Estimation of Vaccination Coverage Using a Constrained Logistic Model

Shannon McClintock

Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University

Nov. 1, 2011

Presenter Disclosure

No relationships to disclose

OUTLINE

1. Introduction
2. New Approach
3. Simulation
4. Analysis of 2003 Kenya DHS
5. Conclusions

Vaccination Coverage

Vaccination has a dual role:

- Protects an individual from vaccine preventable diseases
- Reduces rates of vaccine preventable diseases in a community

Estimation of coverage is useful for:

- Monitoring and evaluation of vaccination programs
- Determining if the population coverage necessary for disease elimination has been achieved
- Assessing the health services available to children in a community

Motivating Data

Demographic and Health Surveys (DHS)

- Mothers provide vaccination information for all children under the age of 5.
- This can be gathered by
  - child's official vaccination card
  - maternal recall
- The 2003 Kenya DHS reports that 60% of children had vaccination cards available.
- We want to assess the coverage of the the combined diphtheria, pertussis, and tetanus vaccine (DPT) via the 2003 Kenya DHS, which is recommended at 6, 10, and 14 weeks.

Available Methods

- Simple point estimation of proportion vaccinated at specific age intervals
- Survival Analysis:
  - Uses time to vaccination as an outcome
  - Considers children unvaccinated at the time of interview to be right-censored
  - Obtains vaccination coverage by 1 minus the Kaplan-Meier survival function
  - Uses the Cox proportional hazards model to evaluate factors affecting the timeliness of vaccination

◂ Example

Limitations of Available Methods

Limitations of Survival Analysis for Vaccination Data:

- Requires exact data on the date of birth and date of vaccination of the child
  - Some impute date of vaccination if missing
  - Some only analyze data for which date of vaccination is available
  - This can bias the estimate of vaccination coverage
- Does not directly model the vaccination coverage (uses empirical estimates of the tail end of the "inverse" Kaplan-Meier curve, which can be unstable)

New Approach

1. Introduction
2. New Approach
3. Simulation
4. Analysis of 2003 Kenya DHS
5. Conclusions

New Approach

Goal

Develop methods that provide accurate estimates of vaccination coverage with reliable inference.

Our new approach:

- Utilize data on all children who were vaccinated according to either vaccination card or maternal recall
- Use the age at the time of interview (x) and whether or not the child was vaccinated ($y = 0, 1$), as indicated by either vaccination card or maternal recall
- Do not need dates of vaccination

Challenges

Challenges in the new approach:

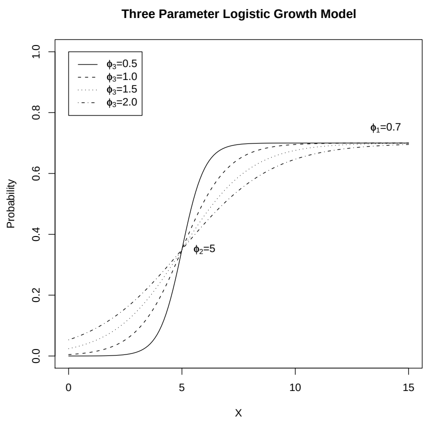
- Binary data is often evaluated through logistic regression, which models the probability of response on the full probability range (0,1) Simulation
- How to estimate a parameter constrained between 0 and 1?
- How to enforce that constraint?

Three Parameter Non-linear Logistic Growth Model

$$M1 : p(x) = \frac{\phi_1}{1 + e^{-(x-\phi_2)/\phi_3}}$$

$$M2 : p(x) = \frac{\frac{1}{1+\exp(-\lambda)}}{1 + e^{-(x-\phi_2)/\phi_3}}$$

- ϕ_1 is the asymptote
- ϕ_2 is the point of inflection
- ϕ_3 is the exponential growth rate parameter (slope)



Examples in literature CI for Model (2)

Methods of Estimation

1. Non-linear least squares NLS
2. Maximum likelihood estimation MLE
   - Nelder-Mead algorithm Nelder-Mead
   - BFGS box-constrained algorithm L-BFGS-S
3. Bayesian estimation Bayesian

Simulation

1 Introduction

2 New Approach

3 Simulation

4 Analysis of 2003 Kenya DHS

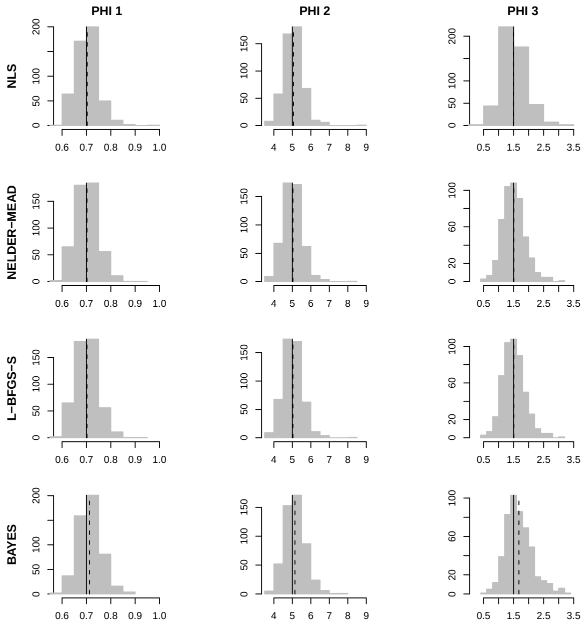
5 Conclusions

Simulation Details

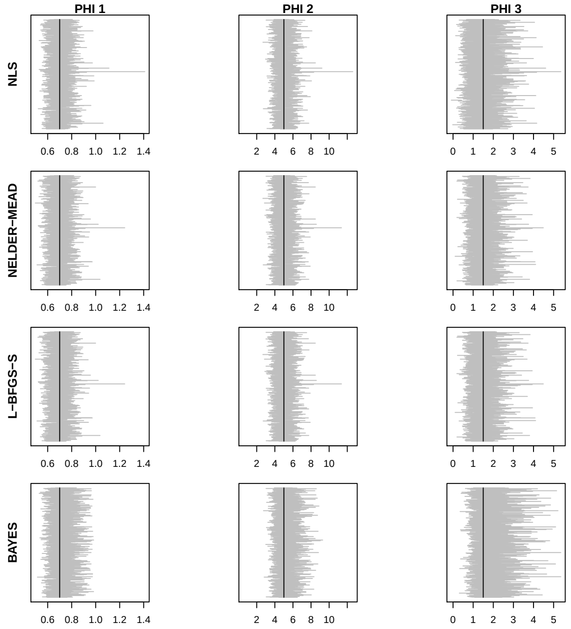
- Compare performance of Models (1) and (2) under different methods of estimation
- True values are ϕ_1=0.70 (λ=0.85), ϕ_2=5.0, and ϕ_3=1.5
- 500 simulations of sample size 350 are generated, where $\mathbf{x} \sim \text{Unif}(0.1, 15)$
- Bayesian estimation was run with 3 chains for 5000 iterations. The first 1000 iterations were discarded for burn-in, and convergence was verified by the Gelman and Rubin statistic $\widehat{R}$

◂ User input

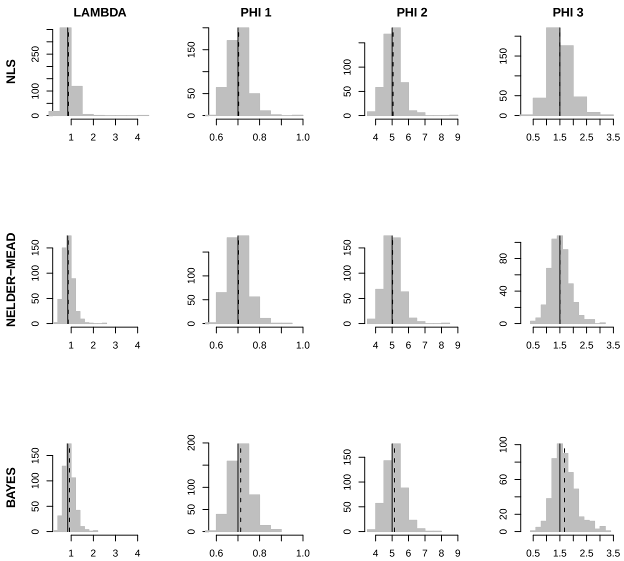
Model 1 Histograms



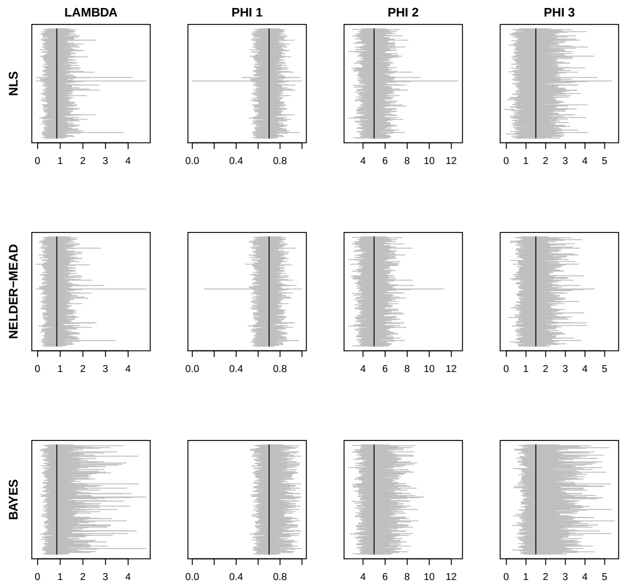
Model 1 CIs



Model 2 Histograms



Model 2 CIs



Simulation Results

	Bias				Coverage				Mean Length			
	λ	ϕ_1	ϕ_2	ϕ_3	λ	ϕ_1	ϕ_2	ϕ_3	λ	ϕ_1	ϕ_2	ϕ_3
Model (1)												
NLS	-	0.003	0.059	-0.005	-	94.4	92.8	94.0	-	0.17	1.89	1.63
NELDER-MEAD	-	0.002	0.032	0.006	-	94.8	94.4	93.2	-	0.18	2.04	1.35
L-BFGS-S	-	0.002	0.032	0.006	-	94.8	94.4	93.0	-	0.18	2.04	1.35
BAYES	-	0.013	0.139	0.175	-	95.4	95.8	91.8	-	0.20	2.39	1.76
Model (2)												
NLS	0.030	0.003	0.059	-0.005	93.4	93.4	92.8	94.0	0.98	0.17	1.89	1.63
NELDER-MEAD	0.022	0.002	0.032	0.006	94.4	94.4	94.4	93.0	0.89	0.18	2.04	1.35
BAYES	0.074	0.013	0.140	0.175	95.0	95.0	95.6	91.4	1.21	0.20	2.40	1.76

◂ Example of erratic simulation

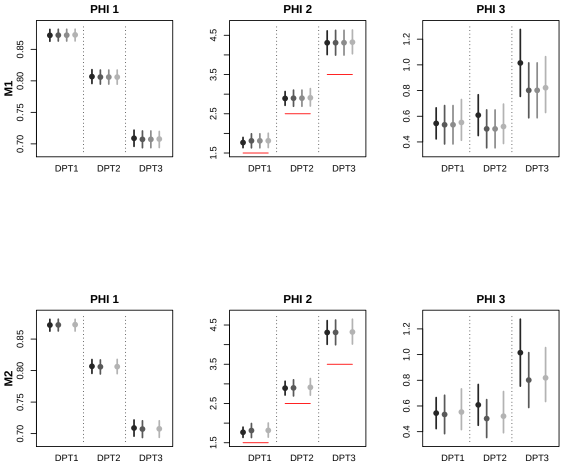
Data Analysis

1. Introduction
2. New Approach
3. Simulation
4. Analysis of 2003 Kenya DHS
5. Conclusions

Summary of DPT Outcomes in 2003 Kenya DHS

		DPT1		DPT2		DPT3	
Entry	Value	N	(%)	N	(%)	N	(%)
No	0	881	(16.2)	1323	(24.4)	1930	(35.6)
Vacc. date on card	1	2580	(47.5)	2396	(44.1)	2157	(39.7)
Vacc. marked on card	1	20	(0.4)	20	(0.4)	18	(0.3)
Reported by mother	1	1949	(35.9)	1689	(31.1)	1323	(24.4)

Analysis of 2003 Kenya DHS



Point estimates and 95% confidence intervals/credible sets for DPT1, DPT2, and DPT3 coverage from the 2003 Kenya DHS. The four lines in decreasing gray scale indicate:

(1) nonlinear least squares,
(2) Nelder-Mead,
(3) L-BFGS-S, and
(4) Bayesian estimates.

L-BFGS-S was not used for Model (2) as it would produce the same results as the Nelder-Mead algorithm.

The red lines on ϕ_2 indicate target vaccination age.

◂ Numeric Results

Conclusions

1. Introduction
2. New Approach
3. Simulation
4. Analysis of 2003 Kenya DHS
5. Conclusions

Conclusions

- Model (2) appropriately constrains the numerator, but may be unstable in certain data configurations
- NLS not be robust to all situations
- Bayesian framework is attractive:
  - naturally restrict parameter estimates through prior distributions
  - inference does not depend on asymptotic rates of convergence
  - stability in the infrequent but not entirely rare data settings yielding unstable MLEs

Conclusions

- The nonlinear logistic model can be used to estimate an asymptote less than 1 when the outcome of interest is binary
- We used this model to estimate vaccination coverage, which also estimates two other meaningful parameters in this context
- This model is most applicable to vaccination research in which respondents are unable to estimate age at the time of vaccination
- This model enables researchers to base inference regarding vaccination coverage on all respondents regardless of whether or not they retained their vaccination cards, hereby eliminating possible bias due to only analyzing complete data cases

Future Work for Logistic Growth Model

- Explore both model-based and design-based approaches to account for the survey design in the analysis
- Accommodate effects of other covariates ◂ Details
- Investigate the effect of study design on parameter estimation
- Investigate sensitivity of the analysis to starting points for estimation algorithms and prior distributions for ϕ
- Investigate behavior of the analysis with regards to the true value of ϕ

Thanks

Thank you

Dr. Lance Waller
Dr. Andrew Hill
Dr. Qi Long
Dr. Matthew Strickland
Dr. Rick Rheingans
Measure DHS at ICF Macro
Environmental Biostatistics Training Grant

Laubereau et al., 2002

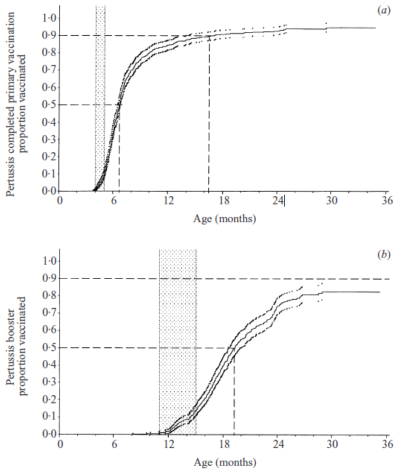


Fig. 1 (*a*) Pertussis vaccination in Germany (completed primary) in 782 children aged 0–3 years. Inverse Kaplan–Meier curves (1 – s(t)) with 95 % confidence interval. The shaded area marks the nationally recommended age-periods for vaccination (3rd to 5th month of life). (*b*) Pertussis vaccination in Germany (completed primary + booster) in 782 children aged 0–3 years. Inverse Kaplan–Meier curves (1 – s(t)) with 95 % confidence interval. The shaded area marks the nationally recommended age-periods for vaccination (12th–15th month of life).

◂ Back

Non-linear functions

- Non-linear functions can be used to estimate the upper bound for unknown quantities
- Applications include:
  - **Ecologic population growth model**
    Pearl and Reed (1920) estimate the carrying capacity of the United States human population
  - **Bioassay (quantal or quantitative)**
    Rodbard and Frazier (1975) estimate antigen counts in radioimmunoassay

◂ Back

CI for λ in Model (2)

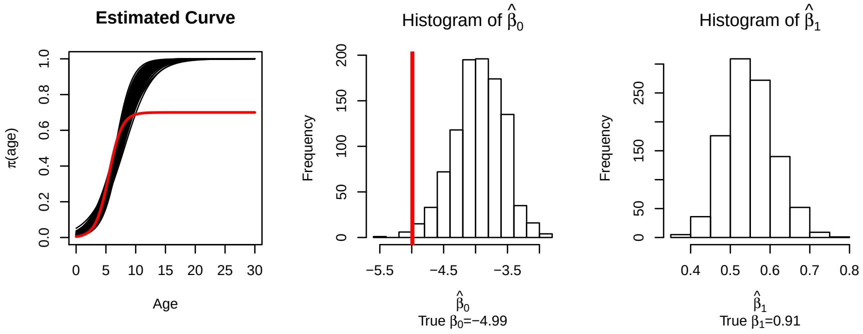
From Model (2), $\hat{\phi}_1 = \frac{1}{1+\exp(-\hat{\lambda})}$.

Asymptotic confidence intervals for ϕ_1 can be created by first calculating asymptotic confidence intervals for λ via $\hat{\lambda} \pm z_{1-\frac{\alpha}{2}} * SE(\hat{\lambda})$ resulting in the interval $(\hat{\lambda}_L, \hat{\lambda}_U)$.

Then apply the transformation $\left(\frac{1}{1+\exp(-\hat{\lambda}_L)}, \frac{1}{1+\exp(-\hat{\lambda}_U)}\right)$ to create a confidence interval for ϕ_1.

◂ Back

Constrained Simulation Results



◂ Back

Non-linear Least Squares

For Model 1, NLS minimizes $\sum_i \left(Y_i - \frac{\phi_1}{1+\exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right)} \right)^2$

- Even though Y_i are not normally distributed in our application, nonlinear least-squares estimates are consistent as long as Models (1) and (2) are correctly specified
- Can be calculated by the `nls` function in R, which uses the Gauss-Newton algorithm
- No inherent upper bound on the the estimate of ϕ_1
- In our application, ϕ_1 should never exceed one

◂ Back

Maximum Likelihood Estimation

The outcome is binary, following the form

$$y_i \sim Bern(p_i), \quad p_i = \frac{\phi_1}{1 + e^{\frac{-(x_i-\phi_2)}{\phi_3}}}$$

with likelihood given by

$$\begin{aligned}
\log \mathrm{L}(\theta) &= \sum_{i=1}^{n} \log \left[p_i^{y_i} (1-p_i)^{1-y_i} \right] \\
&= \sum_{i=1}^{n} \log \left[\left(\frac{\phi_1}{1+\exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right)} \right)^{y_i} \left(1 - \frac{\phi_1}{1+\exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right)} \right)^{1-y_i} \right] \\
&= \sum_{i=1}^{n} y_i \log \phi_1 - \log \left\{ 1 + \exp\left(-\frac{(x_i-\phi_2)}{\phi_3} \right) \right\} \\
&\quad + (1-y_i) \log \left\{ 1 + \exp\left(-\frac{(x_i-\phi_2)}{\phi_3} \right) - \phi_1 \right\}
\end{aligned}$$

◂ Back

Maximum Likelihood Estimation

Let $\Phi = (\phi_1, \phi_2, \phi_3)$, $a_i = \dfrac{-(x_i - \phi_2)}{\phi_3}$, $b_i = 1 + e^{a_i} - \phi_1$, and $c_i = 1 + e^{a_i}$.

$$I_n(\mathbf{\Phi}) = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^{n} \frac{1}{\phi_1 b_i} & \sum_{i=1}^{n} -\frac{e^{a_i}}{\phi_3 b_i c_i} & \sum_{i=1}^{n} \frac{a_i e^{a_i}}{\phi_3 b_i c_i} \\ & \sum_{i=1}^{n} \frac{\phi_1 e^{2a_i}}{\phi_3^2 b_i c_i^2} & \sum_{i=1}^{n} -\frac{\phi_1 a_i e^{2a_i}}{\phi_3^2 b_i c_i^2} \\ & & \sum_{i=1}^{n} \frac{\phi_1 a_i^2 e^{2a_i}}{\phi_3^2 b_i c_i^2} \end{pmatrix}$$

Asymptotic distribution of the three parameters in the logistic growth model:

$$\sqrt{n}(\hat{\Phi}_{MLE} - \Phi) \rightarrow_d N_3 \left(0, \{nI_n(\Phi)\}^{-1}\right)$$

◂ Back

MLE: Nelder-Mead

- Derivative-free minimization algorithm that can be used to estimate parameters from the negative log-likelihood
- Estimates n parameters by forming an n-dimensional simplex using $n + 1$ points
- Does not implicitly yield variance-covariance estimates of the parameters, though they can be estimated by the diagonal of the inverse of the Hessian matrix
- Default optimization algorithm in the `optim` function in R

◂ Back

MLE: BFGS Box-constrained

- Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm
- Quasi-Newton method that uses function values and gradients to build up a picture of the surface to be optimized
- Can be modified to incorporate box constraints on parameter estimates, known as L-BFGS-B algorithm
- Can forcibly constrain $\hat{\phi}_1 \in (0, 1)$
- Must specify constraints for other parameters in the model as well
- Available in the `optim` function in R.

◂ Back

Bayesian Estimation

- Seek inference on the distribution of parameter estimates, given the data $\mathbf{y}$: $p(\mathbf{\Phi}|\mathbf{y}) \propto p(\mathbf{\Phi})p(\mathbf{y}|\mathbf{\Phi})$
- In our application,

$$p(\mathbf{y}|\mathbf{\Phi}) = \prod_{i=1}^{n} \log\left[\xi_i^{y_i}(1-\xi_i)^{1-y_i}\right]$$

$$\xi_i = \frac{\phi_1}{1 + e^{\frac{-(x_i-\phi_2)}{\phi_3}}}$$

  and $p(\mathbf{\Phi})$ is given by the density of the prior distributions of the parameters $\mathbf{\Phi}$
- Can use the prior distribution of Φ to coerce parameter estimates to adhere to their logical constraints
- Parameter estimates are obtained by Markov Chain Monte Carlo techniques using the `bugs` function in the `R2WinBUGS` package in R, which calls WinBUGS 1.4

◂ Back

User input

Box-constraints in the BFGS algorithm:

- $0.01 \leq \phi_1 \leq 0.99$
- $0.10 \leq \phi_2 \leq 100$
- $0.10 \leq \phi_3 \leq 100$

The prior distributions for the Bayesian simulation:

- $\phi_1 \sim \text{Unif}(0.01, 0.99)$
- $\phi_2 \sim \text{Unif}(0.1, 20)$
- $\phi_3 \sim \text{Unif}(0.1, 7)$
- $\lambda \sim$ standard logistic (corresponds to a uniform distribution for ϕ_1)

◂ Back

Example of unstable MLEs

Point estimates and 95% confidence intervals/credible sets for one simulation

Model (1)	$\hat{\lambda}$		$\hat{\phi}_1$		$\hat{\phi}_2$		$\hat{\phi}_3$	
NLS	-	-	0.99	(0.56, 1.41)	8.89	(5.16, 12.62)	3.45	(1.54, 5.36)
NELDER-MEAD	-	-	0.93	(0.61, 1.24)	8.39	(5.41, 11.36)	3.04	(1.59, 4.49)
L-BFGS-S	-	-	0.93	(0.61, 1.24)	8.39	(5.41, 11.36)	3.04	(1.59, 4.49)
BAYES	-	-	0.89	(0.70, 0.98)	7.99	(6.07, 9.29)	2.94	(1.98, 4.07)
Model (2)								
NLS	4.29	(-27.31, 35.89)	0.99	(0.00, 1.00)	8.89	(5.16, 12.62)	3.45	(1.54, 5.36)
NELDER-MEAD	2.54	(-2.10, 7.19)	0.93	(0.11, 1.00)	8.39	(5.42, 11.35)	3.04	(1.60, 4.48)
BAYES	2.05	(0.84, 4.78)	0.89	(0.70, 0.99)	7.98	(6.14, 9.50)	2.96	(2.00, 4.08)

◂ Back to simulation results

Results from Analysis of 2003 Kenya DHS

	$\hat{\lambda}$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$
Model 1, DPT1				
NLS	-	0.8723 (0.8629,0.8817)	1.7666 (1.6356,1.8975)	0.5445 (0.4236,0.6654)
NELDER-MEAD	-	0.8725 (0.8632,0.8819)	1.8109 (1.6320,1.9897)	0.5337 (0.3842,0.6832)
L-BFGS-S	-	0.8725 (0.8632,0.8819)	1.8111 (1.6323,1.9900)	0.5338 (0.3843,0.6832)
BAYES	-	0.8730 (0.8633,0.8820)	1.8125 (1.6389,2.0040)	0.5518 (0.4132,0.7304)
Model 2, DPT1				
NLS	1.9212 (1.8370,2.0054)	0.8723 (0.8626,0.8814)	1.7666 (1.6356,1.8975)	0.5445 (0.4236,0.6654)
NELDER-MEAD	1.9233 (1.8393,2.0072)	0.8725 (0.8629,0.8816)	1.8114 (1.6325,1.9904)	0.5339 (0.3844,0.6835)
BAYES	1.9270 (1.8380,2.0055)	0.8729 (0.8627,0.8814)	1.8150 (1.6425,2.0030)	0.5531 (0.4159,0.7330)
Model 1, DPT2				
NLS	-	0.8068 (0.7958,0.8178)	2.8911 (2.7152,3.0670)	0.6086 (0.4497,0.7675)
NELDER-MEAD	-	0.8060 (0.7949,0.8172)	2.8971 (2.6913,3.1028)	0.5019 (0.3543,0.6494)
L-BFGS-S	-	0.8060 (0.7949,0.8172)	2.8974 (2.6916,3.1031)	0.5016 (0.3542,0.6491)
BAYES	-	0.8060 (0.7947,0.8172)	2.9080 (2.6969,3.1416)	0.5200 (0.3877,0.6947)
Model 2, DPT2				
NLS	1.4292 (1.3587,1.4998)	0.8068 (0.7955,0.8175)	2.8911 (2.7152,3.0670)	0.6086 (0.4497,0.7675)
NELDER-MEAD	1.4244 (1.3531,1.4958)	0.8060 (0.7946,0.8169)	2.8980 (2.6921,3.1038)	0.5019 (0.3543,0.6496)
BAYES	1.4260 (1.3560,1.5010)	0.8063 (0.7951,0.8178)	2.9125 (2.7125,3.1360)	0.5206 (0.3916,0.7125)
Model 1, DPT3				
NLS	-	0.7089 (0.6961,0.7218)	4.3102 (4.0079,4.6125)	1.0150 (0.7544,1.2755)
NELDER-MEAD	-	0.7072 (0.6939,0.7205)	4.3106 (3.9952,4.6260)	0.8017 (0.5880,1.0154)
L-BFGS-S	-	0.7072 (0.6939,0.7205)	4.3115 (3.9960,4.6270)	0.8018 (0.5881,1.0155)
BAYES	-	0.7077 (0.6942,0.7198)	4.3250 (4.0289,4.6350)	0.8218 (0.6291,1.0640)
Model 2, DPT3				
NLS	0.8901 (0.8278,0.9525)	0.7089 (0.6959,0.7216)	4.3102 (4.0079,4.6125)	1.0150 (0.7544,1.2755)
NELDER-MEAD	0.8819 (0.8177,0.9461)	0.7072 (0.6938,0.7203)	4.3120 (3.9965,4.6275)	0.8016 (0.5880,1.0153)
BAYES	0.8839 (0.8191,0.9458)	0.7076 (0.6940,0.7203)	4.3230 (4.0150,4.6485)	0.8189 (0.6346,1.0545)

◂ Back to figure

Accommodate effects of other covariates

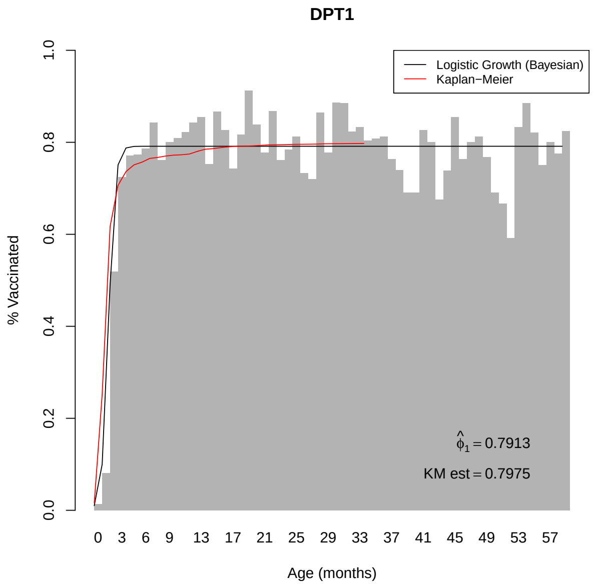
Covariates can be easily included in the model to affect the asymptote, inflection point, or slope. For example, if rural areas are thought to have a lower probability of vaccination than urban, then we could model the probability of vaccination as

$$f(x) = \frac{\phi_1 + \gamma x_{rural}}{1 + \exp[-(x_{age} - \phi_2)/\phi_3]}$$

where x_{rural} is an indicator and the parameter γ represents the increase or the decrease in the vaccination coverage for rural areas compared to urban.

◂ Back

Comparison of logistic growth curve to survival analysis



◂ Back