apc: A Package for Age-Period-Cohort Analysis

B. Nielsen[1]

6 November 2014

Abstract: The apc package includes functions for age-period-cohort analysis based on the canonical parametrisation of Kuang et al. (2008). The package includes functions for organizing the data, descriptive plots, a deviance table, estimation of (sub-models of) the age-period-cohort model, a plot for specification testing, plots of estimated parameters, and sub-sample analysis.

1 Introduction

Age-period-cohort models are extensively used in actuarial sciences, demography, epidemiology and social sciences. They have an identification problem in that the predictor is defined from time effects for age, period and cohort, but these time effects cannot be fully recovered from the predictor. The apc package, see Nielsen (2014a), implements the solution proposed by Kuang et al. (2008), which is to abandon the time effects and reparametrise the predictor in terms of freely varying parameters.

The age-period-cohort model has three time scales: age i, period j and cohort k. These are linked through the identity $i+k=j-1$. The data are two dimensional arrays of responses and doses indexed by two of the three time scales. The statistical model is a generalized linear model with a predictor of the form

$$\mu_{ik} = \alpha_i + \beta_j + \gamma_k + \delta, \tag{1.1}$$

The levels and linear slopes of the individual age, α_i, period, β_j and cohort time effects, γ_k, are not identifiable. The problem is discussed for instance by Clayton and Schifflers (1987a,b), Holford (1985), O'Brien (2011), Luo (2013), while Nielsen and Nielsen (2014) describe the problem algebraically.

The identification problem is addressed by reparametrising the model along the lines of Kuang et al. (2008), see also Nielsen (2014b). This exploits that the second differences of the time effects are identified and that the predictor itself is also identifiable. As an example, define the second difference of the age effect by $\Delta^2\alpha_i = \Delta\alpha_i - \Delta\alpha_{i-1}$ where $\Delta\alpha_i = \alpha_i - \alpha_{i-1}$. In an age-cohort coordinate system it is identifiable from the predictor through

$$\Delta^2\alpha_i = \mu_{ik} - \mu_{i-1,k+1} - \mu_{i-1,k} + \mu_{i-2,k+1}. \tag{1.2}$$

Kuang et al. (2008) show that the predictor satisfies the representation

$$\mu_{ik} = \mu_{11} + (i-1)(\mu_{21} - \mu_{11}) + (k-1)(\mu_{12} - \mu_{11}) \\ + \sum_{t=3}^{i}\sum_{s=3}^{t}\Delta^2\alpha_s + \sum_{t=3}^{j}\sum_{s=3}^{t}\Delta^2\beta_s + \sum_{t=3}^{k}\sum_{s=3}^{t}\Delta^2\gamma_s, \tag{1.3}$$

[1] Nuffield College & Department of Economics, University of Oxford & Programme on Economic Modelling, INET, Oxford. Address for correspondence: Nuffield College, Oxford OX1 1NF, UK. E-mail: bent.nielsen@nuffield.ox.ac.uk.

when the data is organized in an age-cohort rectangle where $i = 1, \ldots, I$ and $k = 1, \ldots, K$ so that $j = 1, \ldots, J$ with $J = I + K - 1$. The parameter

$$\xi = (\mu_{11}, \mu_{12}, \mu_{21}, \Delta^2\alpha_3, \ldots, \Delta^2\alpha_I, \Delta^2\beta_3, \ldots, \Delta^2\beta_J, \Delta^2\gamma_3, \ldots, \Delta^2\gamma_K)' \tag{1.4}$$

is freely varying. It is identified since different values $\xi^\dagger \neq \xi^\ddagger$ implies different predictors $\mu^\dagger \neq \mu^\ddagger$. In the context of an exponential family ξ is then the canonical parameter and family is regular. In other words the identification problem is addressed by working with the canonical parameter ξ instead of the time effects $\alpha_i, \beta_j, \gamma_k, \delta$. We can therefore consider any inference on the time effects that can be expressed in terms of the canonical parameter. Nielsen (2014b) generalizes this representation to a large class of data arrays.

The package epi has a series of functions for demographic and epidemiological analysis as well as some functions for age-period-cohort analysis. There are two major differences between the packages apc and epi. First, apc uses the canonical parametrization of Kuang et al. (2008), whereas epi does not. Secondly, age-period-cohort data come in matrix formats and have to be vectorized before fitting the generalized linear model. The package apc takes data in a variety of matrix formats and vectorizes internally, while epi takes vectorized data in a data frame format.

2 The apc package

The package includes functions for organizing the data, descriptive plots, a deviance table, estimation of (sub-models of) the age-period-cohort model, a plot for specification testing, plots of estimated parameters, and sub-sample analysis. These are described in turn.

The data example for this analysis is a data set for annual mesothelioma deaths in the UK taken from Martínez Miranda et al. (2014). It is though that most mesothelioma deaths are caused by exposure to asbestos. The data set has counts of male deaths by age 25–89 and by 1967–2007. There is no direct measure for the exposure to asbestos.

2.1 Organizing the data

Age-period-cohort data may include doses and responses or just responses. They come in different types of data arrays. apc allows for the six matrix formats arising from the choice of two indices from the age, period, and cohort time scales, as well as a triangular format for chain-ladder analysis. All these are special cases of a generalized trapezoid format. A special data format `apc.data.list` is used to keep track of the data format and the time scales. An artificial response-only data set organized in age-period format can be coded as follows

```
> library(apc)
> m.data <- matrix(data=seq(12),nrow=3,ncol=4)
> data.artificial <- apc.data.list(m.data,"AP",age1=25,per1=1990,unit=5)
> data.artificial$response
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

The function `apc.data.list` assigns ten objects to the variable `data.artificial`. The first argument defines the response data, while the second argument signifies that the response matrix is rectangular with coordinates in age-period format. The remaining arguments are optional. In this case information about the times scales have been given. This shows that the real time scales are $25, 30, 35$ for age and $1990, 1995, 2000, 2005$ for period, which in turn implies that the cohorts are $1955, 1970, \cdots, 1980$. When reporting estimators apc will keep track of these real time scales. No dose data are defined, so `data.artificial$dose` will be `NA`.

A variety of data from the literature are pre-coded including the asbestos data from Martínez Miranda et al. (2014). The available information for that data set is exactly as in the previous example: a data matrix for responses in age-period format, though much larger, along with information about the time scales. It can be called through

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

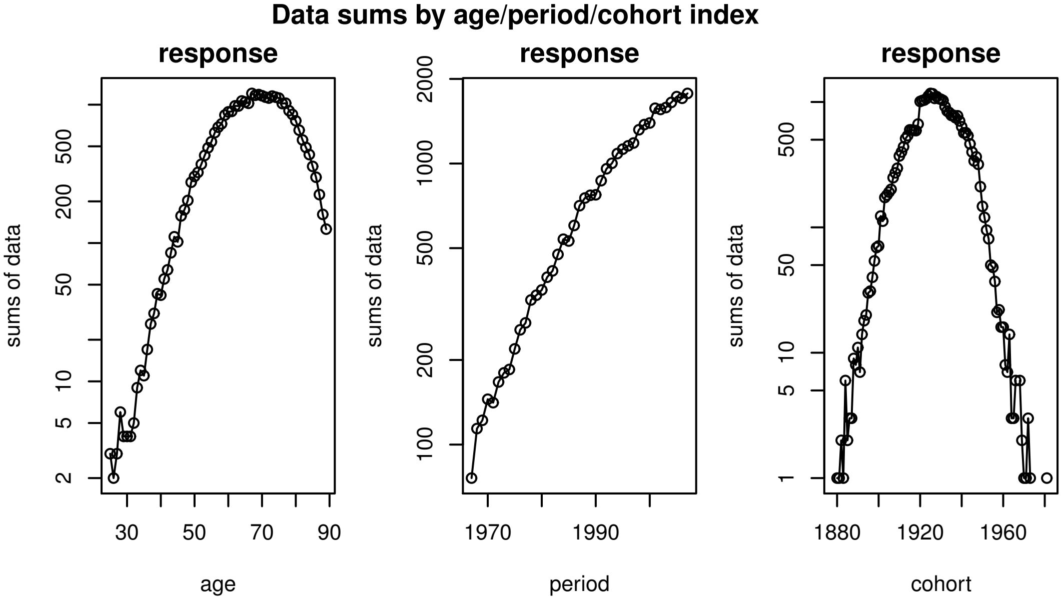

Figure 1: Data sums by age, by period and by cohort.

2.2 Descriptive plots

apc has a variety of plots for descriptive analysis. These include plots of sums of the data by age, period or cohort to get an idea of the aggregate development. Plots of the data matrix against two of the three time indices to spot patterns in the data. Sparcity plots that indicate if some entries in the data matrix are very small. For instance, there are very few mesothelioma deaths for young people. These plots can be called and manipulated individually or they can be called with a single command.

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

Figure 1 show the plots of data sums. The responses are seen to be sparse for young people and for old and recent cohorts. The sparcity plot, which is not reported here, illustrates this in more detail.

2.3 Deviance analysis

At this point the distribution is chosen. Currently four distributions are implemented: A Poisson response model, a Poisson dose-response model, a logistic dose-response model, and a Gaussian model giving least squares regression. The sampling theory for the two Poisson models is described in Martínez Miranda et al. (2014), Nielsen (2014b), respectively.

The age-period-cohort model has a variety of interesting sub-models. These arise by setting some of the coordinates of the canonical parameter ξ to zero. An age-cohort model ”AC” arises by setting the period double-differences to zero, so $\Delta^2\beta_j = 0$ for $j = 1, \ldots, J$. The drift models of Clayton and Schifflers (1987a,b) arise by setting two sets of double-differences to zero. An age-drift model ”Ad” arises as a sub-model of ”AC” by setting $\Delta^2\gamma_k = 0$ for $j = 1, \ldots, K$. For instance, an age model ”A” arises as a sub-model of ”Ad” by setting the cohort slope to zero. A pure trend model ”t” arises as a sub-model of ”Ad” by setting $\Delta^2\alpha_i = 0$ for $i = 1, \ldots, I$. A deviance table gives an overview of the relative performance of the different models. For the mesothelioma data we get the following output.

```
> apc.fit.table(data.asbestos,"poisson.response")
       -2logL df.residual prob(>chi_sq) LR.vs.APC df.vs.APC prob(>chi_sq)      aic
APC  2384.923        2457         0.848        NA        NA            NA 10805.81
AP   5336.034        2560         0.000  2951.111       103         0.000 13550.92
AC   2441.728        2496         0.778    56.805        39         0.033 10784.61
PC   8265.746        2520         0.000  5880.823        63         0.000 16560.63
Ad   5912.422        2599         0.000  3527.499       142         0.000 14049.31
Pd  23461.384        2623         0.000 21076.461       166         0.000 31550.27
Cd   8494.658        2559         0.000  6109.735       102         0.000 16711.54
A   21948.036        2600         0.000 19563.113       143         0.000 30082.92
P   34391.044        2624         0.000 32006.121       167         0.000 42477.93
C   28415.983        2560         0.000 26031.060       103         0.000 36630.87
t   24037.772        2662         0.000 21652.849       205         0.000 32048.66
tA  40073.386        2663         0.000 37688.463       206         0.000 48082.27
tP  34967.432        2663         0.000 32582.509       206         0.000 42976.32
tC  50558.531        2663         0.000 48173.607       206         0.000 58567.42
1   51003.046        2664         0.000 48618.123       207         0.000 59009.93
```

The first column in the table has the heading `-2logL`, noting that the deviance equals minus twice the log likelihood for Poisson and logistic models, but not for Gaussian models. The deviance table indicates that the reduction worth considering is an age-cohort model, which is denoted ”AC”. Moreover, the quality of the unrestricted model is quite good, with a deviance smaller than the degrees of freedom. Nielsen (2014b) gives a detailed discussion of the interpretation of the sub-models.

2.4 Estimation of a particular model

We can look a bit closer at a particular sub-model. For instance, in the case of the asbestos data the unrestricted age-period-cohort model is estimated as follows. The canonical parameter has 208 parameters, so only a few of the estimates are reported here.

```
> fit.apc<- apc.fit.model(data.asbestos,"poisson.response","APC")
> fit.apc$coefficients.canonical[1:8,]
                 Estimate Std. Error      z value     Pr(>|z|)
level         1.041126756         NA           NA           NA
age slope     0.379386996  0.1115535  3.400941274 0.0006715425
cohort slope  0.358297074  0.1125026  3.184789061 0.0014485956
DD_age_27     1.029446394  1.6467618  0.625133761 0.5318832712
DD_age_28     0.065309039  1.4311381  0.045634337 0.9636017004
DD_age_29    -1.097279478  1.1180554 -0.981417831 0.3263867366
DD_age_30     0.414467808  1.1902557  0.348217448 0.7276768856
DD_age_31     0.003217972  1.2247555  0.002627441 0.9979036081
```

No standard errors are reported for the level. This is because a multinomial sampling scheme conditioning on level is used in the case of the Poisson response model. Note that the names for the parameters utilize the information about the real time scales coded through `apc.data.list()`.

2.5 A probability transform plot for the fit

The quality of the fit can be illustrated using a probability transform plot. Using the estimates it plots probability transforms of responses given the fitted value. In other words, do the actual observations seem probably given the estimated model. The plot is given in the original coordinate system. Colours and symbols are used to indicate whether responses are central to the fitted distribution or in the tails of the fitted distribution. The intention of the plot is to reveal if there are particularly many extreme observations given the fit and if they form a particular pattern.

For the asbestos data the probability transform plot is coded as

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

Figure 2 shows the result. For instance, all red point triangles indicates observations in the extreme 1 % of the distribution. Since they are all point down they are in lower end of the distribution. The number of red triangles is not particular large given the number of observations, $n = 2665$, but, they form a pattern among the most recent cohorts. Therefore, a recursive analysis is performed below.

2.6 Plots of the estimates

Figure 3 shows plots of the estimates generated by the code

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

The object `fit.apc` includes information about which sub-model has been estimated. This is used to address the identification issues, which vary according to which components are present in a given sub-model, see Nielsen (2014b) for further details.

Figure 3 (a)-(c) shows the estimated second difference parameters $\Delta^2\alpha_i$, $\Delta^2\beta_j$, $\Delta^2\gamma_k$. The estimates are plotted with pointwise confidence bands centered around zero. The age double differences are noisy for young ages while the cohort double differences are noisy for young and old cohorts. This is due to the sparcity of observations for those age and cohort groups. This calls for a sub-sample analysis, which is described below. For further discussion see Martínez Miranda et al. (2014).

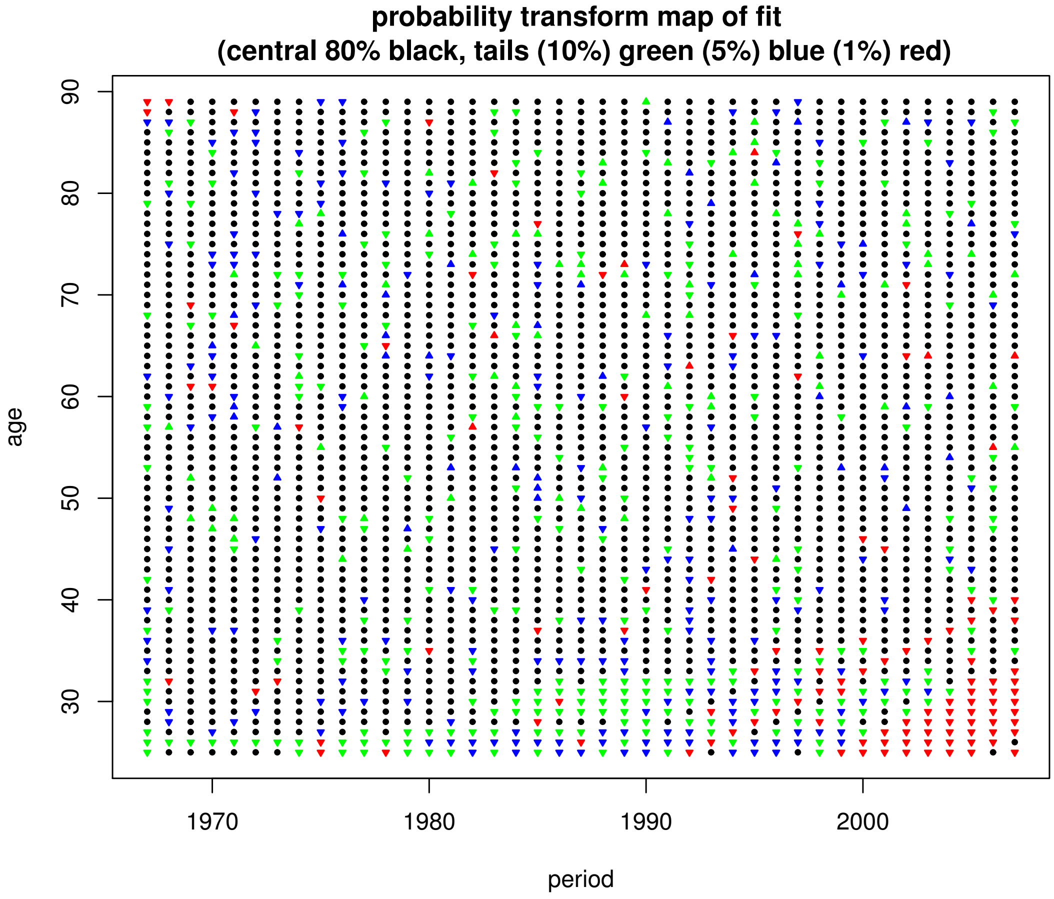

Figure 2: Probability transform plot of age-period-cohort fit to asbestos data.

Given the original model formulation in (1.1) it is of interest to represent the time effects for age, period and cohort somehow. The representation (1.3) implies that these time effects are given as double sums of double differences combined with an ad hoc chosen linear trend, see see Clayton and Schifflers (1987b) for an illustration. Figure 3(g)-(i) shows double sums of double difference combined with a linear trend chosen so that the series start and end in zero. The idea is to give a good visual impression of variation over and above a linear trend and at the same time preserving the degrees of freedom in panels (a)-(c). Figure 3(d)-(f) indicates the linear plane arising from the particular identification choice. Nielsen (2014b) discusses these choices in further detail.

2.7 Sub-sample analysis

The asbestos data is sparse for low ages and for old and young cohorts. A recursive analysis can be used to check how sensitive the above analysis is to this. The idea is to cut parts of observations away and redo the analysis. This is done through

```
> data.asbestos.subset <- apc.data.list.subset(data.asbestos,10,0,0,0,3,16)
```

which cuts the lower 10 age groups, the lower 3 cohort groups and the upper 16 groups. The subset of the data is no longer a rectangle in the age-period coordinate system, but

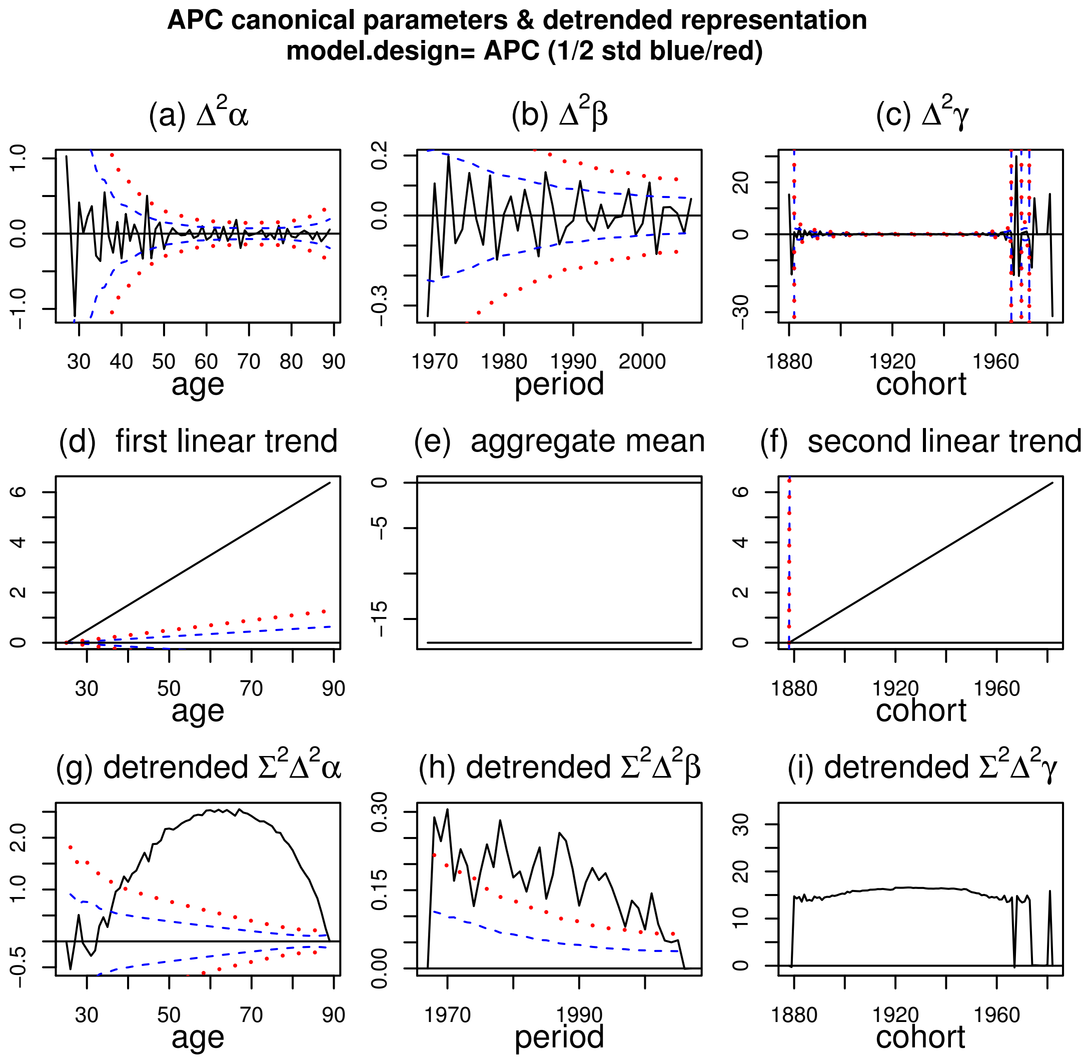

Figure 3: Plots of the fitted values.

rather a rectangle with some corners cut off. This is called a generalized trapezoid in Kuang et al. (2008). The above analysis can now be redone.

Instead of cutting the oldest cohorts one could cut the oldest periods. Doing that implies that both the minimum period and minimum cohort change. This has implications for which double difference parameters are eliminated and which three points should be used to measure the linear plane. This is taken into acoount, but a warning is issued.

```
> data.asbestos.subset <- apc.data.list.subset(data.asbestos,10,0,3,0,0,16)
[1] "apc.data.list.subset WARNING:"
[1] "cuts in argument are:"
[1] 10  0  3  0  0 16
[1] "have been modified to:"
[1] 10  0  3  0  3 16
```

3 Summary

This article describes the apc package for age-period-cohort modelling. It implements the canonical parametrisation of Kuang et al. (2008). The package includes functions for organizing the data, a descriptive plot, a deviance table, estimation of sub-models of the age-period-cohort model, a plot for specification testing, plots of estimated parameters, and sub-sample analysis.

References

D. Clayton and E. Schifflers. Models for temperoral variation in cancer rates. i: age-period and age-cohort models. *Statistics in Medicine*, 6:449–467, 1987a.

D. Clayton and E. Schifflers. Models for temporal variation in cancer rates. ii: Age-period-cohort models. *Statistics in Medicine*, 6:469–481, 1987b.

T. R. Holford. An alternative approach to statistical age-period-cohort analysis. *Journal of Chronic Diseases*, 38:831–836, 1985.

D. Kuang, B. Nielsen, and J. P. Nielsen. Identification of the age-period-cohort model and the extended chain ladder model. *Biometrika*, 95:979–986, 2008.

L. Luo. Assessing validity and application scope of the intrinsic estimator approach of the age-period-cohort problem. *Demography*, 50:1945–1967, 2013.

M. D. Martínez Miranda, B. Nielsen, and J. P. Nielsen. 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*, to appear, 2014.

B. Nielsen. *apc: A Package for Age-Period-Cohort Analysis*, 2014a. URL `http://CRAN.R-project.org/package=apc`. R package version 1.0.

B. Nielsen. Deviance analysis of age-period-cohort models. Nuffield College Discussion Paper, 2014b.

B. Nielsen and J. P. Nielsen. Identification and forecasting in mortality models. *The Scientific World Journal*, pages ID 347043, 24 pages, 2014.

R. M. O'Brien. Constrained estimators and age-period-cohort models. *Sociological Methods & Research*, 40:419–452, 2011.