INSTITUTE AND FACULTY OF ACTUARIES

CURRICULUM 2019

SPECIMEN SOLUTIONS

Subject CS2B – Risk Modelling and Survival Analysis

Question 1

(i) Clustering algorithms require a measure of the distance between each observation on each feature. If the data are not scaled, the weight given to a feature by the algorithm will depend on the units of measurement, as the absolute distances between observations will vary more for some features than others.

We can illustrate this with the data on the population density and the proportion of adult males working in agriculture.

```
> summary(machinelearn[2:4])

   PopDensity          SexRatio        PropMAgric
   Min.   :  0.0300   Min.   :0.6230   Min.   :0.0070
   1st Qu.:  0.2310   1st Qu.:0.9350   1st Qu.:0.2060
   Median :  0.3375   Median :0.9800   Median :0.4345
   Mean   :  5.7153   Mean   :0.9755   Mean   :0.3823
   3rd Qu.:  0.7325   3rd Qu.:1.0183   3rd Qu.:0.5503
   Max.   :219.6570   Max.   :1.5540   Max.   :0.7700
```

PopDensity varies from 0.03 to 219.66, whereas PropMAgric varies only from 0.007 to 0.770. [6]

(ii) *A VALID SOLUTION*

```
> data_z <- as.data.frame(lapply(machinelearn[2:7],
     scale))
> cluster_z <- kmeans(data_z[1:6], 6)
> cluster_z$size


[1]  70  12 141 257  35  73

  > cluster_z$centers

   PopDensity   SexRatio PropMAgric PropMMining PropFManuf PropFDomServ
1 -0.18880421  0.5087740 -0.8760983   2.3154198 -0.3193629 -0.7524701
2  6.10859337 -0.4442488 -1.8425129  -0.5423765  0.8846691  0.3678774
3 -0.12072467 -0.3614010 -0.2812440  -0.2734648 -0.1970575  0.4691578
4 -0.22262935  0.4970005  0.8208959  -0.3508762 -0.4419909 -0.2128162
5  0.58006280 -1.9655087 -1.2771535  -0.4514210 -0.2515955  2.6259378
6 -0.08426169 -0.5241366 -0.5914717  -0.1511976  2.2181094 -0.7548896
```

There are other valid solutions, which will vary according to the seed used for the initial assignment to clusters. [8]

(iii) Cluster 1 is composed of mining districts.

Cluster 2 is a small number of urban districts with a high proportion of females in employment.

Cluster 3 consists of areas close to the average on all measures, without obvious distinguishing features.

Cluster 4 is the largest cluster, forming close to half of the 588 areas. It consists of areas with a rather high proportion of males in agriculture and a low population density.

Cluster 5 consists of areas with a high proportion of females compared with males, and those females are working in domestic service.

Cluster 6 consists of areas where a high proportion of females were employed in manufacturing.

Answers which are sensible given the clusters actually identified in part (ii) will be given credit. Some mention of the relative sizes of the clusters is required for full credit. [6]

(iv)
```
> machinelearn$cluster <- cluster_z$cluster
> aggregate (data = machinelearn, DRTuberculosis ~
        cluster, mean)

  cluster DRTuberculosis
  1       2.313857
  2       3.201000
  3       2.389794
  4       2.112284
  5       2.429371
  6       2.467863

  > aggregate (data = machinelearn, DRLung ~
        cluster, mean)

  cluster DRLung
  1       3.054143
  2       5.183750
  3       2.686007
  4       2.421432
  5       3.149829
  6       3.388260
```

[5]

(v) The clustering algorithm has identified a small number of districts with very high death rates from diseases of the lungs (cluster 2).

It has also identified a group of agricultural districts with low death rates from both causes (cluster 4).

But the differences in the mean death rates between the other clusters are small.

It might be worth trying a different number of clusters to see whether the results are more satisfactory.

Techniques to measure within-cluster homogeneity could be used to assess the validity of the results.

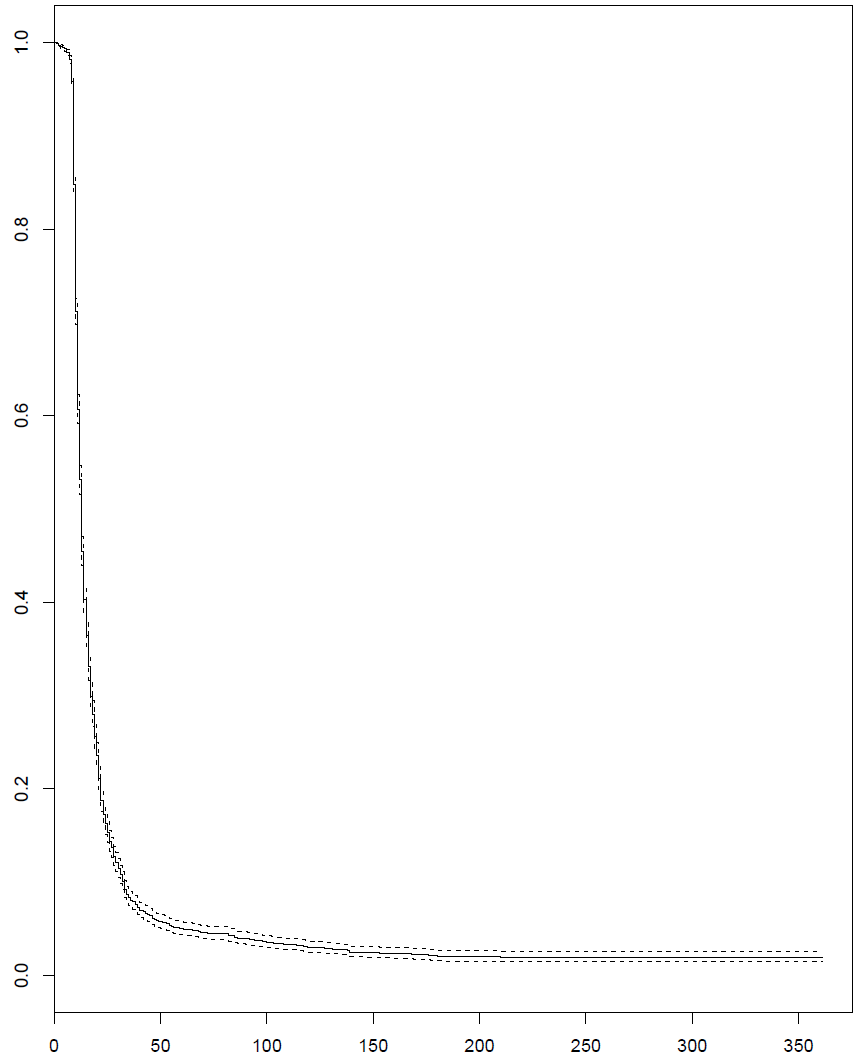
The answers to this part will depend on the actual clusters identified, but should be consistent with the latter.

[5]
[Total 30]

Question 2

```
> library(survival)
```

(i)
```
> fit <- survfit(Surv(survival$DUR,
     survival$EVENT ~ 1)
> plot(fit)
```



[6]

(ii)
```
> coxfit1 <- coxph(Surv(survival$DUR, survival$EVENT) ~
      AGE + URBAN + POOREST + POOR + MIDDLE + RICH + YEAR
      + LOWER, data = survival)
> coxfit1

            coef    exp(coef) se(coef)  z       p
AGE     -0.01951   0.98068  0.00470 -4.15 3.3e-05
URBAN   -0.05431   0.94714  0.04487 -1.21    0.23
POOREST  0.04304   1.04398  0.06961  0.62    0.54
POOR     0.09119   1.09548  0.05916  1.54    0.12
MIDDLE  -0.02115   0.97907  0.05509 -0.38    0.70
RICH     0.01753   1.01769  0.05517  0.32    0.75
LOWER   -0.08929   0.91458  0.03478 -2.57    0.01
YEAR     0.00102   1.00102  0.00194  0.53    0.60

Likelihood ratio test=31.66  on 8 df, p=1e-04
n= 4091, number of events= 3776
```

The only covariates that have a significant effect on the hazard of proceeding to the first child are AGE and LOWER.

As age at cohabitation increases by one year, the hazard of giving birth to the first child reduces by 1.932%.

Women with lower levels of education have a hazard of giving birth to their first child 8.542% lower than do women with higher levels of education. [14]

(iii) We can start by trying a model with just AGE and LOWER, as these were the two most significant covariates in part (ii).

```
> coxfit2 <- coxph(Surv(survival$DUR, survival$EVENT) ~
      AGE + LOWER, data = survival)
> coxfit2

          coef    exp(coef) se(coef)  z       p
AGE   -0.01972   0.98048  0.00456 -4.33 1.5e-05
LOWER -0.06888   0.93343  0.03353 -2.05    0.04

Likelihood ratio test=20.45 on 2 df, p=4e-05
n= 4091, number of events= 3776
```

Comparing the likelihood ratio between this model and the model in part (ii) we have 31.66 – 20.45 = 11.21 with 6 degrees of freedom. Since 11.21 < 12.59 we do not reject the null hypothesis that the model with just 2 covariates is as good a fit to the data as the model with 8 covariates estimated in part (ii).

However, looking at the results from part (ii) we might try adding the next most significant covariate, POOR.

```
> coxfit3 <- coxph(Surv(survival$DUR, survival$EVENT) ~
      AGE + LOWER + POOR, data = survival)
> coxfit3
```

```
          coef    exp(coef) se(coef)  z       p
AGE   -0.01936   0.98083  0.00456 -4.25 2.2e-05
LOWER -0.07341   0.92922  0.03358 -2.19   0.029
POOR   0.09257   1.09699  0.03905  2.37   0.018

Likelihood ratio test=25.98 on 3 df, p=1e-05
n= 4091, number of events= 3776
```

Comparing the likelihood ratio between this model and the previous model with just AGE and LOWER as covariates we have 25.98 – 20.45 = 5.53 with 1 degree of freedom. Since 5.53 > 3.84 we reject the null hypothesis that the model with just 2 covariates (AGE and LOWER) is as good a fit to the data as the model with 3 covariates (AGE, POOR and LOWER). [15]

[Total 35]

Question 3

(i) *A VALID SOLUTION*

```
> MA1=function(N,b,sigma){
     if(N>0 & sigma >0 & b<=1 & b>=-1)
     {
          e=rnorm(n=N+1,sd=sigma)
          e[1]=0
          y=e[2:(N+1)]-e[1:N]
     return(list(b = b, sigma = sigma, y = y))
     }
     else
     {
     cat("Error")
     }
     }
```

Other valid solutions are possible. [10]

(ii) > `par(mfrow=c(2,2))`

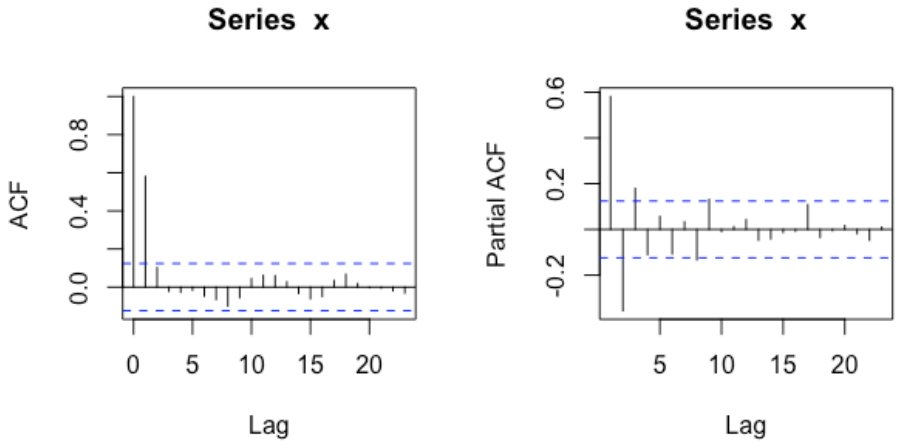
```
> plot.ts(MA1(100,0.5,2)$y)
> plot.ts(MA1(100,0.5,2)$y)
> plot.ts(MA1(100,0.5,2)$y)
> plot.ts(MA1(100,0.5,2)$y)
```
[4]

(iii) *See code submitted.* [3]

(iv) The first line sets the seed of the random number generator to a fixed value equal to that of "`num`".

The second line generates a random sequence of $n = 250$ consecutive observations from the time series model ARMA(1,1) with parameters $a = 0.3$, beta = 0.6 and sigma=0.5. [4]

(v) The following output is generated for `num = 1234`



ACF indicates the exponential decay of its values with a possibly unusual spike at lag 1. So the MA order could be 1.

PACF has similar patterns with possibly two distinct spikes so the AR order could be 1 or 2? [7]

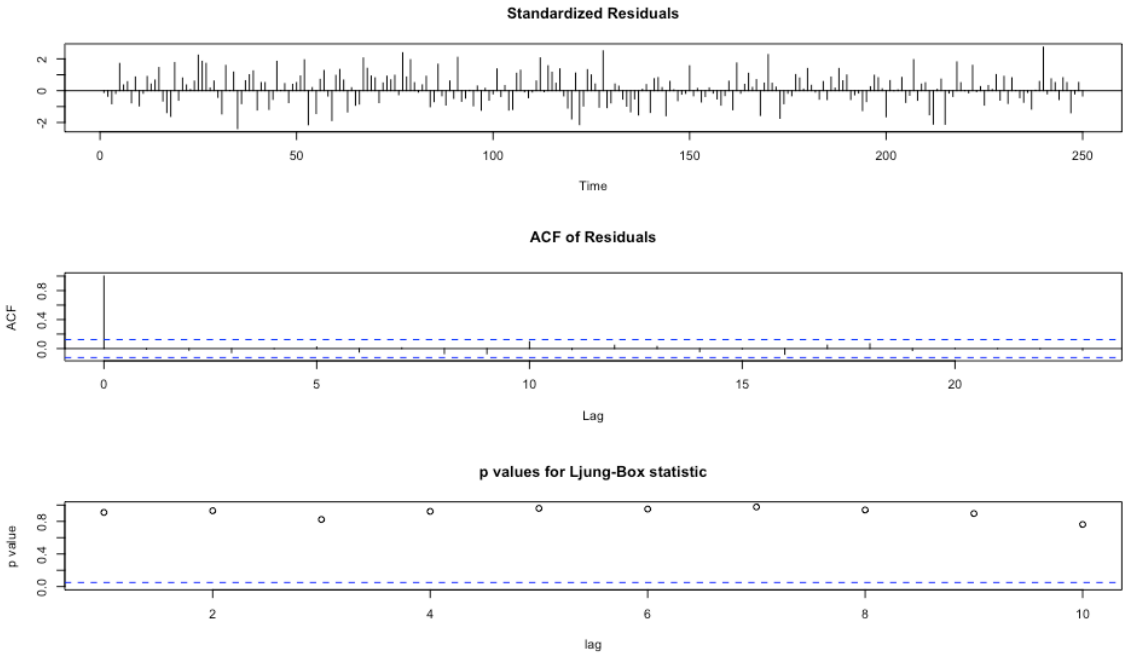
(vi) Try

```
> Model <- arima(x,c(2,0,2),include.mean=F)
> Model

Call:
arima(x = x, order = c(2, 0, 2), include.mean = F)
Coefficients:
         ar1     ar2     ma1     ma2
      -0.4338  0.1984  1.3503  0.4230
s.e.   0.3732  0.0924  0.3731  0.2829

sigma^2 estimated as 0.2746:  log likelihood = -193.72,  aic = 397.44.
```

One could run the diagnostics for this model as

```
> tsdiag(Model)
```

the three plots below are generated from R where the Ljung-Box test. p values are also listed:



Looking that the standard errors for the `ar1` parameter there is some indication that the model ARMA(2,2) could be over parameterised. [7]

[Total 35]