Reproducing Martínez Miranda, Nielsen and Nielsen (2015) using the apc package

10 September 2016, version 2

Bent Nielsen Department of Economics, University of Oxford
& Nuffield College
& Institute for Economic Modelling
bent.nielsen@nuffield.ox.ac.uk
http://users.ox.ac.uk/~nuff0078

Contents

1 Introduction

The purpose of this vignette is to use the `apc` package version 1.2.2. to reproduce some of the result in Martínez Miranda, Nielsen and Nielsen (2015): *Inference and forecasting in the age-period-cohort model with unknown exposure with an application to mesothelioma mortality*, published in *Journal of the Royal Statistical Society* A 178, 29-55. The `apc` package builds on the identification analysis in Kuang, Nielsen and Nielsen (2008a), the development of deviance analysis for general data arrays in Nielsen (2014). The package is discussed in Nielsen (2015).

The data consists of counts of mesothelioma deaths in the UK by age, $25-89$, and period $1967-2007$. This is modelling using a response-only Poisson regression using an age-period-cohort structure. The purpose of analysis is to forecast the future burden of mesothelioma deaths.

The data are available in the `apc` package. They can be called with the command

```
> library(apc)
> data <- data.asbestos()
```

2 Figure 1: Summary of data

The data is organised as a matrix with period as row index and age as column index. Figure 1(a,b,c) in the paper shows sums of the data by age, period and cohort. Figure 1(d) shows log-cumulative deaths by 5-year age and cohort group.

A range of plots illustrating the data can be generated by the command

```
> apc.plot.data.all(data)
```

This command calls a range of particular commands. Some warnings are reported. This is because one of the plots, `apc.plot.data.within` groups the indices and the index ranges are not divisible by the default group size.

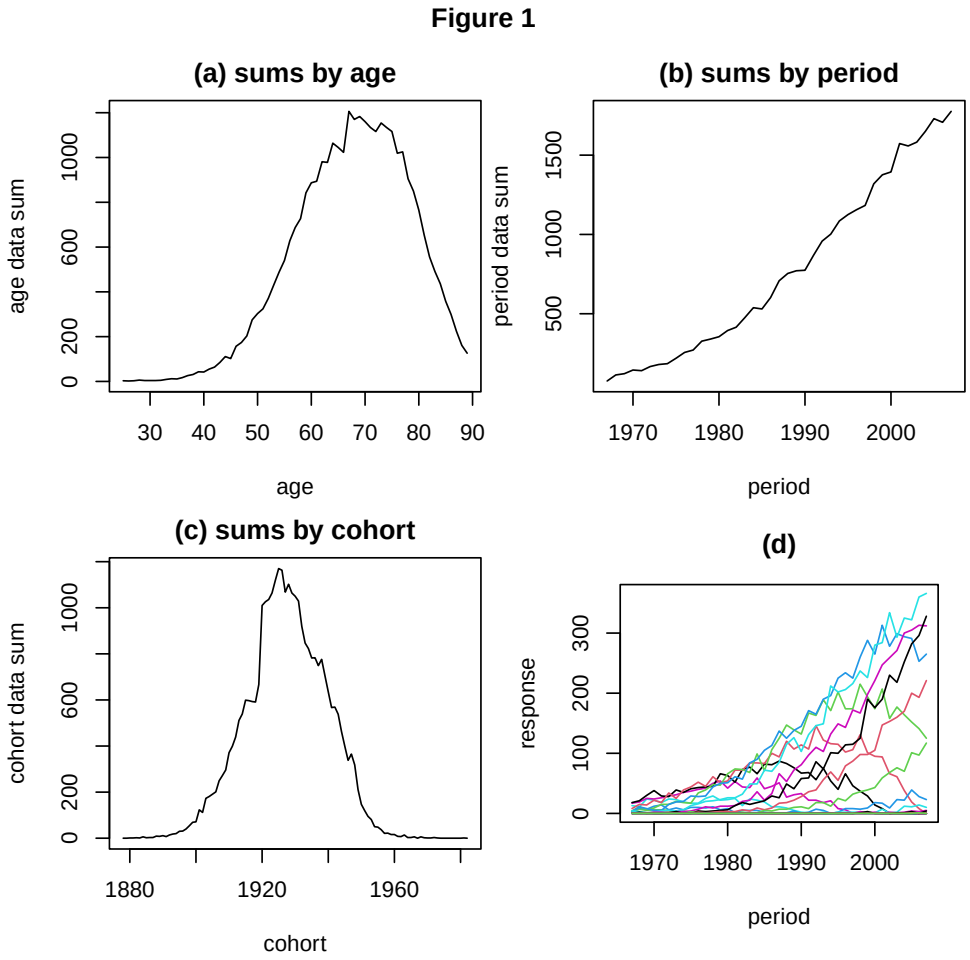
We can also generate plots in a more basic way so as to allow more customization. In particular, Figure 1(a,b,c) can be reproduced by

```
> apc.plot.data.sums(data)
```

To get individual plots one can generate the data sums and plot them as desired. For instance,

```
> data.sums <- apc.data.sums(data)
> par(mfrow=c(2,2),oma=c(0,0,2,0),mar=c(4,4,2,0)+0.1)
> plot(seq(25,89),data.sums$sums.age,
+     main="(a) sums by age",
+     type="l",xlab="age",ylab="age data sum")
> plot(seq(1967,2007),data.sums$sums.per,
+     main="(b) sums by period",
+     type="l",xlab="period",ylab="period data sum")
> plot(seq(1878,1982),data.sums$sums.coh,
```

```
+     main="(c) sums by cohort",
+     type="l",xlab="cohort",ylab="cohort data sum")
> apc.plot.data.within(data,plot.type="pwc",
+     thin=5,type="l",main="(d)",lty=1,legend=FALSE)
> title("Figure 1",outer=TRUE)
```



3 Table 1: Deviance analysis

The deviance Table 1 can be reproduced by a single command

```
> apc.fit.table(data,"poisson.response")[1:4,1:6]
```

```
    deviance df.residual prob(>chi_sq) LR vs.APC df vs.APC prob(>chi_sq)
APC 2384.923        2457         0.848       NaN       NaN           NaN
AP  5336.034        2560         0.000  2951.111       103         0.000
AC  2441.728        2496         0.778    56.805        39         0.033
PC  8265.746        2520         0.000  5880.823        63         0.000
```

4 Figure 3: The standardized residuals

In the first instance we consider the unrestricted age-period-cohort model. This is fitted by the command

```
> fit.apc <- apc.fit.model(data,"poisson.response","APC")
```

A range of plots illustrating the data can be generated by the command

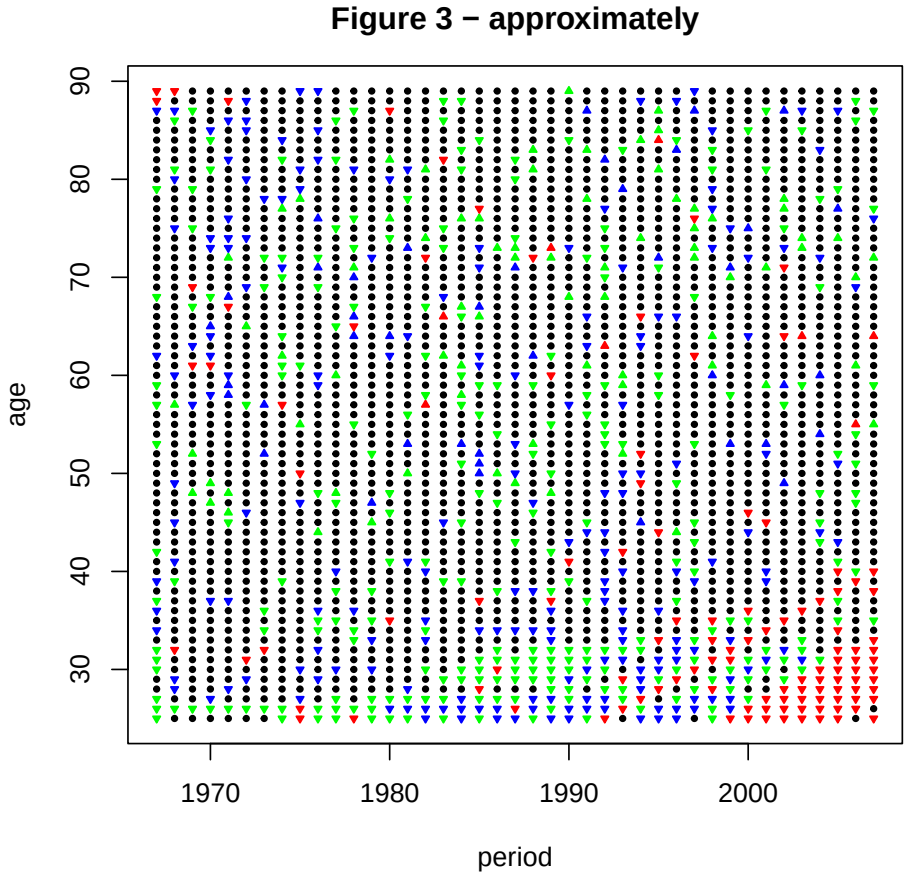
```
> apc.plot.fit.all(fit.apc)
```

```
WARNING apc.plot.fit: sdv large for plot 5  because constant not treated as parameter
WARNING apc.plot.fit: sdv large for plot 9 - possibly not plotted
```

This command calls a range of particular commands. A warning is reported. This is because one of the plots, `apc.plot.fit` may not show the standard errors one would expect.

We can also generate individual plots The package does not exactly reproduce Figure 2. Rather than looking at standardized residuals, we can look at probability transforms of the data given the fitted values through the command

```
> apc.plot.fit.pt(fit.apc,main="Figure 3 - approximately")
```



5 Figure 6: Forecasts based on full sample analysis, but decomposed by cohort

Figure 6 presents forecasts for particular cohorts based on an age-cohort model. The age-cohort model is fitted as follows

```
> fit.ac <- apc.fit.model(data,"poisson.response","AC")
```

We now generate the forecasts for particular cohorts. We need to truncate the range of cohorts when forecasting. This requires a little calculation.

In the paper the range for the cohorts is denoted 1878-1982. In the `apc` package, version 1.2, the range of cohorts is denoted 1879-1983. In any case the index for these cohorts is 1-105. Note that there are 65 age groups and 41 period groups, so that the number of cohorts is 65+41-1=105.

The first 41 cohorts are not going to be extrapolated in any case. Thus, we can potentially forecast 105-41=64 cohorts without having to extrapolate cohort parameters.

In Figure 6 the cohorts are truncated by 1966/1952/1937, in the notation of the paper. This corresponds to truncating the last 16/30/45 cohorts.

We get the truncated forecasts as follows

```
> forecast.16 <- apc.forecast.ac(fit.ac,sum.per.by.coh=c(42,89))
> forecast.30 <- apc.forecast.ac(fit.ac,sum.per.by.coh=c(42,75))
> forecast.45 <- apc.forecast.ac(fit.ac,sum.per.by.coh=c(42,60))
```

Some warnings are produced. These relate to the command `apc.data.list.subset`, which is used for truncating the point forecasts to the desired cohorts.

We need to sum the data by period to show the actual outcomes.

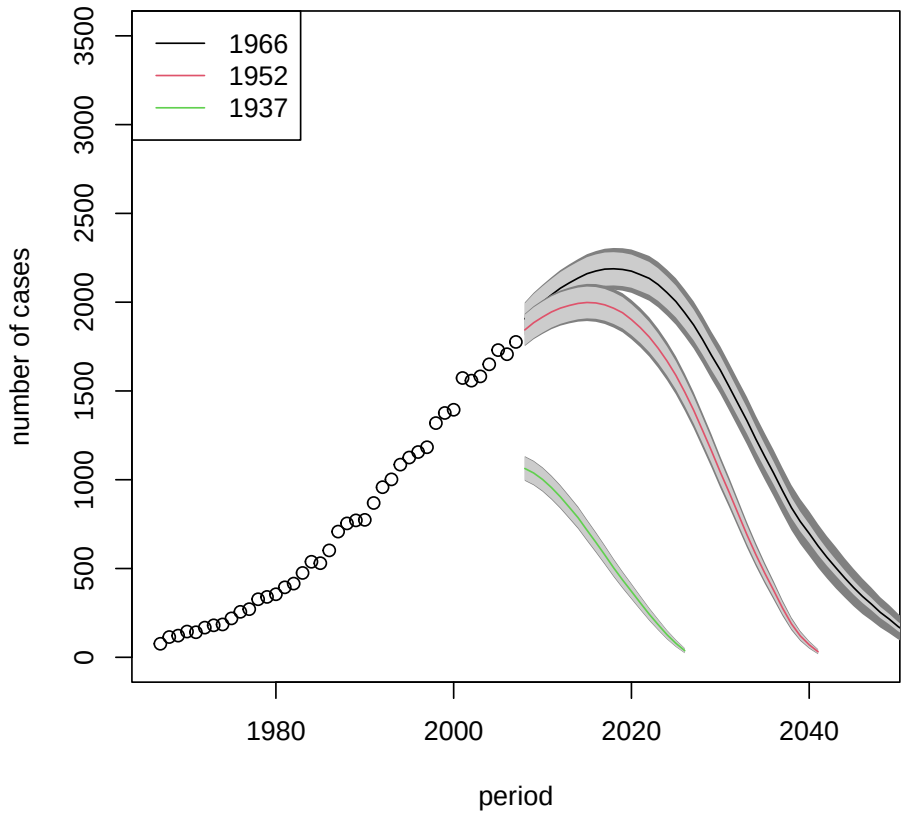
```
> data.sum.per <- apc.data.sums(data.asbestos())$sums.per
```

Figure 6 can now be reproduced as follows. The command `apc.polygon` allows easy plotting of forecast with confidence bands. The function uses `lines` function from the `graphics` package to plot the point forecasts. It also uses the `polygon` function from the `graphics` package to draw up shaded areas for the forecast standard error, and possibly also for the process standard error and the estimation standard error. The darker shaded area represents plus/minus twice the overall forecast standard deviation. The lighter aerea represents plus/minus twice the process error forecast standard deviation, that is the estimates are taken is parameters without estimation uncertainty.

```
> plot(seq(1967,2007),data.sum.per,xlim=c(1967,2047),ylim=c(0,3500),
+     xlab="period",ylab="number of cases",
+     main="Figure 6")
> apc.polygon(forecast.16$response.forecast.per.by.coh,
+     2007,TRUE,TRUE,col.line=1)
> apc.polygon(forecast.30$response.forecast.per.by.coh,
+     2007,TRUE,TRUE,col.line=2)
> apc.polygon(forecast.45$response.forecast.per.by.coh,
```

```
+     2007,TRUE,TRUE,col.line=3)
> legend("topleft",legend=c("1966","1952","1937"),
+     col=c(1,2,3),lty=1)
```

Figure 6



6 Figure 7: Recursive forecasts

Figure 7 presents recursive forecasts using age-cohort models. The darker shaded area represents plus/minus twice the overall forecast standard deviation. The lighter aerea represents plus/minus twice the process error forecast standard deviation, that is the estimates are taken is parameters without estimation uncertainty.

To produce the forecasts we start by extracting a subset of the data array. Then we rerun the age-cohort model and finally produce the forecasts.

```
> data.1991 <- apc.data.list.subset(data.asbestos(),0,0,0,16,0,0)

WARNING apc.data.list.subset: cuts in arguments are:
[1]  0  0  0 16  0  0
have been modified to:
[1]  0  0  0 16  0 16
WARNING apc.data.list.subset: coordinates changed to "AC" & data.format changed to "t
```

```
> fit.ac.1991 <- apc.fit.model(data.1991,"poisson.response","AC")
> forecast.1991 <- apc.forecast.ac(fit.ac.1991)
```

There are two warnings relating to the command `apc.data.list.subset` The first warning concerns the truncation of the data array. Truncating the last 16 periods implies that we also truncate the last 16 cohorts. The second warning shows that the coordinate system has been changed from the original per-age coordinates to age-cohort coordinates.

We start by generating the other forecasts.

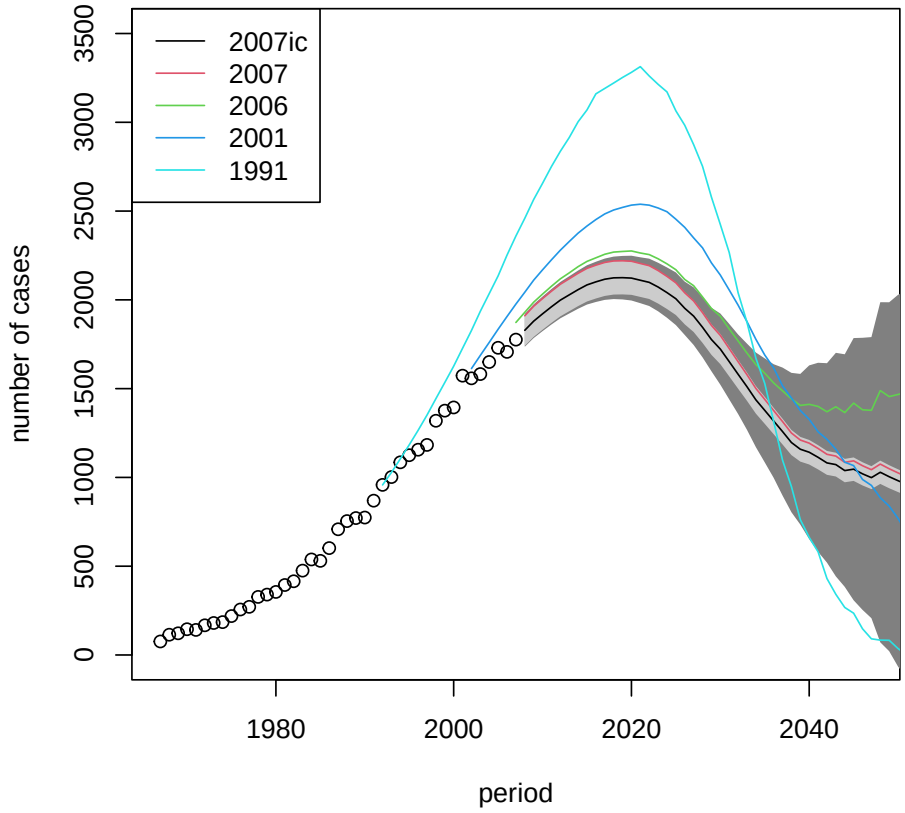
```
> data.2001 <- apc.data.list.subset(data.asbestos(),0,0,0,6,0,0)

WARNING apc.data.list.subset: cuts in arguments are:
[1] 0 0 0 6 0 0
have been modified to:
[1] 0 0 0 6 0 6
WARNING apc.data.list.subset: coordinates changed to "AC" & data.format changed to "t

> fit.ac.2001 <- apc.fit.model(data.2001,"poisson.response","AC")
> forecast.2001 <- apc.forecast.ac(fit.ac.2001)
> data.2006 <- apc.data.list.subset(data.asbestos(),0,0,0,1,0,0)

WARNING apc.data.list.subset: cuts in arguments are:
[1] 0 0 0 1 0 0
have been modified to:
[1] 0 0 0 1 0 1
WARNING apc.data.list.subset: coordinates changed to "AC" & data.format changed to "t

> fit.ac.2006 <- apc.fit.model(data.2006,"poisson.response","AC")
> forecast.2006 <- apc.forecast.ac(fit.ac.2006)
> fit.ac.2007 <- apc.fit.model(data.asbestos(),"poisson.response","AC")
> forecast.2007 <- apc.forecast.ac(fit.ac.2007)
```

This is followed by the plot.

```
> plot(seq(1967,2007),data.sum.per,xlim=c(1967,2047),ylim=c(0,3500),
+     xlab="period",ylab="number of cases",
+     main="Figure 7")
> apc.polygon(forecast.2007$response.forecast.per.ic,
+     2007,TRUE,TRUE,col.line=1)
> apc.polygon(forecast.2007$response.forecast.per   ,
+     2007,FALSE    ,col.line=2)
> apc.polygon(forecast.2006$response.forecast.per   ,
+     2006,FALSE    ,col.line=3)
> apc.polygon(forecast.2001$response.forecast.per   ,
+     2001,FALSE    ,col.line=4)
> apc.polygon(forecast.1991$response.forecast.per   ,
```

```
+     1991,FALSE    ,col.line=5)
> legend("topleft",legend=c("2007ic","2007","2006","2001","1991"),
+     col=c(1,2,3,4,5),lty=1)
```

Figure 7



7 Table 2: Peaks from recursive analysis

We can find the peaks by inspecting the output as follows.

```
> forecast.2007$response.forecast.per[10:14,1]

per_2017 per_2018 per_2019 per_2020 per_2021
2210.916 2219.080 2220.054 2217.510 2204.961

> forecast.2007$response.forecast.per.ic[10:14,1]

per_2017 per_2018 per_2019 per_2020 per_2021
2116.182 2123.996 2124.929 2122.494 2110.483
```

The peak is 2019 in both cases - as it must be - because the intercept correction simply scales the forecasts.

8 Figure 8: Sensitivity analysis

Figure 8 includes forecasts from different models. One forecast is based on the unrestricted age-period-cohort model. This is not implemented in the `apc` package as yet. The other forecasts use age-cohort models, but for different subsets of the data. The forecasts are generated as follows.

```
> data.coh.1966 <- apc.data.list.subset(data.asbestos(),0,0,0,0,0,16)

WARNING apc.data.list.subset: coordinates changed to "AC" & data.format changed to "t

> fit.ac.coh.1966 <- apc.fit.model(data.coh.1966,"poisson.response","AC")
> forecast.coh.1966 <- apc.forecast.ac(fit.ac.coh.1966)
> data.age.35 <- apc.data.list.subset(data.asbestos(),10,0,0,0,0,0)

WARNING apc.data.list.subset: cuts in arguments are:
[1] 10  0  0  0  0  0
have been modified to:
[1] 10  0  0  0  0 10
WARNING apc.data.list.subset: coordinates changed to "AC" & data.format changed to "t

> fit.ac.age.35 <- apc.fit.model(data.age.35,"poisson.response","AC")
> forecast.age.35 <- apc.forecast.ac(fit.ac.age.35,sum.per.by.coh=c(42,89))
```

Finally, a forecast from an age-period-cohort is needed. This requires an extrapolation of the period parameters, see Kuang, Nielsen and Nielsen (2008b). The data appear to be fairly smooth, so an "I0" forecast is chosen. This is the default for `apc.forecast.apc`, so the forecast is generated as follows.

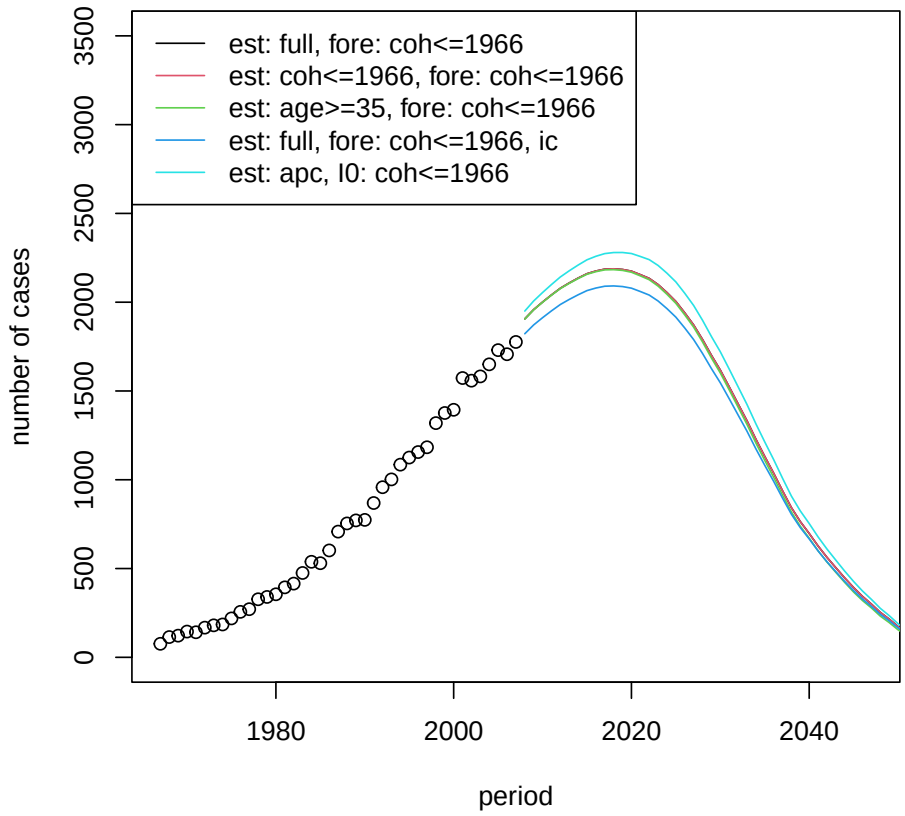
```
> fit.apc.1966 <- apc.fit.model(data.coh.1966,"poisson.response","APC")
> forecast.apc.1966   <- apc.forecast.apc(fit.apc.1966)
```

The plot is then generated as follows. Note that the plot in the paper has no standard deviations, so it would be nearly as easy to use the `lines` function from the `graphics` package as the `apc.polygon` function

```
> plot(seq(1967,2007),data.sum.per,xlim=c(1967,2047),ylim=c(0,3500),
+     xlab="period",ylab="number of cases",
+     main="Figure 8")
> apc.polygon(forecast.16$response.forecast.per.by.coh    ,
+     2007,FALSE,col.line=1)
> apc.polygon(forecast.coh.1966$response.forecast.per     ,
+     2007,FALSE,col.line=2)
> apc.polygon(forecast.age.35$response.forecast.per.by.coh,
+     2007,FALSE,col.line=3)
> apc.polygon(forecast.16$response.forecast.per.by.coh.ic ,
+     2007,FALSE,col.line=4)
```

```
> apc.polygon(forecast.apc.1966$response.forecast.per     ,
+     2007,FALSE,col.line=5)
> legend("topleft",legend=c("est: full, fore: coh<=1966",
+                           "est: coh<=1966, fore: coh<=1966",
+                           "est: age>=35, fore: coh<=1966",
+                           "est: full, fore: coh<=1966, ic",
+                           "est: apc, I0: coh<=1966"),
+     col=c(1,2,3,4,5),lty=1)
```

Figure 8

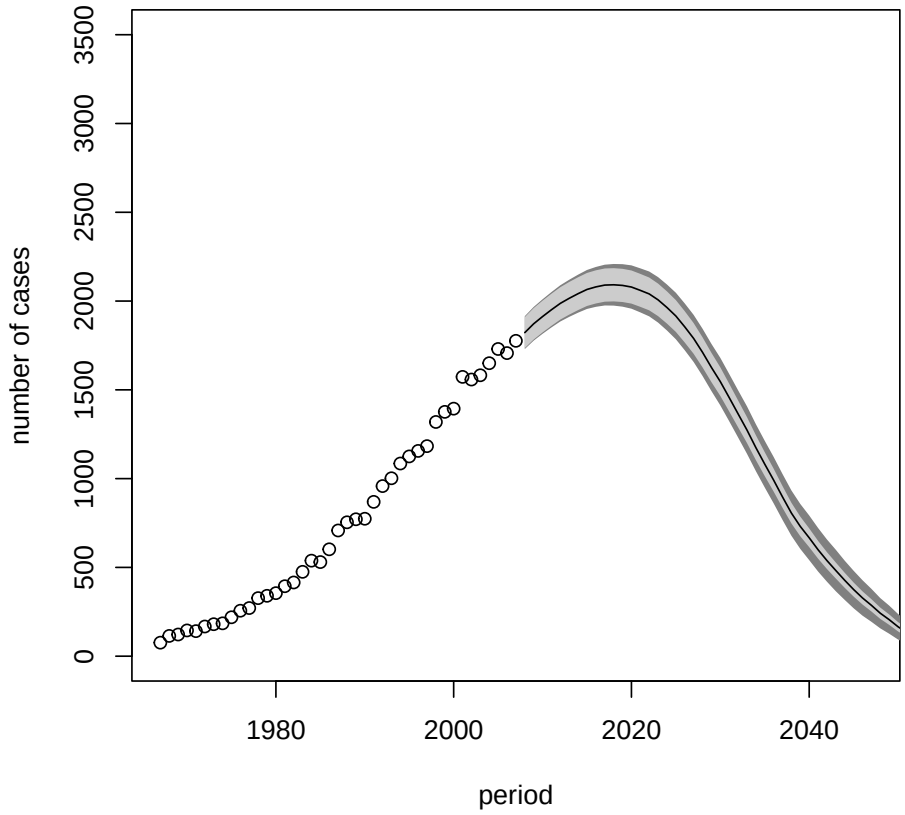


9 Figure 9: Preferred forecast

The preferred forecast is now generated at follows.

```
> plot(seq(1967,2007),data.sum.per,xlim=c(1967,2047),ylim=c(0,3500),
+     xlab="period",ylab="number of cases",
+     main="Figure 8")
> apc.polygon(forecast.16$response.forecast.per.by.coh.ic,2007,TRUE,TRUE)
```

Figure 8



References

Kuang, D., Nielsen, B. and Nielsen, J.P. (2008a) Identification of the age-period-cohort model and the extended chain ladder model. *Biometrika* 95, 979-986. *Download*: Earlier version: http://www.nuffield.ox.ac.uk/economics/papers/2007/w5/KuangNielsenNielsen07.pdf.

Kuang, D., Nielsen, B. and Nielsen, J.P. (2008b) Forecasting with the age-period-cohort model and the extended chain-ladder model. *Biometrika* 95, 987-991. *Download*: Earlier version: http://www.nuffield.ox.ac.uk/economics/papers/2008/w9/KuangNielsenNielsen_Forecast.pdf.

Martínez Miranda, M.D., Nielsen, B. and Nielsen, J.P. (2015) Inference and forecasting in the age-period-cohort model with unknown exposure with an application to mesothelioma mortality. *Journal of the Royal Statistical Society* A 178, 29-55. *Download*: http://www.nuffield.ox.ac.uk/economics/papers/2013/Asbestos8mar13.pdf.

Nielsen, B. (2014) Deviance analysis of age-period-cohort models. *Download*:

http://www.nuffield.ox.ac.uk/economics/papers/2014/apc_deviance.pdf.

Nielsen, B. apc: An R package for age-period-cohort analysis. To appear in *R Journal*. *Download*: https://journal.r-project.org/archive/accepted/nielsen.pdf.