Limburgs Universitair Centrum

Center for Statistics

Modelling the Herd-specific Force Of Infection for BoHV-1
The Influence of Missing Data

Niel Hens[1], Christel Faes[1], Marc Aerts[1], Ziv Shkedy[1], Helena Geys[2], Koen Mintiens[3], Hans Laevens[3] and Frank Boelaert[4].

contact: niel.hens@luc.ac.be; christel.faes@luc.ac.be; marc.aerts@luc.ac.be

[1]Center For Statistics, Limburgs Universitair Centrum, Universitaire Campus, 3590 Diepenbeek, Belgium

[2]Johnson and Johnson, PRD Biometrics and Reporting, Belgium; [3]Veterinary and Agrochemical Research Centre, Belgium; [4]European Food Safety Authority, Italy

Abstract

Modelling infectious diseases often has to deal with missing values in the data. If the missingness is ignorable as defined by Little and Rubin (1987), the analysis can be based on the complete cases only. If however the missingness is non-ignorable, analyses can be affected by merely using the complete cases. This poster shows the effect of ignoring missing data to model the force of infection of the bovine herpesvirus-1 in Belgian Cattle. We propose the use of weighted generalized estimating equations to deal with both clustering and missingness in the data.

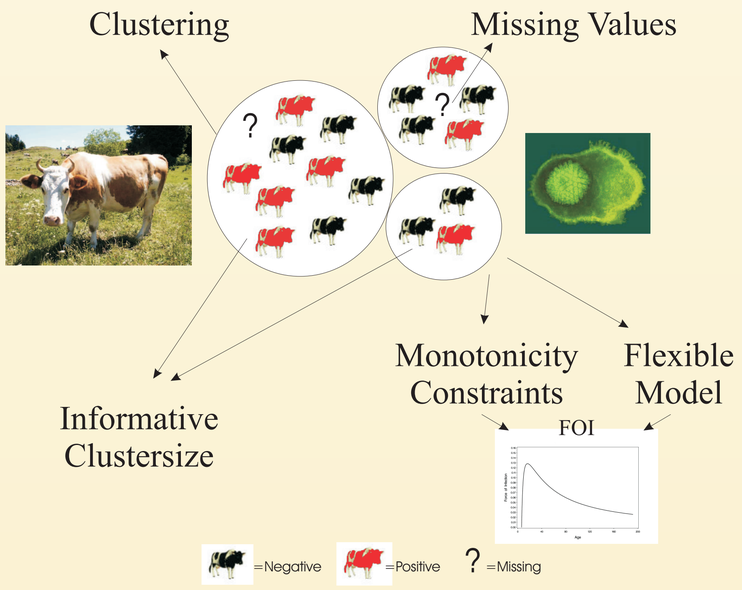
Bovine HerpesVirus-1 (BoHV-1)

- transmissible disease
- serological survey 1997-1998 in Belgium
- province-stratified random sample
- 11284 cows in 309 herds
- ELISA-test for BoHV-1 glycoprotein B
- restriction to 10363 cases for which age $>$ 6 months
- interest in age-specific sero-prevalence

$$\pi(a) = P(gB = 1|a)$$

- interest in age-specific force of infection (FOI), which is the rate at which susceptible animals get infected

$$\ell(a) = \frac{\pi'(a)}{1 - \pi(a)}$$



Missingness in BoHV-1

- 2148 missing values
- *Complete Cases:* Cows for which all variables are observed (=9136 cases)
- *Available Cases:* Cows for which the test result, age and herdsize are observed (=10363 cases)

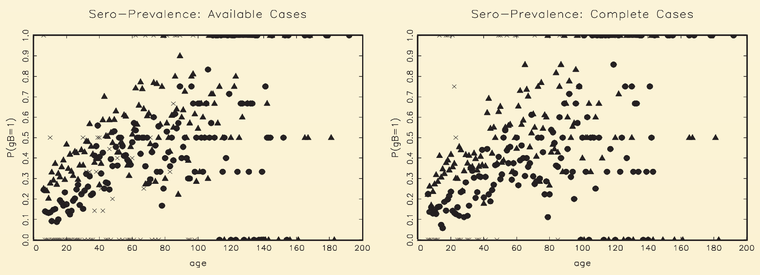


Figure 1: Sero-prevalence plot as a function of age based on the available cases (left) and on the complete cases (right) for small (circles), medium (stars) and large herds (triangles).

Analyses

- Generalized Estimating Equations
  (Clustering; Liang and Zeger, 1986)
- Fractional Polynomial of Age
  (Flexibility; Royston and Altman, 1994)
- Herdsize as a covariate
  (Informative Clustersize; Faes *et al.*, 2005)
- Constrained Optimization
  (Positive Force of Infection)

$$\text{logit}(P(gB = 1)) = \beta_0 + \beta_1 \text{age}^{p_1} + \beta_2 \text{age}^{p_2} + \beta_3 \text{herdsize},$$

with $p_1, p_2 \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}, p_1 \leq p_2$.

- Best Fractional Polynomial Selected Using Deviance-criterion (Royston and Altman, 1994)
- Methods:
  - Complete Cases (CC)
  - Available Cases (AC)
  - Weighted Complete Cases (WCC)

Complete Cases (long dashed line)

- Generalized Estimating Equations
- complete cases only
- best fractional polynomial: $(p_1, p_2) = (-1, -0.5)$

Available Cases (full line)

- Generalized Estimating Equations
- available cases
- best fractional polynomial: $(p_1, p_2) = (-2, -1)$

Weighted Complete Cases (short dashed line)

- Inverse Probability Weighted Generalized Estimating Equations
- complete cases
- animal specific weights $= 1/\hat{\pi}$
- $\hat{\pi}$ = estimated probability for an animal to be observed
- animals unlikely to be observed gain more influence
- implicit imputation of missing values
- π estimated using a generalized additive model (Wood and Augustin, 2002)
- best fractional polynomial: $(p_1, p_2) = (-2, -0.5)$

Results

- age of maximal FOI

	herdsize	15	45	80	120
CC	$\text{age}_{\max}$(years)	3.32	3.09	2.85	2.60
AC	$\text{age}_{\max}$(years)	1.91	1.86	1.80	1.73
WCC	$\text{age}_{\max}$(years)	1.72	1.67	1.62	1.56

- sero-prevalence and force of infection

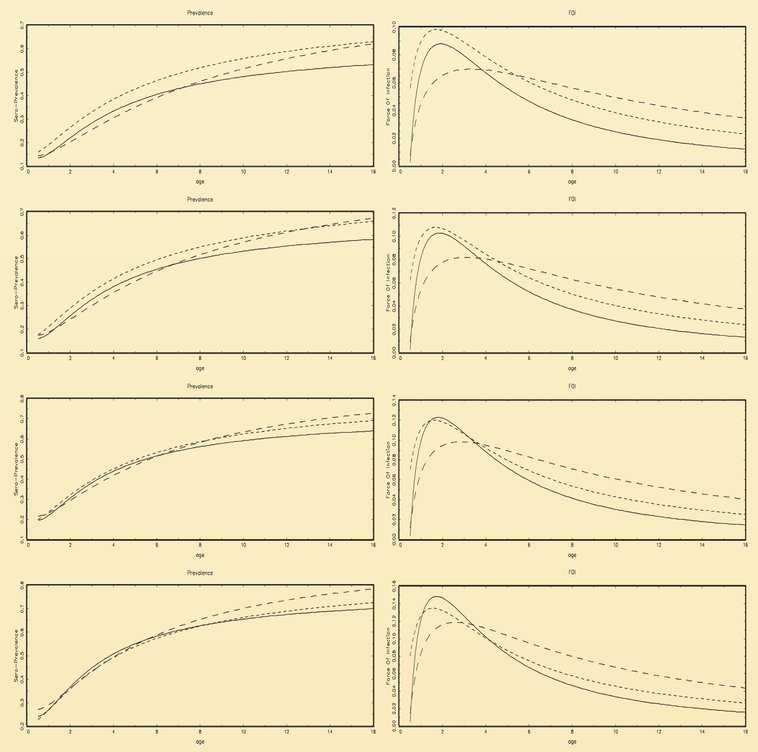


Figure 2: Age-specific sero-prevalence fits together with the age-specific force of infection for the available cases (full line), the complete cases (long dashed line) and weighted complete cases (short dashed line) for herdsizes 15, 45, 80 and 120.

Conclusion

The BoHV-1 data were analyzed to determine the FOI. It is clear from the presented results that the dataset suffers from several complications. To overcome the complication of some values to be missing an inverse probability weighted analysis is proposed. The age of maximal FOI is well estimated based on this weighted analysis. The use of complete cases gives a large overestimation. Merely using the complete cases results in wrong conclusions !