Backcalculating HIV incidence and predicting AIDS in Australia, Cambodia and Vietnam

The aim of today's practical is to give you some 'hands-on' experience with a nonparametric method for backcalculating HIV incidence from observed AIDS incidence data, assuming a known incubation distribution. We will use Becker *et al*'s (1991) EMS approach as described in today's lecture and investigate how well the method performs at predicting AIDS incidence.

Three datasets are available for this practical: quarterly Australian AIDS incidence from 1981 to 1998, and yearly AIDS incidence from 1993 to 1997 in Cambodia and Vietnam.

For Australia, we will backcalculate estimates of HIV incidence from cases of AIDS observed to the end of 1993, then project from the estimated infection curve to predict future AIDS incidence for the next five years. Since we have observed AIDS data to the end of 1998, we can compare the AIDS predictions with the values actually observed.
Only a few years of data are available for Vietnam and Cambodia, where the HIV epidemics are still in relatively early, albeit explosive, phases. For comparison with the Australian data, we will use the EMS backcalculation method to reconstruct past HIV incidence from the yearly AIDS data then predict future AIDS incidence.

Throughout, remember to use the *S-PLUS* help facility `help.start()`.

Australia

1. The file `ozaids.dat` contains quarterly AIDS incidences in Australia from the first quarter of 1981 to the last quarter of 1998. The data have been adjusted for reporting delays, although not for under-reporting as reporting is believed to be relatively complete.

   The variables are:

`time`	year and quarter of diagnosis
incidence by state:	
`nsw`	New South Wales
`vic`	Victoria
`qld`	Queensland
`sa`	South Australia
`wa`	Western Australia
`other`	other Australian States and Territories
incidence by selected risk groups:	
`fhet`	female heterosexual contact
`idu`	injecting drug use contact
`mhomo.idu`	male homosexual contact, with or without injecting drug use
`Total`	Total Australian incidence

Note that data on risk groups within regions are not available.

(a) *Plot quarterly Australian AIDS incidence over time, firstly for total incidence and then by the major regions and risk groups.*

To read the data into *S-PLUS* all we need is the usual command `read.table()`

```
> aids <- read.table("ozaids.dat", header = T, row.names = NULL)
> aids$yrqtr <- seq(81.25, 99.00, 0.25)  ##change the times to numerical
> trellis.device()
> names(aids)
 [1] "nsw"       "oth"       "qld"       "sa"        "vic"       "wa"
 [7] "fhet"      "idu"       "mhomo.idu" "Total"
```

These are the names of the columns in your data.frame called ”aids”.

There are a number of ways of plotting the data. For the column by column approach you can use `plot` or `xyplot`:

```
> plot(aids$yrqtr, aids$Total, type = "l", ylab = "Total", xlab = "Time")
```

will plot lines.

Similarly for `xyplot`

```
> xyplot(aids$Total ~ aids$yrqtr, type ="b") ## points and lines
                                             ## note the change in xy
> xyplot(aids$Total ~ aids$yrqtr, type ="l", ylab = "Total",
+ xlab = "Time")
```

For plotting incidence by regions and risk groups on one set of axes, set up the names for the legend to ensure that the lines on the plot can be determined by sight

```
> legend.names <- names(aids[,2:ncol(aids)])
```

The following will plot AIDS incidence for the different subpopulations

```
> plot(aids$yrqtr, aids$Total, type = "n", xlab = "Time",
+  ylab = "quarterly AIDS incidence")
> for(i in 2:ncol(aids)){
+   lines(aids$yrqtr, aids[,i], lty = 11 - i)
}
> legend(81, 250, legend.names, lty = 10:1)
```

For Trellis to handle the data in the same way, we need to get the data into the right form. To begin, we take the variable names from data.frame ”aids” and replicate them to create factors for Trellis to read

```
> nams <- rep(names(aids[2:ncol(aids)]), rep(nrow(aids), 10))
```

Then we need to vectorize the ”numeric” part of the data.frame

```
> vec <- c(as.matrix(aids[,2:ncol(aids)]))
```

Now put these into a data.frame for plotting

```
> aids.tr <- data.frame(time = rep(aids$yrqtr, 10), aids = vec,
+           group = nams)
> xyplot(aids.tr$aids ~ aids.tr$time | aids.tr$group, type = "l")
```

Note that the plots are all on the same scale. This is not necessarily a pre-requisite of Trellis and it can be helpful allow the plots to choose their own scales

```
> xyplot(aids.tr$aids ~ aids.tr$time | aids.tr$group, type =
+ "l", scales = list(relation = "free"))
```

Trellis uses `loess()` to nonparametrically smooth the data in each ”panel”

```
xyplot(aids.tr$aids ~ aids.tr$time | aids.tr$group, type ="l",
        scales = list(relation = "free"), panel = function(x, y, ...)
{
        panel.xyplot(x, y, type = "l")
        panel.loess(x, y, ...)
})
```

Note that you can decrease the degree of smoothing using a smaller “span” (the default is span = 2/3).

(b) *Comment on the observed trends in AIDS incidence in the different subgroups of the population and for Australia as a whole. When did AIDS incidence peak in Australia?*

What is the most likely explanation of the decrease in the total number of AIDS cases observed since 1994?

2. Weibull distributions

The EMS backcalculation method requires an assumed distribution for the incubation period.

The *Weibull distribution function*

$$P(T \leq t) = F(t) = 1 - \mathrm{e}^{(-bt)^c}$$

for $t > 0$, and scale and index parameters $b > 0$ and $c > 0$, has been widely used in backcalculation. For $c > 1$ the hazard of progression to AIDS is monotone increasing with time since infection.

The incubation distribution for HIV / AIDS depends on age and other factors, especially available treatments, and is widely held to have a median time to progression of about 10 years. We initially assume stationary Weibull distributions with index parameter c = 2.516 and median times to progression of 6, 8, 10 and 12 years.

You can find the scale parameters b corresponding to the median times as follows. Set the parameter c

```
> cc <- 2.516
```

The median survival times are (up to 20 years)

```
> medTime <- seq(6,20, 2)
```

Then b is found by

```
> b <- exp(log(-log(1 - 0.5))/2.516)/medTime
```

Plot the different cumulative distribution functions for the different median survival times

```
> time <- seq(0, 20, 0.2)
> plot(time, 1 - exp(-(b[[1]]*time)^cc), type = "n", xlab =
+ "Time", ylab = "Cumulative Probability")
> for(i in 1:length(b)){
+ lines(time, 1 - exp(-(b[i]*time)^cc), lty = i + 1, col = i + 1)
+ }
```

Then give the matching legend for the plot

```
> legend(0, 1.0, paste("Med. Surv. Time", seq(6, 20, 2)),lty = 2:9,
+  col = 2:9)
```

For each distribution, what is the probability of remaining AIDS free after 5 years? 10 years? 15 years?

3. The EMS backcalculation program

Open a second window within the directory you are working in (but do not quit *S-PLUS*).

Check that you have a dataset called `oz.dat` as well as files called `lynems.f` and `lynems.inc`.

`oz.dat` contains the total quarterly Australian AIDS data as input for the 'stand-alone' backcalculation program `lynems.f`

The program will prompt you with a series of questions and choices of input. I suggest that in the first instance you follow the attached handout which runs through the program and backcalculates Australian HIV incidence estimates using AIDS data to the end of 1993, a Weibull median time to AIDS of 12 years, a smoothing window of $k = 2$ (a low degree of smoothing because it is quarterly data), and predicts AIDS incidence to the end of 1998. The annotated output briefly explains the main choices to be made.

Compile the program by typing

```
f77 lynems.f
```

Ignore the various warnings (!) then type

```
a.out
```

to run the program.

`lynems.f` produces three output files, where you specify the `filename` (`ozout` in the attached example):

`ozout.lis` records the questions, your inputs and output printed to the screen

`ozout.cub` contains the cdfs and pdfs for the chosen incubation distribution, and

`ozout.dat` contains columns for the time intervals, the observed AIDS data, estimated HIV incidence, AIDS predictions, then various further HIV estimates and AIDS predictions resulting from options selected in the program, and residuals.

(a) Now read `ozout.dat` into *S-PLUS*. This requires names for each column of data. If names are not given, *S-PLUS* will choose defaults such as 'V1', 'V2', etc. The following adds names to the columns of the data.frame

```
> lyn <- read.table("ozout.dat", header = F) ## read in with defaults
> names(lyn) <-  c("Time","ObsAids","HIVest","ExpAids",
+   paste("other", 1:8, sep = ""), "Residuals"))
> attach(lyn)
```

(b) *Plot the estimated HIV incidence curve for Australia*

```
> plot(Time, HIVest, type = "n")
> lines(Time, HIVest, lty = 1)
```

The estimated HIV incidence from about 1986 is close to zero here. Is this plausible? When did the HIV epidemic peak in Australia?

(c) Add the observed AIDS counts to the plot. Note that the Observed Aids are the actual AIDS counts until the end of 1993, then the values are predicted. Therefore we must use the actual Totals from `oz.dat`, plus the 1979 and 1980 data from `ozout.dat`

```
> points(Time, c(ObsAids[1:8], aids$Total), pch = "o")
```

The lines for the Expected Aids are the first 55 points i.e. from 1979-1993

```
> lines(Time[1:55], ExpAids[1:55], lty = 2)
```

and adding lines for the predicted points i.e. 1993-1998

```
> lines(Time[55:80], ExpAids[55:80], lty = 2, lwd = 2)
> abline(v = 92.75, lty = 3)
```

The legend or key can be added very easily knowing the names and the parameters used in the plotting

```
> key(x = 94, 850,text = list(c("HIV est.", "Exp. Aids",
+ "Pred. Aids", "End of 1993"), adj = 1, cex = 0.75), lines =
+ list(lty = c(1,2,2,3), lwd = c(1,1,3,1)))
```

How well do the observed and expected AIDS values agree?

How well does the model predict AIDS incidence over the period 1994–1998?

(d) **Performance measures**

The chi-squared statistic $\sum(O-E)^2/E$ gives a comparative measure of the goodness of fit. The *prediction performance* of the models can be assessed informally by plotting and comparing the observed and predicted incidence patterns as above.

Alternatively, a simple quantitative method for assessing prediction performance is one based on a suggestion by Bacchetti (1995). For each time period (here a quarter) divide the predicted number of diagnoses by the observed number, subtract one, then multiply by 100% to obtain a percentage error. Average the absolute values of the percentage errors to obtain an overall performance measure for the model, called the *mean absolute percentage error*. Small values suggest better performance.

Find the mean absolute percentage error for your fitted model.

(e) Run `lynems.f` again, this time without smoothing (choose bandwidth $k = 1$). *Remember to choose a different filename, or delete the old one* `ozout`.

Compare the estimated HIV incidence curves using smoothing ($k = 3$) and no smoothing ($k = 1$). How sensitive is the peak in infections to the degree of smoothing assumed? How do the goodness of fits compare?

Estimates of the total number of HIV infected individuals are of interest and these are given by areas under the curve. Since the estimates near the endpoint of 1993 are increasingly uncertain, *compare estimates of the total number of HIV infections in Australia to the end of 1991*.

(f) **Investigating treatment effects**

The files `treat.cdf` and `treat.pdf` contain cdfs and pdfs for Brookmeyer and Liao's 'fixed treatment initiation time' two-stage incubation distribution. Obtain HIV estimates and AIDS predictions for Australia using this non-stationary incubation distribution. (*Note that you should choose the program option that allows you to read the incubation pdfs and cdfs from a file.*)

Vietnam and Cambodia

The HIV / AIDS epidemic in Vietnam is primarily amongst mostly male injecting drug users, whereas HIV in Cambodia is spread primarily via heterosexual contact.

The observed annual AIDS counts from 1989 to 1997 are given in the files `cambodia.dat` and `vietnam.dat`. Note that AIDS surveillance data from these countries suffer massive under-reporting and reporting delays and that the true incidence is believed to be much higher.

1. Read the data into *S-PLUS*.

   To plot the points on the same axes, do the following

```
> attach(asia)
> plot(year, Vietnam, type = "n")
> points(year, Vietnam, pch = "V")
> points(year, Cambodia, pch = "C")
> detach()
```

   If you prefer to use Trellis, you might like to try the following; otherwise go to 2

```
> asia.tr <- data.frame(year = rep(year, 2), country =
+ c(Cambodia, Vietnam),  group = rep(c("Cambodia", "Vietnam"),rep(6, 2)))
```

```
> asia.tr
   year country    group
 1 1992       0 Cambodia
 2 1993       1 Cambodia
 3 1994      14 Cambodia
```

```
 4 1995      91 Cambodia
 5 1996     300 Cambodia
 6 1997     572 Cambodia
 7 1992       0  Vietnam
 8 1993     106  Vietnam
 9 1994     116  Vietnam
10 1995     201  Vietnam
11 1996     380  Vietnam
12 1997     669  Vietnam

> attach(asia.tr)
```

Plot the data using `xyplot()`

```
> xyplot(country ~ year, groups = group, type = "b",
+ panel = panel.superpose)
```

To put the symbols on the plot we can get at the parameters that Trellis uses in its plotting

```
> sps <- trellis.par.get("superpose.symbol")
> sps$pch[1:2] <- c("C","V")
```

Then plot with the new symbols

```
> trellis.par.set("superpose.symbol", sps)
> xyplot(country ~ year, groups = group, type = "b",
+ panel = panel.superpose)
> detach()
> attach(asia)
```

2. Starting with the AIDS data from Vietnam, backcalculate past HIV incidence using the EMS program. Since few effective treatments are available, try the Weibull incubation distribution with median 10 years and a low degree of smoothing, or no smoothing. Note that `nperyear = 1`.

   Again, plot the estimated HIV infection curve, observed and fitted AIDS incidence, and annual predicted AIDS incidence to 2003.

   The WHO projects that by 2000, the cumulative number of HIV infections will reach about 133,000 to 160,000 and among them, 14,000 to 21,000 will have developed AIDS. *How do your results compare with these predictions?*

3. Now analyse the Cambodian AIDS data, and compare the resulting HIV prevalence estimates and AIDS predictions with those given in the WHO Report on Cambodia discussed in yesterday's tutorial. *How do the predictions from Cambodia and Vietnam compare?*

You could now proceed in one of several directions

- estimate the infection distribution and predict AIDS incidence for one or more risk groups in Australia, *e.g.* male homosexuals
- compare the results from question 3 with those from assuming a Weibull median time to progression of 10 years
- change the degree of aggregation in the Australian data *e.g.* use six-monthly or yearly incidence, and compare the results.

References

Bacchetti, P. (1995) Historical assessment of some specific methods for projecting the AIDS epidemic. *American Journal of Epidemiology* **141**, 776-781.

Becker, N.G., Watson, L.F. & Carlin, J.B. (1991) A method of non-parametric back-projection and its application to AIDS data. *Statistics in Medicine* **10**, 1527-1542.

Acknowledgements

Thanks to Niels Becker for making the EMS programs available to the School, to the National Centre in HIV Epidemiology and Clinical Research for the Australian AIDS data, and to Julian Taylor for help with *S-PLUS*.

Patty Solomon
September 1999