STAT 5703 Homework 2 Exercise 3

Shijie He(sh3975), Yunjun Xia(yx2569), Shuyu Huang(sh3967)

3/1/2020

Part 1

Let X_t be the precipication state at time t. Let state space be $S = \{0, 1\} = \{\text{"rainy day"}, \text{"no rain"}\}$. By first-order Markov chain model and stationarity,

$$
\begin{aligned}
a_1 = p_{00} = p(X_1 = 0|X_0 = 0) = p(X_{t+1} = 0|X_t = 0) \\
a_2 = p_{01} = p(X_1 = 1|X_0 = 0) = p(X_{t+1} = 1|X_t = 0) \\
a_3 = p_{10} = p(X_1 = 0|X_0 = 1) = p(X_{t+1} = 0|X_t = 1) \\
a_4 = p_{11} = p(X_1 = 1|X_0 = 1) = p(X_{t+1} = 1|X_t = 1)
\end{aligned}
$$

That is, a_1 is the probability that the actual day will be rainy given the previous day is also rainy; a_2 is the probability that the actual day will be rainy given the previous has no rain; a_3 is the probability that the actual day will have no rain given the previous is rainy; a_1 is the probability that the actual day will have no rain given the previous has no rain.

Part 2

Let the stationary distribution be $\pi = \{\pi_0, \pi_1\}$. Since we have $\pi^T \mathbf{T} = \pi^T$ and $\pi^T \mathbf{1}_S = \mathbf{1}_S$, so

$$
\begin{bmatrix} \pi_0 & \pi_1 \end{bmatrix} \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} = \begin{bmatrix} a_1\pi_0 + a_3\pi_1 & a_2\pi_0 + a_4\pi_1 \end{bmatrix} = \begin{bmatrix} \pi_0 & \pi_1 \end{bmatrix}
$$

$$
\begin{cases}
a_1\pi_0 + a_3\pi_1 = \pi_0 \\
a_2\pi_0 + a_4\pi_1 = \pi_1 \\
\pi_0 + \pi_1 = 1
\end{cases}
$$

By solving this system of equations, we have

$$
\begin{cases}
\pi_0 = \frac{a_3}{a_3+1-a_1} \\
\pi_1 = \frac{1-a_1}{a_3+1-a_1}
\end{cases}
$$

Therefore, the long-term probability of observing a rainy day in Central Park is $\pi_0 = \frac{a_3}{a_3+1-a_1}$.

Part 3

Note: I extracted the data of July of all years from CentralPark.csv. There are totally 119 years and each year has 31 days of precipitation data. So for this markov chain problem, I only consider 30 transitions for each year. That is, within one year, I only consider the transition state from (7/1 to 7/2) to (7/30 to 7/31).

```r
data <- read.csv('CentralPark.csv')
levels(data$NAME)
```

```
## [1] "NY CITY CENTRAL PARK, NY US"
```

```r
july_date <- c()
july_prcp <- c()
for (i in 1:nrow(data)){
  if (substr(data$DATE[i],1,1) == 7){
    july_date <- c(july_date, toString(data$DATE[i]))
    july_prcp <- c(july_prcp, data$PRCP[i])
  }
}
prcp <- data.frame(DATE = july_date, PRCP = july_prcp, RAIN = numeric(length(july_prcp)))
```

```r
for (i in 1:nrow(prcp)){
  if (prcp$PRCP[i] > 1.5){
    prcp$RAIN[i] = 0
  }
  else{
    prcp$RAIN[i] = 1
  }
}
```

```r
count00 <- 0
count01 <- 0
count10 <- 0
count11 <- 0

for (i in 1:(nrow(prcp)-1)){
  this_year <- str_sub(prcp$DATE[i],-2,-1)
  next_year <- str_sub(prcp$DATE[i+1],-2,-1)
  if (this_year == next_year){
      if (prcp$RAIN[i] == 0 && prcp$RAIN[i+1] == 0){
        count00 <- count00 + 1
      }
      else if (prcp$RAIN[i] == 0 && prcp$RAIN[i+1] == 1){
        count01 <- count01 +1
      }
      else if (prcp$RAIN[i] == 1 && prcp$RAIN[i+1] == 0){
        count10 <- count10 + 1
      }
      else if (prcp$RAIN[i] == 1 && prcp$RAIN[i+1] == 1){
        count11 <- count11 + 1
      }
  }
}
```

```r
c(count00, count01, count10, count11)
```

```
## [1]  257  608  605 2100
```

Therefore, for the historial Central Park data in July, we have

$$n_{00} = 257$$
$$n_{01} = 608$$
$$n_{10} = 605$$
$$n_{11} = 2100$$

where n_{rs} is the number of observations from precipitation state r to state s.
Then we can estimate the a_is using $a_i = \hat{p}_{rs} = \frac{n_{rs}}{n_{r\cdot}}$

$$\hat{a}_1 = \hat{p}_{00} = \frac{n_{00}}{n_{0\cdot}} = \frac{257}{257+608} = 0.2971098$$
$$\hat{a}_2 = \hat{p}_{01} = \frac{n_{01}}{n_{0\cdot}} = \frac{608}{257+608} = 0.7028902$$
$$\hat{a}_3 = \hat{p}_{10} = \frac{n_{10}}{n_{1\cdot}} = \frac{605}{605+2100} = 0.2236599$$
$$\hat{a}_4 = \hat{p}_{11} = \frac{n_{11}}{n_{1\cdot}} = \frac{2100}{605+2100} = 0.7763401$$

Part 4

We have the following hypothesis test:

$$H_0 : X_{t+1}|X_t = 1 \text{ and } 1 - X_{t+1}|X_t = 0 \text{ have the same distribution}$$
$$H_1 : \text{Otherwise}$$

We have the following distributions,

$$X_{t+1}|X_t = 1 = \begin{cases} 1 \text{ with prob. } p_{11} \\ 0 \text{ with prob. } p_{10} \end{cases}$$
$$1 - X_{t+1}|X_t = 0 = \begin{cases} 1 \text{ with prob. } p_{00} \\ 0 \text{ with prob. } p_{01} \end{cases}$$

where the probability boundary is $p_{11} + p_{10} = 1$ and $p_{00} + p_{01} = 1$. Then we can derive,

$$X_{t+1}|X_t = 1 \sim Binomial(n_{1\cdot}, p_{11})$$

$$1 - X_{t+1}|X_t = 0 \sim Binomial(n_{0\cdot}, p_{00})$$

This is a two-sample binomial test, then we can rewrite the hypothesis as follows:

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

where from part 3, $n_{0\cdot} = 257 + 608 = 865$ and $n_{1\cdot} = 605 + 2100 = 2705$. Also, $\hat{p}_{00} = \frac{257}{257+608} = 0.2971098$ and $\hat{p}_{11} = \frac{2100}{605+2100} = 0.7763401$.

So by normal approximation (since sample size is large), we have the test statistic:

$$TS = \frac{\hat{p}_{00} - \hat{p}_{11}}{\sqrt{\frac{n_{0\cdot}\hat{p}_{00}+n_{1\cdot}\hat{p}_{11}}{n_{0\cdot}+n_{1\cdot}}(1 - \frac{n_{0\cdot}\hat{p}_{00}+n_{1\cdot}\hat{p}_{11}}{n_{0\cdot}+n_{1\cdot}})(\frac{1}{n_{0\cdot}} + \frac{1}{n_{1\cdot}})}}$$

Then we can calculate the p-value:

```
p00 <- 0.2971098
p11 <- 0.7763401
p01 <- 0.7028902
p10 <- 0.2236599
n0 <- 865
n1 <- 2705
pp <- (n0*p00+n1*p11)/(n0+n1)
pnorm((p00-p11)/sqrt(pp*(1-pp)*(1/n0+1/n1)))
```

```
## [1] 3.034169e-148
```

Since the p-value is very small, we reject the null hypothesis. That is, the given two random variables are significantly different in Central Park data in July.

Part 5

We have the following hypothesis test:

H_0 : First order chain model is better than second order chain model.

H_1 : Otherwise.

By using the likelihood ratio test for testing a first order v.s. a second order chain model, the test statistic is

$$\begin{aligned}
\Lambda_n &= 2\{\ell(\hat{\mathbf{P}})_{\text{second order}} - \ell(\hat{\mathbf{P}})_{\text{first order}}\} \\
&= 2\{\sum_{u=0}^{1}\sum_{r=0}^{1}\sum_{s=0}^{1} n_{urs}\log(\hat{p}_{urs}) - \sum_{r=0}^{1}\sum_{s=0}^{1} n_{\cdot rs}\log(\hat{p}_{rs})\} \\
&= 2\sum_{u=0}^{1}\sum_{r=0}^{1}\sum_{s=0}^{1} n_{urs}\log(\frac{\hat{p}_{urs}}{\hat{p}_{rs}})
\end{aligned}$$

```
count000 <- 0
count001 <- 0
count010 <- 0
count011 <- 0
count100 <- 0
count101 <- 0
count110 <- 0
count111 <- 0

for (i in 1:(nrow(prcp)-2)){
  this_year <- str_sub(prcp$DATE[i],-2,-1)
  next_year <- str_sub(prcp$DATE[i+1],-2,-1)
  one_more_year <- str_sub(prcp$DATE[i+2],-2,-1)
  if (this_year == next_year && next_year == one_more_year){
      if (prcp$RAIN[i] == 0 && prcp$RAIN[i+1] == 0 && prcp$RAIN[i+2] == 0){
        count000 <- count000 + 1
      }
      else if (prcp$RAIN[i] == 0 && prcp$RAIN[i+1] == 0 && prcp$RAIN[i+2] == 1){
```

```r
        count001 <- count001 + 1
      }
      else if (prcp$RAIN[i] == 0 && prcp$RAIN[i+1] == 1 && prcp$RAIN[i+2] == 0){
        count010 <- count010 +1
      }
      else if (prcp$RAIN[i] == 0 && prcp$RAIN[i+1] == 1 && prcp$RAIN[i+2] == 1){
        count011 <- count011 + 1
      }
      else if (prcp$RAIN[i] == 1 && prcp$RAIN[i+1] == 0 && prcp$RAIN[i+2] == 0){
        count100 <- count100 + 1
      }
      else if (prcp$RAIN[i] == 1 && prcp$RAIN[i+1] == 0 && prcp$RAIN[i+2] == 1){
        count101 <- count101 + 1
      }
       else if (prcp$RAIN[i] == 1 && prcp$RAIN[i+1] == 1 && prcp$RAIN[i+2] == 0){
        count110 <- count110 + 1
      }
      else if (prcp$RAIN[i] == 1 && prcp$RAIN[i+1] == 1 && prcp$RAIN[i+2] == 1){
        count111 <- count111 + 1
      }
  }
}
```

```r
p000 <- count000/(count000+count001)
p001 <- count001/(count000+count001)
p010 <- count010/(count010+count011)
p011 <- count011/(count010+count011)
p100 <- count100/(count100+count101)
p101 <- count101/(count100+count101)
p110 <- count110/(count110+count111)
p111 <- count111/(count110+count111)
c(count000,count001,count010,count011,count100,count101,count110,count111)
```

```
## [1]   66  182  123  467  184  401  466 1562
```

```r
c(p000,p001,p010,p011,p100,p101,p110,p111)
```

```
## [1] 0.2661290 0.7338710 0.2084746 0.7915254 0.3145299 0.6854701 0.2297830
## [8] 0.7702170
```

```r
obs_TS <- 2 * (count000 * log(p000/p00) + count001 * log(p001/p01) +
                 count010 * log(p010/p10) + count011 * log(p011/p11) +
                 count100 * log(p100/p00) + count101 * log(p101/p01) +
                 count110 * log(p110/p10) + count111 * log(p111/p11))
obs_TS
```

```
## [1] 3.236928
```

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

Then we can calculate the p-value,

```r
pchisq(obs_TS, df = 2)
```

```
## [1] 0.801797
```

Since the p-value is 0.801797 which is very large, we fail to reject the null hypothesis. That is, a higher order chain model will not improve the fit of the data.