Occupancy Modeling Species-Environment Relationships with Non-ignorable Survey Designs

Ecological Applications

Kathryn M. Irvine[1,4], Thomas J. Rodhouse [2], Wilson J. Wright[1], & Anthony R. Olsen[3]

[1] U.S. Geological Survey, Northern Rocky Mountain Science Center,
Bozeman, MT 59715, USA

[2] U.S. National Park Service, Upper Columbia Basin Network,
Bend, OR 97701, USA

[3] U.S. Environmental Protection Agency, Western Ecology Division,
Corvallis, OR 97333, USA

[4]E-mail: kirvine@usgs.gov

Appendix S1: *Diagnostics for assessing ignorability of the survey design*

Here we present the mathematical details that underlie the recommendation that a comparison of the ML and P-ML estimates are an approach to assess whether a survey design is ignorable after fitting an occupancy model. The diagnostic is based on the independence condition $w \perp\!\!\!\perp y|(\mathbf{X})$ (Bollen et al. 2016). For each unit i, we represent whether it was included in the sample S by an indicator variable, $I_i = 1$ if $i \in S$. Then a vector of indicator variables for all N sample units in the defined sample frame can be constructed. The independence condition holds if and only if the probability of sample unit i being in S is **not** related to the response y_i given a set of covariates x_i,

$$Pr(I_i = 1|y_i, x_i) = Pr(I_i = 1|x_i) \;\; \forall y_i. \qquad \text{(Eq. S1)}$$

When this independence relationship is true, the design is ignorable or non-informative (Pfeffermann 2011).

Following Pfeffermann (2007) and references therein, by definition the sample model accounts for the model parameters (p, $\boldsymbol{\beta}$) and can be redefined based on the conditional probability density function (pdf) for y_i that includes the set of indicators that represent the design,

$$f_s(y_i|x_i; p, \boldsymbol{\beta}) \stackrel{\text{def.}}{=} f_s(y_i|x_i, I_i = 1; p, \boldsymbol{\beta}),$$

where f_s denotes the sample pdf. As shown in Pfeffermann (2011) by applying Bayes theorem

$$f_s(y_i|x_i, I_i = 1; p, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{Pr(I_i = 1 \mid x_i, y_i; \boldsymbol{\gamma}) f_p(y_i|x_i; p, \boldsymbol{\beta})}{Pr(I_i = 1 \mid x_i; p, \boldsymbol{\beta}, \boldsymbol{\gamma})} \qquad \text{(Eq. S2)}$$

where $f_p(y_i|x_i; p, \boldsymbol{\beta})$ is the population pdf for i and $\boldsymbol{\gamma}$ are the parameters related to the sample weights, if needed. Notice this shows the implicit assumption made with model-based inferences is that Eq. S1 is true, otherwise the sample and population pdfs in Eq. S2 will differ, i.e. the sample is not representative of the population. One pseudo-likelihood estimator is based on rewriting the sample likelihood using the expectations of the sample weights

$$L_s(p, \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}_s, \mathbf{x}_s) = \prod_{i \in S} \frac{E_s(w_i \mid x_i; p, \boldsymbol{\beta}, \boldsymbol{\gamma}) f_p(y_i \mid x_i; p, \boldsymbol{\beta})}{E_s(w_i \mid y_i, x_i; \boldsymbol{\gamma})} \qquad \text{(Eq. S3)}$$

(Equation 3.19 in Pfeffermann 2011). Eq. S3 accounts for both the population values based on a model for the data-generating process and the sample selection process, but assumes that sample units are fixed (Pfefferman & Sverchkov 2003).

Formally, we used the estimating equation approach based on weighting the score functions to account for the mismatch between the census and sample score equations, as opposed to maximizing Eq. S3. In our case, we approximated $\frac{E_s(w_i|x_i;p,\boldsymbol{\beta},\boldsymbol{\gamma})}{E_s(w_i|y_i,x_i;\boldsymbol{\gamma})}$ by using the adjusted weights $\tilde{w}_i$ and used these adjusted weights in the score equations based on the site-occupancy model (Equation 2). More complicated procedures can be used for approximating these expectations based on the observed sample (Pfefferman & Sverchkov 2003; Pfeffermann 2011; Skinner & Mason 2012). Previous work compared these two estimators (maximizing Eq. S3 versus Equation 2) and suggested that weighting the score function performed similarly (Pfefferman & Sverchkov 2003).

An alternative to a comparison of the confidence intervals from the P-ML and ML estimated models is the following:

1. fit the unweighted model with explanatory variables $\mathbf{X}$, $y \sim \mathbf{X}$, and construct residuals,
2. plot the residuals from step 1 versus the sample weights.

If the plot and an appropriate correlation metric suggests there is no association between the residuals and the sample weights, the design can be considered not informative or ignorable (e.g., Bollen et al. 2016). This approach should be similar to comparing the weighted (P-MLE) and unweighted estimates (MLE). We used the comparison of P-ML and ML estimates because constructing occupancy model residuals for step 1 is not trivial at this point (Warton et al. 2017). For a more thorough review of the suggested diagnostic tests and relevant literature see Bollen et al. (2016).

Literature Cited

Bollen, K. A., P. P. Biemer, A. F. Karr, S. Tueller, & M. E. Berzofsky. 2016. Are survey weights needed? A review of diagnostic tests in regression analysis. Annual Review of Statistics and Its Application 3:375–392.

Pfefferman, D. & M. Y. Sverchkov. 2003. "Fitting generalized linear models under informative sampling". *Analysis of Survey Data*. Ed. by R. L. Chamber & C. J. Skinner. Chichester, West Sussex, England: Wiley. Chap. 12175–195.

Pfeffermann, D. 2007. Comment: Struggles with survey weighting and regression modeling. Statistical Science 22:179–183.

Pfeffermann, D. 2011. Modelling of complex survey data; Why model? Why is it a problem? How can we approach it? Survey Methodology 37:115–136.

Skinner, C. & B. Mason. 2012. Weighting in the regression analyis of survey data with a cross-national application. The Canadian Journal of Statistics 40:697–711.

Warton, D. I., J. Stoklosa, G. Guillera-Arroita, D. I. MacKenzie, & A. H. Welsh. 2017. Graphical diagnostics for occupancy models with imperfect detection. Methods in Ecology and Evolution 8:408–419.