Revista Colombiana de Estadística
January 2015, Volume 38, Issue 1, pp. 163 to 179
DOI: http://dx.doi.org/10.15446/rce.v38n1.48808

Accounting for Model Selection Uncertainty: Model Averaging of Prevalence and Force of Infection Using Fractional Polynomials

Un método para la inclusión de la incertidumbre en la selección del modelo: promedio de modelos para la prevalencia y la fuerza de infección usando polinomios fraccionarios

Javier Castañeda[1,a], Marc Aerts[2,b]

[1]Medtronic Bakken Research Center, Maastricht, Netherlands
[2]CenStat, Universiteit Hasselt, Diepenbeek, Belgium

Abstract

In most applications in statistics the true model underlying data generation mechanisms is unknown and researchers are confronted with the critical issue of model selection uncertainty. Often this uncertainty is ignored and the model with the best goodness-of-fit is assumed as the data generating model, leading to over-confident inferences. In this paper we present a methodology to account for model selection uncertainty in the estimation of age-dependent prevalence and force of infection, using model averaging of fractional polynomials. We illustrate the method on a seroprevalence cross-sectional sample of hepatitis A, taken in 1993 in Belgium. In a simulation study we show that model averaged prevalence and force of infection using fractional polynomials have desirable features such as smaller mean squared error and more robust estimates as compared with the general practice of estimation based only on one selected "best" model.

***Key words*:** Bias, Mean Squared Error, Multimodel Estimation, Seroprevalence.

Resumen

En la mayoría de aplicaciones en estadística se desconoce el verdadero modelo que determina el mecanismo de generación de los datos, y los investigadores deben confrontarse con la incertidumbre en la selección del modelo. En muchas ocasiones esta incertidumbre es ignorada cuando solo se

[a]Principal Statistician. E-mail: javier.castaneda@medtronic.com
[b]Director. E-mail: marc.aerts@uhasselt.be

usa el modelo que mejor ajusta los datos observados, lo cual conlleva a estimaciones con nivel de confianza menor a los deseados. Las enfermedades infecciosas pueden ser estudiadas por medio de parámetros tales como la prevalencia dependiente de la edad y la fuerza de infección. En este trabajo nosotros estimamos estos dos parámetros mediante polinomios fraccionarios y proponemos el uso de promedio de modelos para incluir la variabilidad debida a la incertidumbre en la selección del modelo. Nosotros ilustramos esta metodología usando una muestra de seroprevalencia de hepatitis A en Bélgica en 1993. Por medio de simulaciones mostramos que la metodología propuesta en este artículo tiene atractivas propiedades tales como menor erro cuadrado medio y estimaciones más robustas comparado con la frecuente práctica de estimación basada en un único modelo.

***Palabras clave*:** error cuadrado medio, estimación multi-modelo, seroprevalencia, sesgo.

1. Introduction

The process of understanding and explaining mechanisms and relationships in diverse scientific areas is a very complex one. A large number of observable and non-observable factors govern truth in natural phenomena, and statistical models are used in many situations to understand and represent natural relationships. In the best cases we can hope to make meaningful inferences about truth based on a good approximating model. Likelihood and least squares methods provide a rigorous inference theory if the model structure is given, however, in most practical scientific problems the model structure is unknown (Burnham & Anderson 2002, Castañeda & Gerritse 2010). A typical strategy used when analyzing data is to identify a model from a class of models and utilize the selected model for estimation purposes. This approach assumes that the data has been generated from the selected model, ignoring model selection uncertainty which leads to over-confident inferences (Hoeting, Madigan, Raftery & Volinsky 1999). This raises an important concern regarding model selection uncertainty and the need to account for this uncertainty in model selection and estimation of parameters of interest. One possibility to account for model selection uncertainty is the multimodel inference based on model averaging, in which frequentist information-theoretic (Burnham & Anderson 2002) and Bayesian (Hoeting et al. 1999) perspectives provide a well defined methodology. In this paper we focus on the frequentist information-theoretic approach based on the estimated relative Kullback-Leibler distance (Kullback & Leibler 1951).

Infectious diseases, including Hepatitis A, can be studied by a model that allows the estimation of parameters such as age-dependent prevalence and force of infection, which are susceptible to model selection uncertainty. An example of model selection uncertainty is when two or more models fit the data equally well according to a given criterion, but do not give the same estimate of the age-dependent prevalence and force of infection. Small differences from one model to another might not severely affect the estimation of prevalence, but in the case

of force of infection, a primary epidemiological parameter, these differences can largely affect the estimation of such a sensitive parameter.

Shkedy, Aerts, Molenberghs, Beutels, & Damme (2006) described the use of fractional polynomials to estimate age-dependent prevalence and force of infection on seroprevalence samples of hepatitis A, rubella, mumps and varicella. They showed the advantages of this flexible modeling technique and advised its use in the estimation of force of infection.

In this paper we extend the work from Shkedy et al. (2006) by proposing a methodology that accounts for model selection uncertainty.

Faes, Aerts, Geys & Molenberghs (2007) explained that normally it is not clear what models should be used for model averaging and suggested that fractional polynomials are a natural and flexible family of parametric models, lending itself nicely as the set of models to be used in model averaging. They also showed the application of fractional polynomials for the estimation of a safe dose level of exposure in the framework of model averaging. Goeyvaerts, Hens, Ogunjimi, Aerts, Shkedy, Damme & Beutels (2010) also considered multimodel inference in the context of estimation of infectious disease parameters.

In this paper we propose the use of the frequentist information-theoretic approach for model averaging of fractional polynomials to account for model selection uncertainty when estimating age-dependent prevalence and force of infection. These two age-dependent parameters are defined in Section 2. In Section 3 we present fractional polynomials in the context of models for binary responses. The use of model averaging of fractional polynomial estimates to account for model selection uncertainty is described in Section 4. In Section 5 we illustrate the application of model averaging using fractional polynomials to estimate age-dependent prevalence and force of infection of hepatitis A in Belgium (Beutels, Damme & Aelvoet 1997) with models assuming a logistic form for the fraction of disease-susceptible individuals at age a. In Section 6 we present a simulation study that shows the advantages of the proposed methodology over the traditional strategy of selecting one single "best" model. We finish with a discussion in Section 7.

2. Age-Dependent Prevalence and Force of Infection

Mathematical models consisting of a set of differential equations which aim to describe the flow of individuals from one disease stage to the other are often used to describe the process of infectious diseases. Under some conditions (lifelong immunity, disease irreversible and negligible mortality caused by the infection) the partial differential equation which describes the change in the susceptible fraction at a certain age is called force of infection (Shkedy et al. 2006). In epidemiology the force of infection is the rate at which susceptible individuals become infected by an infectious disease. The force of infection can be used to compare the rate of transmission between different groups of the population for the same infectious disease or even between different infectious diseases. Several authors have proposed different methods for estimation. Farrington (1990) used nonlinear

models to estimate force of infection. For the estimation of both the prevalence and the force of infection Keiding (1991) used isotonic regression and Shkedy, Aerts, Molenberghs, Beutels & Damme (2003) used local polynomials. None of these authors accounted for model selection uncertainty in the estimation of prevalence and force of infection.

Typically the force of infection can be estimated from a seroprevalence age-specific cross-sectional sample of size N with a_j the age of the j^{th} subject. Seroprevalence data are binary data defined as: $Y = 1$ if $Z > \zeta$ or $Y = 0$ if $Z \leq \zeta$; where Z is the (log of the) antibody level and ζ a certain threshold value. The binary variable Y_j will take value 1 if subject j had experienced infection before age a_j and value 0 otherwise.

Shkedy et al. (2006) assumed that the disease is irreversible, meaning that the immunity is lifelong, and that the mortality caused by the infection is negligible and can be ignored. Under these assumptions the partial differential equation describing the change in the susceptible fraction at age a and time t is given by:

$$\frac{\partial}{\partial a}q(a,t) + \frac{\partial}{\partial t}q(a,t) = -\ell(a,t)q(a,t)$$

where $q(a,t)$ is the fraction of susceptible individuals at age a and time t, and $\ell(a,t)$ is the hazard or force of infection describing the rate at which susceptible individuals become infected. In a steady state, the term involving the time derivative is equal to zero and the partial differential equation reduces to the following ordinary differential equation:

$$\frac{d}{da}q(a) = -\ell(a)q(a).$$

This representation of the model is known as the static model and describes the change in the susceptible fraction with the host age. The age-dependent prevalence is given by $\pi(a) = 1 - q(a)$.

If $q(a)$ is the fraction of susceptible individuals at age a, $\pi(a) = 1 - q(a)$ can be defined as the probability to be infected at age a and the log-likelihood is given by (Shkedy et al. 2006):

$$L = \sum_{j=1}^{N} \log(\pi(a_j)) + (1 - Y_j)\log(1 - \pi(a_j)).$$

Here $\pi(a) = g^{-1}(\eta(a))$, where $\eta(a)$ is the linear predictor and g is the link function. For binary responses, g is often taken to be a logit link function, $\log(\pi/(1-\pi))$.

Using the static model and assuming a logit link function, the force of infection can be written as:

$$l(a) = -\frac{q'(a)}{q(a)} = \frac{\pi'(a)}{1-\pi(a)} = \eta'(a)\frac{e^{\eta(a)}}{1+e^{\eta(a)}}$$

Shkedy et al. (2006) derived the expression of the force of infection for other link functions.

3. Fractional Polynomials

The motivation to model the force of infection with fractional polynomial as linear predictors, given a link function, is to allow for flexible changes on the force of infection over the age of the host (Shkedy et al. 2006). Conventional polynomials do not have asymptotes and fit the data poorly whenever asymptotic behavior of the infection is expected. High order conventional polynomials offer a wide range of curve shapes but often fit the data badly at the extremes of the observed age and fit the data poorly whenever asymptotic behavior of the infection process is expected.

Fractional polynomials are a generalization of the conventional polynomial class of functions (Royston & Altman 1994). Fractional polynomials can be described as sums of m terms of the form a^{p_l} or $a^{p_l}\log(a)$. In the context of binary responses, a fractional polynomial of degree m for the linear predictor is defined as (Shkedy et al. 2006),

$$\eta_m(a, \boldsymbol{\beta}, p_1, p_2, \ldots, p_m) = \sum_{l=0}^{m} \beta_l H_l(a),$$

where m is an integer, $p_1 \leq p_2 \leq \ldots \leq p_m$ is a sequence of powers, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_m)$ are regression parameters and $H_l(a)$ is a transformation function given by

$$H_l(a) = \begin{cases} a^{p_l} & p_l \neq p_{l-1} \\ \\ H_{l-1}(a) \times \log(a) & p_l = p_{l-1} \end{cases}$$

With $p_0 = 0$ and $H_0 = 1$. The definition includes possible "repeated powers" that involve powers of $\log(a)$, as in the case of having a fractional polynomial of degree $m = 2$ with "repeated powers" $(1, 1)$ which takes the form $\beta_0 + \beta_1 a + \beta_2 a \ln(a)$.

In this paper we focus on generalized linear models with logit link function and fractional polynomials of degree $m = 2$ as linear predictors for which the only covariate in the model is the host age. Royston & Altman (1994) argued that, fractional polynomials of order higher than two are rarely needed in practice and suggested to choose the value of the powers from a set similar to $\{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, \max(3, m)\}$.

One problem that arises when a higher order polynomial model is fitted is that the estimate for the force of infection can become negative. To avoid this, one can fit fractional polynomials over the sequence of powers, and check for each fitted model if $\eta'_m(a, \boldsymbol{\beta}, p_1, p_2, \ldots, p_m) \geq 0$, for all ages a. In case that a given sequence of powers leads to a negative derivative of the linear predictor, the model is not considered an appropriate model.

Even though the selection of a single model (e.g. the best fitting model) for estimation purposes is very common, this approach ignores model selection uncertainty. Shkedy et al. (2006) selected the model with the best goodness-to-fit among all fractional polynomials for which $\eta'_m(a, \boldsymbol{\beta}, p_1, p_2, \ldots, p_m) \geq 0$, which ignores model selection uncertainty. In the next section we present a methodology that accounts for model selection uncertainty.

4. Model Selection Uncertainty

Likelihood and least squares methods provide a rigorous inference theory if the model structure is given. However, in many practical scientific problems there is not a given pre-specified model structure and there is uncertainty about the correct model to be used. Model selection uncertainty arises when the data are used for both, model selection and parameter estimation.

Denote the sampling variance of an estimator $\hat{\theta}$, given a model, by var($\hat{\theta}$|model). More generally, the sampling variance of $\hat{\theta}$ should have two components:

1) var($\hat{\theta}$|model)

2) a variance component due to not knowing the best approximating model to be used and therefore having to estimate it (Burnham & Anderson 2002).

Failure to allow for model selection uncertainty often results in estimating sampling variance that is too low and confidence interval coverage below the nominal value.

4.1. Multimodel Inference: Model Averaging

Traditionally when a set of fractional polynomials is used, attention is restricted to the so called "best" fitting model. This approach assumes that the data has been generated from the selected fractional polynomial model, ignoring model selection uncertainty which leads to over-confident inferences (Hoeting et al. 1999). Instead of focusing on one single model, one can treat each model in the set of fractional polynomial models as a possible model of interest (Faes et al. 2007).

Given a set of R fractional polynomial models that lead to a non-negative force of infection, specified independently of the sample data, formal inferences can be made based on the entire set of models. In this paper our focus is on models assuming a logistic form to estimate the prevalence of infection where the fractional polynomial is the linear predictor and the only covariate in the model is the host age. The linear predictor is not of main interest here, but rather the resulting estimate of the prevalence, and the force of infection. Each one of the R possible models yields an estimate of the age-dependent prevalence. We define a model-averaged (MA) estimate of the age-dependent prevalence of infection, at age a, as

$$\hat{\pi}_{MA}(a) = g^{-1}(\hat{\eta}_{MA}(a)) = \frac{e^{\hat{\eta}_{MA}(a)}}{1 + e^{\hat{\eta}_{MA}(a)}}, \tag{1}$$

where $\hat{\eta}_{MA}(a) = \sum_{i=1}^{R} w_i \hat{\eta}_i(a)$, w_i is the weight of evidence in favor of model i and $\hat{\eta}_i(a)$ is the estimated fractional polynomial for model i. Burnham & Anderson (2002) proposed the weights w_i defined as the Akaike weights (Akaike 1974),

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{r=1}^{R} \exp\left(-\frac{1}{2}\Delta_r\right)}$$

where $\Delta_i = AIC_i - AIC_{\min}$ the Akaike Information criterion (AIC) difference between the AIC of model i and the minimum AIC calculated in the entire set of

models. Here $AIC_i = 2k_i - 2L_i$, with k_i the number of parameters in model i and L_i the log-likelihood for model i. The w_i depends on the entire set of R logistic fractional polynomial models and add to 1. Part of multimodel inference includes first ranking the fitted logistic fractional polynomial models from best to worst, based on the AIC differences $\Delta_i = AIC_i - AIC_{\min}$. After ranking the models the second step is scaling them to obtain the relative plausibility of each fitted model by a weight of evidence (w_i) relative to the selected best model. The AIC is an asymptotically unbiased estimator of the relative, expected Kullback-Leibler (K-L) distance (Faes et al. 2007). As a result, the weight w_i is the weight of evidence in favor of model i for being the K-L best model, given the design and sample size. In this way, the best model k has $\Delta_k = 0$, and thus $\exp\left(-\frac{1}{2}\Delta_k\right) = 1$. The larger Δ_i, the smaller $\exp\left(-\frac{1}{2}\Delta_i\right)$, and the smaller weight w_i given to the model.

Unconditional inferences about precision can be made over the entire set of models by using the Akaike weights (w_i) and the sampling variance of the estimator $\hat{\eta}$ given logistic fractional polynomial model i. The latter is also called the conditional sampling variance, $\text{var}(\hat{\eta}|\text{model}_i)$.

Faes et al. (2007) derived the unconditional variance of the model-averaged estimate as follows. When η is estimated from a specific logistic fractional polynomial model with parameters $\boldsymbol{\beta}$, then the variance of $\hat{\eta}$ is estimated as

$$\widehat{\text{Var}}(\hat{\eta}) = \left(\frac{\partial \eta}{\partial \boldsymbol{\beta}}\right)^T \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) \left(\frac{\partial \eta}{\partial \boldsymbol{\beta}}\right) \Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}},$$

With $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$. However, when interested in the model-average estimator $\eta_{MA} = \sum_{i=1}^{R} w_i \eta_i$, which is estimated as $\hat{\eta}_{MA} = \sum_{i=1}^{R} w_i \hat{\eta}_i$, the variance of $\hat{\eta}_{MA}$ can be expressed as:

$$\begin{aligned}\text{Var}(\hat{\eta}_{MA}) = &\sum_{i=1}^{R} w_i^2 \boldsymbol{E}\left((\hat{\eta}_i - \eta_{MA})^2 \mid M_i\right) \\ &+ \sum_{i=1}^{R}\sum_{\substack{j=1 \\ i\neq j}}^{R} w_i w_j \boldsymbol{E}\left((\hat{\eta}_i - \eta_{MA})(\hat{\eta}_j - \eta_{MA}) \mid M_i, M_j\right)\end{aligned}$$

The first term, the mean squared error of $\hat{\eta}_i$ given model M_i, can be written as

$$\begin{aligned}\boldsymbol{E}\left((\hat{\eta}_i - \eta_{MA})^2 \mid M_i\right) &= \boldsymbol{E}\left((\hat{\eta}_i - \eta_i + \eta_i - \eta_{MA})^2 \mid M_i\right) \\ &= \boldsymbol{E}\left((\hat{\eta}_i - \eta_i) \mid M_i\right)^2 + (\eta_i - \eta_{MA})^2 \\ &+ 2(\eta_i - \eta_{MA})\boldsymbol{E}\left((\hat{\eta}_i - \eta_i) \mid M_i\right) \\ &= \text{Var}(\hat{\eta}_i \mid M_i) + (\eta_i - \eta_{MA})^2,\end{aligned}$$

which is the sum of the conditional variance $\text{Var}(\hat{\eta}_i \mid M_i)$ of $\hat{\eta}_i$ given model M_i and the squared bias of η_i with the model-averaged parameter η_{MA}. The covariance term can be written as

$$\begin{aligned}\boldsymbol{E}\left((\hat{\eta}_i-\eta_{MA})(\hat{\eta}_j-\eta_{MA})\mid M_i,M_j\right)&\\ =\rho_{ij}\sqrt{\boldsymbol{E}\left((\hat{\eta}_i-\eta_{MA})^2\mid M_i\right)\boldsymbol{E}\left((\hat{\eta}_j-\eta_{MA})^2\mid M_j\right)}&\\ =\rho_{ij}\sqrt{\left(\mathrm{Var}(\hat{\eta}_i\mid M_i)+(\eta_i-\eta_{MA})^2\right)\left(\mathrm{Var}(\hat{\eta}_j\mid M_j)+(\eta_j-\eta_{MA})^2\right)},&\end{aligned}$$

where ρ_{ij} represents the across-model correlation of $\hat{\eta}_i$ with $\hat{\eta}_j$, with respect to η_{MA}. As a result,

$$\begin{aligned}\mathrm{Var}(\hat{\eta}_{MA})=&\sum_{i=1}^{R}w_i^2\left(\mathrm{Var}(\hat{\eta}_i\mid M_i)+(\eta_i-\eta_{MA})^2\right)\\ &+\sum_{\substack{i=1\\ i\neq j}}^{R}\sum_{j=1}^{R}w_iw_j\rho_{ij}\\ &\times\sqrt{\left(\mathrm{Var}(\hat{\eta}_i\mid M_i)+(\eta_i-\eta_{MA})^2\right)\left(\mathrm{Var}(\hat{\eta}_j\mid M_j)+(\eta_j-\eta_{MA})^2\right)}.\end{aligned}$$

Often, it is assumed that there is an almost perfect correlation between the estimates from different models, and ρ_{ij} is conservatively assumed to be equal to 1 (Buckland, Burnham & Augustin 1997). In this case the unconditional variance can be written as in Burnham & Anderson (2002),

$$\mathrm{Var}(\hat{\eta}_{MA})=\left(\sum_{i=1}^{R}w_i\sqrt{\mathrm{Var}(\hat{\eta}_i\mid M_i)+(\eta_i-\eta_{MA})^2}\right)^2 \tag{2}$$

This unconditional variance can be estimated by use of estimates $(\hat{\eta}_i,\hat{\eta}_{MA})$ instead of parameters (η_i,η_{MA}).

For large samples, $\hat{\eta}_{MA}(a)\pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\eta}_{MA}(a))}$ is a confidence interval for the true model-averaged logit at age a. The unconditional variance can be used to estimate unconditional confidence limits for the age-dependent prevalence of infection. Using the transformation $\hat{\pi}_{MA}(a)=\frac{e^{\hat{\eta}_{MA}(a)}}{1+e^{\hat{\eta}_{MA}(a)}}$, the endpoints of the confidence interval for the estimated model-averaged logit invert to a corresponding confidence interval for the estimated age-dependent model-averaged prevalence (Agresti 2002).

Multimodel inference using model averaging is recommended because it accounts for the additional variability that is induced by the model selection process. If the purpose is to compare different established models with well-known physical interpretations, and the investigator is not confronted with model selection uncertainly, then multimodel inference may not be required.

5. Application to a Seroprevalence Sample of Hepatitis A

In the previous sections we illustrated the methodology of model averaging using fractional polynomials to account for model selection uncertainty. In this section this methodology is applied to the estimation of age-dependent prevalence and force of infection of hepatitis A with models assuming a logistic form for the fraction of disease-susceptible individuals at age a and fractional polynomials that lead to a non-negative force of infection. We restrict the discussion to models for which the only covariate in the model is the host age.

Viral hepatitis is a serious health problem throughout the world. In Flanders, Belgium, a sero-epidemiological study was undertaken between April 1993 and February 1994 in a sample of the general population. A total of 4,058 blood samples were drawn and collected in 10 hospitals in Flanders. Beutels et al. (1997) published the data of this study in which the objective was to calculate the number of persons in the population who tested positive for hepatitis A based on serology (blood serum) specimens. The purpose of this study was to obtain a clear understanding of the prevalence of hepatitis A. This data has also been analyzed by Shkedy et al. (2006) and Hens, Shkedy, Aerts, Faes, Van Damme & Beutels (2012), among others.

5.1. Model Averaging of Age-Dependent Prevalence Using Fractional Polynomials

Using the methods described in section three and the set of powers $\{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 3\}$, fractional polynomials of degree one and two are used as linear predictors to estimate the prevalence of hepatitis A. A total of 55 generalized linear models with logit link function are fitted and 32 of these models are used for analysis as they lead to non-negative force of infection. Each of the $R = 32$ models corresponds to a different fractional polynomial used as linear predictor based on the host age. Considering the AIC criterion, the "best" fractional polynomial is the one with the smallest AIC. Typically the "best" fractional polynomial would be used for estimation of prevalence but this procedure ignores model selection uncertainty. Model averaging is used to account for this uncertainty in the estimation of age-dependent prevalence.

The first step in the model averaging process is the ranking of the 32 models from best to worst, based on $\Delta_i = AIC_i - AIC_{\min}$. The "best" model is the one with AIC=$AIC_{\min}$ and AIC difference $\Delta_i = 0$. After ranking the models, weights of evidence relative to the selected best model are used to scale and obtain the relative plausibility of each fitted model. To do this Akaike weights w_i are calculated for each model. The ranking and scaling of the complete set of fractional polynomials is presented in Table 1. The best and the second best fitting fractional polynomials (logistic models with linear predictors of the form $\beta_0 + \beta_1 a + \beta_2 a^{1.5}$ and $\beta_0 + \beta_1 a + \beta_2 a \log(a)$, respectively) are equally parsimonious and have similar maximized log-likelihood, -184.1 and -188.0 respectively. The third and fourth

models also appear to have an appropriate fit to the data. Thus, several models are nearly equally good for fitting the data, illustrating model selection uncertainty. Figure 1 shows the nearly equal estimated age-specific prevalence using the best fitting four fractional polynomial logistic regression models.

TABLE 1: Ranking and scaling of 32 fractional polynomials that lead to non-negative force of infection. Powers (p1, p2), AIC value, and AIC weights (w).

p1	p2	AIC	w	p1	p2	AIC	w
1	1.5	374.27	0.9740	-0.5	1.5	427.54	0.0000
1	1	382.00	0.0204	-0.5	2	443.30	0.0000
0.5	2	385.05	0.0045	-1	1.5	447.06	0.0000
0.5	1.5	388.32	0.0009	-1.5	1.5	462.90	0.0000
0.5	1	391.33	0.0002	-0.5	3	467.78	0.0000
0.5	0.5	393.53	0.0001	-2	1.5	474.39	0.0000
-0.5	0.5	394.40	0.0000	-1	2	477.05	0.0000
0	0.5	394.53	0.0000	-0.5	0	499.57	0.0000
0	1	401.27	0.0000	-1.5	2	506.57	0.0000
0	1.5	406.84	0.0000	-2	2	529.47	0.0000
-0.5	1	410.43	0.0000	-1	3	530.65	0.0000
0	2	410.80	0.0000	-1.5	3	590.59	0.0000
0	3	414.62	0.0000	-2	3	639.14	0.0000
-1	1	417.56	0.0000	-1	0	654.24	0.0000
-1.5	1	422.25	0.0000	-1.5	0	855.24	0.0000
-2	1	424.89	0.0000	-2	0	1043.96	0.0000

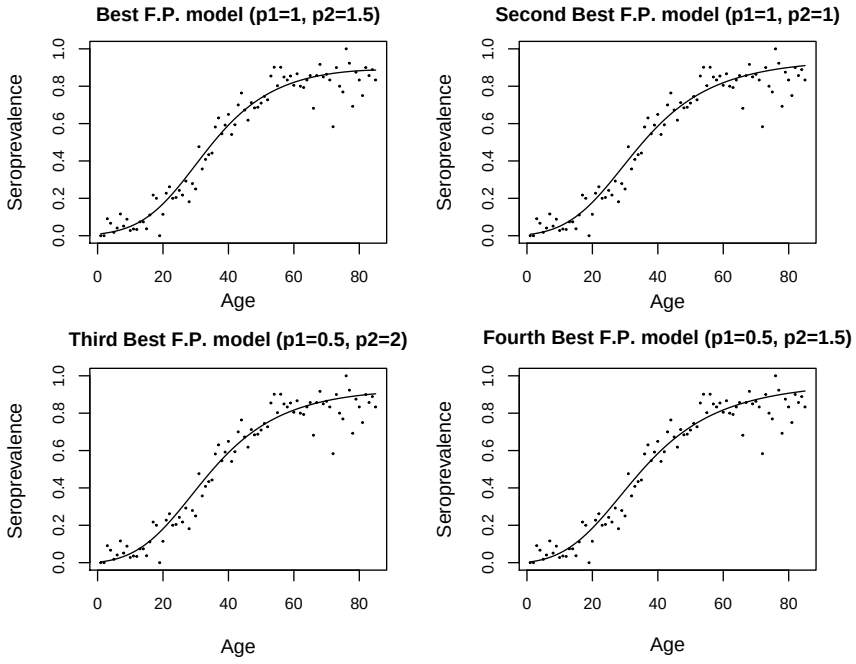

FIGURE 1: Estimated best four fractional polynomial models for prevalence.

Once the model-averaged estimate of the age-dependent prevalence of infection is estimated using equation (1) and the calculated Akaike weights, model selection uncertainty is accounted for by the unconditional variance function in equation (2) using the set of 32 logistic fractional polynomial models. The unconditional variance is used to estimate unconditional confidence limits for the age-dependent prevalence of infection as shown at the end of Section 4.1. Model averaged and unconditional confidence limits for the age-dependent prevalence of infection at ages from 1 to 85 are presented in Figure 2.

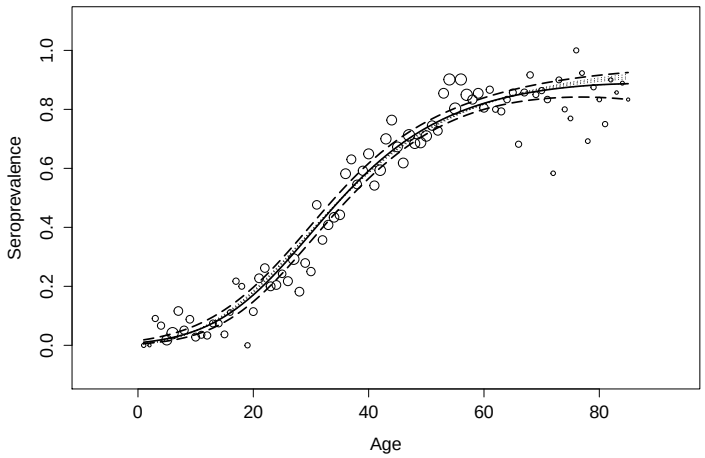

FIGURE 2: Model averaged estimated age-dependent prevalence (solid line) and unconditional 95% confidence limits (dashed line). Dotted lines correspond to the estimated prevalence with each of the four best-fitting fractional polynomials. The size of the circles is proportional to the number of observed subjects at each age.

5.2. Model Averaging of Age-Dependent Force of Infection Using Fractional Polynomials

In the previous section we used model averaging of logistic fractional polynomial models for estimating the age-dependent prevalence of hepatitis A. Using the same fractional polynomial models we can proceed to estimate the force of infection.

Using the static model and assuming a logit link function, the force of infection can be estimated as $\hat{l}(a) = \hat{\eta}'(a)\frac{e^{\hat{\eta}(a)}}{1+e^{\hat{\eta}(a)}}$, where $\hat{\eta}(a)$ is an estimated (fractional polynomial) linear predictor and $\hat{\eta}'(a)$ is its first derivative with respect to age. Using $\hat{l}_i(a)$, the estimated force of infection from fractional polynomial i, the model averaged force of infection is estimated as $\hat{l}_{MA}(a) = \sum_{i=1}^{R} w_i \hat{l}_i(a)$.

The unconditional variance of the force of infection can be estimated as $\widehat{\text{Var}}(\hat{l}_{MA})$ $= \left(\sum_{i=1}^{R} w_i \sqrt{\widehat{\text{Var}}(\hat{l}_i \mid M_i) + (\hat{l}_i - \hat{l}_{MA})^2}\right)^2$. Using the delta method, $\widehat{\text{Var}}(\hat{l}_i \mid M_i)$

can be calculated as $\widehat{\text{Var}}(\hat{l}_i \mid M_i) = \left[\hat{\varphi}_i(\hat{\eta}_i'' + \hat{\eta}_i'^2(1-\hat{\varphi}_i))\right]^2 \widehat{\text{Var}}(\hat{\eta}_i \mid M_i)$ with $\hat{\varphi}_i = \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}}$, and $\hat{\eta}_i'$ and $\hat{\eta}_i''$ the first and second derivatives with respect to age of the estimated logit, respectively.

Figure 3 shows the estimated model averaged force of infection (continuous line [M.A.F.I]) and unconditional 95% confidence limits (95% C.l1. for F.I.) using the 32 fractional polynomials. It also shows the estimated force of infection (F.I.) from the best four fractional polynomials (F.I.- 1st, F.I.- 2nd, F.I.- 3rd and F.I.- 4th). Figure 3 displays an important issue when working with the force of infection; despite the almost equal fit of the best four models for the prevalence (Figure 1), the estimated force of infection using each of these models shows clear differences and model selection uncertainty is even more evident for the force of infection. Small differences from one model to another could slightly affect the estimation of prevalence, but in the case of the force of infection, a primary epidemiological parameter, small differences in these models can severely affect the estimation of such a sensitive parameter.

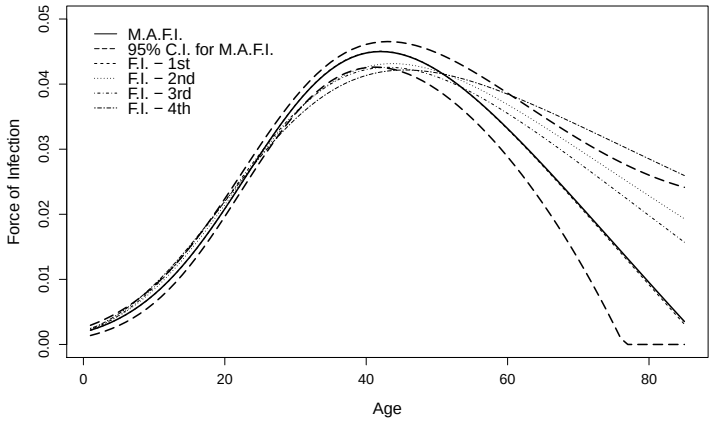

FIGURE 3: Estimated model averaged force of infection, unconditional 95% confidence limits and estimated force of infection from the four best-fitting fractional polynomial models.

6. Simulation Study

We have applied the methods discussed in previous sections to the single cross sectional seroprevalence dataset of hepatitis A. In order to assess the methodology of model averaging using fractional polynomials to estimate age-dependent prevalence and force of infection, we present information on true expected values, such as variance, bias and mean squared error of the multimodel estimator based on a simulation study of $B = 500$ runs. Using fractional polynomials that lead to non-negative force of infection, here we concentrate on the comparison of the estimates using only the best fitting fractional polynomial model (which ignores model selection uncertainty) and the model averaged estimate which accounts for

model selection uncertainty using fractional polynomials based on the set of powers $\{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 3\}$.

To study the characteristics of the model averaged estimate for prevalence and force of infection, we assume that we know the underlying data generation mechanism. Thus, it is possible to generate several independent datasets from the same underlying process. Five hundred independent seroprevalence datasets are generated assuming an underlying process based on $\eta_{true}(a) = -3.0 + 0.02(age)^{1.5} - 0.02(age)^{1.3}$. The fractional polynomial class does not encompass this linear predictor, as the power 1.3 is not included in the set of powers.

The *true* prevalence and *true* force of infection at age a are calculated as $\pi_{true}(a) = \frac{e^{(\eta_{true}(a))}}{1+e^{(\eta_{true}(a))}}$ and ${}_{true}(a) = \eta'_{true}(a)\pi_{true}(a)$, respectively. Therefore the true prevalence and the true force of infection are known and comparisons can be done based on the asymptotic true variance, bias and mean squared error (MSE).

For each age (from 1 to 85 years old) the number of positive cases (number of infected cases at age a) is generated using a binomial distribution with parameters $\pi_{true}(a)$ and n_i = total number of people at age a sampled in 1993 in Belgium. Figure 4 shows one of such a generated seroprevalence datasets (circles) and the 'true' age-dependent prevalence (solid line).

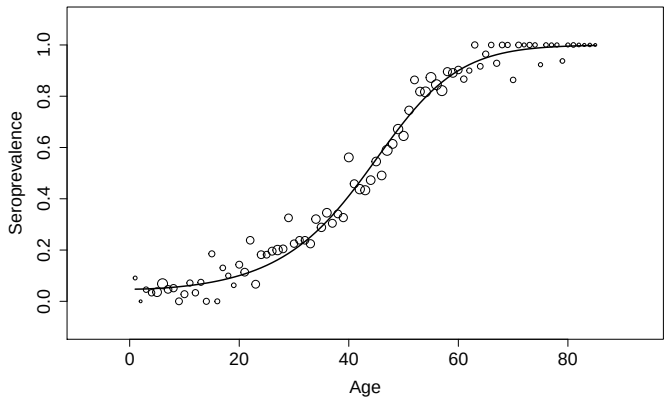

FIGURE 4: A simulated seroprevalence dataset (circles) and the *true* age-dependent prevalence $\pi_{true}(a)$ (solid line) based on the polynomial with $p_1 = 1.5$ and $p_2 = 1.3$. The size of the circles is proportional to the number of subjects at each age.

Fractional polynomials of degree one and two are fitted to each of the 500 generated prevalence datasets. For each generated dataset, the logistic models with fractional polynomials as linear predictors are ranked and scaled based on the AIC weights and the model averaged prevalence and force of infection are estimated. Thus, 500 model averaged prevalence and force of infection are estimated. Similarly, for each one of the 500 generated datasets, the "best" fitting fractional polynomial within the set of fractional polynomials is identified and used to estimate the prevalence and force of infection. Therefore, model averaged and "best" single model estimates can be compared with the *true* prevalence and *true* force of

infection to assess the characteristics of the unconditional and model-conditional estimates.

The asymptotic true variance, bias and mean squared error (MSE) can be estimated respectively as $\hat{\nu} = \frac{\sum_{b=1}^{B}(t_b^* - \bar{t}^*)^2}{B-1}$, $\hat{\beta} = \bar{t}^* - t$ and $\widehat{\text{MSE}} = \hat{\nu} + \hat{\beta}^2$; with $b = 1, \ldots, B$, t the actual value of the parameter of interest (the *true* prevalence or the *true* force of infection), t_b^* the estimated parameter of interest from the b generated seroprevalence dataset, and $\bar{t}^* = \frac{\sum_{b=1}^{B} t_b^*}{B}$.

Two estimation methods are compared, estimation using model averaging which uses all fractional polynomial models to account for model selection uncertainty, and estimation using only the best fractional polynomial model (the model with the minimum AIC among the set of fractional polynomials). The two estimation methods are compared on their variance, bias and mean squared error (MSE) for the estimation of the age-dependent prevalence and the force of infection.

The true generating model (the fractional polynomial with powers $p_1 = 1.5$ and $p_2 = 1.3$) was not selected as the best model since the fractional polynomial class does not encompass the power 1.3. The fractional polynomial with powers $p_1 = 2$ and $p_2 = 3$ was selected as the best fitting model in 34% of the B runs. Fractional polynomials with powers $p_1 = 1$ and $p_2 = 3$, and powers $p1 = -2$ and $p2 = 1.5$ appeared as the best fitting models in 17% and 16% of the times, respectively. Although the underlying process is the same, there is considerable variation in model selection across datasets and it shows model selection uncertainty.

Table 2 shows the true age-dependent prevalence calculated from the data-generating model, the sign of the bias, the squared bias (×10000), the variance (×10000) and MSE (×10000) of the estimated prevalence based on the "best" fractional polynomial only and based on model averaging.

Table 2: Simulation results for 500 runs. The true prevalence based on the simulating setting, prevalence based on the "best" fractional polynomial, model averaged prevalence with the fractional polynomials as a set of candidate models. The sign of the bias, the squared bias (×10000), the variance (×10000) and MSE (×10000) over all simulation runs.

	True prevalence	Prevalence based on the "best" fractional polynomial				Model averaged prevalence			
Age		(sign)	Bias^2	Var	MSE	(sign)	Bias^2	Var	MSE
10	0.0592	+	0.0154	0.8082	0.8236	+	0.0134	0.7167	0.7301
20	0.1003	+	0.0815	0.9591	1.0405	+	0.1120	0.7574	0.8694
30	0.2013	+	0.0264	1.6565	1.6829	+	0.0516	1.2328	1.2844
40	0.4110	-	0.0112	2.2208	2.2320	-	0.0209	2.0575	2.0784
50	0.6979	-	0.0015	1.8262	1.8277	-	0.0205	1.5902	1.6107
60	0.8999	+	0.0152	1.0378	1.0531	+	0.0223	0.9849	1.0073
70	0.9760	+	0.0007	0.4050	0.4057	+	0.0146	0.2804	0.2950
80	0.9953	-	0.0031	0.1057	0.1087	+	0.0010	0.0387	0.0397

Comparing the two methods presented in Table 2, the estimates using only the "best" fractional polynomial are less robust as compared with the estimates using model averaging. In general, model averaging prevalence estimates have smaller variance and MSE as compared with the estimated prevalence based only on the

"best" model. This indicates that model averaging is recommended over the use of one single "best" fractional polynomial, since it reduces the MSE and yields more precise estimates for the prevalence.

Table 3 presents the true age-dependent force of infection, the squared bias, the sign of the bias, the variance and MSE for the estimate of the force of infection based on the "best" fractional polynomial only and based on model averaging. Also for the force of infection, model averaging using the full set of fractional polynomials shows consistently superior results as compared with the estimation based on one single best fitting model. These results are comparable to the findings reported by Faes et al. (2007) in the framework of the estimation of a safe level of exposure with a continuous response.

TABLE 3: Simulation results for 500 runs. The true force of infection based on the simulating setting, force of infection based on the best fractional polynomial, model averaged force of infection with the fractional polynomials as set of candidate models. The sign of the bias, the squared bias (×10000), the variance (×10000) and MSE (×10000) over all simulation runs.

	True force of infection	Force of infection based on the "best" fractional polynomial				Model averaged force of infection			
Age		(sign)	Bias^2	Var	MSE	(sign)	Bias^2	Var	MSE
10	0.0025	+	0.0016	0.0088	0.0104	+	0.0021	0.0037	0.0059
20	0.0071	+	0.0000	0.0078	0.0078	+	0.0001	0.0044	0.0045
30	0.0186	-	0.0007	0.0183	0.0190	-	0.0009	0.0154	0.0163
40	0.0457	-	0.0012	0.0715	0.0727	-	0.0040	0.0477	0.0517
50	0.0894	+	0.0072	0.3245	0.3317	+	0.0056	0.2866	0.2922
60	0.1292	+	0.0723	1.9356	2.0079	+	0.1775	1.3926	1.5701
70	0.1542	+	0.1109	8.1142	8.2251	+	0.5805	3.7186	4.2990
80	0.1707	+	0.0630	24.3592	24.4222	+	1.1253	8.1216	9.2469

We also run simulations in the case the fractional polynomial class encompassed the true generating model (using a fractional polynomial with powers $p_1 = 1.5$ and $p_2 = 3$). In this case the true generating model was selected more frequently (44% of runs) and the model averaged prevalence and force of infection estimates were also more robust, in this instance with smaller MSE, smaller bias and smaller variance, as compared with estimations based on a single "best" model.

7. Conclusion and Discussion

Model uncertainty is an issue in applications where model selection, based on observed data, is required. Researchers are frequently confronted with the critical issue of model selection uncertainty when using statistical modeling and often this uncertainty is ignored leading to over-confident inferences. Data analysis for epidemiological parameters of infectious diseases relies frequently on model selection. Small differences from one model to another might not severely affect the estimation of the prevalence, but in the case of the force of infection, a primary epidemiological parameter, small differences in models can largely affect the esti-

mation of such a sensitive parameter. The model selection process itself induces additional variability that is not taken into account when estimation of the parameters is based only on the selected best fitting model. Thus, model selection uncertainty is of concern and model averaging is a simple alternative of estimation that takes all considered models into account by a ranking and scaling based on weights of evidence.

The use of fractional polynomials is a well-defined and flexible methodology for the construction of a set of candidate models. For model averaging it is important to start from a good set of candidate models and the use of fractional polynomials is an attractive alternative. In the R-project, the package *mfp* can be used to fit first and second order fractional polynomials. It is worth noticing that for some applications assessment of goodness-of-fit is potentially useful in deciding how appropriate the chosen model class is.

Model averaging using fractional polynomials to estimate the prevalence and the force of infection is a promising methodology to account for additional variability induced by the model selection process. In our simulation study we show that model averaged age-dependent prevalence and force of infection have desirable features such as smaller variance and smaller mean squared error as compared with the estimated prevalence and force of infection based, as frequently done, only on one selected "best" model. The use of model averaging, based on a large flexible set of predictor models, such as a set of fractional polynomials, is recommended and yields more robust estimates as compared with the use of a single selected "best" model, since outlying estimates are down weighted by the averaging process.

There are some considerations that can be addressed in future studies. One such consideration is about the correlation among the estimates from different models and its impact on the estimation of the variance. This can be studied by a bootstrap simulation to estimate the variance when correlation different from one is considered.

Acknowledgments

The authors gratefully acknowledge the VLIR-UOS scholarships program in Belgium for funding the execution of this project. The authors wish to thank Dr. Kukatharmini Tharmaratnam, Dr. Baerbel Maus and Dr. Amparo Castro, for providing insightful suggestions to an early version of this paper and to the referees for their valuable comments.

[Received: October 2013 — Accepted: November 2014]

References

Agresti, A. (2002), *Categorical data analysis*, 2nd edition, John Wiley & Sons, New York.

Akaike, H. (1974), 'A new look at the statistical identification model', *IEEE transactions on automatic control* **19**, 716–723.

Beutels, M., Damme, P. V. & Aelvoet, W. (1997), 'Prevalence of hepatitis A, B and C in the flemish population', *European Journal of Epidemiology* **13**, 275–280.

Buckland, S., Burnham, K. & Augustin, N. (1997), 'Model selection: An integral part of inference', *Biometrics* **53**, 603–618.

Burnham, K. & Anderson, D. (2002), *Model selection and multi model inference. A practical information-theoretic approach*, 2, Springer, New York.

Castañeda, J. & Gerritse, B. (2010), 'Appraisal of several methods to model time to multiple events per subject: Modelling time to hospitalizations and death', *Revista Colombiana de Estadística* **11**, 43–61.

Faes, C., Aerts, M., Geys, H. & Molenberghs, G. (2007), 'Model averaging using fractional polynomials to estimate a safe level of exposure', *Risk Analysis* **27**(1), 111–123.

Farrington, C. (1990), 'Modeling forces of infection for measles, mumps and rubella', *Statistics in Medicine* **9**, 953–967.

Goeyvaerts, N., Hens, N., Ogunjimi, B., Aerts, M., Shkedy, Z., Damme, P. V. & Beutels, P. (2010), 'Estimating infectious disease parameters from data on social contacts and serological status', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **59**(2), 255–277.

Hens, N., Shkedy, Z., Aerts, M., Faes, C., Van Damme, P. & Beutels, P. (2012), *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data*, 1st edition, Springer.

Hoeting, J., Madigan, D., Raftery, A. & Volinsky, C. (1999), 'Bayesian model averaging: A tutorial', *Statistical Science* **14**(4), 382–401.

Keiding, N. (1991), 'Age-specific incidence and prevalence: A statistical perspective', *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **154**(3), 371–412.

Kullback, S. & Leibler, R. A. (1951), 'On information and sufficiency', *Annals of Mathematical Statistics* **22**(1), 79–86.

Royston, P. & Altman, D. G. (1994), 'Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **43**(3), 429–467.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, P., & Damme, P. V. (2006), 'Modelling age-dependent force of infection from prevalence data using fractional polynomials', *Statistics in Medicine* **25**(9), 1577–1591.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, P. & Damme, P. V. (2003), 'Modelling forces of infection by using monotone local polynomials', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **52**(4), 469–485.