Allowing Left Truncated and Censored Fertility Data in the Normal Approximated Waiting Model

Per Arnqvist

Institute of Mathematical Statistics
Umeå University
S-901 87 Umeå, Sweden

Abstract

Models describing marital fertility are under consideration. In Arnqvist [2], a normal approximation of the Waiting model was introduced. In this report a modification of the normal approximation is suggested. This specification allows the data to be left truncated and censored, which gives the possibility to apply the normally approximated waiting model in datasets as from the United Nations World Fertility Services. The model is appropriate except for extremely high fertility intensities, when it gives rise to bias in the parameter estimations. In this case, therefore, a bootstrap method is suggested to estimate and correct the bias. This means that the normal approximated waiting model is a good competitor to the well known Poisson or Coale-Trussell model. It also uses an understandable fertility specification.

Key words and phrases: Coale-Trussell model, Poisson model, waiting model, normal approximated waiting model.

1991 AMS subject classification: Primary 62F30, 62P10; secondary 62-07, 62F11.

1 Introduction

1.1 Background

In Arnqvist [1], it was shown that individual fertility data could be better described by a "waiting model" than by the Poisson model as suggested in Broström [4], Trussel [9], Coale & Trussell [5]. The waiting model adds a waiting time after each pregnancy, and this modification gives a much better fit than the Poisson model assumption. However, when data is in the same form as UN:s World Fertility Surveys [10], which are for population age-specific grouped data, usually of the form given in Table 1, the model derived in Arnqvist [1] is not possible

Table 1: *Design of the age-specific grouped data. Here* $B_a = \sum_{j=1}^{n} b_{aj}$ *and* b_{aj} *equals the number of children the* j*th woman in the* a^{th} *interval has given birth to, and* $E_a = \sum_{j=1}^{n} e_{aj}$ *where* e_{aj} *equals the time the* j*th woman has been under exposure in the* a^{th} *interval,* $a = 1, 2, \ldots, 6$.

Age interval	20-24	25-29	30-34	35-39	40-44	45-49
Number of births	B_1	B_2	B_3	B_4	B_5	B_6
Exposure Time	E_1	E_2	E_3	E_4	E_5	E_6

to apply directly. This suggested the introduction of an approximation of the waiting model, and in Arnqvist [2], a normal approximation of the waiting model was derived. This lead to the existence of two models; the Poisson model and the normal approximated waiting model, both describing marital fertility.

The main difference between the two models is the interpretation of fertility. λ_a denotes the fertility measure in the Poisson model. It is estimated by B_a/E_a, and it is some artificial measure of fertility. θ_a is the fertility measure for the normal approximated waiting model. θ_a here means possibility to be pregnant and it is the intensity measure during the active exposure time. The exposure time for one individual in the a^{th} interval denoted e_{aj}, is divided in two parts. One active part, where the individual is assumed to have the possibility to become pregnant, and a nonactive part, where the individual cannot be pregnant. θ_a was introduced in Arnqvist [1], and it is specified in the same way as Coale-Trussell specified λ_a in their intensity model,

$$\theta_a = n_a \cdot e^{(k + m \cdot v_a)}.$$

This is more like the mother natures idea of pregnancy, during the time a woman is pregnant she cannot be pregnant again.

In order to compare the two different models with respect to the intensity θ_a,

a transformation of the intensity λ_a in the Poisson model has to be made. The transformation can be made if the expectation that is calculated under the normal approximation of the waiting model

$$\mathrm{E}[N_W(t)] = \frac{\theta_a t}{1 + \theta_a W},$$

is set equal to the expectation of the Poisson model,

$$\mathrm{E}[N_W(t)] = \lambda_a t,$$

and solved for θ_a. Here $N_W(t)$ denotes the number of pregnancies one individual receives in an interval of length t years.

In Arnqvist [2], the normal approximation was compared with the Poisson model for simulated populations. The comparison was made for the intensity estimation θ_a and the estimation of the mean $\mathrm{E}[N_W(t)]$, and the variance $Var[N_W(t)]$ for the number of the births for the simulated populations.

The simulated populations were generated by assuming that the exposure times for one individual consists of two parts; one active part, which is exponentially distributed, where the individual is assumed to have the possibility to become pregnant, and a nonactive part, denoted waiting time W of one year, where the individual cannot be pregnant. Further, it was assumed that every individual in the population was exposed during her whole reproductive time, meaning [19, 51) years of age, so there were no late marriages and early deaths or migrations.

Why should we consider the newly invented fertility measure? One reason can be that the estimation of fertility, here denoted θ_a, really gives an estimation of the opportunity to become pregnant. The understanding of λ_a under the Poisson model approach is somewhat more difficult. One possible interpretation in that model is that B_a/E_a is a rate. Another reason is that when comparing the different models ability to predict the intensity or the number of births, B_a in the interval I_a, it was an improvement to use the normal approximated waiting model. If the variance or the standard deviation are also estimated, the Poisson model gives very biased estimates. The conclusion of Arnqvist [2] is that the Poisson model is as good as the normal approximated waiting model in the sence of predicting intensity or predicting the number of births for low and moderate intensities, but for larger intensities it does not work.

To really convince ourselves that this is a good approach, one essential step needs to be carried out. We need to investigate how the normal approximated waiting model behaves when the summerized data consists of truncated and censored individual data, due to late marriages and migrations or death. First, however, the normal approximated waiting model needs to be reformulated so it covers this situation.

2 Approximation of the waiting model

In Arnqvist [2], the following approximation was suggested. It is assumed that the number of pregnancies B_a, for a population within the six intervals $20-24$, $25-29, \ldots, 44-49$ is distributed independently, asymptotically normal when the number of individuals is increasing, and we denote it $B_a \rightsquigarrow AsN(\mu_a, \sigma_a^2)$, so the approximated likelihood function of the data can be expressed as

$$L_A(\mu_1, \ldots, \mu_6, \sigma_1^2, \ldots, \sigma_6^2) = \prod_{a=1}^{6} \frac{1}{\sqrt{2\pi\sigma_a^2}} exp\left(\frac{-(B_a - \mu_a)^2}{2\,\sigma_a^2}\right),$$

where

$$\mu_a = N_a \sum_k k\, P(N_W(t) = k)$$

and

$$\sigma_a^2 = N_a \sum_k (k - \mu_a)^2\, P(N_W(t) = k).$$

N_a is the number of individuals that have complete exposure times in the interval a. The index A in $L_A(\cdot)$ will from now on be dropped. The mean and the variance of the number of births for an individual $N_W(t)$, in an interval of length t years and a waiting time of length W years, was suggested to be approximated as

$$\mathrm{E}[N_W(t)] = \frac{\theta t}{1 + \theta\, W}, \tag{1}$$

and

$$Var[N_W(t)] = \frac{\theta t}{(1 + \theta\, W)^3}. \tag{2}$$

This can further be used to estimate the parameters (k, m), together with the asymptotic variance of the parameters in the model. If μ_a and σ_a^2 are replaced with the normal approximations, they become

$$\mu_a = N_a \frac{\theta_a\, t}{1 + \theta_a\, W}$$

and

$$\sigma_a^2 = N_a \frac{\theta_a\, t}{(1 + \theta_a\, W)^3}.$$

They can be simplified if the assumption that $W = 1$ is used, since then

$$\mu_a = N_a \frac{\theta_a\, t}{1 + \theta_a} \tag{3}$$

and

$$\sigma_a^2 = N_a \frac{\theta_a\, t}{(1+\theta_a)^3}. \tag{4}$$

So, (3) and (4) are used, and if we set $\boldsymbol{\psi} = (k, m)$, and $\mathbf{B}$= $(B_1, \ldots, B_6)$ then the approximated likelihood function becomes

$$L(\boldsymbol{\psi}, \mathbf{B}) = \prod_{a=1}^{6} \frac{1}{\sqrt{2\,\pi\, N_a \dfrac{\theta_a\, t}{(1+\theta_a)^3}}} \exp\left(- \frac{\left(B_a - N_a \dfrac{\theta_a\, t}{1+\theta_a} \right)^2}{2\, N_a \dfrac{\theta_a\, t}{(1+\theta_a)^3}} \right).$$

Furthermore, if we use a parameterization of intensity, simular to that suggested by Coale-Trussell,

$$\theta_a = n_a\, e^{k + m\, v_a}$$

which describes the fertility as a product of natural fertility and fertility control, (where the two parameters k and m are the parameters that need to be estimated), then the approximated likelihood function, that should be maximized, can be written as

$$L(\boldsymbol{\psi}, \mathbf{B}) =$$

$$\prod_{a=1}^{6} \frac{1}{\sqrt{2\,\pi\, \dfrac{n_a\, e^{k+m\, v_a}\, N_a\, t}{\left(1 + n_a\, e^{k+m\, v_a}\right)^3}}} exp\left(- \frac{\left(B_a - \dfrac{n_a\, e^{k+m\, v_a}\, N_a\, t}{1 + n_a\, e^{k+m\, v_a}} \right)^2}{\dfrac{2\, n_a\, e^{k+m\, v_a}\, N_a\, t}{(1 + n_a\, e^{k+m\, v_a})^3}} \right). \tag{5}$$

This likelihood covers the case when the individuals are under exposure over whole time intervals I_a, but the information contained in Table 1 are not N_a, the number of individuals per interval. Instead it is E_a, the total exposure time for the women in the population under study. This means that the exposure time e_{aj} for a specific individual in one interval can vary between 0 and 5 years.

However, if instead we consider that the number of births B_a in the intervals depends on the given exposure time E_a, this suggests another specification of the normal approximation of the waiting model. Assume that the number of births B_a in an interval I_a for a population is related to stationary pieces of exposure times for individuals, then the expectation of the number of births can be written as

$$\begin{aligned} \mathrm{E}[B_a] &= \sum_{i=1}^{N_a} \mathrm{E}[b_{ai}|E_a] = \frac{\theta_a}{1+\theta_a} \sum_{i=1}^{N_a} \mathrm{E}[e_{ai}] \\ &= \frac{\theta_a}{1+\theta_a} \mathrm{E}[\sum_{i=1}^{N_a} e_{ai}] = \frac{\theta_a}{1+\theta_a} \mathrm{E}[E_a]. \end{aligned}$$

We can use E_a as an unbiased estimator of $\mathrm{E}[E_a]$, and then $\mathrm{E}[B_a]$ can be approximated by $E_a\,\theta_a/(1+\theta_a)$. In the same manner, if we assume that exposure intervals for the individuals are not to short, then the linear approximation (4) together with the estimator can be used again to approximate the variance according to

$$Var[B_a] = \frac{\theta_a E_a}{(1+\theta_a)^3}.$$

(For a discussion of this approximation of the variance, see the appendix). This can be rephrased as: for a given population and a given time interval I_a, the number of births B_a given the exposure time E_a of the married women is distributed independently, asymptotically normal, i.e $B_a|E_a \rightsquigarrow AsN(\mu, \sigma^2)$, together with the assumption that the B_a's are independent for different intervals, meaning that our asymptotic likelihood can be specified according to

$$L(\boldsymbol{\psi}, \mathbf{B}|\mathbf{E}) =$$

$$\prod_{a=1}^{6} \frac{1}{\sqrt{2\,\pi\,\dfrac{n_a\,e^{k+m\,v_a}\,E_a}{\left(1+n_a\,e^{k+m\,v_a}\right)^3}}} exp\left(-\frac{\left(B_a - \dfrac{n_a\,e^{k+m\,v_a}\,E_a}{1+n_a\,e^{k+m\,v_a}}\right)^2}{\dfrac{2\,n_a\,e^{k+m\,v_a}\,E_a}{(1+n_a\,e^{k+m\,v_a})^3}} \right), \qquad (6)$$

where $\mathbf{E} = E_1, \ldots, E_6$.

What happens is that E_a replaces $N_a\,t$ in the likelihood specification. This means that we need the new formula (6) for the estimation of the fertility measure θ_a.

In the previous specification of the likelihood (5), we could consider the fertility measure applied to individual data. Here, maybe we need to be somewhat more careful. This is due to the fact that the exposure time consists of pieces of different length, and the fertility measure can be understood as some mean fertility measure for the individuals in the population.

The derivatives of the model developed in Arnqvist [2] apply almost directly. The only change is the replacement of E_a for $N_a\,t$. This means that the numerical iterative routine that was used in the previuous study also can be used here.

3 Simulation study

To see whether or not this specification of the normal approximated likelihood is good or not, a simulation study has been performed. Five different parameter settings of (k, m) have been chosen according to Table 1, see also Arnqvist [2],

specifying the fertility intensity in the populations. Throughout the simulations it is assumed that each population consists of 100 individuals.

Table 1: *The parameter values that specify the intensity used in the simulation of the population data generated according to the waiting model.*

Parameter value	
k	m
0.00	0.00
-0.75	-0.75
-0.75	0.75
0.75	-0.75
0.75	0.75

3.1 Populations with truncations and censorings

In the simulation study performed in Arnqvist [2], it was assumed that each woman was under observation over the whole time interval of interest, meaning from twenty to fifty years. Here, in this simulation study, we try to make more realistic assumptions. This means that first, it is assumed that the marriage times follow a distribution giving left truncated entering times of study, and second that there are censorings due to movement or deaths of individuals in the population.

The problem now is to find the patterns for the marriage times and the censoring or end of study times. By using a real dataset, here from Västanfors, it is possible to at least approximate this pattern. (For a demographic discussion and treatment of the Västanfors data see Bengtsson [3]). In Figure 1, the distribution for the marriage times, together with the *loglogistic* distribution estimated from the marriage times in the Västanfors data are shown. It can be seen that a distribution starting at 0 and with a peak around 25 is suitable for this. Several possible distributions can be chosed for this purpose, both discrete and continuous distributions, but the choice landed on the continuous family, here represented with the *lognormal-*, *gamma-*, *weibull-*, and the *loglogistic*-distribution. The *loglogistic* distribution is chosen due to its good performance and simple form. The fit between the *loglogistic* distribution and the Västanfors data is not perfect, but for this simulation study it seems sufficient. The *loglogistic* distribution is parameterizised according to

$$f(x|\alpha,\beta) = \frac{b\, e^{\alpha\,\beta}\, x^{\beta} - 1}{(1 + e^{\alpha\,\beta}\, x^{\beta})^2},$$

where the parameter estimates were obtained through the MLE principle. They are $\alpha = -3.20, \beta = 8.90$.

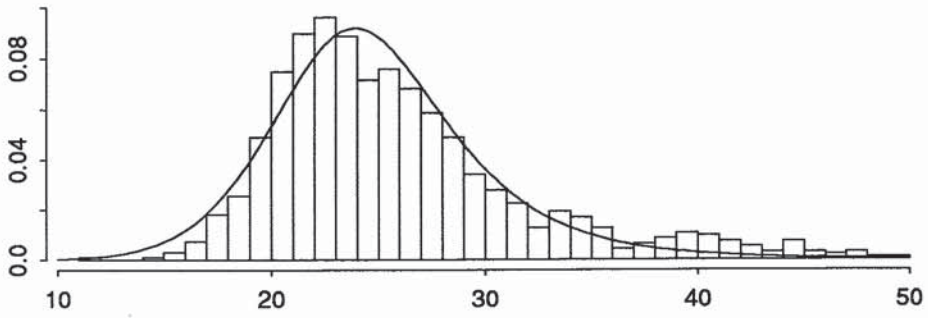

Figure 1: *The histogram of the marriage times in the Västanfors data together with the estimation of the loglogistic density distribution. The Västanfors data consists of 915 married woman.*

Figure 2 shows the censoring times or end of study times for the Västanfors data. Maybe, this could be described by a mixture of two distributions, where the first one should cover the age interval [20,50) and the second distribution [50,100). But since every distribution with positive continuous density can be approximated with the *uniform* distribution on a short interval we have chosen it to cover the death and/or immigration times.

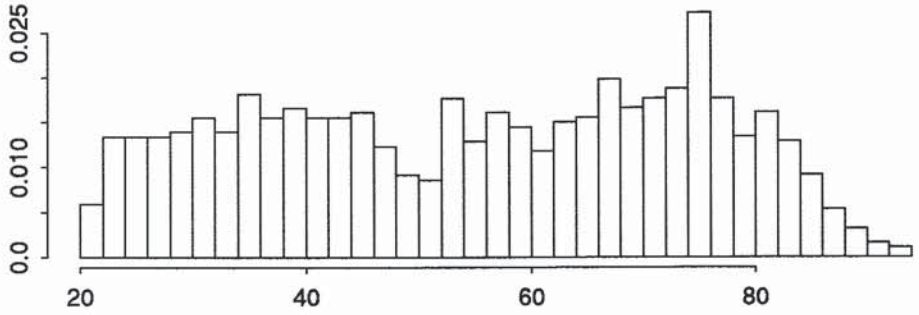

Figure 2: *The distribution of the migration and death times for the Västanfors data.*

This means that in this simulation for each individual in the population first a *loglogistic* entering time m is generated and then a *uniform* leaving time $l \sim U(m, 90)$ is added to the entering time. The number of pregnancies within

the intervals is determined in the same way as in the simulation with no truncation and no censoring, see Arnqvist [2]. Therefore, the summerized data will consist of stationary pieces following the assumptions that are made within the derivation of the normal approximation of the waiting model in section 2.

3.2 A preliminary investigation

To have an idea of what is going on, three pictures are given with 10 simulated populations on each parameter setting according to Table 1. Each population is assumed to consist of 100 individuals. The estimation of the parameter values (k, m) are given under three different models here. The first model is the Poisson model or Coale-Trussel model, the second model is the normal approximated waiting model and the third model is the waiting model which was derived in Arnqvist [1]. The waiting model is the model that uses individual data while the other two use summerized data. It is possible to estimate the parameters within the waiting model since the data are generated as individual data. In Figure 3, the parameter estimates for the three models are plotted. Figure 3a) are the parameter estimates of (k, m) in the Poisson model, Figure 3b) are the parameter estimates of (k, m) under the normal approximation of the waiting model and in Figure 3c) the parameter estimates under the waiting model using individual data are given. We can see in Figure 3 that for very extreme populations with unusually high fertility levels, there is bias in the parameter estimations. This is the case in the parameter estimations for all three models. It also shows that the waiting model that uses individual data has the least bias, but since it is also an approximation the predictions are a little off.

3.3 Simulation results

Tables 2 and 3 give the result for the Monte Carlo study. 10 000 repetitions has been made on each parameter setting according to Table 1. Table 2 shows the parameter estimates of (k, m) for the two models, the Poisson model and the normal approximated waiting model. The waiting model that uses individual data is not considered here since it takes to long to find the parameter estimates in this model. Table 3 gives the estimated mean and standard deviation of the intensity given the two different models. If we compare these results with the results from the simulation study that was performed in Arnqvist [2], we see that the estimates are more variable here. That is due to, of course, the censoring and the truncation of the data. However, in the mean estimation of the intensity we do no worse here. By inspection of Table 2, we notice that there is still bias in the

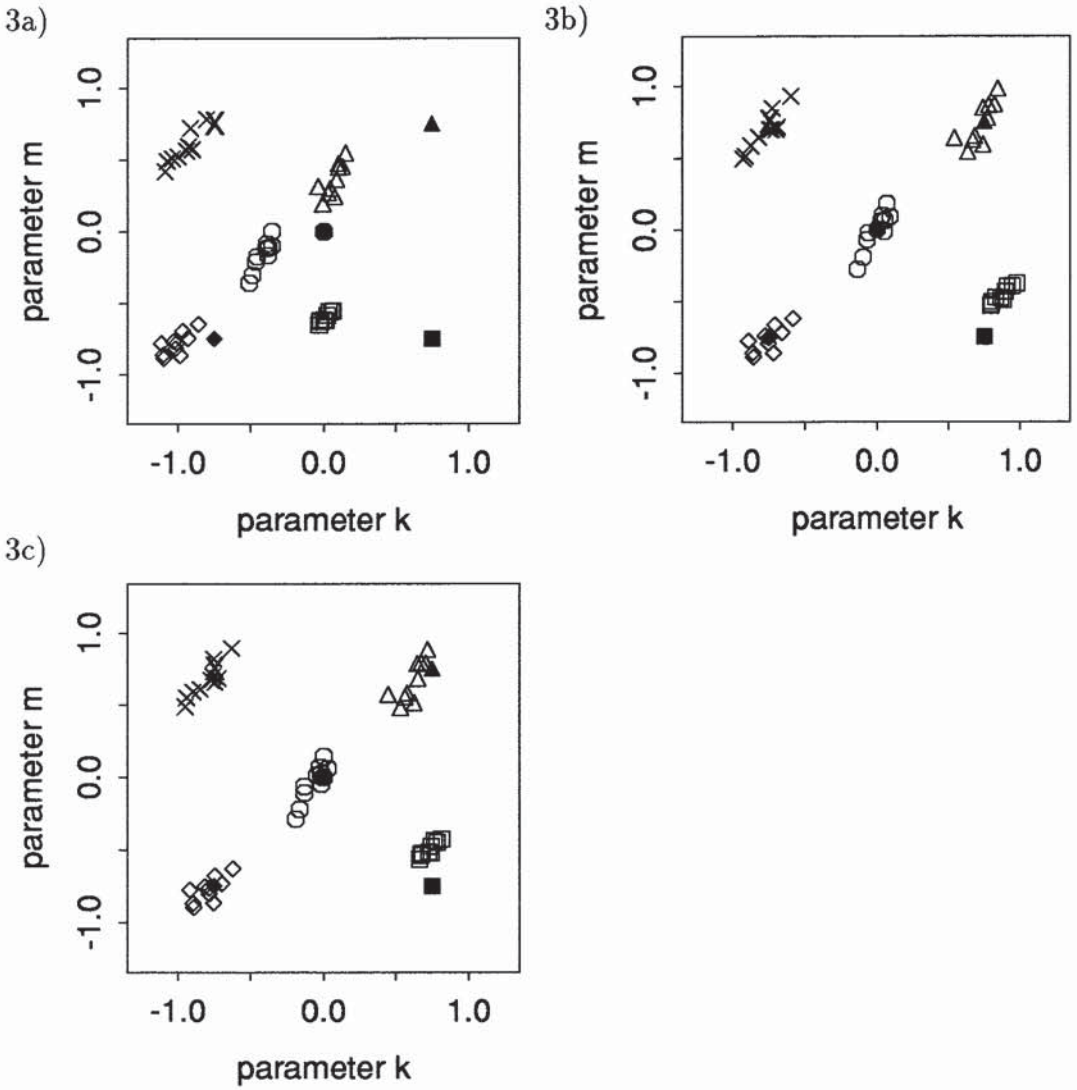

Figure 3: *Three pictures showing the parameter estimates of* (k, m) *in the intensity specification of five different parameter settings under three different models. Picture a) is the result of the analysis under the Poisson model, picture b) is for the normal approximation and picture c) is the result of the analysis in the waiting model using individual data. The simulated number of individuals is 100. It is assumed that the entering time is loglogistic distributed and the leaving time uniformly distributed for each individual.*

Table 2: *The parameter estimates from the two different models; The Poisson model (PM) and the normal approximated waiting model (NAWM). The number of replicated populations is* 10 000 *and* (k, m) *are the parameters in the intensity specification in the simulation of the waiting model.*

		PM			NAWM		
			mean	s.d.		mean	s.d.
k	0.00	$\hat{k}_p$	-0.411	0.068	$\hat{k}_w$	0.001	0.092
m	0.00	$\hat{m}_p$	-0.146	0.089	$\hat{m}_w$	0.010	0.111
k	-0.75	$\hat{k}_p$	-1.000	0.090	$\hat{k}_w$	-0.746	0.108
m	-0.75	$\hat{m}_p$	-0.758	0.099	$\hat{m}_w$	-0.741	0.115
k	-0.75	$\hat{k}_p$	-0.947	0.116	$\hat{k}_w$	-0.758	0.137
m	0.75	$\hat{m}_p$	0.632	0.178	$\hat{m}_w$	0.759	0.201
k	0.75	$\hat{k}_p$	-0.087	0.040	$\hat{k}_w$	0.771	0.068
m	-0.75	$\hat{m}_p$	-0.809	0.046	$\hat{m}_w$	-0.725	0.072
k	0.75	$\hat{k}_p$	0.081	0.048	$\hat{k}_w$	0.754	0.082
m	0.75	$\hat{m}_p$	0.375	0.079	$\hat{m}_w$	0.772	0.111

estimates when the fertility intensities are very large. Large, in this study, means the cases when $(k, m) = (0.75, -0.75)$ and $(k, m) = (0.75, 0.75)$. Of course, it is not possible to compare $\hat{k}_p$ with $\hat{k}_w$ or $\hat{m}_p$ with $\hat{m}_w$. They are different estimates within two different model formulations. However, if we use the transformation that was suggested in Arnqvist [2], it is possible to compare the intensity estimate θ_a, $a = 1, \ldots, 6$, within the two models. In Table 3, we can see that the normal approximation of the waiting model still outperforms the Poisson model for larger intensities. This means that the estimation of the intensity is closer to the true intensity when the normal approximation of the waiting model is used. So, again it can be stressed that the parameter estimates within the Poisson model/Coale Trussel model are not reliable for very productive populations. In summary, we can say that the normal approximation of the waiting model beats the Poisson model when we want to estimate the intensity θ_a, even if the estimation is a little biased.

Table 3: *This table gives the mean and the standard deviation of the intensity estimates from the two different models. First, in the third column is the true intensity used in the simulation (TI) given, next, columns four and five, give the mean and standard deviation from the intensity estimation using the Poisson model (PM), and the last two columns give the results from the estimation when using the normal approximated model (NAWM). The number of replicated populations is* 10 000 *and* (k, m) *are the parameters in the intensity specification used in the simulation of the data.*

Parameters		Interval	TI	PM		NAWM	
			θ	$\hat{\theta}_p^*$	st.dev.	$\hat{\theta}_w$	st.dev.
k	m	20-24	0.460	0.442	0.043	0.463	0.043
0.00	0.00	25-29	0.431	0.425	0.029	0.431	0.029
		30-34	0.395	0.407	0.021	0.393	0.020
		34-39	0.322	0.332	0.023	0.320	0.020
		40-44	0.167	0.158	0.014	0.166	0.015
		45-49	0.024	0.021	0.002	0.024	0.003
k	m	20-24	0.217	0.205	0.022	0.219	0.024
-0.75	-0.75	25-29	0.251	0.244	0.020	0.252	0.021
		30-34	0.308	0.318	0.019	0.308	0.017
		35-39	0.332	0.354	0.023	0.331	0.020
		40-44	0.228	0.220	0.020	0.227	0.020
		45-49	0.040	0.032	0.003	0.040	0.004
k	m	20-24	0.217	0.220	0.031	0.217	0.030
-0.75	0.75	25-29	0.165	0.164	0.016	0.164	0.016
		30-34	0.113	0.112	0.010	0.112	0.010
		35-39	0.070	0.070	0.009	0.069	0.009
		40-44	0.027	0.028	0.005	0.027	0.005
		45-49	0.003	0.003	0.001	0.003	0.001
k	m	20-24	0.974	0.732	0.050	0.997	0.068
0.75	-0.75	25-29	1.125	0.984	0.055	1.142	0.059
		30-34	1.379	1.645	0.073	1.386	0.049
		35-39	1.489	2.195	0.141	1.483	0.055
		40-44	1.021	0.929	0.060	1.008	0.054
		45-49	0.178	0.093	0.004	0.175	0.012
k	m	20-24	0.974	1.002	0.096	0.981	0.080
0.75	0.75	25-29	0.740	0.729	0.042	0.740	0.044
		30-34	0.507	0.501	0.025	0.502	0.024
		35-39	0.312	0.310	0.022	0.307	0.021
		40-44	0.122	0.120	0.011	0.120	0.012
		45-49	0.015	0.014	0.001	0.014	0.002

4 Bias correction of the parameter estimates

For the very extreme cases with high fertility intensities it is apperent that the parameter estimates start to become biased; this happens in both in the situation with the waiting model using individual data, and in the case where we used the normal approximation of the waiting model, and had the summerized data according to Table 1. However, if the model we are assuming is the correct one, by using bootstrap it is possible to bias correct the estimates. Below is a way to do that in the normal approximated waiting model.

Assume that a probability distribution $\mathbf{F}$ has given us data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Here x_i is a path of realizations from our underlying model. Each realization means a pregnancy or birth for the individual $N_W(t)$ in her productive lifetime. It is also assumed that the individuals are productive in the age interval [20,50] years. Each realisation is given to us by random sampling, $\mathbf{F} \rightarrow \mathbf{x}$. The parameter of interest $\boldsymbol{\psi} = \mathbf{t}(\mathbf{F})$, is four dimensional $\boldsymbol{\psi} = (k, m, \alpha, \beta)$. Here (k, m) are the parameters that specify the intensity θ that the population produces children, and (α, β) are the parameters in the loglogistic distribution that specify the marriage times. Let the estimator be $\hat{\boldsymbol{\psi}} = \mathbf{s}(\mathbf{x})$. Then, the bias can be defined to

$$\mathbf{b}_{\mathbf{F}} = bias_{\mathbf{F}}(\hat{\boldsymbol{\psi}}, \boldsymbol{\psi}) = \mathrm{E}_{\mathbf{F}}[\mathbf{s}(\mathbf{x})] - \mathbf{t}(\mathbf{F}),$$

and the definition of the bootstrap bias is

$$\mathbf{b}_{\hat{\mathbf{F}}} = \mathrm{E}_{\hat{\mathbf{F}}}[\mathbf{s}(\mathbf{x}^*)] - \mathbf{t}(\hat{\mathbf{F}}),$$

where $\hat{\mathbf{F}}$ is substituted for $\mathbf{F}$, and $\mathbf{x}^*$ is for the bootstrap samples. In this case, the $bias_{\hat{\mathbf{F}}}$ must be approximated, and this is done through Monte Carlo simulation. By producing independent bootstrap samples $\mathbf{x}^{*1}, \ldots, \mathbf{x}^{*B}$ the bootstrap expectation $\mathrm{E}_{\hat{\mathbf{F}}}[\mathbf{s}(\mathbf{x}^*)]$ is approximated by the average

$$\boldsymbol{\psi}^*(\cdot) = \sum_{b=1}^{B} \boldsymbol{\psi}^*(b)/B = \sum_{b=1}^{B} \mathbf{s}(\mathbf{x}^{*b})/B.$$

Finally, the bootstrap estimate of bias, which is based on the B replicates, is

$$\widehat{bias}_B = \boldsymbol{\psi}^*(\cdot) - \mathbf{t}(\hat{\mathbf{F}}).$$

This is saying that if we can reproduce the model under investigation by using our estimated values $\hat{\boldsymbol{\psi}}$, then the deviation, or bias, in the parameter estimates will be mimicked again, but this time it is possible to estimate the deviation.

4.1 Numerical illustration

The most interesting situation where the bias is essential is when the populations have extremely high fertility rates. To find out the bias, one bootstrap scheme could be as following:

We do not have real data, so to illustrate this we start with the simulation of 10 populations in the same way as in the previous section 3.

- Use $\psi_0 = (k_0, m_0, \alpha_0, \beta_0)$ as start parameters.

Here (k_0, m_0), are for the intensity estimation and (α_0, β_0) are used in the loglogistic distribution which specifies the marriage times. Simulate 10 populations using ψ_0 with 100 individuals, giving

$$\left\{ \begin{array}{c} \begin{bmatrix} B_{11} & \dots & B_{16} \\ E_{11} & \dots & E_{16} \end{bmatrix} \\ \begin{bmatrix} B_{21} & \dots & B_{26} \\ E_{21} & \dots & E_{26} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} B_{101} & \dots & B_{106} \\ E_{101} & \dots & E_{106} \end{bmatrix} \end{array} \right.$$

and take out 50 marriage times. This number can be larger, but we assume that a lot of work is involved in finding this by reading church books, and therefore, we only try to find out a small number.

Estimate the parameters using the previous table and the 50 collected marriage times. This gives us

$$\hat{\psi} = (\hat{k}, \hat{m}, \hat{\alpha}, \hat{\beta}).$$

Use $\hat{\psi}$ to generate B new generations, summerize them according to Table 1, and generate B times 50 marriage times to estimate $\hat{\psi}$. This gives

$$\begin{array}{c} \hat{\psi}^{*1} = (\hat{k}^{*1}, \hat{m}^{*1}, \hat{\alpha}^{*1}, \hat{\beta}^{*1}) \\ \vdots \\ \hat{\psi}^{*B} = (\hat{k}^{*B}, \hat{m}^{*B}, \hat{\alpha}^{*B}, \hat{\beta}^{*B}). \end{array}$$

Then estimate the bias for the different parameters according to

$$bias_{\hat{\psi}}(\hat{k}) = \frac{1}{B}\sum_{b=1}^{B}\hat{k}^{*b} - \hat{k},$$

$$bias_{\hat{\psi}}(\hat{m}) = \frac{1}{B}\sum_{b=1}^{B}\hat{m}^{*b} - \hat{m},$$

$$bias_{\hat{\psi}}(\hat{\alpha}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\alpha}^{*b} - \hat{\alpha},$$

and

$$bias_{\hat{\psi}}(\hat{\beta}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^{*b} - \hat{\beta}.$$

These biased estimations can now be used to correct the original estimates. Actually, since we do not know anything of these new bias corrected estimators, one can use double bootstrap see Hjort [8] page 99, to find out the performans of them. This is not presented here, but preliminary investigations indicate no or very small increase of the variance.

The bootstrap method is illustrated for the case when the intensity in the population is assumed to be given by $(k = 0.75, m = -0.75)$. This means that the mean number of births denoted E$[B]$, for populations consisting of 100 individual becomes according to Table 3, if it is assumed that the individuals have complete exposure times in each interval.

Table 3: *The table gives the fertility intensity and the mean number of births in the five year intervals* $[20,24],[25,29],\ldots,[44,49]$ *for a population with 100 individuals with the parameters* $(k = 0.75, m = -0.75)$. *It is assumed that the individuals have complete exposure times in each interval.*

θ	0.974	1.125	1.379	1.489	1.021	0.178
$E[B]$	487	562	690	745	510	89

The start vector is

$$\boldsymbol{\psi_0} = (k = 0.75, m = -0.75, \alpha = -3.20, \beta = 8.90).$$

This gave 10 populations with 100 individuals. In this case, the populations are merged together into one big population with 1000 individuals and summerized according to Table 1. The estimation of $\boldsymbol{\psi_0}$ gave

$$\hat{\boldsymbol{\psi}} = (\hat{k} = 0.84, \hat{m} = -0.65, \hat{\alpha} = -3.21, \hat{\beta} = 7.90),$$

where $(\hat{\alpha}, \hat{\beta})$ are estimated by taking a random sample of 50 marriage times from the merged population. $\hat{\boldsymbol{\psi}}$ will be used as the starting value of the bootstrap simulation. Now $B = 999$ bootstrap replicates were made and the results are given in Table 4. The mean value of the bootstrap estimates are now used to estimate the bias that the model gives.

Table 4: *Summary statistics of the bootstrap estimates of the parameter vector* $\hat{\boldsymbol{\psi}}$.

parameter	1st.quant.	median	2nd.quant.	mean	st.dev.
$\hat{k}$	0.921	0.936	0.950	0.935	0.021
$\hat{m}$	−0.576	−0.562	−0.547	−0.562	0.021
$\hat{\alpha}$	−3.229	−3.211	−3.191	−3.209	0.029
$\hat{\beta}$	7.430	8.014	8.713	8.112	0.968

$$bias_{\hat{\psi}}(\hat{k}) = \frac{1}{B}\sum_{b=1}^{B} \hat{k}^{*b} - \hat{k} = 0.935 - 0.838 = 0.097$$

$$bias_{\hat{\psi}}(\hat{m}) = \frac{1}{B}\sum_{b=1}^{B} \hat{m}^{*b} - \hat{m} = -0.562 - (-0.649) = 0.087$$

$$bias_{\hat{\psi}}(\hat{\alpha}) = \frac{1}{B}\sum_{b=1}^{B} \hat{\alpha}^{*b} - \hat{a} = -3.209 - (-3.210) = 0.001$$

$$bias_{\hat{\psi}}(\hat{\beta}) = \frac{1}{B}\sum_{b=1}^{B} \hat{\beta}^{*b} - \hat{b} = 8.112 - 7.904 = 0.208$$

This suggests the need to bias adjust the parameter vector $\hat{\psi}$ according to

$$\hat{k} = 0.838 - 0.097 = 0.741$$

$$\hat{m} = -0.649 - 0.087 = 0.736$$

$$\hat{\alpha} = -3.21 - (-0.001) = -3.20$$

$$\hat{\beta} = 7.904 - 0.208 = 7.696$$

In this case, we see that we improve our estimates by using the bias correction for three of the parameters but not for β. If we look at the standard deviation estimate of $\hat{\beta}$ in Table 4, it can be seen that it is very large compared to the others.

This method can be performed for the different parameter values of (k, m) but the gain will not be so much in the other parameter settings, (see Table 1 and Figure 3).

Table 5: *Estimation of the parameter vector ψ with $\hat{\psi}$ which is not bias corrected and $\hat{\psi}^*$ which is bias corrected.*

	k	m	α	β
ψ_0	0.75	−0.75	−3.20	8.90
$\hat{\psi}$	0.84	−0.65	−3.21	7.90
$\hat{\psi}^*$	0.74	−0.74	−3.20	7.70

5 Summary

In this report, the normal approximated waiting model which was derived in Arnqvist [2] is modified. Here, it allows the summerized fertility data according to Table 1 to consist of both censored and left truncated observations. It is further stressed again that the normal approximated waiting model better describes fertility intensity θ, expressed in the same way as Coale-Trussells specification of fertility intensity, than the Coale-Trussell model or Poisson model.

It can be seen that for extreme populations with very high fertility intensity the parameter estimations becomes biased. This is the case for both models, normal approximated waiting model and Poisson model. However, for the Poisson model the bias is severe for the extreme populations, while in the normal approximated waiting model the bias is more moderate. Further, in this report a bootstrap method is suggested to bias correct the parameter estimates in the normal approximated waiting model.

It is shown that by using a better approximation of the variance it is possible to minimize this bias. But to use this better approximation, the derivatives of the new approximation of the variance with respect to (k, m) are needed. This has not been done and this suggests further research. Also ,of course, it will be interesting to apply the model to real data from the UN fertility services.

Acknowledgements

The author is very greatful to his supervisor Professor Yuri Belyaev for his support and helpful ideas during the development of this report.

6 Appendix

In section 3, we could see that the estimation of the parameters (k, m), that specifies the intensity θ, became somewhat biased for the very extreme intensities. This bias is mainly because of the approximation of the variance. The problem is that in the normal approximation of the waiting model, we use the asymptotic normal approximation of the mean and the variance, see (3) and (4), and this variance approximation causes the bias in the parameter estimates in the model derivation (6). This is due to the fact that the exposure times are not five years for all of the individuals. Actually, the variance approximation (4) was not correct when all of the individuals where assumed to be exposed five years, but when the exposure times become shorter the approximation becomes worse.

So, this suggests a finer approximation of the variance. One such approximation can be found in a textbook of renewal theory, [6]. It can be derived in the following way. Let

$$\psi(t) = \mathrm{E}[N_W(t)(N_W(t)+1)],$$

then

$$Var[N_W(t)] = \psi(t) - \mathrm{E}[N_W(t)] - (\mathrm{E}[N_W(t)])^2.$$

Now we can write

$$\begin{aligned} \psi(t) &= \sum_{b=0}^{\infty} b(b+1)P(N_W(t)=b) \\ &= \sum_{b=0}^{\infty} b(b+1)\{K_b(t) - K_{b+1}(t)\}, \end{aligned}$$

where $K_b(t)$ equals the cumulative distribution of $S_b = W+x_1+\ldots+W+x_b$. Here, as before, W is the waiting time after pregnancy and x_i is the active exposure time which is expontially distributed with intensity θ. Let $k_b(x)$ denote the p.d.f. of S_b, then the Laplace transform of $\psi(t)$ becomes

$$\begin{aligned} \psi^*(s) &= \frac{1}{s}\sum_{b=0}^{\infty} b(b+1)\{k_b^*(s) - k_{b+1}^*(s)\} \\ &= \frac{2}{s}\sum_{b=1}^{\infty} b\, k_b^*(s). \end{aligned}$$

For an ordinary renewal process, $k_b^*(s) = \{f^*(s)\}^b$, but for the equilibrium renewal process, which is the case here, we have that $k_b^*(s) = \{f^*(s)\}^{b-1}\{(1-f^*(s))/m_1\, s\}$ so

$$\psi^*(s) = \frac{2}{s^2\, m_1\, \{1 - f^*(s)\}}. \tag{7}$$

By expanding (7) near $s = 0$, we can find the asymptotic result, since first

$$\psi^*(s) = \frac{2}{s^3\, m_1^2}\left[1 + s\frac{m_1^2 + m_2}{2\, m_1} + s^2\left(\frac{m_1^2}{12} + \frac{m_2^2}{4\, m_1^2} - \frac{m_3}{6\, m_1}\right)\right] + o\left(\frac{1}{s}\right).$$

Here m_a, $a = 1, 2, 3$, means the central moments of the distribution of the complete exposure time between pregnancies . By taking an inversion of the Laplace transform, using Tauberian theorem, we have that

$$\psi(t) = \frac{t^2}{m_1^2} + t\frac{m_1^2 + m_2}{m_1^3} + \left(\frac{1}{6} + \frac{m_2^2}{2\, m_1^4} - \frac{m_3}{3\, m_1^3}\right) + o(1).$$

Now, $\mathrm{E}[N_W(t)] = t/m_1$ so we have that

$$Var[N_W(t)] = \frac{m_2\, t}{m_1^3} + \left(\frac{1}{6} + \frac{m_2^2}{2\, m_1^4} + \frac{m_3}{3\, m_1^3}\right) + o(1), \tag{8}$$

where $o(1)$ denotes a function of t tending to 0 as $t \to \infty$. The moments can easely be derived. If we denote the complete exposure time with S, we know that $S = W + x$, where we had that $x \sim exp(\theta)$. Then,

$$F(S) = P(S \leq s) = P(W + x \leq s) = 1 - e^{-\theta(s-W)},\ s \geq W,$$

and

$$f(s) = \frac{\partial F(s)}{\partial s} = \theta e^{-\theta(s-W)}.$$

This gives us

$$\mathrm{E}[S] = m_1 = \int_W^\infty s\,\theta\, e^{-\theta(s-W)}\, ds = W + \frac{1}{\theta},$$

$$m_2 = \int_W^\infty (s - m_1)^2\,\theta\, e^{-\theta(s-W)}\, ds = \frac{1}{\theta^2},$$

and finally

$$m_3 = \int_W^\infty (s - m_1)^3\,\theta\, e^{-\theta(s-W)}\, ds = \frac{2}{\theta^3}.$$

If these moments are substituted into (9), then we have after simplifications, that

$$Var[N_W(t)] \approx \frac{\theta\, t}{(1 + \theta\, W)^3} + \left(\frac{1}{6} + \frac{1}{2\,(1 + \theta\, W)^4} - \frac{2}{3\,(1 + \theta\, W)^3}\right). \tag{9}$$

The first fraction in (9) is the approximation that is used in (4). The correction, therefore, is the part between the brackets. In Figure 4, we see the approximations as functions of the intensity for 4 different interval lengths, $t = 1, 2, 3$ and 4 years of exposure. The approximations are the asymptotic normal approximation (4), the approximation using the third moment of the failure times (9) and the

"true variance" derived from the waiting model Arnqvist [1]. The formulae of the true variance is based on the exact probabilities of the number of births. It is a rather complicated form to be used in estimation of the true variance, and therefore, it is advisible to have an approximation of it. We can also see that (9) follows the true variance very well. This suggests using (9) instead of (4) as the variance approximation. By inspection of Figure 4, it can clearly be seen why we get bias in the estimation of the intensity. However, in this paper it is suggested to use the bootstrap bias correction and not the formulae (9).

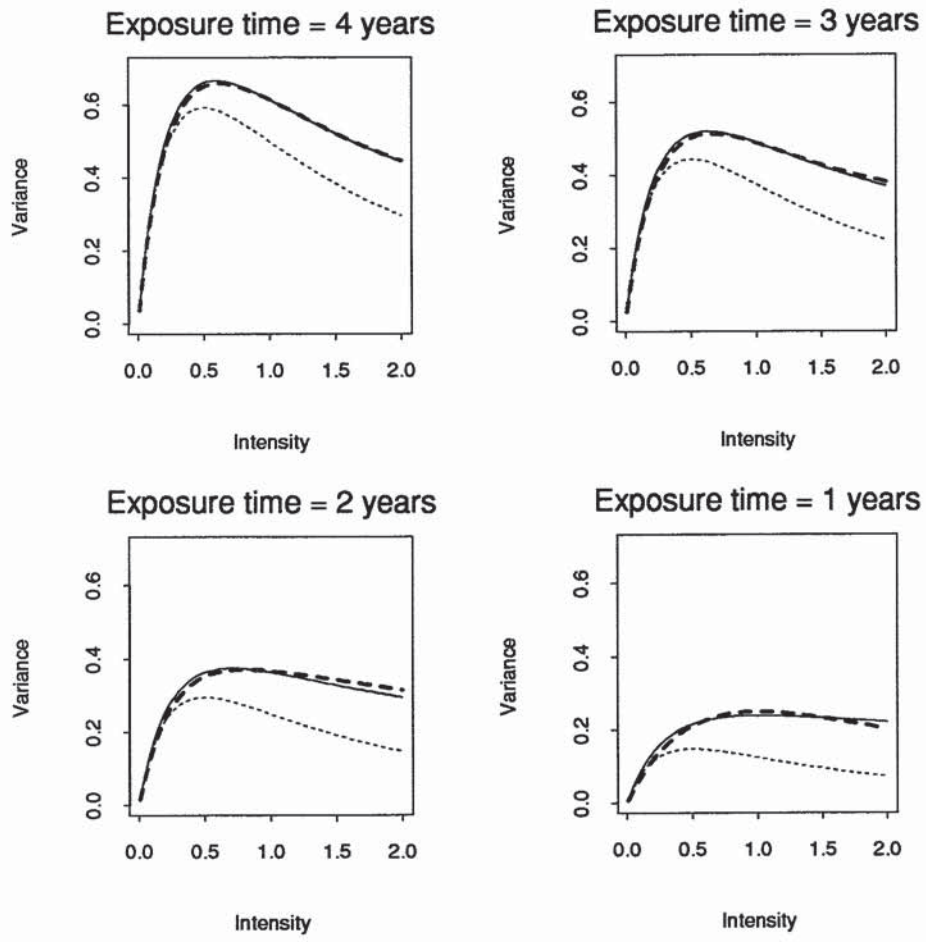

Figure 4: *The four pictures give the variance of the number of births as a function of the intensity for four different exposure times* $4, 3, 2, 1$ *years. The bold dashed line represents the variance as calculated from the waiting model derived in Arnqvist [1]. It is based on the exact formulae for the probabilities of the number of births. The dotted line is for the asymptotic normal approximated variance (4) and the solid line is for the approximation using higher moments (9).*

References

[1] Arnqvist Per (1995). Aspects on the Coale-Trussell model. *Statistical Research Report, no 1*, University of Umeå.

[2] Arnqvist Per (1995). Approximation of the waiting model. *Statistical Research Report, no 2*, University of Umeå.

[3] Bengtsson T. (1989). Mortality and causes of death in Västanfors parish, Sweden, 1700-1900. In Brändström, A. and Tedebrand, L-G (editors): Society, Health, and Population during the Demographic Transition. Umeå.

[4] Broström Göran (1985). Practical aspects on the estimation of the parameters in Coaleś model for marital fertility. *Demography*, **22**, 625-631.

[5] Coale, A.J. & Trussell, T.J. (1974). Model fertility schedules: Variations in the age structure of childbearing in human populations. *Population Index*, **40**, 185-258.

[6] Cox, D.R. (1970). *Renewal theory*, Chapman & Hall, London.

[7] Cox, D.R. & Isham, V. (1980). *Point processes*, Chapman & Hall, London.

[8] Hjorth, Urban J.S. *Computer intensive statistical methods* Chapman & Hall, London.

[9] Trussell, T.J. (1985). Mm (Computer program), Princeton, NJ: Princeton university office of population research.

[10] United Nations (1966). *demographic yearbook 1965*, New York: United Nations, Department of Economics and Social Affairs.