tb$age) <- "years"
units(colon_tb$time) <- "days"

table1(~ rx + sex + age + obstruct + perfor + adhere + differ + extent + surg + nodes + node4| status, c
```

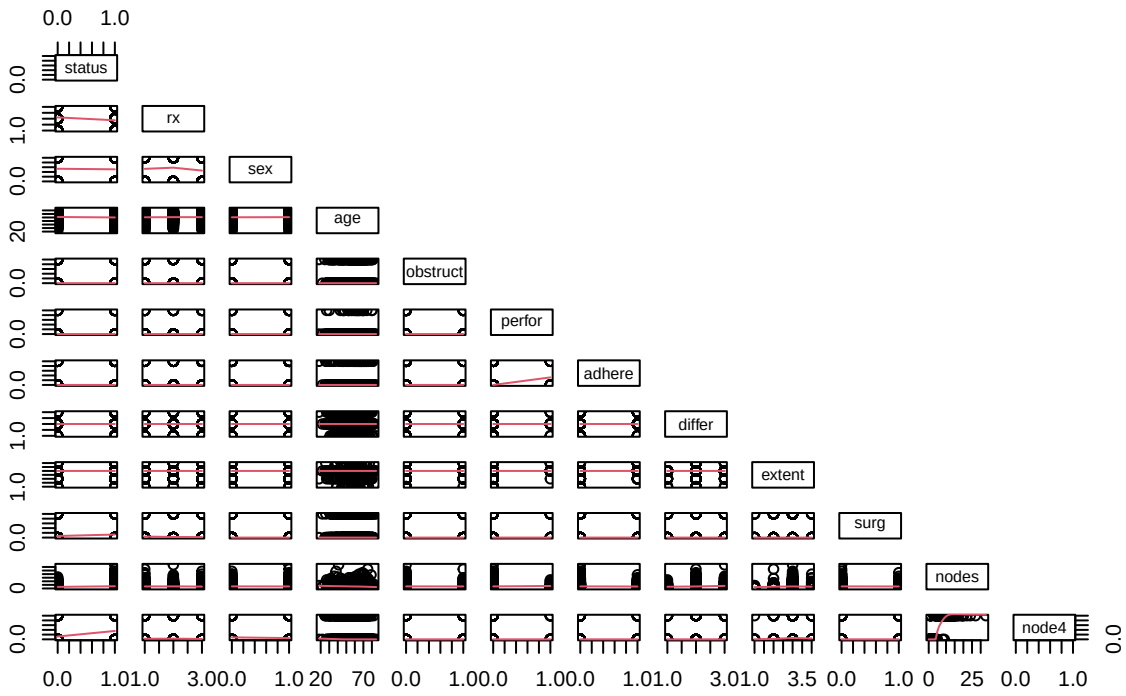
| | Censored | Dead | Total |
|---|---|---|---|
| | (N=461) | (N=468) | (N=929) |
| **Treatment** | | | |
| Obs | 138 (29.9%) | 177 (37.8%) | 315 (33.9%) |
| Lev | 138 (29.9%) | 172 (36.8%) | 310 (33.4%) |
| Lev+5FU | 185 (40.1%) | 119 (25.4%) | 304 (32.7%) |
| **Sex** | | | |
| Female | 216 (46.9%) | 229 (48.9%) | 445 (47.9%) |
| Male | 245 (53.1%) | 239 (51.1%) | 484 (52.1%) |
| **Age (years)** | | | |
| Mean (SD) | 60.5 (11.5) | 59.0 (12.4) | 59.8 (11.9) |
| Median [Min, Max] | 61.0 [22.0, 83.0] | 61.0 [18.0, 85.0] | 61.0 [18.0, 85.0] |
| **Obstruction of colon by tumour** | | | |
| No | 380 (82.4%) | 369 (78.8%) | 749 (80.6%) |
| Yes | 81 (17.6%) | 99 (21.2%) | 180 (19.4%) |
| **Perforation of colon** | | | |
| No | 451 (97.8%) | 451 (96.4%) | 902 (97.1%) |
| Yes | 10 (2.2%) | 17 (3.6%) | 27 (2.9%) |
| **Adherence to nearby organs** | | | |
| No | 408 (88.5%) | 386 (82.5%) | 794 (85.5%) |
| Yes | 53 (11.5%) | 82 (17.5%) | 135 (14.5%) |
| **Differentiation of tumour** | | | |
| Well | 49 (10.6%) | 44 (9.4%) | 93 (10.0%) |
| Moderate | 337 (73.1%) | 326 (69.7%) | 663 (71.4%) |
| Poor | 62 (13.4%) | 88 (18.8%) | 150 (16.1%) |
| Missing | 13 (2.8%) | 10 (2.1%) | 23 (2.5%) |
| **Extent of local spread** | | | |
| Submucosa | 16 (3.5%) | 5 (1.1%) | 21 (2.3%) |
| Muscle | 72 (15.6%) | 34 (7.3%) | 106 (11.4%) |
| Serosa | 359 (77.9%) | 400 (85.5%) | 759 (81.7%) |
| Contiguous structures | 14 (3.0%) | 29 (6.2%) | 43 (4.6%) |
| **Time from surgery to registration** | | | |
| Short | 355 (77.0%) | 327 (69.9%) | 682 (73.4%) |
| Long | 106 (23.0%) | 141 (30.1%) | 247 (26.6%) |
| **Number of lymph nodes with detectable cancer** | | | |
| Mean (SD) | 2.76 (2.43) | 4.56 (4.25) | 3.66 (3.57) |
| Median [Min, Max] | 2.00 [1.00, 16.0] | 3.00 [0, 33.0] | 2.00 [0, 33.0] |
| Missing | 6 (1.3%) | 12 (2.6%) | 18 (1.9%) |
| **More than 4 positive lymph nodes** | | | |
| No | 386 (83.7%) | 288 (61.5%) | 674 (72.6%) |
| Yes | 75 (16.3%) | 180 (38.5%) | 255 (27.4%) |

```
table1(~ rx + sex + age + obstruct + perfor + adhere + differ + extent + surg + nodes + node4| status, c
```

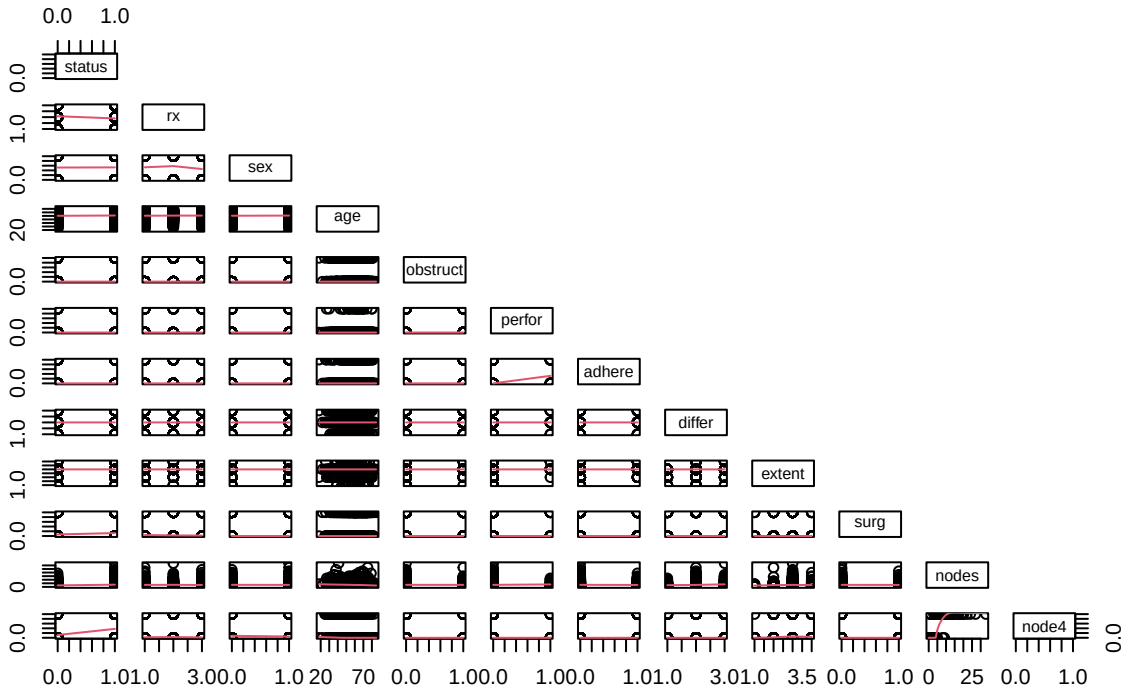
| | Censored | Dead | Total |
|---|---|---|---|
| | (N=477) | (N=452) | (N=929) |
| **Treatment** | | | |
| Obs | 147 (30.8%) | 168 (37.2%) | 315 (33.9%) |
| Lev | 149 (31.2%) | 161 (35.6%) | 310 (33.4%) |
| Lev+5FU | 181 (37.9%) | 123 (27.2%) | 304 (32.7%) |
| **Sex** | | | |
| Female | 230 (48.2%) | 215 (47.6%) | 445 (47.9%) |
| Male | 247 (51.8%) | 237 (52.4%) | 484 (52.1%) |
| **Age (years)** | | | |
| Mean (SD) | 59.6 (11.6) | 59.9 (12.3) | 59.8 (11.9) |
| Median [Min, Max] | 60.0 [22.0, 83.0] | 62.0 [18.0, 85.0] | 61.0 [18.0, 85.0] |
| **Obstruction of colon by tumour** | | | |
| No | 395 (82.8%) | 354 (78.3%) | 749 (80.6%) |
| Yes | 82 (17.2%) | 98 (21.7%) | 180 (19.4%) |
| **Perforation of colon** | | | |
| No | 465 (97.5%) | 437 (96.7%) | 902 (97.1%) |
| Yes | 12 (2.5%) | 15 (3.3%) | 27 (2.9%) |
| **Adherence to nearby organs** | | | |
| No | 421 (88.3%) | 373 (82.5%) | 794 (85.5%) |
| Yes | 56 (11.7%) | 79 (17.5%) | 135 (14.5%) |
| **Differentiation of tumour** | | | |
| Well | 51 (10.7%) | 42 (9.3%) | 93 (10.0%) |
| Moderate | 352 (73.8%) | 311 (68.8%) | 663 (71.4%) |
| Poor | 62 (13.0%) | 88 (19.5%) | 150 (16.1%) |
| Missing | 12 (2.5%) | 11 (2.4%) | 23 (2.5%) |
| **Extent of local spread** | | | |
| Submucosa | 17 (3.6%) | 4 (0.9%) | 21 (2.3%) |
| Muscle | 70 (14.7%) | 36 (8.0%) | 106 (11.4%) |
| Serosa | 376 (78.8%) | 383 (84.7%) | 759 (81.7%) |
| Contiguous structures | 14 (2.9%) | 29 (6.4%) | 43 (4.6%) |
| **Time from surgery to registration** | | | |
| Short | 366 (76.7%) | 316 (69.9%) | 682 (73.4%) |
| Long | 111 (23.3%) | 136 (30.1%) | 247 (26.6%) |
| **Number of lymph nodes with detectable cancer** | | | |
| Mean (SD) | 2.72 (2.44) | 4.66 (4.25) | 3.66 (3.57) |
| Median [Min, Max] | 2.00 [1.00, 19.0] | 3.00 [0, 33.0] | 2.00 [0, 33.0] |
| Missing | 7 (1.5%) | 11 (2.4%) | 18 (1.9%) |
| **More than 4 positive lymph nodes** | | | |
| No | 403 (84.5%) | 271 (60.0%) | 674 (72.6%) |
| Yes | 74 (15.5%) | 181 (40.0%) | 255 (27.4%) |

## 2. Corrlation between Variables

```
pairs(~ status + rx + sex + age + obstruct + perfor + adhere + differ + extent + surg + nodes + node4,
```

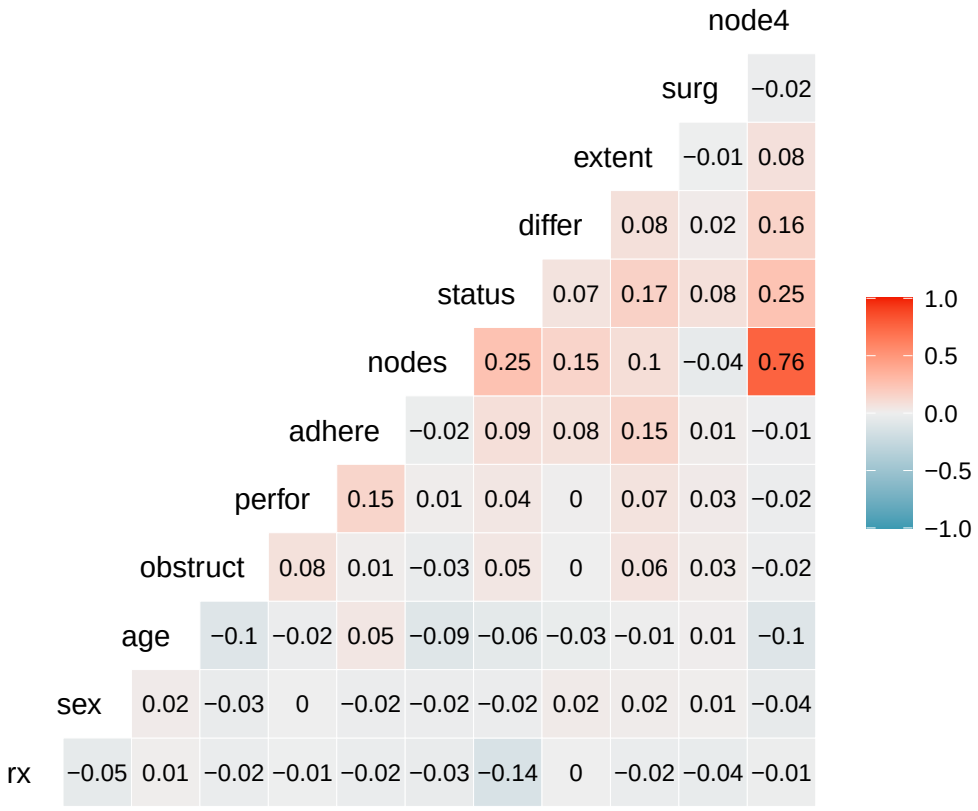



```
pairs(~ status + rx + sex + age + obstruct + perfor + adhere + differ + extent + surg + nodes + node4,
```

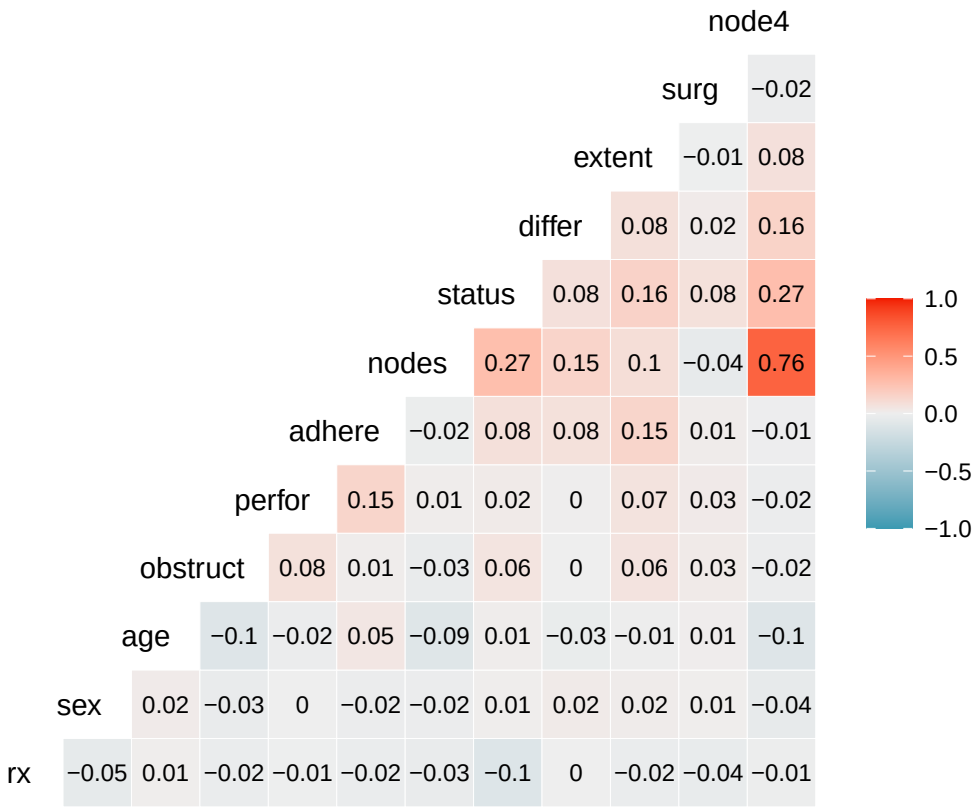



```
colon %>%
  subset(etype == 1) %>%
```

```
  mutate(rx = as.numeric(rx)) %>%
  dplyr::select(-id, -study,-etype, -time) %>%
  ggcorr(label = TRUE, hjust = 0.9, layout.exp = 2, label_size = 3, label_round = 2)
```




```
colon %>%
  subset(etype == 2) %>%
  mutate(rx = as.numeric(rx)) %>%
  dplyr::select(-id, -study,-etype, -time) %>%
  ggcorr(label = TRUE, hjust = 0.9, layout.exp = 2, label_size = 3, label_round = 2)
```


node4
surg −0.02
extent −0.01 0.08
differ 0.08 0.02 0.16
status 0.08 0.16 0.08 0.27
nodes 0.27 0.15 0.1 −0.04 0.76
adhere −0.02 0.08 0.08 0.15 0.01 −0.01
perfor 0.15 0.01 0.02 0 0.07 0.03 −0.02
obstruct 0.08 0.01 −0.03 0.06 0 0.06 0.03 −0.02
age −0.1 −0.02 0.05 −0.09 0.01 −0.03 −0.01 0.01 −0.1
sex 0.02 −0.03 0 −0.02 −0.02 0.01 0.02 0.02 0.01 −0.04
rx −0.05 0.01 −0.02 −0.01 −0.02 −0.03 −0.1 0 −0.02 −0.04 −0.01
1.0
0.5
0.0
−0.5
−1.0