### Question 1

#### (a)
# read data

d <- read.table(file = "http://www.public.iastate.edu/~maitra/stat579/mRNA.dat",
    header = F, sep = ",")

# select those genes with maximum association >5.3
selected.indices <- apply(d, 2, max)>5.3
d.sel <- d[, selected.indices]

#### (b)

# store row and column names, and selected column names

col.names <- 1:dim(d)[2]
row.names <- 1:dim(d)[1]
selected.col.names <- col.names[selected.indices]

# identify most strongly associated marker for each remaining gene
m <- as.vector(apply(d.sel, 2, which.max))

#### (c)

# create object with gene, best marker combinations
o <- data.frame(gene=selected.col.names, marker = m)

## getting the ten strongest associations

# get all the strongest associations
strong.ass <- as.vector(apply(d.sel, 2, max))

o2 <- o[order(strong.ass, decreasing = T)[1:10], ] # provide the 10 strongest
# genes and markers

write.table(o2, file = "gene.table") #using write.table to write to file

## output of file follows:

"gene" "marker" "strong.ass"
"V4726"  4726  51 21.96
"V378" 378  23 14.79
"V7813"  7813  69 13.49
"V1568"  1568  8 13.38
"V3166"  3166  51 11.96
"V4459"  4459  23 11.46
"V8958"  8958  26 11.45
"V4387"  4387  75 9.954
"V9339"  9339  19 9.852
"V4715"  4715  19 9.65

#### (d) table(o$m) # tabulate the frequency of occurrence of each strong marker

```
   1  3  5  6  8 12 14 15 17 19 21 23 24 25 26 27 29 30 31
   3  1  2  3  4  2  4  4  3  1  9  1  5  1  4  3  1  1  2  5
   32  33  34  35  37  38  39  40  41  42  43  45  46  48  49  50  51  52  53  54
   1  3  8  1  1  5  1  1  3  1  2  1  2  2  3  1  4  6  2  2
   57  58  59  60  61  62  64  65  67  69  70  72  75  77  78  79  81  82  83  86
   3  1  1  1  7  1  1 11 1 1  1  1  4  1  2  1  1  1  1  3  2
   88  89  91  92  96  97 100
   2  1  1  3  3  1  1
```
## (a)

```r
# read data
s <- read.table(file = "http://www.public.iastate.edu/~maitra/stat579/SEAir.dat", header = T)

# include date as per calendar
x <- tapply(rep(1, nrow(s)), list(s$mo, s$yr), FUN = sum) # provides matrix of year/month with number of days
cal.dt <- unlist(apply(as.matrix(as.vector(x)), 1, FUN = seq)) # provides calendar date

xx <- vector(mode = "character", length = nrow(s)) # create empty vector of character strings
xx[(s$mo == 1) | (s$mo == 2) | ((s$mo == 3) & (cal.dt < 21)) | ((s$mo == 12) & (cal.dt >= 22))] <- "Winter"
xx[(s$mo == 4) | (s$mo == 5) | ((s$mo == 6) & (cal.dt < 22)) | ((s$mo == 3) & (cal.dt >= 21))] <- "Spring"
xx[(s$mo == 7) | (s$mo == 8) | ((s$mo == 9) & (cal.dt < 21)) | ((s$mo == 6) & (cal.dt >= 22))] <- "Summer"
xx[xx == ""] <- "Fall"

swseas <- cbind(s, season = xx)
```

```r
table(xx)
```

```
   Fall Spring Summer Winter
   736    744    728    714
```

## (b)

```r
# report quantiles (0.25, 0.50, 0.75) for admissions in each year
dow.stats <- tapply(s$admyng, s$dow, quantile, probs=c(0.25, 0.50, 0.75))

# plot statistics as line plot
v <- as.vector(unlist(dow.stats))
plot(x=1:7, y=v[seq(from=1, to=length(v), by=3)+1], xlab="Day of Week", ylab="Number of Admissions", type="l", ylim=c(2,10))
lines(x=1:7, y=v[seq(from=1, to=length(v), by=3)], lty=2)
lines(x=1:7, y=v[seq(from=1, to=length(v), by=3)+2], lty=2)
```

## (c)

```r
# remove leap days
s1 <- s[!((cal.dt == 29) & (s$mo == 2)),]

# use array to create new array and then aperm to permute the indices
s2 <- aperm(array(as.vector(t(s1)), dim = c(ncol(s), 365, 8)), c(2, 1, 3))
```

## (d)

```r
all.means <- apply(s2[, -(1:4), ], c(2, 3), mean, na.rm = T)
```
rownames(all.means) <- names(s1)[-(1:4)]
colnames(all.means) <- as.character(1987:1994)

all.means

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>admyn</td>
<td>5.0657</td>
<td>5.2329</td>
<td>6.0027</td>
<td>5.6027</td>
<td>5.8959</td>
</tr>
<tr>
<td>asyng</td>
<td>2.3945</td>
<td>2.4739</td>
<td>2.6767</td>
<td>2.6684</td>
<td>2.4136</td>
</tr>
<tr>
<td>pm25</td>
<td>19.7255</td>
<td>19.9038</td>
<td>18.0398</td>
<td>15.5188</td>
<td>15.8613</td>
</tr>
<tr>
<td>temp</td>
<td>53.8904</td>
<td>52.6493</td>
<td>52.7945</td>
<td>52.6356</td>
<td>53.0520</td>
</tr>
<tr>
<td>o3max8</td>
<td>0.0323</td>
<td>0.0313</td>
<td>0.0264</td>
<td>0.0276</td>
<td>0.0335</td>
</tr>
<tr>
<td>so2avg</td>
<td>0.0065</td>
<td>0.0067</td>
<td>0.0067</td>
<td>0.0088</td>
<td>0.0095</td>
</tr>
<tr>
<td>coavg</td>
<td>2.2230</td>
<td>1.9624</td>
<td>2.0209</td>
<td>1.7519</td>
<td>1.7900</td>
</tr>
<tr>
<td>stagno</td>
<td>9.0939</td>
<td>8.4716</td>
<td>8.5489</td>
<td>4.2197</td>
<td>4.8939</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1992</th>
<th>1993</th>
<th>1994</th>
</tr>
</thead>
<tbody>
<tr>
<td>admyn</td>
<td>6.0411</td>
<td>7.5233</td>
<td>6.5643</td>
</tr>
<tr>
<td>asyng</td>
<td>2.7753</td>
<td>3.2712</td>
<td>2.7780</td>
</tr>
<tr>
<td>pm25</td>
<td>15.4917</td>
<td>15.1201</td>
<td>12.2066</td>
</tr>
<tr>
<td>temp</td>
<td>54.5452</td>
<td>52.4849</td>
<td>53.7123</td>
</tr>
<tr>
<td>o3max8</td>
<td>0.0326</td>
<td>0.0287</td>
<td>0.0323</td>
</tr>
<tr>
<td>so2avg</td>
<td>0.0097</td>
<td>0.0094</td>
<td>0.0072</td>
</tr>
<tr>
<td>coavg</td>
<td>1.8099</td>
<td>1.6952</td>
<td>1.3918</td>
</tr>
<tr>
<td>stagno</td>
<td>5.9605</td>
<td>6.2197</td>
<td>5.0797</td>
</tr>
</tbody>
</table>

#Comment: Average hospitalization for respiratory diseases largely increased over all years, but considerably in 1989 and in 1993. There was a similar trend in the average numbers of hospitalization for asthma. Mean 24-hour averages of concentration of fine particulate matter decreased all through, with substantial increases reported in 1989 and 1994. Similar trends were noted for average daily concentrations of carbon monoxide. Average daily temperatures remained steady on the whole. Averages of maximum ozone concentrations fluctuated from year to year. The air stagnation index was considerably lower after the first three years, but did pick up again towards the end.

## (e)

summary(lm(admyng ~ pm25 + temp + o3max8 + so2avg + coavg + stagno, data = s))

Call:
lm(formula = admyng ~ pm25 + temp + o3max8 + so2avg + coavg + stagno, data = s)

Residuals:
   Min     1Q Median     3Q    Max
-7.4062 -1.8840 -0.2893  1.6549 10.2975

Coefficients:        Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.189800  0.535531  20.895  < 2e-16 ***
  pm25       0.019627   0.010817   1.814   0.0698 .
   temp     -0.122000   0.009603  -12.704  < 2e-16 ***
o3max8     0.353405   5.326815   0.066   0.9471
   so2avg    77.375755  17.414474  4.443  9.47e-06 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.606 on 1614 degrees of freedom

(1301 observations deleted due to missingness)
Multiple R-squared: 0.1543,   Adjusted R-squared: 0.1511
F-statistic: 49.07 on 6 and 1614 DF,  p-value: < 2.2e-16

## Comment: Hospitalization for respiratory diseases has significance
## linear associations with the daily temperature (p > 2e-16) and the
## 24-hour average of sulfur dioxide concentration (p = 9.47e-6).