Environ Ecol Stat
DOI 10.1007/s10651-013-0257-0

Model selection criteria for overdispersed data and their application to the characterization of a host-parasite relationship

Hyun-Joo Kim · Joseph E. Cavanaugh · Tad A. Dallas · Stephanie A. Foré

Received: 6 January 2012 / Revised: 1 June 2013

Abstract In the statistical modeling of a biological or ecological phenomenon, selecting an optimal model among a collection of candidates is a critical issue. To identify an optimal candidate model, a number of model selection criteria have been developed and investigated based on estimating Kullback's (Information theory and statistics. Dover, Mineola, 1968) directed or symmetric divergence. Criteria that target the directed divergence include the Akaike (2nd international symposium on information theory. Akadémia Kiadó, Budapest, Hungary, pp 267–281, 1973, IEEE Trans Autom Control AC 19:716–723, 1974) information criterion, AIC, and the "corrected" Akaike information criterion (Hurvich and Tsai in Biometrika 76:297–307, 1989), AICc; criteria that target the symmetric divergence include the Kullback information criterion, KIC, and the "corrected" Kullback information criterion, KICc (Cavanaugh in Stat Probab Lett 42:333–343, 1999; Aust N Z J Stat 46:257–274, 2004). For overdispersed count data, simple modifications of AIC and AICc have been increasingly utilized: specifically, the quasi Akaike information criterion, QAIC, and its corrected version, QAICc (Lebreton et al. in Ecol Monogr 62(1):67–118 1992). In this paper, we propose

Handling Editor: Pierre Dutilleul.

H.-J. Kim (✉)
Department of Statistics, Truman State University, Kirksville, MO 63501, USA
e-mail: hjkim@truman.edu

J. E. Cavanaugh
Department of Biostatistics, The University of Iowa, Iowa city, IA 52242, USA

T. A. Dallas
Odum School of Ecology, The University of Georgia, Athens, GA 30602, USA

S. A. Foré
Department of Biology, Truman State University, Kirksville, MO 63501, USA

Published online: 30 June 2013

analogues of QAIC and QAICc based on estimating the symmetric as opposed to the directed divergence: QKIC and QKICc. We evaluate the selection performance of AIC, AICc, QAIC, QAICc, KIC, KICc, QKIC, and QKICc in a simulation study, and illustrate their practical utility in an ecological application. In our application, we use the criteria to formulate statistical models of the tick (*Dermacentor variabilis*) load on a white-footed mouse (*Peromyscus leucopus*) in northern Missouri.

Keywords AIC · KIC · Model selection criterion · Overdispersion · Quasi likelihood · QAIC

1 Introduction

Recently, researchers in ecology and biology have expanded their approach to data analysis and inference. Rather than relying upon traditional hypothesis testing as the primary inferential paradigm, model selection is being increasingly adopted to deal with several competing hypotheses simultaneously (Johnson and Omland 2004). Model selection criteria provide useful tools for choosing a suitable model from a candidate family to characterize the underlying data. Ideally, a criterion will identify candidate models that are either too simplistic to adequately accommodate the data or unnecessarily complex.

Much research has been conducted on the use of model selection criteria and strategies to choose the "best" model among a potentially large candidate collection (Tibshirani 1996; Meinshausen 2007; Wang and Leng 2008; Min et al. 2010). The first model selection criterion to gain widespread acceptance was the Akaike (1973, 1974) information criterion, AIC. AIC is justified as a large-sample estimator of Kullback's (1968, p. 5) directed divergence between the generating model and a fitted candidate model, also known as I-divergence. Since the justification of AIC is based on the conventional asymptotic properties of maximum likelihood estimators, the criterion is applicable in a broad array of modeling frameworks. However, when the sample size is small, AIC tends to favor inappropriately high dimensional candidate models (Hurvich and Tsai 1989); this limits its effectiveness as a model selection criterion.

The "corrected" AIC, AICc, is an adjusted version of AIC originally proposed for linear regression with normal errors (Sugiura 1978; Hurvich and Tsai 1989). In the linear regression framework, for fitted models in the candidate class which are correctly specified or overfit, AIC is asymptotically unbiased and AICc is exactly unbiased as an estimator of Kullback's directed divergence.

For alternative regression models such as Poisson, binomial, or multinomial regression models, where the outcome is a discrete count, model selection is still an important yet challenging issue. The problem becomes more challenging when the degree of variation that exists in the modeled outcome is greater than that which is accommodated by the postulated distribution. Such a phenomenon is known as *overdispersion*, and may arise due to missing covariates, parameter heterogeneity, or partial dependence in the outcome (Burnham and Anderson 2002; Hilbe 2008). The criteria QAIC and QAICc (Lebreton et al. 1992) have been proposed as the quasi-likelihood analogues of AIC and AICc for modeling overdispersed count or binary data.

As an alternative to estimating the directed divergence, model selection criteria can be developed by estimating Kullback's (1968, p. 6) symmetric divergence between the generating model and a fitted candidate model. The symmetric divergence, also known as the J-divergence, is the the sum of two directed divergences. KIC was proposed as an asymptotically unbiased estimator of the J-divergence (Cavanaugh 1999). KIC has been justified under the same general conditions as AIC, and is therefore also broadly applicable. When used to evaluate fitted candidate models, the symmetric divergence is arguably more sensitive than the directed divergence towards detecting improperly specified models (Cavanaugh 2004).

KICc is a corrected variant of AICc that targets the symmetric divergence in the same manner that AICc targets the directed divergence. As with AICc, KICc has been justified for linear regression (Cavanaugh 2004) and nonlinear regression (Kim and Cavanaugh 2005) with normal errors.

In this paper, we discuss the aforementioned criteria in detail and propose analogues of QAIC and QAICc that target the symmetric as opposed to the directed divergence: the quasi Kullback information criterion, QKIC, and the corrected quasi Kullback information criterion, QKICc. We evaluate the selection performance of AIC, AICc, QAIC, QAICc, KIC, KICc, QKIC, and QKICc in a simulation study. The results are presented in Sect. 3.

In Sect. 4, we illustrate the practical utility of AIC, AICc, QAIC, QAICc, KIC, KICc, QKIC, and QKICc in an ecological application. The host-parasite relationship is a complex ecological interaction that is at the base of many epidemiological studies. We examine the effect of habitat, the number of nymphal ticks and the intraspecific factors of body mass, tail length, hind right foot length, and gender that potentially influence larvae *Dermacentor variabilis* (American dog tick) burdens on *Peromyscus leucopus* (white-footed mouse). We employ model selection criteria to identify an optimal model based on a subset of covariates. The data were collected from Adair County, Missouri, between 2006 and 2008.

The simulation studies and data analyses are conducted using R, version 2.9.0 (R Development Core Team 2009). Estimation is based on the iteratively reweighted least squares (IWLS) method, as implemented in the R functions glm, gamlss, and glm.nb.

2 Selection criteria based on Kullback information measures

Let Y be the response vector, and θ_k be an unknown parameter vector, where k refers to the dimension. Let $f(Y\,|\,\theta_k)$ denote the joint density of Y, or equivalently, the likelihood function. We will use $\hat{\theta}_k$ to denote the maximum likelihood estimator (MLE) for θ_k. Accordingly, $f(Y\,|\,\hat{\theta}_k)$ will represent the empirical likelihood corresponding to $f(Y\,|\,\theta_k)$.

Let θ_o denote the true parameter. Thus, $f(Y\,|\,\theta_o)$ will represent the "true" or generating model.

Consider a family of fitted models $\mathcal{F}(k) = \{f(Y\,|\,\hat{\theta}_{k_1}),\ f(Y\,|\,\hat{\theta}_{k_2}), \ldots,\ f(Y\,|\,\hat{\theta}_{k_R})\}$. To determine which of these fitted models best resembles $f(Y\,|\,\theta_o)$, we require a measure that provides a suitable reflection of the disparity between the true model $f(Y\,|\,\theta_o)$

and a candidate model $f(Y\,|\,\theta_k)$. Kullback's directed and symmetric divergence both fulfill this objective.

For two arbitrary parametric densities $f(Y\,|\,\theta)$ and $f(Y\,|\,\theta_*)$, Kullback's directed divergence between $f(Y\,|\,\theta)$ and $f(Y\,|\,\theta_*)$ with respect to $f(Y\,|\,\theta)$ is defined as

$$I(\theta, \theta_*) = \mathrm{E}_\theta \left[\ln \left\{ \frac{f(Y\,|\,\theta)}{f(Y\,|\,\theta_*)} \right\} \right], \tag{1}$$

and Kullback's symmetric divergence between $f(Y\,|\,\theta)$ and $f(Y\,|\,\theta_*)$ is defined as

$$J(\theta, \theta_*) = \mathrm{E}_\theta \left[\ln \left\{ \frac{f(Y\,|\,\theta)}{f(Y\,|\,\theta_*)} \right\} \right] + \mathrm{E}_{\theta_*} \left[\ln \left\{ \frac{f(Y\,|\,\theta_*)}{f(Y\,|\,\theta)} \right\} \right].$$

Here, E_θ denotes the expectation under $f(Y\,|\,\theta)$. Note that $J(\theta, \theta_*)$ is symmetric in its arguments whereas $I(\theta, \theta_*)$ is not. Thus, an alternate directed divergence, $I(\theta_*, \theta)$, may be obtained by switching the roles of $f(Y\,|\,\theta)$ and $f(Y\,|\,\theta_*)$ in (1). The sum of the two directed divergences yields the symmetric divergence: $J(\theta, \theta_*) = I(\theta, \theta_*) + I(\theta_*, \theta)$. The AIC and the KIC family of criteria are formulated based on Kullback's directed and symmetric divergence, respectively.

For the purpose of assessing the proximity between a certain fitted candidate model $f(Y\,|\,\hat{\theta}_k)$ and the true model $f(Y\,|\,\theta_o)$, we consider the measures

$$I(\theta_o, \hat{\theta}_k) = I(\theta_o, \theta_k)|_{\theta_k = \hat{\theta}_k} \qquad \text{and} \qquad J(\theta_o, \hat{\theta}_k) = J(\theta_o, \theta_k)|_{\theta_k = \hat{\theta}_k}.$$

Of these two, Cavanaugh (1999, 2004) conjectures that $J(\theta_o, \hat{\theta}_k)$ may be preferred, since it combines $I(\theta_o, \hat{\theta}_k)$ with its counterpart $I(\hat{\theta}_k, \theta_o)$, a measure which serves a related yet distinct function. To gauge the disparity between $f(Y\,|\,\hat{\theta}_k)$ and $f(Y\,|\,\theta_o)$, $I(\theta_o, \hat{\theta}_k)$ assesses how well samples generated under the true model $f(Y\,|\,\theta_o)$ conform to the fitted candidate model $f(Y\,|\,\hat{\theta}_k)$, whereas $I(\hat{\theta}_k, \theta_o)$ assesses how well samples generated under the fitted candidate model $f(Y\,|\,\hat{\theta}_k)$ conform to the true model $f(Y\,|\,\theta_o)$. As a result of these contrasting roles, $I(\theta_o, \hat{\theta}_k)$ tends to be more sensitive towards reflecting overfit models, whereas $I(\hat{\theta}_k, \theta_o)$ tends to be more sensitive towards reflecting underfit models. Accordingly, $J(\theta_o, \hat{\theta}_k)$ may be more adept at detecting misspecification than either of its components.

In what follows, we will show how the AIC family of model selection criteria arises through estimating a variant of $I(\theta_o, \hat{\theta}_k)$. The KIC family criteria arises through estimating a variant of $J(\theta_o, \hat{\theta}_k)$ in similar manner (Kim and Cavanaugh 2005).

For two arbitrary parametric densities $f(Y\,|\,\theta)$ and $f(Y\,|\,\theta_*)$, let

$$d(\theta, \theta_*) = \mathrm{E}_\theta[-2 \ln f(Y\,|\,\theta_*)]. \tag{2}$$

From (1) and (2), note that we can write

$$2I(\theta_o, \theta_k) = d(\theta_o, \theta_k) - d(\theta_o, \theta_o). \tag{3}$$

Since $d(\theta_o, \theta_o)$ does not depend on θ_k, any ranking of a set of candidate models corresponding to values of $I(\theta_o, \theta_k)$ would be identical to a ranking corresponding to values of $d(\theta_o, \theta_k)$. Hence, for the purpose at hand, $d(\theta_o, \theta_k)$ serves as a valid substitute for $I(\theta_o, \theta_k)$.

Now for a given set of MLEs $\hat{\theta}_k$,

$$d(\theta_o, \hat{\theta}_k) = d(\theta_o, \theta_k)|_{\theta_k = \hat{\theta}_k}$$

would provide a meaningful measure of separation between the true model and a fitted candidate model. Evaluating $d(\theta_o, \hat{\theta}_k)$ is not possible since doing so requires knowledge of θ_o. However, the work of Akaike (1973, 1974) suggests that the goodness-of-fit term $-2 \ln f(Y | \hat{\theta}_k)$ serves as a negatively biased estimator of $d(\theta_o, \hat{\theta}_k)$, and that under appropriate assumptions, this bias can be asymptotically approximated by twice the dimension of θ_k. With this motivation, AIC is defined as

$$\text{AIC} = -2 \ln f(Y | \hat{\theta}_k) + 2k.$$

As the sample size increases, the difference between the expected value of AIC and the expected value of Kullback's directed divergence should tend to zero. Accordingly, we may regard AIC as an asymptotically unbiased estimator of Kullback's directed divergence.

When the sample size n is large and k is comparatively small, the degree of bias incurred in estimating Kullback's directed divergence with AIC is negligible. However, when n is small and k is relatively large (e.g., $k \simeq n/2$), AIC severely underestimates Kullback's directed divergence for high dimensional fitted models in the candidate family. As a result, the criterion may inappropriately favor unnecessarily large models (Hurvich and Tsai 1989).

AICc was proposed to serve as an estimator of Kullback's directed divergence which is less biased in small-sample applications than traditional AIC (Hurvich and Tsai 1989; Hurvich et al. 1990). However, since the justification of AICc is contingent upon the structure of the candidate modeling framework, this criterion is less generally applicable than AIC.

AICc is defined as

$$\text{AICc} = -2 \ln f(Y | \hat{\theta}_k) + \frac{2nk}{(n - k - 1)}.$$

The criterion was originally proposed by Sugiura (1978) in the setting of linear regression models with normal errors. In this framework, the penalty term can be evaluated exactly for correctly specified and overfit models. AICc is justified as an approximately unbiased estimator of the directed divergence in various frameworks, including normal nonlinear regression (Hurvich and Tsai 1989).

In this paper, model selection for count data with overdispersion will be considered. Model selection criteria that serve as counterparts to AIC and AICc will be discussed, and new criteria will be proposed based on targeting Kullback's symmetric divergence.

Count data with overdispersion is common in wildlife and ecological studies (Eberhardt 1978). Poisson regression is a widely used statistical modeling framework when the response variable represents a discrete count. One of the main characteristics of the Poisson distribution is that the variance is the same as the mean. When the empirical variance exceeds the mean, the data exhibits *overdispersion* (Agresti 2002). Overdispersion can be tested using various inferential procedures including the score test, the likelihood ratio test, or the Wald test. A simple and effective approach for accommodating overdispersion is to estimate the excess variation using the reduced chi-square statistic:

$$\hat{c} = \chi^2/d.f. = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}/d.f.$$

Here, μ_i and $V(\mu_i)$ respectively represent the mean and variance functions for y_i under the postulated model, and $d.f.$ denotes the model degrees of freedom. The reduced chi-squared is normalized for the number of data points and accounts for model complexity via the degrees of freedom. When $\hat{c}$ is approximately 1, no overdispersion is evident. Overdispersion is reflected by values of $\hat{c}$ larger than 1. The larger the value of $\hat{c}$, the more extreme the degree of overdispersion.

The estimated variances of the model parameters may be obtained by multiplying the variances under the ordinary Poisson model by $\hat{c}$. As a rule of thumb, values of $\hat{c}$ in the range of 1 to 4 imply that the model structure is adequate (Burnham and Anderson 2002).

Model selection criteria have been adapted to deal with overdispersion based on the concepts of quasi-likelihood and variance inflation (Wedderburn 1974; Hurvich and Tsai 1995). As overdispersion becomes more pronounced, $-2 \ln f(Y \mid \hat{\theta}_k)$, the goodness-of-fit term, becomes larger. An inappropriately large goodness-of-fit term may overwhelm the contribution of the penalty term, and thereby have a dramatic influence on selection patterns. As an alternative, the goodness-of-fit term can be computed as $-2 \ln f(Y \mid \hat{\theta}_k)/\hat{c}$ to compensate for the variance inflation represented by $\hat{c}$ (Burnham and Anderson 2002).

We note that the ordinary likelihood can be regarded as the joint probability distribution for the data. The division of the log-likelihood by the variance inflation factor $\hat{c}$ results in a measure that cannot be back-transformed (via division by -2 and exponentiation) to a proper probability distribution. However, the rescaled log-likelihood leads to the same estimating equations for the regression coefficients as the ordinary log-likelihood, and leads to a Fisher information matrix rescaled in the same manner as the log-likelihood. The latter results in estimated variances for the regression parameter estimates that are multiplied by the variance inflation factor. Therefore, the rescaled log-likelihood can be properly viewed as a (log) quasi-likelihood (Agresti 2002).

With the preceding motivation, QAIC and QAICc (Lebreton et al. 1992; Burnham and Anderson 2002) are defined as

$$\text{QAIC} = \frac{-2 \ln f(Y \mid \hat{\theta}_k)}{\hat{c}} + 2k,$$

and

$$\text{QAICc} = \frac{-2\ln f(Y\,|\,\hat{\theta}_k)}{\hat{c}} + \frac{2nk}{(n-k-1)}.$$

Here, $\hat{c}$ can be obtained from the global model; i.e., the candidate model that contains all covariates of interest, and thereby subsumes all of the models in the candidate family.

We note that quasi-likelihood methods of variance inflation are most appropriate only after a reasonable degree of structural adequacy has been achieved in characterizing the mean (Burnham and Anderson 2002). In addition to Poisson regression models, quasi-likelihood criteria perform well in product multinomial regression models of capture-recapture data in the presence of differing levels of overdispersion (Anderson et al. 1994).

KIC was proposed as an analogue of AIC that targets the symmetric as opposed to the directed divergence (Cavanaugh 1999). (As in the development of AIC, the constant $d(\theta_o, \theta_o)$ is discarded from the estimated measure). KIC is defined as

$$\text{KIC} = -2\ln f(Y\,|\,\hat{\theta}_k) + 3k. \tag{4}$$

As the sample size increases, the difference between the expected value of KIC and the expected value of Kullback's symmetric divergence should tend to zero. Accordingly, we may regard KIC as an asymptotically unbiased estimator of the J-divergence.

Cavanaugh (2004) proposed an analogue of AICc for normal linear regression models based on estimating Kullback's symmetric divergence. The criterion is defined as

$$\text{KICc} = -2\ln f(Y\,|\,\hat{\theta}_k) + n\ln\left(\frac{n}{n-k+1}\right) + \frac{n\,\{(n-k+1)(2k+1)-2\}}{(n-k-1)(n-k+1)}.$$

KICc is an exactly unbiased estimator of symmetric divergence in the normal linear regression framework and an approximately unbiased estimator in the normal nonlinear regression framework (Kim and Cavanaugh 2005).

Similar to QAIC and QAICc, quasi-likelihood criteria that serve as counterparts to KIC and KICc can be proposed by suitably modifying the goodness-of fit term:

$$\text{QKIC} = \frac{-2\ln f(Y\,|\,\hat{\theta}_k)}{\hat{c}} + 3k,$$

$$\text{QKICc} = \frac{-2\ln f(Y\,|\,\hat{\theta}_k)}{\hat{c}} + n\ln\left(\frac{n}{n-k+1}\right) + \frac{n\,\{(n-k+1)(2k+1)-2\}}{(n-k-1)(n-k+1)}.$$

These criteria target the symmetric divergence in the same manner that QAIC and QAICc target the directed divergence.

Certain simulation studies have shown that the criteria in the KIC family outperform their AIC counterparts; see, for instance, Cavanaugh (1999, 2004), Kim and Cavanaugh (2005). QKIC and QKICc are based on the same rationale as KIC and KICc: when data is overdispersed, the symmetric divergence based on the quasi-likelihood may serve as a more sensitive discrepancy measure than the directed divergence.

3 Simulation sets

3.1 Fitting negative binomial regression models to an overdispersed Poisson outcome

Assume a collection of data Y has been generated according to an unknown parametric density $f(Y \mid \theta_o)$, one which corresponds to the Poisson regression model,

$$ln(\lambda_{oi}) = X'_{oi}\beta_o, \quad y_i \sim \texttt{Poisson}(\lambda_{oi}).$$

Here, y_i is a response variable, β_o is a $(p_o + 1) \times 1$ parameter vector, and X_{oi} is $(p_o + 1) \times 1$ vector of covariates. The parameter λ_{oi} denotes the mean of y_i.

Overdispersion commonly arises when relevant covariates are not included among the set of explanatory variables used to formulate a candidate model. With this motivation, the following scenario is considered: the outcome is generated from a Poisson model with a particular set of covariates, yet some of these covariates are omitted from the mean structure for each of the models in the candidate family. Such a setting is practically appealing, since it is rarely possible to identify and measure all of the covariates that may be pertinent to the outcome.

When overdispersion is severe, a structural change in the fitted model is warranted. In this simulation study, data with various levels of overdispersion are considered. The candidate models are based on the negative binomial distribution, which accommodates a specific type of overdispersion in the variance structure. Specifically, we use the NEGBIN2 class of models proposed by Cameron and Trivedi (1986). Thus, the candidate models postulated for the data are of the form

$$ln(\lambda_i) = X'_i\beta, \qquad y_i \sim \texttt{NEGBIN2}(\lambda_i, \alpha),$$

where β is a $(p + 1) \times 1$ parameter vector, and X_i is $(p + 1) \times 1$ vector of covariates. The parameter λ_i denotes the mean of y_i, and α represents the overdispersion parameter.

The negative binomial model adjusts for variance inflation through the incorporation of the overdispersion parameter. For the NEGBIN2 class, the variance function is of the form $\lambda_i + \alpha\lambda_i^2$, where λ_i is the mean of y_i. Thus, the variance function is quadratic in the mean, allowing the variance to exceed the mean by the additive factor $\alpha\lambda_i^2$.

From a practical perspective, this simulation setting can be described as one in which the investigator recognizes that overdispersion might be present due to inaccessible yet relevant covariates. The investigator decides to use candidate models based on the NEGBIN2 class. Any excess variation that cannot be characterized by the variance function of the NEGBIN2 model will be captured via the reduced chi-square statistic $\hat{c}$.

We examine the behavior of AIC, AICc, QAIC, QAICc, KIC, KICc, QKIC and QKICc as order selection criteria in the preceding framework.

In the first nine simulation sets, 1000 samples are generated from a true model where the true parameter vector is of the form $\beta_o = (1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, \gamma, \gamma)'$. In scalar form, the true model can be written as

$$ln(\lambda_{oi}) = 1 + x_{1i} + x_{2i} + x_{3i} + \gamma x_{11i} + \gamma x_{12i}, \quad y_i \sim iid \ \texttt{Poisson}(\lambda_{oi}), \quad (5)$$

where γ is the common coefficient of x_{11i} and x_{12i}.

Nested negative binomial regression models are fit to the data. However, in fitting the models, only x_1 through x_{10} are assumed to be available. The regressors for all models are generated using a Uniform(−1, 1) distribution. Three different values for the coefficient γ are considered: 0.1, 0.5, and 0.9. For each of the true models, three different sample sizes n are employed: 50, 100, and 200. Note that when the coefficient γ is close to 0, the role of the missing covariates x_{11} and x_{12} is minimal, and the absence of the covariates will only induce a minor degree of overdispersion. As γ grows larger, the degree of overdispersion becomes more pronounced.

Nested models of orders 1 through 10 are entertained. Specifically, the model of order 1 includes only x_1, the model of order 2 includes x_1 and x_2, and the largest model of order 10 includes all of the variables $x_1, x_2, \ldots, x_{10}$. For every sample in a set, the fitted model favored by each criterion is recorded. The excess variation measure $\hat{c}$ is also recorded. Over the 1000 samples, the order selections of the eight criteria are tabulated and summarized.

The order selection results for the nine sets corresponding to generating model (5) are featured in Table 1. The candidate models are nested, but the true model is not nested in any of the candidate models due to the omission of x_{11} and x_{12}. Thus, none of the candidate models are properly specified. However, it is clear that the "best" candidate model is the model of order 3, based on x_1, x_2, and x_3. The remaining covariates x_4 through x_{10} are generated independently of the inaccessible covariates (x_{11} and x_{12}), so that the addition of the remaining covariates cannot possibly offer a meaningful improvement to the fit of the model.

In Table 1, a "correct model" selection corresponds to a criterion selection of the fitted candidate model of order 3; i.e., the model based on x_1, x_2, and x_3. A "smaller model" selection corresponds to a choice of a model of order 1 or 2, and a "larger model" selection corresponds to a choice of one of the orders 4 through 10.

The order selection results lead to some clear conclusions regarding the selection patterns of the criteria. First, each J-divergence criterion generally obtains more correct selections than its I-divergence counterpart. Second, each "corrected" criterion outperforms its non-adjusted counterpart. Third, when a substantial degree of overdispersion exists ($\gamma = 0.5$, $\gamma = 0.9$), each quasi-likelihood criterion generally outperforms the corresponding criterion based on the ordinary likelihood. With minimal overdispersion ($\gamma = 0.1$), the performance of each quasi-likelihood criterion is comparable to that of the corresponding ordinary likelihood criterion. QKICc generally outperforms the other criteria, followed by KICc and QKIC. When the sample size is small and the degree of overdispersion is severe, QKIC and QKICc tend to choose underfit models more often than KIC and KICc.

For this simulation set, note that the extra variation measure $\hat{c}$ tends to be of moderate size (1.013–1.282). This implies that the most of the variation caused by the inaccessible covariates is captured by the variance function of the NEGBIN2 model.

Table 1 Order selections

Set	n	Selections	Criterion							
	γ m($\hat{c}$)		AIC	AICc	QAIC	QAICc	KIC	KICc	QKIC	QKICc
1	50	Smaller model	0	0	0	0	0	0	0	0
	0.1	Correct model	723	834	713	833	872	925	853	918
	1.119	Larger model	277	166	287	167	128	75	147	82
2	50	Smaller model	0	0	0	1	0	0	2	5
	0.5	Correct model	583	757	716	859	804	885	872	925
	1.282	Larger model	417	243	284	140	196	115	126	70
3	50	Smaller model	3	6	8	15	12	17	22	34
	0.9	Correct model	588	747	707	843	785	866	853	902
	1.256	Larger model	409	247	285	142	203	117	125	64
4	100	Smaller model	0	0	0	0	0	0	0	0
	0.1	Correct model	713	775	705	775	873	902	875	901
	1.009	Larger model	287	225	295	225	127	98	125	99
5	100	Smaller model	0	0	0	0	0	0	0	0
	0.5	Correct model	652	725	698	790	837	876	881	910
	1.125	Larger model	348	275	302	210	163	124	119	90
6	100	Smaller model	0	0	3	4	0	0	11	13
	0.9	Correct model	657	737	721	789	848	881	877	913
	1.128	Larger model	343	263	276	207	152	119	112	74
7	200	Smaller model	0	0	0	0	0	0	0	0
	0.1	Correct model	716	738	718	749	885	905	881	902
	1.013	Larger model	284	262	282	251	115	95	119	98
8	200	Smaller model	0	0	0	0	0	0	0	0
	0.5	Correct model	696	729	723	759	865	883	884	898
	1.054	Larger model	304	271	277	241	135	117	116	102
9	200	Smaller model	0	0	0	0	0	0	0	0
	0.9	Correct model	647	681	687	730	834	857	856	880
	1.085	Larger model	353	319	313	270	166	143	144	120

The generating model is (5); n is the sample size, γ is the regression coefficient for x_{11} and x_{12}, and m($\hat{c}$) is the average of 1000 variance inflation factors calculated using the largest negative binomial candidate regression model. Order selections for candidate models of order 1 or 2 are reported in the rows labeled “Smaller model”; selections for models of orders 4 through 10 are reported in the rows labeled “Larger model”. Correct order selections (order 3) are reported in the rows labeled “Correct model”

3.2 Fitting Poisson and negative binomial regression models to a negative binomial outcome

In many data analyses, overdispersion in the outcome might not be recognized or suitably characterized, leading to an inappropriate formulation of the variance structure for the family of candidate models. For instance, in modeling overdispersed count data, Poisson regression might be employed since the Poisson distribution is often viewed as the default for count outcomes. Alternatively, the overdispersion in the count data could be acknowledged, leading to the formulation of a candidate family based on the negative binomial distribution, yet a variance structure could be chosen that does not conform to the nature of the overdispersion.

In this section, we investigate the criterion performance when response data arising from a NEGBIN2 distribution is fit using Poisson, NEGBIN1, and NEGBIN2 regression models. As previously mentioned, with the NEGBIN2 model, the variance function is quadratic in the mean; with the NEGBIN1 model, the variance function is linear in the mean. For NEGBIN2 response data, Poisson and NEGBIN1 regression models have incorrect variance structures. However, the variance structure under a NEGBIN1 distribution is less erroneous than that under a Poisson distribution.

Assume a collection of data Y has been generated according to an unknown parametric density $f(Y \mid \theta_o)$, one which corresponds to the NEGBIN2 model:

$$ln(\lambda_{oi}) = X'_{oi}\beta_o, \qquad y_i \sim \mathtt{NEGBIN2}(\lambda_{oi}, \alpha_o).$$

Suppose that the candidate models postulated for the data are of the following three forms.

	Fitted model	Model equation	Distribution	Variance function
(3.2.1)	Poisson	$ln(\lambda_i) = X'_i\beta$	$y_i \sim$ Poisson (λ_i)	λ_i
(3.2.2)	NEGBIN1	$ln(\lambda_i) = X'_i\beta$	$y_i \sim$ NEGBIN1 (λ_i, α)	$\lambda_i + \alpha\lambda_i$
(3.2.3)	NEGBIN2	$ln(\lambda_i) = X'_i\beta$	$y_i \sim$ NEGBIN2 (λ_i, α)	$\lambda_i + \alpha\lambda_i^2$

Here, y_i is a response variable, β_o and β are $(p_o + 1) \times 1$ and $(p + 1) \times 1$ parameter vectors, and X_{oi} and X_i are $(p_o + 1) \times 1$ and $(p + 1) \times 1$ vectors of covariates. The parameters λ_{oi} and λ_i represent the means of y_i under the generating and candidate models, and α_o and α denote the corresponding overdispersion parameters. When α_o is close to 0, the generating distribution is close to a Poisson distribution; when α_o is larger, the variance of y_i exceeds that which is accommodated by the Poisson or NEGBIN1 distribution. Under the Poisson distribution, the overdispersion is ignored; under the NEGBIN1 distribution, the overdispersion is mischaracterized.

We examine the behavior of AIC, AICc, QAIC, QAICc, KIC, KICc, QKIC and QKICc as order selection criteria in settings where Poisson, NEGBIN1, and NEGBIN2 candidate families are used to model NEGBIN2 data.

In each of these three frameworks, we compile nine simulation sets based on the same generating model, as well as the same configurations of overdispersion parameters and sample sizes. For each set, 1000 samples are generated from a true model where $\beta_o = (1, 1, 1, 1)'$. Thus, in scalar form, the true model can be written as

$$ln(\lambda_{oi}) = 1 + x_{1i} + x_{2i} + x_{3i}, \quad y_i \sim iid \ \mathtt{NEGBIN2}(\lambda_{oi}, \alpha_o). \tag{6}$$

Table 2 Order selections

Set	p_o	Selections	Criterion							
	n $\alpha_o,\ m(\hat{c})$		AIC	AICc	QAIC	QAICc	KIC	KICc	QKIC	QKICc
1	3	Underfit	1	3	21	45	3	9	54	78
	50	Correctly specified	31	59	262	402	66	108	423	540
	1.0, 3.385	Overfit	968	938	717	553	931	883	523	382
2	3	Underfit	0	0	2	3	0	0	8	15
	50	Correctly specified	68	136	332	493	181	283	525	675
	0.5, 2.181	Overfit	932	864	666	504	819	717	467	310
3	3	Underfit	0	0	0	2	0	1	2	2
	50	Correctly specified	243	385	392	567	447	568	601	718
	0.2, 1.370	Overfit	757	615	608	431	553	431	397	280
4	3	Underfit	0	0	0	0	0	0	0	3
	100	Correctly specified	12	16	332	418	43	55	554	604
	1.0, 4.140	Overfit	988	984	668	582	957	945	446	393
5	3	Underfit	0	0	0	0	0	0	0	0
	100	Correctly specified	59	88	374	446	161	199	572	634
	0.5, 2.570	Overfit	941	912	626	554	839	801	428	366
6	3	Underfit	0	0	0	0	0	0	0	0
	100	Correctly specified	211	277	436	521	403	463	642	708
	0.2, 1.586	Overfit	789	723	564	479	597	537	358	292
7	3	Underfit	0	0	0	0	0	0	0	0
	200	Correctly specified	13	15	304	349	31	34	517	553
	1.0, 4.625	Overfit	987	985	696	651	969	966	483	447
8	3	Underfit	0	0	0	0	0	0	0	0
	200	Correctly specified	49	53	340	377	112	136	570	602
	0.5, 2.859	Overfit	951	947	660	623	888	864	430	398
9	3	Underfit	0	0	0	0	0	0	0	0
	200	Correctly specified	191	218	446	495	373	399	689	722
	0.2, 1.725	Overfit	809	782	554	505	627	601	311	278

The generating model is (6); n is the sample size, p_o is the true order, and α_o is the overdispersion parameter. m($\hat{c}$) is the average of 1000 variance inflation factors calculated using the largest Poisson candidate regression model. Order selections for candidate models of order 1 or 2 are reported in the rows labeled "Underfit"; selections for models of orders 4 through 10 are reported in the rows labeled "Overfit". Correct order selections (order 3) are reported in the rows labeled "Correctly specified"

The regressors for all models are generated from a Uniform(−1, 1) distribution. Three different overdispersion parameters α_o are considered: 1.0, 0.5, and 0.2. In addition, three different sample sizes n are employed: 50, 100, and 200. Candidate models of orders 1 through 10 are entertained. For every sample in a set, the fitted model favored by each criterion is recorded. The excess variation measure $\hat{c}$ is also recorded. Over the 1000 samples, the order selections are tabulated and summarized.

Table 3 Order selections

Set	p_o n α_o, $m(\hat{c})$	Selections	Criterion							
			AIC	AICc	QAIC	QAICc	KIC	KICc	QKIC	QKICc
1	3	Underfit	70	107	121	193	151	207	239	318
	50	Correctly specified	441	581	526	636	581	626	607	596
	1.0, 1.321	Overfit	489	312	353	171	268	167	154	86
2	3	Underfit	6	15	17	33	25	31	43	69
	50	Correctly specified	418	609	575	764	645	782	776	835
	0.5, 1.323	Overfit	576	376	408	203	330	187	181	96
3	3	Underfit	1	1	1	2	2	2	3	5
	50	Correctly specified	524	715	692	871	759	871	877	941
	0.2, 1.329	Overfit	475	284	307	127	239	127	120	54
4	3	Underfit	8	10	11	15	21	24	24	30
	100	Correctly specified	531	619	504	628	734	774	735	778
	1.0, 1.011	Overfit	461	371	485	357	245	202	241	192
5	3	Underfit	0	0	0	0	0	0	0	0
	100	Correctly specified	494	565	511	620	693	765	721	797
	0.5, 1.052	Overfit	506	435	489	380	307	235	279	203
6	3	Underfit	0	0	0	0	0	0	0	0
	100	Correctly specified	502	594	559	671	732	796	778	840
	0.2, 1.109	Overfit	498	406	441	329	268	204	222	160
7	3	Underfit	0	0	0	0	0	0	0	0
	200	Correctly specified	713	752	971	976	885	898	994	995
	1.0, 2.651	Overfit	287	248	29	24	115	102	6	5
8	3	Underfit	0	0	0	0	0	0	0	0
	200	Correctly specified	689	732	936	945	870	889	978	982
	0.5, 2.129	Overfit	311	268	64	55	130	111	22	18
9	3	Underfit	0	0	0	0	0	0	0	0
	200	Correctly specified	692	735	863	895	859	879	953	962
	0.2, 1.570	Overfit	308	265	137	105	141	121	47	38

The generating model is (6); n is the sample size, p_o is the true order, and α_o is the overdispersion parameter. m($\hat{c}$) is the average of 1000 variance inflation factors calculated using the largest NEGBIN1 candidate regression model. Order selections for candidate models of order 1 or 2 are reported in the rows labeled "Underfit"; selections for models of orders 4 through 10 are reported in the rows labeled "Overfit". Correct order selections (order 3) are reported in the rows labeled "Correctly specified"

The order selection results for the Poisson (3.2.1), NEGBIN1 (3.2.2), and NEGBIN2 (3.2.3) candidate families are featured in Tables 2, 3, and 4, respectively. As with the simulation results reported in Sect. 3.1, each quasi-likelihood criterion outperforms the corresponding criterion based on the ordinary likelihood. This tendency is especially evident in those sets where the degree of excess dispersion is pronounced. In addition, each J-divergence criterion obtains more correct selections than its I-divergence counterpart, and each "corrected" criterion outperforms its non-adjusted counterpart.

Table 4 Order selections

Set	p_o	Selections	Criterion							
	n α_o, $m(\hat{c})$		AIC	AICc	QAIC	QAICc	KIC	KICc	QKIC	QKICc
1	3	Underfit	38	60	55	91	82	120	116	160
	50	Correctly specified	587	724	662	786	754	789	774	793
	1.0, 1.184	Overfit	375	216	283	123	164	91	110	47
2	3	Underfit	3	6	7	19	14	24	27	37
	50	Correctly specified	629	786	736	876	827	889	876	917
	0.5, 1.244	Overfit	368	208	257	105	159	87	97	46
3	3	Underfit	1	1	1	3	2	2	5	7
	50	Correctly specified	607	761	728	861	790	894	874	934
	0.2, 1.283	Overfit	392	238	271	136	208	104	121	59
4	3	Underfit	0	0	0	0	0	0	0	0
	100	Correctly specified	675	761	703	794	864	893	874	908
	1.0, 1.095	Overfit	325	239	297	206	136	107	126	92
5	3	Underfit	0	0	0	0	0	0	0	0
	100	Correctly specified	664	749	719	791	851	886	879	903
	0.5, 1.095	Overfit	336	251	281	209	149	114	121	97
6	3	Underfit	0	0	0	0	0	0	0	0
	100	Correctly specified	670	741	725	805	845	889	885	917
	0.2, 1.119	Overfit	330	259	275	195	155	111	115	83
7	3	Underfit	0	0	0	0	0	0	0	0
	200	Correctly specified	706	735	713	750	870	893	878	895
	1.0, 1.017	Overfit	294	265	287	250	130	107	122	105
8	3	Underfit	0	0	0	0	0	0	0	0
	200	Correctly specified	700	733	709	752	860	882	867	888
	0.5, 1.037	Overfit	300	267	291	248	140	118	133	112
9	3	Underfit	0	0	0	0	0	0	0	0
	200	Correctly specified	715	761	737	784	894	909	899	918
	0.2, 1.046	Overfit	285	239	263	216	106	91	101	82

The generating model is (6); n is the sample size, p_o is the true order, and α_o is the overdispersion parameter. m($\hat{c}$) is the average of 1000 variance inflation factors calculated using the largest NEGBIN2 candidate regression model. Order selections for candidate models of order 1 or 2 are reported in the rows labeled "Underfit"; selections for models of orders 4 through 10 are reported in the rows labeled "Overfit". Correct order selections (order 3) are reported in the rows labeled "Correctly specified"

In the NEGBIN2 framework, where the variance structure is correctly specified, the selection results in Table 4 are comparable to those in Table 1. However, in the Poisson and NEGBIN1 frameworks, the frequencies of correct order selections in Table 2 are smaller than those reported in Table 1. With the Poisson simulations, the correct selection rates for the likelihood-based criteria are often quite poor. These patterns might be expected since the Poisson models ignore the overdispersion and the NEGBIN1 models mischaracterize it. In particular, with severe overdispersion, it is clear that Poisson regression models are inadequate. Yet even in such settings, the quasi-likelihood criteria still choose the correct variable sets more often than the other criteria. In general, over all of the sets, QKICc obtains the most correct selections, followed by QKIC and QAICc.

In the Poisson setting, a value of 1.0 for α_o often leads to mean values for the variance inflation factor $\hat{c}$ that exceed 4. From a practical perspective, such a large value for $\hat{c}$ indicates the need for a structural change in the model. In the NEGBIN1 and NEGBIN2 settings, the majority of the mean $\hat{c}$ values are very close to 1 and all are in the 1–4 range. Such values imply that the degree of excess dispersion is not too severe, and that the variance structure is acceptable. Of course, in the NEGBIN2 framework, the variance structure of the candidate family is properly specified, meaning that the values of $\hat{c}$ should be quite close to 1. Although the mean values of $\hat{c}$ are near 1 for the sets based on larger sample sizes ($n = 100, 200$), the mean values are slightly larger than 1 for the sets based on the smaller sample size ($n = 50$), reflecting a positive bias that dissipates as the sample size increases.

With smaller sample sizes, we note that the glm, gamlss, and glm.nb functions in R may exhibit convergence problems for certain samples. The use of standard nonlinear optimization algorithms for GLM maximum likelihood estimation is usually justified, yet issues may arise in simulation studies based on a large number of replications. A small percentage of generated samples led to fitted candidate models that did not converge. Such samples were detected and omitted from the compilation of the results. We note that the selection patterns of the criteria do not seem to be affected by the infrequent deletion of these samples.

4 Application

4.1 Introduction

The host-tick relationship is a complex interaction that can result in a differential distribution of ticks within a single host population. The factors contributing to variation in tick burden among individuals within a population are likely cryptic. Attempts to model these factors have elucidated some general trends, but have also failed to find the specific attribute or suite of attributes that cause the heterogeneities seen in many host-parasite systems (Brunner and Ostfeld 2008). Commonly examined host attributes potentially important in predicting tick burdens are host sex (Zuk and McKean 1996; Klein 2000), body mass (Mooring et al. 2000), and age (Schalk and Forbes 1997). Other abiotic factors such as seasonality (Krasnov et al. 2005) and meteorological variables (Harlan and Foster 1990) have also been shown to affect the host-tick relationship, likely through differences in life history strategies or behavior at different times of the year.

In Adair County, Missouri, the white-footed mouse (*Peromyscus leucopus*) and the immature stages (larval and nymphal) of the American dog tick (*Dermacentor variabilis*) form a host-specific relationship despite the presence of other potential host species (e.g. *Microtus ochrogaster, Cryptotis parva, Tamias striatus, Reithrodontomys megalotis*) (Kollars 1996). Within this host-parasite system, there is a large degree of heterogeneity in tick burdens on particular individuals. This may result in an increased probability of pathogen transmission (Woolhouse et al. 1997; Perkins et al. 2003). Additionally, differential parasitism can affect host population dynamics through mortality and selective predation of infested individuals (Anderson and May 1978). Despite some studies indicating that differential tick burdens did not negatively influence host fitness (Brunner and Ostfeld 2008), other studies point to the energetic cost of blood regeneration and increased likelihood for pathogen transmission as fitness costs (Musante et al. 2007; Perkins et al. 2003).

Determining the cause of the heterogeneities in tick burdens could yield predictive disease models and a better understanding of host-parasite interactions. The current study aims to examine the factors that influence larval *D. variabilis* burdens on *P. leucopus*. The effect of habitat, number of nymphal *D. variabilis* present, host body mass, host sex, tail length and right hind foot length will be examined using negative binomial regression. Candidate models will be evaluated based on the selection criteria previously introduced.

4.2 Methods

4.2.1 Study site

The current study was conducted at long term monitoring sites in Adair County, Missouri in two different habitats. The habitats consisted of an early successional forest and an old field dominated by nonnative grasses. These sites were located approximately 300 meters apart and were therefore exposed to similar abiotic conditions throughout the study period.

4.2.2 Trapping

Trapping in each site was performed on a permanent trapping grid, containing 104 sampling points. Each sampling point contained a Sherman live trap (H. B. Sherman, Tallahassee, Florida, USA) baited with a mixture of peanut butter and rolled oats. Trapping was performed concurrently on both grids every other month from June 2006 to December 2008 for approximately 9,500 trap nights. Trapping sessions consisted of 3-4 days of mark-recapture sampling with traps checked each morning. Each animal captured from both sites (X_1, 1: forest, 2: field) was subjected to determination of weight (X_2, gram), sex (X_3, 1: female, 2: male), right hind foot (X_4, millimeter) and tail length (X_5, millimeter), and removal of ticks. Ticks were stored in ethanol for later identification of life stage (X_6: the number of nymphal ticks). Animals were marked with a toe-clip number for the purpose of identifying recapture in accordance with the Animal Care and Use Committee standards of the American Society of Mammalogists (American Society of Mammalogists (1998)). From the total data, 172 first capture *P. leucopus* individuals were used for this analysis. The data is available at `http://hjkim.sites.truman.edu/research`.

Table 5 features basic descriptive statistics for the independent variables: specifically, proportions for habitat and gender, as well as means, variances, and pairwise correlation coefficients for the quantitative variables.

Table 5 Descriptive statistics for independent variables

Site	Forest (76.16 %)			Field (23.84 %)		
Sex	Female (42.44 %)			Male (57.56 %)		
	Mean	Standard deviation	Correlation coefficients			
			Weight	Foot length	Tail length	Nymphal ticks
Weight	20.312	4.0911		0.2408	0.5845	−0.0959
Foot length	20.5232	1.1210			0.1877	0.0127
Tail length	73.5174	7.0433				0.0621
Nymphal ticks	0.5523	1.3476				

4.2.3 Analysis

The analysis of the data proceeded by first selecting a suitable variance structure. Poisson, NEGBIN1, and NEGBIN2 regression models containing all 6 variables were fit to the data, and a distribution was chosen after an inspection of the values of the variance inflation factor $\hat{c}$ and AIC. With the chosen distribution, fitted candidate models based on all possible combinations of the 6 variables were compared relative to one another using values of AIC, AICc, QAIC, QAICc, KIC, KICc, QKIC, and QKICc. An optimal fitted model was determined based on the model selection criteria.

4.3 Model selection results

The average larval tick burden per *P. leucopus* individual was 3.24 and the variance was 106.06. As the variance is substantially larger than the mean, overdispersion may be present. The variance inflation factor $\hat{c}$ and AIC values were calculated for Poisson, NEGBIN1, and NEGBIN2 regression models with all 6 variables. The results are as follows.

Fitted model	Variance inflation factor ($\hat{c}$)	AIC
Poisson	20.55	1808.21
NEGBIN1	1.78	635.00
NEGBIN2	1.23	633.21

The variance inflation factor $\hat{c}$, reflecting the excess dispersion that is not accommodated by the variance structure, was 20.55 when fitting the largest Poisson regression model. When fitting the largest NEGBIN1 and NEGBIN2 regression models, the values of $\hat{c}$ were 1.78 and 1.23, respectively. These values of $\hat{c}$ suggest that excess dispersion exists under all three models, yet points to NEGBIN2 regression as the most suitable modeling framework. The NEGBIN2 model also has the smallest AIC value, 633.21, lending additional support to the NEGBIN2 framework. The propriety

Table 6 Best subsets among candidate models based on model selection criteria

Number of variables	Variable(s) included	Criterion							
		AIC	AICc	QAIC	QAICc	KIC	KICc	QKIC	QKICc
1	x_2	637.83	638.07	520.62	520.86	641.83	643.14	524.62	525.93
	x_3	639.61	639.85	522.07	522.31	643.61	644.93	526.07	527.39
2	x_2, x_6	634.68	635.04	518.43	518.79	639.68	641.15	523.43	524.90
	x_2, x_3	637.31	637.67	520.57	520.93	642.31	643.78	525.57	527.04
3	x_1, x_2, x_6	633.10	633.61	517.52	518.03	639.10	640.76	523.52	525.17
	x_2, x_3, x_6	635.54	636.05	519.50	520.01	641.54	643.20	525.50	527.16
4	x_1, x_2, x_4, x_6	633.01	633.69	517.81	518.50	640.01	641.88	524.81	526.69
	x_1, x_2, x_5, x_6	633.40	634.08	518.13	518.81	640.40	642.27	525.13	527.00
5	x_1, x_2, x_4, x_5, x_6	634.10	634.98	519.07	519.96	642.10	644.23	527.07	529.20
	x_1, x_2, x_3, x_4, x_6	634.20	635.08	519.16	520.04	642.20	644.33	527.07	529.20
6	$x_1, x_2, x_3, x_4, x_5, x_6$	635.21	636.32	520.35	521.46	644.21	646.62	529.35	531.76

of NEGBIN2 models has been extensively supported by analyses that attempt to characterize factors influencing parasite infestations (Brunner and Ostfeld 2008; Shaw et al. 1998).

Among the NEGBIN2 candidate models summarized in Table 6, the model containing habitat, host body mass, tail length, and the number of nymphal ticks had the lowest AIC value. The model containing habitat, host body mass, and the number of nymphal ticks had the lowest AICc, KIC, and KICc values. The model containing host body mass and nymphal load had the lowest values of QKIC and QKICc.

It is not clear which model is "best" since the criteria choose different candidate models. However, based on the sample size ($n = 172$) and the value of the variance inflation factor ($\hat{c} = 1.23$), the simulation results provide some guidance. In settings where the sample size is large and the degree of overdispersion is modest, AIC and AICc tend to choose over parameterized models. Also, the quasi-likelihood criteria outperform the ordinary likelihood criteria, and each J-divergence criterion outperforms its I-divergence counterpart. Based on these considerations, the model chosen by QKIC and QKICc based on host body mass (x_{2i}) and nymphal load (x_{6i}) might be considered as the optimal candidate. This fitted model can be written as

$$ln(\hat{\lambda}_i) = -1.5229 + 0.1154x_{2i} + 0.3027x_{6i}, \quad (7)$$

implying that there is positive relationship between larval *D. variabilis* burdens and the body mass of *P. leucopus* and between larval and nymphal *D. variabilis* burdens. For this model, the MLE of overdispersion parameter, $\hat{\alpha}$, is 4.78.

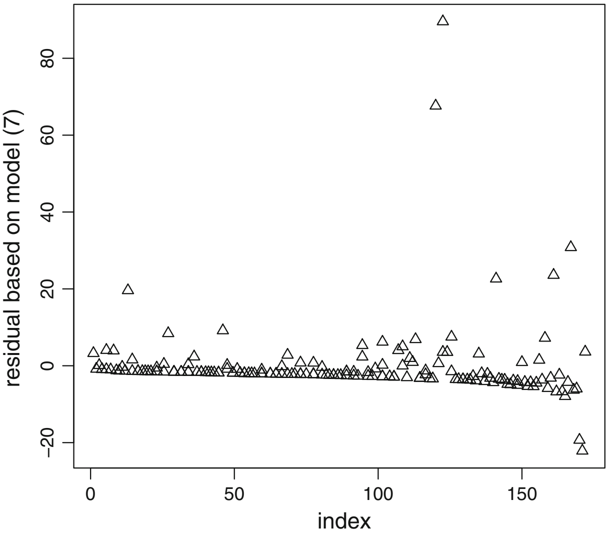

Fig. 1 Residuals versus index based on the size of the predicted value (model (7))

To illustrate the predictive properties of model (7), Figure 1 features a plot of the residuals versus an index based on the size of the predicted value: i.e., index 1 represents the smallest predicted value, index 2 represents the next largest predicted value, . . ., index *n* represents the largest predicted value. The figure illustrates a couple of important properties regarding the predictive efficacy of the fitted model. First, most of the residuals are small. In fact, 40% of the residuals are less than 2 in magnitude, 65% of the residuals are less than 3 in magnitude, and 85% of the residuals are less than 5 in magnitude. Second, the residuals tend to increase as the index for the predicted values increases, consistent with the notion of overdispersion. Thus, the model tends to predict the response more accurately when the response is small.

4.4 Discussion

Our study found evidence that larval *D. variabilis* burdens were strongly associated with host body mass. The positive relationship between larval tick burdens and body mass could be a function of increased surface area to mass ratio resulting in higher encounter probabilities as seen in South African ungulates (Gallivan and Horak 1997), differential energetic demands causing the hosts to forage more, and/or decreased grooming rates in larger individuals (Mooring et al. 2000).

Model selection criteria also suggested the positive relationship between larval and nymphal tick load. Nymphal ticks are larger and inflict a higher energetic cost to the host relative to larvae, as they take a larger bloodmeal. The presence of nymphal ticks could cause the host to enter a negative feedback cycle, in which it needs to forage more extensively to attempt to mount an immune response to the ticks present, but acquires more ticks through the extensive foraging.

The lack of the significance of host sex in our study is a surprising finding, as analyses based on data from the same location found a significant effect of sex in predicting tick burdens (Dallas et al. 2012). However, these analyses were based on conventional model selection criteria such as AIC, AICc, KIC, and KICc using larval and nymphal ticks combined. The final model based on the quasi-likelihood selection criteria QKIC and QKICc suggests that sex is not a major factor influencing the number of larval ticks. While sex biased parasitism is a common trend (Zuk and McKean 1996; Klein 2000), it is not ubiquitous (Brunner and Ostfeld 2008; Wilson et al. 2002). It is possible that larval ticks cannot discern between male and female hosts after host acquisition, as Dukes and Rodriquez (1976) suggest based on data relating to host odor. Additionally, larval ticks are opportunistic in their host selection, rarely moving more than a few meters away from the site of oviposition (Sonenshine 1991).

The non-significant effect of tail and right hind foot length is not surprising, as these measures are traditionally used to determine species and are not likely related to individual susceptibility to parasites.

5 Conclusion

Our results support three conclusions regarding model selection criteria based on Kullback information measures. First, when dealing with overdispersed count data, the quasi-likelihood criteria tend to outperform the ordinary likelihood criteria. Second, for the purpose of delineating between correctly specified and misspecified models, the symmetric divergence may serve as a more sensitive discrepancy measure than the directed divergence. Third, the performance of a criterion appears to be dictated by how well its penalty term corrects for the negative bias in the goodness-of-fit term.

The ambiguity of an optimal model is illustrated by differences in the models preferred by various selection criteria. This phenomenon occurs both in our simulation studies and in our application. Studies relying solely on a classical criterion such as AIC may therefore misinterpret the importance of the variables considered in candidate models.

In settings where overdispersion is suspected, formulating a family of candidate models based on a distribution that accommodates overdispersion may be a better analytical approach than formulating a family based on a conventional distribution (e.g., Poisson, binomial), and attempting to capture the excess variation via a simple variance inflation factor. The former setting was considered in the first simulation study (Table 1); the latter setting was considered in the first part of the second study (Table 2). The superiority of the selection results in the first study indicates that appropriate mean structures may be more easily identified when the analyst employs a distribution designed for overdispersed data. An extensive collection of simulation sets not featured here convey similar selection patterns.

Acknowledgments The authors wish to extend their appreciation to the anonymous reviewers for carefully reading the previous versions of this manuscript, and for providing constructive suggestions that served to improve the simulation study and the application.

References

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, Hoboken

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csáki F (eds) 2nd international symposium on information theory. Akadémia Kiadó, Budapest, Hungary, pp 267–281

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control AC 19:716–723

American Society of Mammalogists (1998) Guidelines for the capture, handling, and care of mammals as approved by the American Society of Mammalogists. J Mammal 79(4):1416–1431

Anderson DR, Burnham KP, White G (1994) Akaike information criterion model selection in overdispersed capture-recapture data. Ecol Soc Am 75:1780–1793

Anderson RM, May RM (1978) Regulation and stability of host-parasite population interactions. J Animal Ecol 47:219–267

Brunner JL, Ostfeld RS (2008) Multiple causes of variable tick burdens on small-mammal hosts. Ecology 89(8):2259–2272

Burnham KP, Anderson DR (2002) Model selection and multimodel inference. Springer, New York

Cavanaugh JE (1999) A large-sample model selection criterion based on Kullback's symmetric divergence. Stat Probab Lett 42:333–343

Cavanaugh JE (2004) Criteria for linear model selection based on Kullback's symmetric divergence. Aust N Z J Stat 46:257–274

Dallas T, Foré SA, Kim H-J (2012) Modeling the influence of *Peromyscus leucopus* body mass, sex, and habitat on immature *Dermacentor variabilis* burden. J Vector Ecol 37(2):338–341

Dukes JC, Rodriquez J (1976) A bioassay for host-seeking responses of tick nymphs (ixodidae). J Kansas Entomol Soc 49(4):562–566

Eberhardt LL (1978) Appraising variability in population studies. J Wildl Manag 42:207–238

Gallivan G, Horak I (1997) Body size and habitat as determinants of tick infestations of wild ungulates in south Africa. S Afr J Wildl Res 27(2):63–70

Harlan H, Foster WA (1990) Micrometeorologic factors affecting field host-seeking activity of adult Dermacentor variabilis (acari:ixodidae). J Med Entomol 27(4):471–479

Hilbe JM (2008) Negative binomial regression. Cambridge University Press, UK

Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika 76:297–307

Hurvich CM, Tsai CL (1995) Model selection for extended quasi-likelihood models in small samples. Biometrika 51:1077–1084

Hurvich CM, Shumway RH, Tsai CL (1990) Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples. Biometrika 77:297–307

Johnson J, Omland K (2004) Model selection in ecology and evolution. TRENDS Ecol Evol 19(2):101–108

Kim H-J, Cavanaugh JE (2005) Model selection criteria based on Kullback information measures for nonlinear regression. J Stat Plan Inference 134(2):332–349

Klein S (2000) The effects of hormones on sex differences in infection: from genes to behavior. Neurosci Biobehav Rev 24(6):627–638

Kollars T (1996) Interspecific differences between small mammals as hosts of immature Dermacentor variabilis (Acari: Ixodidae) and a model for detection of high risk areas of rocky mountain spotted fever. J Parasitol 82(5):707–710

Krasnov BR, Morand S, Hawlena H, Khokhlova IS, Shenbrot GI (2005) Sex-biased parasitism, seasonality, and sexual size dimorphism in desert rodents. Oecologia 146(2):209–217

Kullback S (1968) Information theory and statistics. Dover, Mineola

Lebreton J, Burnham K, Clobert J, Anderson D (1992) Model survival and testing biological hypotheses using marked animals: a unified approach with case studies. Ecol Monogr 62(1):67–118

Meinshausen N (2007) Relaxed Lasso. Comput Stat Data Anal 52:374–393

Min A, Holzmann H, Czado C (2010) Model selection strategies for identifying most relevant covariates in homoscedastic linear model. Comput Stat Data Anal 54(12):3194–3211

Mooring M, Benjamin J, Harte C, Herzog N (2000) Testing the interspecific body size principle in ungulates: the smaller they come, the harder they groom. Animal Behav 60(1):35–45

Musante AR, Pekins PJ, Scarpitti DL (2007) Metabolic impacts of winter tick infestations on calf moose. Alces 43:101–107

Perkins SE, Cattadori IM, Tagliapietra V, Rizzoli AP, Hudson PJ (2003) Empirical evidence for key hosts in persistence of a tick-borne disease. Int J Parasitol 33:909–917

R Development Core Team (2009) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria

Schalk G, Forbes MR (1997) Male biases in parasitism of mammals: effects of study type, host age, and parasite taxon. Oikos 78:67–79

Shaw D, Grenfell B, Dobson A (1998) Patterns of macroparasite aggregation in wildlife host populations. Parasitology 117:597–610

Sonenshine D (1991) The biology of ticks. Oxford University Press, New York

Sugiura N (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. Commun Stat A7:13–26

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc B 58:267–288

Wang H, Leng C (2008) A note on adaptive group Lasso. Comput Stat Data Anal 52:5277–5286

Wedderburn RWM (1974) Quasilikelihood functions, generalized linear models and the Gauss-Newton method. Biometrika 61:43–47

Wilson K, Bjornstad O, Dobson A, Merler S, Poglayen G, Randolph S, Read A, Skorping A (2002) Heterogeneities in macroparasite infections: patterns and processes. In: Hudson PJ, Rizzoli A, Grenfell BT, Heesterbeek H, Dobson AP (eds) The ecology of wildlife diseases. Oxford University Press, New York

Woolhouse M, Dye C, Etard J, Smith T, Charlwood J, Garnett G, Hagan P, Hii J, Ndhlovu P, Quinnell R, Watts C, Chandiwana S, Anderson R (1997) Heterogeneities in the transmission of infectious agents: implications for the design of control programs. Natl Acad Sci USA 94:338–342

Zuk M, McKean KA (1996) Sex differences in parasite infections: patterns and processes. Int J Parasitol 26(10):1009–1024

Author Biographies

Hyun-Joo Kim received a Ph.D. degree in Statistics from the University of Missouri-Columbia in 2000. In 2000, she joined the Department of Mathematics and Computer Science at Truman State University, Missouri. She is currently a full professor of Statistics at the Department of Statistics. In 2010, she founded the Center for Applied Statistics and Evaluation and has been the director of the center. Her current research interests include environmental statistics, model selection, and statistical consulting.

Joseph E. Cavanaugh received B.S. degrees in Computer Science and Mathematics from Montana Tech in 1986, an M.S. degree in Statistics from Montana State University in 1988, and a Ph.D. degree in Statistics from the University of California, Davis, in 1993. From 1993 to 2003, he was on the faculty of the Department of Statistics at the University of Missouri—Columbia. He is currently Professor and Director of Graduate Studies in the Department of Biostatistics at the University of Iowa. He holds a secondary appointment in the Department of Statistics and Actuarial Science. His methodological research interests include model selection and time series analysis.

Tad A. Dallas graduated from Truman State University, Missouri, in 2010 with his M.Sc. in biology. In 2011, he worked in a United States Department of Agriculture laboratory in Fort Pierce, Florida, examining pathogens of citrus crops. Since 2011, he has been working towards his Ph.D. at the University in Georgia's Odum School of Ecology in the lab of Dr. John Drake, where his research interests focus on host-pathogen interactions in a population and community context.

Stephanie A. Foré graduated from Miami University, Ohio, USA with a Ph.D. in Botany in 1991. She is currently a full professor of biology at Truman State University, Missouri, USA. Her current research is focused on the ecology of ticks.