Prediction and Error Propagation in Cohort Diffusion Models

Mikko Myrskylä[1]
Joshua R. Goldstein[2]

February 6, 2009

Abstract

We study prediction and error propagation in the Gompertz, logistic, and Hernes cohort diffusion models. We show that the models can be treated in a unifying framework in which the models are linearized with respect to cohort age and predictions and prediction variance are derived from the underlying linear process. We develop and compare different methods for deriving predictions from the underlying linear process and show that a midpoint method, which has not been used in cohort diffusion models, improves accuracy over standard methods. For an important special case, random walk with drift, we develop an analytical prediction variance estimator and study its accuracy with respect to a Monte Carlo estimator. Simulation studies and empirical applications to first births and marriages show that the analytical estimator is accurate, allowing forecasters to make precise the level of within-model prediction uncertainty.

[1] Population Studies Center, 239 McNeil Building, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, USA. Email: myrskylm@pop.upenn.edu, phone: +1 267 235 7257.

[2] Director; Head of the Laboratory of Economic and Social Demography, Max Planck Institute for Demographic Research, and Professor of Sociology and Public Affairs, Princeton University. Konrad-Zuse-Straße 1, 18057 Rostock, Germany. Email: goldstein@demogr.mpg.de, phone +49 (0)381 2081 107, fax:+49 (0)381 2081 407

1 Introduction

Diffusion models have proven to be useful in forecasting uncompleted cohort experience. Goldstein and Kenney (2001) and Li and Wu (2008) show that the Hernes model (Hernes 1972) can be used for predicting marriage rates. It was long believed that the Gompertz model was inadequate for predicting fertility (Hoem, Madsen et al. 1981; Pollard and Valkovics 1992), but recent research (Goldstein 2008) suggests that if fit to the cohort rates, instead of fitting the model to period rates as was common in the early literature, the Gompertz model actually performs quite well. In principle, also the logistic model can be also be used to forecast cohort experience, but while the model has been used to explain fertility patterns (Ike 2002), it has not been used for forecasting purposes in the cohort context. The model has, however, been used extensively in the economic literature to forecast the diffusion of innovation (Mar-Molinero 1980; Harvey 1984; Gruber and Verboven 2001; Meade and Islam 2006).

Irrespective of the context, it is a common practice to linearize the diffusion model before estimation (Harvey 1984; Frances 1994; Li and Wu 2008). When forecasting is the goal, this approach has obvious advantages over some other methods such as fitting the diffusion curve to observed cumulative proportions (Hernes 1972; Goldstein and Kenney 2001; Martin 2004; Billari and Toulemon 2006). In prior research, the linear processes have been usually modeled as static time trends (Frances 1994; Li and Wu 2008) or, in the rare cases where the process has had a dynamic, autoregressive structure, no attempt to derive prediction variance has been made (Harvey 1984).

Our aim is to provide a unified framework for time series based estimation, prediction and prediction error estimation in the Gompertz, logistic, and Hernes cohort diffusion models. We build on prior research on cohort diffusion models by i) treating the underlying linear process as a dynamic time series process; ii) showing how predictions based on the underlying linear process can be improved using the midpoint method, a method is often used in the numerical analysis of differential equations; and iii) deriving an analytical variance estimator for the predictions in an important special case, random walk with drift. Empirical applications to first births and marriages suggest that the random walk based

cohort diffusion models may be useful in predicting the future experience of a cohort and in quantifying the prediction uncertainty.

The paper is organized as follows. In Section 2, we briefly introduce our approach in a non-technical way. In Sections 3-5, we show how estimation, prediction and prediction error estimation can be done in the Gompertz, logistic and Hernes cohort diffusion models using the time series approach. In Section 6 we apply the models to simulated and empirical data. Section 7 discusses the results. The Appendix provides certain equations and formulas which are used throughout the paper and a summary table of the most important results.

2 Overview of the time series approach

The idea of linearizing a diffusion[3] model, fitting a regression model to the underlying linear process, and deriving predictions from the linear process is not new. For example, Winsor (1932) shows how the logistic and Gomperts models can be linearized with respect to time, and Harvey (1984) takes the next step by showing how the predictions of a logistic model can be constructed from an autoregressive integrated moving average (ARIMA) time series model fit to the linearized part. More recently, Li and Wu (2008) use the Hernes model to predict first births, and follow Winsor and Harvey by first linearizing the model and then fitting a regression model to the underlying linear process.

The steps in the process of obtaining predictions and prediction error estimates from an underlying linear process are as follows. Let P_t denote the proportion in a cohort "infected" by age t – that is the proportion of those who, depending on the application, have married, have experienced a first birth, or more generally have adopted the innovation that is being modeled. We assume that P_t depends on age t through a monotonic increasing function F: $P_t = F(t)$. The following steps are needed to produce a time series modeling based prediction and prediction intervals of P at age $t+k$, given observations up to t:

1. Find a linearization H so that $H(P_t) \equiv g_t$ is linear in cohort age t. We call g_t the underlying linear process.

2. Model g_t as time series process (e.g., ARIMA model), and estimate the parameters of the model using standard techniques as detailed in for example Hamilton (Hamilton 1994).

Repeat steps 3-4 for $i = 1,...,k$:

3. Construct a one-step ahead prediction $\hat{g}_{t+i}$ for the underlying linear process and derive a one-step ahead prediction $\hat{P}_{t+i}$ from $\hat{g}_{t+i}$ using the inverse of H.

[3] Depending on the context, these models are also called growth curve models, or growth models.

4. Estimate the variance of $\hat{P}_{t+i}$. The source of the variance is the randomness in the underlying linear process identified in step 2.

A few comments are in place here. First, the linearization of the model in step 1 may not be unique. Further, the linearization is constructed using a continuous notation for the diffusion model. Data, however, is invariable discrete. The way continuous notation is translated to accommodate discrete data, most importantly the way derivatives are treated, has implications on the predictions. Second, our empirical analysis indicates that the underlying linear process may be accurately described by a simple model such as random walk with drift (ARIMA(0,1,0)). Third, transforming the predictions $\hat{g}_{t+i}$ for the underlying linear process into predictions $\hat{P}_{t+i}$ may not be straightforward because H is defined for continuous time but the observations are in discrete time. Moreover, the predictions $\hat{P}_{t+i}$ invariably involve the past value $\hat{P}_{t+i-1}$, therefore the predictions need to proceed recursively. Finally, when estimating the variance in step 4, the errors cumulate rather than fade away if the model for g_t includes unit roots (as does the random walk with drift model).

The Sections 3-5 show how this approach is operationalized for the Gompertz, logistic and Hernes models. The Section 3 for the Gompertz model is the most detailed, since the logistic and Hernes cases are very much analogous to the Gompertzian case. To anticipate the results, Appendix Table 1 summarizes the model equations, linearizations, models for the underlying linear process, prediction equations and analytical prediction variance estimators for the Gompertz, logistic and Hernes models.

3 The Gompertz diffusion model

3.1 The model

Let P_t be the proportion in a cohort that has by age t adopted the innovation under study. Throughout the paper we assume that we have observed $P_0, P_1, ..., P_t$ and that $P_{t+1}, P_{t+2}, ..., P_{t+k}$ are being predicted. The Gompertz growth model for a proportion P_t is

$$P_t = k \exp\left[-\exp\left(a - bt\right)\right]. \tag{3.1}$$

For a behavioral interpretation of the Gompertz model see Goldstein (2008). Log of the log-derivative linearizes the model to $\ln b + a - bt$. To accommodate the model for discrete data, we use the discretization $\frac{d \ln P_t}{dt} \approx \frac{1}{P_t} \frac{P_{t+1} - P_{t-1}}{2}$ (see Appendix (8.2)), proposed by Li and Wu (2008) in the context of the Hernes model. With this linearization we have

$$\ln b + a - bt \approx \ln\left(\frac{1}{P_t} \frac{P_{t+1} - P_{t-1}}{2}\right) \equiv g_t. \tag{3.2}$$

We model the underlying linear process g_t as a time series process. In the case of a random walk with drift, the model is

$$g_t = g_{t-1} + \delta + \varepsilon_t = g_0 + \delta t + \sum_{i=1}^{t} \varepsilon_i, \quad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right) \tag{3.3}$$

and the model parameters $\left(\delta, \sigma_\varepsilon^2\right)$ are estimated by[4]

$$\hat{\delta} = \frac{g_{t-1} - g_1}{t-2} \quad \text{and} \quad \hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{t-1}\left(g_i - g_{i-1} - \hat{\delta}\right)^2}{t-3}. \tag{3.4}$$

[4] In (3.2), the number of observations drops from $t+1$ to $t-1$.

3.2 Prediction

One-step and k-step ahead predictions $\hat{P}_{t+1}$ and $\hat{P}_{t+k}$ are based on predictions for the underlying linear process. For the case of a random walk with drift, the predictions are $\hat{g}_{t+1} = g_t + \hat{\delta}$ and $\hat{g}_{t+k} = g_t + \hat{\delta}k$. To derive the predictions $\hat{P}_{t+1}$ and $\hat{P}_{t+k}$ from the underlying linear process we need the approximation (8.3), $0.5 \cdot (P_{t+1} - P_{t-1}) \approx P_t - P_{t-1}$. This is done as follows. First note that for a Gompertz model, $\exp(g_t)$ describes proportional change. This can be approximated by

$$\exp(g_t) = \frac{1}{P_t}\frac{P_{t+1} - P_{t-1}}{2} \approx \frac{1}{P_t}(P_t - P_{t-1}) = 1 - \frac{P_{t-1}}{P_t}. \tag{3.5}$$

Using the right hand side expression for g_t in (3.5) we can approximate P_t in terms of the previous observed proportion, P_{t-1}, and current value of the underlying process, g_t: $P_t \approx P_{t-1}/[1 - \exp(g_t)]$. Similarly, one-step ahead prediction $\hat{P}_{t+1}$ can be expressed in terms of the last observed proportion P_t and predicted value of the underlying linear process $\hat{g}_{t+1}$:

$$\hat{P}_{t+1} = \frac{P_t}{1 - \exp(\hat{g}_{t+1})}. \tag{3.6}$$

By applying (3.6) recursively we get the k-step ahead predictions. These predictions, however, are still preliminary: predictions based on (3.6) will underestimate P_{t+k} because a discrete growth factor $\exp(\hat{g}_{t+1})$ is applied to P_t, whereas optimally one would apply a continuous growth factor to all values between $\hat{P}_{t+1}$ and P_t. Obviously, if the step length is small enough the problem is negligible. We reduce the bias by splitting the step into two parts and applying the factor $\exp(\hat{g}_t)$ to the first part and the factor $\exp(\hat{g}_{t+1})$ to the second part. The method is analogous to the midpoint method which is a refinement of the Euler method for solving differential equations numerically (Griffiths and Smith 1991). The method can be applied in two steps or by taking the average of $\exp(\hat{g}_{t+1})$ and $\exp(\hat{g}_t)$ and

applying that to P_t.[5] For simplicity, we use the latter approach. The mid-point modified one-step and k-step ahead predictions are

$$\hat{P}_{t+1} = \frac{P_t}{1-\exp\left[0.5\cdot\left(\hat{g}_{t+1}+g_t\right)\right]} \quad \text{and} \quad \hat{P}_{t+k} = \frac{\hat{P}_{t+k-1}}{1-\exp\left[0.5\cdot\left(\hat{g}_{t+k}+\hat{g}_{t+k-1}\right)\right]}. \tag{3.7}$$

3.3 Prediction variance

We develop an analytical and a Monte Carlo estimator for the variance $V\left(\hat{P}_{t+j}\right)$ for $j=1,...,k$.

3.3.1 An analytical variance estimator

The analytical variance estimator is based on two approximations; first we approximate the predictions and then we approximate the variance using the delta method (8.4) and the Taylor series approximation (8.6). For small $\exp\left(\hat{g}_{t+j}\right)$ (that is large, negative $\hat{g}_{t+j}$) the predictions (3.7) can be approximated as

$$\hat{P}_{t+1} \approx P_t + \exp\left(\hat{g}_{t+1}\right) \quad \text{and} \quad \hat{P}_{t+k} \approx P_t + \sum_{i=1}^{k}\exp\left(\hat{g}_{t+i}\right). \tag{3.8}$$

These predictions are linear in $\exp\left(\hat{g}_{t+j}\right)$, so their variance is easier to derive than the variance of the predictions (3.7). We derive the one-step and k-step ahead prediction variances as follows.

Variance for one-step ahead predictions

For the one-step ahead prediction $\hat{P}_{t+1} = P_t + \exp\left(\hat{g}_{t+1}\right)$ the variance is

$$V\left(\hat{P}_{t+1}\right) = V\left[\exp\left(\hat{g}_{t+1}\right)\right] \tag{3.9}$$

[5] This is not exactly the same as dividing the step into two parts and applying two separate growth factors to each part, but empirically the difference is negligible.

because P_t is a constant. The delta method approximation for $V\left[\exp\left(\hat{g}_{t+1}\right)\right]$ is given by

$$V\left[\exp\left(\hat{g}_{t+1}\right)\right] = V\left(\hat{g}_{t+1}\right)\left[\frac{d\exp\left[E\left(\hat{g}_{t+1}\right)\right]}{dx}\right]^2. \tag{3.10}$$

We assume that the contribution of the uncertainty in the drift estimate to the prediction variance is negligible. Then

$$V\left(\hat{g}_{t+1}\right) = E\left(g_t + \hat{\delta} - g_t - \delta - \varepsilon_{t+1}\right)^2 \approx E\left(\varepsilon_{t+1}\right)^2 = \sigma_\varepsilon^2, \tag{3.11}$$

and

$$\frac{d\exp\left[E\left(\hat{g}_{t+1}\right)\right]}{dx} = \exp\left[E\left(\hat{g}_{t+1}\right)\right] = \exp\left(g_t + \delta\right). \tag{3.12}$$

Plugging (3.11) and (3.12) into (3.10) we get the variance for the one-step ahead prediction:

$$V\left(\hat{P}_{t+1}\right) = \sigma_\varepsilon^2 \exp\left(2g_t + 2\delta\right). \tag{3.13}$$

The variance (3.13) is estimated by replacing σ_ε^2 and δ by their estimators, given in (3.4).

Variance for k-step ahead predictions

The variance of $\hat{P}_{t+k} = P_t + \sum_{i=1}^{k}\exp\left(\hat{g}_{t+i}\right)$ is a double sum of the covariances:

$$V\left[\sum_{i=1}^{k}\exp\left(\hat{g}_{t+i}\right)\right] = \sum_{i=1}^{k}\sum_{j=i}^{k}\operatorname{cov}\left[\exp\left(\hat{g}_{t+i}\right), \exp\left(\hat{g}_{t+j}\right)\right]. \tag{3.14}$$

The diagonal elements of the covariance matrix can be estimated using the delta method as

$$V\left[\exp\left(\hat{g}_{t+i}\right)\right] = i\sigma_\varepsilon^2 \exp\left(2g_t + 2i\delta\right). \tag{3.15}$$

Simulation experiments indicated that the off-diagonal elements $\operatorname{cov}\left[\exp\left(\hat{g}_{t+i}\right),\exp\left(\hat{g}_{t+j}\right)\right]$, $i \neq j$, contribute significantly to the variance. The reason for this is the double-counting of the errors: shocks ε_t up to $t = i$ are both in g_{t+i} and g_{t+j}, provided $j \geq i$. These off-diagonal elements can be approximated using the first order Taylor series approximation as

$$\operatorname{cov}\left[\exp\left(\hat{g}_{t+i}\right),\exp\left(\hat{g}_{t+j}\right)\right] \approx \min\left(i,j\right)\cdot\sigma_{\varepsilon}^{2}\cdot\exp\left(g_t + i\delta\right)\exp\left(g_t + j\delta\right). \tag{3.16}$$

The interpretation for (3.16) is the following. There are $\min\left(i,j\right)$ common shocks ε_t in g_{t+i} and g_{t+j}, each contributing σ_{ε}^{2} to the covariance, and the exponential terms of the form $\exp\left(g_t + i\delta\right)$ which are present both in the diagonal terms in (3.15) and in the off-diagonal terms in (3.16) scale the covariance proportionally to the size of the terms $\exp\left(\hat{g}_{t+i}\right)$. Note that for $i = j$, the equation for off-diagonal elements (3.16) reduces to the equation (3.15) for the diagonal elements.

The k-step ahead prediction variance is obtained by plugging (3.15) and (3.16) into (3.14):

$$V\left(\hat{P}_{t+k}\right) = \sigma_{\varepsilon}^{2}\exp\left(2g_t\right)\sum_{i=1}^{k}\sum_{j=1}^{k}\min\left(i,j\right)\cdot\exp\left[\delta\left(i+j\right)\right]. \tag{3.17}$$

First order Taylor series approximation applied directly to (3.14) would deliver the same variance estimator (3.17).

The estimators (3.13) and (3.17) reveal important facts about the nature of prediction uncertainty in cohort diffusion models. First, the multiplying factor σ_{ε}^{2} shows that the prediction variance grows linearly with the variance of the error term ε. Second, the factor $\exp\left(2g_t\right)$ implies that if the predictions are made at a late age (so t is large and g_t negative and large, as the drift δ in g is always negative), the prediction variance is small. If the predictions are made at an early age, then t is small, g_t is less negative, and the variance is large. Finally, the term $\exp\left(\delta\right)$ in (3.13) and (3.17) implies that if the drift in g is large (the drift is always negative), meaning that diffusion takes place soon, the prediction variance is

small. If, however, the drift is closer to 0 and diffusion happens at a slow pace and, the prediction variance is large. The same remarks apply also to the logistic and Hernes models (see Sections 4 and 5).

3.3.2 Monte Carlo variance estimator

A simple Monte Carlo variance estimator can be based on simulated paths of the underying linear process. In the case of a random walk with drift, we simulate $K = 1,000$ sample paths $g_{t+1}, g_{t+2}, ..., g_{t+k}$ using the equation

$$g_{t+j} = g_t + \hat{\delta} j + \sum_{i=1}^{j} \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \hat{\sigma}_\varepsilon^2\right). \tag{3.18}$$

The simulated paths of g are transformed to predictions $\hat{P}$ using the prediction equation (3.7). The variance and parametric or non-parametric confidence intervals can be calculated from the simulated realizations of P. We use the 0.025 and 0.975 percentiles of the simulated prediction distribution as the lower and upper bounds for the 95 % confidence interval for the predictions.

Appendix Table 1 summarizes the important results of the Section 3: The Gompertz model.

4 The logistic diffusion model

4.1 The model

As in the Gompertz case, let P_t be the proportion in a cohort that has by age t adopted the innovation, $P_0, P_1, ..., P_t$ the observed proportions and $P_{t+1}, P_{t+2}, ..., P_{t+k}$ the yet to be observed proportions we wish to predict. The logistic diffusion model for a proportion P_t is

$$P_t = \frac{a}{1+\exp(a-bt)}. \tag{4.1}$$

For a behavioral interpretation of the logistic diffusion model see Mansfield (1963). The model is linearized by $\ln\left(\frac{dP_t}{dt}\frac{1}{P_t^2}\right) = \ln b + a - bt$. To accommodate the model for discrete data, we use the discretization $dP_t/dt \approx 0.5\cdot(P_{t+1} - P_{t-1})$ (see Appendix (8.1)). This gives us

$$\ln b + a - bt \approx \ln\left(\frac{P_{t+1}-P_{t-1}}{2}\frac{1}{P_t^2}\right) \equiv g_t. \tag{4.2}$$

We model the underlying linear process g_t as a time series process. In the case of a random walk with drift, the model is given by (3.3) and the model parameters are estimated by (3.4).

4.2 Prediction and variance estimation

Predictions for the underlying linear process are used to derive predictions $\hat{P}_{t+j}$. In order to be able to express P_t in terms of P_{t-1} and g_t, we use the approximation

$$\frac{P_{t+1}-P_{t-1}}{2}\frac{1}{P_t^2} \approx (P_t - P_{t-1})\frac{1}{P_{t-1}^2}. \tag{4.3}$$

Noting that $\exp(g_t) = \frac{P_{t+1}-P_{t-1}}{2}\frac{1}{P_t^2}$, we how have an approximate expression for P_t in terms of P_{t-1} and g_t: $P_t = P_{t-1} + P_{t-1}^2\exp(g_t)$. The predictions can then be constructed as

(4.4) $$\hat{P}_{t+1} = P_t + P_t^2 \exp\left(\hat{g}_{t+1}\right) \quad \text{and} \quad \hat{P}_{t+k} = \hat{P}_{t+k-1} + \hat{P}_{t+k-1}^2 \exp\left(\hat{g}_{t+k}\right).$$

Harvey (1984) presents the same prediction equations for the logistic diffusion model. Predictions based on (4.4), however, underestimate P_{t+k} for the same reason the predictions (3.6) underestimates P_{t+k} in the Gompertz case: The growth factor is applied to P_t, instead of applying a continuous growth factor to all values between $\hat{P}_{t+1}$ and P_t. We use the same midpoint technique to reduce the bias as we did in the Gompertz case: we split the steps into two parts, and to apply the growth factor $\exp\left(\hat{g}_t\right)$ to the first part, and growth factor $\exp\left(\hat{g}_{t+1}\right)$ to the second part. We do this by taking the mean of the two successive growth factors and applying that to P_t. Thus the one-step ahead and k-step ahead predictions are

(4.5) $$\hat{P}_{t+1} = P_t + P_t^2 \exp\left[0.5 \cdot \left(\hat{g}_{t+1} + g_t\right)\right] \quad \text{and} \quad \hat{P}_{t+k} = \hat{P}_{t+k-1} + \hat{P}_{t+k-1}^2 \exp\left[0.5 \cdot \left(\hat{g}_{t+k} + \hat{g}_{t+k-1}\right)\right].$$

The prediction variance for the logistic model is analogous to the prediction variance for the Gompertz model, the difference being that in the logistic model we have multipliers $\hat{P}_{t+i}^2$ and $\hat{P}_{t+j}^2$ entering the covariance term (3.16). Therefore the approximation for the covariances is

(4.6) $$\operatorname{cov}\left[\hat{P}_{t+i}^2 \exp\left(\hat{g}_{t+i}\right), \hat{P}_{t+j}^2 \exp\left(\hat{g}_{t+j}\right)\right] \approx \sigma_\varepsilon^2 \cdot \exp\left(2g_t\right) \min\left(i, j\right) \cdot \exp\left[\left(i + j\right)\delta\right] \hat{P}_{t+i}^2 \hat{P}_{t+j}^2$$

and the estimator for the variance of a k-step ahead prediction is

(4.7) $$V\left(\hat{P}_{t+k}\right) = \sigma_\varepsilon^2 \exp\left(2g_t\right) \sum_{i=1}^{k} \sum_{j=1}^{k} \min\left(i, j\right) \cdot \exp\left[\delta\left(i + j\right)\right] \cdot \hat{P}_{t+i}^2 \hat{P}_{t+j}^2.$$

Monte Carlo variance estimation for the logistic model is constructed the same way the Monte Carlo variance estimator is constructed in the Gompertz case. Appendix Table 1 summarizes the results of the Section 4: The Logistic diffusion model.

5 The Hernes diffusion model

As in the Gompertz case, let P_t be the proportion in a cohort that has by age t adopted the innovation, $P_0, P_1, ..., P_t$ the observed proportions and $P_{t+1}, P_{t+2}, ..., P_{t+k}$ the yet to be observed proportions we wish to predict. The Hernes diffusion model for a proportion P_t is

$$P_t = \frac{1}{1 + \frac{1-P_0}{P_0}\exp\left(\frac{a - ab^t}{\ln b}\right)}. \tag{5.1}$$

For a behavioral interpretation of the model, see Hernes (1972). The model is linearized as $\ln\left(\frac{dP_t}{dt}\frac{1}{P_t(1-P_t)}\right) = \ln a + bt$. To accommodate the model for discrete data, we use discretization $dP_t / dt \approx 0.5 \cdot \left(P_{t+1} - P_{t-1}\right)$ (see Appendix (8.1)). This gives us

$$\ln a + bt \approx \ln\left(\frac{P_{t+1} - P_{t-1}}{2}\frac{1}{P_t(1-P_t)}\right) \equiv g_t. \tag{5.2}$$

We model the underlying linear process g_t as a time series process. In the case of a random walk with drift, the model is given by (3.3) and the model parameters are estimated using (3.4).

5.2 Prediction and variance estimation

Li and Wu (2008) propose the equation

$$\hat{P}_{t+k} = \frac{1}{1 + \frac{1-P_t}{P_t}\exp\left[-\exp\left(\sum_{i=t+1}^{k}\hat{g}_{t+k}\right)\right]} \tag{5.3}$$

for predicting P_{t+k}. In our simulation experiments, however, (5.3) severely underestimated P_{t+k} for large k. Better predictions were obtained using any of the following three equations:

(5.4) $$\hat{P}_{t+k} = \frac{1}{1 + \frac{1-\hat{P}_{t+k-1}}{\hat{P}_{t+k-1}} \exp\left[-\exp\left(\hat{g}_{t+k}\right)\right]},$$

(5.5) $$\hat{P}_{t+k} = \hat{P}_{t+k-1} + \hat{P}_{t+k-1}\left(1-\hat{P}_{t+k-1}\right)\exp\left(\hat{g}_{t+k}\right),$$

(5.6) $$\exp\left(g_t\right)\hat{P}_{t+k}^2 + \left[1-\exp\left(g_t\right)\right]\hat{P}_{t+k} = \hat{P}_{t+k-1}.$$

The equation (5.4) is a simple modification of Li and Wu's equation (5.3), the difference being that (5.3) is not recursive, whereas (5.4) is. The equation (5.5) is obtained using the approximation

(5.7) $$\frac{P_{t+1}-P_{t-1}}{2}\frac{1}{P_t\left(1-P_t\right)} = \exp\left(g_t\right) \approx \left(P_t - P_{t-1}\right)\frac{1}{P_{t-1}\left(1-P_{t-1}\right)}$$

and solving P_t in terms of P_{t-1} and g_t. The third prediction equation (5.6) is quadratic and arises from the approximation

(5.8) $$\frac{P_{t+1}-P_{t-1}}{2}\frac{1}{P_t\left(1-P_t\right)} = \exp\left(g_t\right) \approx \left(P_t - P_{t-1}\right)\frac{1}{P_t\left(1-P_t\right)}.$$

Simulation experiments indicated that the prediction equations (5.4)-(5.6) produce almost identical results for large and small k, and estimate P_{t+k} markedly better than (5.3). Because of its simplicity and linearity in $\exp\left(g_t\right)$, we use equation (5.5). As in the Gompertz and logistic models, we use the midpoint method to correct the downward bias that arises from the fact that the growth factor $\exp\left(\hat{g}_{t+1}\right)$ is applied to P_t, instead of applying a continuous growth factor continuously to values between $\hat{P}_{t+1}$ and P_t by splitting the step into two parts and applying the growth factor $\exp\left(\hat{g}_t\right)$ to the first part, and growth factor $\exp\left(\hat{g}_{t+1}\right)$ to the second part. We do this by taking the mean of the two successive growth factors and applying that to P_t. Thus the k-step ahead prediction in the Hernes model is

$$\hat{P}_{t+k} = \hat{P}_{t+k-1} + \hat{P}_{t+k-1}\left(1-\hat{P}_{t+k-1}\right)\exp\left[0.5\cdot\left(\hat{g}_{t+k}+\hat{g}_{t+k-1}\right)\right]. \tag{5.9}$$

The prediction variance for the Hernes model is similar to the prediction variance for the Gompertz model. The difference is that we have multipliers $\hat{P}_{t+i}\left(1-\hat{P}_{t+i}\right)$ and $\hat{P}_{t+j}\left(1-\hat{P}_{t+j}\right)$ which enter the covariance term (3.16). Therefore the approximation for the covariances is

$$\begin{aligned}&\operatorname{cov}\left[\hat{P}_{t+i}\left(1-\hat{P}_{t+i}\right)\exp\left(\hat{g}_{t+i}\right),\hat{P}_{t+j}\left(1-\hat{P}_{t+j}\right)\exp\left(\hat{g}_{t+j}\right)\right]\\&\quad\approx\sigma_{\varepsilon}^{2}\cdot\exp\left(2g_{t}\right)\cdot\min\left(i,j\right)\cdot\exp\left[\left(i+j\right)\delta\right]\cdot\hat{P}_{t+i}\left(1-\hat{P}_{t+i}\right)\hat{P}_{t+j}\left(1-\hat{P}_{t+j}\right)\end{aligned} \tag{5.10}$$

and the estimator for the variance of a k-step ahead prediction is

$$V\left(\hat{P}_{t+k}\right)=\sigma_{\varepsilon}^{2}\exp\left(2g_{t}\right)\sum_{i=1}^{k}\sum_{j=1}^{k}\min\left(i,j\right)\cdot\exp\left[\delta\left(i+j\right)\right]\cdot\hat{P}_{t+i}\left(1-\hat{P}_{t+i}\right)\hat{P}_{t+j}\left(1-\hat{P}_{t+j}\right). \tag{5.11}$$

Monte Carlo variance estimation for the Hernes model is constructed the same way the Monte Carlo variance estimator is constructed in the Gompertz case. Appendix Table 1 summarizes the important results of the Section 5: Hernes diffusion model.

6 Simulation experiments and empirical applications

In this section we put the stochastic Gompertz, logistic and Hernes diffusion models described in Sections 3-5 into work. In Section 6.1 we conduct simulation experiments where the data generating process can be controlled and compare different methods for deriving predictions from the underlying linear process and the accuracy of the analytical variance estimator. In Sections 6.2 and 6.3 we apply the methods to predict marriage rates in France (Section 6.2) and first births in the Netherlands (Section 6.3).

6.1 Simulation experiments

We construct artificial data sets using the Gomperts, logistic, and Hernes model formulations. For each model, the values P_t are derived from an artificially generated g_t using the model equations shown on row 1 of Appendix Table 1. The underlying process g_t is for all models random walk with drift with normal, independent shocks with zero mean and variance σ_ε^2. For the Gompertz model, we use as the drift and variance parameters $\delta = -0.2, \sigma_\varepsilon^2 = 0.015^2$, for logistic model, they are $\delta = -0.2, \sigma_\varepsilon^2 = 0.025^2$, and for the Hernes model the parameters are $\delta = -0.2, \sigma_\varepsilon^2 = 0.030^2$. The starting value P_0 is 0.001 for all models. As the process g_t is a random walk, the shocks cumulate over time in g_t also in the proportion P_t.

For each of the three models, Gompertz, logistic, and Hernes, we generate data $P_0, P_1, ..., P_{35}$ using the process described above. This data is then "observed" up to ages 16 and 26. Using the observed data (up to age 16 or 26), we fit the correct models (Gompertz model for the Gompertz data, logistic model for the logistic data, and Hernes model for the Hernes data) and use the models to predict the values of P up to age 35. We also estimate the prediction variances and corresponding 95% confidence intervals using both the analytical variance estimator and the Monte Carlo based estimator. When using the Monte Carlo estimator, we calculate confidence intervals non-parametrically, using the percentiles of the prediction distribution rather than multiples of standard error as the basis for confidence interval.

We start by considering the prediction accuracy with and without the midpoint correction. Figure 1 shows one simulated path $P_0, P_1, ..., P_{35}$ for the Gompertz model and predictions with and without the midpoint correction when predictions start at age 16. The figure indicates that the predictions not using midpoint correction may be downward biased, whereas the midpoint corrected predictions may be approximately unbiased.

Figure 1. Simulated diffusion data using the Gompertz model and predictions with and without midpoint correction. Data used up to age 16.

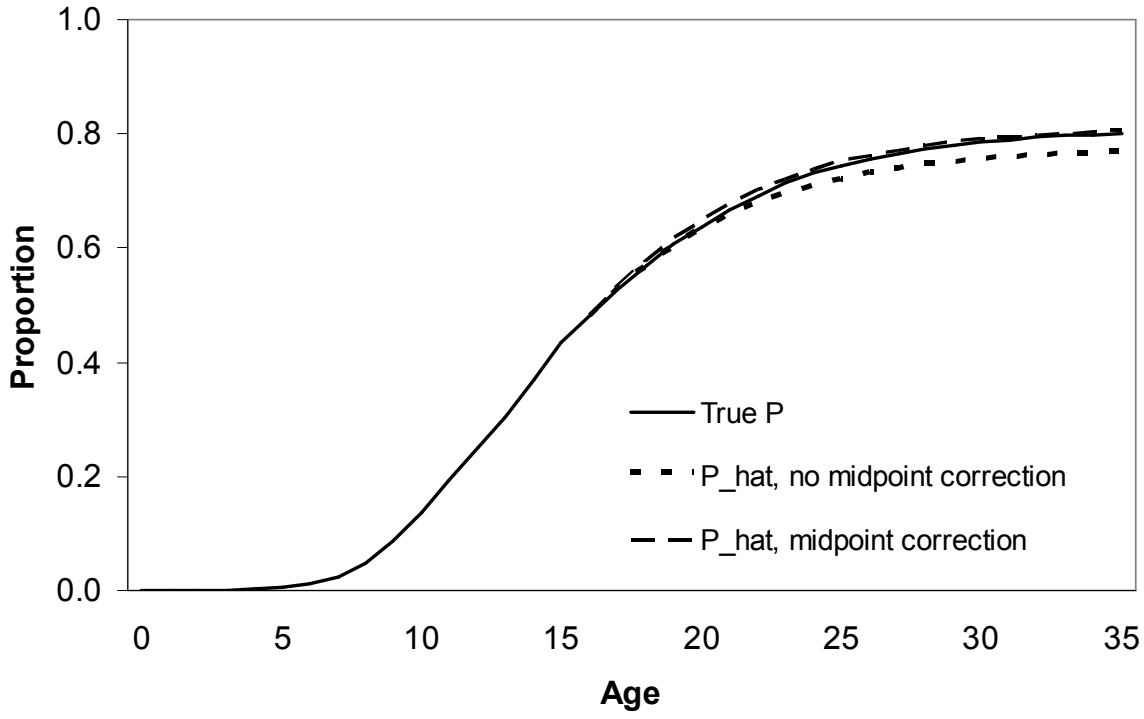


Table 1 shows the estimated bias for Gompertz, logistic and Hernes models from 1,000 simulated samples at ages 20, 25, 30 and 35. The data confirms what the Figure 1 suggested: The predictions not using the midpoint correction are downward biased, and the longer the prediction horizon, the larger the bias. This holds for the Gompertz, logistic, and Hernes models. The midpoint correction, however, significantly reduces the bias for all models, to less than one percentage in all cases.

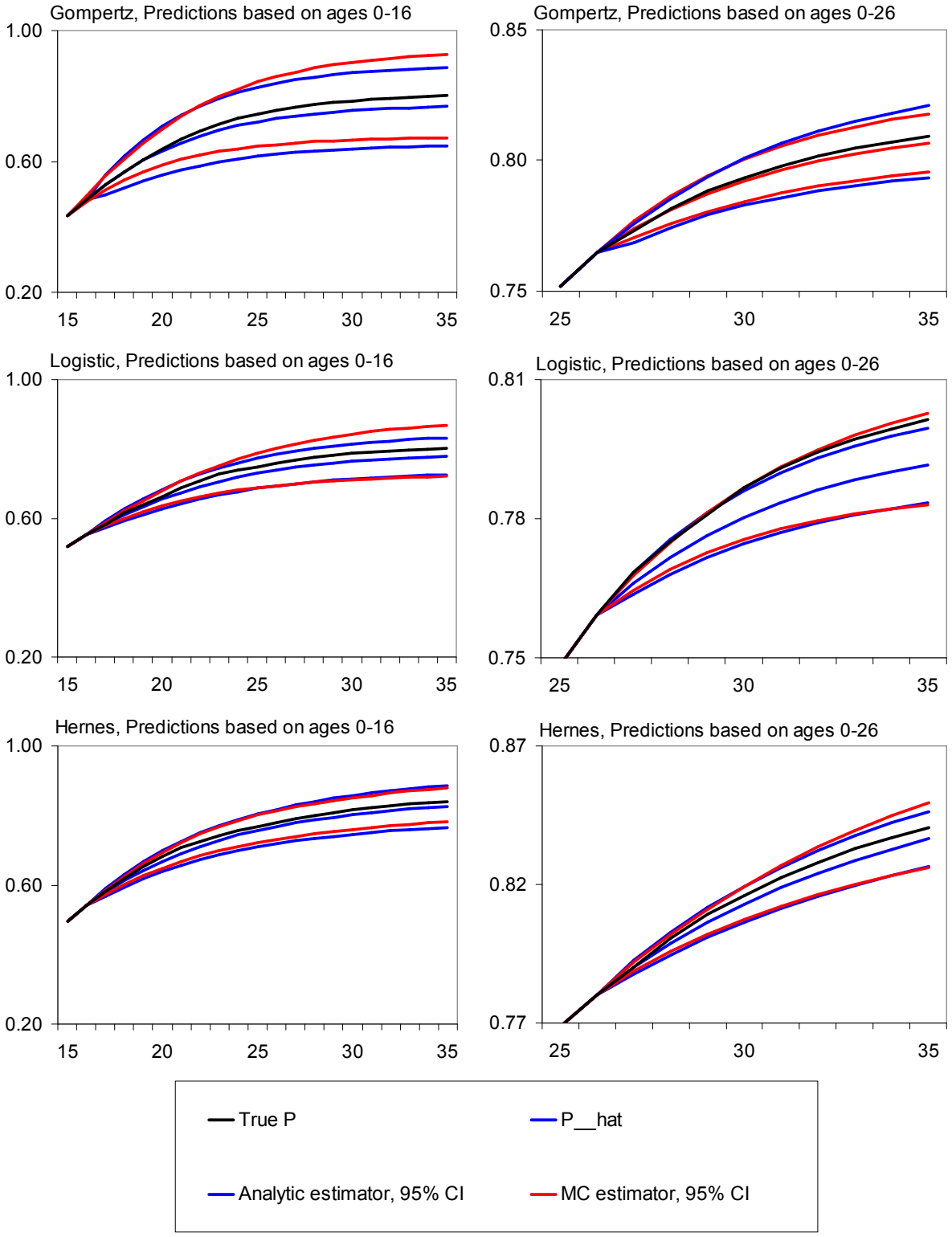
Table 1. Estimated relative bias* (%) for forecasts with and without midpoint correction for Gompertz, logistic and Hernes models at selected ages. Number of sample paths 1,000; data used up to age 16.

Model	Midpoint correction	Age 20	25	30	35
Gompertz	No	-0.8	-3.1	-3.8	-4.2
	Yes	0.9	0.9	0.7	0.5
Logistic	No	-1.6	-3.6	-3.1	-3.0
	Yes	0.5	0.0	0.3	0.8
Hernes	No	-1.9	-2.1	-2.1	-2.3
	Yes	-0.2	-0.3	-0.5	-0.5

* Relative bias calculated as the average of $(\hat{P} - P)/P$ over simulated samples.

Next we consider prediction variance estimation. In order to assess the accuracy of our delta-method approximations, we compared the confidence intervals obtained with the variability obtained by Monte Carlo simulation. Figure 2 shows comparisons of the analytical and Monte Carlo confidence intervals for the Gompertz, logistic, and Hernes models for cases where the predictions start at age 16 and at age 26. For all models the analytical and Monte Carlo estimator produce fairly similar confidence intervals. The Monte Carlo estimator should be accurate since it uses the same model as the data generating process. Thus the fact that the confidence intervals for the analytical variance estimator closely track the Monte Carlo estimator indicates that the analytical, delta method and Taylor series approximation based variance estimator works reasonably well.

Figure 2. Comparison of the analytical and Monte Carlo variance estimators. Simulated data; Gompertz, logistic and Hernes models; predictions with midpoint correction and confidence interval estimates use data up to age 16 (left) and age 26 (right hand side).

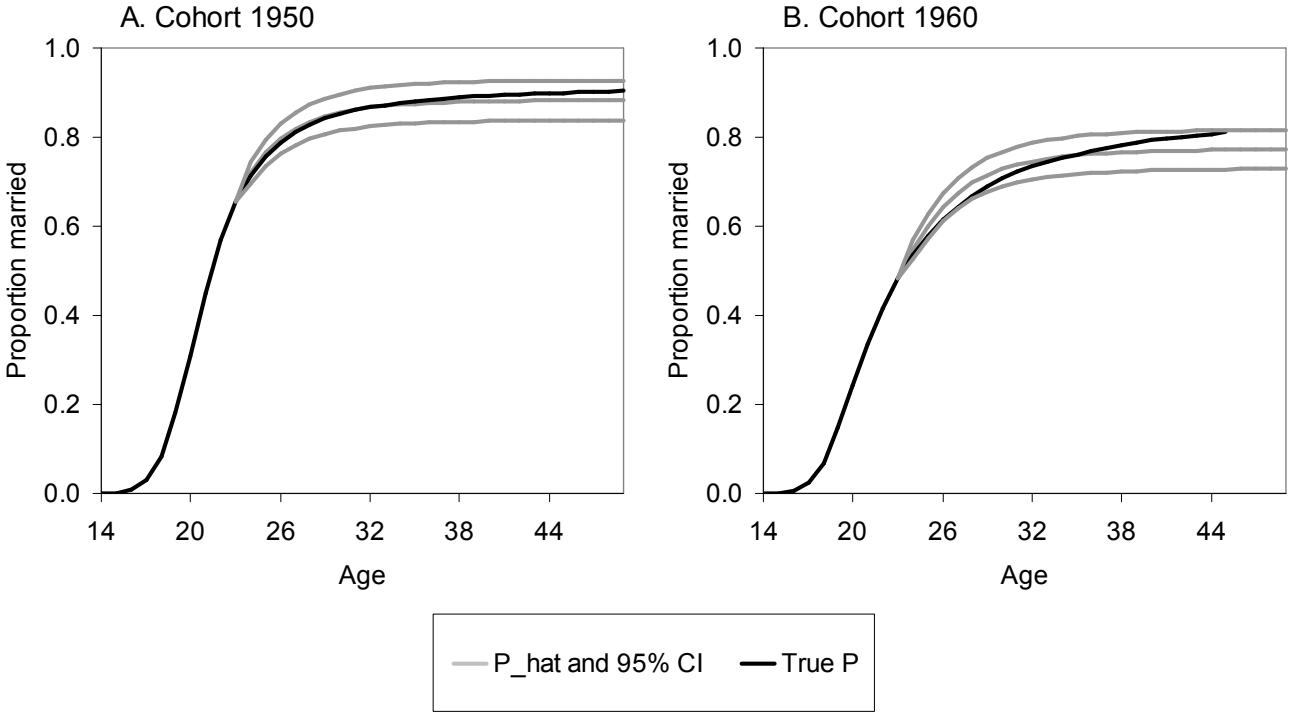


6.2 Application I: French first marriages and the Hernes model

In prior research the Hernes model has been used to predict proportion married within a cohort (Goldstein and Kenney 2001; Li and Wu 2008). Goldstein and Kenney (2001), however, do not provide any bounds of uncertainty for their predictions, and Li and Wu (2008) use a prediction method that produces severely biased estimates. Here we use the Hernes model discussed in Section 5 to predict the proportion married using French data, and use the estimated prediction intervals to assess the likelihood that younger cohorts would catch up to the older cohort's marriage rates. We start by fitting the Hernes model to the 1950 and 1960 cohorts. For both cohorts, we estimate the parameters of the underlying random walk with drift model using data up to age 23 (starting from age 14), and then predict the marriage rates up to age 50.

Results for the 1950 cohort are shown in Figure 3, Panel A. Results for the 1960 cohort are shown in Figure 3, Panel B. Panel A shows that the Hernes model produces reasonable predictions for the future experience for cohort 1950 when data is observed only up to age 23. The maximum prediction error (at age 50) is only 2.2 percentage points. The difference between the predictions and observed data emerge quite late, after age 33.

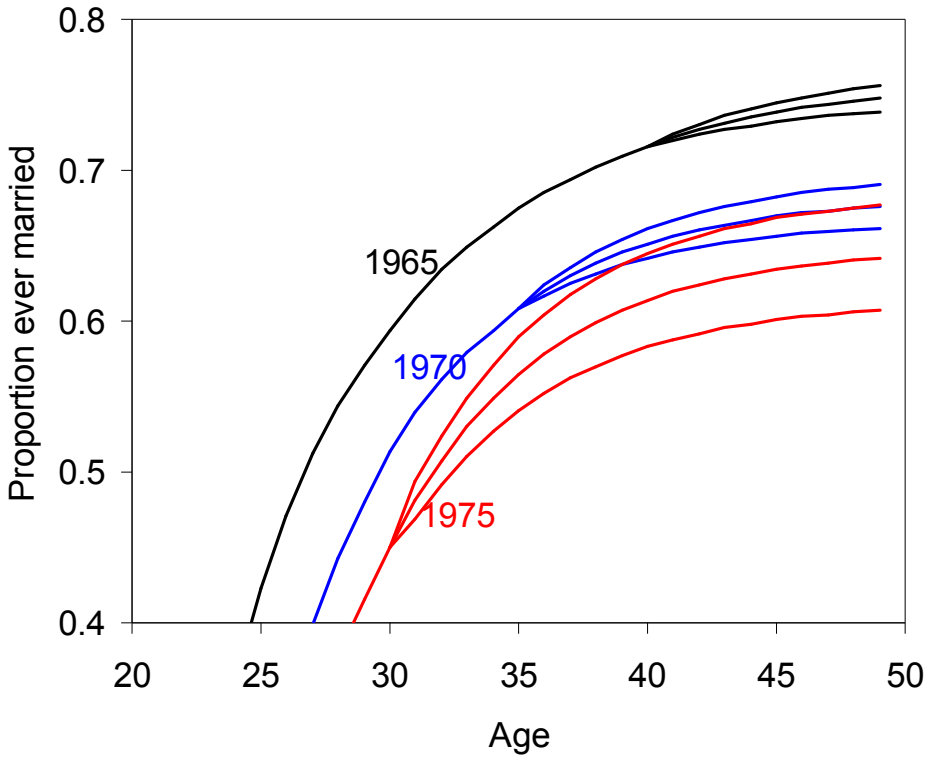
Figure 3. Proportion having married by age; French female cohorts 1950 and 1960. Predictions and 95% confidence interval are based on the Hernes model with underlying random walk with drift model. The predictions use the midpoint correction.



Panel B of Figure 3 shows the results for the 1960 cohort. Again, we have used data up to age 23 when estimating the random walk with drift model, and have then used this estimated model to derive predictions and prediction errors. The Hernes model predicts reasonably well the cohort's experience up to age 45, which is the oldest age for which data was available at the time of modeling. The observed data may, however, be reaching outside the 95% confidence interval, potentially implying that at these ages the reality may not be exactly Hernesian.

Figure 4. Proportion having married by age; French female cohorts 1965, 1970 and 1975. Predictions and 95 % confidence interval are based on the Hernes model with underlying random walk with drift model. The predictions use the midpoint correction.

BLACK = 1965 COHORT; BLUE = 1970 COHORT; RED = 1975 COHORT



Next we compare the cohorts born in 1965, 1970 and 1975, and analyze the likelihood that the younger cohorts' proportion ever married would catch up with the older cohorts' proportion ever married. We do this by constructing for each cohort predictions and 95% confidence intervals (using the analytical estimator) for proportion ever married by age. Figure 4 shows the predictions. The lower bound of the predictions for the 1965 cohort is higher than the upper

bound of the predictions for the 1970 and 1975 cohorts. Thus it is extremely unlikely that the 1970 or 1975 cohorts would catch up with the 1965 cohort. The prediction interval for the 1975 cohort, however, overlaps with the prediction interval of the 1970 cohort, suggesting that the 1975 cohort's proportion ever married might catch up with 1970 cohort.

It is important to note, however, that in the predictions shown in Figure 4 the shocks in the underlying random walk with drift model which ultimately give rise to the uncertainty in the predictions are assumed to be independent. This is may not be an accurate description of reality: period fluctuations which influence marriage rates (or the underlying random walk process) may do so for all cohorts. Therefore we have also used the Monte Carlo method to construct predictions for the 1970 and 1975 cohorts using the same shocks in the random walk processes that give raise to the uncertainty in the predictions (not shown). As the shocks are the same, the correlation between the shocks in for the 1970 cohort and 1975 cohort is one. When the probability of catching up is evaluated assuming this extreme correlation in the shocks, none of the 1,000 simulated paths for the 1970 and 1975 cohorts resulted in overlap in the proportion ever married, suggesting that also for the 1975 cohort, catching up with the 1970 cohort is unlikely. However, the assumption that the shocks are perfectly correlated may be too strong; thus in future research, we will use historical data to estimate the correlations across the cohorts' underlying linear processes and use the estimated correlation in the Monte Carlo simulations in order to get a more accurate view of the likelihood of the younger cohorts catching up to the older cohorts' rates.

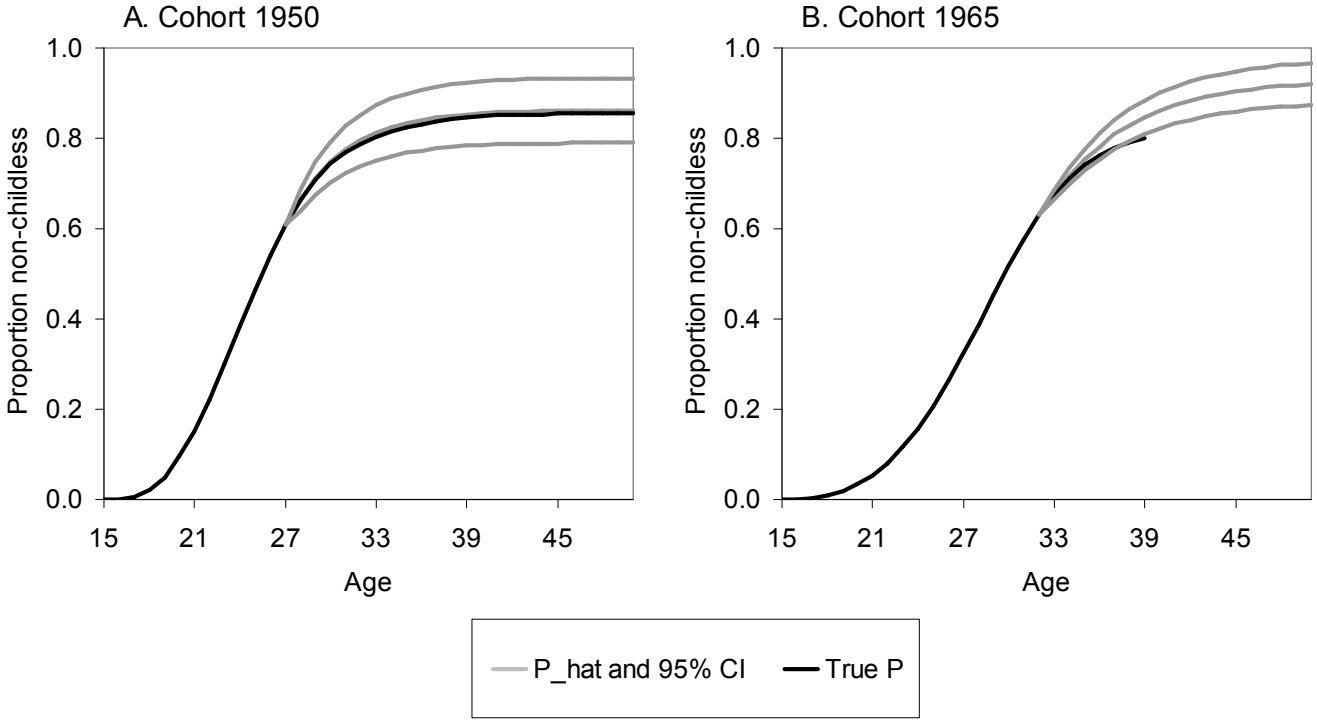
6.3 Application II: Dutch first births and the Gompertz model

Kohler (Kohler 2001) and Bernardi (Bernardi 2003) have shown that social interaction is a key variable influencing fertility decision. Consistent with the social interaction theories, Goldstein's recent results (Goldstein 2008) indicate that the Gompertz model may work well in predicting first birth and childlessness if applied to cohort data, but at older ages and especially for the later cohorts there may be departures from the model. Without confidence intervals, however, it is difficult to assess what is a departure from the model and what is within-model fluctuation. Here we fit the Gompertz model to Dutch data, and predict, with confidence intervals, the proportion not childless for 1950 and 1965 cohorts. Experiments with the Gompertz model (not shown) suggested that the proportion should be close to 2/3 before reasonable fit can be expected.

Therefore we use data up to age 28 for the 1950 cohort (by this age 66 % of the cohort had had a first birth) and for the 1965 cohort we use data up to age 34 (by this age 67 % of the cohort had had a first birth.

Results for the Gompertz model for the cohort 1950 are shown in Figue 5, Panel A. For the 1950 cohort, Gompertz model produces very accurate predictions (maximum error in the predictions is 1.1 percentage points). Panel B of Figure 5 shows the results for the 1965 cohort. The figure shows that almost immediately after we start predicting the data, the observations tend outside the 95% confidence interval. If the model holds, one should expect to see the true data be outside the 95 % confidence interval on average every twentieth time, and this may be what is happening in Panel B of Figure 5. A potentially more likely explanation is that the cohort 1965 has postponed their childbearing so late that the behavioral assumptions on which the Gompertz model is built are not anymore the only driving forces behind P_t. At ages above 30 biology inevitably starts to enter the equation, and fecundity starts to decline; this may be the factor explaining the low first birth proportion compared to the forecasts. This potential explanation is discussed in more detail in Goldstein (Goldstein 2008).

Figure 5. Proportion non-childless by age; Dutch female cohorts 1950 and 1965. Predictions and 95% confidence interval are based on the Gompertz model with underlying random walk with drift model. The predictions use the midpoint correction.



7 Discussion

In this paper we studied prediction and error propagation in the Gompertz, logistic, and Hernes cohort diffusion models. We showed that for all these models predictions can be derived from an underlying linear process. We compared different methods for deriving the predictions and found that the midpoint correction, which has not been used in cohort diffusion models before, improves the accuracy significantly with respect to previously used methods. We also derived an analytical variance estimator for the predictions. This closed form estimator reveals important facts about the sources of uncertainty in cohort diffusion models, most importantly that the earlier the predictions are made and the slower the diffusion, the larger the uncertainty in the predictions.

Simulation studies and empirical applications to first births and marriages showed that the developed methods are useful in quantifying uncertainty in the predictions: They give a precise sense of the within-model error, and allow the forecasters a new ability to characterize the uncertainty. When the model assumptions hold less than perfectly, as in the case of first births for the Dutch 1965 cohort whose fertility may be constrained by extra-model factors such as biology (Goldstein 2008), the constructed confidence intervals give a lower bound for the total uncertainty.

In summary, this paper developed methods for predicting the diffusion of an innovation within cohorts. The new methods improve accuracy in point forecasts and allow the researcher to quantify the uncertainty in the predictions.

The developed methods give raise to several future research questions. First, we will explore the potential of the cohort diffusion models for predicting period fertility rates by combining a large number of adjoining cohorts. The Lee model for fertility (Lee 1993), which can be considered the gold standard for stochastic fertility forecasting, has the potential drawback that it forces the age-shape of fertility to be constant across time. Given recent developments in fertility, especially the postponement of having children (Sobotka 2004), the constant shape of age-specific fertility rates seem unrealistic. Period fertility forecasts based on cohort fertility patterns would automatically have realistically changing age-patterns of childbearing.

Second, we will expand the range of fitted populations, incorporating cohort fertility and marriage rates from the United States and other European countries, including Eastern and Mediterranean Europe in order to study how generally applicable the methods are. It is especially interesting to see where the methods do not work – for example in the case of postponement of childbearing, the tendency of the model to overpredict fertility at oldest ages for the youngest cohorts is likely to be an indication of sterility, a phenomenon the model is not built to capture. Departures from the model may provide means of indirectly estimating the magnitude of lost fertility due to sterility.

Third, we will study the correlation in the underlying time series processes across cohorts in order to gain knowledge on first, how the cohorts' marital and childbearing decisions are correlated, and second, in order to be able to accurately estimate the likelihood of the younger cohorts catching up to the older cohorts' rates.

Finally, we will look at the variability in the drift parameters over time and place in order to provide a richer description of past marriage and fertility changes and to inform forecasts of future developments.

Appendix. Often used equations and summary of the results

Some identities, approximations and discretizations which are used often:

(8.1) Discretization 1:
$$\frac{dP_t}{dt} \approx \frac{P_{t+1} - P_{t-1}}{2}$$

(8.2) Discretization 2:
$$\frac{d \ln P_t}{dt} \approx \frac{1}{P_t} \frac{P_{t+1} - P_{t-1}}{2}$$

(8.3) Approximating change:
$$\frac{P_{t+1} - P_{t-1}}{2} \approx P_t - P_{t-1}$$

(8.4) The delta method:
$$V\left[H\left(X\right)\right] \approx V\left(X\right)\left[\frac{dH\left(\mu_X\right)}{dX}\right]^2$$

(8.5) Variance of a sum:
$$V\left(\sum_{i=1}^{k} X_i\right) = \sum_{i=1}^{k}\sum_{j=1}^{k} \operatorname{cov}\left(X_i, X_j\right)$$
$$= \sum_{i=1}^{k} V\left(X_i\right) + 2\sum_{i=1}^{k}\sum_{j\neq i}^{k} \operatorname{cov}\left(X_i, X_j\right)$$

(8.6) First order Taylor series approximation:

$$V\left(\sum_{i=1}^{k} f_i\left(X_i\right)\right) \approx \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{df_i\left(\mu_{X_i}\right)}{dX_i} \frac{df_i\left(\mu_{X_j}\right)}{dX_j} \operatorname{cov}\left(X_i, X_j\right)$$

Appendix Table 1. Summary of the Gompertz, logistic and Hernes models with a random walk with drift as the underlying linear process.

	Gompertz	**Logistic**	**Hernes**
1. Model equation	$P_t = k\exp\left[-\exp\left(a-bt\right)\right]$	$P_t = \dfrac{a}{1+\exp\left(a-bt\right)}$	$P_t = \dfrac{1}{1+\dfrac{1-c}{c}\exp\left(\dfrac{a-ab^t}{\ln b}\right)}$
2. Linearization (g)	$\ln\left(\dfrac{d\ln P_t}{dt}\right) = \ln b + a - bt = g_t$	$\ln\left(\dfrac{dP_t}{dt}\dfrac{1}{P_t^2}\right) = \ln b + a - bt = g_t$	$\ln\left(\dfrac{dP_t}{dt}\dfrac{1}{P_t\left(1-P_t\right)}\right) = \ln a + bt = g_t$
3. Model for g; estimators for δ and σ_ε^2; predictions $\hat{g}_{t+k}$	Model: $g_t = g_0 + \delta t + \sum_{i=1}^{t}\varepsilon_i$ Estimators: $\hat{\delta} = \dfrac{g_{t-1}-g_1}{t-2}$; $\hat{\sigma}_\varepsilon^2 = \dfrac{\sum_{i=1}^{t-1}\left(g_i - g_{i-1} - \hat{\delta}\right)^2}{t-3}$ Predictions: $\hat{g}_{t+k} = g_t + \hat{\delta}k$		
4. Predictions $\hat{P}_{t+k}$	$\hat{P}_{t+k} = \dfrac{\hat{P}_{t+k-1}}{1-\exp\left[0.5\cdot\left(\hat{g}_{t+k}+\hat{g}_{t+k-1}\right)\right]}$	$\hat{P}_{t+k-1} + \hat{P}_{t+k-1}^2\exp\left[0.5\cdot\left(\hat{g}_{t+k}+\hat{g}_{t+k-1}\right)\right]$	$\hat{P}_{t+k-1} + \hat{P}_{t+k-1}\left(1-\hat{P}_{t+k-1}\right)\exp\left[0.5\cdot\left(\hat{g}_{t+k}+\hat{g}_{t+k-1}\right)\right]$
5. Variance $V\left(\hat{P}_{t+k}\right)$	$\sigma_\varepsilon^2\exp\left(2g_t\right)\cdot$ $\sum_{i,j=1}^{k}\min\left(i,j\right)\exp\left[\delta\left(i+j\right)\right]$	$\sigma_\varepsilon^2\exp\left(2g_t\right)\cdot$ $\sum_{i,j=1}^{k}\min\left(i,j\right)\exp\left[\delta\left(i+j\right)\right]\hat{P}_{t+i}^2\hat{P}_{t+j}^2$	$\sigma_\varepsilon^2\exp\left(2g_t\right)\cdot$ $\sum_{i,j=1}^{k}\min\left(i,j\right)\cdot\exp\left[\delta\left(i+j\right)\right]\cdot\hat{P}_{t+i}\left(1-\hat{P}_{t+i}\right)\hat{P}_{t+j}\left(1-\hat{P}_{t+j}\right)$

References

Bernardi, L. (2003). "Channels of Social Influence on Reproduction." Population Research and Policy Review **22**(5-6): 527-555.

Billari, F. C. and L. Toulemon (2006). Cohort Childlessness Forecasts and Analysis Using the Hernes Model. European Population Conference 2006. Liverpool, UK.

Frances, P. H. (1994). "A Method to Select Between Gompertz and Logistic Trend Curves." Technological Forecasting and Social Change **46**: 45-49.

Goldstein, J. R. (2008). A Behavioral Gompertz Model for Cohort Fertility Schedules in Low and Moderate Fertility Populations. Population Association of America Annual Conference. New Orleand, LA, USA.

Goldstein, J. R. and C. T. Kenney (2001). "Marriage Delayed or Marriage Forgone? New Cohort Forecasts of First Marriage for U.S. Women." American Sociological Review **66**(4): 506-519.

Griffiths, D. V. and I. M. Smith (1991). Numerical methods for engineers: a programming approach. . Boca Raton, CRC Press.

Gruber, H. and F. Verboven (2001). "The diffusion of mobile telecommunications services in the European Union." European Economic Review **45**(3): 577-588.

Hamilton, J. D. (1994). Time Series Analysis. Princeton Princeton University Press.

Harvey, A. C. (1984). "Time series forecasting based on the logistic curve." Journal of the Operational Research Society **35**: 641-646.

Hernes, G. (1972). "The Process of Entry into First Marriage." American Sociological Review **37**(2): 173-182.

Hoem, J. M., D. Madsen, et al. (1981). "Experiments in Modelling Recent Danish Fertility Curves." Demography **18**(2): 231-244.

Ike, S. (2002). "A non-stationary stochastic process model of completed marital fertility in Japan." The Journal of Mathematical Sociology **26**(1-2): 35-55.

Kohler, H.-P. (2001). Fertility and Social Interaction: An Economic Perspective, Oxford University Press.

Lee, R. D. (1993). "Modeling and Forecasting the Time Series of U.S. Fertility: Age Patterns, Range, and Ultimate Level." International Journal of Forecasting **9**: 187-202.

Li, N. and Z. Wu (2008). Modeling and Forecasting First Marriage: A Latent Function Approach. Population Association of America Annual Conference. New Orleans, LA, USA.

Mansfield, E. (1963). "The speed of response of firms to new techniques." Quarterly Journal of Economics **77**: 290–311.

Mar-Molinero, C. (1980). "Tractors in Spain: A logistic analysis." Journal of the Operational Research Society **31**: 141-152.

Martin, S. P. (2004). Reassessing delayed and foregone marriage in the United States. Russel Sage Working Paper.

Meade, N. and T. Islam (2006). "Modelling and forecasting the diffusion of innovation - A 25-year review." International Journal of Forecasting **22**(3): 519-545.

Pollard, J. and E. Valkovics (1992). "The Gompertz distribution and its applications." Genus **48**(3-4): 15-28.

Sobotka, T. (2004). Postponement of Childbearing and Low Fertility in Europe. Amsterdam, Dutch University Press.

Winsor, C. P. (1932). "The Gompertz Curve as a Growth Curve." Proceedings of the National Academy of Sciences of the United States of America **18**(1): 1-8.