ESTADÍSTICA (2011), 63, 181, pp. 95-112

STATISTICAL METHODS TO QUANTIFY THE EFFECT OF MITE PARASITISM ON THE PROBABILITY OF DEATH IN HONEY BEE COLONIES

CHELSEA DEROCHE
Department of Experimental Statistics, Louisiana State University
cderoc2@lsu.edu

JOSÉ D. VILLA
USDA-ARS Honey Bee Breeding, Genetics and Physiology Laboratory
Jose.Villa@ars.usda.gov

LUIS A. ESCOBAR
Department of Experimental Statistics, Louisiana State University
luis@lsu.edu

ABSTRACT

Varroa destructor is a mite parasite of European honey bees, *Apis mellifera*, that weakens the population, can lead to the death of an entire honey bee colony, and is believed to be the parasite with the most economic impact on beekeeping. The purpose of this study was to estimate the probability of death for a honey bee colony as a function of the mite-infestation level present in the colony and to explore the influence of other variables (such as genetic origin of the colony and season of the year) on the relationship. Preliminary analyses showed that there was an association between season and mite infestation that needed to be considered in later analyses. Two analytical approaches were considered to account for the lack of deaths in colonies from two genetic origins which led to divergence of the maximum likelihood method when including origin as one of the variables in the logistic regression model. In the first approach, we used Firth's penalized likelihood method which has the double effect of correcting the bias of maximum likelihood (ML) estimates and providing estimates of the parameters. The second approach consists of forcing a death at the largest mite infestation for each of the two genetic origins without deaths. This approach, in general, would tend to provide slightly larger colony-death probability estimates. Because there were multiple observations on the same colony over a period of time, the data are longitudinal and the observations may not be independent. For this reason, we used a Generalized Estimating Equations (GEE) approach, which considers the dependency among the observations and compared it with the simple logistic regression that ignores the dependency. The GEE analysis showed increasing odds of death with increasing mite infestation and found no influence of season or

genetic origin on the relationship. The results of the analysis using simple logistic regression are similar to those obtained using the more complex GEE analysis, suggesting that, for the data set considered, the longitudinal observations can be treated as statistically independent.

Key words

Generalized estimating equations; logistic regression; *Varroa destructor*.

RESUMEN

El acaro *Varroa destructor* es un parásito de las abejas europeas (*Apis mellifera*) que reduce el tamaño de la población de abejas, puede llevar a la muerte de colonias y es considerado como el parásito de mayor impacto económico sobre la apicultura. El objetivo de este estudio fue estimar la probabilidad de muerte de las colonias en función del nivel de infestación presente en la colonia e investigar la influencia de otras variables sobre esta relación (como el origen genético de la colonia y la estación del año). Análisis preliminares mostraron una asociación entre la estación del año y la infestación, requiriendo ser considerada en los análisis posteriores. Debido a la ausencia de muertes en colonias de dos orígenes genéticos lo cual produjo divergencia en el modelo de estimación usando máxima verosimilitud cuando se incluyó la variable origen genético en la regresión logística, se utilizaron dos metodologías alternativas. La primera, fue el uso del método penalizado de Firth el cual produce el doble efecto de corregir el sesgo de los estimadores de máxima verosimilitud y el producir estimadores de los parámetros. La segunda metodología, consistió en considerar como muertas aquellas colonias con el nivel máximo de infestación observado en los dos origenes sin mortalidad. Este método tiende a producir probabilidades de muerte más altas. Debido al uso de observaciones múltiples sobre el tiempo para cada colonia, las observaciones son longitudinales y dependientes. Por esto, se utilizó el método de ecuaciones de estimación generalizadas (GEE), el cual incorpora la dependencia entre observaciones y permite comparaciones con funciones de regresión logística que ignoran la dependencia. El método de las GEE indica probabilidades de muerte mas altas con aumentos en el nivel de infestación, pero los efecto de origen genético y de estación del año no fueron estadísticamente importantes. Los resultados con regresión logística fueron similares a los obtenidos con el método de GEE, sugiriendo, que para este grupo de datos, las observaciones longitudinales pueden ser consideradas como independientes.

Palabras clave

Ecuaciones generalizadas de estimación; regresión logistica; *Varroa destructor*.

1. Introduction

Varroa destructor is an obligate mite parasite of honey bees with particularly devastating effects on colonies of the western honey bee (*Apis mellifera*). These mites

co-evolved with Asian honey bees, primarily *Apis cerana*, leading to a stable host-parasite relationship in which both survive (Oldroyd 1999). In contrast, a number of introductions of the parasite unto *Apis mellifera* the last four decades have led to problems in Europe, the Americas, Africa, and some Caribbean and Pacific Islands. Mortality of most unmanaged native or feral *Apis mellifera* colonies increases dramatically after these introductions (Kraus and Page 1995, Fries et al. 2006, Villa et al. 2008). As an example, a feral undesirable population of colonies on the nature preserve of Santa Cruz Island, California was eradicated with the purposeful introduction of mites (Wenner et al. 2009). To avert mortality and weakening of colonies, beekeepers rely on scheduled repeated treatments with acaricides. Poor scheduling or ineffectiveness of treatments can lead to highly infested colonies which tend to have the highest mortalities (e.g. Genersch et al. 2010, Guzman–Novoa et al. 2010). Despite obvious mortality in western honey bees, there is a paucity of statistical analyses to quantify the strength of the relationship between infestation level and probability of colony mortality. Likewise, the influence of seasonal, regional and stock differences upon the effects of mites on colony mortality is poorly understood.

We conducted a longitudinal study in which the infestation with *V. destructor* and the fate of honey bee colonies were monitored through time. The populations of bees and their parasites in colonies were allowed to fluctuate without providing treatment for mite control or routine management other than supplying adequate hive volumes for colonies. These observations produced a data set in which the survival of colonies could be compared to the level of infestation at an earlier time. The analytical approaches applied to this data set provide a method of quantifying risk of colony death in relation to the density of parasitic mites as well as estimating the variability of this relationship. Such knowledge can be applied in designing programs for integrated pest management and in setting goals in bee breeding for economically useful levels of bee resistance to the mites. Additionally, the analytical approaches presented here could be applied to other data sets from different beekeeping and geographic situations.

The information collected in this study fits the structure of longitudinal data with binary data corresponding to the dead or alive condition of a colony at each measurement time. The structure of the data is complicated because it represents longitudinal observations. The data are unbalanced because colonies are observed different number of times. Additionally, the number of colonies being observed at any one time is relatively small, and the number of repeated measures per colony is small. One possible analytical approach for this type of study is to model the data using a Markov chain where the conditional probability of success is a function of the experimental variables and previous responses (Bonney 1987). As indicated by Fitzmaurice and Lipsitz (1995), this approach is not appropriate, in this case, when the primary interests are the regression parameters for the expectation. Another model approach is the generalized estimating equations model as developed in Lian and Zeger (1986), Prentice (1988), Lipsitz, Laird, and Harrington (1991), Carey, Zeger, and Diggle (1993), and well described in Fitzmaurice, Laird, and Ware (2004). We compared the use of a simple logistic

regression approach in which the dependency among the observations are ignored with a GEE approach which models the association between a pair of binary responses in terms of their correlation.

The Data

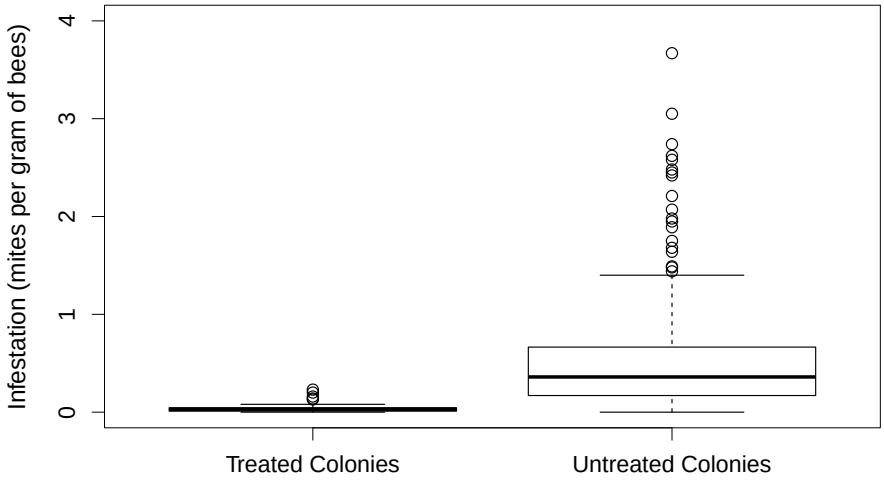
The data were obtained from a longitudinal study consisting of the monitoring and sampling of colonies maintained in up to 4 apiaries between the Winter of 1999–2000 and Spring of 2007. Most colonies ($k = 55$) were left untreated to allow the levels of *V. destructor* to develop. As colonies died or were removed from observation (censored), others were moved into the same apiaries to be monitored and sampled. On average, 16 untreated colonies (range 8–34) were monitored each year. Four additional colonies were treated yearly with an acaricide (and survived) from 2000 to 2004 to verify that untreated colonies were developing much higher mite infestations (see Figure 1). The data from these four colonies, however, were not used in assessing the relationship between mite infestation and survival times in this research.

Sample collections and observations of the status of each colony were made approximately every three months, coinciding with four locally relevant colony condition/growth periods: Spring (March to May), Summer (June to August), Fall (September to November) and Winter (December to February). Samples of approximately 150g of adult workers were taken by shaking at least two frames with bees covering developing brood or from the center of the colony into an empty box and then transferring the approximate amount of bees into a collection jar. The sample was weighed and then soaked at least overnight in ethanol (70% by volume). Mites were removed by shaking the sample and washing them off the bees with more alcohol solution through a mesh screen. Samples were washed repeated times until no additional mites were found. Number of mites and total weight of bees were converted to mites per gram of bees. The database of observations used for analysis included a colony number, genetic origin (Origin), year, season (Season), a measurement of infestation (Infestation), measured in mites per gram of bees, and the status (Status) of the colony (Dead or Alive) at the next observation. Colonies that were not observed further were categorized as censored. Figure 1 is a box plot of the variable Infestation for the 331 and 35 observations from the untreated and treated colonies, respectively.

2. Methods

A primary interest in this study is to assess the effect of the continuous variable Infestation and some categorical variables (Origin and Season) on the survival of colonies. In particular, an important objective is to identify a relationship between probability of death and the observed experimental variables. For this purpose, logistic regression is a natural method to model the data. An alternative to the logistic regression model is the method of generalized estimating equations.

Figure 1. Levels of *Varroa destructor* in Untreated ($k = 55$) and Treated ($k_1 = 4$) Colonies Sampled at Different Times of the Year from 1999–2007



As discussed earlier, there are other possible models for this type of data, but in view of the size of the study and the structure of the data we did not attempt to use the alternative models. As we discuss later, the methods used arrive at similar conclusions and these approaches are useful in answering biological questions.

2.1 Logistic Regression

Logistic regression models are used to predict the probability of an event that has a binary outcome (Dead or Alive) and to explore the relationship between the response and the explanatory variables. The likelihood function for the n_i observations from the ith colony is given by

$$L(\boldsymbol{\beta}|\boldsymbol{Y}_i) = \prod_{j=1}^{n_i} [\pi(\boldsymbol{x}_{ij})]^{y_{ij}} [1 - \pi(\boldsymbol{x}_{ij})]^{1-y_{ij}}$$

where the y_{ij} are Bernoulli distributed with $\Pr(y_{ij} = 1) = \pi(\boldsymbol{x}_{ij})$, y_{ij} takes a value of 1 (for Dead) or 0 (for Alive), $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{in_i})'$, and $\pi(\boldsymbol{x}_{ij})$ represent the probability of dead as a function of the explanatory variables levels

$\boldsymbol{x}_{ij} = (1, x_{ij1}, \ldots, x_{ijp})'$ (i.e., Infestation, Origin, Season) associated with the observation y_{ij} for colony i and time j. The logistic regression model for the predicted probability of death is

$$\pi(\boldsymbol{x}_{ij}) = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})}{1 + \exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})} \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is a vector of $p + 1$ unknown parameters. When appropriated, we use the simplified notation $\pi_{ij} = \pi(\boldsymbol{x}_{ij})$. From model (1), the odds of death are given by

$$\gamma(\boldsymbol{x}_{ij}) = \left[\frac{\pi(\boldsymbol{x}_{ij})}{1 - \pi(\boldsymbol{x}_{ij})}\right], \tag{2}$$

or equivalently

$$\log\left[\gamma(\boldsymbol{x}_{ij})\right] = \boldsymbol{\beta}'\boldsymbol{x}_{ij} = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp}. \tag{3}$$

When there is a single explanatory variable x in the model, $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, $\boldsymbol{x}_{ij} = (1, x_{ij1})'$, and $\boldsymbol{\beta}'\boldsymbol{x}_{ij} = \beta_0 + \beta_1 x_{ij1}$, where x_{ij1} is the level of x for the ith colony at time j. In this case, β_1 is the difference of log-odds due to an increase of 1 unit in x. That is

$$\begin{aligned} \beta_1 &= \log[\gamma(\boldsymbol{x}_{ij}^{+})] - \log[\gamma(\boldsymbol{x}_{ij})] \\ &= \beta_0 + \beta_1(x_{ij1} + 1) - \beta_0 - \beta_1 x_{ij1}, \end{aligned} \tag{4}$$

where $\boldsymbol{x}_{ij} = (1, x_{ij1})'$, $\boldsymbol{x}_{ij}^{+} = (1, x_{ij1} + 1)'$, and x_{ij1} is any arbitrary level of the explanatory variable x. For the general case of two or more explanatory variables, the interpretation of the kth regression coefficient β_k is similar to (4); that is, β_k is the difference in log-odds when the kth explanatory variables x_{ijk} increases by 1 unit and all the other explanatory variables are held at the same level.

The total likelihood, $L(\boldsymbol{\beta})$, is obtained as the product of the $L(\boldsymbol{\beta}, \boldsymbol{Y}_i)$ in the study. That is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} L(\boldsymbol{\beta}|\boldsymbol{Y}_i) \tag{5}$$

where k is the number of colonies in the study.

Data Patterns that Yield Non-convergent Likelihoods

In the data, all of the colonies from two of the genetic origins had no deaths during the observation period. This complicates the analysis when the categorical

variable Origin is included in the model through the explanatory variables $\boldsymbol{x}_{ij}$ because the data structure takes a quasi-complete separation pattern which implies that the ML estimates of the parameters do not exist (Albert and Anderson, 1984 and Santner and Duffy, 1986). There are alternative approaches to recover the estimability of the parameters. We now describe two approaches that allow the inclusion of the variable Origin in the model.

1. **Penalized maximum likelihood:** Firth (1993) proposed the penalized ML estimation as a method for bias reduction. For the binomial logistic regression problem considered in Section 2.1 above, Firth's method provides finite and unique ML parameter estimates, see Firth (1993, Section 3.3). Firth's method replaces the usual gradient equation

$$g(\beta_r) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \pi_{ij}) x_{ijr} = 0, \quad r = 0, 1, \ldots, p$$

with the following modified gradient equation

$$g^*(\beta_r) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[y_{ij} - \pi_{ij} + (0.5 - \pi_{ij}) h_{ij} \right] x_{ijr} = 0, \quad r = 0, 1, \ldots, p$$

where the h_{ij}'s are the diagonal elements of the matrix

$$\sqrt{W}\, X (X' W X)^{-1} X' \sqrt{W}$$

with

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix}, \quad X_i = [\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{in_i}]', \quad W = \operatorname{diag}[W_1, \ldots, W_k]$$

and

$$W_i = \operatorname{diag}[\pi_{i1}(1 - \pi_{i1}), \ldots, \pi_{in_i}(1 - \pi_{i,n_i})], \quad i = 1, \ldots, k.$$

2. **Data pattern modification** A simple and ad hoc approach to ensure convergence of the ML likelihood method is by modifying the data to obtain the next worse case scenario. The original idea came from reliability studies where it is common to change situations with no failures to a failure at the longest survival time(s) to estimate parameters or to obtain bounds on failure time estimates. This practice of inducing failures might be "conservative" because it has the potential of giving failure estimates higher than

what the observed process yields. In our case, this is achieved by forcing a death at the largest mite infestation for each of the two origins with no deaths. For each genetic origin with no deaths, the data modification consists of switching the Status variable from "Alive" to "Dead" for the observation with the largest mite-infestation level. As we discuss in Section 4.1, this ad hoc method gives reasonable answers when compared with the penalized ML approach.

2.2 Generalized Estimating Equations

The data in this study are longitudinal because colonies were observed multiple times. Longitudinal data can be correlated, and if this possible correlation is ignored, the standard errors might be downward biased. The generalized estimating equations (GEE) approach is a quasi-likelihood estimation method that is an alternative to ML for longitudinal data. The GEE approach has frequently been used in biomedical and health sciences. A characteristic of the GEE approach is that there is no need for a distributional assumption of the observations, and the entire approach is based on the regression model (3) and an approximate variance matrix for $\text{Var}(\boldsymbol{Y}_i), i = 1, \ldots, k$. Fitzmaurice, Laird, and Ware (2004, Chapter 11) provide a complete description of the GEE approach and illustrate its application with several examples.

Using the notation of Section 2.1, the GEE estimate of $\boldsymbol{\beta}$, say $\widehat{\boldsymbol{\beta}}_G$, is obtained from the iterated solution of the estimating equation (Liang and Zeger, 1986)

$$S(\boldsymbol{\beta}, R) = \sum_{i=1}^{k} D_i V_i^{-1}(\widehat{\boldsymbol{\alpha}})(\boldsymbol{Y}_i - \boldsymbol{\pi}_i) = 0, \tag{6}$$

where $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{in_i})'$, $D_i = \partial \boldsymbol{\pi}_i / \partial \boldsymbol{\beta}$, $V_i(\boldsymbol{\alpha}) = \sqrt{A_i}\, R_i(\boldsymbol{\alpha}) \sqrt{A_i}$, A_i is an $n_i \times n_i$ matrix with $\pi_{ij}(1 - \pi_{ij})$ as the jth diagonal element, $R_i(\boldsymbol{\alpha})$ is a $n_i \times n_i$ "working" correlation matrix fully specified by the parameters $\boldsymbol{\alpha}$, and $\widehat{\boldsymbol{\alpha}}$ is a consistent estimate of $\boldsymbol{\alpha}$. The estimation method is iterative because $\widehat{\boldsymbol{\alpha}}$ depends on $\widehat{\boldsymbol{\beta}}_G$ and then both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ must be estimated simultaneously.

For example, for an autoregressive correlation matrix $R_i(\alpha)$, the correlation between Y_{ij} and $Y_{i,j+t}$ is modeled by $\text{Corr}(Y_{ij}, Y_{i,j+t}) = \alpha^t$, for $t = 0, 1, \ldots, n_i - j$. At convergence, a consistent estimate for the correlation parameter α is

$$\widehat{\alpha} = \left[\frac{1}{\sum_{i=1}^{k}(n_i - 1) - (p + 1)} \right] \sum_{i=1}^{k} \sum_{j=1}^{n_i - 1} \widehat{e}_{ij} \widehat{e}_{i,j+1},$$

where $\widehat{e}_{ij} = (y_{ij} - \widehat{\pi}_{ij}) / \sqrt{\widehat{\pi}_{ij}(1 - \widehat{\pi}_{ij})}$ and $\widehat{\pi}_{ij}$ is the estimate of π_{ij} in (1) using the GEE $\widehat{\boldsymbol{\beta}}_G$ estimate for $\boldsymbol{\beta}$ obtained in (6).

The GEE method does not completely specify a parametric model; consequently no likelihood function exists. As a result, the Akaike information criterion (AIC) goodness of fit statistic is not available. A similar statistic, the QIC (quasi-likelihood) is useful in identifying important model variables. The QIC is defined as follows

$$\mathrm{QIC}(R) = -2\log\left[L(\widehat{\boldsymbol{\beta}}_G)\right] + 2\,\mathrm{trace}(\widehat{\Omega}_I\widehat{V}_R),$$

where $\widehat{\boldsymbol{\beta}_G}$ is the estimate of $\boldsymbol{\beta}$ obtained from the GEE, $\widehat{V}_R$ is a robust estimator of the covariance matrix R in the GEE method, and

$$\widehat{\Omega}_I = -\left.\frac{\partial^2 S(\boldsymbol{\beta}, I)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right|_{\widehat{\boldsymbol{\beta}}_G}$$

is the estimate of the inverse of $\widehat{V}_R$.

When a model is approximately correctly specified, $\Omega_I\widehat{V}_R$ should be well approximated by a $p \times p$ identity matrix, where p is the number of explanatory variables in the model. Then $\mathrm{trace}(\widehat{\Omega}_I\widehat{V}_R) \approx \mathrm{trace}(I) \approx p$. This motivates the definition

$$\mathrm{QIC}_{\mathrm{c}} = -2\log\left[L(\widehat{\boldsymbol{\beta}}_G)\right] + 2p.$$

$\mathrm{QIC}_{\mathrm{c}}$ is useful in variable selection. Also, when $\mathrm{QIC} \approx \mathrm{QIC}_{\mathrm{c}}$, it is an indication of a model correctly specified (Pan 2001), confirming that an appropriate correlation structure $R(\boldsymbol{\alpha})$ and the important experimental variables have been included in the model.

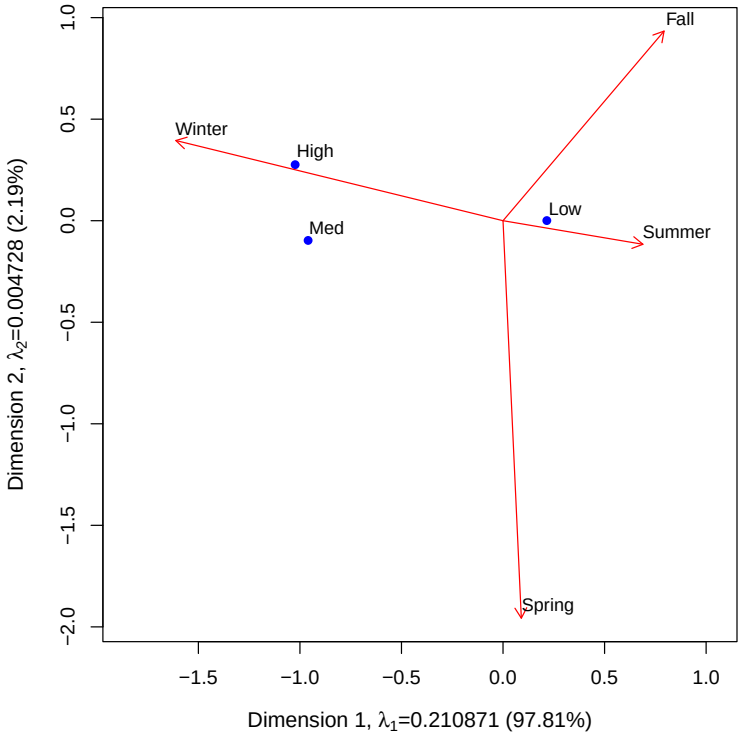
3. An Exploratory Analysis of the Associations Among the Variables

In an initial analysis of the association between the dependent variable Status (Dead or Alive) and the continuous independent variable Infestation, and the categorical independent variables (Origin and Season), we defined discrete "Infestation Class" of the continuous variable as follows: a Low Infestation Class, L, corresponding to mite infestations below 0.50 mites per gram of bees, a High Infestation Class, H, corresponding to infestations exceeding 1.0 mites per gram, and the rest of the observations were classified at the Medium Infestation Class M. Table 1 shows the coefficients of association between Status and the variables Origin, Season, and Infestation Class. Because the classification table of Status versus Origin has a large number of cells with expected counts less than 5, the χ^2_{df} test may not be valid. The exact Fisher test, F, for Status versus Origin gives $F = 5.35 \times 10^{-7}$ with a p-value of 0.17. In conclusion, there is no discernible association between the variables Status and Origin in the data. The associations between Status and the variables Season and Infestation Class are much stronger.

Table 1. Coefficients of Association Between Status (Dead or Alive) and Other Observed Variables

Variable	df	χ^2_{df}	$Pr > \chi^2_{df}$
Origin	8	10.74	0.22
Season	3	11.08	0.01
Infestation Class	2	16.96	< 0.01

Figure 2. Correspondence Analysis for Season and Infestation



The lack of association between Origin and Status was expected, particularly because of the low number of colonies for a large number of states of the variable Origin ($n = 9$).

In view of the results in Table 1, it is necessary to explore the association between the last two variables in the table. A Chi-Square test for the association between

Season and Infestation Class gives a test value of $\chi^2_6 = 82.85$ which indicates a strong association between these variables. The nature of this association can be illustrated using correspondence analysis (CA). Figure 2 shows a "row principal asymmetric biplot" for the CA (Greenacre 2008, Chapter 11) between Infestation Class and Season. Infestation Class is plotted in principal coordinates and Season is shown in standard coordinates. The eigenvalues for the CA analysis are $\lambda_1 = 0.210871\,(97.81\%)$ and $\lambda_2 = 0.004728\,(2.19\%)$. The plot shows that the Winter season is mainly associated with the High Infestation Class, the Fall and the Summer seasons are associated with values in the Low Infestation Class, and the Spring is associated with the Medium Infestation Class. In Section 4, we use the association between the categorical variable Season and the continuous variable Infestation to explain some of the logistic regression results.

4. Results

4.1 Results with Logistic Regression

Firth's Fits

Firth's fit for the model including the variables Infestation, Season, and Origin, using the entire original data are shown in Table 2. This table suggests that Season and Origin are not important variables to describe status of the colonies. Three other models were considered and fit using Firth's method: Infestation as

Table 2. Analysis of Effects Using Firth's Method and the Original Data

Variable	df	Wald	$\Pr > \chi^2_{\text{df}}$
Infestation	1	13.7992	0.0002
Season	3	2.4599	0.4826
Origin	8	11.7510	0.1627

the only explanatory variable and each variable alone with Infestation. The AICs and the SC (Schwartz Bayesian Criterion) for all models are shown in Table 3. According to the AIC criteria, the model including all three variables is the best model, but as shown in Table 2, Season and Origin are statistically non-significant in that model. On the other hand, the Schwartz criterion statistic suggests that the best model includes only the Infestation variable. Using parsimony in the model selection, this suggests that the model with just Infestation should be adequate.

Firth's fit for the entire data and the variable Infestation is shown in Table 4. Firth's probability of death estimates $\widehat{\pi}_F(x)$, as a function of mite Infestation, x, are obtained from (1) using the parameter estimates in Table 4. The probability of death estimate, $\widehat{\pi}_F(x)$, as a function of Infestation, x, is shown in Figure 3.

Table 3. Comparison of Models Using Firth's Method and the Original Data

Model	Variables in the Model	AIC	SC
1	Infestation	221.64	229.25
2	Infestation and Season	217.31	236.32
3	Infestation and Origin	216.87	254.89
4	Infestation, Season, and Origin	212.87	262.29

Table 4. Firth's Fit Using the Original Data

Parameter	Estimate	Std Err	$\Pr > \chi^2_1$
Intercept	−2.748	0.2615	<0.0001
Infestation	1.229	0.2594	<0.0001

Data Modified Fits

The model including the variables Infestation, Season, and Origin was fit using ML and the modified data (forcing a death for the largest infestation for the two categories of the variable Origin without deaths), as explained at the end of Section 2.1). Again, the variables Season and Origin were not statistically significant, and the model using just the variable Infestation seems adequate. This model is also suggested because it has the lowest AIC score in Table 5 where all four models were fit using logistic regression with the modified data.

Table 5. Comparison of Models Using Logistic Regression and the Modified Data

Model	Variables in Model	AIC
1	Infestation	232.51
2	Infestation and Season	236.25
3	Infestation and Origin	238.80
4	Infestation, Season, and Origin	242.80

Simple Logistic Regression Fit

The analyses presented above suggest omitting the variables Origin and Season from the model. Table 5 suggests that in a logistic regression fit, the variable Season has more ability than the variable Origin to describe the variability in the data. We therefore proceeded to fit logistic regression models that include only the variables Infestation and Season. In this case, because Origin is not included in the model, there is convergence of the ML method for both models, and one does not need Firth's method or the modified data that results from forcing deaths. Table 6 shows a test for the effect of Season in a logistic regression model including

Infestation and Season as variables. This test indicates that when Infestation is considered, Season is a non-significant effect. The association analysis in Section 3 showed that Status is associated with both Season and Infestation and that Season is associated with Infestation. The CA in that section also showed that the Winter Season is associated with high levels of Infestation. These generally higher infestation levels in the Winter season are caused by smaller amounts of developing brood to harbor mites and a concentration of mites in adult worker bees (Rinderer et al. 2001). Thus we conclude that the dominating effect is Infestation and that the effect of Infestation is consistent across Seasons. The variable Season is confounded with the variable Infestation. Table 7 shows the ML estimates when just the variable Infestation is included in the logistic regression fit.

Table 6. Season and Infestation Effect in a Logistic Regression with the Original Data

Effect of	df	Wald	$\Pr > \chi^2_{\text{df}}$
Season	3	2.8532	0.4148
Infestation	1	15.8189	<0.0001

Table 7. ML Estimation for Infestation Using Logistic Regression with the Original Data

Parameter	Estimate	Std Err	Wald	$\Pr > \chi^2_1$
Intercept	−2.776	0.2638	110.72	<0.0001
Infestation	1.249	0.2608	22.93	<0.0001

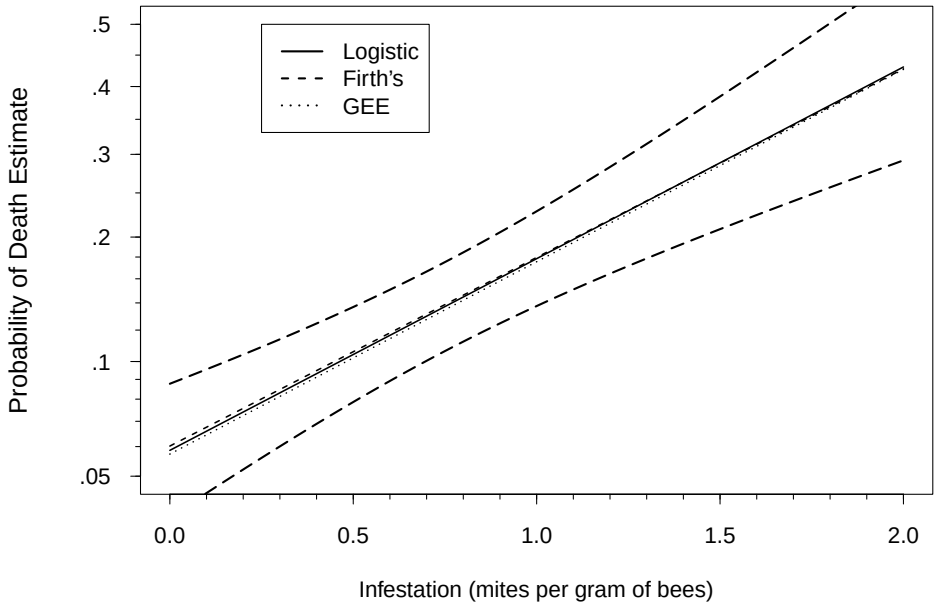
Using the parameter estimates in Table 7 and (1), the probability of death estimate, $\widehat{\pi}_L(x)$, as a function of Infestation, x, for the simple logistic model is

$$\widehat{\pi}_L(x) = \frac{\exp(-2.776 + 1.249x)}{1 + \exp(-2.776 + 1.249x)}. \tag{7}$$

The probability of death estimate for a colony with no infestation (that is, $x = 0$) is $\widehat{\pi}_L(0) = 6\%$. Also from (1) and (7), the approximate increase in odds of death for every one unit increase in the variable Infestation (measured in mites per gram of bees as a unit) is ≈ 2.49 with a 95% confidence interval of $[1.09, 4.81]$.

Figure 3 shows the probability of death estimates $\widehat{\pi}_L(x)$ and point-wise approximate 90% confidence intervals for $\pi(x)$, based on the simple logistic model. The slope of the logistic line, 1.249, is the ML estimate of log-odd increase for every one unit increase of mite infestation. The plot also shows Firth's estimate, $\widehat{\pi}_F(x)$, from Section 4.1, and the GEE estimate $\widehat{\pi}_G(x)$, of Section 4.2. The range $[0, 2]$ for mite infestation was chosen to include most of the mite-infestation values observed in the data (see Figure 1).

Figure 3. Logistic $\widehat{\pi}_L(x)$, Firth's $\widehat{\pi}_F(x)$, and GEE $\widehat{\pi}_G(x)$ Probabilities of Death Estimates in the Logistic Scale $\log[\pi(x)/(1-\pi(x))]$. The Non-Linear Curves are Point-wise Approximate 90% Confidence Intervals for $\pi(x)$ Based on the Logistic Model



4.2 Results with GEE

Since the data were collected on the same colonies over time, the data are correlated. The GEE analysis was done with the full dataset. There was a total of $k = 55$ clusters corresponding to the number of untreated colonies that were sampled. Table 8 shows the results obtained applying the GEE approach using the auto-regressive "working" correlation structure.

The estimate for α is $\widehat{\alpha} = -0.0298$. With this small value for $\widehat{\alpha}$, observations that are two or more periods apart have absolute correlation estimates smaller than 10^{-3}. This indicates that the data show little correlation among the measurements. Analyses with other "working" correlation structures lead to the same conclusion. The probability of death estimate, $\widehat{\pi}_G(x)$, as a function of Infestation, x, is shown in Figure 3.

To check the goodness of fit, we examined the QIC and QIC_c. As seen in Table 7, $\text{QIC}_\text{c} \approx \text{QIC}$, which implies that the model is approximately correctly specified.

Table 8. Analysis of Effects Using Generalized Estimating Equations

Parameter	Estimate	Std Err	$\Pr > \chi^2_1$
Intercept	−2.8004	0.2883	<0.0001
Infestation	1.2533	0.2377	<0.0001

Table 9. Analysis of Effects Using Generalized Estimating Equations

Variable	QIC	QIC_c
Infestation	227.6246	227.8295

4.3 Comparison of Firth's, Logistic, and GEE Analysis

Figure 3 shows that the estimates of probability of death as a function of Infestation, $\widehat{\pi}_F(x)$, $\widehat{\pi}_L(x)$, and $\widehat{\pi}_G(x)$ are basically indistinguishable for the mite-infestation values shown in the figure. Note that Figure 1 suggests that comparisons beyond 2 mites per gram of bees are infrequent.

Standard error estimates not shown here indicate that ignoring the dependency on the longitudinal data has the effect of providing smaller standard errors estimates. For these data, however, the differences in results with these three methods are small.

5. Conclusions

To ensure proper estimation with data that are dependent, the suggested model for examining the increasing risk with mite infestation is the GEE model. The model suggests that with every one unit increase in mite infestation (measured as mites per gram of bees), there is a 250% increase in odds of death of the colony. This direct relationship between the infestation with mites and the probability of colony death is suggested by numerous empirical observations and anecdotes. The statistical analyses here show that this relationship and the confidence limits around it can be quantified.

Using our local climate and geographical conditions, we quantified the expected probability of mortality at different levels of infestation while incorporating other potential variables affecting the relationship (colonies of different genetic origins and at different times of the year). Sample sizes were probably inadequate to detect possible minor differences between different genetic origins in the risk versus infestation relationship. While Season and Infestation had strong effects on mortality when considered as single variables, several analyses confirmed that they were confounded. High and medium levels of Infestation were more common in the Winter and Spring, respectively.

The results we report are a baseline against which to compare other possible data sets representing different climatic or geographic situations. A broader set of data and analyses like the ones presented would be valuable for a number of applications:

1. Data with measurements of individual colony infestations followed by evaluations of mortality in the subsequent quarter could provide some guidelines on which to base more conservative economic thresholds for treatment decisions. For example, an apiculturist could use a threshold mite level that produces a manageable (and recoverable) level of mortality the next quarter as a guideline for when to apply treatments for mites.

2. As genetically resistant types of bees are incorporated into beekeeping operations, monitoring reduced mite levels could indicate reduced risk and need for reliance on acaricide treatments.

3. There may be differences between genetic types of bees in the effects of mites once they attain certain potentially dangerous levels. Statistical analyses that compare strata as we used here could possibly be applied to datasets with larger sample sizes for groups of colonies of each genetic origin.

6. Acknowledgments

The authors thank the Executive Editor of Estadística for her encouragement and timely handling of this manuscript. We thank Dr. James Silva for his help with references and explanations about correspondence analysis. We thank two anonymous referees for comments that were helpful in improving the clarity and presentation of this manuscript.

References

ALBERT, A. and ANDERSON, J. A. (1984). "On the existence of maximum likelihood estimates in logistic regression models." *Biometrika.* **71**: 1–10.

BONNEY, G. (1987). "Logistic regression for dependent binary observations." *Biometrics.* **43**: 951–973.

CAREY, V., ZEGER, S., and DIGGLE, P. (1993). "Modelling multivariate binary data with alternating logistic regressions." *Biometrika.* **80**: 517–526.

FIRTH, D. (1993). "Bias Reduction of Maximum Likelihood Estimates." *Biometrika.* **80**: 27–38.

FITZMAURICE, L., LAIRD, N. M., and WARE, J. H. (2004). *Applied Longitudinal Analysis.* John Wiley & Sons, Hoboken NJ.

FITZMAURICE, G., and LIPSITZ, S. (1995). "A model for binary time series data with serial odds ratio patterns." *Applied statistics.* **44**: 51–61.

FRIES, I., IMDORF, A., and ROSENKRANZ, P. (2006). "Survival of mite infested (*Varroa destructor*) honey bee (*Apis mellifera*) colonies in Nordic climate." *Apidologie.* **37**: 564–570.

GENERSCH, E., OHE, W. von der, KAATZ, H., SCHROEDER, A., OTTEN, C., BUCHLER, R., BERG, S., RITTER, W., MUHLEN, W., GISDER, S., MEIXNER, M., LIEBIG, G., and ROSENKRANZ, P. (2010). "The German bee monitoring project: a long term study to understand periodically high winter losses of honey bee colonies." *Apidologie.* **41**: 332–352.

GREENACRE, M. (2008). *La Práctica del Análisis de Correspondencias.* Fundación BBVA, Barcelona.

GUZMAN-NOVOA, E., ECCLES, L., CALVETE, Y., MCGOWAN, J., KELLY, P. G., and CORREA–BENITEZ, A. (2010). "*Varroa destructor* is the main culprit for the death and reduced populations of overwintered honey bee (*Apis mellifera*) colonies in Ontario, Canada." *Apidologie.* **41**: 443–450.

KRAUS, B. and PAGE, R. E. JR. (1995). "Effect of *Varroa jacobsoni* (Mesostigmata: Varroidae) on feral *Apis mellifera* (Hymenoptera: Apidae) in California." *Environmental Entomology.* **24**: 1473–1480.

LIANG, K. Y. and ZEGER, S. L. (1986). "Longitudinal data analysis using generalized linear models." *Biometrika.* **73**: 13–22.

LIPSITZ, S., LAIRD, N., and HARRINGTON, D. (1991). "Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association." *Biometrika.* **78**: 153–160.

OLDROYD, B. P. (1999). "Coevolution while you wait: *Varroa jacobsoni*, a new parasite of western honeybees." *TREE.* **14**: 312–315.

PAN, W. (2001). "Akaike's information criterion in generalized estimating equations." *Biometrics.* **57**: 120–125.

PRENTICE, R. (1988). "Correlated binary regression with covariates specific to each binary observation." *Biometrics.* **44**: 1033–1048.

RINDERER, T. E., DE GUZMAN, L. I, DELATTE, G. T., STELZER, J. A., LANCASTER, V. A., KUSNETSOV, V., BEAMAN, L., WATTS, R., and HARRIS, J. W. (2001). "Resistance to the parasitic mite *Varroa destructor* in honey bees from far-eastern Russia." *Apidologie.* **32**: 381–394.

SANTNER, T. J. and DUFFY, D. (1986). "A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models." *Biometrika.* **73**. 755-758.

VILLA, J. D., BUSTAMANTE, D. M., DUNKLEY, J. P., and ESCOBAR, L. A. (2008). "Changes in honey bee (Hymenoptera: Apidae) colony swarming and survival pre- and post-arrival of *Varroa destructor* (Mesostigmata: Varroidae) in Louisiana." *Annals of the Entomological Society of America.* **101**: 867–871.

WENNER, A. M., THORP, R. W., and BARTHELL, J. F. (2009). "Biological control and eradication of feral honey bee colonies on Santa Cruz Island, California: A summary." In Damiani, C. C. and D. K. Garcelon (eds.). *Proceedings of the 7th California Islands Symposium.* Institute for Wildlife Studies, Arcata, CA.

Received November 2011
Revised August 2012