Ecology, 96(9), 2015, pp. 2370–2382

Model averaging and muddled multimodel inferences

Brian S. Cade[1]

U.S. Geological Survey, 2150 Centre Avenue, Building C, Fort Collins, Colorado 80526 USA

Abstract. Three flawed practices associated with model averaging coefficients for predictor variables in regression models commonly occur when making multimodel inferences in analyses of ecological data. Model-averaged regression coefficients based on Akaike information criterion (AIC) weights have been recommended for addressing model uncertainty but they are not valid, interpretable estimates of partial effects for individual predictors when there is multicollinearity among the predictor variables. Multicollinearity implies that the scaling of units in the denominators of the regression coefficients may change across models such that neither the parameters nor their estimates have common scales, therefore averaging them makes no sense. The associated sums of AIC model weights recommended to assess relative importance of individual predictors are really a measure of relative importance of models, with little information about contributions by individual predictors compared to other measures of relative importance based on effects size or variance reduction. Sometimes the model-averaged regression coefficients for predictor variables are incorrectly used to make model-averaged predictions of the response variable when the models are not linear in the parameters. I demonstrate the issues with the first two practices using the college grade point average example extensively analyzed by Burnham and Anderson. I show how partial standard deviations of the predictor variables can be used to detect changing scales of their estimates with multicollinearity. Standardizing estimates based on partial standard deviations for their variables can be used to make the scaling of the estimates commensurate across models, a necessary but not sufficient condition for model averaging of the estimates to be sensible. A unimodal distribution of estimates and valid interpretation of individual parameters are additional requisite conditions. The standardized estimates or equivalently the t statistics on unstandardized estimates also can be used to provide more informative measures of relative importance than sums of AIC weights. Finally, I illustrate how seriously compromised statistical interpretations and predictions can be for all three of these flawed practices by critiquing their use in a recent species distribution modeling technique developed for predicting Greater Sage-Grouse (*Centrocercus urophasianus*) distribution in Colorado, USA. These model averaging issues are common in other ecological literature and ought to be discontinued if we are to make effective scientific contributions to ecological knowledge and conservation of natural resources.

Key words: generalized linear models; Greater Sage-Grouse; model averaging; multicollinearity; multimodel inference; partial effects; partial standard deviations; regression coefficients; relative importance of predictors; species distribution models; zero-truncated Poisson regression.

Introduction

Considering multiple statistical models and incorporating model uncertainty into analyses has gained considerable traction in biological and ecological literature since the seminal work of Burnham and Anderson (1998, 2002). There is a growing recognition that we should not put too much analytical faith in any single model given the presence of reasonable competing models (the multiple working hypotheses of Chamberlin [1890]). The approach to incorporating estimates from multiple candidate models into multimodel inferences advocated by Burnham and Anderson (1998, 2002) is to summarize the information by model averaging, weighting model estimates with Akaike information criterion (AIC) weights. This approach is easily accomplished for most regression models and there have been numerous reviews discussing this approach (e.g., Johnson and Omland 2004, Hobbs and Hilborn 2006, Burnham et al. 2011, Grueber et al. 2011). However, these reviews and syntheses fail to mention that one form of model averaging advocated by Burnham and Anderson (2002, 2004), for individual regression coefficients (parameter estimates for predictor variables), can be seriously flawed whenever there is multicollinearity among the predictor variables in the candidate models, rendering this procedure of limited practical utility.

Multicollinearity among predictor variables (defined as any linear relationship among them) is a common condition for observational studies (Graham 2003,

Manuscript received 28 August 2014; revised 25 February 2015; accepted 3 March 2015. Corresponding Editor: B. D. Inouye.

[1] E-mail: cadeb@usgs.gov

Dormann et al. 2013) and, thus, this issue is pervasive whenever model-averaged regression coefficients have been used for model interpretations or predictions. I will demonstrate that multicollinearity among the predictor variables implies that the scaling of the units for the regression coefficients for a given predictor variable (X_i) may not be constant among candidate models with different combinations of predictors. Thus, regression coefficients as a measure of the change in the response per unit change in the predictor variable ($\Delta y/\Delta X_i$) cannot legitimately be averaged because a unit change in the predictor variable does not necessarily mean the same thing across all candidate models. I will demonstrate the nature of this issue using an example data set extensively analyzed by Burnham and Anderson (2002:224–238) and suggest the genesis of a possible solution based on standardization by partial standard deviations of predictors. Two related procedures will also be addressed: determining relative importance of predictors from accumulated AIC model weights, and computing model-averaged predictions of the response variable from model-averaged regression coefficients in nonlinear models. Finally, I will discuss an example of the flawed use of model-averaged regression coefficients and related procedures in an analysis for a species distribution model developed for Greater Sage-Grouse (*Centrocercus urophasianus*; see Plate 1) in Colorado (Rice et al. 2013) to demonstrate that these are not minor, esoteric statistical issues but major issues that can have profound impacts on ecological interpretations and the utility of statistical relationships.

While I use Rice et al. (2013) as a recent and particularly transparent ecological example of the perils of model averaging regression coefficients, many other publications have similar flaws. The extent to which this flawed model averaging occurs in recent ecological literature is indicated by a list of publications provided in Appendix A (not intended to be an exhaustive list) that used model-averaged regression coefficients. The unfortunate consequence of using these approaches is that we are often left with a poorer understanding of important statistical relationships that might strongly impact our ecological understanding and conservation decisions while unreasonably feeling satisfied that model uncertainty has been addressed. All literature published using any of these procedures associated with model averaging regression coefficients has unreliable statistical results. It is impossible to know whether the faulty model averaging practices had minor or major consequences for interpretations of modeled relationships without examining the individual model results and data.

Model Averaging

The model averaging approach to incorporate model uncertainty into inferences formulated by Burnham and Anderson (1998, 2002) relies on weighting various model estimates with AIC weights obtained across a set of candidate models. The issues I discuss here apply equally to the large-sample AIC $= -2\log L(\hat{\theta} \mid g_j, \text{data}) + 2K$ and the small-sample adjusted version $\text{AIC}_c = \text{AIC} + (2K(K+1))/(n-K-1)$, where g_j is the jth model in a set of candidate models, n is sample size, and K is the number of parameters estimated in model g_j (Burnham and Anderson 2002, Lukacs et al. 2010). I am considering typical regression applications including least squares mean regression ($\boldsymbol{\theta} = E[\mathbf{y} \mid \mathbf{X}] = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$), quantile regression ($\boldsymbol{\theta} = Q_y(\tau \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}(\tau)$), generalized linear models ($\boldsymbol{\theta} = E[\mathbf{y} \mid \mathbf{X}] = h^{-1}(\mathbf{X}\boldsymbol{\beta})$, where h is a link function), and their mixed-effects counterparts; where $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times p$ matrix of predictor variables, $\boldsymbol{\beta}$ is a $p \times 1$ matrix of parameters ($p \leq K$ and p may include an intercept term), and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors with variance σ^2 when required by the model form. Estimates of the K parameters are maximum likelihood (L) estimates. Differences in AIC (or AIC_c) between models are the currency for this multimodel inference approach, where $\Delta\text{AIC}_j = \text{AIC}_j - \min \text{AIC}$, and the AIC weights $w_j = (\exp(-0.5\Delta\text{AIC}_j))/(\sum_{i=1}^{R} \exp(-0.5\Delta\text{AIC}_i))$. Note that while I have followed the convention of scaling the ΔAIC_j as differences from the model with the minimum AIC, it is possible to scale them from any model desired that aids interpretations, e.g., a null model with just an intercept term.

Two related forms of model-averaged parameter estimates can be obtained using the AIC weights. The first involves averaging across the estimated response parameter $\hat{\theta}_j$ derived from a function of the entire regression model, all its estimated parameters, and some data, where the model-averaged parameter estimates across g_j models are

$$\hat{\bar{\boldsymbol{\theta}}} = \sum_{j=1}^{R} w_j \hat{\boldsymbol{\theta}}_j \tag{1}$$

i.e., the weighted-average of the estimates is the model-averaged estimate. This appears to be a reasonable approach for incorporating model uncertainty into estimated parameters that are derived from an entire model and that have the same units and interpretations across all models, e.g., the predicted mean response $\hat{\theta}_j = \hat{E}_j[y \mid \mathbf{X}, g_j]$ in a regression model, since AIC and the derived weights apply to an entire model.

The second form of model averaging used for constructing averages across estimates of parameters for individual predictor variables within models, i.e., regression coefficients that are partial rates of change, is

$$\hat{\bar{\beta}}_i = \frac{\sum_{j=1}^{R} w_j I_i(g_j) \hat{\beta}_{ij}}{w_+(i)} \tag{2}$$

where $\hat{\beta}_{ij}$ denotes the estimate for the ith of p parameters in model g_j, $I_i(g_j)$ is an indicator function taking the value 1 if predictor X_i is in model g_j (0 otherwise), and

$w_+(i)$ is the sum of AIC weights for all models in the set where the parameter for ith predictor variable is estimated (Burnham and Anderson 2002, 2004, Lukacs et al. 2010). The purpose of this model averaging is to address the uncertainty associated with estimates for individual predictor variables that change when they are combined with other predictors in different models. This model averaging of parameter estimates for predictor variables also has been modified to a sort of shrinkage estimate by averaging across all models in the candidate set for all predictors, where estimates are forced to zero for parameters not explicitly included in a particular model (Burnham and Anderson 2002, Lukacs et al. 2010).

Similar multicollinearity issues apply to either form of model averaging regression coefficients. The model averaging of regression coefficients ends up being done across estimates ($\hat{\beta}_i = \Delta y/\Delta X_i$) without common denominators and is nonsensical because a unit change in the predictor variable (ΔX_i) is not the same across all models. This issue was noted previously by Candolo et al. (2003) for AIC model averaging. Draper (1999) similarly criticized Bayesian model averaging of regression parameter estimates. I will elaborate on the roots of this issue and the genesis of a solution in the next section.

One justification sometimes offered for ignoring that the individual regression coefficients have units and interpretations that are not identical across the R candidate models when using AIC model-averaged estimates is that they provide a sort of shrinkage estimate (Burnham and Anderson 2002, Lukacs et al. 2010, Giudice et al. 2012). The focus is really on estimates for the full set of predictor variables in the model and the degree that they shrink towards zero, similar to justifications provided by Hoeting et al. (1999) for Bayesian model averaging of individual coefficients. Shrinkage estimators allow more uncertain estimates of parameters for the full set of predictors that are less strongly related to the response to shrink towards zero, thereby reducing model complexity and providing better predictions to new samples (Copas 1983, Tibshirani 1996, Harrell 2001). Regularization procedures for shrinkage estimates incorporate a penalty term on the estimation function (e.g., penalized likelihood methods) that will force estimates for weakly supported predictors towards zero. But AIC model-averaged estimates for regression coefficients are not obtained by the simultaneous estimation of all parameters and their shrinkage such that the estimates and their standard errors correctly reflect the conditional nature of their partial effects associated with multicollinearity. Given the availability of valid statistical procedures for obtaining shrinkage estimates, there is little reason to rely on AIC model-averaged regression coefficients to provide something similar to shrinkage estimates (Burnham and Anderson 2002, Lukacs et al. 2010, Giudice et al. 2012), especially given that these model-averaged estimates have no defined units in the presence of multicollinearity.

Multicollinearity and Model-Averaged Regression Coefficients

The parameter estimates for a given predictor variable can have different units and interpretations among the R candidate models as they are rates of change in the responses (e.g., E[y]) given a unit change in the predictor conditional on what other predictors are in the model. These parameter estimates as rates of change are not guaranteed to have the same scaling of units or interpretations across models with different combinations of predictor variables unless all of them are uncorrelated with each other, an unlikely situation outside of a well-controlled, balanced, randomized experiment. It is common to see parameter estimates for a predictor that change in sign or order of magnitude depending on what other predictors and covariance structure are in the candidate models (e.g., Neter et al. 1996:291). That this is due to changing multicollinearity is easily demonstrated by algebraic manipulation of an individual regression coefficient in a multiple regression model.

The Frisch-Waugh theorem (Frisch and Waugh 1933) as generalized by Lovell (1963) tells us that parameters and estimates for any single predictor variable in a multiple linear regression model can be obtained by partialing out the effects of all other predictors by linear projections and then estimating a simple regression model. So the linear mean regression model $E[\mathbf{y} \mid \mathbf{X}] = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ can be partitioned into $E[\mathbf{y} \mid \mathbf{X}] = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_C\boldsymbol{\beta}_C + \boldsymbol{\varepsilon}$, where $\mathbf{X}_1$ is an $n \times 1$ matrix for an individual predictor, $\boldsymbol{\beta}_1$ is the 1×1 parameter vector for this predictor, $\mathbf{X}_C$ is an $n \times (p-1)$ matrix of the additional predictors, and $\boldsymbol{\beta}_C$ is the $(p-1) \times 1$ vector of parameters for the additional predictors, and $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ are dimensioned as before for the full model. Extending notation to incorporate estimation across $j = 1$ to r candidate models with potentially different numbers and combination of predictors in $\mathbf{X}_{Cj}$ and their parameters $\boldsymbol{\beta}_{Cj}$, the parameter $\boldsymbol{\beta}_1$ (or its estimate) for the single predictor $\mathbf{X}_1$ can be expressed in several equivalent forms

$$\begin{aligned}\beta_{1j} &= (\mathbf{X}_1'\mathbf{M}_{Cj}'\mathbf{M}_{Cj}\mathbf{X}_1)^{-1}\,\mathbf{X}_1'\mathbf{M}_{Cj}'\mathbf{M}_{Cj}\mathbf{y} \\ &= \mathrm{Cov}(\mathbf{M}_{Cj}\mathbf{X}_1, \mathbf{y})/\mathrm{Var}(\mathbf{M}_{Cj}\mathbf{X}_1) \\ &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\,\mathbf{X}_1'\mathbf{X}_{Cj}\beta_{Cj} \qquad (3)\end{aligned}$$

where $\mathbf{M}_{Cj}$ is the residualizing linear projection matrix $\mathbf{I} - \mathbf{X}_{Cj}\,(\mathbf{X}_{Cj}'\mathbf{X}_{Cj})^{-1}\,\mathbf{X}_{Cj}'$ for the additional predictor variables in the jth model, $\mathbf{I}$ is the $n \times n$ identity matrix, and Cov and Var denote the covariance and variance functions, respectively. Note that $\mathbf{M}_{Cj}\mathbf{X}_1$ are residuals denoting the part of $\mathbf{X}_1$ not linearly related to $\mathbf{X}_{Cj}$. Anytime the linear relation between $\mathbf{X}_1$ and $\mathbf{X}_{Cj}$ is not constant across the $j = 1$ to r models, the estimates $\hat{\beta}_{1j}$ and parameters $\boldsymbol{\beta}_{1j}$ will be based on changing

TABLE 1. Correlation matrix of the first-year college grade point average (GPA) example ($n = 20$) from Burnham and Anderson (2002:225–238) with four predictors; Scholastic Aptitude Test (SAT) math score, SAT verbal score, high school math GPA, and high school (HS) English GPA.

Predictor	GPA (y)	SAT math (X_1)	SAT verbal (X_2)	HS math (X_3)	HS English (X_4)
GPA (y)	1.000	0.850	0.653	0.695	0.606
SAT math (X_1)		1.000	0.456	0.559	0.663
SAT verbal (X_2)			1.000	0.434	0.417
HS math (X_3)				1.000	0.272
HS English (X_4)					1.000

denominators, $(\mathbf{X}_1'\mathbf{M}_{Cj}'\mathbf{M}_{Cj}\mathbf{X}_1)^{-1}$ or $\text{Var}(\mathbf{M}_{Cj}\mathbf{X}_1)$ in Eq. 3, precluding sensible averaging of them because the unit scales are not the same. Eq. 3 also indicates that the only time these estimates are guaranteed to have the same denominator is when there is no linear relation between $\mathbf{X}_1$ and $\mathbf{X}_{Cj}$. The bottom equality of Eq. 3 implies that the parameter or its estimate is equivalent to the simple regression on $\mathbf{X}_1$, i.e., $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$, when there is no multicollinearity, so there is no model uncertainty in the regression coefficients to estimate regardless of what other predictors are included in $\mathbf{X}_{Cj}$. Although I have demonstrated this issue with the linear mean regression model because the algebraic manipulation is straightforward, similar principles apply to other forms of linear models.

To demonstrate impacts on estimates and model-averaged estimates for predictor variables due to changes in the scaling of predictors under different model covariance structures, I will use the college grade point average (GPA) example data analyzed by Burnham and Anderson (2002:225–238). Statistical code is provided in Supplement 1. The response variable in their least squares regression model, first year college GPA (y), is positively correlated with four predictor variables; math score on the Scholastic Aptitude Test (SAT) (X_1), verbal score on the SAT (X_2), high school math GPA (X_3), and high school English GPA (X_4). These predictors have low to moderate positive correlations with each other (Table 1). Estimates for each predictor variable in the 8 of 15 candidate models that included the variable and their model-averaged estimates are provided in Tables 2–5, and correspond to values provided by Burnham and Anderson (2002:229–230, Table 5.14). It is apparent from the discussion of their analyses of these data and related simulations that Burnham and Anderson (2002:225–238) recognized that most of the variation in parameters and estimates for predictors among the 15 possible candidate models was due to multicollinearity, even though multicollinearity was low as indicated by variance inflation factors (VIF) that only ranged 1.08–2.49 across the candidate models. What they failed to address was that the numerical differences in these parameters and estimates were thoroughly confounded with changes in their scaling, i.e., the denominators $[\text{Var}(\mathbf{M}_{Cj}\mathbf{X}_1)]$ were not constant across the $j = 15$ models. Their AIC model averaging of regression coefficients acts as if they are just numbers without any units attached to them.

An easy way to visualize the change in scaling of the estimates is to view them as partial regression plots (Neter et al. 1996:361–368), which are equivalent to plotting $\mathbf{M}_{Cj}\mathbf{y}$ vs. $\mathbf{M}_{Cj}\mathbf{X}_1$. The estimate for the partial effect ($\hat{\boldsymbol{\beta}}_1$) of $\mathbf{X}_1$ then is the rate of change in the part of the response not linearly related to the other predictors ($\mathbf{M}_{Cj}\mathbf{y}$) for a unit change in the part of $\mathbf{X}_1$ not linearly related to the other predictors ($\mathbf{M}_{Cj}\mathbf{X}_1$). The scale of $\mathbf{M}_{Cj}\mathbf{X}_1$ will be compressed compared to the scale of the original $\mathbf{X}_1$ depending on the degree of multicollinearity with the other $p - 1$ predictor variables and, thus, a unit change is not the same quantity for $\mathbf{X}_1$ and $\mathbf{M}_{Cj}\mathbf{X}_1$.

Partial regression plots for X_1 (SAT math score) in the college GPA example for two models, model 15 that included all four predictors and model 11 that included X_2 and X_3, indicate that the residualized predictor X_1 has less variability under model 15 than model 11 (Fig. 1). The change in scaling of any given predictor ($\text{Var}(\mathbf{M}_{Cj}\mathbf{X}_1)$, which are partial variances) under different model covariance structures can be computed directly by matrix algebra or more conveniently by calculating partial standard deviations (Bring 1994), where the partial standard deviation for the ith of p predictors in model j is given by $s_{ij}^* = s_i \text{VIF}_{ij}^{-0.5}([n-1]/[n-p])^{0.5}$, s_i is the sample standard deviation of the ith predictor X_i and VIF_{ij} is the variance inflation factor for the ith predictor in model j. The variance inflation factor $\text{VIF}_{ij} = 1/(1 - R^2_{p-1,\,j})$, where $R^2_{p-1,\,j}$ is the coefficient of determination of the regression of the ith predictor in model j on the remaining $p - 1$ predictors in model j. The inverse of the variance inflation factor, VIF_{ij}^{-1}, then is the proportion of the variance in the ith predictor of model j that is not linearly related to the remaining $p - 1$ predictors. The variance inflation factor equals one when there is no correlation among predictors and, thus, the partial standard deviation of predictor X_{ij} equals the standard deviation of X_i when it is the single predictor in model j. Variance inflation factors are readily computed for individual continuous predictors in linear or generalized linear models and also have been generalized to

TABLE 2. Parameter estimates $\hat{\beta}_{1j}$ (where j is model number) in the eight models that included X_1, mathematics score on the SAT, in the college grade point average example ($n = 20$) of Burnham and Anderson (2002:229, Table 5.14).

Model j	Predictors used	Unstandardized estimates		w_j	VIF_{1j}^{-1}	Partial SD	Standardized estimates	
		$\hat{\beta}_{1j}$	$\widehat{\text{SE}}(\hat{\beta}_{1j} \mid g_j)$			X_{1j}	$\hat{\beta}^*_{1j}$	$\widehat{\text{SE}}(\hat{\beta}^*_{1j} \mid g_j)$
11	X_1, X_2, X_3	0.002185	0.0004553	0.454	0.6314	139.6188	0.305049	0.0635716
5	X_1, X_2	0.002606	0.0004432	0.269	0.7921	151.9735	0.395981	0.0673588
6	X_1, X_3	0.002510	0.0004992	0.103	0.6874	141.5742	0.355294	0.0706686
15	X_1, X_2, X_3, X_4	0.002010	0.0005844	0.066	0.4022	114.8592	0.230886	0.0671286
12	X_1, X_2, X_4	0.002586	0.0005631	0.044	0.5213	126.8592	0.328084	0.0714370
13	X_1, X_3, X_4	0.002129	0.0006533	0.028	0.4055	111.8821	0.238237	0.0730926
1	X_1	0.003178	0.0004652	0.027	1.0000	166.2004	0.528232	0.0773137
8	X_1, X_4	0.002987	0.0006357	0.006	0.5603	127.8154	0.381811	0.0812455
$\bar{\hat{\beta}}_1$†		0.002368	0.0005350	0.997‡			0.335535	0.0851482

Notes: The models, predictors used, unstandardized estimates $\hat{\beta}_{1j}$ and their standard errors, Akaike information criterion (AIC) weights w_j, and model-averaged estimates, $\bar{\hat{\beta}}_1$ and $\widehat{\text{SE}}(\bar{\hat{\beta}}_1)$), correspond to those in Table 5.14 of Burnham and Anderson (2002). Also provided are the inverse of the variance inflation factors VIF_{1j}^{-1}, partial standard deviations of X_{1j}, parameter estimates $\hat{\beta}^*_{1j}$ standardized by their partial standard deviations and their standard errors, and the model-averaged standardized parameter estimate and its standard error. The variable g_j is the jth model in a set of candidate models.

† SE in this row are $\widehat{\text{SE}}(\bar{\hat{\beta}}_1)$.

‡ The parameter in this cell is $w_+(1)$, the sum of AIC weights for all models in the set where the parameter for the $i = 1$ predictor variable was estimated.

categorical predictors and ≥2 continuous predictors (Fox and Monette 1992).

Inverses of the variance inflation factors and partial standard deviations for all four predictors in the college GPA models are provided in Tables 2–5. The partial standard deviations for X_1 for the models shown in Fig. 1 are 114.9 SAT math scores for model 15 and 139.6 SAT math scores for model 11, and the standard deviation of X_1 is 166.2 (Table 2). Thus, equivalent rates of change in X_1 in both models would imply a parameter $\beta_{1,11}$ in model 11 that is 82.3% (114.9/139.6) of the parameter $\beta_{1,15}$ in model 15 simply because of the scaling changes due to multicollinearity. There is considerably more variation in the partial standard deviations of X_1 among models (Table 2) than for any of the other three predictors (Tables 3–5), especially for models with greater AIC_c weights that contribute most to the model-averaged estimates. Thus, averaging across estimates $\hat{\beta}_{1j}$ for X_1 is likely to be much more misleading because of changes in the scaling of units than is averaging estimates for the other predictors. The real variation in rates of change in $\mathbf{y}$ for a given predictor $\mathbf{X}_1$ (slopes) across models that have different combinations of multicollinear predictors are thoroughly confounded with the fixed scaling changes in the different models [$\text{Var}(\mathbf{M}_{Cj}\mathbf{X}_1)$ is not constant]. It is impossible to interpret the model-averaged regression coefficients (Tables 2–5) in terms of a $\Delta y/\Delta X_i$ because we do not know what units should apply to the denominator because it no longer refers to any specific covariance structure among the predictor variables. Because all the simulations in Burnham and Anderson (2002) for this and similar

TABLE 3. Parameter estimates $\hat{\beta}_{2j}$ in the eight models that included X_2, verbal score on the SAT, in the college grade point average example ($n = 20$) of Burnham and Anderson (2002:229, Table 5.14).

Model j	Predictors used	Unstandardized estimates		w_j	VIF_{2j}^{-1}	Partial SD	Standardized estimates	
		$\hat{\beta}_{2j}$	$\widehat{\text{SE}}(\hat{\beta}_{2j} \mid g_j)$			X_{2j}	$\hat{\beta}^*_{2j}$	$\widehat{\text{SE}}(\hat{\beta}^*_{2j} \mid g_j)$
11	X_1, X_2, X_3	0.001312	0.0005252	0.454	0.7453	121.0400	0.158844	0.0635716
5	X_1, X_2	0.001574	0.0005555	0.269	0.7921	121.2645	0.190888	0.0673588
14	X_2, X_3, X_4	0.001423	0.0007113	0.002	0.7151	118.5605	0.168662	0.0843321
15	X_1, X_2, X_3, X_4	0.001252	0.0005515	0.066	0.7094	121.7154	0.152412	0.0671286
12	X_1, X_2, X_4	0.001568	0.0005811	0.044	0.7687	122.9238	0.192778	0.0714370
7	X_2, X_3	0.002032	0.0007627	<0.001	0.8115	122.7352	0.249355	0.0936057
9	X_2, X_4	0.002273	0.0008280	<0.001	0.8263	123.8504	0.281459	0.1025484
2	X_2	0.003063	0.0008367	<0.001	1.0000	132.6165	0.406203	0.1109608
$\bar{\hat{\beta}}_2$†		0.001405	0.0005558	0.835‡			0.170502	0.0674792

Notes: The models, predictors used, unstandardized estimates $\hat{\beta}_{2j}$ and their standard errors, AIC weights w_j, and model-averaged estimates, $\bar{\hat{\beta}}_2$ and $\widehat{\text{SE}}(\bar{\hat{\beta}}_2)$, correspond to those in Table 5.14 of Burnham and Anderson (2002). Also provided are the inverse of the variance inflation factors VIF_{2j}^{-1}, partial standard deviations of X_{2j}, parameter estimates $\hat{\beta}^*_{2j}$ standardized by their partial standard deviations and their standard errors, and the model-averaged standardized parameter estimate and its standard error.

† SE in this row are $\widehat{\text{SE}}(\bar{\hat{\beta}}_2)$.

‡ The parameter in this cell is $w_+(2)$, the sum of AIC weights for all models in the set where the parameter for the $i = 2$ predictor variable was estimated.

TABLE 4. Parameter estimates $\hat{\beta}_{3j}$ in the eight models that included X_3, high school math grade point average, in the college grade point average example ($n = 20$) of Burnham and Anderson (2002:230, Table 5.14).

		Unstandardized estimates				Partial SD	Standardized estimates	
Model j	Predictors used	$\hat{\beta}_{3j}$	$\widehat{SE}(\hat{\beta}_{3j} \mid g_j)$	w_j	VIF_{3j}^{-1}	X_{3j}	$\hat{\beta}^*_{3j}$	$\widehat{SE}(\hat{\beta}^*_{3j} \mid g_j)$
11	X_1, X_2, X_3	0.1799	0.0877	0.454	0.6468	0.725052	0.130415	0.0635716
7	X_2, X_3	0.3694	0.1186	<0.001	0.8115	0.789211	0.291512	0.0936057
6	X_1, X_3	0.2331	0.0973	0.103	0.6874	0.726397	0.169328	0.0706686
15	X_1, X_2, X_3, X_4	0.1894	0.0919	0.066	0.6183	0.730707	0.138425	0.0671286
3	X_3	0.5066	0.1236	<0.001	1.0000	0.852750	0.431974	0.1054146
13	X_1, X_3, X_4	0.2474	0.0990	0.028	0.6701	0.737961	0.182578	0.0730926
14	X_2, X_3, X_4	0.3405	0.1045	0.002	0.8014	0.807047	0.274774	0.0843321
10	X_3, X_4	0.4171	0.1054	0.001	0.9259	0.843055	0.351652	0.0888936
$\bar{\hat{\beta}}_3$†		0.1930	0.0932	0.654‡			0.140389	0.0679552

Notes: The models, predictors used, unstandardized estimates $\hat{\beta}_{3j}$ and their standard errors, AIC weights w_j, and model-averaged estimates, $\bar{\hat{\beta}}_3$ and $\widehat{SE}(\bar{\hat{\beta}}_3)$, correspond to those in Table 5.14 of Burnham and Anderson (2002). Also provided are the inverse of the variance inflation factors VIF_{3j}^{-1}, partial standard deviations of X_{3j}, parameter estimates $\hat{\beta}^*_{3j}$ standardized by their partial standard deviations and their standard errors, and the model-averaged standardized parameter estimate and its standard error.

† SE in this row are $\widehat{SE}(\bar{\hat{\beta}}_3)$.

‡ The parameter in this cell is $w_+(3)$, the sum of AIC weights for all models in the set where the parameter for the $i = 3$ predictor variable was estimated.

regression examples rely on averaging across coefficient estimates with different units because the denominators differ, the statistical performance suggested by distributions of their simulated model-averaged estimates is of questionable merit. Similar concerns apply to simulations done by Freckleton (2011) for evaluating the bias of model-averaged coefficients with increasing multicollinearity among predictors. Performance evaluations by Lukacs et al. (2010) used simulations for completely uncorrelated predictors but these are not relevant to the more common situation of multicollinear predictors. Indeed, Lukacs et al. (2010) mentioned in their discussion that it is problematic to extend model averaging to multicollinear predictors.

One possible way forward for model averaging regression coefficients in the presence of changing multicollinearity among models is to equate the scaling of parameter estimates for predictors by standardizing estimates such that $\text{Var}(\mathbf{M}_{Cj}\mathbf{X}_1)$ is constant across the j candidate models prior to model averaging them. We can easily standardize them to $\text{Var}(\mathbf{M}_{Cj}\mathbf{X}_1) = 1.0$ across the j models by computing the partial standard deviations associated with the covariance structure of the predictors included in the models because $s^{*2}_{1j} = \text{Var}(\mathbf{M}_{Cj}\mathbf{X}_1)$. This is identical to the standardization approach recommended by Bring (1994) for comparing different predictors within a single model. In the linear model this can be accomplished by either transforming the predictors prior to obtaining estimates, $X^*_{ij} = X_{ij}/s^*_{ij}$, or more conveniently by transforming the estimates after the models are estimated, $\hat{\beta}^*_{ij} = \hat{\beta}_{ij}s^*_{ij}$. The first method should be used for generalized linear models with nonlinear link functions (e.g., logistic and Poisson regression). These standardized estimates now provide a rate of change per one partial standard deviation of the predictor given the covariance with whatever other

TABLE 5. Parameter estimates $\hat{\beta}_{4j}$ in the eight models that included X_4, high school English grade point average, in the college grade point average example ($n = 20$) of Burnham and Anderson (2002:230, Table 5.14).

		Unstandardized estimates				Partial SD	Standardized estimates	
Model j	Predictors used	$\hat{\beta}_{4j}$	$\widehat{SE}(\hat{\beta}_{4j} \mid g_j)$	w_j	VIF_{4j}^{-1}	X_{4j}	$\hat{\beta}^*_{4j}$	$\widehat{SE}(\hat{\beta}^*_{4j} \mid g_j)$
15	X_1, X_2, X_3, X_4	0.0876	0.1765	0.066	0.5198	0.380340	0.033304	0.0671286
12	X_1, X_2, X_4	0.0112	0.1893	0.044	0.5438	0.377397	0.004206	0.0714370
13	X_1, X_3, X_4	0.1756	0.1932	0.028	0.5462	0.378230	0.066426	0.0730926
8	X_1, X_4	0.0989	0.2182	0.006	0.5603	0.372302	0.036831	0.0812456
14	X_2, X_3, X_4	0.4533	0.1824	0.002	0.8160	0.462327	0.209586	0.0843321
10	X_3, X_4	0.5790	0.1857	0.001	0.9259	0.478607	0.277123	0.0888936
9	X_2, X_4	0.5195	0.2207	<0.001	0.8263	0.452110	0.234859	0.1025484
4	X_4	0.7790	0.2407	<0.001	1.0000	0.484110	0.377103	0.1165274
$\bar{\hat{\beta}}_4$†		0.0902	0.1989	0.147‡			0.034757	0.0760897

Notes: The models, predictors used, unstandardized estimates $\hat{\beta}_{4j}$ and their standard errors, AIC weights w_j, and model-averaged estimates, $\bar{\hat{\beta}}_4$ and $\widehat{SE}(\bar{\hat{\beta}}_4)$, correspond to those in Table 5.14 of Burnham and Anderson (2002). Also provided are the inverse of the variance inflation factors VIF_{4j}^{-1}, partial standard deviations of X_{4j}, parameter estimates $\hat{\beta}^*_{4j}$ standardized by their partial standard deviations and their standard errors, and the model-averaged standardized parameter estimate and its standard error.

† SE in this row are $\widehat{SE}(\bar{\hat{\beta}}_4)$.

‡ The parameter in this cell is $w_+(4)$, the sum of AIC weights for all models in the set where the parameter for the $i = 4$ predictor variable was estimated.

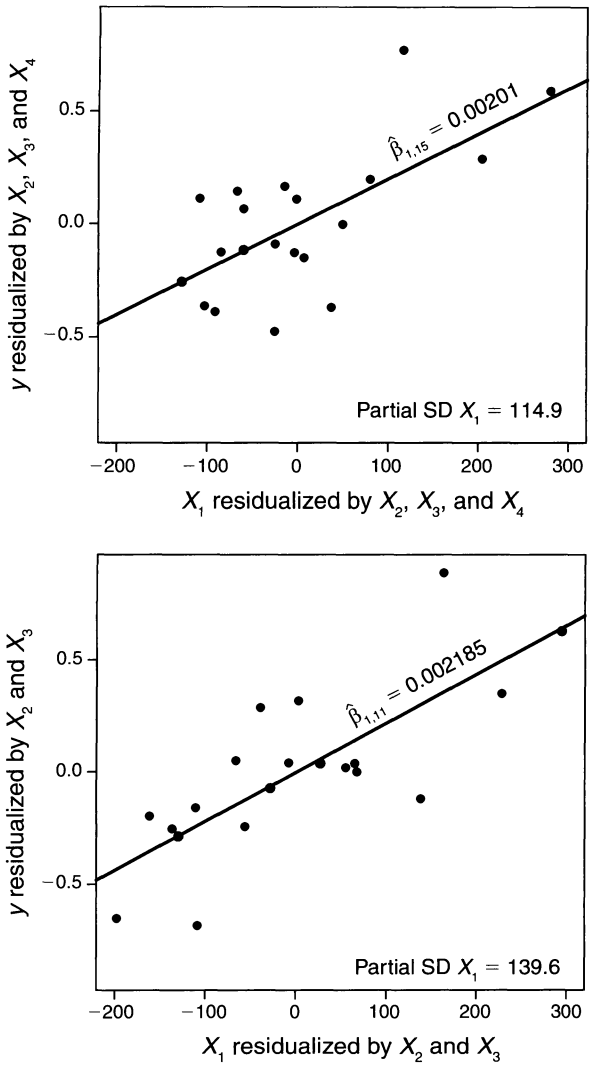

FIG. 1. Partial regression plots (n = 20) for estimated parameters $\hat{\beta}_{1j}$ (where j is model number) for X_1 in models 15 (top) and 11 (bottom) for the college grade point average (GPA) example of Burnham and Anderson (2002). Note that the partial standard deviation for residualized X_1 in model 11 (with X_2 and X_3) is greater than for the residualized X_1 in model 15 (with X_2, X_3, and X_4). The response (y) is first year college GPA, X_1 is Scholastic Aptitude test (SAT) math score, X_2 is SAT verbal score, X_3 is high school math GPA, and X_4 is high school English GPA. Line slopes are reported on the line.

variables were included in the model. It often is useful for numerical stability or ease of interpretation to also center predictor variables to have mean = 0 so that the regression intercept corresponds to the mean of **X**. This also can be done when standardizing by partial standard deviations, but is not required to deal with the changing scales associated with multicollinearity among predictor variables.

I provide standardized estimates for the college GPA example by model and their model-averaged estimates and standard errors (Tables 2–5) obtained by using Eq. 2 and variance computations in Burnham and Anderson (2002: 180), where standardized estimates are substituted for the usual unstandardized estimates (Supplement 1). The model-averaged standardized estimates indicated there was one-third of a unit change in college GPA (0.3355) for SAT math scores, one-half (0.1705) that much change for SAT verbal scores, less than one-half that (0.1404) much change for high school math GPA, and one-tenth (0.0348) that much change with high school English GPA when expressed as partial standard deviations of the corresponding predictors. The coefficient of variation of the model-averaged standardized estimate for SAT math scores was 12% greater than the model-averaged unstandardized estimate (Table 2), whereas the coefficients of variation of model-averaged standardized and unstandardized estimates for the other three predictors (Tables 3–5) were nearly identical (ratios of 0.993–1.002). This is consistent with the greater variation in partial standard deviations of SAT math scores among models compared to the variation in partial standard deviations for the other three predictors.

We also can standardize estimates so that $\text{Var}(\mathbf{M}_{Cj}\mathbf{X}_1) = s_{1j}^{*2}$ for partial standard deviations of the predictor variables associated with a model having a specific covariance structure, providing model-averaged standardized estimates with units scaled to that covariance structure. A logical covariance structure to use is the one associated with all predictor variables in the full model, which is model 15 for the college GPA example (Tables 2–5). These standardized estimates are $\hat{\beta}_{ij}^{*} = \hat{\beta}_{ij}(s_{ij}^{*}/s_{i,15}^{*})$. This yields model-averaged estimates of $\bar{\hat{\beta}}_1 = 0.0029$ [$\widehat{\text{SE}}(\bar{\hat{\beta}}_1) = 0.00074$], $\bar{\hat{\beta}}_2 = 0.0014$ [$\widehat{\text{SE}}(\bar{\hat{\beta}}_2) = 0.00055$], $\bar{\hat{\beta}}_3 = 0.1921$ [$\widehat{\text{SE}}(\bar{\hat{\beta}}_3) = 0.09299$], and $\bar{\hat{\beta}}_4 = 0.0914$ [$\widehat{\text{SE}}(\bar{\hat{\beta}}_4) = 0.20006$]. Only the model-averaged standardized estimate for predictor X_1 (SAT math scores) differed substantially (23% greater slope estimate and 38% greater standard error) from the model-averaged unstandardized estimates (Tables 2–5) given by Burnham and Anderson (2002, Table 5.14). Again, this is consistent with the greater impacts of changing multicollinearity on partial standard deviations of SAT math scores compared to the other predictors. More importantly there are well defined units associated with the model-averaged standardized estimates that don't exist for the model-averaged unstandardized estimates. Here they are rates associated with a one unit change in the part of the predictor that is not linearly related to the other three predictors.

It is important to emphasize that equating scales (common denominators) of regression coefficient estimates is a necessary condition for model averaging to be sensible but other conditions need to be met too. The distribution of the standardized estimates should be unimodal so that an average of them coupled with its standard deviation is an informative description of the distribution of the individual model estimates. Individual regression coefficients also must provide interpretable rates of change in the response **y** in the context of

the candidate model structure. This precludes sensible model averaging of individual regression coefficients in most models with interactions or polynomial structures for the predictors that require the simultaneous interpretation of two or more regression coefficients. These two additional conditions were satisfied for the college grade point average example but will often not be met in many ecological analyses.

AIC Weights and Relative Importance of Predictors

The sum of AIC (or AIC_c) model weights, w_j, across candidate models often used with the model-averaged regression coefficients provides a relatively uninformative assessment of the relative importance of predictors. The AIC model weights apply to an entire model and not any individual predictor variables within that model and, thus, have minimal information content about contributions of individual predictor variables to predicted responses, the objective function minimized in model estimation, or predictor effect size. Ratios of AIC weights for the jth model compared to the model with highest weight, $w_j/\max(w_j)$, are equivalent to the inverse of evidence ratios (Burnham and Anderson 2002:77–79). These ratios are logically interpreted as relative importance of models, indicating the proportionate reduction in likelihoods between the best model (relative importance of 1.0) and the jth model when adjusted for number of parameters. The poor performance of the sum of AIC weights as a measure of relative importance for individual predictors found by Murray and Conner (2009) and Galipaud et al. (2014) can be directly traced to their property of indicating relative importance of models. Simulations by Doherty et al. (2012) found variable importance approached 1.0 for all predictors with increasing sample size regardless of whether or not they were related to the true modeled relationship. This also is a logical consequence of the sum of AIC weights being a measure of relative importance of models rather than of individual predictors. Having balanced subsets of candidate models with and without a given predictor (Burnham and Anderson 2002, Doherty et al. 2012) has no relevance to this fundamental issue. The simulations conducted by Burnham and Anderson (2002:227) suggest that at best the sum of AIC weights as a measure of relative importance may indicate the proportion of times a given predictor would be selected in repeated random sampling, i.e., how many times it occurs in models but no information about its relative contributions in any model.

Relative importance is a slippery concept with many interpretations, but an interpretation based on proportion of times a predictor is selected for a model is of limited utility. More useful interpretations for predictor variables within a given regression model typically are related to contributions to reducing the objective function used in statistical estimation and expected change in the response variable given a unit change in the predictor (Bring 1994). In essence, the relative importance of individual predictors across models should involve the relative importance within models, e.g., how much it contributes to the likelihood that is maximized or the equivalent minimization of the objective function relative to the other predictors in that model. Kruskal and Majors (1989) and MacNally (2000) provide some overview of the issues related to relative importance of individual predictors.

There are existing procedures to compute relative importance of individual predictors within a model based on ratios of parameter estimates for standardized predictors (based on partial standard deviations) or equivalently the ratio of t statistics (estimate divided by its standard error) for unstandardized predictors (Bring 1994); variance decomposition (Grömping 2007); and more involved hierarchical partitioning approaches to variance decomposition (Chevan and Sutherland 1991, Christensen 1992). The ratio of absolute values of standardized predictors based on setting partial standard deviations equal to one or the equivalent ratio of absolute values of t statistics for unstandardized predictors within individual candidate models can readily be incorporated into a model averaging scheme, where these ratios are weighted by the AIC model weights identical to Eq. 2 for parameter estimates (Supplement 1). These ratios within a model are scaled relative to a maximum of 1.0 for the strongest predictor within a model and then those ratios are substituted for the parameter estimates in Eq. 2 to obtain model-averaged estimates. Alternatively, one could just take the ratios of the absolute value of the model-averaged standardized estimates for predictors.

Model-averaged estimates of relative importance for the four predictors in the college GPA example made by either using ratios of absolute values of t statistics, $|t_{ij}|$, for models with unstandardized estimates or equivalently ratios of absolute values of standardized estimates, $|\hat{\beta}^*_{ij}|$, provided relative importance values of 1.00 for X_1 (SAT math score), 0.52 for X_2 (SAT verbal score), 0.47 for X_3 (high school math GPA), and 0.14 for X_4 (high school English GPA). Simply taking the ratios of absolute values of model-averaged standardized estimates from Tables 2–5 provided model-averaged relative importance values of 1.00 for X_1, 0.51 for X_2, 0.42 for X_3, and 0.10 for X_4. Both of these measures of relative importance related to effect sizes and variance reduction indicated that X_2 had half the effect, X_3 had slightly less than half the effect, and X_4 had around 10% of the effect compared to the most important predictor X_1. Relative importance found by summing AIC_c weights (Tables 2–5) ranked the predictors similarly to the previous procedures; 1.00 for X_1, 0.83 for X_2, 0.65 for X_3, and 0.15 for X_4 (Burnham and Anderson 2002:227). These values indicated far greater relative importance for X_2 and X_3 based solely on weights of models in which they occurred, even though the estimates were small and contributed little to maximizing the likelihoods in those models. The former measures of relative importance

based on standardized effects size and variance reduction are more useful than the latter measure based on proportion of times a variable is included in models and more consistent with an intuitive interpretation of what is meant by relative importance.

Model-Averaged Regression Coefficients and Predictions

Model-averaged estimates of parameters for the response of a regression model cannot be correctly obtained from model-averaged estimates of parameters for the predictors in the regression model if the model is not linear in the predictors, i.e., $\hat{\bar{E}}[y \mid \mathbf{X}] = \hat{\bar{\beta}}_0 + \hat{\bar{\beta}}_1 X_1 + \hat{\bar{\beta}}_2 X_2 + \ldots + \hat{\bar{\beta}}_k X_k$ only if $E[\mathrm{y} \mid \mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$. This computational convenience for models that are linear in the parameters for the predictors may be the only real value of the simple model-averaged regression coefficients. But predicted responses based on the average of the R set of regression coefficients for predictors (incorrect) will not be the same as the average of the predicted responses for the R candidate models (correct) for nonlinear models, including generalized linear models with nonlinear link functions, e.g., logistic regression as used in species occupancy modeling. This is a logical consequence of Jensen's inequality (Jensen 1906). The magnitude of the differences between the correct and incorrect model-averaged predictions will depend on the magnitude of the predicted responses, magnitude and variation of parameter estimates across models and their interaction with AIC weights, and nonlinear model forms actually being used.

Example Issues in a Sage-Grouse Distribution Model

I illustrate the severity of these issues of model averaging and flawed interpretations with a recent species distribution modeling approach by Rice et al. (2013). Species distribution models as large-scale extensions of habitat or resource selection models are widely used as a tool to aid conservation and land-use planning decisions (Elith and Leathwick 2009, Franklin 2009, 2013). Rice et al. (2013) proposed a novel approach for species distribution modeling that estimates large-scale (1-km^2 unit of resolution) changes in the number of locations of radio-marked animals as a function of landscape cover types across multiple studies in a large geographic extent of a species range. Their example application was for Greater Sage-Grouse (*Centrocercus urophasianus*) in northwestern Colorado, a species of conservation concern because of historical declines in populations and range contraction (Schroeder et al. 2004), coupled with increasing human development and activities that are potential stressors (Knick and Connelly 2011). The Rice et al. (2103) approach to species distribution modeling was based on using zero-truncated Poisson regressions on telemetry location information collected on individual Greater Sage-Grouse. The statistical units for analysis were mapped 1-km^2 grid cells across the range of Greater Sage-Grouse in northwestern Colorado; counts ($\geq$1) of telemetry locations for individual grouse were the response variable (y); and the proportions of area of selected land cover types within the 1-km^2 grid cells were the potential predictor variables ($\mathbf{X}$). Identifications of individual telemetered Sage-Grouse were incorporated in the model as random-effects on the intercept. Hundreds to thousands of candidate models based on combinations of up to 12 predictor variables were estimated for breeding, summer, or winter season models. Model-averaged predictions, $\hat{\bar{E}}[y \mid X]$, were based on model-averaged parameter estimates, $\hat{\bar{\beta}}_i$, using $\mathrm{AIC_c}$ model weights, and relative importance of predictors in the three seasonal models were also based on accumulating $\mathrm{AIC_c}$ model weights across the multiple models (2049 for breeding, 257 for summer, and 513 for winter season models).

Multicollinearity and model-averaged regression coefficients for predictors

The issues with model-averaged parameter estimates are particularly transparent in the Rice et al. (2013) analyses because the 12 candidate predictor variables used (the small proportion of urban cover type was excluded) were proportions of different cover types in fixed areas (1-km^2 mapped grid cells) and, thus, form a multi-part composition (Aitchison 1986) with components that sum to a constant quantity (1.0 in this case). These compositional predictors have an inherent negative covariance structure (Aitchison 1986, Aitchison and Egozcue 2005). Because of this strong multicollinearity among predictors, the model-averaged estimates for predictors across the R candidate models (Rice et al. 2013; Tables 3, 5, and 6) are unreliable for interpreting partial effects associated with the cover types. The impact of the negative covariance structure is readily apparent in the model-averaged estimates; e.g., for the breeding season model (Rice et al. 2013; Table 3) they ranged from -5.45 to -1.20 per unit standard deviation of proportions for cover types that are not Sage-Grouse breeding habitat and ranged from 0.31 to 2.14 per unit standard deviation for proportions of cover types that are Sage-Grouse breeding habitat. Although Rice et al. (2013) standardized their predictors using sample standard deviations of the cover type proportions, this standardization does not eliminate the scaling issues with multicollinear predictors similar to standardization by partial standard deviations, nor does it eliminate the negative covariance structure and interpretation issues for compositional predictors. A linear combination of the habitat subset (e.g., proportions of sagebrush, salt desert shrub, shrubland, mountain shrub, grassland, agriculture, riparian, pinyon juniper) of cover types must be inversely related with a linear combination of the non-habitat subset (e.g., proportions of alpine, forest, bareground, forest shrub) simply based on the compositional structure of these proportions of fixed areas. It is mathematically impossible to have Sage-Grouse counts increase with one subset without them having a corresponding decrease with the complemen-

PLATE 1. Male Greater Sage-Grouse (*Centrocercus urophasianus*) on a lek in open sagebrush steppe typically used as breeding habitat. Photo credit: Tatiana Gettelman.

tary subset. Thus, we might expect the magnitude of the parameter estimates obtained for candidate models that only included cover types from the habitat subset to be much greater than when candidate models included cover types from both the habitat and non-habitat subsets of predictors. This would occur because the redundant subset of predictors are inversely related, and ultimately would result in attenuation of model-averaged parameter estimates, potentially confounding the effects of important covariates. An example simulation of compositional predictors for zero-truncated count regression models similar to the Rice et al. (2013) data for breeding Sage-Grouse that demonstrates these properties is included in Appendix B and Supplement 2.

Even my suggested approach of standardizing predictors by their partial standard deviations was not effective at dealing with the multicollinearity induced by compositional predictors. The redundant subsets of linear combinations of predictor variables forces even the standardized estimates to have a multimodal distribution of individual estimates across models with different combinations of predictors such that their average is uninformative (Table B2). Furthermore, because the compositional predictors are inherently ratios of a part to the sum of the parts, it is not obvious that interpretations of individual regression coefficients for a single part (individual predictor) are sensible in the additive model. Indeed, this has been the motivation for the development of specialized transformations for compositional predictors based on the log ratio transformation (Aitchison 1986, Aitchison and Egozcue 2005) such as the isometric log ratio (Hron et al. 2012). This log ratio transformation approach creates orthonormalized predictors that are log-contrasts between parts of the composition that correctly allow estimated increases in mean counts of Sage-Grouse locations with increasing proportions of habitat cover types and the corresponding decreasing proportions of the complementary non-habitat cover types. Dealing with zero proportions can be problematic with the log ratio

approach but is not insurmountable (Aitchison and Egozcue 2005).

AIC weights and relative importance of predictors

We do not know which cover types were more strongly related to the predicted distribution of Sage-Grouse based on Rice et al. (2013) use of AIC_c weights to compute relative importance. For example, they found that sagebrush and mountain shrub cover types had a relative importance of 1.0 for their breeding season model but for sagebrush this simply reflects the fact that they forced it into all candidate models. The implication here is that mountain shrub and sagebrush had equivalent relative importance simply because they both occurred in all models with great AIC weight. Riparian (0.74), other shrublands (0.86), and grassland (0.86) had lower relative importance based on their sum of AIC_c weights. However, if Rice et al. (2013) had used an alternative measure of relative importance reflecting effects size such as ratios of t statistics for unstandardized predictors or ratios of standardized predictors (based on partial standard deviations), then it still would have been possible to provide useful measures of relative importance for all the predictors even when all models were forced to include the sagebrush predictor. Some indication of what alternative measures of relative importance might indicate for predictors in the breeding season model (Rice et al. 2013, Table 3) can be obtained by looking at ratios of the model-averaged estimates divided by their standard errors (t statistics), suggesting that sagebrush ($t = 0.80$) had lower relative importance compared to riparian ($t = 1.28$), other shrublands ($t = 1.30$), grassland ($t = 1.34$), or mountain shrub ($t = 1.47$). A definitive answer requires new analyses with estimates that better address the multicollinearity of the compositional predictors.

Model-averaged regression coefficients and model-averaged predictions

The model-averaged estimates for predictors in Rice et al. (2013, Tables 3, 5, and 6) should not have been used for computing model-averaged predictions of mean counts, $\hat{E}[y \mid X]$, because the count model used by Rice et al. (2013) was not linear in the parameters for the predictors. The zero-truncated Poisson regression model used was $E[y \mid \mathbf{X}] = \exp(\mathbf{X\beta})/(1 - \exp(-\exp(\mathbf{X\beta})))$, where $\exp(\mathbf{X\beta})$ is the mean λ of a Poisson distribution as a nonlinear function of the predictors. Even if a model is a linear combination of parameters in the logarithmic scale, it still would be incorrect to make predictions in the logarithmic scale based on the average of the R set of parameter estimates for the predictors and then back-transform (exponentiate) those averaged predictions to the count scale. The predictions are for mean counts and means are not equivariant to nonlinear transformations like the logarithmic, i.e., $\exp[\text{mean}(\log(\text{count}_i))] \neq \text{mean}[\exp(\log(\text{count}_i))]$. Correctly making model-averaged predictions, $\hat{E}[y \mid \mathbf{X}]$, for the R candidate model sets in this case requires averaging across R sets of predictions as in Eq. 1. The actual difference between the correct and incorrect model-averaged predictions may be small in some instances, but why estimate them incorrectly given the chance that they might deviate greatly from correct estimates? Appendix B and Supplement 2 provide an example when the correct and incorrect model-averaged predictions are similar and when they differ.

Rice et al. (2013) compounded their mistake in estimating model-averaged estimates of mean counts of telemetry locations by using them to make predictions for the probability of any count ≥ 1, i.e., probability of occupancy. They apparently made these predictions by using the relationship between predicted means from a zero-truncated Poisson distribution and predicted means from a conventional Poisson distribution, which allowed for computing probabilities of counts ≥ 1 (W. Thogmartin, *personal communication*). This transformation makes the predictions extremely sensitive to unverifiable assumptions about the proportion of zeros associated with means of the Poisson distribution for areas with very low to no Sage-Grouse use (mean counts of locations <2). Rates of change in the predicted probabilities of occupancy also are greatest in those regions of low to no Sage-Grouse use, which are inherently extrapolations outside of the majority of the predictor sample space because samples only occurred where there were ≥ 1 grouse locations (Appendix B: Fig. B1).

Conclusion

The issue of model uncertainty in regression coefficients is inherently an issue of multicollinearity. There is no model uncertainty in parameters for predictor variables when they are all uncorrelated as they have the same value regardless of which combinations are included in a model. Any variation in estimates among models with uncorrelated predictors (e.g., as in Lukacs et al. 2010) is just usual sampling variation not model uncertainty. Uncorrelated predictor variables are unlikely in most observational studies, although they might be obtained by transformations, e.g., with orthonormalizing log ratios for compositional variables or by principal components. Because multicollinearity implies there is different scaling of units for the regression coefficients of a given predictor across candidate models with different combinations of predictors, some method for removing the scaling differences across models is required for averaging to provide sensible summaries for multimodel inference. The use of partial standard deviations for predictors is one viable method for standardization that will work for some predictor covariance structures as demonstrated with the college GPA example of Burnham and Anderson (2002). Other useful standardization approaches may exist and more investigation is required to establish the statistical performance of these model-averaged standardized

estimates before they are recommended for routine use. The choice of an appropriate covariance structure among predictor variables to be used for standardizing will be critical. The simple averaging of regression coefficients recommended by Burnham and Anderson (2002, 2004) that ignores the multicollinear covariance structures should be discontinued immediately. Multicollinearity, of course, is a source of related issues (e.g., increased sampling variability of estimates) in individual regression models as summarized in Graham (2003) and Dormann et al. (2013).

The use of estimates for predictors that are standardized by their partial standard deviations (or the equivalent t statistics of unstandardized estimates) directly translates into estimates of relative importance that incorporate information on relative effect sizes and variance reduction within models that can be averaged across all candidate models. While certainly not the only possible measure of relative importance, this measure is likely to be far more useful to most analysts than the sum of AIC weights measure of relative importance that does not address the contributions of individual predictors within any of the candidate models. Furthermore, it is a measure of relative importance that is useful regardless of any constraints imposed on how many models include specific subsets of predictors.

There are many models, perhaps the majority, where standardization will not eliminate interpretation issues for model-averaged regression coefficients. The complex negative covariance structure associated with compositional predictors as in the Rice et al. (2013) analyses resulted in a multimodal distribution of standardized estimates such that their average was of low information content. Averaging individual estimates when simultaneous interpretation of multiple estimates is required to make sensible interpretation of the effects of a predictor variable also will not be useful, e.g., models that have various combinations of polynomial terms for predictors (Blums et al. 2005) or interactions among predictors. A desire to address model uncertainty is commendable, but some aspects of interpreting multiple candidate models cannot be effectively reduced to simplistic statistical summaries such as averages. For many covariate structures the effect of model uncertainty on parameter estimates for the predictor variables may only be interpreted sensibly by looking at the individual estimates across the multiple candidate models. Alternatively, model averaging the predicted responses (Eq. 1) for different combinations of values for predictor variables can be used to indirectly explore model relationships. The issues with model averaging discrete pieces of a model are a lot like the issues with interpreting main effects in an analysis of variance that includes interactions: while it may not be too misleading in some limited circumstances, it is not a good thing to do in general.

This information is intended to help curb the overzealous use of simplistic AIC-based model averaging of regression coefficients, treating it as a panacea for dealing with uncertainty and inference across multiple models. It should be considered along with other recent concerns about AIC and the information-theoretic approach. The initial enthusiasm for the information-theoretic approach as an alternative paradigm for statistical inference is now being tempered by increased recognition of its direct relationship with other likelihood based statistics (Aho et al. 2014, Murtaugh 2014*a*, *b*, Spanos 2014). The currency of the information-theoretic approach is differences in AIC between two models, ΔAIC, which fundamentally is based on the same likelihood ratios associated with likelihood ratio hypothesis tests, confidence intervals based on inversions of those tests, and coefficients of determination (or partial coefficients of determination). These statistics just represent different scaling of the same basic information, likelihood ratios between two models. The different scales can be useful for different purposes, but it seems unreasonable to expect the scaling associated with ΔAIC and the information-theoretic approach to be a fundamentally superior paradigm for improved inferential insights. If model averaging for multimodel inferences is to be one of the principal advantages of the AIC information-theoretic approach to data analysis as Burnham and Anderson (2014) suggested, it must be done in a more enlightened fashion than currently employed, including only averaging estimates with comparable units and interpretations.

Acknowledgments

My thoughts on the modeling issues discussed here benefited from discussions with C. Aldridge, G. Auble, P. Dixon, and D. Draper. The manuscript benefitted from reviews of earlier drafts provided by C. Aldridge, G. Auble, P. Dixon, D. Draper, J. Fieberg, J. Giudice, S. Lele, T. Shaffer, T. Stanley, D. Stauffer, and P. Stevens. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Literature Cited

Aho, K., D. Derryberry, and T. Peterson. 2014. Model selection for ecologists: the worldviews of AIC and BIC. Ecology 95: 631–636.

Aitchison, J. 1986. The statistical analysis of compositional data. Chapman and Hall, London, UK.

Aitchison, J., and J. J. Egozcue. 2005. Compositional data analysis: Where are we and where should we be heading? Mathematical Geology 37:829–850.

Blums, P., J. D. Nichols, J. E. Hines, M. S. Lindberg, and A. Mednis. 2005. Individual quality, survival variation and patterns of phenotypic selection on body condition and timing of nesting in birds. Oecologia 143:365–376.

Bring, J. 1994. How to standardize regression coefficients. American Statistician 48:209–213.

Burnham, K. P., and D. R. Anderson. 1998. Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York, New York, USA.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Springer-Verlag, New York, New York, USA.

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference: understanding AIC and BIC in model selection. Sociological Methods and Research 33:261–304.

Burnham, K. P., and D. R. Anderson. 2014. *P* values are only an index to evidence: 20th- vs. 21st-century statistical science. Ecology 95:627–630.

Burnham, K. P., D. R. Anderson, and K. P. Huyvaert. 2011. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behavioral Ecology and Sociobiology 65:23–35.

Candolo, C., A. C. Davison, and C. G. B. Demetrio. 2003. A note on model uncertainty in linear regression. Statistician 52 (Part 2):165–177.

Chamberlin, T. C. 1890. The method of multiple working hypotheses. Science 15:92–96 (reprinted 1965 Science 148: 754–759).

Chevan, A., and M. Sutherland. 1991. Hierarchical partitioning. American Statistician 45:90–96.

Christensen, R. 1992. Comments on Chevan and Sutherland. American Statistician 46:74.

Copas, J. B. 1983. Regression, prediction, and shrinkage. Journal of the Royal Statistical Society, Series B (Methodological) 45:311–354.

Doherty, P. F., G. C. White, and K. P. Burnham. 2012. Comparison of model building and selection strategies. Journal of Ornithology 152:S317–S323.

Dormann, C. F., et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36:27–46.

Draper, D. 1999. Comment—Bayesian model averaging: a tutorial. Statistical Science 14:405–409.

Elith, J., and J. R. Leathwick. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40:677–697.

Fox, J., and G. Monette. 1992. Generalized collinearity diagnostics. Journal of the American Statistical Association 87:178–183.

Franklin, J. 2009. Mapping species distributions: spatial inference and prediction. Cambridge University Press, Cambridge, UK.

Franklin, J. 2013. Species distribution models in conservation biogeography: developments and challenges. Diversity and Distributions 19:1217–1223.

Freckleton, R. P. 2011. Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. Behavioral Ecology and Sociobiology 64: 91–101.

Frisch, R., and F. V. Waugh. 1933. Partial time regressions as compared with individual trends. Econometrica 1:387–401.

Galipaud, M., M. A. F. Gillingham, M. David, and F.-X. Dechaume-Moncharmont. 2014. Ecologists overestimate the importance of predictor variables in model averaging: a plea for cautious interpretations. Methods in Ecology and Evolution 5:983–991.

Giudice, J. H., J. R. Fieberg, and M. S. Lenarz. 2012. Spending degrees of freedom in a poor economy: a case study of building a sightability model for moose in northeastern Minnesota. Journal of Wildlife Management 76:75–87.

Graham, M. H. 2003. Confronting multicollinearity in ecological multiple regression. Ecology 84:2809–2815.

Grömping, U. 2007. Estimators of relative importance in linear regression based on variance decomposition. American Statistician 61:139–147.

Grueber, C. E., S. Nakagawa, R. J. Laws, and I. G. Jamieson. 2011. Multimodel inference in ecology and evolution: challenges and solutions. Journal of Evolutionary Biology 24:699–711.

Harrell, F. E., Jr. 2001. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer-Verlag, New York, New York, USA.

Hobbs, N. T., and R. Hilborn. 2006. Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. Ecological Applications 16:5–19.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Rejoinder—Bayesian model averaging: a tutorial. Statistical Science 14:412–417.

Hron, K., P. Filzmoser, and K. Thompson. 2012. Linear regression with compositional explanatory variables. Journal of Applied Statistics 39:1115–1126.

Jensen, J. L. W. V. 1906. Sur les fonctions convexes et les inéagalitiés entre les valeurs moyennes. Acta Mathematica 30:175–193.

Johnson, J., and K. Omland. 2004. Model selection in ecology and evolution. Trends in Ecology and Evolution 19:101–108.

Knick, S. T., and J. W. Connelly. 2011. Greater sage-grouse and sagebrush: an introduction to the landscape. Pages 1–12 *in* S. T. Knick and J. W. Connelly, editors. Greater sage-grouse ecology and conservation of a landscape species and its habitats. Studies in avian biology. Volume 38. University of California Press, Berkley, California, USA.

Kruskal, W., and R. Majors. 1989. Concepts of relative importance in recent scientific literature. American Statistician 43:2–6.

Lovell, M. C. 1963. Seasonal adjustment of economic time series and multiple regression analysis. Journal of the American Statistical Association 58:993–1010.

Lukacs, P. M., K. P. Burnham, and D. R. Anderson. 2010. Model selection bias and Freedman's paradox. Annals of the Institute of Statistical Mathematics 62:117–125.

MacNally, R. 2000. Regression model-building in conservation biology, biogeography and ecology: the distinction between—and reconciliation of—'predictive' and 'explanatory' models. Biodiversity and Conservation 9:655–671.

Murray, K., and M. M. Conner. 2009. Methods to quantify variable importance: implications for the analysis of noisy ecological data. Ecology 90:348–355.

Murtaugh, P. A. 2014*a*. In defense of *P* values. Ecology 95:611–617.

Murtaugh, P. A. 2014*b*. Rejoinder. Ecology 95:651–653.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. Applied linear statistical models. Fourth edition. Richard D. Irwin, Chicago, Illinois, USA.

Rice, M. B., A. D. Appa, M. L. Phillips, J. H. Gammonley, B. P. Petch, and K. Eichoff. 2013. Analysis of regional species distribution models based on radio-telemetry datasets from multiple small-scale studies. Journal of Wildlife Management 77:821–831.

Schroeder, M. A., et al. 2004. Distribution on sage-grouse in North America. Condor 106:363–376.

Spanos, A. 2014. Recurring controversies about *P* values and confidence intervals revisited. Ecology 95:645–651.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. Journal Royal Statistical Society B 58:267–288.

Supplemental Material

Ecological Archives

Appendices A and B and Supplements 1 and 2 are available online: http://dx.doi.org/10.1890/14-1639.1.sm