HDS 5230 High performance computing

Homework Week 2

Miao Cai[*]

2019-01-30

You will use these datasets to answer some questions listed below. You must be careful to think about what the appropriate denominator is for each question. As you code the answers, be mindful to use the 'high performance' coding approaches in data.table.

```
library(data.table)
data_path = "healthcare2/"
csv_files = list.files(path = data_path, pattern = "*.csv")

readallcsv = function(i){
  assign(
    gsub(".csv", "", csv_files[i]),
    fread(paste0(data_path, csv_files[i])),
    envir = parent.frame()
  )
}

for (i in seq_along(csv_files))  readallcsv(i)
```

1) Are men more likely to die than women in this group of patients? Assume people without a date of death in the mortality table are still alive.

Here I recode all patients without a `Gender` to `other`.

```
Patient[!Gender %in% c("female", "male"), Gender := "other"]
q1 = Mortality[Patient, on = "PatientID"]
q1[, .(death_percent = sum(!is.na(DateOfDeath))/.N), by = Gender][order(-death_percent)]
```

```
##    Gender death_percent
## 1:   male     0.3594713
## 2: female     0.3511153
## 3:  other     0.3492670
```

According to the returned data, it seems that males do have a little bit higher chance to die than women in this group of patients, although the difference is nominal.

2) I am interested to know if there are patterns in the disease groups across gender. For every patient with at least one outpatient visit, identify if they have been diagnosed with any of the 22 conditions listed in the diseaseMap table at any time point. You will need to consider all three ICD columns in the outpatientVisit file (not just one). Create a table with the rate of disease for each condition for men, women, and all. It should look like this, where the XX% is the percent with the condition:

[*]Department of Epidemiology and Biostatistics, Saint Louis University. Email address miao.cai@slu.edu

```r
OutpatientVisit = melt(
  OutpatientVisit,
  measure.vars = patterns("^ICD10"),
  id.vars = c("VisitID", "PatientID"),
  value.name = "ICD10"
)

num_gender = Patient[, .N, by = Gender]

q2 = DiseaseMap[
  OutpatientVisit, on = "ICD10"
  ][, .N, by = .(PatientID, Condition)][Patient, on = "PatientID"]

q2_1 = q2[,.(condition_N = .N), by = .(Condition, Gender)]
q2_1 = num_gender[q2_1, on = "Gender"][
  , mortalityrate := paste0(round(condition_N*100/N, 2), "%")][
    order(Condition, Gender)][,condition_N := NULL][,N := NULL]
q2_1 = dcast(q2_1, Condition ~ Gender, value.var = "mortalityrate")

q2_2 = q2[
  ,.(condition_N = .N), by = .(Condition)
  ][,Overall := paste0(round(condition_N*100/nrow(Patient), 2), "%")][
    order(Condition)][,condition_N:=NULL]

q2 = q2_2[q2_1, on = "Condition"][!is.na(Condition),]

q2
```

```
##                          Condition Overall female   male  other
##  1:                         Alcohol   7.88%  7.75%  7.99%  8.09%
##  2:                          Cancer   4.97%     5%     5%  4.72%
##  3:       Congestive_heart_failure   4.32%  3.06%  5.61%  4.59%
##  4:                        Dementia   3.08%  3.19%  2.98%  3.06%
##  5:                      Depression  10.53% 12.44%  8.41% 11.03%
##  6:    Diabetes_with_complications   4.05%  4.18%  3.85%  4.46%
##  7: Diabetes_without_complications   9.97% 10.25%   9.8%  9.31%
##  8:                           Drugs   3.94%  4.07%  3.84%  3.76%
##  9:                             HIV   0.62%  0.57%  0.63%  0.96%
## 10:                    Hypertension  30.09% 28.54%  32.1% 28.11%
## 11:                       LiverMild   0.92%  0.95%  0.93%   0.7%
## 12:                     LiverSevere   5.07%  4.84%  5.22%  5.54%
## 13:        Metastatic_solid_tumour   3.29%  3.28%  3.46%  2.42%
## 14:          Myocardial_infarction    4.5%  3.16%  5.89%  4.72%
## 15:                         Obesity  16.21% 18.36% 13.97% 15.93%
## 16:                       Paralysis   1.34%  1.46%  1.17%  1.59%
## 17:           Peptic_ulcer_disease   0.96%  1.02%   0.9%  0.89%
## 18:    Peripheral_vascular_disease   2.38%  2.41%  2.24%     3%
## 19:                       Pulmonary   7.17%  7.07%  7.25%  7.27%
## 20:                           Renal   3.49%  3.63%  3.42%  3.06%
## 21:                       Rheumatic   1.22%   1.3%   1.1%  1.34%
## 22:                          Stroke   2.82%  2.66%  3.04%  2.49%
##                          Condition Overall female   male  other
```

I assume the denominator for each gender is the number of patients in that specific group.Whereas the denominator for overall group is the number of patients : 20000

Out of the 22 conditions, women had 10 conditions with lower rate than man. It seems that there are no significant pattern between man and women.

3) Calculate the mortality rate for every year between 2005 and 2018. Is it generally increasing, or decreasing? Assume patients are only at risk of death as of their first visit (in the outpatient Visit file). Once they have died, they are no longer at risk in subsequent years

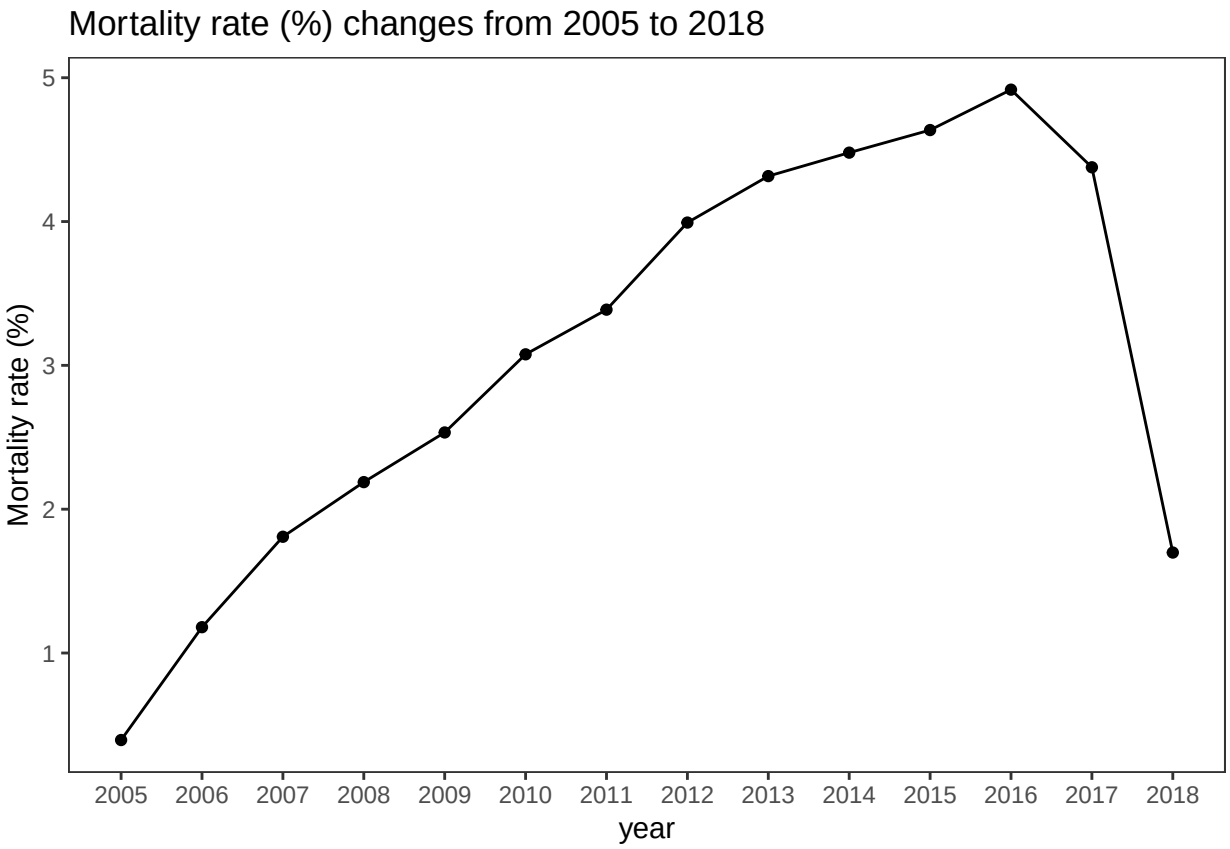
```
q3 = Mortality[Patient, on = "PatientID"]
q3[, year := as.integer(substr(DateOfDeath, 1, 4))]
q3 = q3[, .(N_death = .N), by = year][order(year)][!is.na(year)]

q3[, cum_death := cumsum(N_death)
  ][, atrisk := 20000 - shift(cum_death, fill = 0, type = "lag")
    ][, mortality_rate := N_death*100/atrisk]

q3
```

```
##     year N_death cum_death atrisk mortality_rate
##  1: 2005      79        79  20000       0.395000
##  2: 2006     235       314  19921       1.179660
##  3: 2007     356       670  19686       1.808392
##  4: 2008     423      1093  19330       2.188308
##  5: 2009     479      1572  18907       2.533453
##  6: 2010     567      2139  18428       3.076840
##  7: 2011     605      2744  17861       3.387268
##  8: 2012     689      3433  17256       3.992814
##  9: 2013     715      4148  16567       4.315809
## 10: 2014     710      4858  15852       4.478930
## 11: 2015     702      5560  15142       4.636111
## 12: 2016     710      6270  14440       4.916898
## 13: 2017     601      6871  13730       4.377276
## 14: 2018     223      7094  13129       1.698530
```

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
ggplot(q3, aes(year, mortality_rate)) +
  geom_point() + geom_line() +
  scale_x_continuous("year", labels = 2005:2018, breaks = 2005:2018) +
  labs(title = "Mortality rate (%) changes from 2005 to 2018") +
  ylab("Mortality rate (%)") + theme_test()
```



According to the time trend plot, the mortality rate has been generally increasing, while it experienced a major drop in the recent two years (2017 and 2018).

a. This is a harder question to answer than at first glance. What should the denominator of patients be for every year? How will you calculate it?

From my understanding, the denominator should be the patients at risk in the specific year (who were still alive). I calculated it by excluding the patients who died in the last year.