Prediction and Error Propagation in Cohort Diffusion Models

Mikko Myrskylä[1]

Joshua R. Goldstein[2]

September 15, 2008

Preliminary draft – do not cite without permission

Abstract

We study prediction and error propagation in the Gompertz, logistic, and Hernes cohort diffusion models. We show that the linearized forms of these models can be modeled as a random walk with drift, and that predictions and prediction error estimates can be derived from the random walk model. We develop and compare different methods for deriving predictions from the underlying random walk model. We also develop an analytic variance estimator for the prediction variance and study its accuracy with respect to a Monte Carlo estimator. Simulation studies and empirical applications to first births and marriages show that the analytic estimator is accurate, allowing forecasters to make precise the level of "within-model" uncertainty that should be attached to their forecasts, a level that should be viewed as a lower-bound of the total uncertainty which could include departure from the model.

TABLE OF CONTENTS

[1] Population Studies Center, 239 McNeil Building, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104-6298, USA. Email: myrskylm@pop.upenn.edu, phone: +1 267 235 7257, fax: +1 215 898-2124.

[2] Director; Head of the Laboratory of Economic and Social Demography, Max Planck Institute for Demographic Research. 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 tools in forecasting incomplete cohort experience. For example, Goldstein and Kenney (Goldstein and Kenney 2001) and Li and Wu (Li and Wu 2008) show how the Hernes model (Hernes 1972) can be used to forecast marriage rates. For fertility, it has long looked like the Gompertz model would be inadequate for predicting (Hoem, Madsen et al. 1981; Pollard and Valkovics 1992), but recent research by Goldstein (Goldstein 2008) suggests that if fit to the cohort rates (instead of fitting the model to period rates), the Gompertz model actually performs quite well. A third well known diffusion model, the logistic model, can also be used to forecast cohort experience. This far, however, the logistic model has received more attention in the economic literature, as it has been used to forecast the use of tractors (Mar-Molinero 1980), mobile telecommunications services (Gruber and Verboven 2001), and more generally sales and innovations (Harvey 1984; Meade and Islam 2006).

Despite the longstanding interest in diffusion models, two fundamental questions about their use in the context of cohort prediction remain unanswered. First, how should the models be estimated from incomplete data? Second, how to estimate uncertainty in the predictions? These questions have been partly answered, but no synthesis has been built. In this paper we study prediction and error propagation in the Gompertz, logistic, and Hernes cohort diffusion models, and show that prediction and error propagation can be treated in a unified way: We show that the linearized forms of these models can be modeled as a random walk with drift, and that predictions and prediction error estimates can be derived from the random walk model. We develop and compare different methods for deriving predictions from the underlying time series model. We also develop and compare the accuracy of a closed form analytic estimators and Monte Carlo estimators for the prediction variance. Empirical applications to first births and marriages suggest that the random walk based cohort diffusion models can be highly useful in predicting the future experience of a cohort.

The paper is organized as follows. In Section 2, we describe the steps needed in the time series approach in a heuristic way. In Sections 3-5, we show in a detailed manner how prediction and prediction error estimation is done in the Gompertz, Hernes and Logistic cohort diffusion models when using the time series approach. In Section 6 we apply the models to both simulated and real data. Section 7 discusses the results. The Appendix shows certain formulas which are used throughout the paper.

2 Overview of the time series approach

The idea of linearizing a growth or diffusion model and fitting a time series or some other regression model to the linearized part is, in itself, not new. For example, Winsor (Winsor 1932) shows how the logistic and Gomperts models can be linearized with respect to time, and Harvey (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. Harvey, however, does not discuss the model in cohort context, and leaves also prediction intervals unestimated. Li and Wu (Li and Wu 2008) use the Hernes model in the cohort diffusion context, and follow Winsor and Harvey by linearizing the model and fitting a regression model (without autocorrelation structure) to the linearized part. In this paper, possible autocorrelation in the linear part is not taken into account, and the prediction intervals ignore the cumulating nature of errors.

We build on prior research by providing a unified framework for time series based prediction and prediction error estimation in cohort diffusion models. First, let P_t denote the proportion "infected" – that is those who, depending on the application, have married, have had a first birth, or more generally have adopted the innovation. We assume that P_t depends on time t through a monotonic increasing function F : $P_t = F(t)$. Now 7 steps are needed to produce a time series modeling based prediction and prediction intervals of P for a future time $t+1$, given observations up to t. These are as follows:

1. Find a linearization H so that $H(P_t) = g_t$, where g is linear in t. This often involves taking derivatives and logarithms. One also has to deal with the issue that H is often easiest to find using continuous time, but the observations are by necessity in discrete time.

2. Model the linearized part g_t with an ARIMA model, such as random walk with drift which is the same as ARIMA(0,1,0)

3. Estimate the parameters of the ARIMA model using standard techniques (e.g. Hamilton 1994)

Repeat steps 4-5 for $i = 1,...,k$:

4. Predict $\hat{g}_{t+i}$ using the estimated ARIMA model

5. Derive $\hat{P}_{t+i}$ from $\hat{g}_{t+i}$ using the inverse of H. This, often, is less straightforward than it seems, because H may be a functional, rather than a function and because H may be defined using

continuous time but the observations are in discrete time. The prediction $\hat{P}_{t+i}$ invariably involves the past value $\hat{P}_{t+i-1}$, therefore the prediction up to $\hat{P}_{t+k}$ needs to proceed recursively.

6. Estimate the variance of $\hat{P}_{t+k}$. If the ARIMA model for g_t involves differencing, one has to deal with fact that the shocks in g_t do not fade away but cumulate to P_{t+k}.

The Sections 3-5 show how this approach can be used 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, Table 1 summarizes the model equations, linearizations, models for the linear part, prediction equations and analytical prediction variance estimators.

3 The Gompertz growth model

3.1 The model

Throughout the paper we assume observations $P_0, P_1, ..., P_t$ and we make predictions $P_{t+1}, P_{t+2}, ..., P_{t+k}$. The Gompertz growth model for a proportion P_t is

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

Log of the log-derivative linearizes the model to $\ln b + a - bt$. We use the discretization (8.2), $\frac{d\ln P_t}{dt} \approx \frac{1}{P_t}\frac{P_{t+1}-P_{t-1}}{2}$, proposed by Li and Wu (Li and Wu 2008). With this linearization we get

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

We model the linear term g_t as a random walk with drift:

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

The model parameters $\left(\delta, \sigma_\varepsilon^2\right)$ are estimated as[3]

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

3.2 Prediction

Predictions $\hat{P}_{t+1}$ and $\hat{P}_{t+j}$ are based on predictions $\hat{g}_{t+1} = g_t + \hat{\delta}$ and $\hat{g}_{t+k} = g_t + \hat{\delta}k$. We use the approximation (8.3), $0.5\cdot\left(P_{t+1} - P_{t-1}\right) \approx P_t - P_{t-1}$, to transform g into P. This is done as follows. First note that $\exp\left(g_t\right)$ describes proportional change, which can be approximated as

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

Now P_t can be expressed in terms of P_{t-1} and g_t, and $\hat{P}_{t+1}$ in terms of P_t and $\hat{g}_{t+1}$:

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

(3.6) $$P_t \approx \frac{P_{t-1}}{1-\exp\left(g_t\right)} \quad \text{and} \quad \hat{P}_{t+1} = \frac{P_t}{1-\exp\left(\hat{g}_{t+1}\right)}.$$

Equation (3.6) gives the one-step ahead predictions, and applying (3.6) recursively one gets the arbitrary k-step ahead prediction. This method, however, will underestimate the true P_{t+k}. This is because we are using discrete data, and 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. The downward bias can be reduced by splitting the steps into two parts, and apply the growth factor $\exp\left(\hat{g}_t\right)$ to to the first part, and growth factor $\exp\left(\hat{g}_{t+1}\right)$ to the second part. This can be done in two steps, or one can also simply take the average of $\exp\left(\hat{g}_{t+1}\right)$ and $\exp\left(\hat{g}_t\right)$ and apply that to P_t.[4] This holds for the first step, also for the further steps that are needed to produce the k-step ahead prediction $\hat{P}_{t+k}$. Thus we have the one-step ahead and k-step ahead predictions as follows:

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

3.3 Prediction variance

Here 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 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

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

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). This is done as follows:

[4] 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 so small that one does not need to worry about it.

3.3.1.1 One-step ahead prediction variance

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

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

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

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

Now

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

and

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

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

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

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

3.3.1.2 k-step ahead prediction variance

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:

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

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

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

Simulation experiments indicated that the off-diagonal elements $\operatorname{cov}\left[\exp\left(\hat{g}_{t+i}\right),\exp\left(\hat{g}_{t+j}\right)\right]$ are also important. The reason for this is the double-counting of the errors: innovations up to i are both in g_{t+i} and g_{t+j}, provided $j \geq i$. These off-diagonal elements can be approximated using the Taylor method as

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

The interpretation for (3.16) is the following. There are $\min\left(i,j\right)$ common innovations 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 (3.15) and in the off-diagonal terms (3.16) essentially scale the covariance proportionally to the size of the terms $\exp\left(\hat{g}_{t+j}\right)$. 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):

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

Taylor approximation (8.6) applied directly to (3.14) would deliver the same 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 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 late (so t is large and g_{t} negative and large), the prediction variance is small. If the predictions are made early, 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) and growth takes place soon, the prediction variance is small. If, however, growth is slow and the drift is closer to 0, the prediction variance is large. These remarks apply also to the logistic and Hernes models.

3.3.2 Monte Carlo variance estimator

In the Monte Carlo variance estimation, we simulate $K = 1,000$ sample paths $g_{t+1}, g_{t+2}, ..., g_{t+k}$ using the formula

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

These simulated paths of g are transformed to P using the prediction equation (3.7). The variance is then directly calculable from the 1,000 predictions, as are the non-parameteric confidence intervals. The Monte Carlo point estimate for P_{t+k} is the median of the predictions.

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

4 The Logistic growth model

4.1 The model

The logistic growth model for a proportion P_t is

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

The model is linearized as $\ln\left(\frac{dP_t}{dt}\frac{1}{P_t^2}\right) = \ln b + a - bt$. We use discretization (8.1), $\frac{dP_t}{dt} \approx \frac{P_{t+1}-P_{t-1}}{2}$, so

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

As in (3.3), the g_t is modeled as a random walk with drift and $\left(\delta, \sigma_\varepsilon^2\right)$ is estimated with (3.4).

4.2 Prediction and variance estimation

Using the approximation

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

we get equations for the predictions (Harvey (1984) presents similar results):

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

These, however, underestimate the true P_{t+k} because of the same reasons analogous equations underestimated P_{t+k} in the Gompertz case: 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. We use the same technique to reduce the bias as we did in the Gompertz case, that is to split the steps into two parts, and apply the growth factor $\exp\left(\hat{g}_t\right)$ to 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. The same holds for predictions that go further. 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 the same as it is for the Gompertz model.

Table 1 summarizes the important results of the Section 4: The Logistic growth model.

5 The Hernes growth model

The Hernes growth model for a proportion P_t is

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

.The model is linearized as $\ln\left(\frac{dP_t}{dt}\frac{1}{P_t\left(1-P_t\right)}\right) = \ln a + bt$. We use discretization (8.1), so

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

The g_t is modeled as a random walk with drift as in (3.3), and $\left(\delta,\sigma_\varepsilon^2\right)$ is estimated with (3.4).

5.2 Prediction and variance estimation

Li and Wu (2008) propose the equation

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

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 recursively 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 in (5.3), one jumps to the prediction $\hat{P}_{t+k}$ from observation P_t, whereas in (5.4) one proceeds recursively using predictions $\hat{P}_{t+1},...,\hat{P}_{t+k-1}$. 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_r in terms of P_{t-1} and g_t. The quadratic equation (5.6) arises from the approximation

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

Simulation experiments indicated that the prediction equations (5.4)-(5.6) produce almost identical results, even for large 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 the basis for predictions. We, however, correct the downward bias in (5.5) 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 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. The same holds for predictions that go further. Thus the k-step ahead prediction is

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

The prediction variance for the Hernes model with predictions $\hat{P}_{t+k}=\hat{P}_{t+k-1}+\hat{P}_{t+k-1}\left(1-\hat{P}_{t+k-1}\right)\exp\left(\hat{g}_{t+k}\right)$ 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

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

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

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

Monte Carlo variance estimation for the Hernes model is the same as it is for the Gompertz model.

Table 1 summarizes the important results of the Section 5: The Hernes growth model.

6 Simulation experiments and empirical applications

In this section we put the Gompertz, logistic and Hernes models described in Sections 3-5 into work. We start with simulation experiments where the data generating process can be controlled (Section 6.1), and then apply the methods to real data 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 models formulations. The values P_t are generated from g_t using the model equations (line 1_ of Table 1. The process g_t is subject to shocks: g_t is a random walk with drift δ and shock variance σ_ε^2. Thus the shocks ε_t affect both the linearized part g_t and the proportion P_t, and as the model for g_t is a random walk with drift, these shocks cumulate over time.

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, 21, and 26. Using the observed data (up to age 16, 21, or 26), we fit the right models (Gompertz model for the Gompertz data, logistic model for the logistic data, and Hernes model for the Hernes data) and predict the values up to age 35. We also estimate the prediction variances, confidence intervals, and coefficients of variation (defined as standard error divided by estimate) 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.

Let us first consider the Gompertz case. Figure 1 shows the predicted values, confidence intervals and summary statistics for the variance estimators when data is observed up to age 16 (Panel A), age 21 (Panel B), and age 26 (Panel C). The left hand side of each panel shows the predictions and corresponding confidence intervals. Here the black line is the true data, blue lines represent the results obtained using analytical formulas, and red lines represent the Monte Carlo based estimates. The right hand side of each panel shows summary measures of the variance estimators; again the blue lines represent the analytical estimators and red lines represent the Monte Carlo estimators.

The left hand side of the Panels A-C of Figure 1 show that the point estimates are quite close to the true data, and the later one starts predicting, the smaller the errors. When prediction starts at age 16, the error at age 35 is 3 percentage points; if data is observed up to age 21, the error at age 35 is less than 2 percentage points, and is data is observed up to age 26, the error at age 35 is almost zero. From the 95% confidence intervals one can see that the analytical confidence interval estimator very closely matches the Monte Carlo estimator, which in our case should be very accurate since it uses the same model (albeit estimated, not true parameters) as the data generating process. The right hand side of the Panels A-C confirm that the two variance estimators produce very similar results: the lengths of the confidence intervals are almost equivalent, and so are the coefficients of variation and standard errors (SEs).

Figure 2 shows similar graphs for the logistic case. Again, the point estimates (left hand side of Panels A-C) are reasonably close to the true data, and the prediction errors get smaller as more data is observed. When prediction starts at age 16, the maximum error is 2 percentage points, if data is observed up to age 21, the error at age 35 is about 1 percentage point, and if data is observed up to age 26, the maximum error is essentially zero. The estimated variances and lengths of confidence intervals are slightly larger for the Monte Carlo estimator than for the analytical estimator (right hand side of Panels A-C), but in qualitative terms the estimated magnitude of uncertainty is still approximately the same.

Figure 3 shows the simulation results for the Hernes model. These are very much in line with what was observed for the Gompertz and logistic models: The point estimates are close to the true data, the errors get smaller as more data is observed, and the two variance estimators produce similar estimates.

Figure 4 represents the results shown in Figures 1-3, but zooms in to the predictions so it is easier to see how well the random walk based prediction method and the uncertainty estimators perform. The point estimates show clearly that the random walk based predictions are useful in forecasting, and that the error is smaller if more data is observed. The Figure 4 also confirms what was observed in Figures 1-3: The analytical variance estimator track closely the Monte Carlo estimator, giving a precise sense of the within-model uncertainty.

6.2 Empirical application I: French first marriages and the Hernes model

In prior research the Hernes model has been used to predict proportion married within in cohort (Goldstein and Kenney 2001; Li and Wu 2008). We do the same here, with French data. We fit the Hernes model to 1950 and 1965 cohorts. For both cohorts, we estimate the parameters of the underlying random walk with drift model using data up to age 23, and then predict the marriage rates up to age 50. Results for the Hernes model for this cohort are shown in Figue 5, Panel A. The left hand side figure of the Panel A shown the true data and predictions (analytical and Monte Carlo based) for the whole age range starting from age 14 up to age 50. The middle figure of Panel A zooms into the predictions. This figure shows that the Hernes model produces reasonable predictions for the future experience of the 1950 cohort when data is observed only up to age 23: The maximum prediction error for the analytical estimator is only 2.2 percentage points. The difference between the predictions and reality emerge quite late, after age 33, implying that at these ages the reality may not be exactly Hernesian. The right hand side of Panel A shows that the analytical and Monte Carlo estimators for variance and confidence intervals produce again results which are very close to each other.

Panel B of Figure 5 shows the results for the 1965 cohort. Again, we have used data up to age 23 when estimating the model, and have then used this estimated random walk with drift model to produce predictions and prediction errors. The left hand figure and middle figure of Panel B show that the Hernes model predicts reasonably well the cohort's experience up to age 42. The true rates at these aegs, however, seem to be increasing so rapidly that by age 50, the reality may be out of the estimated 95 % confidence interval. The right hand side figure of Panel B shows that the magnitude of the uncertainty in predictions is slightly larger according to the Monte Carlo estimator than the analytical estimator.

6.3 Empirical application II: Dutch first births and the Gompertz model

Goldsteins recent results (Goldstein 2008) indicate that the Gompertz model may work well in predicting first birth and childlessness if applied to cohort data. Here we fit the Gompertz model to Dutch data, and predict the proportion not childless for 1950 and 1965 cohorts. Experiments with the Gompertz model 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 6, Panel A. The left hand side figure of the Panel A shown the true data and predictions (analytical and Monte Carlo based) for the whole age range starting from age 15 up to age 50. The middle figure of Panel A zooms into the predictions. This figure shows that for this cohort, the Hernes model produces very accurate predictions (maximum error in the predictions is 1.1 percentage points) and that the analytical and Monte Carlo estimators give a similar picture on the uncertainty in the predictions.

Panel B of Figure 6 shows the results for the 1965 cohort. The analytical and Monte Carlo estimators produce almost equivalent results in every sense, but the worrying thing is that the true data is outside the 95 % confidence intervals. We have highlighted this fact by bolding the true data. It is, of course, true that one should expect to see the true data be outside the 95 % confidence interval on average every twentieth time, so it may be that the model is right. A potentially more likely explanation is that the cohort 1965 has pushed 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 – more explicitly, fecundity starts to decline – and this declining fecundity may be the factor explaining to high forecasts. This issue has been dealt with in more detail in Goldstein (2008).

7 Discussion

In this paper we studied the prediction and error propagation in the Gompertz, logistic, and Hernes cohort diffusion models. We showed that the linearized forms of these models can be modeled as a random walk with drift, and that predictions and prediction error estimates can be derived from the random walk model. We compared different methods for deriving predictions from the underlying random walk model, and showed that it is important to correct for the discretization error in predictions.

We also developed and compare the accuracy of a closed form analytic estimators and Monte Carlo estimators for the prediction variance. Simulation studies and empirical applications to first births and marriages showed that the estimates 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 of the Dutch 1965 cohort whose old-age fertility seems to be constrained by extra-model factors such as biology (Goldstein 2008), the random walk based estimates give a lower bound for the total uncertainty. In future work, we will use historical data for first births and marriages to compare the relative importance of the within-model error to the total error. If the within-model error accounts for a large fraction of the total error, then we will recommend our methods as useful gages of the uncertainty in forecasts. If, however, the within-model error is small, then we would recommend characterizing our methods as providing a lower-bound on uncertainty, to which a substantial amount of model specification uncertainty would need to be added.

References

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.

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

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.

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.

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

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.

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.

Appendix. Often used equations

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

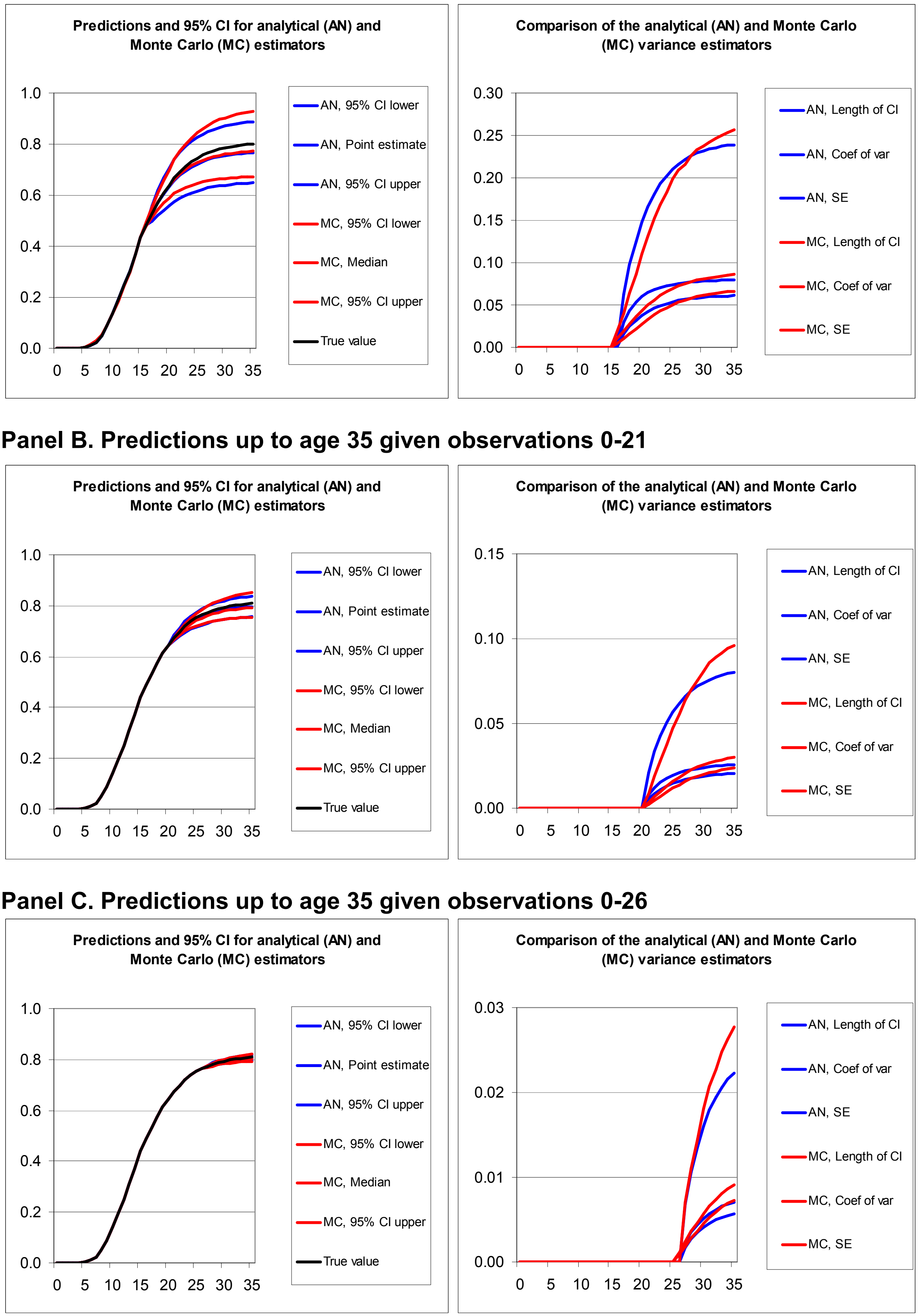
(8.6) Taylor approcimation: $$V\left(\sum_{i=1}^{k} f_i\left(X_i\right)\right) \approx \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{df_i\left(\mu_{X_i}\right)}{dX_i} \frac{df_i\left(\mu_{X_j}\right)}{dX_j} \operatorname{cov}\left(X_i, X_j\right)$$

Tables and figures

Table 1. Summary of the Gompertz, logistic and Hernes models.

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

Figure 1. Simulation results, Gompertz model. ANALYTICAL=BLUE, MONTE CARLO=RED

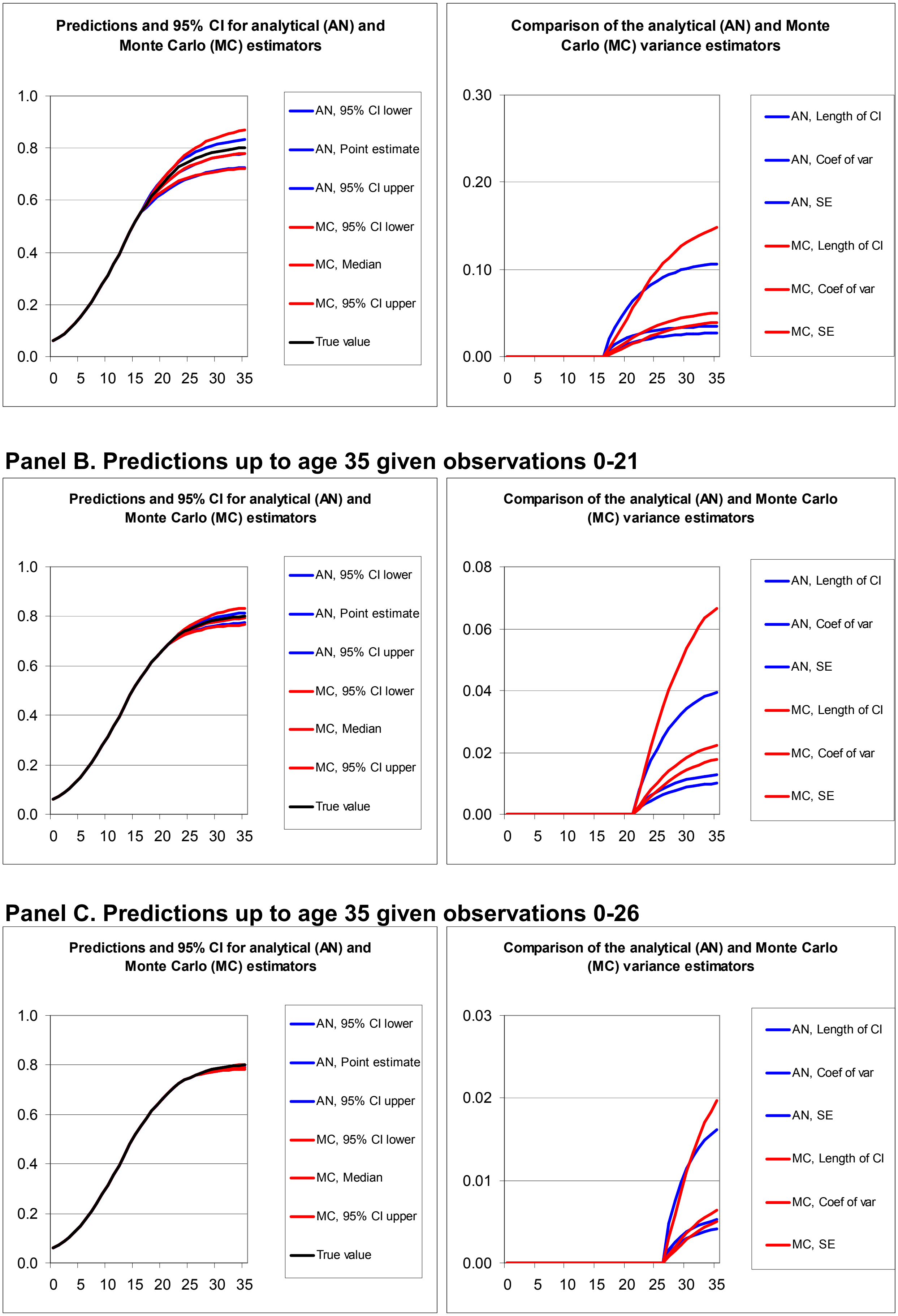


Panel A. Predictions up to age 35 given observations 0-16

Panel B. Predictions up to age 35 given observations 0-21

Panel C. Predictions up to age 35 given observations 0-26

Figure 2. Simulation results, logistic model. ANALYTICAL=BLUE, MONTE CARLO=RED

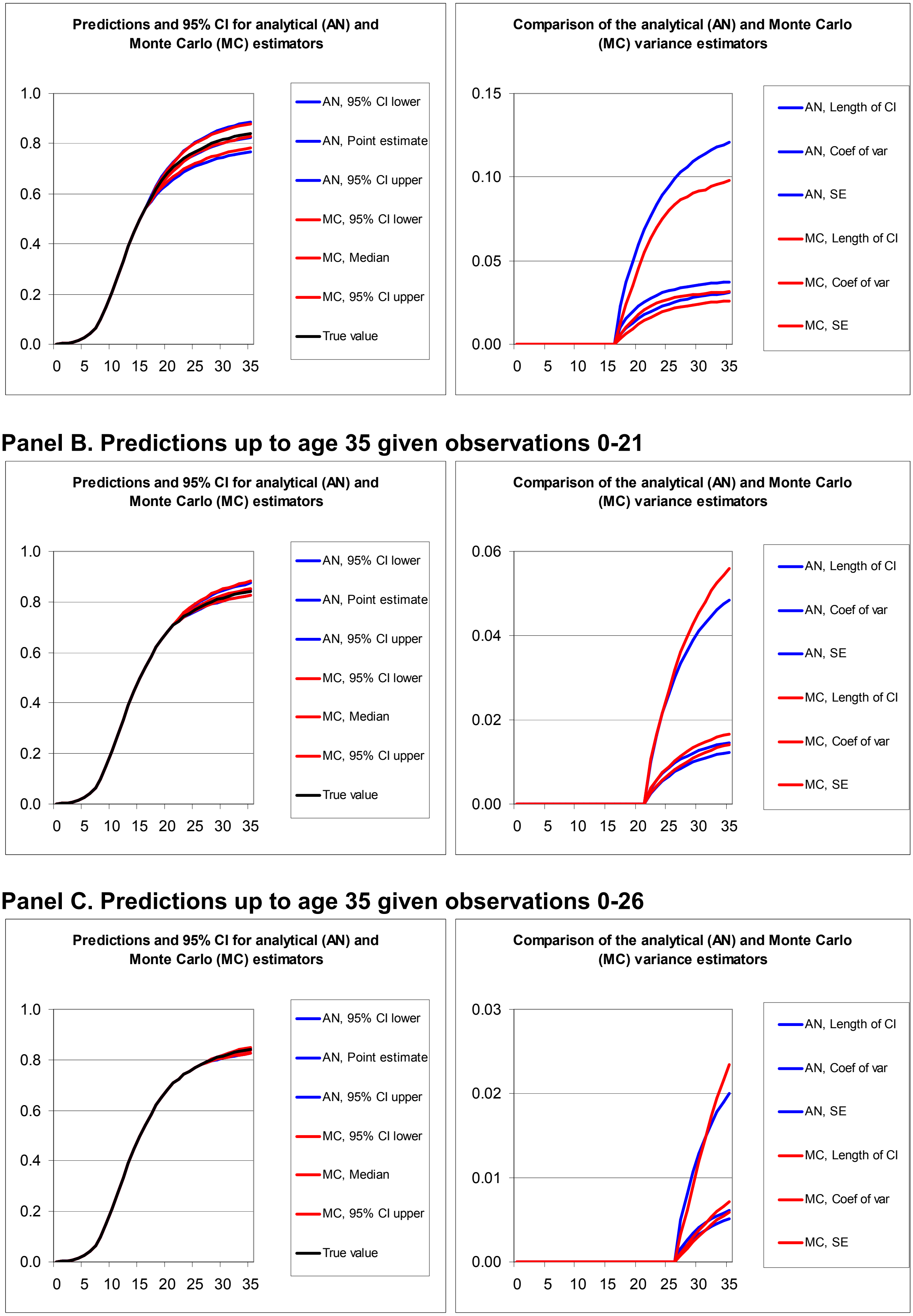


Panel A. Predictions up to age 35 given observations 0-16

Panel B. Predictions up to age 35 given observations 0-21

Panel C. Predictions up to age 35 given observations 0-26

Figure 3. Simulation results, Hernes model. ANALYTICAL=BLUE, MONTE CARLO=RED



Panel A. Predictions up to age 35 given observations 0-16

Panel B. Predictions up to age 35 given observations 0-21

Panel C. Predictions up to age 35 given observations 0-26

Figure 4. Simulation results, Gompertz, Logistic and Hernes models.

Predictions using ages 0-16 (left) and 0-26 (right). Same results as in Sections 7.1-7.3, but zoomed.

ANALYTICAL = BLUE and MONTE CARLO = RED.

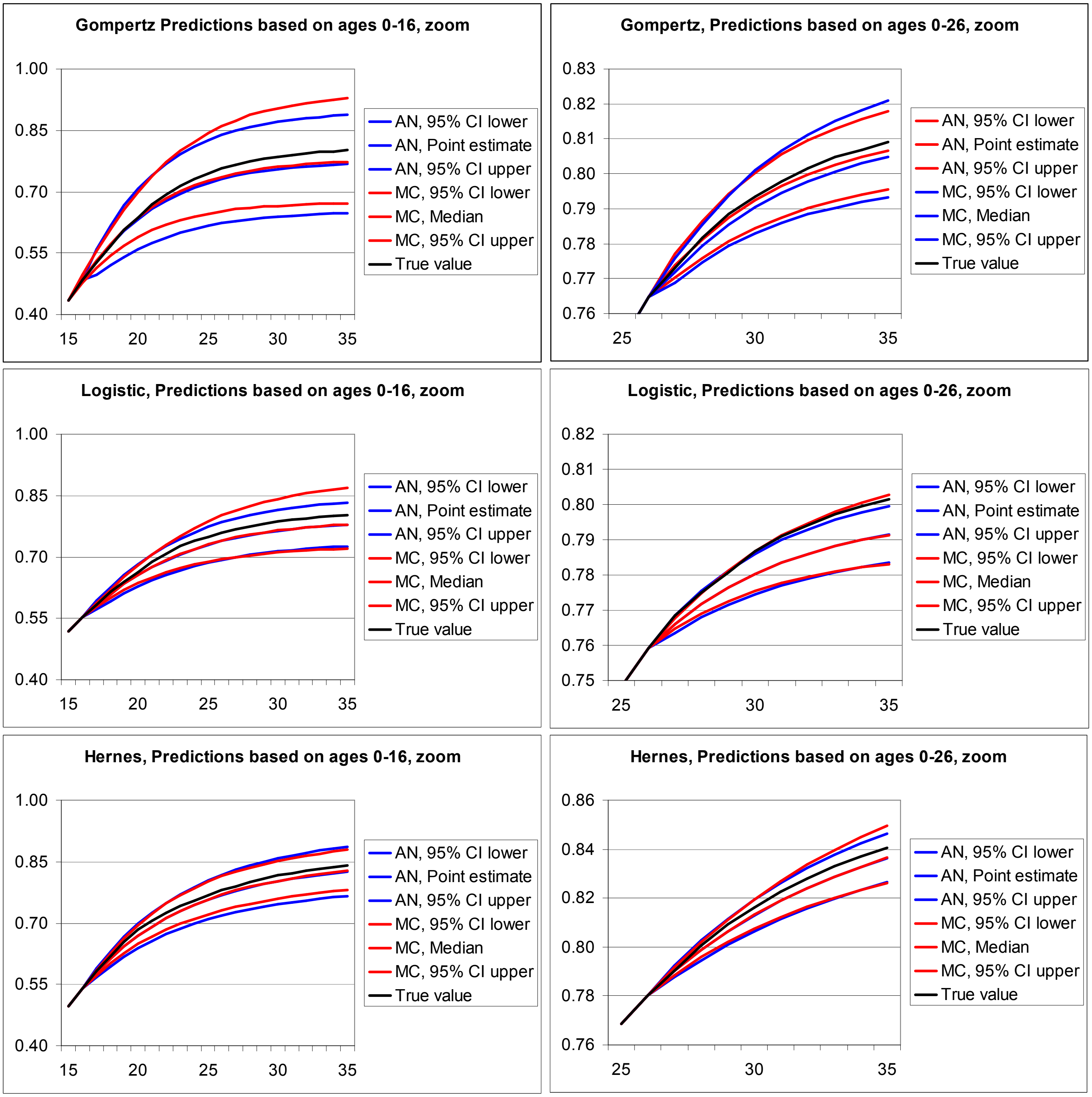
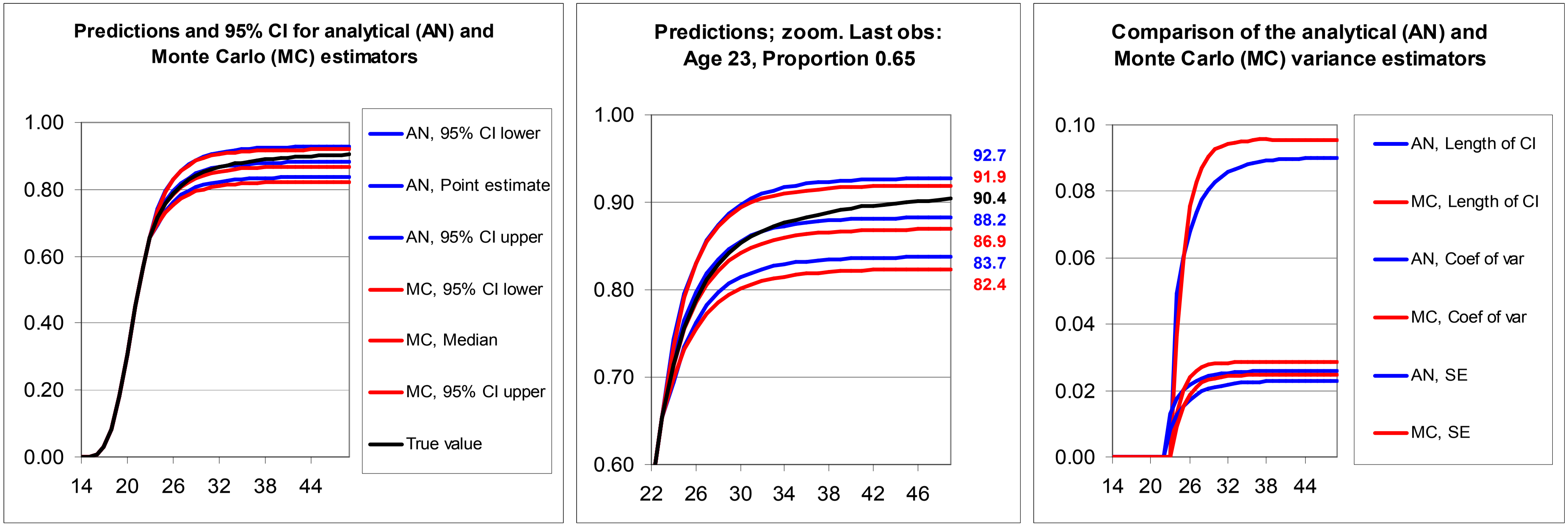


Figure 5. Empirical Application I: French first marriages, Hernes model

Panel A. Cohort 1950. Predictions based on ages 14-23, largest observed P=0.65



Panel B. Cohort 1965. Predictions based on ages 14-23, largest observed P=0.48

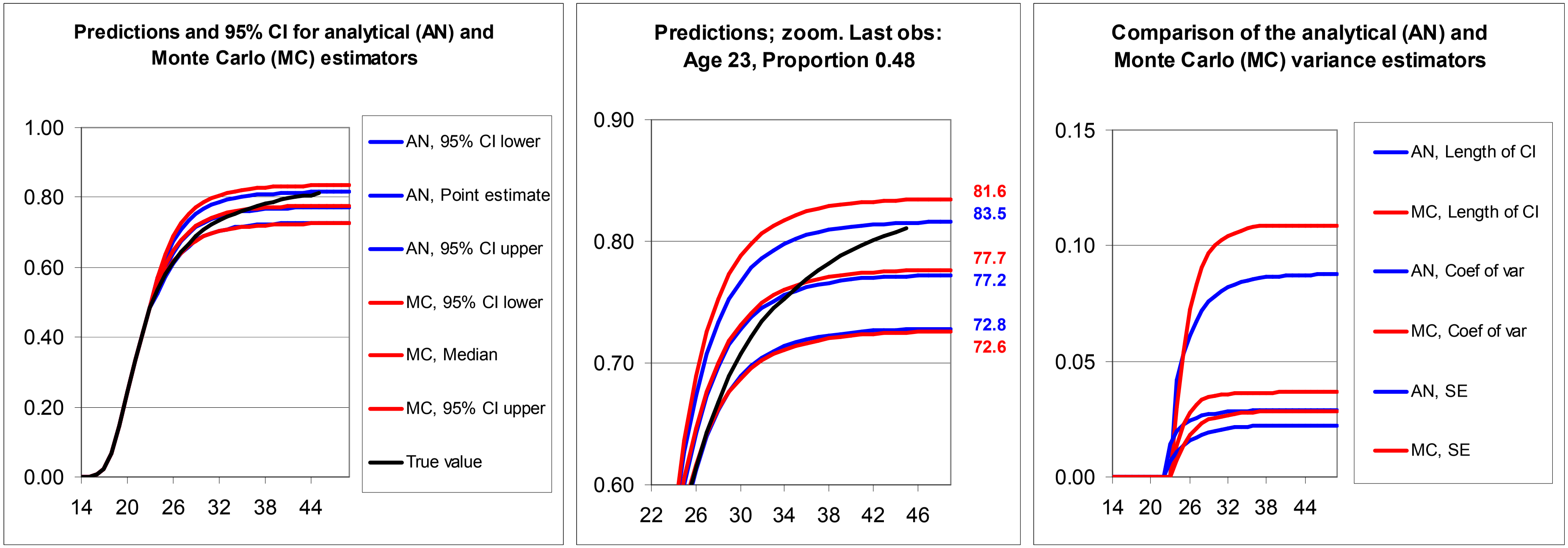
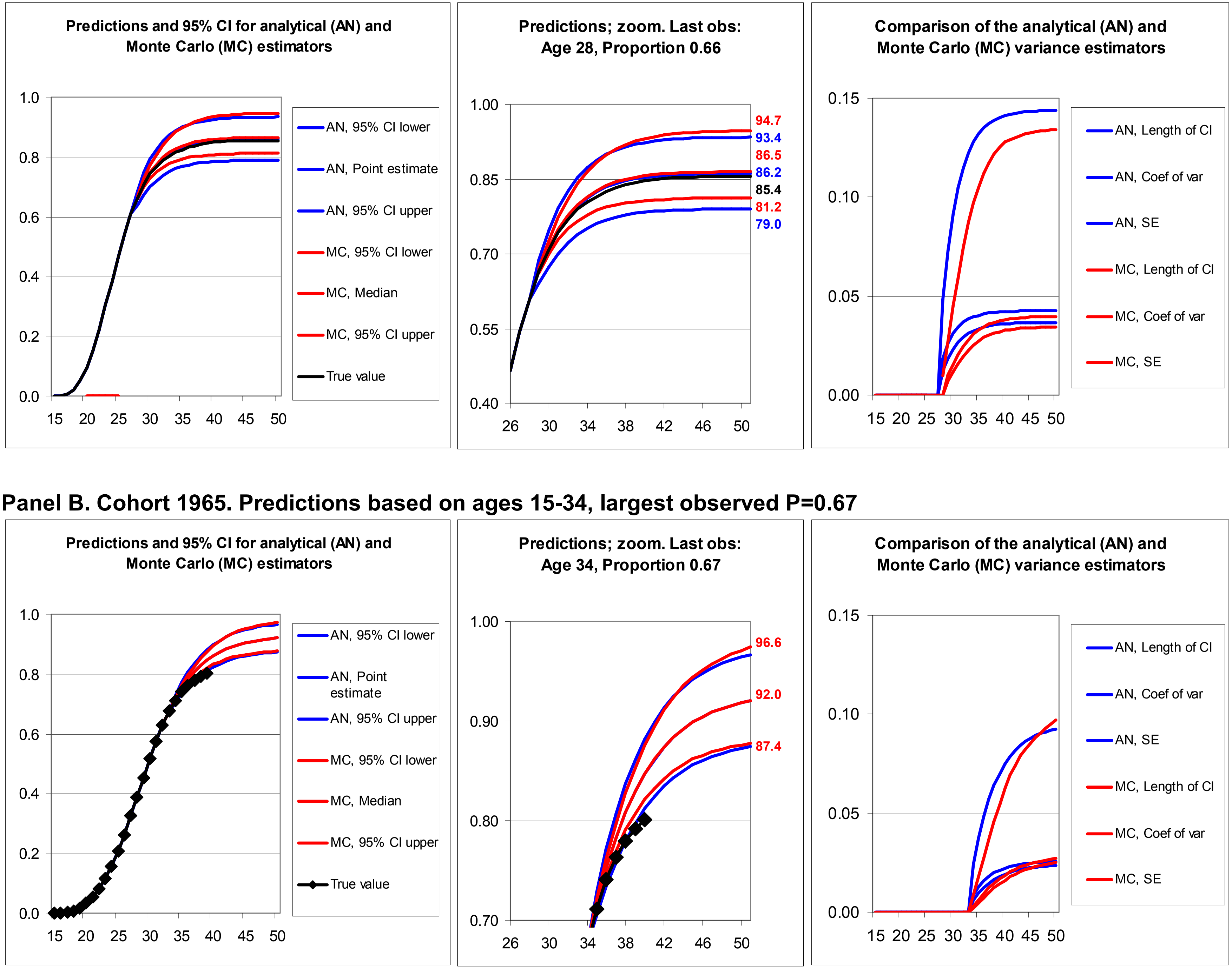


Figure 6. Empirical Application II: Dutch proportion non-childless, Gompertz model



Panel A. Cohort 1950. Predictions based on ages 15-28, largest observed P=0.66

Panel B. Cohort 1965. Predictions based on ages 15-34, largest observed P=0.67