Exercise3

Chutian Chen cc4515; Congcheng Yan cy2550; Mingrui Liu ml4404

3/1/2020

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

(1)

Let X_i represent whether the ith day is rainy.

a_1: $P(X_i = 0|X_{i-1} = 0)$

a_2: $P(X_i = 1|X_{i-1} = 0)$

a_3: $P(X_i = 0|X_{i-1} = 1)$

a_4: $P(X_i = 1|X_{i-1} = 1)$

(2)

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

So $P(X = 0) = \frac{a_3}{1-a_1+a_3}$

(3)

```
data <- read.csv('./CentralPark.csv', header = T)
data$DATE <- as.POSIXct(strptime(as.character(data$DATE), "%m/%d/%y"))
data <- data %>% mutate(rain = if_else(PRCP>=1.5,TRUE,FALSE))
data$month <- month(data$DATE)
data$rain_tomo <- append(data$rain,c(NA))[2:(length(data$rain)+1)]
```

```
rain_day <- nrow(data %>% filter(month == 7, rain))
nrain_day <- nrow(data %>% filter(month == 7, !rain))
r_r <- nrow(data %>% filter(month == 7, rain, rain_tomo))
r_nr <- nrow(data %>% filter(month == 7, rain, !rain_tomo))
nr_r <- nrow(data %>% filter(month == 7, !rain, rain_tomo))
nr_nr <- nrow(data %>% filter(month == 7, !rain, !rain_tomo))
a1 <- r_r/rain_day
a3 <- nr_r/nrain_day
a2 <- r_nr/rain_day
a4 <- nr_nr/nrain_day
print(c(a1,a2,a3,a4))
```

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

(4)

$H_0 : a_0 = a_4$, $H_1 : a_0 \neq a_4$

We can use p_{00} and p_{11} to represent a_0 and a_4.

$\hat{p}_{00} \rightarrow^D N(\hat{p}_{00}, \frac{\hat{p}_{00}(1-\hat{p}_{00})}{n_0})$

$\hat{p}_{11} \to^D N(\hat{p}_{11}, \frac{\hat{p}_{11}(1-\hat{p}_{11})}{n_1})$

Because p_{00} and p_{11} are asymptotic independent, under H_0

$\hat{p}_{00} - \hat{p}_{11} \to^D N(0, \frac{\hat{p}_{00}(1-\hat{p}_{00})}{n_0} - \frac{\hat{p}_{11}(1-\hat{p}_{11})}{n_1})$

```
pnorm((a1-a4)/sqrt(a1*(1-a1)/rain_day+a4*(1-a4)/nrain_day))
```

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

We can reject H_0.

(5)

```
data$rain_tomo2 <- append(data$rain_tomo,c(NA))[2:(length(data$rain_tomo)+1)]

r_r_r <- nrow(data %>% filter(month == 7, rain, rain_tomo, rain_tomo2))
r_r_nr <- nrow(data %>% filter(month == 7, rain, rain_tomo, !rain_tomo2))
r_nr_r <- nrow(data %>% filter(month == 7, rain, !rain_tomo, rain_tomo2))
r_nr_nr <- nrow(data %>% filter(month == 7, rain, !rain_tomo, !rain_tomo2))
nr_r_r <- nrow(data %>% filter(month == 7, !rain, rain_tomo, rain_tomo2))
nr_r_nr <- nrow(data %>% filter(month == 7, !rain, rain_tomo, !rain_tomo2))
nr_nr_r <- nrow(data %>% filter(month == 7, !rain, !rain_tomo, rain_tomo2))
nr_nr_nr <- nrow(data %>% filter(month == 7, !rain, !rain_tomo, !rain_tomo2))

p000 <- r_r_r / (r_r_r + r_r_nr)
p001 <- r_r_nr / (r_r_r + r_r_nr)
p010 <- r_nr_r / (r_nr_r + r_nr_nr)
p011 <- r_nr_nr / (r_nr_r + r_nr_nr)
p100 <- nr_r_r / (nr_r_r + nr_r_nr)
p101 <- nr_r_nr / (nr_r_r + nr_r_nr)
p110 <- nr_nr_r / (nr_nr_r + nr_nr_nr)
p111 <- nr_nr_nr / (nr_nr_r + nr_nr_nr)

p00 <- a1
p01 <- a2
p10 <- a3
p11 <- a4
```

H_0: the first order model holds.

H_1: the first order model doesn't hold.

Use likelihood ratio test:

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

$$= 2\left\{\sum_{r=1}^{S}\sum_{s=1}^{S}\sum_{t=1}^{S} n_{rst}\log\hat{p}_{rst} - \sum_{r=1}^{S}\sum_{s=1}^{S}\sum_{t=1}^{S} n_{rst}\log\hat{p}_{st}\right\} = 2\sum_{r=1}^{S}\sum_{s=1}^{S}\sum_{t=1}^{S} n_{rst}\log\left(\frac{\hat{p}_{rst}}{\hat{p}_{st}}\right)$$

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

```
lambda <- (r_r_r*log(p000/p00) + r_r_nr*log(p001/p01) + r_nr_r*log(p010/p10) + r_nr_nr*log(p011/p11) + r
pchisq(lambda,2)
```

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

We can accept H_0. Higher order chain doesn't improve the fit of the data.