Statistical Models for Presence-Only Data: Finite-Sample Equivalence and Addressing Observer Bias

William Fithian
Department of Statistics
Stanford University
wfithian@stanford.edu

Trevor Hastie
Department of Statistics
Stanford University
hastie@stanford.edu

July 1, 2019

Abstract

Statistical modeling of presence-only data has attracted much recent attention in the ecological literature, leading to a proliferation of methods, including the inhomogeneous poisson process (IPP) model [15], maximum entropy (Maxent) modeling of species distributions [12] [9] [10], and logistic regression models. Several recent articles have shown the close relationships between these methods [1] [15]. We explain why the IPP intensity function is a more natural object of inference in presence-only studies than occurrence probability (which is only defined with reference to quadrat size), and why presence-only data only allows estimation of relative, and not absolute intensities.

All three of the above techniques amount to parametric density estimation under the same exponential family model. We show that the IPP and Maxent models give the exact same estimate for this density, but logistic regression in general produces a different estimate in finite samples. When the model is misspecified, logistic regression and the IPP may have substantially different asymptotic limits with large data sets. We propose "infinitely weighted logistic regression," which is exactly equivalent to the IPP in finite samples. Consequently, many already-implemented methods extending logistic regression can also extend the Maxent and IPP models in directly analogous ways using this technique.

Finally, we address the issue of observer bias, modeling the presence-only data set as a thinned IPP. We discuss when the observer bias problem can solved by regression adjustment, and additionally propose a novel method for combining presence-only and presence-absence records from one or more species to account for it.

1 Introduction

A common ecological problem is estimating the geographic distribution of a species of interest from records of where it has been found in the past. There are many motivations for solving this problem, including planning wildlife man-

agement actions, monitoring endangered or invasive species, and understanding species' response to different habitats.

A great variety of methods for modeling this type of data have been proposed in the ecological literature, including among others the inhomogeneous poisson process (IPP) model [15], maximum entropy (Maxent) modeling of species distributions [12] [9] [10], and the logistic regression model along with its various generalizations such as GAM, MARS, and boosted regression trees [7]. Elith et al. (2006) provide an extensive survey of methods in common use.

In recent years several articles have emerged pointing out connections between the three modeling methods above. Each method requires a presence-only data set along with a set of background points consisting of either a regular grid or random sample of locations in the geographic region of interest. Warton and Shepard (2010) showed that logistic regression estimates converge to the IPP estimate when the size of the presence-only data set is fixed and the background sample grows infinitely large. Aarts et al. (2011) additionally described a variety of models for presence-only and other data sets whose likelihoods may all be derived from the IPP likelihood.

Our primary aim in writing this paper is to provide additional clarity to this topic and extend the results in several directions. We derive and interpret the relationships between between the IPP, Maxent, and logistic regression, and argue that all three methods can be viewed as solutions to the the same parametric density estimation problem (besides this density, the IPP method also estimates an intensity of sightings — the estimated species distribution multiplied by the total number n_1 of sightings — but we argue that this multiplier is typically not of scientific interest). While IPP and Maxent produce exactly the same density estimate given any finite sample of presence and background points, logistic regression does not. However, we introduce "infinitely-weighted logistic regression," which does produce the same estimate as the other two methods. This further elucidates the relationship between the methods and more importantly provides an easy way to implement extensions of the IPP model whenever they have already been implemented for weighted logistic regression.

Additionally, we address the important issue of observer bias in presence-only studies. We propose two methods, both based on modeling the presence samples as the result of a thinned underlying occurrence process. One method amounts to GLM regression adjustment, while the other combines presence-only data with presence-absence or count data to estimate and adjust for observer bias.

1.1 Presence-Only Data

Modeling of species distributions is simplest and most convincing when the observations of species presence are collected systematically. In a typical design, a surveyor visits a one-square-kilometer patch of land for one hour and records whether or not she discovers any specimens in that interval. The records of unsuccessful surveys are called absence records — a mild misnomer since ecologists recognize that specimens could be present but go undetected — and data sets of this form are called presence-absence data.

Unfortunately, presence-absence data are often expensive to collect, especially for rare or elusive species. For many species of interest, the only data available are museum or herbarium records of locations where a specimen was found

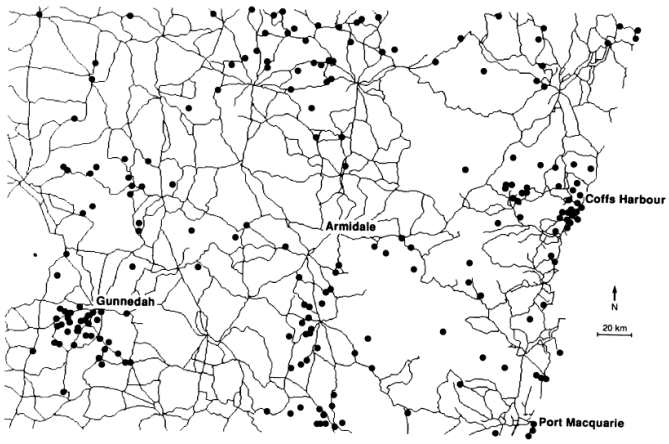

Figure 1. Koala records (courtesy of New South Wales National Parks & Wildlife Service) and the road network on part of the New South Wales north coast.

Figure 1: Observer bias in presence-only data for koalas. Taken from Margules et al. (1994).

and reported, for instance by a motorist or hiker. Typically these presence-only records are collected haphazardly and frequently suffer from unknown observer bias such as that illustrated in Figure 1. The clustering of koala sightings near roads and cities probably has more to do with the behavior of people than of koalas.

In recent years many such presence-only data sets have become available electronically, and geographic information systems (GIS) enable ecologists to remotely measure a variety of geographic covariates without having to visit the actual locations of the observations. As a result presence-only data has become a popular object of study in ecology [3].

1.2 What Should We Estimate?

Before we can sensibly decide how to model presence-only data, we must adress the issue of what it is we are modeling in the first place. How should we think of "species occurrence," the scientific phenomenon nominally under study? This issue arises with presence-only and presence-absence data alike.

1.2.1 Occurrence Probability

Figure 2 is a typical "heat-map" output of a study of the willow tit in Switzerland using count data [13]. The map reveals which locations are more or less favored by the species (in this case, high elevation and moderate forest cover appear to be the bird's habitat of choice). The legend tells us that the color of a region reflects the local probability of "occurrence."

But precisely what event has this probability? Reading the paper, we discover that occurrence means that there is at least one willow tit present on a survey path through a 1 km $\times$ 1 km quadrat of land. In this case, the authors analyze a presence-absence data set using a hierarchical model that explicitly

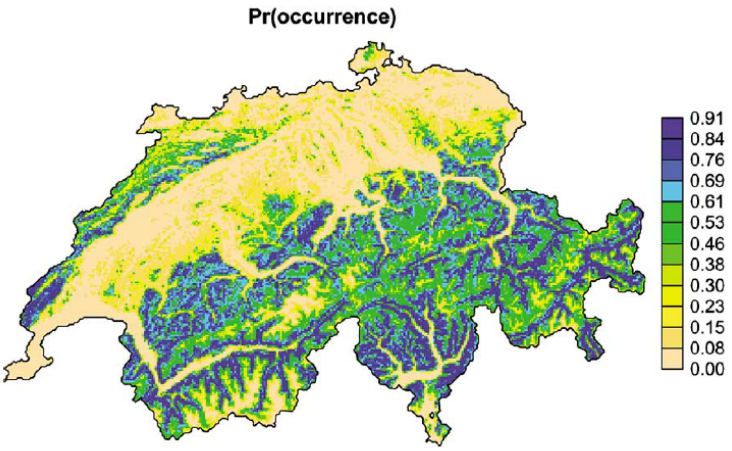

Figure 2: Typical heat map of occurrence probabilities. Taken from Royle et al. (2005)

accounts for the possibility that a bird was present but not detected at the time of the survey.

Because the survey path length varies across sampling units, the authors use it in their model as a predictor of presence probability. It is not specified which value of this predictor is used in generating the heat map, which makes the map difficult to interpret.

Even if we could interpret the heat map as the probability of a bird being present anywhere in the quadrat (not just along a path of unspecified length), the probability of a bird being present would still be larger in a 2 km $\times$ 2 km sampling unit and smaller in a 100 m $\times$ 100 m one. Therefore the very definition of "occurrence probability" in a presence-absence study depends crucially on the specific sampling scheme used to collect the presence-absence data. Consequently, interpreting the legend of such a heat map can only make sense in the context of a specific quadrat size, namely whatever size was used in the study. We would recommend that this information always be displayed alongside the plot to avoid conveying the false impression (suggested by a heat map) that occurrence probability is an intrinsic property of the land, when it is really an extrinsic property.

Though the choice of quadrat size used to define occurrence probability is ecologically arbitrary, it can at least yield estimates with a meaningful interpretation. By contrast, estimating occurrence probability in a presence-only study is a murkier proposition. Any method purporting to do so without reference to quadrat size would seem to be predicting the same probability of occurrence within a large quadrat or a small one, which cannot make sense.

1.2.2 Occurrence Rate

Since occurrence probability is only meaningful with reference to a specific quadrat size, it is a somewhat awkward quantity to model in a presence-only study. In this context it is more natural to estimate an occurrence *rate* or in-

tensity: i.e. a quantity with units of inverse area (e.g. $1/\text{km}^2$) corresponding to the expected number of specimens *per unit area*. Under some simple stochastic models for species occurrence, including the poisson process model considered here, specifying the occurrence rate is equivalent to specifying occurrence probability simultaneously for all quadrat sizes.

Unfortunately, a presence-only data set only affords us direct knowledge of the expected number of specimen *sightings* per unit area. The absolute sightings rate is reflected in the number of records in our data set, but at best, this rate is only *proportional* to the occurrence rate discussed above, which typically is the real estimand of interest. Without other data or assumptions we would have no way of knowing what this constant of proportionality might be. In other words, we can at best hope to estimate a relative, not absolute, occurrence rate.

Even assuming that the rate of sightings is proportional to the rate of occurrence is optimistic, since it rules out observer bias like that in Figure 1, an issue we take up again in Section 2.5.

1.3 Notation

We now introduce notation we will use for the remainder of the article. We begin with some geographic domain of interest $\mathcal{D}$, typically a bounded subset of $\mathbb{R}^2$. If the time of an observation is an important variable, we might alternatively take $\mathcal{D} \subseteq \mathbb{R}^3$, giving a point process in both space and time. Associated with each geographic location $z \in \mathcal{D}$ is a vector $x(z)$ of measured features.

Our presence-only data set consists of n_1 locations of sightings $z_i \in \mathcal{D}$, $i = 1, 2, \ldots, n_1$, accompanied by n_0 "background" observations $z_i, i = n_1 + 1, \ldots, n_1 + n_0$ (typically a simple uniformly random sample from $\mathcal{D}$). Finally let $x_i = x(z_i)$ be the features associated with observation i, and $y_i = \mathbf{1}_{i \leq n_1}$ be an indicator of presence/background status. Our treatment of these data as random or fixed will vary throughout the article.

1.4 Outline

The rest of the paper is organized as follows. In Section 2 we define the log-linear inhomogeneous poisson process (IPP) model and its application to presence-only data, with special focus on interpreting its parameters and their maximum likelihood estimates. In particular, the estimate of the intercept α reflects nothing more than the total number of presence samples, and as such is typically not of scientific interest for the reasons discussed in Section 1.2.2. In fact, estimating the IPP model amounts to parametric density estimation in an exponential family model, followed by multiplying the fitted density by n_1. The density thus obtained reflects the *relative* rate of sightings as a function of geographic coordinates z.

Aarts et al (2011) showed that many methods in species modeling can be motivated by the IPP model. We explicitly draw these connections and generalize them for several illuminating examples. In Section 3 we consider a particularly important example, showing that the popular Maxent method of Phillips et al. (2004) follows immediately from partially maximizing the IPP log-likelihood with respect to the intercept α. It is equivalent to the IPP in the sense that both methods produce exactly the same slope estimates $\hat{\beta}$ in any finite sample; that is, Maxent obtains the exact same density as that obtained in the first step

of IPP estimation. By finite sample, we mean a specific set of presence and background points.

In Section 4 we discuss so-called "naive" logistic regression and its connections to the IPP model. We derive its likelihood from the IPP likelihood, but show that if the log-linear model is misspecified this convergence may not occur until the background sample is quite large. The need for a large background sample is due not only to variance, but also to bias that persists until the proportion n_1/n_0 becomes negligibly small. We show, however, that if we upweight all the background samples by large weight $W \gg 1$ we can use logistic regression to recover the IPP estimate $\hat{\beta}$ precisely with any finite presence and background sample. We call this procedure "infinitely weighted logistic regression." It is best thought of as a means of using GLM software to maximize an IPP likelihood.

One advantage of viewing the IPP as a unifying model for presence-only data is that we can derive from it a variety of other simple parametric models for other types of data, including presence-absence and count data. In Section 5 we consider how we might use this fact to combine disparate data sets into one likelihood function, either to share information across species, to estimate an overall abundance rate, or even to estimate the level of observer bias in presence-only samples. Section 6 contains discussion.

2 The Inhomogeneous Poisson Process Model

The IPP is a simple model for the distribution of a random set of points $\mathbf{Z}$ falling in some domain $\mathcal{D}$. Both the number of points and their locations are random.

An IPP can be defined by its intensity function

$$\lambda : \mathcal{D} \longrightarrow [0, \infty) \tag{1}$$

Informally, λ indexes the likelihood that a point falls at or near z. For subsets $A \subseteq \mathcal{D}$, define

$$\Lambda(A) = \int_A \lambda(z)\, dz \tag{2}$$

and assume $\Lambda(\mathcal{D}) < \infty$.

There are two main ways to formally characterize an IPP with intensity λ. One simple definition is that the total number of points is a Poisson random variable with mean $\Lambda(\mathcal{D})$, and their locations are independent and identically distributed with density $p_\lambda(z) = \lambda(z)/\Lambda(\mathcal{D})$. The only thing distinguishing an IPP from a simple random sample from p_λ is that the size of the IPP sample is itself random.

Alternatively, we can think of an IPP as a continuous limit of a poisson count model in discretized geometric space. Let $N(A) = \#(\mathbf{Z} \cap A)$, the number of points falling in set A. An equivalent characterization of the IPP model is that for any A,

$$N(A) \sim \text{Poisson}(\Lambda(A)) \tag{3}$$

with $N(A)$ and $N(B)$ independent for disjoint sets A and B. For more on the IPP and other point process models, see e.g. [6].

In the case of a finite discrete domain $\mathcal{D} = \{z_1, z_2, \ldots, z_m\}$, the IPP model reduces to a discrete Poisson model, with

$$N(z_i) \sim \text{Poisson}(\lambda(z_i)) \tag{4}$$

In this sense, the IPP model may be seen as a limit of finer and finer discretizations of $\mathcal{D}$. We discuss this connection further in Section 2.4.

2.1 Modeling Presence-Only Data as an IPP

Warton and Shepherd (2010) proposed modeling the species sightings $z_1, \ldots, z_{n_1}$ as arising from an IPP sightings process whose intensity is a log-linear function of the features $x(z)$:

$$\lambda(z) = e^{\alpha + \beta' x(z)} \tag{5}$$

The formal linearity assumption does not impose as many restrictions as it may appear to, since our choice of feature vector $x(z)$ could incorporate polynomial terms, interactions, a spline basis, or other such basis expansions, which substantially broaden the set of allowable functions $\lambda(z)$.

If we adopt the interpretation of an IPP as a simple random sample with random size, we see that α and β play very different roles. Since α only multiplies $\lambda(z)$ by a constant, it has no effect on $p_\lambda(z) = \lambda(z)/\Lambda(\mathcal{D})$. The "slope" parameters β completely determine p_λ, while α merely scales the intensity up or down to attain any expected sample size $\Lambda(\mathcal{D})$ we want.

2.2 Maximum Likelihood for the IPP

Like many exponential family models, the log-linear IPP has simple and enlightening score equations. The log-likelihood is

$$\ell(\alpha, \beta) = \sum_{y_i=1} (\alpha + \beta' x_i) - \int_{\mathcal{D}} e^{\alpha + \beta' x(z)}\, dz \tag{6}$$

Note that at this point we are only using the presence samples. Differentiating with respect to α we obtain the score equation

$$n_1 = \int_{\mathcal{D}} e^{\alpha + \beta' x(z)}\, dz = \Lambda(\mathcal{D}) \tag{7}$$

That is, whatever $\hat{\beta}$ is, $\hat{\alpha}$ plays the role of a "normalizing" constant guaranteeing that $\lambda(z)$ integrates to n_1. This is our first glimpse at why $\hat{\alpha}$ is typically not of scientific interest, since it merely encodes (in a roundabout way) the total number of records we have.

Solving for α in (7) we obtain the partially maximized log-likelihood

$$\ell^*(\beta) = \sum_{y_i=1} \left(\log n_1 - \log \left(\int_{\mathcal{D}} e^{\beta' x(z)}\, dz \right) + \beta' x_i \right) - n_1 \tag{8}$$

Rearranging terms and ignoring constants, we have

$$\ell^*(\beta) = \sum_{y_i=1} \beta' x_i - n_1 \log \left(\int_{\mathcal{D}} e^{\beta' x(z)}\, dz \right) \tag{9}$$

$$= \sum_{y_i=1} \log p_\lambda(z_i) \tag{10}$$

the same log-likelihood we would obtain by conditioning on n_1 and treating the z_i as a simple random sample from the density $p_\lambda = \frac{e^{\beta' x(z)}}{\int_{\mathcal{D}} e^{\beta' x(z)}\, dz}$.

Finally, differentiating (9) with respect to β and dividing by n_1 gives the remaining score equations:

$$\frac{1}{n_1} \sum_{y_i=1} x_i = \frac{\int_{\mathcal{D}} e^{\beta' x(z)} x(z)\, dz}{\int_{\mathcal{D}} e^{\beta' x(z)}\, dz} \tag{11}$$

$$= \mathbb{E}_{p_\lambda} x(z) \tag{12}$$

This amounts to finding β for which the expectation of $x(z)$ under $p_\lambda(z)$ matches the empirical mean of the presence samples.

Maximizing the likelihood of a log-linear IPP model, then, amounts to

1. Choosing β so p_λ matches the means of the features $x(z)$ in the presence sample.

2. Choosing α so that $\lambda(z) = n_1 p_\lambda(z)$.

The first step is really a parametric density estimation problem for the presence sample, and the second step doesn't matter if n_1 has no scientific meaning. Whenever we don't care about n_1, the IPP is at its heart nothing more than density estimation (and only a little more complicated if we do care).

Despite our general skepticism about n_1 (and therefore $\hat{\alpha}$) as a quantity of scientific interest, there is one exception we can think of. When several species are under consideration, it might be interesting that species 1 was sighted twice as often as species 2 — especially if we can obtain an independent estimate of the true abundance level of species 2, say through presence-absence data. We expand upon this idea in Section 5.

2.3 Numerical Evaluation of the Integral

The IPP likelihood and score equations involve integrals that, in general, we cannot evaluate analytically. However, we can use the background samples to evaluate it via Monte Carlo integration. If our background points comprise a simple random sample, we can replace the original log-likelihood (6) with

$$\ell(\alpha, \beta) = \sum_{y_i=1} \alpha + \beta' x_i - \frac{|\mathcal{D}|}{n_0} \sum_{y_i=0} e^{\alpha + \beta' x_i} \tag{13}$$

with $|\mathcal{D}| = \int_{\mathcal{D}} dz$ representing the total area of the region.

Two other options for numerically evaluating the integral are to choose background points in a regular fine grid of $\mathcal{D}$, or to assign quadrature weights to the background points and approximate the integral with a weighted sum. In the first case, the optimization criterion would be the same, and in the second the only difference would be that the second sum would be a weighted sum over the background points.

Repeating the previous derivation gives the numerical version of the score

equations

$$n_1 = \frac{|\mathcal{D}|}{n_0} \sum_{y_i=0} e^{\alpha+\beta' x_i} \tag{14}$$

$$\frac{1}{n_1} \sum_{y_i=1} x_i = \frac{|\mathcal{D}| n_0^{-1} \sum_{y_i=0} e^{\beta' x_i} x_i}{|\mathcal{D}| n_0^{-1} \sum_{y_i=0} e^{\beta' x_i}} = \frac{\sum_{y_i=0} e^{\beta' x_i} x_i}{\sum_{y_i=0} e^{\beta' x_i}} \tag{15}$$

Throughout, we will refer to (13) as the numerical IPP log-likelihood to distinguish it from the true IPP log-likelihood (6). In practice, "fitting" the IPP model generally means maximizing (13). Operationally, this means solving (14-15)

2.4 Connection to the Poisson Log-Linear Model

Suppose that $x(z)$ is a continuous function on $\mathcal{D}$. If we use background points from a regular fine grid, we are essentially discretizing $\mathcal{D}$ into n_0 pixels A_i, each of approximately the same size $\frac{|\mathcal{D}|}{n_0}$ and centered at z_i. If we use the approximation

$$\Lambda(A_i) = \int_{A_i} e^{\alpha+\beta' x(z)}\, dz \tag{16}$$

$$\approx |A_i| e^{\alpha+\beta' x_i} \tag{17}$$

$$\approx \frac{|\mathcal{D}|}{n_0} e^{\alpha+\beta' x_i} \tag{18}$$

then the IPP model implies that the counts in each neighborhood A_i are generated independently via the Poisson log-linear model (LLM):

$$N(A_i) \sim \text{Poisson}\left(\frac{|\mathcal{D}|}{n_0} e^{\alpha+\beta' x_i}\right) \tag{19}$$

The log-likelihood of this model is (up to an additive constant)

$$\ell(\alpha, \beta) = \sum_{y_i=0} N(A_i)(\alpha + \beta' x_i) - \frac{|\mathcal{D}|}{n_0} \sum_{y_i=0} e^{\alpha+\beta' x_i} \tag{20}$$

Since $x(z)$ is continuous,

$$\sum_{y_i=0} N(A_i)(\alpha + \beta' x_i) = \sum_{y_i=0} \sum_{\substack{y_k=1 \\ z_k \in A_i}} \alpha + \beta' x_i \tag{21}$$

$$\approx \sum_{y_i=0} \sum_{\substack{y_k=1 \\ z_k \in A_i}} \alpha + \beta' x_k \tag{22}$$

$$= \sum_{y_k=1} \alpha + \beta' x_k \tag{23}$$

Therefore, (13) is almost exactly the same as the Poisson LLM log-likelihood for this discretized model. The only difference between the two is that in (20) we

have also discretized the location of each presence sample to match its nearest background sample.

We could indeed fit an IPP model in exactly this way, by simply deleting the features of the presence samples and recording only how many fall into each background sample's surrounding pixel. Approximating the model in this way, proposed by Berman and Turner (1992), provides a simple way of accessing the modeling flexibility of already-implemented GLM methods, at the cost of some loss of data, since it effectively replaces the covariate vector x_i for each presence sample with that of its nearest background sample.

As we will see later, this rounding is not really necessary. In Section 4 we propose a different procedure, infinitely weighted logistic regression, that also allows us to fit an IPP model using GLM software, but produces exactly the same estimates we would obtain if we directly maximized the (numerical) IPP likelihood on the original presence and background data.

2.5 Identifiability and Observer Bias

Observer bias poses one of the most serious challenges to valid inference in presence-only studies. Scientifically, we are interested in the *occurrence process* consisting of all specimens of the species of interest. However, our data set consists of what we might call the *observation process*, consisting only of the occurrences observed and reported by people.

We can model the observation process as an occurrence process *thinned* by incomplete observation. That is, suppose that specimens occur with intensity $\tilde{\lambda}(z)$, but that most occurrences go unobserved. Each occurrence is observed with probability $s(z)$, which may depend on features of the geographic location z (for instance, proximity to the road network). If observation is independent across occurrences, then the observation process is an IPP with intensity

$$\lambda(z) = \tilde{\lambda}(z)s(z) \tag{24}$$

It is important to keep in mind that our presence-only data set only directly reflects λ, the intensity of observations.

One (optimistic) assumption we could make about s is that it is an unknown constant — i.e., that there is no observer bias. In that case, by estimating $\lambda(z)$ we are also estimating $\tilde{\lambda}(z)$ up to an unknown constant of proportionality s, hence $p_{\tilde{\lambda}} = p_\lambda$ but $\tilde{\lambda} \neq \lambda$. Even in this optimistic scenario we can only estimate relative occurrence intensities, not absolute intensities.

A bit more realistic is the assumption that s is an unknown function of z, but that s and $\tilde{\lambda}$ are known to depend on z through two disjoint feature sets. For instance, we could model $\tilde{\lambda}$ and s as log-linear in features $x_1(z)$ and $x_2(z)$ respectively

$$\lambda(z) = \tilde{\lambda}(z)s(z) \tag{25}$$

$$= e^{\tilde{\alpha}+\tilde{\beta}'x_1(z)}e^{\gamma+\delta'x_2(z)} \tag{26}$$

Then the observation process follows the log-linear model $\lambda(z) = e^{\alpha+\beta'x(z)}$ with $x = \binom{x_1}{x_2}$ and $\beta = \binom{\beta_1}{\beta_2}$. Note that $\tilde{\alpha}$ and $\tilde{\beta}$ are the quantities of primary scientific interest, whereas α and β are the parameters governing the process we actually observe. Nevertheless, $\tilde{\beta} = \beta_1$ is still identifiable from the data because

β is. Note that this procedure does not in any way require that x_1 and x_2 be independent or uncorrelated. [1]

As $n_0, n_1 \to \infty$, our estimate $\hat{\beta}_1$ converges to the true value of $\tilde{\beta}$, the slope coefficients of $\tilde{\lambda}$. However, $\hat{\alpha}$ will converge not to $\tilde{\alpha}$ but rather to $\tilde{\alpha}+\gamma$. Without knowing γ we have no way of estimating $\tilde{\alpha}$. Similarly, if x_1 and x_2 had overlapping (or linearly dependent) coordinates we could not estimate $\tilde{\beta}$ for those coordinates.

To be more concrete, suppose koala occurrence is known to depend only on elevation (x_1), and that observer bias is known to depend only on proximity to roads (x_2). Then, despite the obvious observer bias in Figure 1 we could still estimate what elevations koalas tend to frequent, by making the correct adjustments for road proximity. By contrast, we could not estimate from this data whether koalas tend to avoid roads, since that is confounded by the obvious observer bias.

Even in the most optimistic scenario, we can estimate $\alpha = \tilde{\alpha} + \gamma$ but it carries no real information about $\tilde{\alpha}$. Indeed, we have already seen that the only role $\hat{\alpha}$ plays in estimation to make λ integrate to n_1.

The distinction between β and $\tilde{\beta}$ is also important, but for most of this paper we will focus on estimation of β, the slope parameters of the process we get to observe. We revisit this distinction and what to do about it in Section 5.

2.6 Geographic Space and Feature Space

In the context of logistic regression, it will be more natural to think of the x_i as a sample of points in "feature space" (i.e. the range of $x(z)$) rather than as the features corresponding to a sample in the geographic domain $\mathcal{D}$. There is no real distinction between these two viewpoints, so long as we adjust for the fact that some values of x are more common in $\mathcal{D}$ than others. This topic is also addressed in Elith et al. (2011), but to unify our presentation we explain it again here.

Suppose the z_i for presence samples ($y_i = 1$) arise from an IPP with intensity $\lambda(x(z))$. We will show that the corresponding x_i are then an IPP with intensity $\lambda_x(x) = \lambda(x)h(x)$, where

$$h(x) = \int_{\{z:\, x(z)=x\}} dz \tag{27}$$

Suppose x were discrete. Then $h(x)$ would be the total area of land with features equal to x, and if a presence sample were taken uniformly at random from $\mathcal{D}$ ($\beta = 0$) its probability of having features x would be proportional to that area. For continuous x, $h(x)$ is proportional to the marginal density of x in our domain $\mathcal{D}$, but the same intuition applies.

Suppose B is some subset of feature space, and consider the number $N_x(B)$ of x_i falling in the set B. This is the same as the number of z_i falling in the

[1] As with any regression adjustment scheme, we should proceed with caution here. If our linear model is misspecified (perhaps we should have included x_2^2) and x_1 is correlated with the missing variables, even regression adjustment will not remove all bias. In perverse situations it could even make the situation worse. Of course, this must be weighed against the fact that if there is observer bias, not accounting for it at all gives biased estimates too. See Section 5 for another option for dealing with observer bias.

inverse image $A = x^{-1}(B) = \{z : x(z) \in B\}$. That is,

$$N_x(B) = N(A) \sim \text{Poisson}\left(\Lambda(A)\right) \tag{28}$$

But

$$\Lambda(A) = \int_A \lambda(x(z))\, dz \tag{29}$$

$$= \int_B \int_{\{z:x(z)=x\}} \lambda(x)\, dz\, dx \tag{30}$$

$$= \int_B \lambda(x)h(x)\, dx \tag{31}$$

Furthermore, if B_1 and B_2 are disjoint sets, then so are $A_1 = x^{-1}(B_1)$ and $A_2 = x^{-1}(B_2)$. It follows that $N_x(B_1) = N(A_1)$ and $N_x(B_2) = N(A_2)$ are independent, so the x_i satisfy our second definition of an IPP.

In terms of our "random sample" view of an IPP, the above derivation implies that λ_x integrates over the whole of feature space to $\Lambda(\mathcal{D})$, and the x_i corresponding to $y_i = 1$ are distributed with density $\lambda(x)h(x)/\Lambda(\mathcal{D})$.

In the case of the log-linear IPP model $\lambda(z) = e^{\alpha+\beta' x(z)}$, this density is $e^{\alpha+\beta' x}h(x)/\Lambda(\mathcal{D})$, an exponentially tilted version of $h(x)$.

3 Maximum Entropy

Another popular approach to modeling presence-only data is the Maxent method, proposed by Phillips et al. (2004). The authors begin by assuming that the presence samples $z_1, \ldots, z_{n_1}$ are a simple random sample from some probability distribution $p(z)$.

The authors adopt the view, inspired by information theory, that the estimate $\hat{p}$ should have large entropy $H(p) = -\int_{\mathcal{D}} p(z) \log(p(z))\, dz$, while also matching certain moments of the sample. Intuitively, the goal is for the estimate to be as "close to uniform" as possible, while still satisfying certain constraints that make it resemble the empirical distribution. Indeed, if we maximized entropy with no constraints, we would take p to be the uniform distribution over $\mathcal{D}$.

Phillips et al. (2004) propose to choose the p which maximizes $H(p)$ subject to the constraint that the expectation of the features $x(z)$ under $\hat{p}$ matches the sample mean of those features, i.e.

$$\frac{1}{n_1} \sum_{y_i=1} x_i = \int_{\mathcal{D}} x(z)\hat{p}(z)\, dz = \mathbb{E}_{\hat{p}} x(z) \tag{32}$$

They show that this criterion is equivalent to maximizing the likelihood of the parametric model

$$p(z) = \frac{e^{\beta' x(z)}}{\int_{\mathcal{D}} e^{\beta' x(u)}\, du} \tag{33}$$

This is exactly the parametric form of p_λ for our log-linear IPP, and its log is exactly the partially maximized log-likelihood $\ell^*(\beta)$. The likelihood (33) is simply the likelihood of a simple random sample from p_λ, i.e. an IPP conditioned

on n_1. Indeed, the constraint (32) is nothing more than the score criterion for β in an IPP. This result may also be found in Appendix A of Aarts et al. (2011).

The popular software package Maxent implements a method slightly more complex than the one originally proposed in 2004. First, it automatically generates a large basis expansion of the original features into many derived features (quadratic terms, interactions, step functions, and hinge functions of the original features). Then, it fits a model by optimizing an ℓ_1-regularized version of the conditional IPP likelihood:

$$\sum_{y_i=1} \beta' x_i - n_1 \log\left(\int_{\mathcal{D}} e^{\beta' x(z)}\, dz\right) - \sum_j r_j|\beta_j| \tag{34}$$

The regularization parameters r_j are chosen automatically according to rules based on an empirical study of numerous presence-only data sets [10]. [2]

Mathematically, the basis expansion only increases the length of the feature vector $x(z)$. Moreover, the ℓ_1 regularization scheme does not constitute an essential difference with the other methods considered here. One could (and often should) regularize the parameters of a fitted IPP process as well, especially if $x(z)$ contains many features resulting from a large basis expansion.

Applying a penalty $J(\beta)$ to the Maxent log-likelihood does not change the equivalence between the two models, so long as α is left unpenalized. Indeed, if we add a penalty term $J(\beta)$ to the IPP log-likelihood (6), we still obtain (7) after differentiating with respect to α. But then, when we partially maximize $\ell(\alpha, \beta) - J(\beta)$ we simply obtain $\ell^*(\beta) - J(\beta)$, the penalized Maxent log-likelihood. Note that this equivalence depends on our not penalizing α in (6).

This argument generalizes immediately to a generic penalized likelihood method with any parametric form for $\log \lambda(z)$. We have established the following general proposition:

Proposition 1. *Given some parametric family of real-valued functions $\{f_\theta : \theta \in \mathbb{R}^p\}$ with penalty function $J(\theta)$, consider the penalized negative log-likelihood for an IPP with intensity $e^{\alpha + f_\theta(x(z))}$*

$$g_1(\alpha, \theta) = -\left(\sum_{y_i=1} \alpha + f_\theta(x_i)\right) + \int_{\mathcal{D}} e^{\alpha + f_\theta(x(z))}\, dz + J(\theta) \tag{35}$$

and the penalized negative log-likelihood for a simple random sample with density proportional to $e^{f_\theta(x(z))}$

$$g_2(\theta) = -\sum_{y_i=1} f_\theta(x_i) + n_1 \log\left(\int_{\mathcal{D}} e^{\alpha + f_\theta(x(z))}\, dz\right) + J(\theta) \tag{36}$$

Then (35) and (36) are equivalent in the sense that if (α, θ) minimizes g_1, θ minimizes g_2, and if θ minimizes g_2, there exists a unique α for which (α, θ) minimizes g_1.

The same applies if we replace the integrals in (35) and (36) with sums over the background sample.

Proof. Partially optimize g_1 over α as in (8) to obtain g_2. □

[2] In the notation of the Maxent papers, λ is what we call β, and β is what we call r.

Thus we see that, while Maxent and the IPP appear to be different models with different motivations, they result in the exact same density $p_\lambda(z)$. Fundamentally, this is a consequence of what we observed in Section 2.2: Maxent solves the same density estimation problem as step 1 of the IPP-fitting procedure, then skips step 2.

4 Logistic Regression

Another ostensibly different model for presence-only data is the so-called "naive" logistic regression. This approach treats presence-only modeling as a problem of classifying points as presence ($y = 1$) or background ($y = 0$) on the basis of their features. The logistic regression model treats n_1, n_0, and the x_i as fixed and the y_i as random with

$$\mathbb{P}(y_i = 1|x_i) = \frac{e^{\eta+\beta' x_i}}{1 + e^{\eta+\beta' x_i}} \tag{37}$$

Superficially this approach may appear ad hoc and unmotivated compared to IPP or Maxent. Weighed against this concern is the fact that logistic regression is an extremely mature method in statistics, enjoying myriad well-understood and already-implemented extensions such as GAM, MARS, LASSO, boosted regression trees, and more.

Logistic regression modeling of presence-only data has often been motivated by analogy to logistic regression for presence-absence data. Since it is not known whether the species is present at or near the background examples, these are sometimes referred to as "pseudo-absences," and the supposed naivete of the method refers to the idea that it treats background samples as actual absences. For instance, Ward et al. (2009) introduced latent variables coding "true" presence or absence and proposed fitting this model via the EM algorithm [14].

This interpretation raises once again the troublesome question of what it would actually mean for one of our randomly sampled background points to be a "true absence" or "true presence." Would there need to be a specimen sitting directly on the location? Or is it enough for it to be within 100 m? 1 km?

Fortunately we can sidestep these concerns entirely, since there are deep connections between the logistic regression and IPP models which yield a more straightforward interpretation. Warton & Shepherd (2010) showed that when the IPP model holds, so does the logistic regression model, and that if n_1 is held fixed and $n_0 \to \infty$, the difference between the fitted $\hat{\beta}$ for the two models converges to 0.

It is true that if the log-linear IPP model is correctly specified then, as we show below, $\mathbb{P}(y_i|x_i)$ is exactly as in (37), with the same slope parameters β. However, if the model is misspecified, the actual estimates $\hat{\beta}$ for the logistic regression in finite sample (finite presence and background samples) may be very different, even if n_1 and n_0 are both very large. Indeed, in an asymptotic regime where n_1 grows along with n_0, the two models' estimates for β generally do not even converge to the same limit. The limiting parameters of the logistic regression in general will depend on the limiting ratio of n_0 to n_1 (Fithian and Hastie, 2012).

Nevertheless, by using a form of logistic regression in which the background is upweighted by a very large number, we will see that we can exactly recover

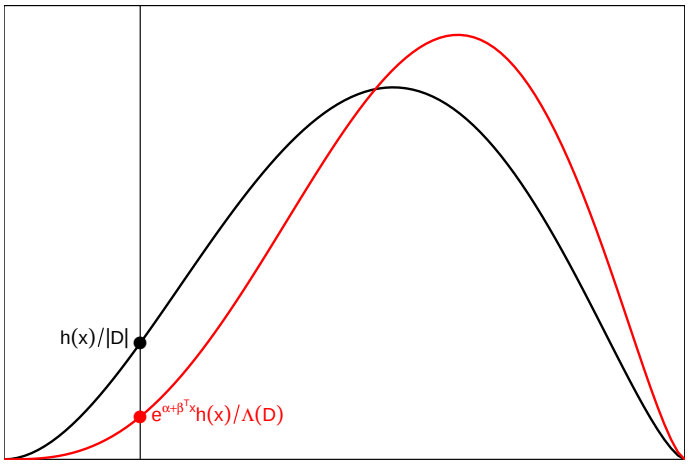

Figure 3: Presence-only sampling as case-control logistic regression.

the IPP estimate for $\hat{\beta}$ even in finite samples. In a sense, our weighting scheme "tricks" logistic-regression software into maximizing the numerical IPP likelihood. This implies in particular that packages implementing extensions of weighted logistic regression can be used to fit analogous extensions of the IPP model.

4.1 Case-Control Sampling

If we begin with a log-linear IPP model and treat as fixed the size of our presence and background samples, the x_i are a mixture of two simple random samples, one from density $e^{\alpha+\beta' x} h(x)/\Lambda(\mathcal{D})$ and the other from density $h(x)/|\mathcal{D}|$ (see Section 2.6). Then by Bayes' rule,

$$\mathbb{P}(y_i = 1|x_i) = \frac{\mathbb{P}(y_i = 1)\mathbb{P}(x_i|y_i = 1)}{\mathbb{P}(y_i = 0)\mathbb{P}(x_i|y_i = 0) + \mathbb{P}(y_i = 1)\mathbb{P}(x_i|y_i = 1)} \tag{38}$$

$$= \frac{n_1 e^{\alpha+\beta' x_i} h(x_i)/\Lambda(\mathcal{D})}{n_0 h(x_i)/|\mathcal{D}| + n_1 e^{\alpha+\beta' x_i} h(x_i)/\Lambda(\mathcal{D})} \tag{39}$$

$$= \frac{e^{\eta+\beta' x_i}}{1 + e^{\eta+\beta' x_i}} \tag{40}$$

with $e^{\eta} = \frac{n_1 e^{\alpha}|\mathcal{D}|}{n_0 \Lambda(\mathcal{D})}$. This logic is depicted in Figure 3.

Thus, the log-linear IPP model implies the individual $y_i|x_i$ follow a logistic regression model with the same slope parameters β. [3]

Thus, given any finite sample of presence and background points, we could either maximize the numerical IPP likelihood or the logistic regression likelihood, and in either case we would be fitting the same model. This fact alone does not guarantee we will obtain the same estimates $\hat{\beta}$ in any given finite sample, but if the log-linear model is correctly specified then maximizing either likelihood gives a consistent estimator of the true β.

[3] The y_i are technically not independent given n_0 and n_1 (if we knew the other $n_1 + n_0 - 1$ labels, we would know the last as well). This is always true in case-control studies, but it is typically ignored since the dependence is weak for large samples.

However, when the log-linear model is misspecified, the fitted slopes for logistic regression and numerical IPP will in general not converge to the same limiting β if n_1 and n_0 grow large together. In fact, the limiting logistic regression parameters depend on the limiting ratio of n_1/n_0.

The reason is that when a model is misspecified (as most parametric models are, such as the linear model), we are estimating the best parametric (linear) approximation in the population to the true function. But when we change the mix n_1/n_0, we are in effect changing the true population, and hence the parametric (linear) approximation [5]. The estimates only converge to the IPP estimate when $n_1/n_0 \to 0$; i.e. when the background sample grows so large that it dwarfs the presence records in the population from which we are sampling.

If we modify the logistic regression procedure a bit, we need not wait for an infinite number of background samples. We can recover the same $\hat{\beta}$ that we would estimate with an IPP using the same presence and background samples.

4.2 Infinitely Weighted Logistic Regression

Typically, although we use a finite background sample, we actually have an infinite (or at least much larger) reservoir of possible background points we could have used. Suppose we view each background point in our sample as a representative of many more background points which we only excluded for the purpose of computational convenience. To reflect this we might assign case weights to the samples

$$w_i = \begin{cases} W & y_i = 0 \\ 1 & \text{otherwise} \end{cases} \tag{41}$$

for some large number W. We would then obtain the weighted log-likelihood function

$$\ell_{\text{WLR}}(\eta, \beta) = \sum_i w_i \left[y_i(\eta + \beta' x_i) - \log\left(1 + e^{\eta + \beta' x_i}\right) \right] \tag{42}$$

$$= \sum_{y_i=1} \eta + \beta' x_i - W \sum_{y_i=0} \log\left(1 + e^{\eta + \beta' x_i}\right) - \sum_{y_i=1} \log\left(1 + e^{\eta + \beta' x_i}\right) \tag{43}$$

If a unique MLE $(\hat{\alpha}_{\text{IPP}}, \hat{\beta}_{\text{IPP}})$ exists for the IPP model with the same features, and $(\hat{\alpha}_W, \hat{\beta}_W)$ solve (42) for weighting factor W, we will show that

$$\lim_{W \to \infty} \hat{\beta}_W = \hat{\beta}_{\text{IPP}} \tag{44}$$

That is, for large W this is really a method for fitting the IPP / Maxent model.[4]

We prove a more general version of this fact, again allowing the possibility of a general penalized likelihood approach.

Proposition 2. *Consider a fixed data set* $\{(x_i, y_i)\}_{i=1,\ldots,n}$ *and convex penalty function* $J(\beta)$.

[4] Although technically $\hat{\beta}_W \neq \hat{\beta}_{\text{IPP}}$ for any finite W (hence the name "infinitely weighted"), in practice $W > 10n_1$ or so should be quite enough for the estimate to converge for all practical purposes.

Suppose the penalized numerical negative log-likelihood for an IPP with intensity $e^{\alpha+\beta' x(z)}$

$$g_1(\alpha,\beta) = -\sum_{y_i=1} \alpha + \beta' x_i + \frac{1}{n_0}\sum_{y_i=0} e^{\alpha+\beta' x_i} + J(\beta) \tag{45}$$

has a unique minimizer $(\hat{\alpha}_{\text{IPP}}, \hat{\beta}_{\text{IPP}})$*. Also, define the penalized weighted logistic regression log-likelihood*

$$\begin{aligned} g_W(\eta,\theta) = &-\sum_{y_i=1} \eta + \beta' x_i + \sum_{y_i=1} \log\left(1 + e^{\eta+\beta' x_i}\right) \\ &+ W\sum_{y_i=0} \log\left(1 + e^{\eta+\beta' x_i}\right) + J(\beta) \end{aligned} \tag{46}$$

Then if $(\hat{\eta}_{W_k}, \hat{\beta}_{W_k})$ *is any sequence of minimizers of* g_{W_k} *with* $W_k \to \infty$*, we also have*

$$\hat{\beta}_{W_k} \to \hat{\beta}_{\text{IPP}} \tag{47}$$

Proof. Define

$$\tilde{g}_W(\alpha,\beta) = g_W(\alpha - \log Wn_0, \beta) - n_1 \log Wn_0 \tag{48}$$

$$\begin{aligned} &= -\sum_{y_i=1} \alpha + \beta' x_i + W\sum_{y_i=0} \log\left(1 + \frac{1}{Wn_0} e^{\alpha+\beta' x_i}\right) \\ &\quad + \sum_{y_i=1} \log\left(1 + \frac{1}{Wn_0} e^{\alpha+\beta' x_i}\right) + J(\beta) \end{aligned} \tag{49}$$

a shifted version of g_W. Since $(\eta_{W_k} + \log Wn_0, \beta_{W_k})$ minimize $\tilde{g}_{W_k}$, it suffices to show that any sequence of minimizers of $\tilde{g}_{W_k}$ also satisfy (47).

In particular, we will show that as $W \to \infty$, $\tilde{g}_W \to g_1$ uniformly on compact subsets of $\mathbb{R}^{p+1}$. Since $\tilde{g}_W$ and g_1 are convex and the minimizer of g_1 is unique, uniform convergence in any neighborhood of $(\hat{\alpha}_{\text{IPP}}, \hat{\beta}_{\text{IPP}})$ is sufficient to prove the claim.

The difference between the two criteria is

$$\begin{aligned} \tilde{g}_W(\alpha,\beta) - g_1(\alpha,\beta) = &\sum_{y_i=0} \left(W\log\left(1 + \frac{1}{Wn_0} e^{\alpha+\beta' x_i}\right) - \frac{1}{n_0} e^{\alpha+\beta' x_i}\right) \\ &+ \sum_{y_i=1} \log\left(1 + \frac{1}{Wn_0} e^{\alpha+\beta' x_i}\right) \end{aligned} \tag{50}$$

For any compact $\Theta \subseteq \mathbb{R}^{p+1}$, we have

$$\sup_{\substack{1\le i\le n_0+n_1 \\ (\alpha,\beta)\in\Theta}} \alpha + \beta' x_i = B < \infty \tag{51}$$

so that $\frac{1}{Wn_0} e^{\alpha+\beta' x_i}$ tends uniformly to 0 for all values of α, β and x_i under consideration.

Using the Taylor expansion $\log(1+u) = u + O(u^2)$, we obtain

$$\sup_{(\alpha,\beta)\in\Theta} |g_1(\alpha,\beta) - \tilde{g}_W(\alpha,\beta)| = O\left(W^{-1}\right) \tag{52}$$

proving the result. □

The above proof goes through with virtually no modification if we substitute for the logistic regression log-likelihood the poisson log-linear model log-likelihood:

$$\ell_{\mathrm{WLLM}}(\eta, \beta) = \sum_i w_i \left[y_i(\eta + \beta' x_i) - e^{\eta + \beta' x_i} \right] \tag{53}$$

$$= \sum_{y_i=1} \eta + \beta' x_i - W \sum_{y_i=0} e^{\eta + \beta' x_i} - \sum_{y_i=1} e^{\eta + \beta' x_i} \tag{54}$$

Proposition 3. *Under the same conditions as Proposition 2, if instead*

$$g_W(\eta, \theta) = -\sum_{y_i=1} \eta + f_\theta(x_i) + \sum_{y_i=1} e^{\eta + f_\theta(x_i)} + W \sum_{y_i=0} e^{\eta + f_\theta(x_i)} + J(\theta) \tag{55}$$

then (47) holds as before for any sequence of minimizers of g_{W_k} with $W_k \to \infty$.

Proof. As before, define

$$\tilde{g}_W(\alpha, \theta) = g_W(\alpha - \log W n_0, \theta) - n_1 \log W n_0 \tag{56}$$

$$\tag{57}$$

Then

$$\tilde{g}_W(\alpha, \theta) - g_1(\alpha, \theta) = \sum_{y_i=0} \left(\frac{1}{n_0} e^{\alpha + f_\theta(x_i)} - \frac{1}{n_0} e^{\alpha + f_\theta(x_i)} \right) + \sum_{y_i=1} \frac{1}{W n_0} e^{\alpha + f_\theta(x_i)} \tag{58}$$

$$= \sum_{y_i=1} \frac{1}{W n_0} e^{\alpha + f_\theta(x_i)} \tag{59}$$

which tends uniformly to zero on compact subsets of $\mathbb{R}^{p+1}$. The rest of the proof is the same. □

These two results also imply that logistic regression and poisson regression converge to each other when we upweight the negative examples. This phenomenon has a very simple explanation. As we upweight the negative examples we drive all the fitted means toward zero, by driving $\hat{\eta}$ to $-\infty$. There is hardly any difference between a Poisson(e^λ) random variable and a Bernoulli $\left(\frac{e^\lambda}{1+e^\lambda}\right)$ random variable for very negative λ, and for that reason there is hardly any difference between the two GLMs.

4.3 Logistic Regression as Density Estimation

One interpretation of the results we have just reviewed is that in the context of presence-only data, logistic regression solves the same parametric density estimation problem as Maxent and the IPP do. Moreover, our infinitely weighted logistic regression even finds the exact same estimates as the numerical IPP and Maxent procedures do.

Using logistic regression for density estimation is not without precedent. For example, see Section 14.2.5 of Hastie et al. (2009) in which it is discussed as a

means for turning the unsupervised problem of density estimation into a well-understood supervised classification problem of samples against background. The specific proposal in that book chooses a different weighting scheme (assigning half the total weight to the presence points and the other half to the background) which, unlike infinitely weighted logistic regression, does not give exactly the IPP solution.

Viewing logistic regression as a density estimation procedure resolves the conceptual misunderstandings that originally led to its labeling as "naive."

4.4 Simulation Study: Weighted vs Unweighted Logistic Regression

We have seen that both infinitely weighted logistic regression (a.k.a. numerical IPP) and unweighted logistic regression estimate the same β parameter of the same log-linear IPP model, and when the background sample is much larger than the presence sample, the estimates $\hat{\beta}$ are close to each other.

However, the infinitely weighted logistic regression estimate can converge much faster to the large-background-sample limit if the linear model is misspecified, as we illustrate here with a simulation study.

Consider a geographic region with a single covariate x whose background density is $p_0(x) = N(0, 1)$. Now, suppose a species follows our log-linear IPP model with slope β, so that $\lambda(x) \propto e^{\beta x}$. Then the density of presence samples in feature space is $p_1(x) = e^{\beta x} p_0(x) / \left(\int e^{\beta u} p_0(u)\, du \right) = N(\beta, 1)$.

Suppose our species is in fact a mixture of two subspecies, one of which comprises 95% of the population and prefers x large, while the remaining 5% prefer x small. If each subspecies follows our model with coefficients 1.5 and -2, respectively, then

$$\lambda(x) \propto .95e^{1.5x} + .05e^{-2x} \tag{60}$$

which no longer follows the log-linear model. $p_0(x)$ and $p_1(x)$ are depicted in the upper panel of Figure 4 as the dashed and solid black lines. The black line in the lower panel shows $\lambda(x) = p_1(x)/p_0(x)$, the relative intensity as a function of the covariate x. In the lower panel all the curves have been normalized so that $\Lambda(\mathcal{D}) = \int \lambda(x) p_0(x)\, dx = 1$.

If we fit an infinitely-weighted logistic regression (or equivalently a log-linear IPP) to a large presence and background sample, our fitted $\hat{\beta}^{(\mathrm{IWLR})}$ will tend to $\mu_1 = \mathbb{E}_{p_1}(x) = 1.325$. We have plotted the corresponding large-sample estimates $\hat{\lambda}^{(\mathrm{IWLR})}(x)$ and $\hat{p}_1^{(\mathrm{IWLR})}(x)$ as blue lines in the respective panels of Figure 4.

If alternatively we fit an unweighted logistic regression to the same data set with large $n_0 = n_1$, however, the estimate $\hat{\beta}^{(\mathrm{LR})}$ will tend to roughly 1.04. The reason is roughly that the bump at -2 matters much more to the logistic regression. The resulting large-sample estimates $\hat{p}_1^{(\mathrm{LR})}(x)$ and $\hat{\lambda}^{(\mathrm{LR})}(x)$ are plotted in red. See [5] for a deeper explanation of this phenomenon.

If we fit an unweighted logistic regression to a large sample with a different ratio n_1/n_0, we would get a different estimate, which would tend toward the IPP estimate of 1.325 if and only if this ratio tended to 0. By the same token, when n_1 and n_0 are fixed, the ratio between them can play a significant role in determining the estimated β. In contrast, the IWLR / IPP estimate tends to 1.325 in large samples no matter what the ratio n_1/n_0.

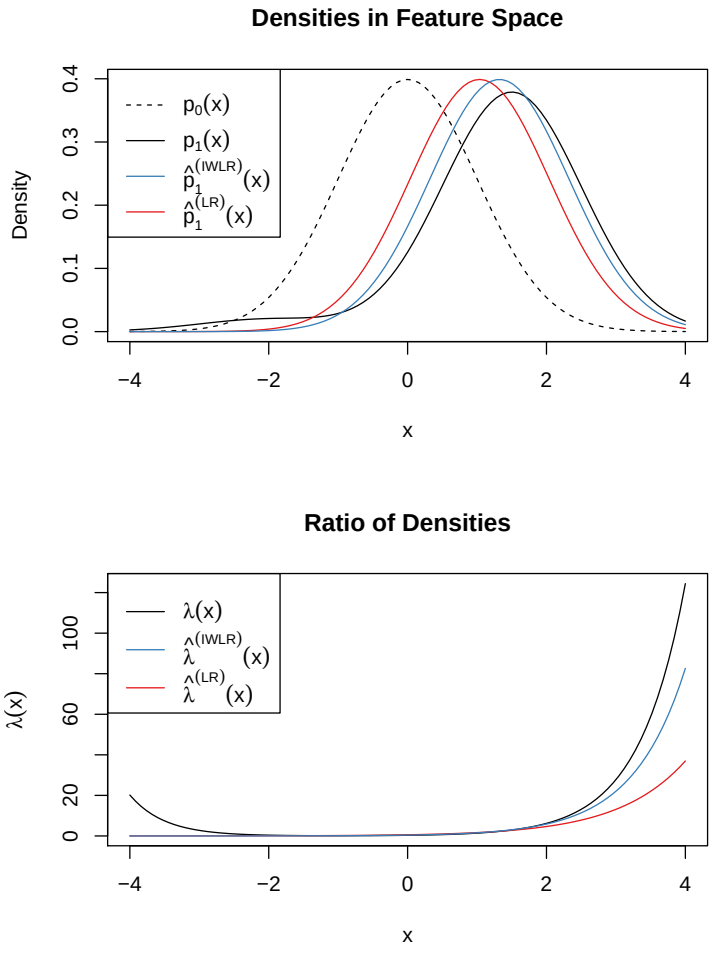

Figure 4: Misspecified log-linear IPP model with limiting estimates for infinitely-weighted logistic regression (IWLR) and standard logistic regression with $n_0 = n_1$.

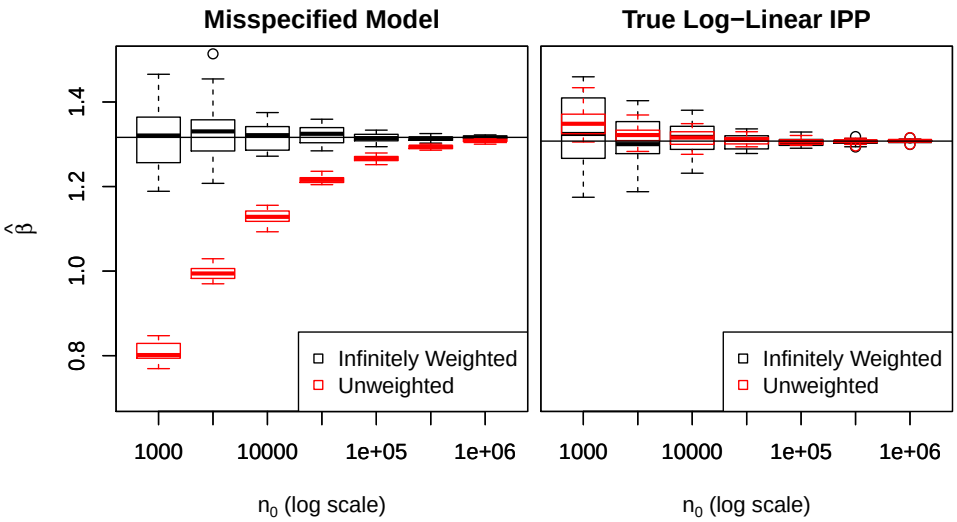

Figure 5: $n_1 = 3000$. Unweighted logistic regression may require a large background sample before convergence when the model is misspecified.

The left panel of Figure 5 illustrates this with a simulation study of the example just discussed. We first generate a single presence sample of size $n_1 = 3000$ from this species, then generate 20 sets of n_0 background samples from $p_0 = N(0, 1)$ for each of a range of values n_0 ranging from 10^3 to 10^6.

For each background sample, we fit both an "infinitely" weighted ($W = 10^4$) and unweighted logistic regression to the combination of presence and background points. For relatively large sizes of background sample, there is very little sampling variability, but the logistic regression estimates carry a large bias that depends greatly on the size of the background sample. The limiting $\hat{\beta}$, to which both methods would converge given an infinite background sample, is depicted with a horizontal line.

In the right panel, we repeat this study but with the presence sample taken from $N(\mu_1, 1)$, the correctly-specified model with the same mean as our misspecified model. Now the situation is very different; no matter what the mix of presence and background samples, the log-odds are truly linear with slope $\beta = \mu_1$. Consequently, $\hat{\beta}^{(\mathrm{LR})} \xrightarrow{p} \beta$ regardless of the limiting ratio n_1/n_0.

Since the choice of background sample size is primarily a matter of convenience, it is preferable to use an estimator that depends on it as little as possible. When the linear model is misspecified (which is nearly always the case), we recommend the infinitely weighted logistic regression over unweighted logistic regression for this reason.

4.5 Extending the IPP Model

Logistic regression is one of the most widely applied methods in statistics. For decades, applied statisticians have been developing, studying, and using variations on logistic regression to solve classification problems in statistics. R packages exist for fitting generalized additive models (GAMs), boosted regres-

sion trees, MARS, and every manner of tailored regularization schemes (see, e.g., [7]).

All of these methods are well understood within the context of logistic regression. We believe that the most important practical implication of the finite-sample equivalence between the IPP model and infinitely weighted logistic regression is that all of these methods can now be equally well understood and easily applied within the context of the IPP model.

For instance, we can fit an IPP / Maxent version of boosted regression trees with the following single line of R:

```
boosted.ipp <- gbm(y~., family="bernoulli", data=dat, weights=1E3^(1-y))
```

For an IPP / Maxent version of LASSO, ridge, or the elastic net:

```
lasso.ipp <- glmnet(dat.x, dat.y, family="binomial", weights=1E3^(1-y))
```

For an IPP GAM:

```
gam.ipp <- gam(y~s(x1)+s(x2), family=binomial, data=dat, weights=1E3^(1-y))
```

This added flexibility promises to provide a powerful tool to modelers of presence-only data.

4.6 Summary of Relationships

Thus far, we have considered several closely related models for a single presence-only sample. In this section, we collect them all in the same place and review their relationships.

Inhomogeneous Poisson Process

The "mother" model, from which all may be derived, is the inhomogeneous poisson process (IPP), whose log-likelihood is

$$\sum_{y_i=1} (\alpha + \beta' x_i) - \int_{\mathcal{D}} e^{\alpha+\beta' x(z)}\, dz \tag{61}$$

which in practice is approximated numerically via

$$\sum_{y_i=1} (\alpha + \beta' x_i) - \frac{|\mathcal{D}|}{n_0} \sum_{y_i=0} e^{\alpha+\beta' x_i} \tag{62}$$

Fitting this model amounts to solving for the density $p_\lambda(z) \propto e^{\beta' x(z)}$ for which the expected features $\mathbb{E}_{p_\lambda} x(z)$ match the empirical mean $\frac{1}{n_1} \sum_{y_i=1} x_i$, then multiplying that density by n_1.

Maxent

Conditioning on n_1 we obtain the exponential family density model $p(z) \propto e^{\beta' x(z)}$, resulting in the log-likelihood

$$\sum_{y_i=1} \beta' x_i - n_1 \log \left(\int_{\mathcal{D}} e^{\beta' x(z)}\, dz \right) \tag{63}$$

Again, in practice we replace the integral with a numeric integral. This is the log-likelihood maximized by Maxent, and it corresponds exactly to the log-likelihood (61) partially maximized over α. Hence, both procedures give exactly the same estimates of β.

Logistic Regression

The logistic regression log-likelihood is

$$\sum_i y_i(\eta + \beta' x_i) - \log\left(1 + e^{\eta + \beta' x_i}\right) \tag{64}$$

When the log-linear IPP model is correctly specified, this model is as well (aside from the fact that the $y_i|x_i$ are only approximately independent), with the same true β as in the IPP model. However, in finite samples the estimates for β given by maximizing (64) instead of (62) may be significantly different.

Infinitely Weighted Logistic Regression

We can solve this problem by upweighting all the background points by $W \gg 1$, obtaining

$$\sum_{y_i=1} (\eta + \beta' x_i) - W \sum_{y_i=0} \log\left(1 + e^{\eta + \beta' x_i}\right) - \sum_{y_i=1} \log\left(1 + e^{\eta + \beta' x_i}\right) \tag{65}$$

In the limit where $W \to \infty$, we recover with this method exactly the same $\hat{\beta}$ as we would by maximizing (62) given the same presence and background samples.

Discretized Poisson LLM

Another means for approximating the IPP log-likelihood with a GLM log-likelihood is the Berman and Turner method, which simply discretizes geographic space into pixels and assigns each presence point to a bin belonging to its nearest background point:

$$\sum_{y_i=0} \sum_{\substack{y_k=1 \\ z_k \in A_i}} \alpha + \beta' x_i - \frac{1}{n_0} \sum_{y_i=0} e^{\alpha + \beta' x_i} \tag{66}$$

This rounding of presence features is unnecessary given that we can exactly fit the IPP likelihood using the infinitely weighted approach of (65).

4.7 Model Selection and Evaluation

Regardless of which of the above likelihood models we choose, there remains the issue of model selection. With the use of geographic information systems, ecologists often have access to a large number of predictor variables, and may wish to winnow the field before modeling to avoid overfitting. Conversely, if some continuous variables are known to be important predictors, assuming a linear effect on the log-intensity may be too restrictive, and we may wish to expand the basis using splines, interactions, wavelets, etc. In either case, some regularization may also be called for.

It would be impossible to give a full treatment here of the many important considerations governing model selection. In any case, these choices need not be governed by which of the methods from Section 4.6 we choose. In particular, the large set of derived features and ℓ_1 regularization used by Maxent software can just as well be applied to the IPP model. Using the infinitely weighted logistic regression method, we can implement the exact loss function used by the Maxent with software for penalized GLMs.

5 Pooling Data from Multiple Sources

We have seen that the IPP model is a unifying framework for understanding several popular approaches to modeling presence-only data. Its simple form induces familiar parametric forms for other quantities we might model as well, for example presence-absence and species count data.

If a team of surveyors exhaustively survey plots of land A_i with areas $|A_i|$ and covariates x_i, then the number N_i of specimens they encounter will be distributed poisson with mean $\Lambda(A_i) = |A_i|e^{\tilde{\alpha}+\tilde{\beta}'x_i}$. Recall that $\tilde{\alpha}$ and $\tilde{\beta}$ are the parameters of the underlying occurrence process — $\tilde{\lambda}$ in (25). This is a poisson log-linear model with offsets $\log|A_i|$. [5]

If instead we only record whether or not at least one specimen was encountered, then we have a bernoulli GLM with complementary log-log link and offset $\log|A_i|$

$$\mathbb{P}(N_i > 0|x_i) = 1 - e^{-|A_i|e^{\tilde{\alpha}+\tilde{\beta}'x_i}} \tag{67}$$

The fact that one model yields coherent likelihoods for all of these sampling schemes means that we can pool together presence-only, presence-absence, and count data into a single log-likelihood.

Why might we want to do this? First, suppose we make the assumption of no selection bias, i.e. $\beta = \tilde{\beta}$. Then if we had access to a small presence-absence data set and a large presence-only data set, the presence-only data would help us to pin down $\tilde{\beta}$ allowing the presence-absence data to more efficiently estimate $\tilde{\alpha}$.

In practice, it may be too much to assume the surveyors see every animal, but we could assume that at least there is no dependence on z, and surveyors detect or fail to detect each animal independently with equal probability. In this case, the intercept for presence-absence data is not $\tilde{\alpha}$ but $\tilde{\alpha} - \varepsilon$ with $\varepsilon > 0$. In that case presence-absence sampling will not tell us the absolute level of abundance but it might still give a useful lower bound if ε is not too large.

5.1 Estimating Sampling Bias

Suppose that we are not willing to assume away sampling bias, but we do believe that it is the same for several species. For instance, proximity to roads and cities may introduce a comparable observer bias on similar species.

With this assumption as motivation, Phillips et al. (2009) proposed using other species' sightings as background observations instead of randomly sampled locations. This method, called the "target-group background" (henceforth TGB) method by the authors, effectively "controls away" any observer bias that affects all species equally.

Unfortunately, the TGB approach also controls away any real environmental factor that affects the overall abundance of all the species together. If half of our environment is a lush rainforest and the other half is a desert wasteland, this method implicitly assumes that the overall abundance in each region is the same, and the only reason we have fewer samples from the desert is observer bias.

[5] We should proceed with caution in modeling count data as Poisson, since the actual counts may be overdispersed, owing for instance to unaccounted-for latent variables or clustering among sightings.

If we have several species with comparable observer bias, we might use an extension of the thinning model (25)

$$\tilde{\lambda}_j(z) = e^{\tilde{\alpha}_j + \tilde{\beta}_j' x(z)} \tag{68}$$

$$\lambda_j(z) = e^{\tilde{\alpha}_j + \gamma + (\tilde{\beta}_j + \delta)' x(z)} \tag{69}$$

where j indexes species.

All the parameters of this model are identifiable once we have

1. Presence-only data for every species.
2. Presence-absence or count data for a single species (say, $j = 1$).

The presence-only data lets us independently estimate α_j and β_j for each species, while the presence-absence or count data lets us estimate $\tilde{\alpha}_1$ and $\tilde{\beta}_1$. This is enough to estimate the other $\tilde{\alpha}_j$ and $\tilde{\beta}_j$ because

$$\tilde{\alpha}_j = \alpha_j - \gamma = \alpha_j + \tilde{\alpha}_1 - \alpha_1 \tag{70}$$

$$\tilde{\beta}_j = \beta_j - \delta = \beta_j + \tilde{\beta}_1 - \beta_1 \tag{71}$$

If there is imperfect but unbiased detection in the presence-absence study, then as before we can only estimate $\tilde{\alpha}_j - \varepsilon$ and not $\tilde{\alpha}_j$. However, this would not affect identifiability for $\tilde{\beta}_j$.

Depending on the species involved and the amount of data available, we may also benefit by shrinking estimates of the β_j toward each other or sharing information in some other way.

5.2 Simulation Study

In this section we simulate a simple fictional landscape with ten species, two environmental features, and a third feature which induces observer bias.

Our geographic domain is the unit square $[0, 1]^2$, and the two environmental variables are identified with the geographic axes. For simplicity we have discretized the geographic space into pixels, so there is no distinction between the LLM and the IPP.

The first variable, "wasteland," is negatively associated with all ten species by varying amounts, and the second, "elevation," is positively associated with some species and negatively associated with others. The particular $\tilde{\beta}_j$ for each species is randomly generated.

The region also contains four "towns" in which human population, the third variable, is large. The human population does not affect the species, but it does have a multiplicative effect on the rate at which sightings occur. Figure 6 displays the relative observance and occurrence intensities for two species.

Our data consist of presence-only data sets for each species, each with 300 points in expectation. First we fit an IPP to each species using the other species' observations as a background sample, following the TGB proposal, looking for contrasts between the observation intensities of the different species. Its estimates are listed in Table 1 under the heading TGB.

The TGB method succeeds in removing the observer bias from the estimated intensities, but it additionally "controls away" the average effect of the wasteland variable on all species, producing severely biased estimates. Given the data

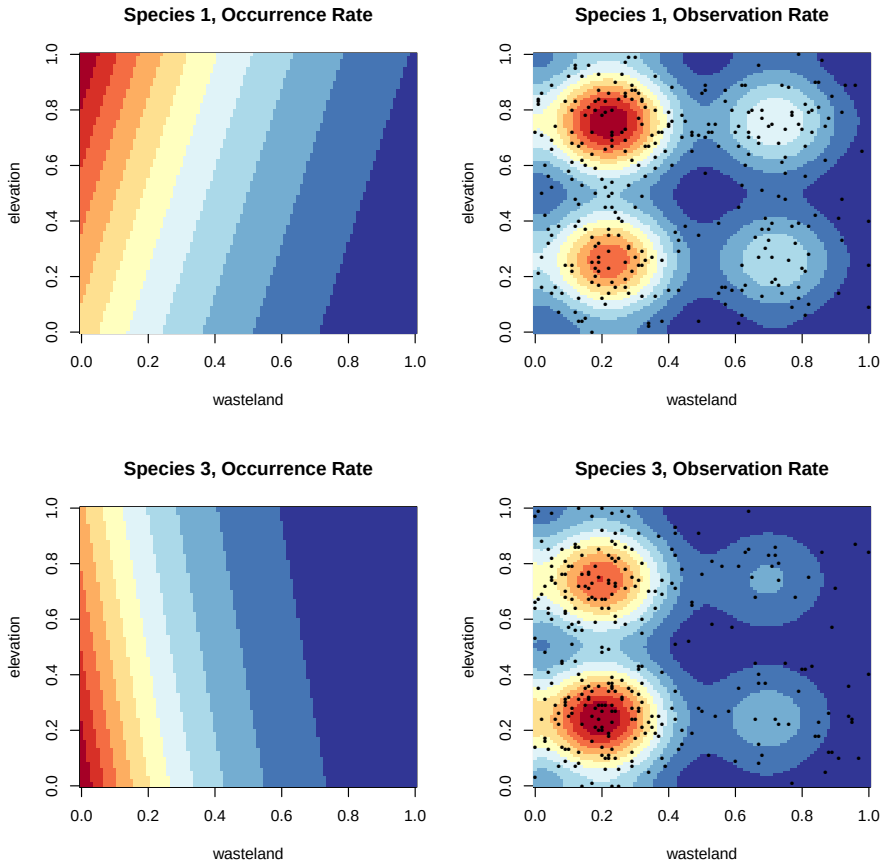

Figure 6: $\tilde{\lambda}$ (occurrence rate) and λ (observation rate) for two species. Red is high, blue is low.

we have and no other information or assumptions, there is simply no way to distinguish between observer bias and a legitimate environmental effect common to all the species.

If we knew ahead of time that the human population variable only affected $s(z)$ and the environmental variables only affected $\tilde{\lambda}$, we could use the regression-adjustment approach of Section 2.5 for each species. Because that assumption is correct in this case, regression adjustment successfully adjusts for the observer bias, but does not remove the true environmental effect. The estimates from this method are in Table 1 under the heading RA (for regression adjustment). Note that we have not actually estimated the effect of human population on $\tilde{\lambda}$, but rather assumed (correctly in this case) that it has no effect.

If we were not willing to make this assumption (for instance, we may believe the human population might have a real effect on the species), we could still account correctly for observer bias by maximizing the likelihood of the model defined by (68-69). We call this the data-pooling (DP) technique. This method requires additional data: we must have unbiased presence-absence or count data for at least one species. Here we independently simulate count data for species 1, collected from 300 randomly chosen sites. Since the distribution of the counts depends on sampling area $|A_i|$ and detection probability ε, we take all $|A_i|$ to be equal and choose $|A_i|$ and ε implicitly so that the average site has expected count of 2. The estimates from this method are in Table 1 under the heading DP (for data pooling).

	Log Human Pop.				Wasteland				Elevation			
Species	True	TGB	RA	DP	True	TGB	RA	DP	True	TGB	RA	DP
1	0.00	0.01	—	-0.08	-1.55	0.93	-1.23	-1.52	0.42	0.40	0.43	0.33
2	0.00	-0.17	—	-0.24	-2.09	-0.07	-2.16	-2.45	0.98	0.87	0.89	0.79
3	0.00	0.05	—	-0.07	-2.79	-0.72	-2.67	-2.96	-0.39	-0.44	-0.29	-0.38
4	0.00	-0.18	—	-0.26	-1.43	0.77	-1.31	-1.60	-1.04	-1.15	-1.03	-1.13
5	0.00	0.33	—	0.22	-1.96	0.15	-1.98	-2.27	1.78	1.52	1.51	1.42
6	0.00	-0.07	—	-0.15	-2.07	0.00	-1.92	-2.21	-2.31	-3.19	-2.90	-3.00
7	0.00	0.24	—	0.13	-2.35	-0.31	-2.34	-2.63	0.88	0.38	0.46	0.36
8	0.00	-0.01	—	-0.10	-1.88	-0.05	-2.11	-2.40	0.04	0.24	0.32	0.22
9	0.00	0.05	—	-0.06	-2.99	-1.12	-3.10	-3.39	1.01	0.85	0.90	0.81
10	0.00	0.01	—	-0.08	-1.93	0.09	-1.99	-2.28	0.43	0.43	0.50	0.41

Table 1: Results of three methods for correcting observer bias. The target-group background method succeeds in controlling for the Human Population variable, but in so doing controls away the Wasteland variable.

We represent this phenomenon pictorially in Figure 7, which displays scatter plots of estimated effect size versus true effect size for the TGB, RA, and DP methods.

6 Discussion

We have shown here that the IPP, Maxent, and infinitely-weighted logistic regression are equivalent in several senses:

1. All may be derived from the IPP model.

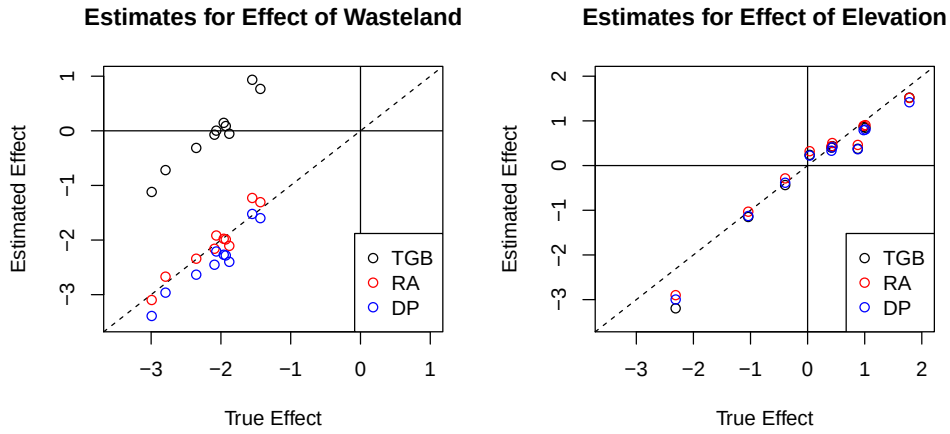

Figure 7: Scatter plots of estimated versus true effect sizes for the two environmental variables. Dotted line is $x = y$. The TGB method inappropriately removes the average effect of wasteland.

2. All may be thought of as performing the same parametric density estimation problem, which amounts to fitting β.

3. All fit the same $\hat{\beta}$ given the same finite presence and background samples.

The only difference between the IPP and the other two methods is that it additionally estimates an intensity equal to n_1 times the fitted density. Since we can only estimate the intensity up to a proportionality constant, this distinction is not essential.

Standard logistic regression shares the first two aspects with the other methods, but is not equivalent in finite samples. If the model is misspecified, then even when the presence and background samples grow infinitely large, the logistic regression and IPP estimates do not converge to each other unless n_1/n_0 becomes negligible. This implies that the size n_0 of the background sample can play an unwanted role in determining the estimate $\hat{\beta}$.

Infinitely-weighted logistic regression provides a general tool for repurposing the many packages that extend logistic regression to fit analagously-extended (numerical) IPP models. This added flexibility promises to provide a powerful tool to modelers of presence-only data.

These findings remain the same if we replace the $\beta' x_i$ term in the exponent with a smooth parametric model that is convex in its parameters θ, or apply a convex regularizing penalty to β.

Finally, the IPP modeling framework induces simple likelihoods for other forms of data which may not have the same biases as presence-only data sets, and may be combined with the presence-only data into a single likelihood function. We have demonstrated one technique for using such heterogeneous data to correct for observer bias.

Acknowledgements

The authors are grateful to Jane Elith for helpful discussions and suggestions. Will Fithian was supported by VIGRE grant DMS-0502385 from the National Science Foundation. Trevor Hastie was partially supported by grant DMS-1007719 from the National Science Foundation, and grant RO1-EB001988-15 from the National Institutes of Health.

References

[1] G. Aarts, J. Fieberg, and J. Matthiopoulos. Comparative interpretation of count, presence–absence and point methods for species distribution models. *Methods in Ecology and Evolution*, 2011.

[2] M. Berman and T.R. Turner. Approximating point process likelihoods with glim. *Applied Statistics*, pages 31–38, 1992.

[3] J. Elith, C.H. Graham, R.P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R.J. Hijmans, F. Huettmann, J.R. Leathwick, A. Lehmann, et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2):129–151, 2006.

[4] J. Elith, S.J. Phillips, T. Hastie, M. Dudík, Y.E. Chee, and C.J. Yates. A statistical explanation of maxent for ecologists. *Diversity and Distributions*, 2011.

[5] W. Fithian and T. Hastie. Local case-control sampling. 2012.

[6] C. Gaetan and X. Guyon. *Spatial statistics and modeling*. Springer Verlag, 2009.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*, volume 1. Springer Series in Statistics, 2009.

[8] CR Margules, MP Austin, D. Mollison, and F. Smith. Biological models for monitoring species decline: The construction and use of data bases [and discussion]. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1307):69–75, 1994.

[9] S.J. Phillips, R.P. Anderson, and R.E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3):231–259, 2006.

[10] S.J. Phillips and M. Dudík. Modeling of species distributions with maxent: new extensions and a comprehensive evaluation. *Ecography*, 31(2):161–175, 2008.

[11] S.J. Phillips, M. Dudík, J. Elith, C.H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1):181–197, 2009.

[12] S.J. Phillips, M. Dudík, and R.E. Schapire. A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning*, page 83. ACM, 2004.

[13] J. Andrew Royle, James D. Nichols, and Marc Kry. Modelling occurrence and abundance of species when detection is imperfect. *Oikos*, 110(2):353–359, 2005.

[14] G. Ward, T. Hastie, S. Barry, J. Elith, and J.R. Leathwick. Presence-only data and the em algorithm. *Biometrics*, 65(2):554–563, 2009.

[15] D.I. Warton and L.C. Shepherd. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, 4(3):1383–1402, 2010.