Ex3_bx2168_hl3339_wf2255

Banruo Xie;Wen Fan;Hanjun Li

3/7/2020

```
library(dplyr)
library(lubridate)
```

(1)

a_1 means the probability of rainy day given the previous day is rainy day

a_2 means the probability of no rain day given the previous day is rainy day

a_3 means the probability of rainy day given the previous day is no rain day

a_4 means the probability of no rain day given the previous day is no rain day

(2)

Let X_i represent whether the ith day is rainy.

By Bayesian formula:

$P(X_n = 0) = P(X_n = 0|X_{n-1} = 0)P(X_{n-1} = 0) + P(X_n = 0|X_{n-1} = 1)P(X_{n-1} = 1) = a_1P(X_n = 0) + a_3(1 - P(X_n = 0))$

Therefore, $P(X_n = 0) = \frac{a_3}{1-a_1+a_3}$

```
data <- read.csv('./CentralPark.csv', header = T)
data$DATE <- as.POSIXct(strptime(as.character(data$DATE), "%m/%d/%y"))
data <- data %>% mutate(is_rain = if_else(PRCP>=1.5,T,F))
data$month <- month(data$DATE)
data$will_rain <- append(data$is_rain,c(NA))[2:(length(data$is_rain)+1)]
print(c((nrow(data %>% filter(month == 7, is_rain, will_rain)))/
           nrow(data %>% filter(month == 7, is_rain)),
         (nrow(data %>% filter(month == 7, is_rain, !will_rain)))/
           nrow(data %>% filter(month == 7, is_rain)),
         (nrow(data %>% filter(month == 7, !is_rain, will_rain)))/
           nrow(data %>% filter(month == 7, !is_rain)),
         (nrow(data %>% filter(month == 7, !is_rain, !will_rain)))/
           nrow(data %>% filter(month == 7, !is_rain))))
```

```
## [1] 0.3107527 0.6892473 0.2308808 0.7691192
```

(4)

Hypothesis test: $H_0 : p_{00} = p_{11}$, $H_1 : p_{00} \neq p_{11}$

p_{00} is the probability of rainy day given the previous day is rainy day

p_{11} is the probability of no rain day given the previous day is no rain day

Since p_{00} and p_{11} are independent, therefore, $\hat{p}_{00} \xrightarrow[\infty]{D} N(\hat{p}_{00}, \frac{\hat{p}_{00}(1-\hat{p}_{00})}{n_0})$

$\hat{p}_{11} \xrightarrow[\infty]{D} N(\hat{p}_{11}, \frac{\hat{p}_{11}(1-\hat{p}_{11})}{n_1})$

$\hat{p}_{00} - \hat{p}_{11} \xrightarrow[\infty]{D} N(0, \frac{\hat{p}_{00}(1-\hat{p}_{00})}{n_0} - \frac{\hat{p}_{11}(1-\hat{p}_{11})}{n_1})$

```
a1 = (nrow(data %>% filter(month == 7, is_rain, will_rain)))/
  nrow(data %>% filter(month == 7, is_rain))
a4 = (nrow(data %>% filter(month == 7, !is_rain, !will_rain)))/
  nrow(data %>% filter(month == 7, !is_rain))
print(pnorm((a1-a4)/sqrt(a1*(1-a1)/nrow(data %>% filter(month == 7, is_rain))
                         +a4*(1-a4)/nrow(data %>% filter(month == 7, !is_rain)))))
```

```
## [1] 2.223776e-157
```

Therefore, we reject H_0

(5)

```
data$will_rain2 <- append(data$will_rain,c(NA))[2:(length(data$will_rain)+1)]
```

H_0: Higher model chain can not improve. H_1: Higher model chain does improve.

Using likelihood ratio test:

$$\Lambda_n = 2\left\{\ell(\hat{\mathbf{P}})_{\text{second order}} - \ell(\hat{\mathbf{P}})_{\text{first order}}\right\} = 2\left\{\sum_{r=1}^{S}\sum_{s=1}^{S}\sum_{t=1}^{S} n_{rst}\log\hat{p}_{rst} - \sum_{s=1}^{S}\sum_{t=1}^{S} n_{.st}\log\hat{p}_{st}\right\}$$
$$= 2\left\{\sum_{r=1}^{S}\sum_{s=1}^{S}\sum_{t=1}^{S} n_{rst}\log\hat{p}_{rst} - \sum_{r=1}^{S}\sum_{s=1}^{S}\sum_{t=1}^{S} n_{rst}\log\hat{p}_{st}\right\} = 2\sum_{r=1}^{S}\sum_{s=1}^{S}\sum_{t=1}^{S} n_{rst}\log\left(\frac{\hat{p}_{rst}}{\hat{p}_{st}}\right)$$

By asymptotic theory, $\Lambda_n \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_{(S-1)^2}$

```
p00 <- (nrow(data %>% filter(month == 7, is_rain, will_rain)))/
  nrow(data %>% filter(month == 7, is_rain))
p01 <- (nrow(data %>% filter(month == 7, is_rain, !will_rain)))/
  nrow(data %>% filter(month == 7, is_rain))
p10 <- (nrow(data %>% filter(month == 7, !is_rain, will_rain)))/
  nrow(data %>% filter(month == 7, !is_rain))
p11 <- (nrow(data %>% filter(month == 7, !is_rain, !will_rain)))/
  nrow(data %>% filter(month == 7, !is_rain))

r000 <- nrow(data %>% filter(month == 7, is_rain, will_rain, will_rain2))
r001 <- nrow(data %>% filter(month == 7, is_rain, will_rain, !will_rain2))
r010 <- nrow(data %>% filter(month == 7, is_rain, !will_rain, will_rain2))
r011 <- nrow(data %>% filter(month == 7, is_rain, !will_rain, !will_rain2))
r100 <- nrow(data %>% filter(month == 7, !is_rain, will_rain, will_rain2))
r101 <- nrow(data %>% filter(month == 7, !is_rain, will_rain, !will_rain2))
r110 <- nrow(data %>% filter(month == 7, !is_rain, !will_rain, will_rain2))
r111 <- nrow(data %>% filter(month == 7, !is_rain, !will_rain, !will_rain2))

p000 <- r000/(r000 + r001)
p001 <- r001/(r000 + r001)
p010 <- r010/(r010 + r011)
p011 <- r011/(r010 + r011)
p100 <- r100/(r100 + r101)
p101 <- r101/(r100 + r101)
p110 <- r110/(r110 + r111)
p111 <- r111/(r110 + r111)
```

```r
result <- 2* (r000*log(p000/p00) + r001*log(p001/p01) + r010*log(p010/p10)
              + r011*log(p011/p11) + r100*log(p100/p00)
              + r101*log(p101/p01) + r110*log(p110/p10)
              + r111*log(p111/p11))


pchisq(result,2)
```

```
## [1] 0.8286566
```

Therefore, we fail to reject H_0, higher model chain does not improve fit of the data.