STATISTICS IN ECOLOGICAL MODELING; PRESENCE-ONLY DATA AND BOOSTED MARS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Gillian Ward

September 2007

© Copyright by Gillian Ward 2007
All Rights Reserved

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Trevor Hastie) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Rob Tibshirani)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(David Rogosa)

Approved for the University Committee on Graduate Studies.

Abstract

The research presented in this thesis was primarily motivated by current problems in ecological modeling. One such major problem is the so-called presence-only problem, the analysis of which forms the core of this thesis. The remainder of the thesis introduces boosted MARS, a new flexible modeling procedure.

In ecological modeling of the habitat of a species, it can be prohibitively expensive to determine species absence. Presence-only data consists of a sample of locations with known presences and a separate group of locations sampled from the population, with unknown presences. We propose four different approaches, all of which estimate an underlying presence-absence model for presence-only data. In the first, an application of the EM algorithm can be used with almost any off-the-shelf logistic model. In the other three models, gradient boosted regression trees are used to maximize different loss functions arising from different model assumptions. Simulations and analyses based on sampling from presence-absence records of fish in New Zealand rivers illustrates that these new procedures can reduce both deviance and the shrinkage of marginal effect estimates that occur in the naive model often used in practice. It is shown that the population prevalence of a species is only identifiable when there is some unrealistic constraint on the structure of the logistic model. In practice, it is strongly recommended that an estimate of population prevalence be provided.

Boosted MARS is a modeling procedure which aims to mimic the flexibility and performance of gradient boosted trees, but while providing a smoother fitted model. Refitting the boosted MARS basis functions using L_1 regularization sometimes improves predictive performance. However, boosted regression trees usually outperform either boosted MARS mdoel. We also investigate the effects of smoothing and thresholding the MARS basis functions on predictive and descriptive performance.

Acknowledgments

In my first statistics class, as an undergraduate at Heriot-Watt University in Scotland, Ian Currie opened a new world to me as he demonstrated how math can be used to solve real-world problems. Later in my undergraduate career, Denis Mollison and Stan Zachary provided funding and supervision for research projects, and encouraged me to continue my studies to a postgraduate level. I will always be grateful for these initial opportunities in learning and becoming comfortable with statistics.

Over the years there have been many other people who have supported and inspired me. I learnt much from my teachers at the University of Cambridge and UC Berkeley. Paul Holland, who has since left Berkeley, gave me a different perspective of statistics and encouraged me to think more critically. In my first foray into the working world I met Tony Thrall who continues to be a mentor to me today. I have also appreciated the challenges of teaching statistics, working with a committed team at the University of Brighton in England.

Of course I could not have done this without my advisor Trevor Hastie. He has used both carrot and stick to motivate me to both complete this thesis and to be more confident in my statistical skills. I am also very grateful to other professors at Stanford who have inspired and guided me, in particular my reading and thesis defense committee members

Rob Tibshirani, David Rogosa, Chris Manning and Art Owen. Susan Holmes deserves a special thank you for getting me a last minute place in the program.

On a more personal side, I am indebted to Kerry Rodden for her emotional and financial support through these four years. Although I was the "stats guru" for her thesis, she was so much more for mine. My father sadly died before I could tell him I was returning to my PhD studies, but financial support from him and my mother have helped assuage some of the pressures of graduate school.

I am also very grateful to the Stanford Red Barn, and all the wonderful horses and instructors there who gave me both a space to relax and another very different arena in which to compete. To Chili, Jack, Razzi, Max and now Cash 'N' Diamonds, thank you for your patience.

Contents

Abstract **v**

Acknowledgments **vii**

1 Introduction **1**

I Presence-Only Modeling **4**

2 Introduction to the Presence-Only Problem **5**

3 An Example: the Longfin Eel **8**

4 The Presence-Only Problem **10**

4.1 The landscape 11

4.2 Defining species presence 12

4.3 Presence-only data sampling 13

4.4 Revisiting the longfin eel 14

5 Existing Presence-Only Models **16**

5.1 The naive presence-only model 18

5.2 Manly's exponential model 21

5.3 Maxent 22

5.4 Gradient boosting and regression trees 25

6 The Logistic Presence-Only Model **27**

6.1 The presence-only logistic likelihood 28

6.2 Modified case-control sampling 30

6.3 The full likelihood 33

6.4 Model 1: the EM model 34

6.5 The EM model and stepwise procedures 37

6.6 Bias in the naive presence-only model 38

6.7 Estimating the popoulation prevalence π 39

7 Presence-Only Models and Gradient Boosting **43**

7.1 Model 2: the sample logistic model 43

7.2 Model 3: the population logistic model 45

7.3 The population logistic model as a limit 48

7.4 Model 4: the population log model 50

7.5 Adjusting for shrinkage inherent in boosting 52

7.6 Comparing the presence-only models 53

8 Assessing the Presence-Only Models **55**

8.1 Simulation studies 57

8.1.1 Modeling via known basis functions 59

8.1.2 Modeling via boosted regression trees 61

8.2 The longfin eel; generating presence-only data 67

9 The Presence-Only Problem: Conclusions **71**

II Boosted MARS **74**

10 Introduction to Boosted MARS **75**

10.1 Motivation 76

10.2 MARS 78

10.3 Adaptations of MARS basis functions 80

11 Boosted MARS **83**

11.1 Step 1: Boosted MARS 84

11.2 Step 2: L_1 regularization on MARS basis functions 86

11.3 Comparing boosted MARS and L_1 boosted MARS 87

11.4 Orthogonalization 88

12 Evaluation of Boosted MARS **89**

12.1 Simulation studies 89

12.2 The longfin eel 95

13 Boosted MARS: Conclusions **99**

A Miscellaneous calculations **101**

A.1 Manly's exponential model and the naive logistic model 101

A.2 Estimation of π for presence-only logistic regression 103

A.3 Proof of penalized likelihood 105

B Implementation of the presence-only models **107**

B.1 The `ecogbm` package 107

Chapter 1

Introduction

Statistical techniques are widely used in ecology in order to model and analyze organisms and ecosystems. Those ecological models vary considerably in scope and complexity. A behavioral ecologist might study interactions between individuals of the same species (Buston, 2003), or growth dynamics of a single-species population (Gotelli, 1995). In community ecology, the focus is on interactions between species, such as predation (Kerfoot and Sih, 1987) or competition (Roughgarden, 1983). On a larger scale, environmental ecology attempts to study the relationship between organisms and their environment, in part to assess the impacts of pollution and other environmental changes on ecosystems (Freedman, 1994)

The statistical focus of this thesis is prediction and classification. In particular, we consider the use of such models in ecological niche or *habitat modeling* Guisan and Zimmermann (2000). This is the modeling of species presence or abundance with respect to environmental variables. These models estimate the niche in environmental space where the species

can be found. Such models have many applications in ecological management, for example in indicating suitable areas for habitat protection (Milsom et al., 2000). Peterson and Vieglais (2001) use habitat models to predict potential invasions of non-native species or they may be used to increase resolution of species atlases used in many other applications Araújo et al. (2005). More recently, focus has turned to predicting the effect of climate change on species viability and extinction Thomas et al. (2004). This problem brings a whole new set of problems, as we must extrapolate beyond the range of current data to predict the effects of future environmental changes.

Even the amount of available data can vary considerably between problems. In some areas of research, large amounts of data are collected electronically, for example through remote sensing of environmental measurements (Lillesand et al., 2004) or GPS tracking of mobile species (Weimerskirch et al., 2002). At the other extreme, data collection can be time-consuming and expensive, especially in studying for rare species (Engler et al., 2004). In this latter case, there has been a push toward fully utilizing currently available data, such as the rich sources of species presence records in herbaria and museums. However, these ad-hoc records provide no specific knowledge of where the species is not present. The statistical challenge of using these *presence-only data* in order to estimate a presence-absence habitat model is the impetus behind Part I of this thesis. We critique current approaches to this problem and then provide two types of solutions to this problem. The first works as a wrapper around almost any logistic modeling procedure while the second uses special implementations of gradient boosted regression trees.

Ecologists have begun to incorporate recent statistical and machine learning techniques in their analyses. Support vector machines have been used successfully in both habitat modeling (Guo et al., 2005) and classification of remote sensing data Zammit et al. (2006). Leathwick et al. (2006) and De'ath (2007) have reported increased predictive accuracy in habitat modeling using boosted regression trees. Additionally, the simple structure of multivariate adaptive regression splines (MARS) models allows ecologists to create smooth

but flexible models and subsequently generate species distribution maps in other software (Leathwick et al., 2005). In Part II of this thesis, we investigate how a gradient boosting implementation of MARS can retain the smoothness and simplicity of MARS, while increasing predictive accuracy.

The statistical models presented here may also be useful in other fields. Presence-only data also arises in text classification in the form of "positive and unlabeled examples" (Liu et al., 2003), and in economics as "contaminated controls" (Lancaster and Imbens, 1996). Boosted MARS is a general modeling technique that can be used in place of generalized linear models and for more general loss functions in many situations.

Part I

Presence-Only Modeling

Chapter 2

Introduction to the Presence-Only Problem

Accurate modeling of wildlife habitat selection is important for effective ecological management of a species. Logistic models are often used to estimate species distribution, ideally based on recorded presences or absences at randomly sampled locations, and given environmental variables as predictors. However, such *presence-absence data* is expensive to collect and for many species the only available data are of recorded or *observed presences*. For example, herbaria and museums have extensive historical occurrence records (Elith et al., 2006) and radiotelemetry provides a rich source of range locations for mobile species (Frair et al., 2004). Additionally, such data is becoming increasingly available online (Stein and Wieczorek, 2004).

Environmental information can be collected for these locations where the species was observed, using geographical information system (GIS) technology. Some methods use only these observed presences to estimate a species range; for example HABITAT (Walker and

Cocks, 1991) estimates the range of a species via a convex hull defined in environmental space. However, many methods also require a *background sample* or *pseudo-absences* — a random sample of locations taken from the full landscape of interest. Although the presence or absence of a species at the background locations is unknown, these locations provide a summary of the landscape against which we can compare the observed presences. Indeed, using the background data can considerably increase prediction accuracy (Elith et al., 2006). This combination of a sample of locations with known presences, and a background sample of locations from the whole landscape of interest is what we will refer to as *presence-only data*.

The modeling of presence-only data in ecology is reviewed in Keating and Cherry (2004) and Pearce and Boyce (2006). One method is to fit a logistic model to the recorded presences and some background data, ignoring the existence of true presences in the background data. We refer to this as the *naive model*, and discuss it in more detail in Section 5.1. It suffices to mention here that the naive model can be highly biased, especially for more common species. An alternate procedure, the exponential model of Manly et al. (2002, Section 5.4), exploits the result that the fitted naive logistic regression model provides estimates for a log model of the presence-absence probability. However, this approach does not result in valid estimates of probabilities (Keating and Cherry, 2004), a point we discuss further when we introduce the model in more detail in Section 5.2.

Lancaster and Imbens (1996) provide an exact method for estimating a presence-absence logistic regression model using presence-only data. They use direct maximization of a presence-only likelihood to estimate the linear predictor for the presence-absence logistic model. However, this technique is rarely used in presence-only modeling as there is no easily-available implementation, and convergence problems have been reported when the population prevalence of the species is assumed unknown. Indeed, we will show that without an independent estimate of overall species prevalence, estimation of this model is

unstable due to over-reliance on the logistic form of the model. We also provide a more accessible EM algorithm for maximizing this presence-only likelihood, that can be used with a wide-ranging group of models. As part of this thesis we compare the performance of this EM model with a nonparametric boosted regression trees approach based on the same likelihood. Both the EM and gradient boosting models assume that the overall prevalence of the species is known, or can be estimated e.g. using expert knowledge.

Maximum entropy modeling (Maxent, Phillips et al., 2006) approaches the problem from a different perspective. Logistic models aim to estimate the conditional probabilities of presence given environmental variables, but Maxent models the density of environmental variables given the presence of a species (relative to an empirical reference distribution). We will discuss this model and introduce two models that approach the problem in a similar way. A recent extension of Maxent (Phillips et al., 2006) deals with bias in the sampling distribution, which we touch on in Section 4.3.

The presence-only problem is also referred to as use-availability modeling (e.g. Keating and Cherry, 2004). This distinction in terminology is often used to discriminate between wide-ranging and localized species, and the resulting definitions of presence. We will generally use the presence-only terminology, but may refer to use-availability when discussing research that uses that term. To overcome ambiguities in the presence-only problem and bring together the concepts of presence-only and use-availability, we start by carefully specifying the presence-only model and then clarify sampling assumptions (Chapter 4). After describing and critiquing existing models (Chapter 5) we then present four new models (Chapters 6 and 7). The methods are compared using simulations and artificially generated presence-only data of a species of eel in New Zealand rivers in Chapter 8.

Chapter 3

An Example: the Longfin Eel

The longfin eel, *Anguilla dieffenbachii*, is a freshwater eel from New Zealand (New Zealand Statistics, 2005). When mature, it migrates hundreds of kilometers out to sea where it breeds and subsequently dies. The larvae travel back to New Zealand on ocean currents, feeding on plankton. The eels then swim upstream, and may climb waterfalls or pass by large hydroelectric dams when appropriate eel passage facilities are provided. In 2000 and 2004, fishing quotas were introduced to protect the population from over-fishing.

Imagine that we wish to model the type of environment that the longfin eel favors, or to assess the impact of dams or other barriers. One way to do this is to model the probability that at least one eel is present at a particular location, given certain environmental covariates. These environmental covariates might include average summer or winter air temperature at the site, distance from the coast and the presence of a dam downstream. In particular, we will ignore spatial information, so that we can apply this model to other freshwater locations. We would ideally collect data to build this model by randomly sampling locations along rivers in New Zealand, and then visiting each site to identify whether

the species was present or not. However, this may be prohibitively expensive or time-consuming. Instead there may be ad-hoc records of eel presence freely available, for example from ecologists studying other fish species. Thus we know in which river locations an eel has been observed — these are our observed presences. However, we don't know for sure where the eel does *not* live.

As the rivers of New Zealand are well-studied, a wealth of ecological measurements are readily available. Thus it is straightforward to access environmental measurements for our observed presence locations. Similarly, we can generate a random sample of freshwater locations, using GIS, for which we also access environmental information. These locations form our background sample.

As we introduce further concepts, and turn to assessing the proposed and existing models, we will continue to reference this example.

Chapter 4

The Presence-Only Problem

A precise framework is necessary to enable a clear discussion of the presence-only problem and potential models. The aim is to use presence-only data to build a statistical model of the probability of species presence across a landscape, given the values of environmental variables. To do this, we first need clear definitions of the landscape and what we mean by the presence of a species. We also need to state our assumptions about the sampling structure of the presence-only data. We do this after first introducing some notation to discriminate between true and observed presences.

We wish to predict the probability of a true presence ($y = 1$) or absence ($y = 0$) of a species, given environmental variables or covariates $\mathbf{x}$. However, we know only the observed presences ($z = 1$) and the background data ($z = 0$). Note that an observed presence implies there must be a true presence, so $z = 1$ implies $y = 1$. For the background data ($z = 0$), we do not know the true presence or absence, and so y is unknown when $z = 0$.

For some models, we will sometimes also refer to the sampled locations, i.e. all locations

for which we have either an observed presence or which are in the background data, using the notation $s = 1$. Then $s = 0$ indicates that a location is in the landscape we are interested in, but is not part of our data set. We will also use n_p and n_a to indicate the number of observed presences and background data respectively. The aim of presence-only modeling is to use the observed z and $\mathbf{x}$ to estimate the model of interest, $\mathbb{P}(y = 1|\mathbf{x})$. However, our data are generated instead by the probability $\mathbb{P}(z = 1|\mathbf{x}, s = 1)$.

4.1 The landscape

We consider a landscape or region, which we denote $\mathcal{L}$. This landscape is a set of locations suggested by the problem and defined by the ecologist. It may be limited by political or geographic boundaries, or by environment e.g. all rivers within a region. We will also refer to $\mathcal{L}$ as the *population* of locations, so that population refers to the area that we are interested in. As with all statistical models, one implication of the choice of a landscape is that the resulting estimated model may not be valid outside that landscape.

We define the landscape $\mathcal{L}$ as a discrete set of pixels covering the region of interest. This landscape induces the multivariate density function $f(\mathbf{x})$ of the covariates $\mathbf{x}$ for that landscape. (Although $\mathcal{L}$ is discrete, we assume the scale is small enough that we can approximate the probabilities with a continuous density function.) We can also restrict $\mathcal{L}$ to a subset $\mathcal{L}_1$ of the landscape that includes only locations where this species is present. This induces $f_1(\mathbf{x}) = f(\mathbf{x}|y = 1)$, the density of the covariates $\mathbf{x}$ for locations where the species is found. A similar argument gives us $f_0(\mathbf{x}) = f(\mathbf{x}|y = 0)$. We show later that it is the relative values of $f_1(\mathbf{x})$ and $f(\mathbf{x})$ that provide information about the probability of the presence of the species at a location with covariate values $\mathbf{x}$.

In the definition of $\mathcal{L}_1$ we assumed that at each location the species is either present or absent. Within the lifetime of a sedentary species this may be true, but wide-ranging species may only visit a given location occasionally. To be more specific, we define a new time-indexed landscape $\mathcal{L}^{\mathcal{T}}$ that is our current landscape $\mathcal{L}$ considered at all time points within a time frame $\mathcal{T}$ say, i.e. $\mathcal{L}^{\mathcal{T}} = \mathcal{L} \times \mathcal{T}$. We assume both that the values of the covariates $\mathbf{x}$ are constant and that the model is stationary across $\mathcal{T}$. In other words, the probability of a presence given the covariates $\mathbf{x}$ does not depend on the time t: $\mathbb{P}(y = 1|\mathbf{x}, t) = \mathbb{P}(y = 1|\mathbf{x})$. Then the true probability of presence for a location is the proportion of times $t \in \mathcal{T}$ that the species is present at that location. This stationarity assumption is similar in flavor to an explicit assumption of Guisan and Thuiller (2005), that the species is in pseudo-equilibrium with its environment during the time frame of the data collection.

4.2 Defining species presence

In order to create and interpret our models, it is important that we have a precise definition of species presence. We provide a framework for that definition, because it will necessarily be problem-specific, given the nature of the species. Consider two species, a wide-ranging bird of prey and a long-lived tree species. The bird's range may cover a wide area, so to define its presence at a location we need to state a fixed time period within which the bird must pass through the location. This time period will be related to the assumed sampling scheme, and should be chosen so that altering the start of the time period will not significantly affect the model. For example, birds of prey have different day and night behaviors, so this time period should be at least one day.

For both sedentary and wide-ranging species we must also define the spatial scale: the size of the pixels in the landscape $\mathcal{L}$. These are determined by the ecologist and may depend on

the mobility of the species under study, or the resolution of available environmental measurements (Guisan and Thuiller, 2005). It is important to make this spatial scale explicit, as it is necessary for calculating the external estimate of species population prevalence, and in interpreting model predictions. In our discussion we will assume that both of these scales have been defined for the problem. We then define a presence of a species at a location to mean that at least one organism of that species has appeared within the boundaries of that pixel at some point during the specified time window.

This definition of species presence, including both temporal and spatial scales, is important when interpreting the estimated model (Guisan and Thuiller, 2005). However, its most immediate use is as a scaling of the model probabilities, via the population prevalence of the species π. This is the proportion of all locations in the region of interest where we would expect to record the species as present, given the spatial and temporal scales. The spatial scale and π are intrinsically linked, in that changing the spatial scale may change the proportion of locations with a presence. At the extreme, a very large scale would mean that the whole landscape was a single pixel, and we would have $\pi = 1$, assuming there was at least one individual present somewhere.

4.3 Presence-only data sampling

In practice, presence-only data are not collected according to a known sampling scheme. There is no random selection of locations, or any consideration of spatial/temporal scales. However, in order to create models of species presence, we need to make some assumptions about the sampling scheme. We start by setting out those assumptions, and then discuss a method of dealing with sampling bias.

Ideally the observed presences and the background data are representative of the true

presences for that species and the whole landscape respectively. This is achievable for the background data, which can be sampled at random using GIS. For the observed presences, we assume that for each location and time in $\mathcal{L}^{\mathcal{T}}$ where a species is present, it is equally likely that presence is observed. Statistically this means that locations with covariate value $\mathbf{x}$ occur in the sample of observed presences with probability proportional to $f_1(\mathbf{x}) = f(\mathbf{x}|y = 1)$. However, there is often no known sampling scheme for the observed presences. Records at herbaria or museums are typically ad-hoc records of species occurrence collected by different individuals, over long periods of time. They may be biased, for example, towards more accessible locations (Reese et al., 2005), but the major difficulty here is that we do not know the exact structure of that bias.

One way ecologists attempt to overcome this bias is by choosing the background data in a way that echoes the bias in sampling the observed presences, so that these two biases cancel each other out. This is true for presence-absence logistic modeling, given certain regularity conditions (Zadrozny, 2004), and holds for the presence-only problem for the same reasons. One approach (Dudík et al. (2006)) is to generate background data from observed presences collected across many species. This assumes that the data collected for other species in the same landscape has a similar sampling bias as that collected for the species of interest. Further, we must assume that those other species are sufficiently randomly dispersed across the landscape so that they accurately represent the overall sampling bias. In a similar approach, Zaniewski et al. (2002) models the other species' presence-only data to provide weights for the random sampling of the background data from the landscape.

4.4 Revisiting the longfin eel

In our example of the longfin eel, a natural definition of the landscape would be all freshwater within New Zealand. Alternatively, one could restrict the area to be freshwater

within only the North Island, or within a certain watershed. In our example, we will restrict this to rivers above a threshold of mean annual flow. The locations then are pixels or sections of all rivers, and the size of those sections may be determined by the pixelation in the data stored in GIS. As the eels do not have yearly migratory patterns, we can define the temporal scale to be 24 hours, and this will inform our estimation of the overall population prevalence π. In particular, π is the proportion of all locations in the landscape of rivers that we would expect to have any eel pass through within a given 24 hour period.

To illustrate how sampling bias may affect the estimated model, consider a scenario where our observed data are generated from ad-hoc sightings of the eel by ecologists as part of a number of other studies. The sampling locations for these studies will be chosen to suit the purpose of each study. If most studies are of large river or estuarine ecology, and there are no studies of mountain freshwater ecology, then we will not receive any ad-hoc observations of the longfin eel in rivers far from the sea. However, longfin eels are legendary climbers (New Zealand Statistics, 2005) and so may be found high up rivers. Using these data, we would assume that eels are not found far from the sea, and so our model would be biased.

If we knew the form of the bias in the sampling locations of the other studies, and the propensity of those ecologists to record ad-hoc observations, then we could adjust the model to account for this. However, we do not often know the form of this bias, especially if observed presences are collected from many different studies. One way to approximate this bias is by using ad-hoc observations of other species found in freshwater as the background data; they provide evidence that someone was there looking for something. However, they do not tell us where people were looking but found no species, and so this may lead to overestimation of prevalence at such locations.

Chapter 5

Existing Presence-Only Models

A comparison of many models that have been applied to presence-only data is given in Elith et al. (2006). Of these, two of the best performing models were Maxent and a naive boosted trees model, which together form a starting point for the work presented in this thesis. In this section we introduce these models, plus the exponential model of Manly, which is related to Maxent. In each method, the aim is to ultimately create a model of $\mathbb{P}(y = 1|\mathbf{x})$, the probability of a true presence y at a location, given covariates $\mathbf{x}$. However, the presence-only data are generated by $\mathbb{P}(z = 1|\mathbf{x}, s = 1)$, the conditional probability of an observed presence $z = 1$. It is this observation that illustrates the core of the presence-only problem.

To illustrate connections and differences between methods, we start by providing a categorization in terms of the choice of link function and the selection of the background data (Table 5.1). The link function defines how the probability of interest $\mathbb{P}(y = 1|\mathbf{x})$ is related to the linear predictor $\eta(\mathbf{x})$. In this thesis we consider two link functions, the logistic link

	Sample likelihood of $P(z = 1|s = 1, \mathbf{x})$	**Population** density estimation of $f_1(\mathbf{x})$
Logistic link $\mathbb{P}(y = 1|\mathbf{x}) = \frac{e^{\eta(\mathbf{x})}}{1+e^{\eta(\mathbf{x})}}$	Naive model Lancaster & Imbens	
Log link $\mathbb{P}(y = 1|\mathbf{x}) = e^{\eta(\mathbf{x})}$	Manly's exponential model	Maxent

Figure 5.1: *The modeling procedures used in the presence-only problem can be categorized in two ways. The first is by the use of a log or logistic link function relating* $\mathbb{P}(y = 1|\mathbf{x})$ *and* $\eta(\mathbf{x})$*. The second is whether the background data are assumed to be a sample or the full population of locations.*

function

$$\text{logit}\left(\mathbb{P}(y = 1|\mathbf{x})\right) = \eta(\mathbf{x}) \qquad \Longrightarrow \qquad \mathbb{P}(y = 1|\mathbf{x}) = \frac{e^{\eta(\mathbf{x})}}{1 + e^{\eta(\mathbf{x})}} \tag{5.1}$$

and the log link function

$$\log(\mathbb{P}(y = 1|\mathbf{x})) = \eta(\mathbf{x}) \qquad \Longrightarrow \qquad \mathbb{P}(y = 1|\mathbf{x}) = e^{\eta(\mathbf{x})} \tag{5.2}$$

The generation of the background data is assumed to occur in two ways. The background data for the *sample* method is taken to be a simple random sample of all locations in the landscape. The *population* method is to take the background data to be the full population of locations, equivalently all pixels in the landscape. This distinction leads to two slightly different approaches in the modeling procedures. In the sample method we take the familiar presence-absence modeling approach of maximizing a conditional likelihood, which here is based on $\mathbb{P}(z = 1|\mathbf{x}, s = 1)$. However, when we know the full landscape of locations,

this implies we know the density $f(\mathbf{x})$, and hence the variability in the data is in the form of $f_1(\mathbf{x}) = \mathbb{P}(\mathbf{x}|y = 1)$. Thus the problem becomes one of density estimation, where the form of $f_1(\mathbf{x})$ depends on $f(\mathbf{x})$ and the link function. To keep the notation simple, we make no distinction in notation between these two different types of background data. For both we use $z = 0$ to denote the background data, and let n_a be the total number of background observations; for the population models, n_a is equal to the cardinality of the landscape $\mathcal{L}$.

5.1 The naive presence-only model

Logistic models are widely used in habitat modeling when the true presences and absences are known for a sample of locations (Keating and Cherry, 2004). This presence-absence logistic model is

$$\mathbb{P}(y = 1|\mathbf{x}) = \frac{e^{\eta(\mathbf{x})}}{1 + e^{\eta(\mathbf{x})}},$$

for some function η of the covariates $\mathbf{x}$. For example, $\eta(\mathbf{x})$ can be linear in $\mathbf{x}$ (as in logistic regression) or a non-linear function e.g. GAMs (Hastie and Tibshirani, 1990) or boosted trees (Friedman, 2001). Our aim in presence-only modeling is to estimate the η from this presence-absence model, using presence-only data.

Some ecologists (e.g. Parker and Anderson, 2003) treat the observed presences and background data $\mathbf{z}$ as if they are the true presences and absences $\mathbf{y}$. We call this the *naive model*, and it fails in two ways. Firstly, the observed presences and background data are not sampled at the same rate as true presences and absences occur in the population. We will show that this can be overcome using a case-control model, and if the overall proportion of presence locations π is known, probability estimates can be corrected. The second problem is that some of the background data are really presences. This is sometimes referred to as *contamination of the controls* (Lancaster and Imbens, 1996), and can lead to a large bias in the case-control adjusted naive model.

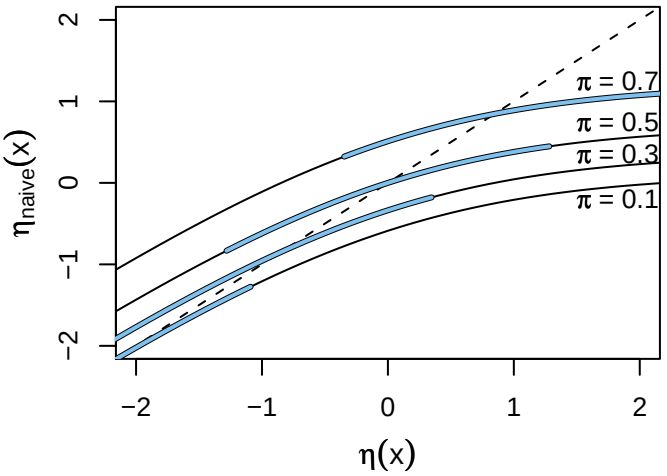


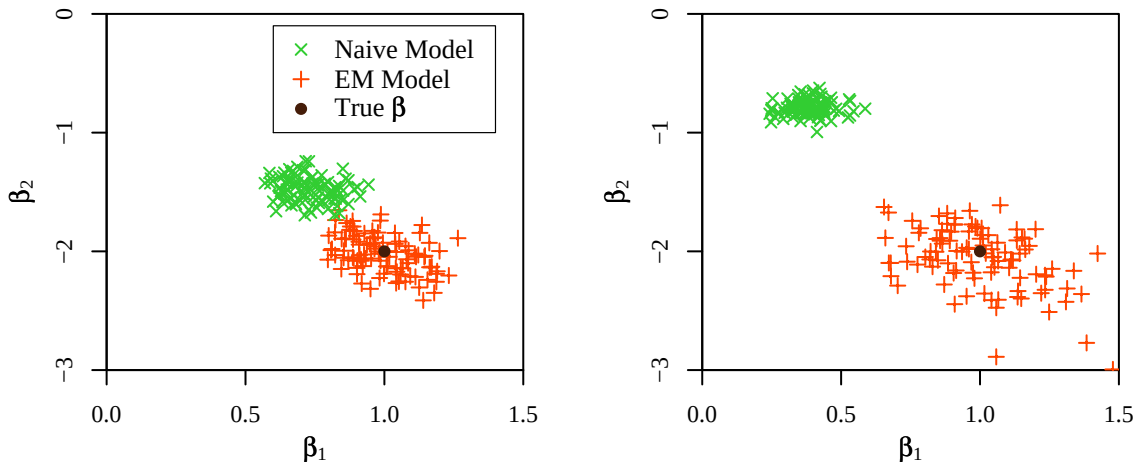
Figure 5.2: *The case-control adjusted naive model $\eta^{\star}_{naive}$ is an increasing but non-linear function of the true model of interest, η, and the population prevalence π. The blue lines indicate the typical range of η that occurs for each level of π.*

We will analyze the form of the bias in the case-control adjusted naive model in Section 6.6. There we show that the case-control adjusted naive model $\eta^{\star}_{\text{naive}}$ can be written in terms of the presence-absence model η as follows:

$$\eta^{\star}_{\text{naive}}(\mathbf{x}) = -\log(1-\pi) - \log\left(e^{-\eta(\mathbf{x})} + 1\right) \tag{5.3}$$

Thus $\eta^{\star}_{\text{naive}}$ is increasing in the true η, but nonlinearly (Figure 5.2). Although the true and naive predictors are similar up to a constant for $\eta \ll 0$, when $\eta > 0$ the naive model considerably underestimates the rate at which η is increasing. In particular, the naive linear predictor is bounded above: $\eta_{\text{naive}} \leq -\log(1-\pi)$. Hence the estimated probabilities for the naive model will be underestimated for locations with higher probabilities of presence.

This bias has considerable effect on logistic regression estimates, where the true η is assumed linear in $\mathbf{x}$, i.e. $\eta(\mathbf{x}) = \mathbf{x}^T\beta$. In this case, any naive logistic regression model must be biased. In particular, the estimates for the naive model β_{naive} will tend to underestimate the slopes β. This is illustrated in simulations of presence-only data (Figure 5.3). Although the estimates for the EM model (introduced in Section 6.4) are approximately unbiased,



(a) π=0.1 **(b)** π=0.4

Figure 5.3: *Parameter estimates for the naive logistic regression model are biased toward the origin, with increased bias for larger π. The EM procedure (Section 6.4) produces approximately unbiased estimates, but with increasing variance for larger π. These estimates are from 100 simulations of presence-only data, generated from the model $\eta(\mathbf{x}) = \alpha + x_1\beta_1 + x_2\beta_2$, where $\beta_1 = 1$ and $\beta_2 = -2$. The $\mathbf{x}$ are i.i.d. standard normals and $n_p = 300$ and $n_a = 1000$.*

the naive model estimates are shrunk to zero, with more shrinkage for larger π. However, note that even for $\pi = 0.1$, the naive estimates are noticeably different from the truth. It is easy to see how this shrinkage may impact variable selection, e.g. important variables may not appear statistically significant.

At this point it may seem tempting to use the relationship between the true and naive models (5.3) to make a post-hoc transformation of the naive model. Certainly, for a flexible fitting procedure where the relationship between η and $\mathbf{x}$ is not predetermined, this is theoretically possible. However, the form of the relationship introduces complex interactions between variables in the naive model. Additionally, from (5.3) we see there is an implicit constraint of $\eta_{\text{naive}} < -\log(1-\pi)$; exceeding this value results in undefined η. In practice this constraint is not observed, nor is there an easy way to enforce this. Alternatively, we can think of the undefined η as representing a probability greater than 1. To see this, we derive the form of the true probability $\mathbb{P}(y = 1|\mathbf{x})$, for the range of values $\eta^{\star}_{\text{naive}} < -\log(1-\pi)$,

given $\eta^{\star}_{\text{naive}}$ from (5.3):

$$e^{-\eta^{\star}_{\text{naive}}(\mathbf{x})} = (1-\pi)\left(e^{-\eta(\mathbf{x})}+1\right)$$

$$\Rightarrow \mathbb{P}(y=1|\mathbf{x}) = \frac{1}{e^{-\eta(\mathbf{x})}+1} = (1-\pi)\,e^{-\eta^{\star}_{\text{naive}}(\mathbf{x})}. \tag{5.4}$$

Thus even when $\eta_{\text{naive}} > -\log(1-\pi)$, we can get an estimate of $\mathbb{P}(y=1|\mathbf{x})$ that is greater than 1. Although this is not a valid probability, we can at least use this in assessing the models in Chapter 8. We will see that this method does not perform well in practice.

5.2 Manly's exponential model

We briefly discuss this model as it assumes a similar model structure to Maxent. Further, while Maxent has not yet generated much response in the ecological modeling literature, the form of Manly's model has been well critiqued. The exponential model, or equivalently log-probability model, is $\mathbb{P}(y=1|\mathbf{x}) = e^{\eta(\mathbf{x})}$, which is equivalent to the log link (5.2). The primary benefit of using this model is that η from this *log* model is equivalent, up to an additive constant of $\log(\frac{\pi n_a}{n_p})$, to the η obtained by fitting the naive *logistic* model. This follows immediately from the form for the probability

$$\mathbb{P}(z=1|s=1,\mathbf{x}) = \frac{\frac{\pi n_a}{n_p}e^{\eta(\mathbf{x})}}{1+\frac{\pi n_a}{n_p}e^{\eta(\mathbf{x})}} \tag{5.5}$$

which is provided in Manly et al. (2002, Section 5.4), and reproduced in Appendix A.1 using our notation for clarity. If we know π we can make the additive adjustment to the naive logistic η. However, even if we don't know π, we can estimate the probability $\mathbb{P}(y=1|\mathbf{x})$ up to a multiplicative constant.

Although the exponential model is easy to fit, the form of the model is problematic. Keating and Cherry (2004) use simulations to illustrate that this model does not guarantee valid

probability estimates; an average of 6% of the estimated probabilities exceeded 1 in over 50% of simulations. In addition, the estimated probabilities, referred to as the resource selection function (RSF), were not proportional to the true probabilities. A typical approach for use-availability or presence-only data is to use the RSF in place of actual probabilities when the overall prevalence of a species π is unknown. Ecologists then use the RSF in presenting results, with the assumption that the RSF is proportional to the true probability. At this point we note that in examples based on use-availability data, Johnson et al. (2006) found the bias in the RSF to be negligible unless π is large (> 0.25).

5.3 Maxent

Maxent (Phillips et al., 2006) is a method of estimating species distribution from presence-only data, that has been shown to work well in practice (e.g. Elith et al., 2006). A maximum entropy argument is used to justify a log probability link function $\log \mathbb{P}(y = 1|\mathbf{x}) = \eta(\mathbf{x}) + \delta$. In this section we describe Maxent in statistical language, and compare it to other statistical models. We reference the maximum entropy calculations, and discuss some of the advantages and disadvantages of this model.

The aim of Maxent is to estimate the probability density $f_1(\mathbf{x})$ of the covariates across locations with true presences, by comparing f_1 to f, the density over all locations. We expect these densities to be similar because they have many locations in common, and we use the observed presences to determine how they differ. For example, if one covariate is elevation above sea level, then we might require that the mean elevation for the estimated f_1 is similar to the mean elevation in the locations with observed presences. Thus we could estimate f_1 by minimizing the difference between f_1 and f subject to constraining those mean elevations to be close. In Maxent, this difference is taken to be the relative entropy of f_1 with respect to f (also known as the Kullback-Liebler divergence). To increase model

specificity, we can extend the constraints on the means of covariates to encompass constraints on higher order moments, cross products or quantiles.

To present this modeling in statistical language, we define $h(\mathbf{x})$ to be the vector of basis functions upon which approximate constraints are defined. It can be shown (Della Pietra et al., 1997) that maximizing the relative entropy of f_1 to f subject to those constraints is equivalent to maximum likelihood estimation of f_1, with L_1 constraints on the parameters β, where f_1 is assumed to follow a Gibb's distribution which takes the form

$$f_1(\mathbf{x}) = f(\mathbf{x})\, e^{\eta(\mathbf{x})} \tag{5.6}$$

and where $\eta(\mathbf{x}) = \alpha + h(\mathbf{x})^T\beta$, for a normalizing constant α.

There are several advantages to this approach. Maxent has a simple interpretation in terms of the probability of interest: using Bayes rule and (5.6),

$$\mathbb{P}(y = 1|\mathbf{x}) = \frac{f_1(\mathbf{x})\mathbb{P}(y = 1)}{f(\mathbf{x})} = \pi\, e^{\eta(\mathbf{x})}. \tag{5.7}$$

When π is unknown, the model structure means that probabilities can be estimated up to a multiplicative constant to provide a relative likelihood of species presence. This is a huge practical benefit as the true population prevalence may be hard to estimate in practice. In addition, the fitting procedure automatically performs variable selection via the L_1 regularization. This technique is widely used in statistics (e.g. the lasso, Tibshirani, 1996) as a method of shrinking parameter estimates to decrease variance in predictions. The tuning parameter of the L_1 penalty affects the number of estimated parameters that are set to zero.

However, this model has the same failings as Manly's exponential model, because of the log link function. Here we take time to illustrate the issue more carefully. From (5.7), we see that for $\mathbb{P}(y = 1|\mathbf{x}) \leq 1$ to be true, we require $\eta(\mathbf{x}) \leq -\log(\pi)$. However, the only

constraint on η is given by the requirement that the density integrate to 1:

$$\int_{\mathbf{x}} f_1(\mathbf{x}) = 1 \qquad \Longrightarrow \qquad \int_{\mathbf{x}} f(\mathbf{x})\, e^{\eta(\mathbf{x})} = 1.$$

The finite landscape version of this is $\sum_{\mathcal{L}} \frac{1}{n_a} e^{\eta(\mathbf{x})} = 1$, which implies $\eta(\mathbf{x}) < \log(n_a)$. In practice, $n_a \gg \pi^{-1}$, so the model may generate probabilities that are larger than 1. Another way to view this problem, is via the model for $f_0(\mathbf{x})$. We know that

$$f(\mathbf{x}) = \pi\, f_1(\mathbf{x}) + (1 - \pi)\, f_0(\mathbf{x})$$

and plugging in (5.6) we get

$$f_0(\mathbf{x}) = \frac{1 - \pi\, e^{\eta(\mathbf{x})}}{1 - \pi} f(\mathbf{x}).$$

Thus, the density $f_0(\mathbf{x})$ is negative if $\eta(\mathbf{x}) > -\log(\pi)$.

Phillips et al. (2006) recognize that the model can result in estimated probabilities a lot larger than 1, outside the landscape of the problem. However, they also claim that Maxent models a probability distribution over the landscape. We have already seen that the probability distribution estimated may be invalid, because it may imply that $f_0(\mathbf{x})$ is negative. We will also see (Chapter 8) that the model commonly estimates probabilities greater than 1 for locations *within* the landscape. We agree that Maxent can provide a good ranking of locations with respect to habitat suitability and because of this it performs well when measured using AUC, the area under the ROC curve (Elith et al., 2006). However, this model suffers from many of the same problems as that of Manly et al. (2002), because of the log link structure of the model.

We end this discussion of log link models by noting that for large negative $\eta(\mathbf{x})$, the log and logistic links may give similar results. When $\eta(\mathbf{x}) \ll 0$, $\mathbb{P}(y = 1|\mathbf{x}) = p$ is small, so $(1 - p)$ will be close to 1, and hence $\log(p/(1 - p) \approx \log(p)$. Thus for rare species, with small π, where most of the $\eta(\mathbf{x})$ are negative, the log models, including Maxent, may provide a

good approximation to the logistic models.

5.4 Gradient boosting and regression trees

Many of the models introduced in this thesis use gradient boosted regression trees, so we end this chapter with a brief description of boosting and regression trees. A good summary of boosting can be found in Hastie et al. (2001, chapter 10). Boosting is a statistical method of fitting an additive model of basis functions $h(\mathbf{x}; \gamma)$,

$$\eta(\mathbf{x}) = \sum_{j=1}^{J} \beta_j h(\mathbf{x}; \gamma_j),$$

where we measure the cost of error in the estimate using a loss function $\mathbf{L}(\mathbf{y}, \eta; X)$, for an outcome $\mathbf{y}$. The basis functions h can be any function of the covariates $\mathbf{x}$; in this thesis we will use regression trees which we describe later. To fit the model we will use gradient boosting (Friedman, 2001), a stepwise algorithm that fits a model using a steepest descent method on the loss function with respect to η. The loss function in our examples will be the negative log-likelihood of the model,

$$\mathbf{L}(\mathbf{y}, \eta; X) = -\log L(\mathbf{y}, \eta; X),$$

occasionally with penalization functions that apply to specific examples.

At each iteration of the fitting procedure, a basis function is fit to the current working residuals. Shrinkage is applied to this basis function, which is then added to the current model. The current working residuals are the negative first derivative of the loss function defined for each observation, given the estimated model from the previous step. A basis function is fit to these residuals using squared error loss, and then the optimal parameters

for this basis function are estimated using the original loss function. In our examples, this second step cannot be directly solved, so an approximation is used, which involves taking a single Newton step.

An additional problem arises with two of our three models; the loss function is not convex. Although this is not an issue with the boosting procedure itself, it can cause problems when taking a Newton step, because the gradient may be positive. To circumvent this problem, we adjust the calculations of the gradient so that the gradient calculated for each data point, $\frac{d^2}{d\beta^2} \log L_i$, is forced to be non-negative. Although these adjusted Newton steps, and the Newton step itself, are only approximations, the stepwise property of boosting can auto-correct any errors introduced this way. The amount of shrinkage applied to this model can be adjusted in the implementation of the procedure - typically we use a multiplier between 0.001 and 0.01. As the training set loss is reduced at each step, the optimal number of trees is chosen by minimizing the loss on a separate test set.

One popular basis function to use in boosting is a regression tree, which provides a rectilinear piecewise constant estimate across the space of covariates and includes interactions between variables up to a level specified in the implementation. Boosted regression trees can provide flexible estimates, and have recently been used in estimating logistic models of species presence from presence-absence data (Leathwick et al., 2006; De'ath, 2007).

Chapter 6

The Logistic Presence-Only Model

In this section we approach the problem in the usual conditional likelihood manner of e.g. Lancaster and Imbens (1996). In terms of the categorization of Figure 5.1, the EM model we develop in this section falls into the top left corner. In particular, we model the probability that a species of interest is present, $y = 1$, conditional on some ecological covariates $\mathbf{x}$, $\mathbb{P}(y = 1|\mathbf{x})$, using a logistic link function:

$$\text{logit}\, \mathbb{P}(y = 1|\mathbf{x}) = \eta(\mathbf{x}) \;\; \Rightarrow \;\; \mathbb{P}(y = 1|\mathbf{x}) = \frac{e^{\eta(\mathbf{x})}}{1 + e^{\eta(\mathbf{x})}}. \tag{6.1}$$

In studies where the true presences and absences are known for a random sample of locations, these models are fit using established methods, for a range of different assumptions about the structure of η.

As mentioned in Section 5.1, there are two reasons why we cannot use the established methods directly on the observed presences and background data z. Firstly, there is missing information about the true presence or absence at each location in the background sample. Secondly, even if the true presences/absences are known for the background data, we

do not have a simple random sample of locations; the combination of observed presences and background data have a higher proportion of presences than occur in the population. In Section 6.2 we investigate how to overcome the second problem using a modified case-control methodology, which we then use in Section 6.3 to find the easily-optimized *full likelihood* for the true and observed presences and background data. We then illustrate an application of the EM algorithm to the full likelihood, which helps us overcome the problem of missing information in a simple and elegant way (Section 6.4).

In Section 6.5 we see how computational efficiency of the EM algorithm can be increased for stepwise logistic modeling procedures. Finally, we prove the form of the bias in the naive model (Section 6.6). Throughout this chapter we assume that we know the overall population prevalence of the species $\pi = \mathbb{P}(y = 1)$. In Section A.2 we illustrate under what conditions we can also estimate π directly from these data, and in Chapter 8 look at an example of sensitivity analysis of uncertainty in π. However, we start by deriving the presence-only logistic likelihood.

6.1 The presence-only logistic likelihood

To estimate η for the presence-absence logistic model (6.1) we wish to maximize the likelihood for the presence-only data, with respect to η. This likelihood is based on $\mathbb{P}(z = 1|s = 1, \mathbf{x})$; the probability that a data point in our presence-only data is an observed presence, given environmental covariates $\mathbf{x}$. We refer to this likelihood as the *presence-only logistic likelihood*.

To write $\mathbb{P}(z = 1|s = 1, \mathbf{x})$ in terms of η we start by using a total-probability argument across

$y = 0, 1$:

$$\begin{aligned}\mathbb{P}(z = 1|s = 1, \mathbf{x}) =& \mathbb{P}(z = 1|y = 1, s = 1, \mathbf{x})\mathbb{P}(y = 1|s = 1, \mathbf{x}) \\ &+ \mathbb{P}(z = 1|y = 0, s = 1, \mathbf{x})\mathbb{P}(y = 0|s = 1, \mathbf{x}).\end{aligned} \tag{6.2}$$

We can think of $\mathbb{P}(z = 1|y = 1, s = 1, \mathbf{x})$ as a sampling probability of whether true presences in our presence-only data were observed ($z = 1$) or not ($z = 0$). The random sampling assumptions (Section 4.3) imply that the sampling of the observed presences was assumed independent of $\mathbf{x}$, given the true presence or absence. Thus we use an application of Bayes rule to show

$$\mathbb{P}(z = 1|y = 1, s = 1, \mathbf{x}) = \frac{\mathbb{P}(z = 1, y = 1|s = 1)}{\mathbb{P}(y = 1|s = 1)} = \frac{n_p}{n_p + \pi n_a} \tag{6.3}$$

Further, $\mathbb{P}(z = 1|y = 0, s = 1) = 0$ because all true absences in the data must occur in the background data.

We will show in the next section that $\mathbb{P}(y = 1|s = 1, \mathbf{x})$ is a case-control probability (6.6) with offset given by (6.8). Substituting all these probabilities in (6.2), we have

$$\begin{aligned}\mathbb{P}(z = 1|s = 1, \mathbf{x}) &= \frac{n_p}{n_p + \pi n_a} \frac{e^{\eta^*(\mathbf{x})}}{1 + e^{\eta^*(\mathbf{x})}} + 0 \\ &= \frac{\frac{n_p}{\pi n_a} e^{\eta(\mathbf{x})}}{1 + \left(1 + \frac{n_p}{\pi n_a}\right) e^{\eta(\mathbf{x})}}.\end{aligned} \tag{6.4}$$

Thus the probability of an observed presence is a multiple of the true probability of a presence (recorded or not), with an adjusted intercept of η; it is the combination of these two adjustments that make the presence-only problem difficult to solve.

The presence-only likelihood is the product of the probabilities of the observed presences

and background data $\mathbf{z}$:

$$L(\eta|\mathbf{z}, X) = \prod_i \mathbb{P}(z_i|s_i = 1, \mathbf{x}_i)$$
$$= \prod_i \left(\frac{\frac{n_p}{\pi n_a} e^{\eta(\mathbf{x}_i)}}{1 + \left(1 + \frac{n_p}{\pi n_a}\right) e^{\eta(\mathbf{x}_i)}} \right)^{z_i} \left(\frac{1 + e^{\eta(\mathbf{x}_i)}}{1 + \left(1 + \frac{n_p}{\pi n_a}\right) e^{\eta(\mathbf{x}_i)}} \right)^{1-z_i}. \quad (6.5)$$

This likelihood can be maximized directly in certain situations, and this is the approach that Lancaster and Imbens (1996) took for a model with $\eta(\mathbf{x})$ linear in $\mathbf{x}$. However, for more general η, direct maximization is not always feasible. In the next few sections we will develop an application of the EM algorithm that provides an elegant solution to solving this problem.

6.2 Modified case-control sampling

In this section we assume the hypothetical situation where all true presences and absences are known for the background data. We will show that in this circumstance we can estimate the logistic model (6.1) using a modified case-control argument. We later use this property in an EM algorithm approach (Section 6.4) for solving the presence-only problem. To develop the modified case-control argument, we start by introducing the traditional application of case-control sampling in logistic modeling. We then apply this to our hypothetical situation.

Traditional case-control sampling (e.g. McCullagh and Nelder, 1989, p111) is a retrospective sampling technique whereby cases and controls are sampled separately from their respective populations, at different and unknown rates. A typical application is in the study of a disease, where the cases ($y = 1$) are selected at random from all people with

the disease and the controls ($y = 0$) are sampled at random from all non-diseased individuals. We use $s = 1$ to indicate that a data point is selected to be in that sample of cases and controls. Then, following the usual case-control notation, we write the sampling rates for the cases and controls as $\gamma_1 = P(s = 1|y = 1)$ and $\gamma_0 = P(s = 1|y = 0)$ respectively. We assume that γ_1 and γ_0 are both independent of the covariates $\mathbf{x}$. Let n_1 and n_0 denote the number of cases and controls respectively, and N denote the size of the population. Then if $\pi = \mathbb{P}(y = 1)$ is the proportion of cases in the population, the total number of cases in the population is $N\pi$. We are choosing n_1 of the possible $N\pi$ cases at random, so by definition,

$$\gamma_1 = \frac{n_1}{\pi N} \quad \text{and similarly} \quad \gamma_0 = \frac{n_0}{(1-\pi)N}.$$

As we are only interested in the ratio of γ_1 to γ_0, the true size of N is unimportant.

The case-control probability $P(y = 1|s = 1, \mathbf{x})$ is written conditional on the event that a data point is in the sample of cases and controls. This allows us to adjust for the different sampling rates. Following the usual case-control calculations (McCullagh and Nelder, 1989, p111) an application of Bayes rule gives

$$\begin{aligned}
\mathbb{P}(y = 1|s = 1, \mathbf{x}) &= \frac{\mathbb{P}(s = 1|y = 1, \mathbf{x})\,\mathbb{P}(y = 1|\mathbf{x})}{\mathbb{P}(s = 1|y = 0, \mathbf{x})\,\mathbb{P}(y = 0|\mathbf{x}) + \mathbb{P}(s = 1|y = 1, \mathbf{x})\,\mathbb{P}(y = 1|\mathbf{x})} \\
&= \frac{\gamma_1 e^{\eta(\mathbf{x})}}{\gamma_0 + \gamma_1 e^{\eta(\mathbf{x})}} \\
&= \frac{e^{\eta^*(\mathbf{x})}}{1 + e^{\eta^*(\mathbf{x})}}
\end{aligned} \tag{6.6}$$

where $\eta^*(\mathbf{x}) = \eta(\mathbf{x}) + \log(\gamma_1/\gamma_0)$. Plugging in the sampling rates gives

$$\eta^*(\mathbf{x}) = \eta(\mathbf{x}) + \log\left(\frac{n_1}{n_0}\right) - \log\left(\frac{\pi}{1-\pi}\right). \tag{6.7}$$

Thus the case-control model is equivalent to the true model (6.1), but where the intercept is adjusted to account for the sampling rates. Even when the population prevalence π is unknown, η is identifiable up to a constant.

We now show how this applies in our hypothetical situation, where we assume that we know the true presences and absences for all our presence-only data. We start by calculating γ_1 and γ_0 for this setup, in terms of n_p and n_a, the number of observed presences and background data respectively. The expected number of true presences in our hypothetical situation is $n_p + \pi n_a$, which is all the observed presences plus a proportion π of the background data. As there are $n_p + n_a$ total data points, the expected proportion of true presences in our data is

$$\mathbb{P}(y = 1|s = 1) = \frac{n_p + \pi n_a}{n_p + n_a} \quad \text{and similarly} \quad \mathbb{P}(y = 0|s = 1) = \frac{(1-\pi)n_a}{n_p + n_a}.$$

An application of Bayes rule gives

$$\gamma_1 = \mathbb{P}(s = 1|y = 1) = \frac{\mathbb{P}(y = 1|s = 1)\,\mathbb{P}(s = 1)}{\mathbb{P}(y = 1)} = \frac{n_p + \pi n_a}{\pi\,(n_p + n_a)}\,\mathbb{P}(s = 1)\,, \text{ and}$$

$$\gamma_0 = \mathbb{P}(s = 1|y = 0) = \frac{\mathbb{P}(y = 0|s = 1)\,\mathbb{P}(s = 1)}{\mathbb{P}(y = 0)} = \frac{(1-\pi)n_a}{(1-\pi)(n_p + n_a)}\,\mathbb{P}(s = 1).$$

and thus the offset in the modified case-control model is

$$\log\left(\frac{\gamma_1}{\gamma_0}\right) = \log\left(\frac{n_p + \pi n_a}{(1-\pi)n_a}\right) - \log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{n_p + \pi n_a}{\pi n_a}\right). \tag{6.8}$$

This is equivalent to (6.7) replacing n_1 and n_0 with $n_p + \pi n_a$ and $(1-\pi)n_a$, the expected number of true presences/absences. As we are assuming that π is known, the true model η can be estimated outright.

To summarize, if we knew the true presences/absences for all the data, we could estimate the model of interest (6.1) by fitting the usual model to these data and then subtracting the constant $\log\left(\frac{n_p+\pi n_a}{\pi n_a}\right)$ from $\eta(\mathbf{x})$. This will be the maximization step in an EM algorithm, but to show this we first need to consider the likelihood of the full data.

6.3 The full likelihood

Continuing with the scenario that we know all true presences and absences, we now write out the *full likelihood*. This likelihood is written in terms of the joint probability of the true presences and absences $\mathbf{y}$ and the observed presences and background data $\mathbf{z}$, conditional on X:

$$L(\eta|\mathbf{y},\mathbf{z},X) = \prod_i \mathbb{P}(y_i,z_i|s_i=1,\mathbf{x}_i) = \prod_i \mathbb{P}(y_i|s_i=1,\mathbf{x}_i)\,\mathbb{P}(z_i|y_i,s_i=1,\mathbf{x}_i) \tag{6.9}$$

This factors into the case-control probabilities $\mathbb{P}(y_i|s_i=1,\mathbf{x}_i)$ and the sampling probabilities for the observed presences and background data, conditional on the true presence, $\mathbb{P}(z_i|y_i,s_i=1,\mathbf{x}_i)$. From (6.3), we know $\mathbb{P}(z_i=1|y_i=1,s_i=1,\mathbf{x}) = \frac{n_p}{n_p+\pi n_a}$ and hence $\mathbb{P}(z_i=0|y_i=1,s_i=1,\mathbf{x}_i) = \frac{\pi n_a}{n_p+\pi n_a}$. Further, $\mathbb{P}(z_i=0|y_i=0,s_i=1,\mathbf{x}_i)=1$ because there are no true absences in the sample of observed presences. As none of these sampling probabilities depend on η, they only appear in the likelihood as a constant, and are irrelevant in this scenario. However, we will re-visit these in Section 6.7 when we consider simultaneous estimation of π.

Thus the likelihood depends only on the case-control probabilities:

$$L(\eta|\mathbf{y},\mathbf{z},X) \propto \prod_i \mathbb{P}(y_i|s_i=1,\mathbf{x}_i). \tag{6.10}$$

To estimate η, we fit a model to the true presences $\mathbf{y}$ to obtain $\hat{\eta}^*$, and then subtract the constant $\log\left(\frac{n_p+\pi n_a}{\pi n_a}\right)$. Although we do not know the true presences/absences for the background data, in the next few sections we will see that we can still estimate η this way, using an application of the EM algorithm.

6.4 Model 1: the EM model

The Expectation-Maximization or EM algorithm (Dempster et al., 1977) is a technique for maximizing a likelihood where there are some latent or unknown variables. In this example, the latent variables are the true presences and absences $\mathbf{y}$ for the background data $z = 0$. The likelihood is maximized by alternating between expectation and maximization steps. The expectation of the log-likelihood is taken with respect to the parameter estimates obtained in the previous step. Then that expected log-likelihood is maximized to re-estimate the parameters. The rest of this section develops the EM algorithm for this case; the procedure is summarized in Figure 6.1.

From (6.10), the full likelihood of both $\mathbf{z}$ and $\mathbf{y}$ is

$$L(\eta|\mathbf{y}, \mathbf{z}, X) \propto \prod_i \mathbb{P}(y_i = 1|s_i = 1, \mathbf{x}_i)^{y_i} \left(1 - \mathbb{P}(y_i = 1|s_i = 1, \mathbf{x}_i)\right)^{1-y_i}$$

and so the log-likelihood is

$$\log L(\eta|\mathbf{y}, \mathbf{z}, X) = \sum_i \left(y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right) + \text{ constant} \tag{6.11}$$

where $p_i = \mathbb{P}(y_i = 1|s_i = 1, \mathbf{x}_i)$.

In the E step of the kth iteration, we take the expectation of the log-likelihood, conditional on the estimated model from the previous step $\eta^{(k-1)}$. As the log-likelihood is linear in the y_i, we replace the unknown y_i for the background data with their expectation $\hat{y}_i^{(k)}$:

$$\hat{y}_i^{(k)} = \mathbb{E}[y_i|\eta^{(k-1)}, \mathbf{x}_i] = \mathbb{P}(y_i = 1|\eta^{(k-1)}, \mathbf{x}_i) = \frac{e^{\eta^{(k-1)}(\mathbf{x}_i)}}{1 + e^{\eta^{(k-1)}(\mathbf{x}_i)}}, \quad \text{for } z_i = 0.$$

Further, because we know that $y = 1$ for the observed presences, we have $\hat{y}_i^{(k)} = 1$ for $z_i = 1$.

1. Chose initial estimates $\hat{y}_i^{(0)} = \pi$ for $i : z_i = 0$.

2. Repeat until convergence:

- *Maximization step:*
  - Calculate $\hat{\eta}^{*(k)}$ by fitting either
    1. a logistic model of $\hat{\mathbf{y}}^{(k-1)}$ on X
    2. a logistic model of $\begin{pmatrix} \mathbf{1}_P \\ \mathbf{1_A} \\ \mathbf{0_A} \end{pmatrix}$ on $\begin{pmatrix} X_P \\ X_\mathbf{A} \\ X_\mathbf{A} \end{pmatrix}$ with weights $\begin{pmatrix} \mathbf{1}_P \\ \hat{\mathbf{y}}_\mathbf{A}^{(k-1)} \\ \mathbf{1_A} - \hat{\mathbf{y}}_\mathbf{A}^{(k-1)} \end{pmatrix}$
  - Calculate $\hat{\eta}^{(k)} = \hat{\eta}^{*(k)} - \log(\frac{n_p + \pi n_a}{(1-\pi)n_a}) - \log(\frac{\pi}{1-\pi})$.
- *Expectation step:*

$$\hat{y}_i^{(k)} = \frac{e^{\hat{\eta}^{(k)}} + 1}{1 + e^{\hat{\eta}^{(k)}} + 1} \text{ for } z_i = 0 \quad \text{and} \quad \hat{y}_i^{(k)} = 1 \text{ for } z_i = 1$$

Figure 6.1: *Estimation of the contaminated case-control model via the EM algorithm, when π is known.*

The maximization step involves estimation of a case-control logistic model as described in Section 6.2, fitted to $\hat{y}_i^{(k)}$ of the observed presences and background data. Any logistic model fitting procedure can be used, as long as that procedure can handle either weights, or non-integer responses. That is because the $\hat{y}_i^{(k)}$ for the background data lie in the interval $[0, 1]$ and are not binary. If the procedure can handle non-integer responses, then the maximization step is straightforward. Fit $\hat{\mathbf{y}}^{(k)}$ to X, to obtain $\eta^{*(k)}$. Then $\eta^{(k)} = \eta^{*(k)} - \log(\frac{n_p + \pi n_a}{\pi n_a})$.

If the fitting procedure can handle weights but not non-integer responses, the maximization step involves fitting an augmented $\hat{\mathbf{y}}^{(k)}$ to an augmented X. To see this, we look at the

expected log-likelihood, taking expectation over (6.11) to give

$$\begin{aligned}\mathbb{E}[\log L(\eta; \mathbf{y}, \mathbf{z}, X)|\eta^{(k)}] &= \sum_i \left(\hat{y}_i^{(k)} \log(p_i) + (1 - \hat{y}_i^{(k)}) \log(1 - p_i)\right) + \text{ const} \\ &= \sum_{z_i=1} \log p_i + \sum_{z_i=0} w_{k,i} \log p_i + \sum_{z_i=0} w_{k,i}^* \log(1 - p_i) + \text{ const},\end{aligned}$$

where $w_{k,i} = \hat{y}_i^{(k)}$. It is apparent that this is equivalent to a weighted logistic model with $n_p + 2n_a$ data points, of which $n_p + n_a$ data points have an outcome of 1, and n_a data points have an outcome of 0. The background data now appear twice, with outcomes 1 and 0 and corresponding weights $\hat{y}_i^{(k)}$ and $1 - \hat{y}_i^{(k)}$. More concretely, the maximization step is fitting:

$$\text{a logistic model of } \begin{pmatrix} \mathbf{1}_P \\ \mathbf{1}_\mathbf{A} \\ \mathbf{0}_\mathbf{A} \end{pmatrix} \text{ on } \begin{pmatrix} X_P \\ X_\mathbf{A} \\ X_\mathbf{A} \end{pmatrix} \text{ with weights } \begin{pmatrix} \mathbf{1}_P \\ \hat{\mathbf{y}}_\mathbf{A}^{(k-1)} \\ \mathbf{1}_\mathbf{A} - \hat{\mathbf{y}}_\mathbf{A}^{(k-1)} \end{pmatrix} \tag{6.12}$$

where the rows of X_P and $X_\mathbf{A}$ are the rows of X that correspond to the observed presences and background data. The elements of $\hat{\mathbf{y}}_\mathbf{A}^{(k-1)}$ are the elements of $\hat{\mathbf{y}}^{(k-1)}$ that correspond to the background data, and because $\hat{\mathbf{y}} = 1$ for the observed presences, we can use $\mathbf{1}_P$ in place of $\hat{\mathbf{y}}_P^{(k-1)}$, where $\mathbf{1}_P$ is a vector of ones of length equal to the number of observed presences. Similarly $\mathbf{0}_\mathbf{A}$ is a vector of zeros with length equal to the number of background data. After fitting the model, the resulting $\eta^{*(k)}$ must be adjusted as before to obtain $\eta^{(k)}$.

We have seen that the EM algorithm works by alternately fitting the case-control logistic model, adjusting the intercept and then calculating the expectations of the latent y_i for $z_i = 0$. Theoretical results in Dempster et al. (1977) guarantee that a local maximum of the log-likelihood (6.11) is attained. All that is required are suitable starting values for the latent $\mathbf{y}$; using $\hat{y}_i^{(0)} = \pi,\ \forall i : z_i = 0$, so that the sample mean for the background data equals the population mean, ensures that the case-control offset adjustment is correct. Results of the EM procedure applied to a simple logistic regression example are illustrated in Figure 5.3; the estimates are approximately unbiased even as π increases.

6.5 The EM model and stepwise procedures

Stepwise modeling procedures, such as boosted regression trees (Section 5.4), have started to appear in the ecological modeling literature (Leathwick et al., 2006), and as such are a natural choice for the procedure used in the M-step of the EM model. Further, the stepwise nature of the boosted trees model suggests an improvement; interleaving expectation steps with every few boosting steps may decrease the time to convergence of the EM algorithm. In practice this decrease can be an order of magnitude. This interleaving is straightforward to implement using the existing boosted trees package `gbm` in R (Ridgeway, 2004). Although we illustrate the procedure for a boosted trees model, this could be implemented for any stepwise procedure.

The algorithm follows the same steps as in Figure 6.1, fitting the boosted trees model to the augmented data set (6.12). The only difference is that instead of fitting a whole boosted trees model in the M-step, we take the existing boosted trees model, update the weights $\hat{\mathbf{y}}_{\mathbf{A}}^{(k-1)}$ in the saved model, and then continue to run the fitting procedure to add some more trees. As adding single trees is relatively inexpensive, we run the model to add say 10 or 50 extra trees at each maximization step - this parameter may be adjusted to optimize the fit.

The optimal number of trees is chosen in a similar way to a regular boosted trees model; by minimizing a test-set or cross-validation error. However, we have to be careful here as any built-in error calculations would be based on the estimates $\hat{y}^{(k)}$ from the previous iteration of the fitted model. Instead we use the observed deviance, which is -2 times the log of the presence-only likelihood (6.5). The observed deviance is calculated using predictions based on the case-control adjusted $\eta^{(k)}$, and the observed presences $\mathbf{z}$ for a separate test set or for cross-validation sets. Once the optimal model size has been chosen, predictions and marginal plots of the effects of each variable or groups of variables can be calculated as

usual, via the built-in functions of the boosted trees package.

6.6 Bias in the naive presence-only model

The naive presence-only model was introduced in Section 5.1; it is a logistic model fit directly to the observed presences and background data, ignoring the "contamination" of true presences in the background data. In this section we develop the relationship between the naive model and the true logistic model (6.1) that we are interested in. Let η_{naive} be the linear predictor of the naive logistic model for the presence-only data. We will start by calculating the case-control adjustment for the naive model, and then show how the adjusted η_{naive} is related to the true η.

To obtain the best estimate of η from the naive model, we can do a case-control adjustment to η_{naive} (e.g. Mace *et al.*, 1996). To do this, we assume that the data are random samples of size n_p and n_a from the populations of true presences and absences respectively. Then from (6.7), the case-control offset for the naive model is

$$\log\left(\frac{\gamma_1}{\gamma_0}\right) = \log\left(\frac{n_p}{n_a}\right) - \log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{(1-\pi)n_p}{\pi n_a}\right).$$

We can write out this case-control adjusted η_{naive} in terms of the true η of interest. Using Bayes rule and (6.4),

$$\begin{aligned}
\eta_{\text{naive}}(\mathbf{x}) - \log\left(\frac{(1-\pi)n_p}{\pi n_a}\right) &= \text{logit}\left(\mathbb{P}(z=1|s=1,\mathbf{x})\right) - \log\left(\frac{(1-\pi)n_p}{\pi n_a}\right) \\
&= \log\left(\frac{\frac{n_p}{\pi n_a}e^{\eta(\mathbf{x})}}{1+e^{\eta(\mathbf{x})}}\right) - \log\left(\frac{(1-\pi)n_p}{\pi n_a}\right) \\
&= -\log(1-\pi) - \log\left(e^{-\eta(\mathbf{x})}+1\right).
\end{aligned}$$

We have already discussed the form of this bias in Section 5.1.

6.7 Estimating the popoulation prevalence π

Ideally we would like to be able to estimate the population prevalence π, as well as η, from presence-only data. However, if we make no assumptions about the structure of η (such as in a boosted trees model), then π and η are not identifiable. This means that if we know neither π nor η, then we cannot estimate them accurately. In fact the only way to estimate π is by making unrealistic assumptions about the form of η. In this section we prove conditions for the identifiability of π and η; an illustration of how it may work in practice is given in Appendix A.2.

First we give a heuristic proof that π and η are not identifiable when there is no constraint on the structure of η. Let $f_1(\mathbf{x}) = \mathbb{P}(\mathbf{x}|y=1)$ and $f_0(\mathbf{x}) = \mathbb{P}(\mathbf{x}|y=0)$ be the densities of the covariates $\mathbf{x}$ for the populations of true presences and absences respectively. Then the overall density of $\mathbf{x}$ is $f(\mathbf{x}) = \pi f_1(\mathbf{x}) + (1-\pi) f_0(\mathbf{x})$, and by Bayes rule,

$$\mathbb{P}(y=1|\mathbf{x}) = \frac{\mathbb{P}(y=1)\mathbb{P}(x|y=1)}{\mathbb{P}(x)} = \frac{\pi f_1(\mathbf{x})}{(1-\pi) f_0(\mathbf{x}) + \pi f_1(\mathbf{x})}.$$

The model of interest here is $\text{logit}(\mathbb{P}(y=1|\mathbf{x})) = \eta(\mathbf{x})$, so

$$\eta(\mathbf{x}) = \log\left(\frac{\pi}{1-\pi}\right) + \log\left(\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}\right).$$

Thus η defines, up to a multiplicative constant, the form of the ratio between f_1 and f_0.

The background data provide information about the form of f, as they are a random sample from all locations on the landscape. Similarly we learn about the shape of f_1 from the

observed presences. However we have no information about f_0 directly, so if η is unrestricted and π unknown, there is no restriction on the form of f_0. To see this, note that if (π, f_0, f_1, f) is one representation of the densities, then we can write

$$f = \pi f_1 + (1-\pi) f_0 = \pi^* f_1 + (1-\pi^*) f_0^*$$

where $f_0^* = \frac{1}{1-\rho\pi}\left((1-\pi)f_0 + (1-\rho)\pi f_1\right)$ is a density and $\pi^* = \rho\pi$, $0 < \rho < 1$. As f and f_1 remain the same, (π^*, f_0^*, f_1, f) is a family of possible alternate representations. Thus π and f_0 are not uniquely defined without any assumption on the structure of η.

We next prove identifiability of π and η in a conditional logistic regression model, where we assume $\eta(\mathbf{x}) = \mathbf{x}^T\beta$ is linear. The observed likelihood for the observed presences and background data is

$$\begin{aligned} L(\eta|\mathbf{z}, X) &= \prod_i \mathbb{P}(z_i = 1|s_i = 1, \mathbf{x}_i)^{z_i} \mathbb{P}(z_i = 0|s_i = 1, \mathbf{x}_i)^{1-z_i} \\ &= \exp\left(\sum_i z_i \log\left(\frac{\mathbb{P}(z_i = 1|s_i = 1, \mathbf{x}_i)}{\mathbb{P}(z_i = 0|s_i = 1, \mathbf{x}_i)}\right) + \sum_i \log\left(\mathbb{P}(z_i = 0|s_i = 1, \mathbf{x}_i)\right)\right). \end{aligned}$$

From (6.4), this is the density of a curved exponential family (Lehmann & Casella, 1998) with natural parameter $\zeta(\pi, \eta)$ where

$$\zeta_i = \text{logit}(\mathbb{P}(z_i = 1|s_i = 1, \mathbf{x}_i)) = \log\left(\frac{n_p}{\pi n_a}\frac{e^{\eta(\mathbf{x}_i)}}{1 + e^{\eta(\mathbf{x}_i)}}\right).$$

This family is of full rank if $(\pi, \eta) \neq (\pi', \eta') \iff \zeta \neq \zeta'$, and we note that

$$\zeta(\pi, \eta) = \zeta'(\pi', \eta') \iff \frac{\pi'}{\pi} = \frac{e^{-\eta(\mathbf{x}_i)} + 1}{e^{-\eta'(\mathbf{x}_i)} + 1}. \tag{6.13}$$

We will assume that X includes a column of 1s, so that $\mathbf{x}_i^T\beta$ includes an intercept term, and also that X is of full rank. Because X is full rank, $\beta \neq \beta' \iff X\beta \neq X\beta'$, and if $X\beta \neq X\beta'$ then $\frac{e^{-\mathbf{x}_i^T\beta}+1}{e^{-\mathbf{x}_i^T\beta'}+1}$ can not be constant for all i. Hence, there does not exist π and π' such that

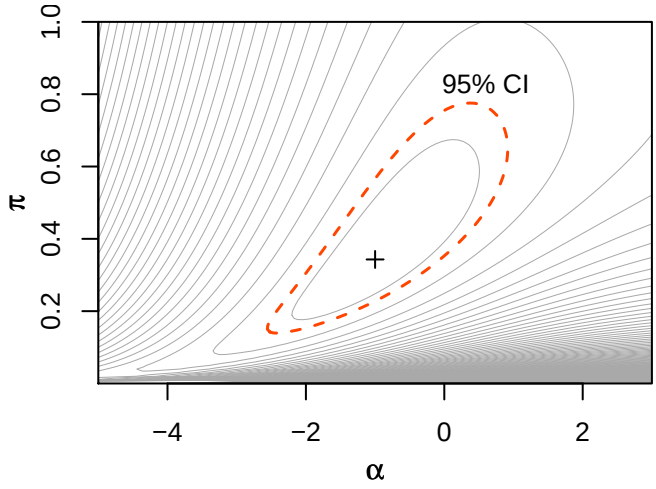


Figure 6.2: *The likelihood surface, and a 95% confidence interval (dashed line), for π and the intercept α at the true level of β for a simulation of 200 observed presences and 1000 background data. The true values of $\pi = 0.34$, and $\beta_0 = -1.0$ are indicated on the plot. The generative model is $\eta(x) = -1 + \beta x$, where $\beta = 2$ and the x are i.i.d. stanford normals.*

$\zeta = \zeta'$. Similarly if $\pi \neq \pi'$, there is no β and β' such that (6.13) holds.

A stricter proof of non-identifiability for general η follows directly from this argument. As η is unconstrained, it is possible to choose η such that the equality (6.13) holds for all $\mathbf{x}_i$, for fixed π,π' and η'. More generally, if we can write $\eta(\mathbf{x}) = h(\mathbf{x})^T\beta$ for some basis functions h of $\mathbf{x}$, such as for GAMs, then π and η are identifiable if and only if $h(\mathbf{x})$ is of full rank. This follows immediately from the argument for the logistic regression model.

We briefly turn our attention to the other models considered in this thesis. It turns out that π is not identifiable in the log probability model (e.g. Section 5.2). This is immediately apparent from the form of $\mathbb{P}(z = 1|\mathbf{x}, s = 1)$ given in (5.5), which depends on π and η through the form $\eta(\mathbf{x}) + \log(\frac{\pi n_a}{n_p})$. Thus π and the intercept of η are not identifiable.

In this thesis, we also approach the presence-only problem by estimating the covariate densities $f_1(\mathbf{x})$ directly (Maxent and Sections 7.2 and 7.4). However, the assumption of a log or logistic probability model implies that the results above also apply in these cases.

Alternatively, if we are willing to assume parametric forms for these densities then we can usually estimate π. For example, assuming f_0 and f_1 are Gaussian, π could be estimated using a method similar to the EM algorithm for Gaussian mixture estimation (e.g. Hastie et al., 2001, p236). However, this estimate may be severely biased if the assumption is not true.

This analysis strongly contra-indicates estimating π from presence-only data. Although it is theoretically possible to estimate π for the linear logistic regression model, there is an implicit assumption that the log of the tilting function f_1/f_0 is linear in $\mathbf{x}$. This may be a useful representation of the true model, but not as an assumption upon which π is estimated. In practice estimates of π are highly variable, and correlated with the intercept, even in simulated models where the logistic model is the truth (Figure 6.2). Where the true densities deviate from the assumed structure, the estimates are highly unstable; Lancaster & Imbens (1996) report failure of convergence of the direct maximization of the presence-only likelihood in several examples.

Chapter 7

Presence-Only Models and Gradient Boosting

In this chapter we introduce 3 new methods of estimating a presence-absence model from presence-only data, all of which use gradient boosting in order to maximize the otherwise intractable likelihoods. In our implementation, we have used gradient boosting of regression trees (Section 5.4), though any base learner may be used in place of regression trees. These three new methods, as well as the EM model of the previous section, fit into the existing framework (Figure 7.1).

7.1 Model 2: the sample logistic model

The first of the boosting models uses a similar approach to the EM model (Section 6.4), in that we wish to maximize the presence-only logistic likelihood (Section 6.1), which is

	Sample likelihood of $P(z = 1\|s = 1, \mathbf{x})$	**Population** density estimation of $f_1(\mathbf{x})$
Logistic link $\mathbb{P}(y = 1\|\mathbf{x}) = \frac{e^{\eta(\mathbf{x})}}{1+e^{\eta(\mathbf{x})}}$	Naive model Lancaster & Imbens EM model Sample logistic	Population logistic
Log link $\mathbb{P}(y = 1\|\mathbf{x}) = e^{\eta(\mathbf{x})}$	Manly's exponential model	Maxent Population log

Figure 7.1: *The modeling procedures used in the presence-only problem can be categorized in two ways. The first is by the use of a log or logistic link function relating $\mathbb{P}(y = 1|\mathbf{x})$ and $\eta(\mathbf{x})$. The second is whether the background data are assumed to be a sample or the full population of locations. Existing models are given in black and new models are highlighted.*

based on the conditional probability of an observed presence z, given $\mathbf{x}$ and $s = 1$. From (6.5) we know that this likelihood can be written

$$\begin{aligned} L(\eta|\mathbf{z}, X) &= \prod_i \mathbb{P}(z_i|s_i = 1, \mathbf{x}_i) \\ &= \prod_i \left(\frac{\frac{n_p}{\pi n_a} e^{\eta(\mathbf{x}_i)}}{1 + \left(1 + \frac{n_p}{\pi n_a}\right) e^{\eta(\mathbf{x}_i)}} \right)^{z_i} \left(\frac{1 + e^{\eta(\mathbf{x}_i)}}{1 + \left(1 + \frac{n_p}{\pi n_a}\right) e^{\eta(\mathbf{x}_i)}} \right)^{1-z_i} . \end{aligned} \tag{7.1}$$

This likelihood is not straightforward to maximize for general η. Lancaster and Imbens (1996) provided code to maximize this likelihood using a generalized method of moments approach , for η linear in $\mathbf{x}$ (or some fixed basis functions of $\mathbf{x}$). A different approach that we have already seen (Section 6.4) is to use an EM algorithm to maximize the full likelihood written also in terms of the missing true presences and absences. Here we consider direct

maximization of the likelihood via gradient boosting, and in particular, its implementation using boosted trees.

The loss function for this problem is the negative log-likelihood. We start by noting that this is a non-convex loss function. We use (7.1) to write out the log-likelihood, with some re-arrangement, to give

$$\log L(\eta|\mathbf{z}, X) = \sum_{z_i=1} \eta(\mathbf{x}_i) + \sum_{z_i=0} \log\left(1 + e^{\eta(\mathbf{x}_i)}\right) - \sum_i \log\left(1 + \left(1 + \tfrac{n_p}{\pi n_a}\right) e^{\eta(\mathbf{x}_i)}\right) + k \qquad (7.2)$$

for some constant k. It is easy to show that the function $f(\eta; \alpha) = \log(1 + \alpha \exp(\eta))$ is convex in η for $\alpha > 0$. This function appears twice in the log-likelihood, but is both added and subtracted for different constants α. Hence there are both convex and concave terms in the log-likelihood, and so the loss function is not necessarily convex in η. To overcome this, we will use the adjusted Newton step in the gradient boosting algorithm (Section 5.4).

7.2 Model 3: the population logistic model

We continue to assume a logistic link function for $\mathbb{P}(y = 1|\mathbf{x})$, but now assume that the background sample of sites is the full population of locations $\mathcal{L}$. Because we know $\mathcal{L}$, by definition we know the density $f(\mathbf{x})$. Hence the randomness in the data is with respect to the density $f_1(\mathbf{x})$ of covariates for the true presences, and the problem becomes one of density estimation. The structure of f_1 is defined in terms of f by the assumption of a logistic probability link. We start by developing this structure and then show how this is used in estimating f_1.

The logistic model (5.1) is

$$\log\left(\frac{\mathbb{P}(y=1|\mathbf{x})}{\mathbb{P}(y=0|\mathbf{x})}\right) = \eta(\mathbf{x}) \qquad \Longrightarrow \qquad \frac{\mathbb{P}(y=1|\mathbf{x})}{\mathbb{P}(y=0|\mathbf{x})} = e^{\eta(\mathbf{x})}. \tag{7.3}$$

We use a standard application of Bayes rule:

$$\mathbb{P}(y=k|\mathbf{x}) = \frac{\mathbb{P}(y=k)f_k(\mathbf{x})}{f(\mathbf{x})} \qquad \Longrightarrow \qquad \frac{\mathbb{P}(y=1|\mathbf{x})}{\mathbb{P}(y=0|\mathbf{x})} = \frac{\pi f_1(\mathbf{x})}{(1-\pi)f_0(\mathbf{x})},$$

and plug in the form of the logistic model (7.3) to get

$$\pi f_1(\mathbf{x}) = e^{\eta(\mathbf{x})}(1-\pi)f_0(\mathbf{x}).$$

Then plugging this in to the total probability identity, $f(\mathbf{x}) = \pi f_1(\mathbf{x}) + (1-\pi)f_0(\mathbf{x})$, and with some manipulation of terms we see

$$f_1(\mathbf{x}) \;=\; \frac{f(\mathbf{x})}{\pi\left(e^{-\eta(\mathbf{x})}+1\right)}. \tag{7.4}$$

Note that this model constrains $f_1(\mathbf{x})$ to be non-negative on the support of $f(\mathbf{x})$, and strictly positive as long as η is finite. Unlike for Maxent, it is also true that the corresponding $f_0(\mathbf{x})$ is non-negative; the form for f_0 is

$$f_0(\mathbf{x}) \;=\; \frac{f(\mathbf{x})}{(1-\pi)\left(e^{\eta(\mathbf{x})}+1\right)}.$$

Thus our problem is one of semi-parametric density estimation, where we are estimating $f_1(\mathbf{x})$ with respect to a baseline density $f(\mathbf{x})$. The likelihood for this density estimation problem is then

$$L(\eta|X) = \prod_{z_i=1} f_1(\mathbf{x}_i) = \prod_{z_i=1} \frac{f(\mathbf{x})}{\pi\left(e^{-\eta(\mathbf{x})}+1\right)} \tag{7.5}$$

where $z_i = 1$ indicates an observed presence. Implicit in the estimation of f_1 is the assumption that it is a true density function: it must be non-negative and integrate to 1

across the full population of locations. The non-negativity holds by the parameterization of the model, but the second requirement generates the constraint on the likelihood that $\int f_1(\mathbf{x})d\mathbf{x} = 1$. In Appendix A.3 we show that maximizing (7.5) subject to this constraint is equivalent to unconstrained maximization of a penalized log-likelihood

$$\log L^*(\eta|X) = \sum_{z_i=1} \log f_1(\mathbf{x}_i) - n_p \int f_1(\mathbf{x})d\mathbf{x} \tag{7.6}$$

where n_p is the number of observed presences. Plugging (7.4) into (7.6) gives us the penalized likelihood for this problem:

$$\log L^*(\eta|X) = \sum_{z_i=1} \left[\log f(\mathbf{x}_i) - \log(\pi) - \log\left(e^{-\eta(\mathbf{x})} + 1\right)\right] - n_p \int \frac{f(\mathbf{x})}{\pi\left(e^{-\eta(\mathbf{x})} + 1\right)} d\mathbf{x}$$

The final term is an integral with respect to $f(\mathbf{x})$, for which we have no closed form. However, f is generated by the finite landscape $\mathcal{L}$, which assigns a probability of $\frac{1}{n_a}$ to each location in $\mathcal{L}$. Therefore, we can replace that integral with a sum over all locations in $\mathcal{L}$, which we indicate using the notation $z_i = 0$. Additionally, the $\log f(\mathbf{x}_i)$ term does not depend on η and so can be ignored as a constant. This gives:

$$\log L^*(\eta|X) = -\sum_{z_i=1} \log\left(e^{-\eta(\mathbf{x})} + 1\right) - \frac{n_p}{\pi n_a} \sum_{z_i=0} \left(e^{-\eta(\mathbf{x})} + 1\right)^{-1} + \text{ const.} \tag{7.7}$$

One interesting point to notice here is that the background data get the same weighting in this log-likelihood, no matter whether there are 100 or 100,000 data points. Thus there is no inherent problem with unbalanced sample sizes.

We note here that the loss function for this problem, which equals the negative log likelihood, is not necessarily convex in η. Although $\log(e^\eta + 1)$ and $e^{-\eta} + 1$ are both convex in η, the inverse of a convex function is not convex, so the second summation on the r.h.s. of (7.7) is not convex. The non-convexity of the loss function is only problematic when taking

the Newton step to estimate the optimal parameters for the fitted regression tree. Hence we use the adjusted Newton step (Section 5.4), to ensure that the gradient is positive.

7.3 The population logistic model as a limit of the sample logistic model

We show heuristically that the population model can be thought of as the limit of the sample model, as the background sample size approaches that of the whole population. In these calculations we will assume an infinite background sample and so look at the case where $n_u \to \infty$. We compare the limit of the log-likelihood for the sample logistic model (7.2) to the log-likelihood for the population logistic model, and use L_S and L_P to denote the likelihoods for the sample and population models respectively. Implicit in the sample logistic model is the assumption that $f_1(\mathbf{x})$ integrates to 1 across $\mathbf{x}$. We make this an explicit constraint in the maximization of L_S, and as $n_u \to \infty$, to match the constraint on the maximization of L_P.

From (7.7) the form of the log-likelihood for the population model, without the penalization term, is:

$$\begin{aligned} \log L_P(\eta|X) &= -\sum_{z_i=1} \log\left(e^{-\eta(\mathbf{x})}+1\right) + \text{ const} \\ &= \sum_{z_i=1} \left[\eta(\mathbf{x}) - \log\left(1+e^{\eta(\mathbf{x})}\right)\right] + \text{ const.} \end{aligned} \tag{7.8}$$

In contrast, the sample model has log-likelihood given by (7.2):

$$\begin{aligned}\log L_S(\eta|\mathbf{z}, X) = &\sum_{z_i=1}\left[\eta(\mathbf{x}_i) - \log\left(1+\left(1+\tfrac{n_p}{\pi n_a}\right)e^{\eta(\mathbf{x}_i)}\right)\right] \\ &+ \sum_{z_i=0}\left[\log\left(1+e^{\eta(\mathbf{x}_i)}\right) - \log\left(1+\left(1+\tfrac{n_p}{\pi n_a}\right)e^{\eta(\mathbf{x}_i)}\right)\right] + \text{ const.}\end{aligned} \tag{7.9}$$

Note that both these likelihoods have the implicit or explicit constraint $\int f(\mathbf{x})d\mathbf{x} = 1$. Thus to show that (7.9) converges to (7.8), we need to show that the second summation of (7.9), over $z_i = 0$, tends to a constant, and the first summation, over $z_i = 1$, tends to (7.8).

Denoting $\gamma = \frac{n_p}{\pi n_a}$, the second summation in (7.9) can be written

$$-\sum_{z_i=0}\log\left(\frac{1+(1+\gamma)\,e^{\eta(\mathbf{x}_i)}}{1+e^{\eta(\mathbf{x}_i)}}\right) = -\sum_{z_i=0}\log\left(1+\gamma\frac{e^{\eta(\mathbf{x}_i)}}{1+e^{\eta(\mathbf{x}_i)}}\right) = -\sum_{z_i=0}\log\left(1+\gamma p_i\right)$$

where $p_i = \mathbb{P}(y_i = 1|\mathbf{x}_i)$. By definition, $\gamma \to 0$ as $n_a \to \infty$, so remembering that $n_a = \#\{i; z_i = 0\}$, we use a Taylor expansion to get

$$-\sum_{z_i=0}\log\left(1+\gamma p_i\right) = \sum_{z_i=0}\left[-\gamma p_i + O(\gamma^2 p_i^2)\right] = -\frac{n_p}{\pi}\frac{1}{n_a}\sum_{z_i=0}p_i + O\left(\frac{1}{n_a}\right).$$

$$\stackrel{n_a\to\infty}{\longrightarrow} -\frac{n_p}{\pi}\mathbb{E}[\mathbb{P}(y=1|\mathbf{x})]$$

Thus the second summation in (7.9) tends to a constant as $n_a \to \infty$.

To show that the first summation in (7.9) tends to (7.8), we will show that

$$\sum_{z_i=1}\log\left(1+\left(1+\tfrac{n_p}{\pi n_a}\right)e^{\eta(\mathbf{x}_i)}\right) - \sum_{z_i=1}\log\left(1+e^{\eta(\mathbf{x}_i)}\right) \stackrel{n_a\to\infty}{\longrightarrow} 0. \tag{7.10}$$

Following the same arguments from above, the l.h.s. of (7.10) can be written

$$\sum_{z_i=1}\log\left(\frac{1+(1+\gamma)\,e^{\eta(\mathbf{x}_i)}}{1+e^{\eta(\mathbf{x}_i)}}\right) = \sum_{z_i=1}\log\left(1+\gamma p_i\right) = \sum_{z_i=1}\left[\gamma p_i - O(\gamma^2 p_i^2)\right].$$

In this case, the summation is of $n_p = \#\{i; z_i = 1\}$ terms, which stays fixed, so

$$\sum_{z_i=1} \left[\gamma p_i - O(\gamma^2 p_i^2)\right] = \frac{n_p}{\pi} \frac{1}{n_a} \sum_{z_i=1} p_i - O\left(\frac{1}{{n_a}^2}\right) \stackrel{n_a \to \infty}{\longrightarrow} 0.$$

Hence (7.10) is true and so (7.9) tends to (7.8) as $n_a \to \infty$.

This result is similar in form to that of Owen (2006). In that paper, the limit of a logistic regression model, as the number of absences tends to infinity, is equivalent to density estimation of the covariate distribution of the presences, where that density is assumed to be an exponential tilt of the density for the absences. This exponential tilt is the exact relation of the densities in the logistic model.

7.4 Model 4: the population log model

The final model we consider is similar to Maxent in that it assumes the same log probability link function (5.2) and the background sample is assumed to be the full landscape of locations $\mathcal{L}$. The difference between this model and Maxent is in the choice of basis functions and the fitting procedure. While Maxent does L_1 constrained fitting assuming an η linear in certain fixed basis functions, this model uses gradient boosting of regression trees. Obviously, the concerns about the specificity of the log link function still apply here, but we were interested to see how the extra flexibility in the boosted trees, over the Maxent basis functions, impacts performance.

As with the population logistic model, we know the density of the background data $f(\mathbf{x})$, so we need to estimate the density $f_1(\mathbf{x})$. The structure of f_1 is defined by our assumption

of a log probability link; using (5.2) and an application of Bayes rule we get

$$f_1(\mathbf{x}) \;=\; \frac{\mathbb{P}(y=1|\mathbf{x})f(\mathbf{x})}{\mathbb{P}(y=1)} \;=\; e^{\eta(\mathbf{x})}f(\mathbf{x}). \tag{7.11}$$

Hence the likelihood for this density estimation problem is then

$$L(\eta|X) = \prod_{z_i=1} f_1(\mathbf{x}_i) = \prod_{z_i=1} e^{\eta(\mathbf{x})}f(\mathbf{x}). \tag{7.12}$$

Again we need to ensure that f_1 is a true density. It is easy to see that f_1 is non-negative, and we can ensure that it integrates to one by using the penalized log-likelihood (7.6) we introduced in Section 7.2. Thus plugging (7.11) into (7.6) gives us the penalized log-likelihood for this problem:

$$\log L^*(\eta|X) = \sum_{z_i=1} \left[\eta(\mathbf{x}_i) + \log f(\mathbf{x}_i)\right] - n_p \int e^{\eta(\mathbf{x}_i)}f(\mathbf{x})d\mathbf{x}.$$

As before (Section 7.2), we can replace the integral with respect to $f(\mathbf{x})$ with a finite sum over all locations in $\mathcal{L}$. Additionally, the $\log f(\mathbf{x}_i)$ term does not depend on η and so can be ignored as a constant, so we get:

$$\log L^*(\eta|X) = \sum_{z_i=1} \eta(\mathbf{x}_i) - \frac{n_p}{n_a} \sum_{z_i=0} e^{\eta(\mathbf{x}_i)} + \text{ const.}$$

This loss function is convex in η and so we use a straightforward implementation of gradient boosting of regression trees to estimate η.

7.5 Adjusting for shrinkage inherent in boosting

In the estimates of η for the boosted trees models, we see a lot of shrinkage relative to η for a model fit to the true presences and absences. This reflects the observations of Rosset et al. (2004) that boosting behaves similarly to L_1 regularization on all possible basis functions, which in our example is all possible regression trees. L_1 regularization is a widely used shrinkage method aimed at reducing model variability. However, in our presence-only models, we are trying to remove the shrinking effect of having unknown presences/absences in our background data. Although boosted trees is a very flexible modeling procedure, the extra shrinkage incurred because of the boosting is not itself desirable.

To overcome this shrinkage we apply a post-hoc linear adjustment to $\hat{\eta}$, giving

$$\hat{\eta}_{\text{adj}}(\mathbf{x}) = \beta_0 + \beta_1 \hat{\eta}(\mathbf{x}).$$

In practice this is done as follows. After fitting the boosted trees model to the training set data, the optimal number of trees is chosen using the test set data. Then we take the predictions $\hat{\eta}(\mathbf{x})$ for the test set data, and use these as the covariates in a linear presence-only model to estimate β_0 and β_1. It is important that we use the test set and not the training set data to estimate these parameters to avoid overfitting.

The presence-only model used in this post-hoc adjustment is ideally chosen to match the form of the boosting model, so e.g. a post-hoc adjustment of the sample logistic boosting model would use a linear EM model. However, we do not have an equivalent form of the population logistic model for η linear in $\mathbf{x}$. In this case, because the results from the sample and population logistic models tend to be similar, we will use the sample logistic based EM model as a stand-in for a linear population logistic model. An application of Maxent, using only linear features, can be used to make this adjustment for the log population model,

although the regularization multiplier (the shrinkage parameter) should be set to zero.

7.6 Comparing the presence-only models

At this point we pause briefly to recap the differences between the existing and new models. At the beginning of Chapter 5 we pointed out the different assumptions that underly the four quadrants in Figure 7.1: the log vs logistic link functions and the sample vs population assumption of the background data. Here we concentrate the differences between models in the top left and bottom right quadrants.

First, in the top left quadrant, are all models that assume a logistic link function and that the background data are a sample of al locations. The naive model is different from the others in that it can be highly biased because it maximizes a different incorrect loss function. The other three models are all attempting to maximize the same presence-only likelihood (6.5), albeit in two different ways, and for different structures for η in terms of $\mathbf{x}$. Lancaster & Imbens assume η is linear in $\mathbf{x}$, whereas the sample logistic model assumes that η is a linear combination of regression trees. However, both of these models attempt direct maximization of the presence-only likelihood. In contrast, the EM model attempts to maximize the presence-only likelihood (6.5) by instead maximizing the full likelihood (6.10) which includes the sometimes missing true presences/absences. This full likelihood, when integrated with respect to the distribution of the unobserved y, results in the presence-only likelihood. The EM model is more flexible than the other models in that almost any logistic fitting procedure may be used, thus allowing many different structures of η relative to $\mathbf{x}$.

In the bottom right corner of Figure 7.1, Maxent and the population log model attempt to maximize the same likelihood (7.12), subject to constraint that f_1 be a density. However, they assume different structures and constraints on η. Maxent uses L_1 regularization of a

large set of basis functions of $\mathbf{x}$, whereas the population log model uses gradient boosting of regression trees.

Chapter 8

Assessing the Presence-Only Models

Analysis and comparison of presence-only models is often problematic, because we wish to use presence-absence data to best assess the models, but there are few data sets where true presence-absence data are available, for the same species and regions, as presence-only data. On the other hand, although we know the truth when using simulated data, it is hard to reproduce the complexity that is inherent in ecological systems, and so results from simulations may not accurately reflect the results we might get in practice. As we do not have access to any paired presence-only and presence-absence data sets, we use two different types of simulations. The first method uses full simulations of data generated from a logistic model. In addition, we use actual presence-absence data of longfin eels in New Zealand rivers to generate presence-only data. Thus we can assess the models estimated from presence-only data, using the true presences and absences.

A typical choice of metric for comparing presence-absence models is the binomial deviance. While we could use this to compare all the logistic link models, this is not an appropriate comparison for the two log link models, as some of the fitted probabilities for

these models are greater than 1. In that case, the binomial deviance is not defined, and even if we replace all those probabilities with a probability of $1 - \epsilon$, the resulting value of the deviance is highly dependent on the choice of ϵ. Instead, we assess these models using two overlapping sets of metrics for the two approaches. In the full simulations, we know the true η and hence the true $\mathbb{P}(y = 1|\mathbf{x})$. We cannot compare the ηs directly, as they have different interpretations for the log and logistic models. Instead, we use the root mean squared error (RMSE) and correlations between the true and predicted probabilities $\mathbb{P}(y = 1|\mathbf{x})$. The RMSE measures how accurately we can estimate the exact probability, and hence in Maxent we use our knowledge of π to convert the given output to a probability estimate. The correlation instead indicates how accurately we know the true probabilities up to a linear transformation. This measure does not penalize for shrinkage, so for example, the naive model will appear to perform better when assessed using correlation rather than RMSE.

In both sets of analyses, we assess fit using AUC, the area under the receiver operating characteristic (ROC) curve (Bradley, 1997). The ROC curve is a plot of sensitivity vs. (1-specificity) for different levels of a classification threshold in a binary classification problem. In our example, sensitivity is the proportion of correctly classified true presences and (1-specificity) is the proportion of incorrectly classified true absences. The AUC lies between 0 and 1, with a high AUC indicating a good classifier. A classifier that randomly assigns true presence and absence categories will have an AUC of 0.5. We note here that the AUC statistic is based solely on the relative ordering of the predicted probabilities, and not their absolute values. This measure benefits Maxent because it is irrelevant whether the estimated probabilities are greater or less than 1, whereas probabilities greater than 1 may be heavily penalized in the calculation of RMSE.

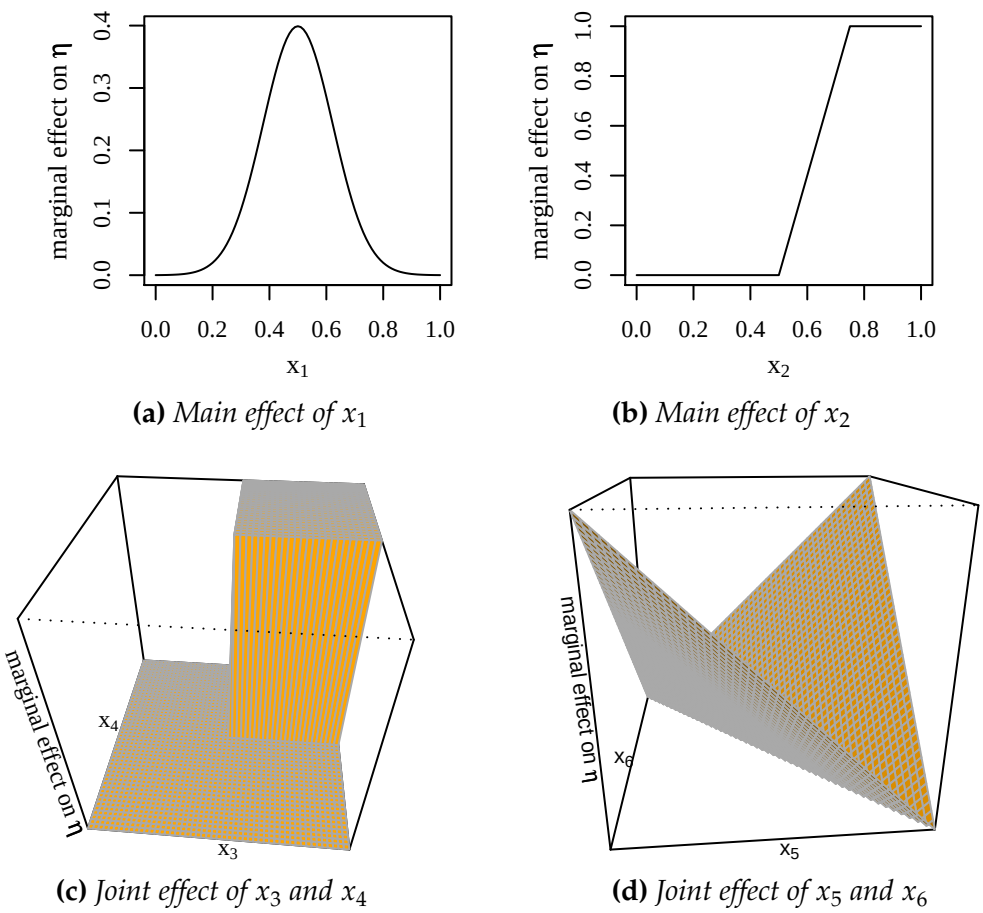
We will compare all the models presented in Figure 7.1, plus the post-hoc adjusted versions of boosted trees logistic models (Section 7.5) and the transformed naive model (at the end of Section 5.1). As there is a lot of variability inherent in most presence-absence

models, and we wish to measure the *additional* variability introduced by the missing true presences/absences, we provide a baseline against which to compare the relative performance of the presence-only models. We call this the *full model* and it uses the same data that are used in fitting the presence-only models, but with the added knowledge of *all* true presence and absences. We assume a logistic link for this model, and we fit the model using usual presence-absence models. The only place where care is needed is that these data have different sampling rates of true presences/absences than occur in the population, so we must make a case-control adjustment as in (6.6). In this case, $\gamma_1 = \frac{n_1}{N\pi}$ and $\gamma_0 = \frac{n_0}{N(1-\pi)}$, so the adjustment is $\log(n_1/n_0) - \log(\pi/(1-\pi))$, where n_1 and n_0 are the number of true presences and absences in these data for the full model.

8.1 Simulation studies

In creating a generative model for the simulation studies, we wanted all modeled probabilities to be no greater than 1, and for the population prevalence to be π. However, for π sufficiently large it was hard to create a generative model with a log link function that satisfied both these conditions. This is because a log link model that is additive in η is multiplicative in the probability, so for a given variable, the effect on the probability will be much larger if the probability was already large. This in itself is a good indicator that the log link is not appropriate in modeling probabilities. Because of this, we assumed a logistic link in the generative model for these simulation studies, even though we knew this would favor the logistic models.

We assume there are 10 independent uniform $[0, 1]$ covariates, that are iid for each location. The generative model for the simulation depends on only 6 of these covariates, while the



(a) *Main effect of* x_1 **(b)** *Main effect of* x_2

(c) *Joint effect of* x_3 *and* x_4 **(d)** *Joint effect of* x_5 *and* x_6

Figure 8.1: *The generative model for the simulation study depends only on the first six variables through main and two-way joint effects.*

remaining 4 are noise. The model, which is also illustrated in Figure 8.1, is:

$$\eta(\mathbf{x}) = \phi\left(\frac{x_1 - 0.5}{0.125}\right) + ((4x_2 - 2 \vee 0) \wedge 1) + 1_{\{x_3 > 0.5, x_4 > 0.5\}} + |x_5 + x_6 - 1| + c(\pi) \qquad (8.1)$$

where the constant $c(\pi)$ is chosen so that the mean presence for this model equals π. Thus there are two main effects on η from x_1 and x_2, and two 2-way effects. The value of π varies between 0.1 and 0.5 to assess the impact of population prevalence on the relative performance of the models.

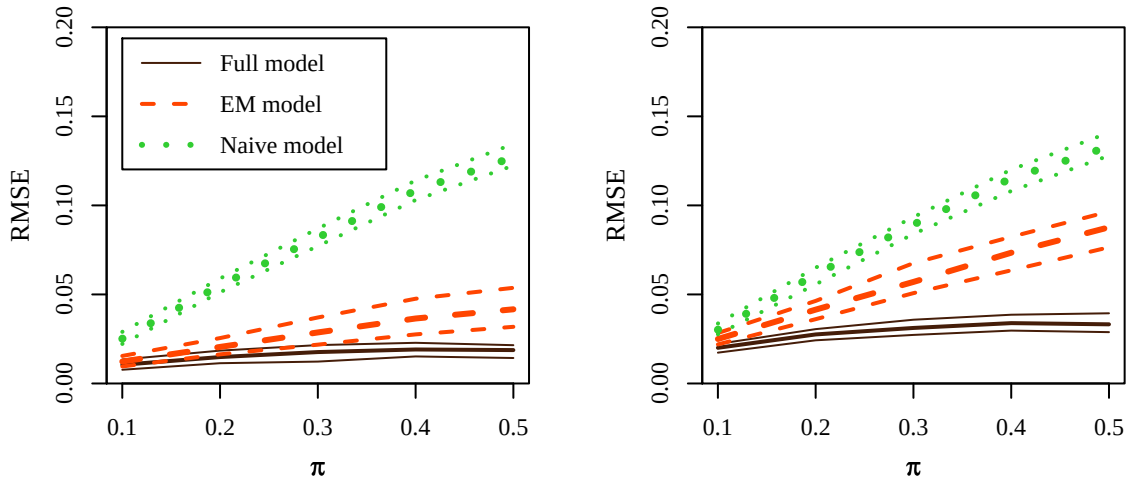
The data were generated following the sampling assumptions set out in Section 4.3. There are 500 observed presences and 2000 background data in each realization of the simulation, and the data generation and model fitting were repeated 100 times. These 2500 observations were used to fit the models, with the test set, if necessary to choose the optimal model, taken from those data. In order to assess the final model, we use a validation set of 10,000 observations randomly sampled from the population.

We approach the fitting of this model in two ways. The first directly uses the basis functions from the generative model when fitting the presence-only models, both with and without the 10 original covariates as noise variables. This provides an insight into when the presence-only models work well. We then look at the typical scenario, where only the 10 original covariates are given, and compare the performance of Maxent and various boosted regression tree models.

8.1.1 Modeling via known basis functions

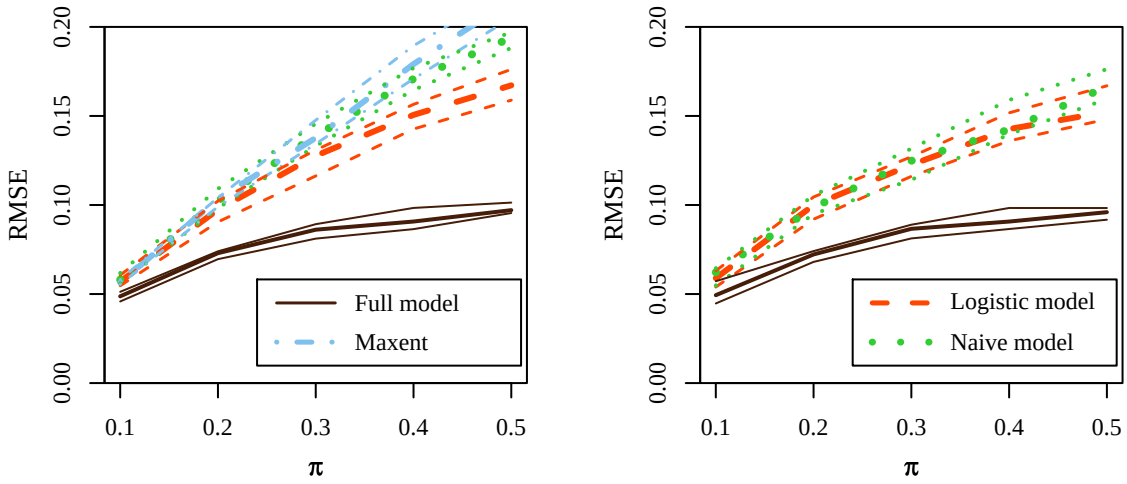
The aim of this part of the simulation was to illustrate how well the presence-only model could reproduce the presence-absence model for a simple model in the absence of excess noise, and then observe the effect of adding in extra noise. The covariates for the first model are the four basis functions $h_j(\mathbf{x})$ for $j = 1, \ldots, 4$ that are implicitly defined in (8.1), so e.g. $h_1(\mathbf{x}) = \phi\left(\frac{x_1 - 0.5}{0.125}\right)$ and $h_4 = 2|x_5 + x_6 - 1|$. In the model with extra noise, we use these same four basis functions, plus all 10 original covariates, thus our set of covariates is $\{h_1(\mathbf{x}), \ldots, h_4(\mathbf{x}), x_1, \ldots, x_{10}\}$. We fit these two sets of covariates to three models, each of which use a logistic regression procedure. For the full model we do a logistic regression fit of the true presences and absences, followed by case-control adjustments. For the other two models we fit the observed presences and background data. In the naive fit we use a case-control adjusted logistic regression, and for the EM model we use logistic regression in each M step of the algorithm.

Unsurprisingly, the RMSE of the predicted probabilities from these presence-only models increases as a function of π, as illustrated in Figure 8.2. When the noise variables are included in the fit, the RMSE is higher for each model, but this is most noticeable for the EM model. For a simple, well defined model with little noise, the EM model performs almost as well as the full model. However, with extra noise, the EM model only manages to reduce RMSE over the naive model by about 50% relative to the full model. The absolute difference in RMSE is negligible for $\pi \approx 0.1$ but is considerable for larger π. As more noise is introduced, for example in the following section, this gain in performance over the naive model decreases further.



(a) *No noise variables* **(b)** *With noise variables*

Figure 8.2: *The root mean squared error (RMSE) across the validation set for the predicted probabilities as a function of π, for models using the basis function from the generative model either with or without extra noise variables. The three lines for each model are the quartiles of the RMSE across 100 realizations. These plots may be directly compared with Figure 8.3, which gives results for boosted regression tree models for the same data.*



(a) *No adjustment* **(b)** *With post-hoc adjustment*

Figure 8.3: *The root mean squared error (RMSE) across the validation set for the predicted probabilities as a function of π, for Maxent, and models using boosted regression trees with or without the post-hoc adjustment. (Maxent does not have a post-hoc adjustment.) The three lines for each model are the quartiles of the RMSE across 100 realizations. The logistic model here is the population logistic model, though the sample and population logistic models perform similarly. These plots can be directly compared with those for models using the known basis functions (Figure 8.2).*

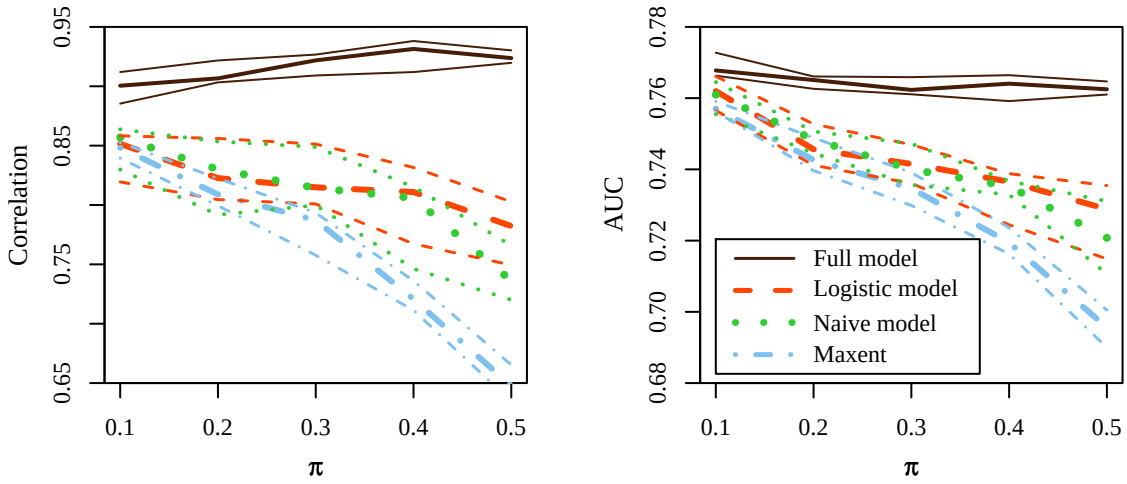
8.1.2 Modeling via boosted regression trees

In this section we look at how accurately each procedure can reproduce the generative model (8.1), given only the covariates $\{x_1, \ldots, x_{10}\}$, and not the form of the basis functions $h_j(\mathbf{x})$. For each model considered in this section, except for Maxent, we use gradient boosting of regression trees. For the full and naive models, we use the Bernoulli distribution of the `gbm` package in R (Ridgeway, 2004), and then make case-control adjustments. The transformed naive model is then obtained by using the transformation (5.4) on the case-control adjusted naive model. The EM model uses the Bernoulli distribution of `gbm` at each M-step, and the three boosting models from Chapter 7 are fit using new distributions that were created for the `gbm` package and are now available as part of the `ecogbm` package. For all models using boosted trees, we also do the post-hoc adjustment (Section 7.5). For the full model, this involves fitting a logistic regression model, and for all other logistic

boosted trees models, we run the EM model.

Now the models must estimate the form of the basis functions of the generative model (8.1), and hence the variability in the estimated models increases, as indicated by the larger RMSE of the probability estimates (Figure 8.3 compared to Figure 8.2). Further, using boosted regression trees has reduced the impact of fitting a correct presence-only model over using the naive model. The reduction in RMSE for the population logistic model, over the naive model, is now about 20% of the possible reduction (as measured by the full model). This is considerably less than the 50% reduction seen for the basis function models. It appears that the benefit of using correct presence-only models decreases as the complexity of the model fitting increases — this is intuitive as higher complexity implies it is harder to discern between true presences and absences for the background data. Using a post-hoc adjustment reduces that difference further; although both adjusted models have improved performance, the adjusted naive model now performs almost as well as the adjusted population logistic model. Maxent has a slightly higher RMSE than the unadjusted naive model for π greater than 0.3, and performs considerably worse than any of the models with post-hoc adjustment.

Looking at the other summary statistics of correlation and AUC (Figure 8.4), we see that Maxent again does considerably worse than the other models for $\pi > 0.2$. However, for these two statistics, the naive model performs comparably with the logistic model for $\pi \leq 0.4$. This indicates that the naive model is managing to order the predicted validation set values about as well as the logistic model, even though the actual probability estimates are not as accurate (Figure 8.3). The reason for this is explained by the better performance of the post-hoc adjusted naive model and illustrated in Figure 8.5; the unadjusted naive model has considerable shrinkage toward π of the predicted probabilities. This shrinkage also occurs to a lesser extent in the logistic population model. Maxent, at the other extreme, tends to have a large variability in predictions, with about 2% of predictions greater than 1.

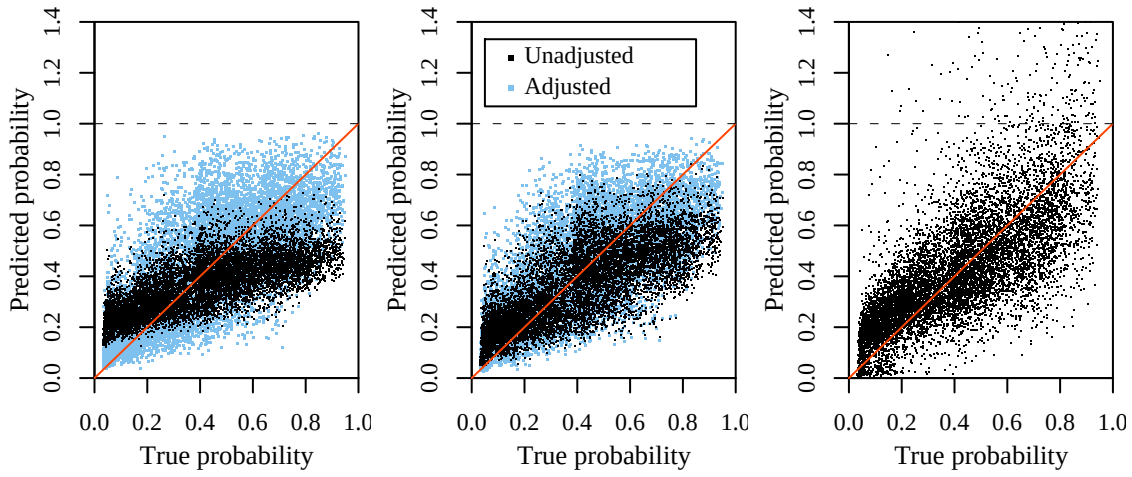


(a) *Correlation* **(b)** *AUC*

Figure 8.4: *The correlation and area under the ROC curve (AUC) across the validation set for the predicted probabilities as a function of π, for Maxent, and the models using boosted regression trees without the post-hoc adjustment. The logistic model here is the population logistic model, though the sample and population logistic models perform similarly. The three lines for each model are the quartiles of the RMSE across 100 realizations.*

To get a more detailed view of the differences between models, we compare the summary metrics of all models for a relatively large population prevalence of $\pi = 0.4$ (Figure 8.6). Post-hoc adjustments are excluded from Figure 8.6(b) as they have little effect on the correlation metric. For a relatively prevalent species, such as this, the logistic models perform considerably better than the log models. In particular, the population log model, along with the transformed naive model, has a high variability in performance, with some occurrences of considerably poor performance. Maxent performs poorly here, primarily because of the large variability in predicted values (Figure 8.5c). The new logistic models have the highest correlations and lowest RMSE, with the adjusted naive model performing almost as well.

We note that the EM model has a higher median RMSE than the sample logistic model, even though they are trying to maximize the same likelihood, whether directly or indirectly. Additionally, in some realizations, the EM model performs particularly poorly. This

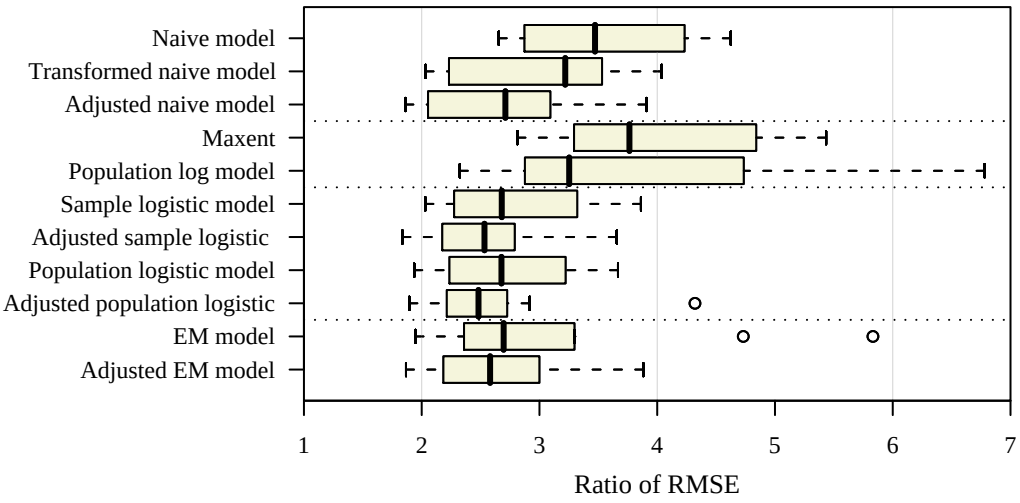


(a) *Naive model* **(b)** *Logistic model* **(c)** *Maxent*

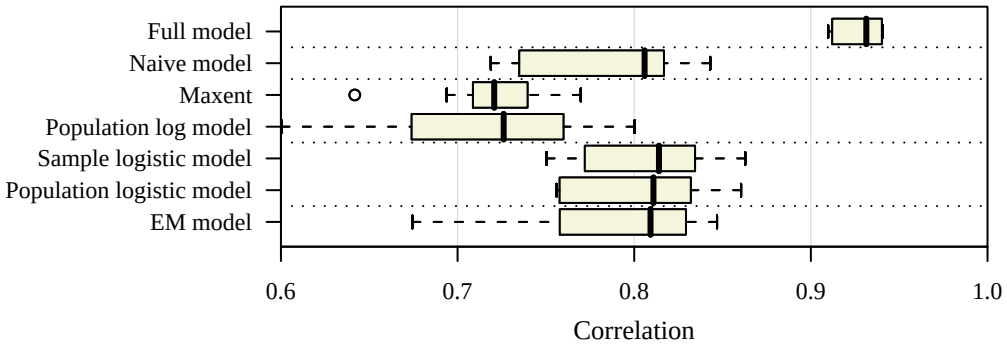
Figure 8.5: *The predicted vs the true probabilities for $\pi = 0.4$ for the validation set. The predicted probabilities are given with or without the post-hoc adjustment. The logistic model is the population logistic model, though the sample logistic model has similar results. At this value of π, approximately 2% of the predicted probabilities from Maxent are greater than 1.*

is possibly because the shape of the optimization surface changes after each E-step for the EM model, whereas the form of the likelihood for the sample logistic model remains fixed. The combination of the greedy stepwise nature of gradient boosting and the changing likelihood appears to result in the EM model being more likely to get trapped in local minima.

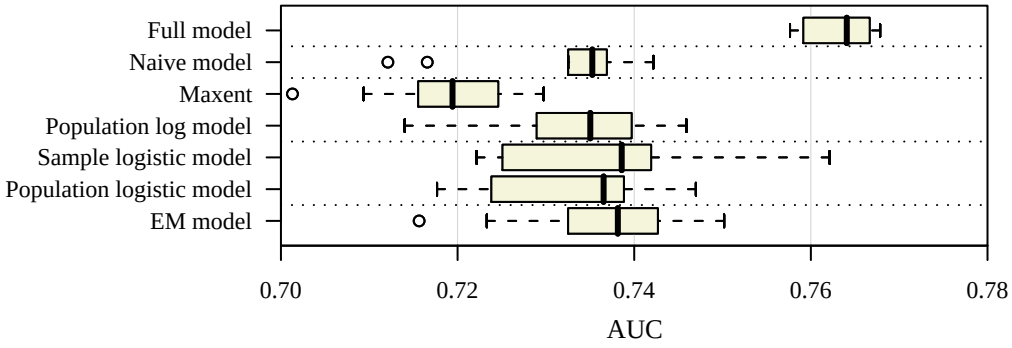
We end this section by looking at how the models are affected when π is mis-specified (Figure 8.7). The case-control adjustments for the naive and full models, plus the normalizations for Maxent and the log model, all depend on this assumed value of π. Hence changing π will have a large impact on the RMSE for these models, but will have no effect on the AUC and only a small effect on the correlation. Thus we use the correlation and AUC to measure the effect of misspecified π. In these analyses we fit the models repeatedly to the same data, while changing the assumed value of π. The AUC of the logistic models is relatively unaffected by the assumed value of π. However, as the assumed π gets smaller, the correlation for the logistic model decreases to be similar to that of the



(a) *RMSE for predicted probabilities as a ratio of full model RMSE*

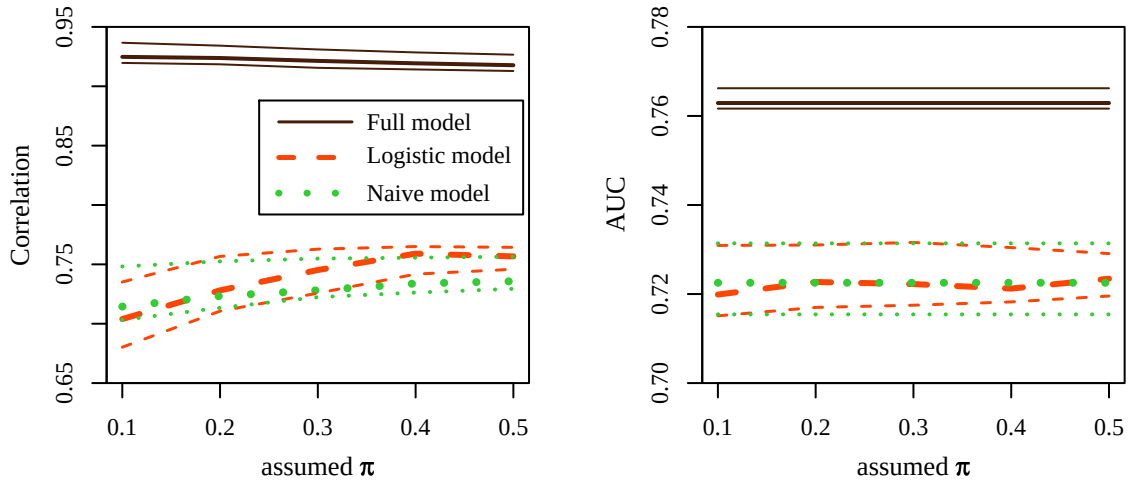


(b) *Correlation between true and predicted probabilities*



(c) *AUC for predicted probabilities*

Figure 8.6: *The root mean squared error (RMSE) as a ratio of full model RMSE, correlation and area under the ROC curve (AUC), for the predicted probabilities of the validation set. All models use boosted regression trees, except Maxent. These data were generated with a population prevalence* $\pi = 0.4$.



(a) *Correlation* **(b)** *AUC*

Figure 8.7: *The validation set correlation and area under the ROC curve (AUC) for the predicted probabilities as a function of the* assumed *value of* π*, for models using boosted regression trees. The three lines for each model are the quartiles of the correlation across 100 realizations. The logistic model here is the population logistic model, though the sample and population logistic models perform similarly. The true value of* π *is* 0.4.

naive model. However, for π between 0.3 and 0,5, the logistic model is still performing better than the naive model, as measured by correlation.

8.2 The longfin eel; generating presence-only data from presence-absence data

We have access to presence-absence data of longfin eel, which have previously been described in Leathwick et al. (2005) as part of a data set of diadromous fish in the rivers of New Zealand. These data were drawn from the New Zealand Freshwater Fish Database (McDowall and Richardson, 1983, and `http://www.niwa.co.nz/services/nzffd/`), which is curated by New Zealand's National Institute of Freshwater and Atmospheric Research. The database contains 11,858 records of sites sampled since 1980. Any locations in tidal rivers or in lakes or other stationary water have been removed, leaving only freshwater river locations. Unlike the data in Leathwick et al. (2005), locations upstream of a dam or other obstruction have not been removed. The longfin eel is relatively populous and occurs at 51.3% of these locations, and in 17.3% of locations occuring upstream of a dam.

We generated presence-only data from these presence-absence data according to the sampling assumptions set out in Section 4.3. We treat this presence-absence data set as being a random sample of all freshwater river locations in New Zealand. First, the background data were generated by sampling randomly without replacement from all available data. Then the sample of observed presences was generated by sampling without replacement from the remaining presence locations. A random sample of one quarter of the data were set aside as a test set and the rest were used to train the models. The training set contains around 4,400 and 1,300 background data and observed presences respectively. The test set contains around one third of these numbers, of which around 1,200 were *true* presences and 700 were *true* absences. To reduce spurious effects, we repeated the presence-only data generation 10 times, with different random seeds; the results provided are summaries across these repetitions. There are 21 environmental covariates describing conditions at the sampled site as well as up- and downstream.

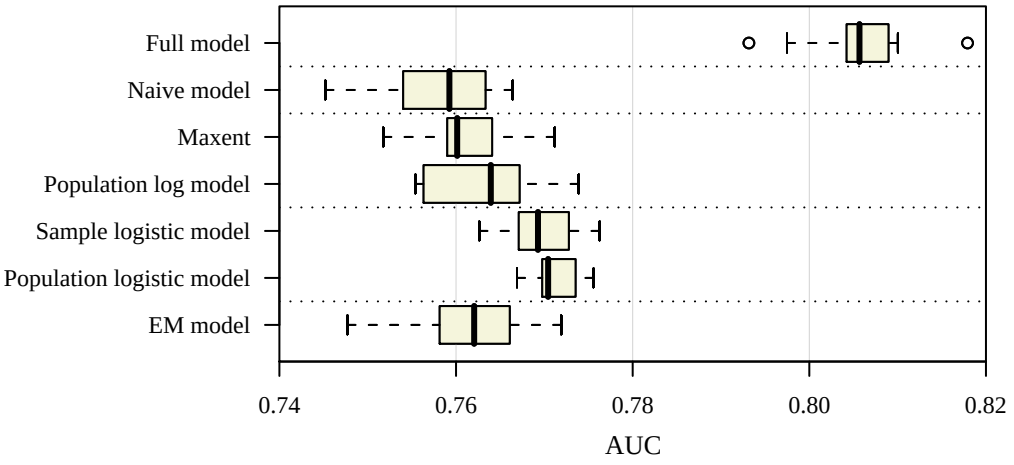
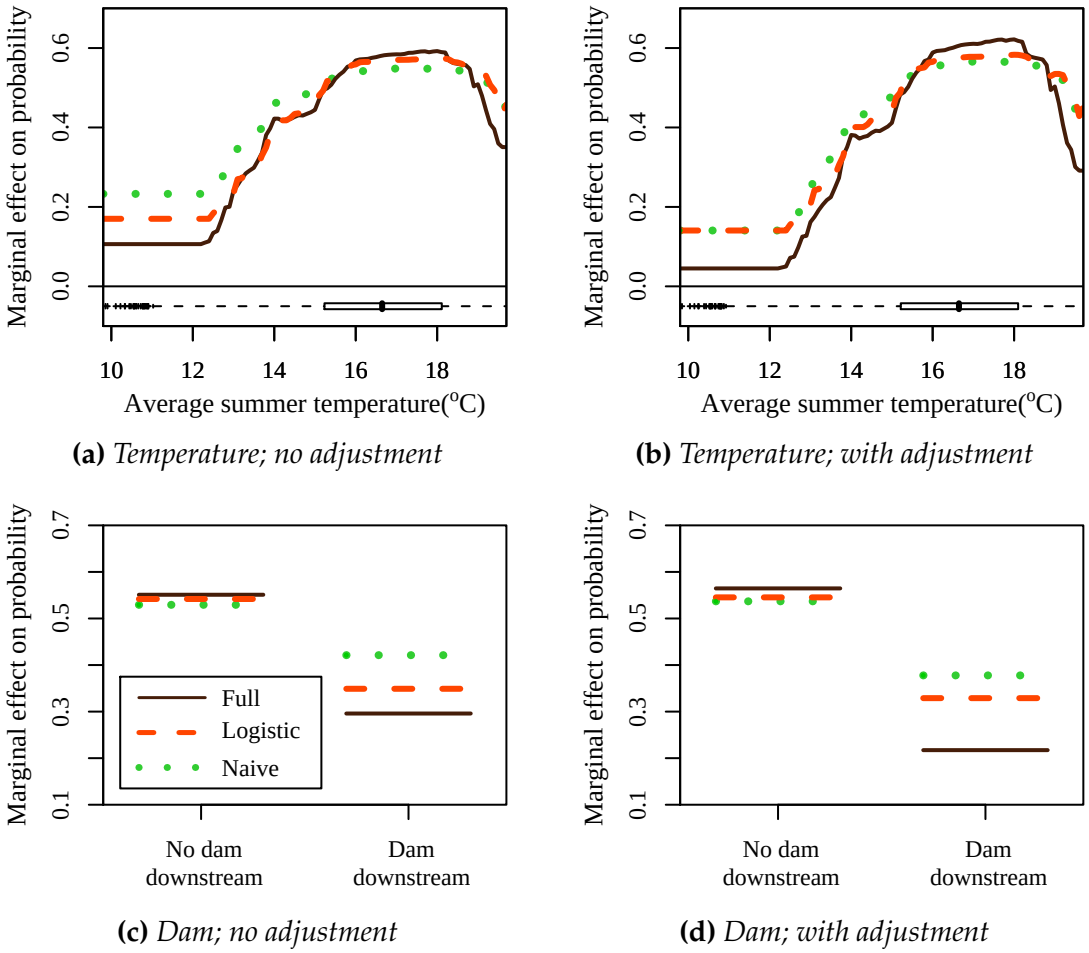


Figure 8.8: *The area under the ROC curve (AUC) for the predicted probabilities of the longfin eel test set using the true value of $\pi = 0.51$. The AUC was calculated using the true presences and absences. All models use boosted regression trees, except Maxent.*

We compare the same models as in Section 8.1.2, namely Maxent and boosted trees implementations of all other models. A total of 10,000 trees were fitted for each boosted trees model, each with a depth of 4, allowing four-way interaction terms. Comparing the models using AUC (Figure 8.8), we see that the population logistic model performs the best of all the presence-only models; this model improves AUC by 20% over the naive model, relative to the full model, which is our best-case scenario. The sample logistic model gives similar results, but the EM model performs considerably worse. The two worst performing models are Maxent and the naive model. Here the population log model performs slightly better than Maxent, which might be because boosted trees can model the complexity of such data better than the basis functions of Maxent.

Two of the most important predictors in modeling the true presences were the summer temperature and the presence of a dam downstream. Figure 8.9 illustrates that the naive model in particular tends to underestimate the range of the effect that each of these predictors has on η. This is particularly noticeable in the binary downstream dam variable, and echoes the patterns in Figure 5.3 and Figure 8.5 that the naive logistic regression model effectively shrinks the parameter estimates and hence the predicted values. We also see



(a) *Temperature; no adjustment*

(b) *Temperature; with adjustment*

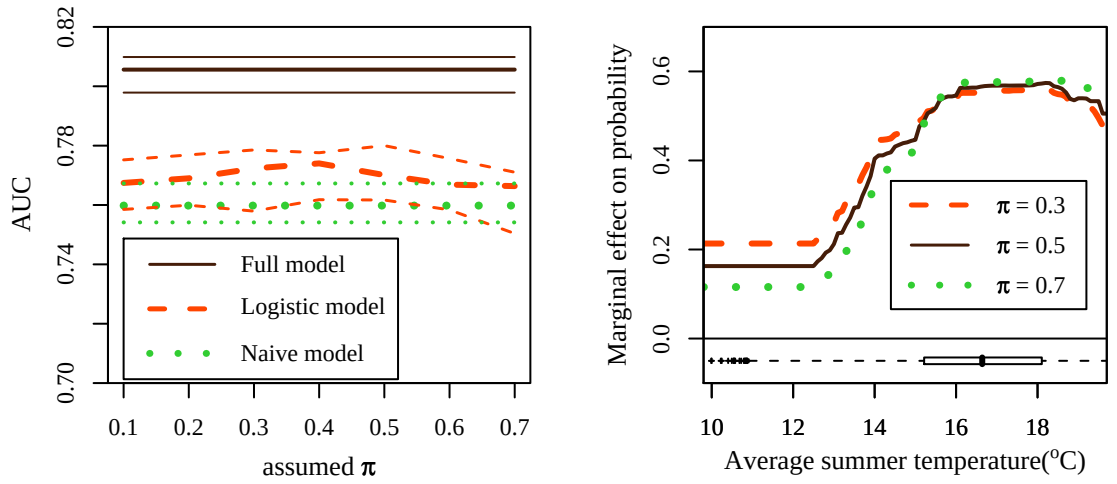
(c) *Dam; no adjustment*

(d) *Dam; with adjustment*

Figure 8.9: *The marginal effects of average summer air temperature (°C) and a downstream dam on the probability of presence of the longfin eel, for models with and without post-hoc adjustments. The pointwise standard deviation of effects across the 10 repetitions is approximately 0.01 for all effects. The boxplots indicate the distribution of the summer temperature across all locations in the sample.*

some shrinkage of the marginal effects for the EM and logistic models. However, while the post-hoc adjustment considerably decreases the shrinkage of the naive model, it has little effect on the EM or logistic models. We note that even the full model experiences some shrinkage because of the use of boosted trees.

Finally, we ran a sensitivity analysis of the model, for each of the 10 sampling repetitions,



(a) *AUC by assumed* π **(b)** *Marginal effect on* $\mathbb{P}(y = 1|\mathbf{x})$

Figure 8.10: *The area under the ROC curve (AUC) for the full, naive and population logistic models, and the marginal effect on average summer air temperature for the population logistic model, by the assumed value of* π*. The three lines in (a) indicate the quartiles across the 10 realizations.*

to determine the effect of differing prior beliefs of π. Each model was fit repeatedly, using assumed values of π between 0.1 and 0.7. The AUC remains constant across π for the full and naive models, as they are only monotonically affected by the value of π. The AUC for the population logistic model appears to be largest for an assumed π of 0.4, which is slightly lower than the true $\pi = 0.51$. Both over- and under-estimation of π decreases the AUC for this model, but for all values of π the median AUC for the population logistic model is higher than the AUC for the naive model (Figure 8.10a) and Maxent (not shown).

The marginal effects on η of all variables for the population logistic model follow a pattern similar to that for mean summer temperature illustrated in Figure 8.10(b). Across different values of π, the shape of the marginal effect is relatively consistent, with the magnitude of the effect increasing with π. Unsurprisingly, as the assumed value of π gets smaller the effect size tends to resemble that of the naive model (not shown); we can think of the naive model as a limit of our sample logistic model as the assumed value of π tends to zero, with an adjustment in the intercept of the model.

Chapter 9

The Presence-Only Problem: Conclusions

We have proposed four modeling procedures that directly estimate the presence-absence probabilities from presence-only data. The first, which uses an application of the EM algorithm, works as a wrapper to (almost) any logistic modeling procedures. For stepwise procedures, fitting time of this EM model may be reduced by interleaving E-steps within the stepwise procedure. The other models all use gradient boosted regression trees of log or logistic link models.

The EM model performs considerably better than the naive model for data where the generative logistic model is linear in the covariates and where there are no noise variables. This gain in performance lessens as the variability inherent in the problem increases, for example by the addition of noise variables to the data or by increasing the complexity of the generative model. The most highly variable problem considered here was the gradient boosted regression trees models based on real ecological data of longfin eel. Even for this

problem, the population logistic model has considerably better performance than the naive model; it increased AUC by about 20% of the difference between the naive model and the best case scenario where all presences and absences are known. However, in the example based solely on simulations, this gain in AUC over the naive model was negligible.

One of the major impacts on model estimates when using a naive model is the shrinkage of the parameter estimates. This occurs because the true presences in the background data are treated as absences. When boosted regression trees are used, this can add some extra shrinkage as well. This shrinkage can be somewhat reduced by a post-hoc adjustment using the EM model where η is linear in a single variable — the fitted values from original fitted model. Although this does not affect AUC, because it is a monotonic transformation of the predictions, it can increase RMSE of the predicted probabilities. Thus this adjustment can help in problems where the absolute (as opposed to relative) predicted values are important.

Previous work in estimating the presence-absence model from presence-only data has included simultaneous estimation of the population prevalence π. However, we have shown that π is not identifiable when no assumptions are made about the structure of η, such as in boosted trees models. In addition, even when some assumption is made about the structure of η, e.g. for logistic regression models with η linear in the covariates $\mathbf{x}$, the resulting estimate of π is highly variable and heavily dependent on that assumption. Although this structure may be a useful modeling approximation, the estimation of π relies too heavily on this assumption. Instead, it is better to obtain an estimate of π from some other source and if desired use a sensitivity analysis to assess the dependence of the resulting model on this estimate.

Some modeling procedures, e.g. Maxent, do not require knowledge of π. The log link

function used in that model implies that probabilities can be estimated up to a multiplicative constant with no known value of π. However, the log probability link is itself problematic in that it may result in predicted probabilities greater than 1, and as such it has been critiqued in ecological modeling literature when used in other models. We have also seen evidence that Maxent does not order the predicted probabilities as well as our new presence-only models for more prevalent species. However, when π is small, most predicted values of η for the logistic model are large and negative and the log link is similar to the logistic link function in this range. Indeed, we have seen that Maxent appears to perform similarly to the presence-only models for $\pi = 0.1$.

Some of the new presence-only models are closely related. The sample and population logistic models give similar results in practice and we can think of the population logistic model as a limit of the sample logistic model as the number of background data increase. The EM model and the sample logistic model share the same logistic link and assumption of a sample of background locations. Theoretical results for the EM algorithm (Dempster et al., 1977) indicate that the two models are maximizing the same likelihood, whether directly or indirectly, but in our analyses the EM model performs worse, especially in the more realistic example. This may be because the greedy approach of the gradient boosting algorithm and the continual updating of the model weights means the EM model is more likely to get stuck in a local but not global minimum of the non-convex loss function. However, the EM model is more flexible, in that it can be used with almost any modeling procedure, and for procedures with single solutions, such as logistic regression or GAM, it may avoid these pitfalls.

These presence-only models provide a flexible way of estimating species distribution from extensive records of species presence. The EM model uses a straightforward wrapper around most logistic procedures, and the boosted trees models are now available through the `ecogbm` package in R. Because of the simplicity of fitting these models, they can be easily adopted by anyone working in this field.

Part II

Boosted MARS

Chapter 10

Introduction to Boosted MARS

We turn now from the specific problem of estimating the presence-absence model from presence-only data, to a more general problem of creating a new modeling procedure that may be used in many situations. Gradient boosted regression trees are becoming popular in ecological modeling (e.g. Leathwick et al., 2006; De'ath, 2007) because they have good predictive power. However, the piecewise constant nature of regression trees often results in discontinuities in the estimated marginal effects, and it is often unclear whether such discontinuities are real or spurious. These marginal effects are important in ecological modeling as we are often also interested in the effects of individual ecological variables on the outcome.

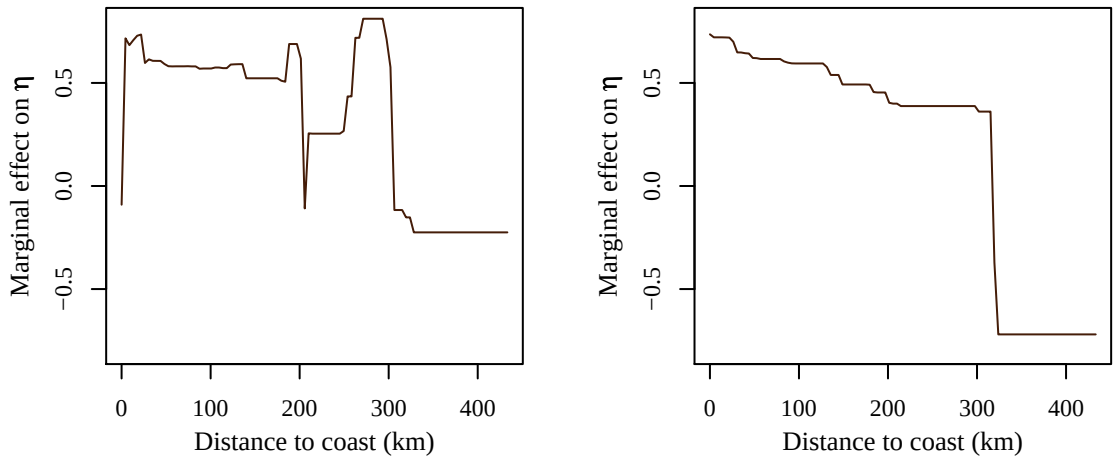
In this part of the thesis, we investigate another modeling technique that aims to echo the flexibility and predictive performance of gradient boosted regression trees, but that results in a smoother model. We continue to use gradient boosting, but now fit MARS models at each step, instead of regression trees. MARS or multivariate adaptive regression splines

(Friedman, 1991) fits piecewise linear basis functions, with optional smoothing, for least-squares loss functions. Boosting provides a way of fitting MARS models for more general loss functions. We also introduce new post-hoc adjustments to the MARS basis functions.

We investigate the use of a subsequent L_1 regularization procedure on the basis functions of the boosted MARS model. Friedman and Popescu (2004) have previously taken a similar approach, using L_1 regularization on general sets of "rules" or basis functions. Recent development of L_1 regularization algorithms for more general loss functions (Park and Hastie, 2006; Friedman et al., 2007) allows us to apply these ideas to logistic loss functions for binomial data. One of the benefits of L_1 regularization is that we can reduce the total number of basis functions in the final model. This provides for fast model predictions, important for applying such models in generating large and detailed species distribution maps. Further, the simple structure of MARS basis functions allows implementation of model predictions within mapping software, such as Geographic Information Systems (Leathwick et al., 2005)

10.1 Motivation

Gradient boosted regression trees (Section 5.4 and Friedman, 2001) provide a highly flexible modeling procedure for a wide range of loss functions. For example, in the longfin eel models, estimated marginal effects of average summer temperature on η have a highly non-parametric form (Figure 8.9a). However, paired with this flexibility is a potential lack of smoothness in the model, illustrated more clearly in the marginal effects of another variable from the same data set: distance from the coast (Figure 10.1a). The discontinuities arise from the large number of piecewise constant basis functions that form the boosted



(a) *Unconstrained basis functions* **(b)** *Monotonic basis functions*

Figure 10.1: *The marginal effect of downstream distance to the coast on η for a logistic model of the probability of presence of the longfin eel. This is a gradient boosted regression trees model on the true presences and absences of one realization of the data generated for the presence-only analyses (Section 8.2). In (a), the model made no assumption about the monotonicity of the effect of distance from sea, while (b) assumes that η decreases with increasing distance. Basis functions for all other covariates are unconstrained.*

trees model. This lack of smoothness can be somewhat reduced by constraining the fitting procedure to only allow monotone effects (Figure 10.1b). However, in practice monotonicity is not often an appropriate assumption, or may not be obvious a priori. Further, even with the assumption of monotonicity, the estimated marginal effects may still not be smooth enough; it is unclear whether the sudden drop in Figure 10.1(b) at around 320km is a true indication of the effect of distance from the coast. Even if it this reflective of the data in the training set, we may believe that the underlying model is more smooth than this suggests.

We have already seen (Section 5.4) that one of the benefits of gradient boosting is that it provides a straightforward fitting procedure for loss functions that are difficult to optimize directly. We keep this same optimization procedure but consider alternate basis functions. These basis functions should be able to create highly flexible models, and be fast and easy

to fit, as with regression trees. They should also be able to automatically choose and fit interaction terms between covariates. However, they should also be continuous and smooth. Thus, instead of the *piecewise constant* form of the regression tree, we look at continuous basis functions that are generated from *piecewise linear* basis functions. We also consider an option of smoothing out the elbows in those basis functions in a way that the first derivative of the basis functions is also continuous. These basis functions are the same as used in multivariate adaptive regression splines (MARS, Friedman, 1991), and we start by introducing MARS in the next section.

10.2 MARS

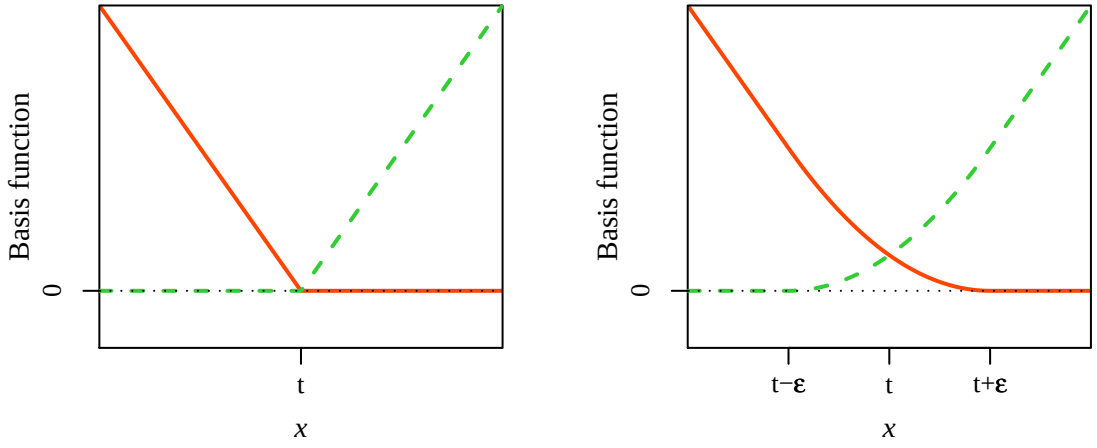
Multivariate adaptive regression splines (MARS, Friedman, 1991) is a flexible regression procedure that automatically models interactions between variables. It fits a model linear in basis functions $\{h_m(\mathbf{x})\}_{m=1}^{M}$:

$$f(\mathbf{x}) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(\mathbf{x}). \tag{10.1}$$

The MARS fitting procedure determines both the basis functions and the parameters $\{\beta_m\}_{m=0}^{M}$. This is similar in style to regression trees, however while regression trees create basis functions based on piecewise constant functions, MARS creates basis functions using combinations of piecewise linear or *hinge functions* (Figure 10.2a). These hinge functions $(x-t)_+$ and $(t-x)_+$ have a knot or hinge at t, are zero on one side of the knot and are linear on the other side:

$$(x-t)_+ = \begin{cases} x-t & \text{for } x > t \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad (t-x)_+ = \begin{cases} t-x & \text{for } x < t \\ 0 & \text{otherwise .} \end{cases} \tag{10.2}$$

This pair of hinge functions is called a *reflected set*. In practice, when fitting a MARS model, we only consider basis functions with knots at an observation.



(a) *Hinge functions* **(b)** *Hinge functions with smoothing*

Figure 10.2: *The hinge function is linear, with a knot or hinge at t (a). A smooth version of the hinge (b) is obtained using a quadratic form within ϵ of the hinge. The solid red line illustrates $(t-x)_+$ and the dashed green line illustrates $(x-t)_+$ for the unsmoothed and smoothed functions.*

Basis functions are added to the model in an iterative manner, starting with $h_0(\mathbf{x}) = 1$. At step k, the algorithm adds two more basis functions to the model, which are a multiple of one previous basis function $h_j(\mathbf{x})$ for $j \leq 2(k-1)$ and a reflected set based on one of the variables x_i:

$$h_{2k-1}(\mathbf{x}) = h_j(\mathbf{x})(x_i - t_k)_+ \quad \text{and} \quad h_{2k}(\mathbf{x}) = h_j(\mathbf{x})(t_k - x_i)_+. \tag{10.3}$$

The prior basis function $h_j(\mathbf{x})$, the new variable x_i, the knot t_k and the coefficients β_{2k-1} and β_{2k} are chosen to minimize the resulting squared error loss. As $h_0(\mathbf{x}) = 1$, the initial basis functions $h_1(\mathbf{x})$ and $h_2(\mathbf{x})$ are reflected hinge functions. Subsequent basis functions may also be reflected hinge functions, or may also be products of hinge functions. Thus at each step MARS may choose to add another main effect to the model, or to include an interaction between variables, the form of which is defined by the products of the hinge functions.

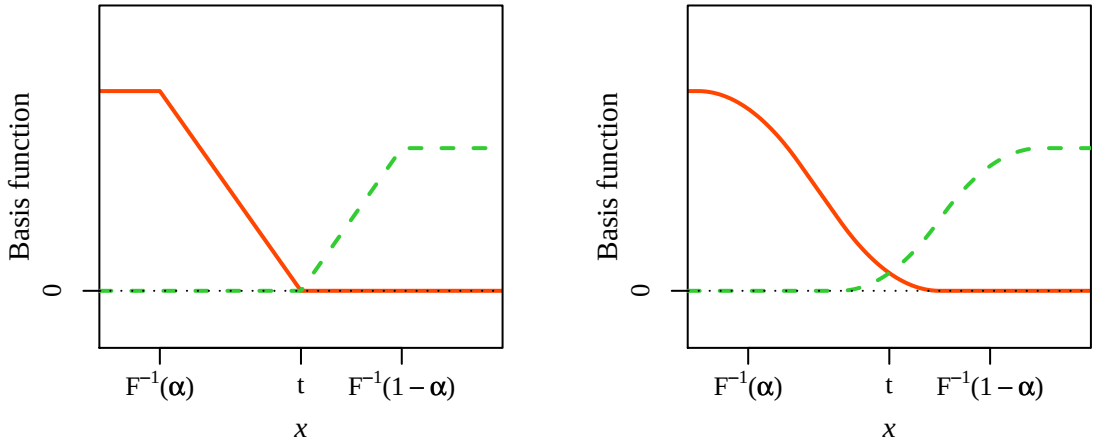
10.3 Adaptations of MARS basis functions

In this section we consider two post-hoc alterations of hinge functions that can be used in constructing a MARS model. The first aims to provide a smoother fitted model by replacing the hinge in the basis functions with a quadratic form. The second alteration thresholds the hinge function so that beyond a given percentile of the data, the MARS basis functions remain constant. Both of these are post-hoc in that the MARS model is estimated from the data based on the original hinge functions (10.2), and these adjustments are made to the resulting estimated model. In our implementation, these adjusted MARS basis functions are used in subsequent fitting procedures, which can correct any inaccuracies created by such post-hoc adjustments. The smoothing of the hinge function has been previously described in Friedman (1991), but to our knowledge the thresholded hinge function has not appeared in statistical literature.

Hinge functions and thus the MARS models constructed of them have discontinuous first derivatives. By replacing the elbow of the hinge function with a smooth function, we can create MARS basis functions with continuous first derivatives. In particular, within ϵ of the elbow t, we use a quadratic form in place of the hinge functions $(x-t)_+$ and $(t-x)_+$. This quadratic form is chosen so that is has the same first derivatives as the hinge function at $|x-t| = \epsilon$. We can write this smoothed hinge function $s((x-t)_+)$ as

$$s((x-t)_+) = \begin{cases} x-t & \text{for } x > t+\epsilon \\ \dfrac{(x-t-\epsilon)^2}{4\epsilon} & \text{for } |x-t| \le \epsilon \\ 0 & \text{for } x < t-\epsilon \end{cases} \tag{10.4}$$

The function $s((t-x)_+)$ is defined similarly and both are illustrated in Figure 10.2(b). The amount of smoothing is determined by ϵ, which is set in advance for each variable. Ideally it is specified as a single multiple of some measure of spread for each variable. In this



(a) *Thresholded hinge functions* **(b)** *Thresholded and smoothed functions*

Figure 10.3: *The thresholded hinge function (a) is linear, with hinges at t and at the quantiles $F^{-1}(\alpha)$ or $F^{-1}(1-\alpha)$. A smooth version of the thresholded hinge (b) is obtained using a quadratic form within ϵ of both t and either $F^{-1}(\alpha)$ or $F^{-1}(1-\alpha)$. The solid red lines and the dashed green lines illustrate the thresholded hinge functions $\left(t-(x \wedge F_x^{-1}(\alpha))\right)_+$ and $\left((x \wedge F_x^{-1}(1-\alpha))-t\right)_+$ respectively, both without and with smoothing.*

implementation we use a fraction of the inter-quartile range (IQR) for each variable.

The second adjustment is in response to a well-known flaw of MARS — that MARS is not robust to extrapolation beyond the variable space covered by the training set. A MARS model includes basis functions that are multiplicative in the covariates across subsets of the variable space, e.g. $h_{10}(\mathbf{x}) = (x_2 - t_2)(x_3 - t_3)(x_6 - t_6)$ for $\{x_2 > t_2\} \times \{x_3 > t_3\} \times \{x_6 > t_6\}$ for some constants t_2, t_3 and t_6. In this case, it is easy to see that as x_2, x_3 and x_6 get larger, the basis function $h_{10}(\mathbf{x})$ increases cubicly. In comparison, a boosted regression tree model is highly robust in extrapolation, as the predicted model is constant as any single variable is increased beyond the convex hull of the training set. Further, for any new data point, the predicted model must take on a value equal to the prediction for a data point in the training set.

Thus we increase the robustness of extrapolation in MARS by adjusting the hinge functions

(10.2) so that they are constant beyond the extreme $\alpha \times 100\%$ of the training set data in each variable. Thus if we define $F_x^{-1}(\alpha)$ to the be the $\alpha \times 100$th percentile of x, the thresholded hinge functions are $\big((x \wedge F_x^{-1}(1-\alpha)) - t\big)_+$ and $\big(t - (x \wedge F_x^{-1}(\alpha))\big)_+$. In particular, this gives

$$\Big((x \wedge F_x^{-1}(1-\alpha)) - t\Big)_+ = \begin{cases} F_x^{-1}(1-\alpha) - t & \text{for } x \geq F_x^{-1}(1-\alpha) \\ x - t & \text{for } t < x < F_x^{-1}(1-\alpha) \\ 0 & \text{for } x \leq t, \end{cases} \tag{10.5}$$

and both thresholded functions are illustrated in Figure 10.3(a).

We can also implement smoothing of these thresholded hinge functions as we did in (10.4) for the original hinge functions. Now we smooth both elbows, at t and either $F_x^{-1}(1-\alpha)$ or $F_x^{-1}(\alpha)$, using the quadratic curve. This is illustrated in Figure 10.3(b) but we do not give an explicit form here as it follows directly from (10.4). In smoothing the thresholded hinge functions, we need to be careful about the relative values of the smoothing parameter ϵ and the position of the elbows. For example, if we are thresholding and smoothing the hinge function $(x-t)_+$, then we create another elbow at $F_x^{-1}(1-\alpha)$. We smooth within ϵ of each elbow, thus we require that $F_x^{-1}(1-\alpha) - t \geq 2\epsilon$ in order to smooth correctly between these two points. We can enforce this by either forcing ϵ to be small enough for this particular hinge function, or by not allowing hinge functions with elbows within 2ϵ of $F_x^{-1}(1-\alpha)$. We implement the latter constraint more generally by discarding basis functions with knots that occur within the $\gamma \times 100\%$ extremes of each variable, where γ is chosen sensibly. This also reduces the possibility of overfitting the MARS model to a few extreme data points.

Chapter 11

Boosted MARS

The aim of this procedure is to find a model that predicts an outcome y given covariates $\mathbf{x}$. This model is chosen to minimize a loss function $\mathrm{L}(y, f(\eta(\mathbf{x})))$, where the linear predictor $\eta(\mathbf{x})$ is linear in certain basis functions $h_j(\mathbf{x})$ of $\mathbf{x}$:

$$\eta(\mathbf{x}) = \sum_{j=0}^{J} \beta_j h_j(\mathbf{x}). \tag{11.1}$$

In this boosted MARS procedure, we take the basis functions to be MARS basis functions (10.3), which are products of hinge functions, with optional smoothing and thresholding. In our implementation and analysis, we look at two loss functions; the Gaussian least-squares loss function and the binomial logistic loss function. This methodology is easy to extend to other loss functions based on the negative log-likelihood of the usual generalized linear models, such as a log-linear Poisson model. Other loss functions may require specially written software.

We consider two different implementations of Boosted MARS for each loss function. In

both implementations we start by using gradient boosting to build a large additive model as in (11.1), where the basis functions $h_j(\mathbf{x})$ are MARS models. The second implementation then takes a further step. We use the basis functions generated in the first step to create a model using the same loss function. In this fitting procedure, we use an L_1 penalization on the β_j to give a more parsimonious model by choosing a subset of those basis functions. To avoid confusion we refer to the one-step and two-step procedures as boosted MARS and L_1 boosted MARS respectively, or as unregularized and regularized boosted MARS.

11.1 Step 1: Boosted MARS

We start with a typical gradient boosting approach to optimize the loss function, as described in Section 5.4. At iteration $k+1$ we update the current linear predictor $\hat{\eta}^{(k)}(\mathbf{x})$ by adding a new set of basis functions. These basis functions $h^{(k+1)}(\mathbf{x})$ are chosen by fitting a MARS model to the current residuals, using squared error loss. We then use these MARS basis functions to estimate the optimal $\beta^{(k+1)}$ using the true loss function. As we want to model the extra information gained over the current model, the current linear predictor $\hat{\eta}^{(k)}(\mathbf{x})$ is included in this model as an offset. For example, if the loss function is the negative of the binomial log-likelihood then we estimate $\beta^{(k+1)}$ via a logistic regression, linear in the new basis functions, with offset $\hat{\eta}^{(k)}(\mathbf{x})$. These $\hat{\beta}^{(k+1)}$ are then multiplied by the step size ϵ, before being used with the new basis functions to update $\hat{\eta}^{(k)}$:

$$\hat{\eta}^{(k+1)}(\mathbf{x}) = \hat{\eta}^{(k)}(\mathbf{x}) + \epsilon \sum_{m=0}^{M} \hat{\beta}_m^{(k+1)} h_m^{(k+1)}(\mathbf{x}).$$

Thus the final estimated model after K boosting steps is

$$\hat{\eta}^{(K)}(\mathbf{x}) = \epsilon \sum_{k=1}^{K} \sum_{m=0}^{M} \hat{\beta}_m^{(k)} h_m^{(k)}(\mathbf{x}). \tag{11.2}$$

For the L_1 boosted MARS implementation it is not necessary to re-estimate the parameters β at each iteration — the parameter estimates from each MARS fit may be used in updating $\hat{\eta}^{(k)}(\mathbf{x})$. Any extra error introduced by using slightly inaccurate parameter estimates is not problematic as the β will be re-estimated in the second, L_1 step. Further, this extra variability may increase the range of basis functions from which we can choose during the second step. However, in our analyses we do re-estimate the β at each boosting step, as we also assess this boosted MARS model before L_1 regularization for each data set. In ad-hoc analyses, we found little difference in L_1 boosted MARS fits with and without this re-estimation procedure.

We also implement stochastic boosting, in order to increase variability at each step and hence increase the number of unique basis functions in the boosted MARS model. In stochastic boosting, a random subset of the data are chosen at each iteration. The MARS fit, along with any re-estimation of parameters, is calculated using only this random subset. However the update (11.2) to the boosted model is calculated using all training data.

We can also consider adaptations of the MARS basis functions, through thresholding and/or smoothing of the hinge functions. As we noted in Section 10.3, these are post-hoc adaptations. Thus unsmoothed and unthresholded basis functions are estimated at each iteration and the post-hoc adjusted basis functions are used in the subsequent fit to estimate $\hat{\beta}^{(k+1)}$. Additionally, the adjusted basis functions are used in updating $\hat{\eta}^{(k)}$ at each iteration. Even when $\hat{\beta}^{(k+1)}$ is not re-estimated, any lack of fit resulting from this smoothing can be accounted for in subsequent boosting iterations.

We can choose the size of each MARS model to imply a maximum level of interaction between variables. This is a maximum, because MARS may fit a model that depends on fewer variables, and in theory may fit all basis functions in terms to the same variable. In practice, we choose a smaller model size than would be used in fitting a single MARS model. By combining many smaller MARS basis functions, we can achieve the same level

of complexity as a single, larger MARS model. Additionally, smaller models will reduce the time needed to calculate the boosted MARS model, and so are more computationally efficient. Such a trade-off between complexity and efficiency is common to statisticians, and may be investigated on a case-to-case basis. In our examples we use a MARS model of size 10, which includes 5 sets of reflected bases, and thus allows up to 5-way interactions.

11.2 Step 2: L_1 regularization on MARS basis functions

In the optional second step, we re-estimate all the parameters β and use shrinkage to build our final model from a subset of the smoothed basis functions generated in step 1. This is done by optimizing the loss function for the model (11.1) with respect to β, with L_1 regularization of β. We can write this as

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \; \mathrm{L}(y, f(\eta(\mathbf{x}))) + \lambda \sum_{k=1}^{K} \sum_{m=1}^{M} |\beta_m^{(k)}| \tag{11.3}$$

where $\eta(\mathbf{x}) = \beta_0 + \sum_{k=1}^{K} \sum_{m=1}^{M} \beta_m^{(k)} h_m^{(k)}(\mathbf{x})$. The regularization parameter is λ; adjusting λ changes the amount of regularization in the model. Note that we have replaced all the $\beta_0^{(k)} h_0^{(k)}(\mathbf{x})$ terms with a single intercept β_0, as $h_0^{(k)}(\mathbf{x}) = 1$ by definition. As usual, the regularization term does not include the intercept β_0.

L_1 regularization is a form of shrinkage of the coefficients β. Shrinkage is typically used to reduce variability in fitted models (e.g. Hastie et al., 2001, p59), by constraining the total size of the β. Although this creates bias in the estimated β, the reduction in variability can often outweigh the effect of increased bias on model fit. The art here is in choosing the optimal amount of bias, which can be controlled by adjusting λ. We can re-write the problem (11.3) as one that minimizes the unpenalized loss $\mathrm{L}(y, f(\eta(\mathbf{x})))$ subject to the constraint $\sum_{k=1}^{K} \sum_{m=1}^{M} |\beta_m^{(k)}| \leq k(\lambda)$ for some function k of λ. Because this constraint is on the sum of

the *absolute* values of β, a subset of the estimated β are zero (Tibshirani, 1996). This is a form of variable selection, where the number of non-zero coefficients can be controlled by the regularization parameter λ. The optimal value of λ is chosen by minimizing the loss function for a separate test set of data.

Software to do this is available in R, either through the `lars` package for least-squares loss (Efron et al., 2004), or through the more recent `glmpath` package which is also available for the binomial and Poisson log-likelihoods (Park and Hastie, 2006). A new faster algorithm by Friedman et al. (2007) should work for more general loss functions. Each of these implementations calculate the whole path for the regularization, for all different values of the regularization, and thus allow easy selection of the optimal λ using test set deviance. In practice, we apply the L_1 procedure to all unique basis functions only — we filter out all replicated bases in the boosted MARS model before the L_1 step.

11.3 Comparing boosted MARS and L_1 boosted MARS

We can also choose the size of the final (unregularized) boosted MARS model by deciding how many boosting steps to retain in the model; we do this by choosing the model size that minimizes the loss function for a separate test set. However, while we can choose model size for models both with and without regularization, using regularization gives us more flexibility in model choice. With the unregularized model the coefficients are pre-set and we must include all estimated basis functions up to a particular iteration. There is no such constraint for the regularized model.

Further, with regularization, the final model contains only a subset of the total basis functions. In our experience, these models often contain one tenth or less of the total number of basis functions. These smaller models allow quicker predictions on other data sets,

which can be important, for example, when creating detailed species distribution maps over large regions. In a more general look at L_1 regularization of sets of "rules" or basis functions, Friedman and Popescu (2004) report performance comparable to the best alternative methods. However, they stress that the primary benefit of this procedure is in providing interpretable models. In our setup, if the fitted L_1 boosted MARS model contains a small number of basis functions, it has similar properties of interpretability to any MARS model (Friedman, 1991). However, the extra flexibility in building and scaling back the MARS basis functions should improve predictive performance over a single MARS model.

11.4 Orthogonalization

Although L_1 regularization is powerful, it may struggle when there is high collinearity between covariates. Thus we also considered using an orthogonalization of the basis matrix before applying the regularization procedure. If we denote the basis matrix for this problem by $H(\mathbf{x})$, then the singular value decomposition (SVD, e.g. Hastie et al., 2001, p487) of H is $H = UDV^T$. If H is an $(n \times p)$ matrix with rank $d \leq n, p$, then U $(n \times d)$ and V $(p \times d)$ are orthogonal matrices and D $(d \times d)$ is a diagonal matrix. We then apply the L_1 regularization procedure using the orthogonal matrix U in place of the original basis matrix H.

We do not further consider this orthogonalization because in practice we found no improvement in fit over the L_1 boosted MARS mdoel. Further, while each MARS basis function depends on only a small subset of covariates, these orthogonalized basis functions each rely on many if not all covariates. Thus the sparsity and simplicity of MARS is lost by using such an orthogonalization.

Chapter 12

Evaluation of Boosted MARS

We evaluate these boosted MARS models for two different loss functions: Gaussian least squares and binomial logistic loss. We use the data previously described in Section 8.2 of presence/absence records of longfin eels to test the logistic loss function. We also use simulation studies to assess both the least squares and logistic loss functions. In all examples we use minimum validation set deviance as an assessment measure. In creating the boosted MARS models, we use the `mars` function in the `mda` package of R at each boosting step. The L_1 regularization is calculated using `lars` for the least-squares loss or `glmpath` for logistic loss.

12.1 Simulation studies

The simulation models we consider in this section are derived from the simulation model used in Section 8.1 and illustrated in Figure 8.1. Let $\eta_(\mathbf{x})$ denote the linear predictor for that

model (8.1):

$$\eta(\mathbf{x}) = \phi\left(\frac{x_1 - 0.5}{0.125}\right) + ((4x_2 - 2 \vee 0) \wedge 1) + 1_{\{x_3 > 0.5, x_4 > 0.5\}} + |x_5 + x_6 - 1| + c(\pi) \quad (12.1)$$

Then the generative distributions of the binomial and Gaussian models are

$$\begin{array}{lll} \text{Binomial:} & y \sim \text{ Binomial } (p(\mathbf{x}), 1) & \text{for } \text{logit}(p(\mathbf{x})) = \frac{1}{2}\eta(\mathbf{x}) \\ \text{Gaussian:} & y = \frac{1}{2}\eta(\mathbf{x}) + \epsilon & \text{for } \epsilon \sim \text{ N}(0, 1). \end{array}$$

In these simulations we use the multiple $\frac{1}{2}$ of η to increase the amount of variability in the data and hence to avoid overfitting. For the binomial model $c(\pi)$ in (12.1) is chosen so that the population mean of y is $\pi = 0.3$. For the Gaussian model we set $c(\pi) = 0$. The covariates are independent and randomly generated from the 10-dimensional unit cube uniform distribution, $\mathbf{x} \sim \text{Uniform}_{10}(0, 1)$. Ten realizations of each model are used to provide an indication of variability in results.

We consider two different sample sizes of 500 and 2,000, using either 40 or 100 boosting steps. For the 100 step implementation, the shrinkage factor is $\epsilon = 0.02$. For 40 boosting steps we used both $\epsilon = 0.01$ and 0.05. Note that we are considerably under-fitting the boosted MARS model to the data for 40 boosting steps with $\epsilon = 0.01$, as $0.01 \times 40 < 1$. However, this has the effect of reducing the total number of unique basis functions available for the subsequent L_1 fit, which may reduce overfitting. We do not provide results for this unregularized model, as the underfitting implies that test and training set deviances are considerably larger than for any other model.

We compare the boosted MARS results with those for gradient boosted regression trees. We use 10,000 trees, each with a maximum depth of 5, and set the shrinkage parameter $\epsilon = 0.001$. As a baseline comparison, we also provide deviances for single MARS models. For the Gaussian model, MARS is fit to the data directly. As there is no logistic loss implementation of MARS readily available, we use least-squares loss to fit a MARS model

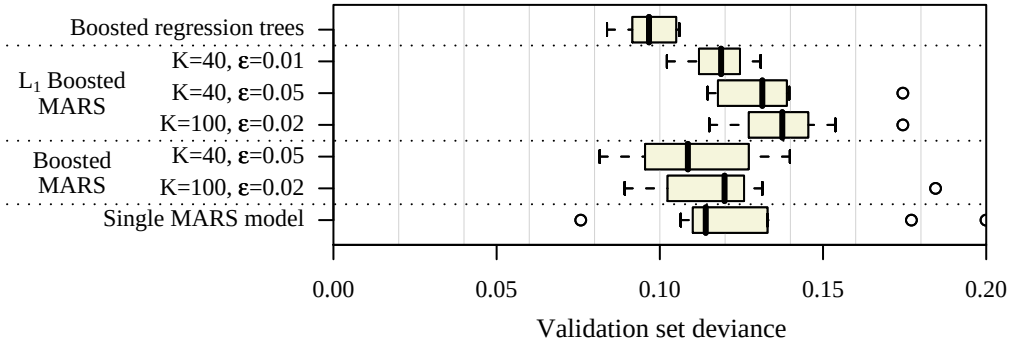
to the binary $\{0,1\}$ data as a first step. The coefficients of the basis functions are then re-estimated using the true loss function to create the final fit. The MARS models for the Gaussian data contain 20 and 10 basis functions for the large and small data sets respectively. MARS models for the binomial data contain 10 and 6 basis functions. These were the optimal model size as calculated for each data set.

Tuning parameters are optimized by minimizing the deviance on a separate test set. This test set is either 250 or 100 observations (one half the size of the training set). The tuning parameters are the number of boosting iterations to include in the boosted trees and boosted MARS models, and the value of the tuning parameter λ for the L_1 boosted MARS (11.3). The results provided here are the deviances of those best-fitting models as measured for separate validation sets of size 4,000.

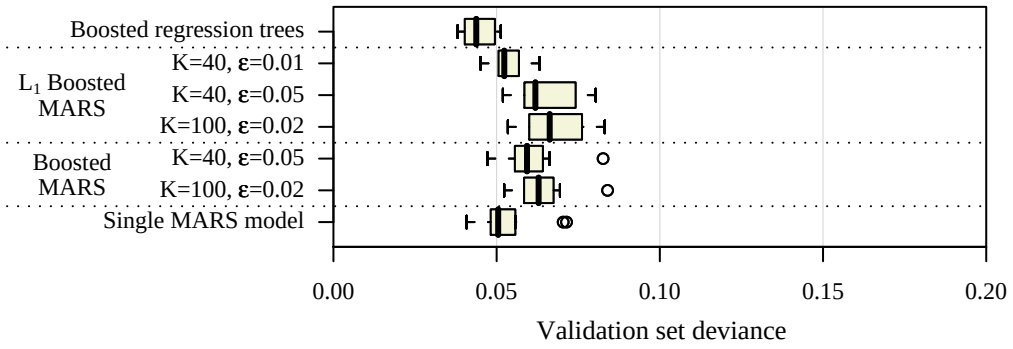
Across all data sets, there was a slight improvement in test set deviance when using thresholding of the hinge functions, but little difference when smoothing was also applied. Thus all the results provided here are for models with smoothing and thresholding. We look at the results for these models fit using both stochastic and non-stochastic boosting.

The results were very different for the Gaussian and logistic data. For both sample sizes of the Gaussian simulations (Figure 12.1) the L_1 boosted MARS models tend to overfit the test set data, and thus have high validation set error. The only exception to this is the L_1 boosted MARS model with $K = 40$ and $\epsilon = 0.01$ which was fit using fewer unique basis functions than the other models. For the larger sample size, the single MARS model does as well as any boosted MARS model, while for $n = 500$, the unregularized boosted MARS model ($K = 40, \epsilon = 0.05$) with no regularization slightly outperforms the single MARS model. However, none of the MARS models has deviance as low as the boosted regression tree model.

We also tried stochastic boosting, building the MARS model at each iteration using half



(a) *Sample size:* $n = 500$

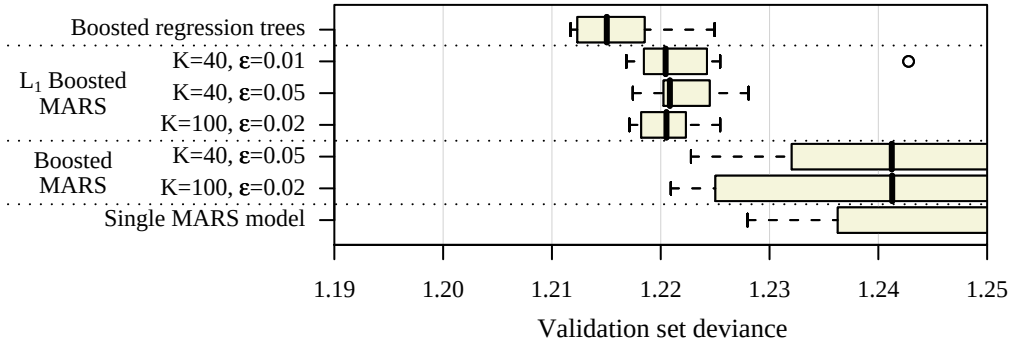


(b) *Sample size:* $n = 2,000$

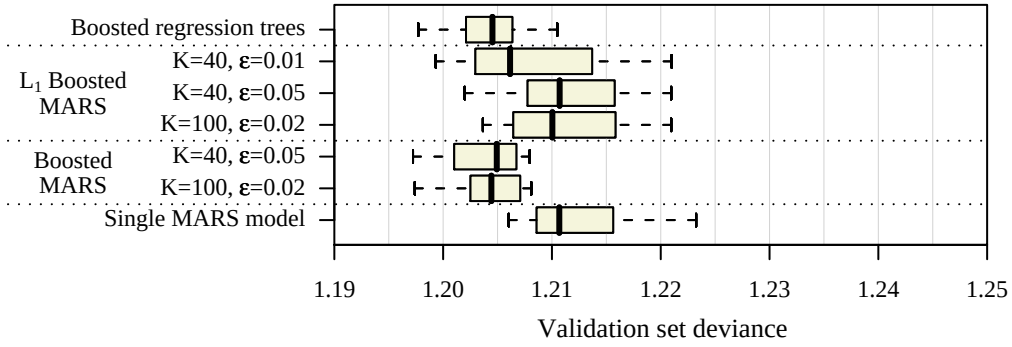
Figure 12.1: *Validation set deviances across 10 realizations for the Gaussian simulation data, using least-squares loss with non-stochastic boosting. For the boosted MARS models, the number of iterations K and the shrinkage parameter ϵ are given.*

the test set data sample at random (results not shown). This made little difference to the boosted MARS models without regularization. However the median deviance increased considerably for the L_1 boosted MARS model ($K = 100, \epsilon = 0.02$). This lack of fit is due to the large number of possible basis functions available for the L_1 regularization in this example — the stochasticity created many more similar but not identical MARS basis functions. In contrast, for non-stochastic boosting, especially when ϵ is small, many of the same basis functions appear in subsequent steps because the fitted model changes slowly.

For the binomial data (Figure 12.2), using boosted MARS can considerably improve fit over using a single MARS model. For the smaller data set, the L_1 regularization improves model



(a) *Sample size:* $n = 500$

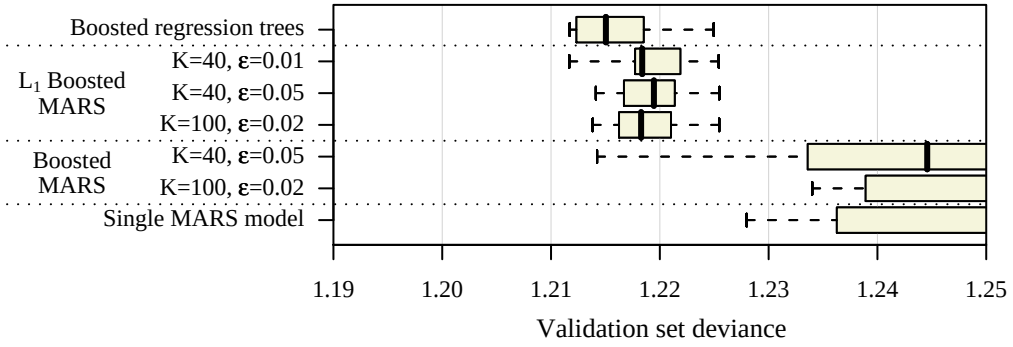


(b) *Sample size:* $n = 2,000$

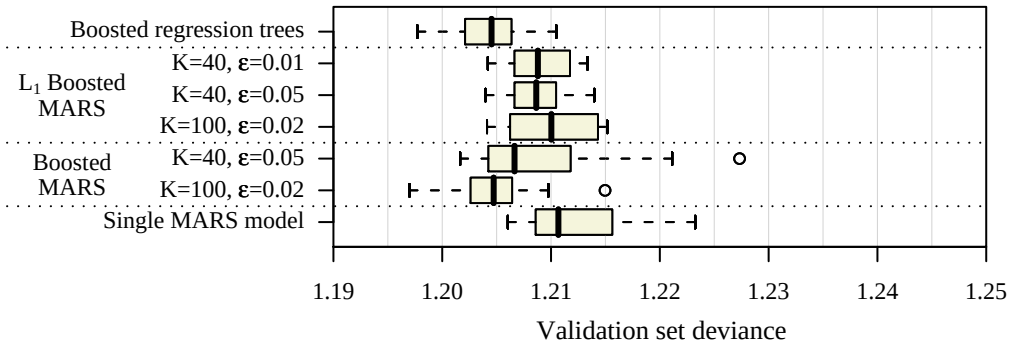
Figure 12.2: *Validation set deviances across 10 realizations for the binomial simulation data using a logistic loss function with non-stochastic boosting. For the boosted MARS models, the number of iterations K and the shrinkage parameter ϵ are given.*

fit over the unregularized models. However, boosted regression trees still out-performs all boosted MARS models. When $n = 2,000$, however, the situation is reversed — the unregularized models perform better than the L_1 boosted MARS, and they also match the performance of boosted regression trees.

Using stochastic boosting has a small effect on these results (Figure 12.3). The deviances for the L_1 boosted MARS models for $n = 500$ are marginally reduced, though not enough to outperform boosted trees. For the larger data set, the variability in results for the L_1 models is reduced when using stochastic boosting. However, the unregularized boosted MARS models still outperform the regularized models.



(a) *Sample size:* $n = 500$



(b) *Sample size:* $n = 2,000$

Figure 12.3: *Validation set deviances across 10 realizations for the binomial simulation data using a logistic loss function with stochastic boosting. For the boosted MARS models, the number of iterations K and the shrinkage parameter ϵ are given.*

In summary, the efficacy of boosted MARS and L_1 boosted MARS appears to be highly model and data specific. However, the models do appear to be more effective, as measured against a single MARS model, for logistic rather Gaussian models. Further, in the case of a large binomial data set, unregularized boosted MARS performs comparably to boosted regression trees.

12.2 The longfin eel

We also assess the boosted MARS models using presence-absence data of longfin eels that we have seen previously in Section 8.2. In these analyses we use the same training/test splits as before, however we use the true presences and absences to build models, rather than the presence-only data. For this problem we use the logistic loss function based on the binomial likelihood. Three of the covariates in this data set have highly skew distributions, so we transformed these covariates prior to model fitting to prevent overfitting of outlying data points. We used a square root transformation rather than log because some of the covariate values were zero. One of these covariates was downstream distance to the sea — in presenting the marginal effects of this covariate we have transformed this covariate back to its original scale.

We provide results for boosted MARS models with and without regularization for 40, 100, 200 and 1,000 boosting iterations with $\epsilon = 0.01, 0.02, 0.01, 0.005$ respectively. As in the simulation study, the first set of parameters will result in a boosted MARS model that is underfit, and so we only show results for the L_1 model for these data. As in the previous section, all MARS models used in the boosting procedures contain 10 basis functions.

A single MARS model with 20 basis functions, the optimal model size for these data, is provided as a baseline comparison. We also compare the boosted MARS models with boosted regression trees. These were built using 10,000 trees, each of depth 5, and using a shrinkage parameter of $\epsilon = 0.001$. The models are compared using minimum test set deviance.

For these data (Figure 12.4) we see the same pattern as for the smaller logistic model in Figure 12.2, in that the L_1 boosted MARS models perform better than the models with no regularization. Further, all boosted MARS models increase model performance over the

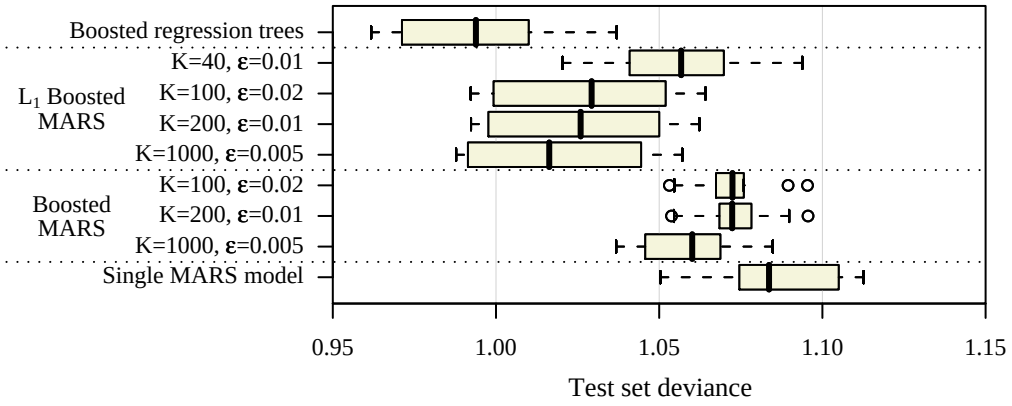
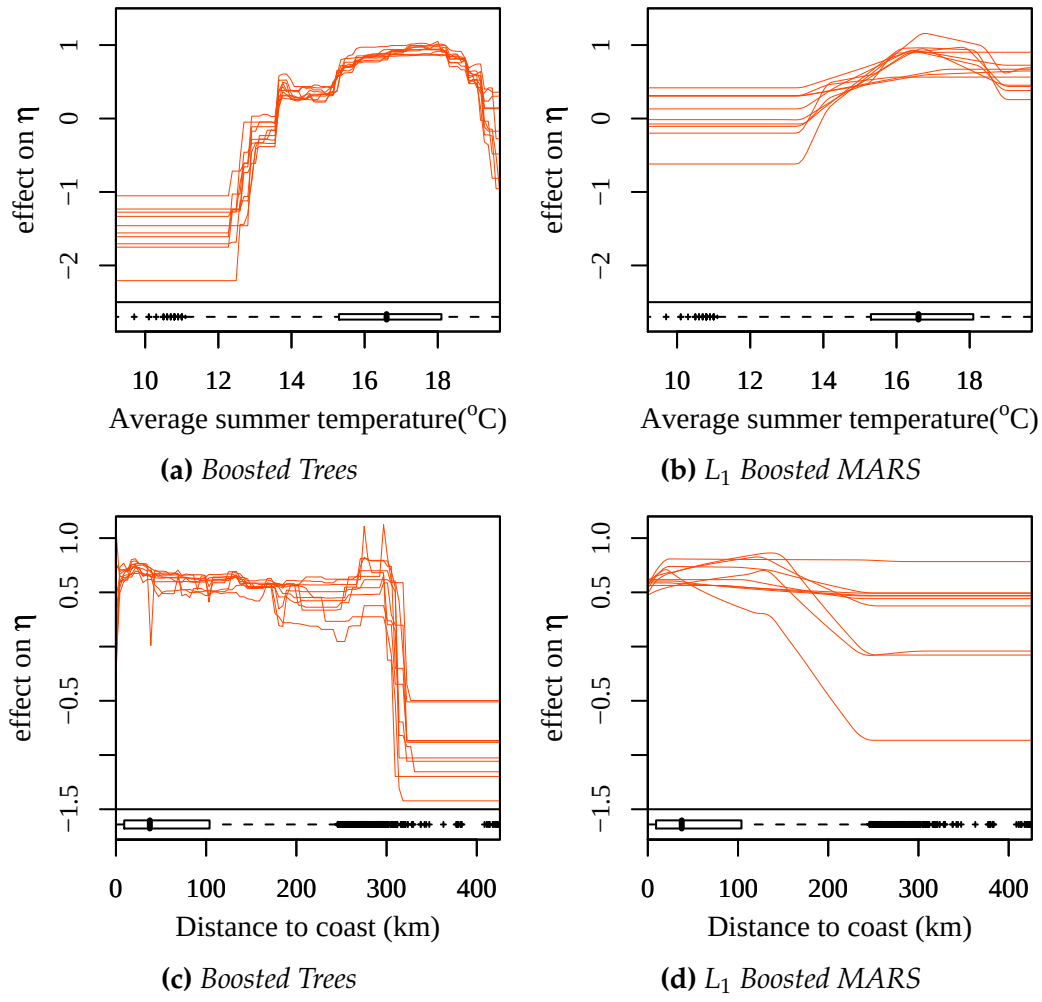


Figure 12.4: *Test set deviances across 10 realizations for the longfin eel data using a logistic loss function, without stochastic boosting. For the boosted MARS models, the number of iterations K and the shrinkage parameter ϵ are given.*

baseline single MARS model. However, boosted regression trees outperform all of our boosted MARS models.

For these data, the better performing L_1 models were those with larger K, which indicates that they were fit using larger numbers of basis functions. This would imply that stochastic boosting might further improve model fit, however in practice it has little effect on the results (not shown), except to reduce test set deviance of the L_1 boosted MARS model ($K = 40, \epsilon = 0.01$) to levels similar to the other L_1 boosted MARS models.

In Figure 12.5 we provide estimates of marginal effects for two of the variables, average summer temperature and distance to coast. These are from the optimal fitted models, as measured by test set deviance, of boosted trees and L_1 boosted MARS. The marginal effects for the L_1 boosted MARS are considerably smoother than those for boosted trees. However, the effects for L_1 boosted MARS have a higher level of variability across the 10 different data sets, especially for larger distances to the coast. For some of the L_1 boosted MARS models, the estimated effect size for distance to coast is negligible. In contrast the boosted trees model consistently estimates a large step in marginal effect around 300km



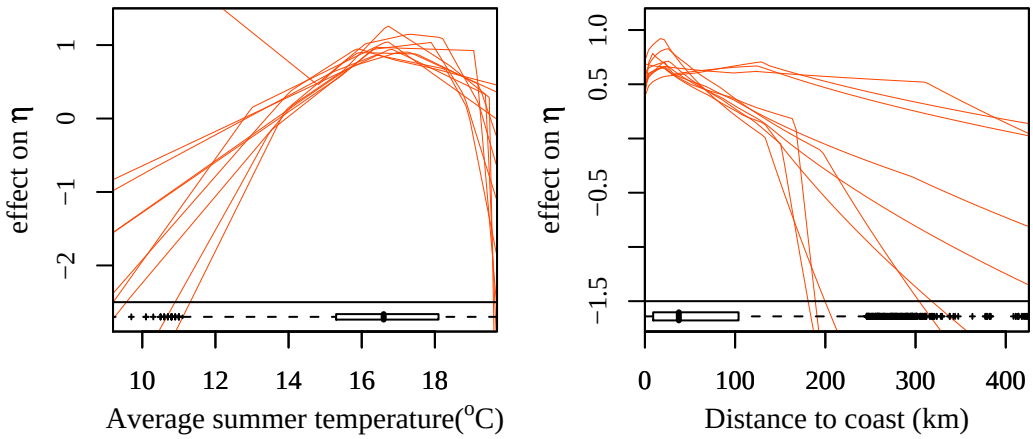
(a) *Boosted Trees* **(b)** L_1 *Boosted MARS*

(c) *Boosted Trees* **(d)** L_1 *Boosted MARS*

Figure 12.5: *Estimated marginal effects on η of average summer temperature and downstream distance to coast for ecah of ten data sets. These are for optimal boosted trees, and L_1 boosted MARS models ($K = 100, \epsilon = 0.02$; thresholding of the extreme 5% of data; smoothing with parameter 5% of IQR). The square root of distance to coast was used in fitting the models.*

from the coast — the 99th percentile of distance. The setting of the thresholding parameter for the MARS hinge functions will have a considerable effect on the boosted MARS marginal effect at this point. More generally, L_1 boosted MARS provides smaller estimated effect sizes than boosted trees, which will be in part due to the L_1 regularization.

We can observe the extrapolation issues of MARS basis functions in Figure 12.5, which



(a) *Summer temperature* **(b)** *Distance to coast*

Figure 12.6: *Estimated marginal effects on η of average summer temperature and downstream distance to coast for L_1 boosted MARS models with no thresholding or smoothing. The boosting parameters are $K = 100, \epsilon = 0.02$. Each line indicates the effect for one of the ten data sets. The square root of distance to coast was used in fitting the models*

provides estimated marginal effects for L_1 boosted MARS with no thresholding. Effects for extremes in average summer temperature and for a high distance to coast head off to $-\infty$ at vastly differing rates. This creates a very high variance in estimated marginal effects when extrapolating beyond the training set, and also within the training data set, especially for high distances to the coast. This indicates that thresholded hinge functions (Figure 12.6) should be used when fitting boosted MARS models, to provide more stable estimates of extremes of marginal effects.

Chapter 13

Boosted MARS: Conclusions

While gradient boosted regression trees provide a very flexible modeling procedure, the resulting fitted model and estimated marginal effects are not smooth. We have outlined a smoother model that still retains the flexibility of boosted trees but that uses MARS basis functions in place of regression trees. By using post-hoc smoothing of MARS basis functions, we can create fitted models with continuous first derivatives. Further, we introduced a post-hoc thresholding of hinge functions to reduce the problems with extrapolation that are common to MARS models. The boosted MARS models can either be used as-is, or with subsequent L_1 regularization.

We assessed the boosted MARS models using simulations of Gaussian and binomial data, and analyses of presence-absence data of fish in New Zealand rivers. Regularized and unregularized boosted MARS models had lower validation and test set deviance than single MARS models for the logistic data, however this was not always true for the Gaussian simulations. Further, the relative performance of boosted MARS and L_1 boosted MARS models varied across loss functions and data sets. The L_1 boosted MARS models appear

to considerably improve model fit over single MARS models for binomial data where the complexity of the model was large relative to the size of the training set. However, even in these situations, the models do not outperform boosted regression trees.

The efficacy of boosted MARS (with or without regularization) has not been consistently proven in these analyses. Part of the problem is that L_1 regularization using many basis functions quickly overfits the test set data. Further, the choice of thresholding parameter for the hinge functions can impact extreme values of marginal effect. It may be that further work can improve upon the performance of this modeling procedure. However, these models do provide smoother estimates of marginal effects, which may aid interpretation in actual modeling problems.

Appendix A

Miscellaneous calculations

A.1 Manly's exponential model and the naive logistic model

Manly et al. (2002) have shown that the log-probability model $\mathbb{P}(y = 1|\mathbf{x})$ can be estimated from presence-only data using an intercept adjustment of the naive logistic model. We repeat this derivation here, using our notation for clarity.

The presence-only data are generated by $\mathbb{P}(z = 1|s = 1, \mathbf{x})$. An application of Bayes Theorem gives

$$\begin{aligned}\mathbb{P}(z = 1|s = 1, \mathbf{x}) =& \mathbb{P}(z = 1|y = 1, s = 1, \mathbf{x})\mathbb{P}(y = 1|s = 1, \mathbf{x}) \\ &+ \mathbb{P}(z = 1|y = 0, s = 1, \mathbf{x})\mathbb{P}(y = 0|s = 1, \mathbf{x}) \\ =& \mathbb{P}(z = 1|y = 1, s = 1, \mathbf{x})\mathbb{P}(y = 1|s = 1, \mathbf{x}).\end{aligned}$$

The second equality follows because $\mathbb{P}(z = 1|y = 0, s = 1, \mathbf{x}) = 0$, i.e. a location cannot have an observed presence ($z = 1$) if there is no true presence ($y = 0$). As the sampling of

observed given true presences is not model dependent, we use the result (6.3) that $\mathbb{P}(z = 1|y = 1, s = 1, \mathbf{x}) = \frac{n_p}{n_p+\pi n_a}$.

Now we consider $\mathbb{P}(y = 1|s = 1, \mathbf{x})$. If we were using a logistic model, this would be the case-control probability, but here we are assuming the log probability model $\mathbb{P}(y = 1|\mathbf{x}) = e^{\eta(\mathbf{x})}$. Bayes' Theorem gives

$$\begin{aligned}\mathbb{P}(y = 1|s = 1, \mathbf{x}) &= \frac{\mathbb{P}(s = 1|y = 1, \mathbf{x})\mathbb{P}(y = 1|\mathbf{x})}{\mathbb{P}(s = 1|y = 0, \mathbf{x})\mathbb{P}(y = 0|\mathbf{x}) + \mathbb{P}(s = 1|y = 1, \mathbf{x})\mathbb{P}(y = 1|\mathbf{x})} \\ &= \frac{e^{\eta(\mathbf{x})}}{\gamma\left(1 - e^{\eta(\mathbf{x})}\right) + e^{\eta(\mathbf{x})}}\end{aligned}$$

where $\gamma = \mathbb{P}(s = 1|y = 0, \mathbf{x})/\mathbb{P}(s = 1|y = 1, \mathbf{x})$. To show that $\mathbb{P}(s = 1|y, \mathbf{x})$ does not depend on $\mathbf{x}$ we use Bayes' Theorem

$$\mathbb{P}(s = 1|y, \mathbf{x}) = \frac{\mathbb{P}(s = 1|y)\mathbb{P}(\mathbf{x}|s = 1, y)}{\mathbb{P}(\mathbf{x}|y)} = \mathbb{P}(s = 1|y).$$

This is true because given the true presence at a location, the locations in our sample are sampled at a rate proportional to $f_1(\mathbf{x}) = \mathbb{P}(\mathbf{x}|y = 1)$, and similarly for true absences. Thus

$$\gamma = \frac{\mathbb{P}(s = 1|y = 0)}{\mathbb{P}(s = 1|y = 1)} = \frac{\mathbb{P}(y = 0|s = 1)\mathbb{P}(s = 1)/\mathbb{P}(y = 0)}{\mathbb{P}(y = 1|s = 1)\mathbb{P}(s = 1)/\mathbb{P}(y = 1)} = \frac{\pi\mathbb{P}(y = 0|s = 1)}{(1 - \pi)\mathbb{P}(y = 1|s = 1)}.$$

The expected number of true presences in the sample is $n_p + \pi n_a$ (all observed presences plus a proportion π of the pseudo-absences), so $\mathbb{P}(y = 1|s = 1) = (n_p + \pi n_a)/(n_p + n_a)$ and hence

$$\frac{\mathbb{P}(s = 1|y = 0)}{\mathbb{P}(s = 1|y = 1)} = \frac{\pi\left(1 - \frac{n_p+\pi n_a}{n_p+n_a}\right)}{(1 - \pi)\frac{n_p+\pi n_a}{n_p+n_a}} = \frac{\pi n_a}{n_p + \pi n_a}.$$

Plugging everything back into to the first equation we get

$$\begin{aligned}\mathbb{P}(z=1|s=1,\mathbf{x}) =& \mathbb{P}(z=1|y=1,s=1,\mathbf{x})\mathbb{P}(y=1|s=1,\mathbf{x}) \\ =& \frac{n_p}{n_p+\pi n_a}\frac{e^{\eta(\mathbf{x})}}{\frac{\pi n_a}{n_p+\pi n_a}\left(1-e^{\eta(\mathbf{x})}\right)+e^{\eta(\mathbf{x})}} \\ =& \frac{\frac{\pi n_a}{n_p}e^{\eta(\mathbf{x})}}{1+\frac{\pi n_a}{n_p}e^{\eta(\mathbf{x})}}\end{aligned}$$

after some manipulation. This looks like the logistic model, with an additive adjustment in η. Thus, if π is known, we can estimate the true η (for the *log* model) by subtracting $\log(\frac{\pi n_a}{n_p})$ from the η estimated via the naive *logistic* model. This is similar in flavor to case-control modeling, but with a very different underlying model.

A.2 Estimation of π for presence-only logistic regression

In a logistic regression model, where $\eta(\mathbf{x}) = \mathbf{x}^T\beta$ is linear and X is of full rank, we can estimate both π and β. To do this we develop a modified version of the EM procedure developed in Section 6.4, where π is also estimated at every step. As π is unknown we need to use the full likelihood (6.10) for the naive and true presences $\mathbf{z}$ and $\mathbf{y}$:

$$\begin{aligned}L(\pi,\eta|\mathbf{y},\mathbf{z},X) &= \left(\prod_i \mathbb{P}(y_i|s_i=1,\mathbf{x}_i)\right)\left(\prod_i \mathbb{P}(z_i|y_i,s_i=1,\mathbf{x}_i)\right) \\ &= L_1(\pi,\eta|\mathbf{y},\mathbf{z},X)\ L_2(\pi|\mathbf{y},\mathbf{z},X)\end{aligned}$$

Note that L_2 depends on π only and L_1 depends on π and η only through $\eta^* = \eta + \log\left(\frac{n_p+\pi n_a}{\pi n_a}\right)$. Thus we can re-parameterize (π,η) to (π,η^*), so that L_1 depends only on η^*, and we can maximize L_1 and L_2 separately.

L_1 is maximized by fitting a logistic regression model to $\mathbf{y}$, to obtain $\eta^*(\mathbf{x})$. Additionally,

the maximum likelihood estimate for π is easily found by maximizing L_2. Using the probabilities (6.3) and noting that $\mathbb{P}(z_i = 0|y_i = 0, s_i = 1, \mathbf{x}_i) = 1$ as all true absences must occur in the background data:

$$\begin{aligned} L_2(\pi|\mathbf{y}, \mathbf{z}, X) &= \prod_i \mathbb{P}(z_i|y_i, s_i = 1, \mathbf{x}_i) \\ &= \left(\prod_{z_i=1} \mathbb{P}(z_i = 1|y_i = 1, s_i = 1, \mathbf{x}_i) \right) \left(\prod_{z_i=0, y_i=1} \mathbb{P}(z_i = 0|y_i = 1, s_i = 1, \mathbf{x}_i)^{y_i} \right) \\ &= \left(\frac{n_p}{n_p + \pi n_a} \right)^{n_p} \left(\frac{\pi n_a}{n_p + \pi n_a} \right)^{\sum_{z_i=0} y_i} \end{aligned}$$

$$\Rightarrow \frac{\partial}{\partial \pi} \log L_1 = -\left(n_p + \sum_{z_i=0} y_i \right) \frac{n_a}{n_p + \pi n_a} + \sum_{z_i=0} y_i \frac{1}{\pi} \Rightarrow \hat{\pi} = \frac{1}{n_a} \sum_{z_i=0} y_i$$

Thus we have an intuitive estimate for π - the fraction of the background data that are true presences. Further, we can retrieve $\hat{\eta}$ from $\{\hat{\pi}, \hat{\eta}^*\}$ by subtracting $\log(\frac{n_p + \hat{\pi} n_a}{\hat{\pi} n_a})$ from the intercept.

Maximizing L_1 and L_2 forms the M-step of the EM algorithm. Because both $\log L_1$ and $\log L_2$ are linear in the y_i, the expectation step of the EM algorithm is still to replace each y_i with its expectation conditional on the parameter estimates from the previous step. So the EM algorithm for estimating η and π follows as in Figure 6.1, except that in the maximization step, before adjusting the intercept of $\hat{\eta}^{*(k)}$, we update the estimate of π: $\pi^{(k)} = \frac{1}{n_a} \sum_{i \in \mathbf{A}} \hat{y}_i^{(k-1)}$.

This algorithm appears to provide an ideal solution for estimating π for presence-only data. However, we have already seen in Figure 6.2 that even when the logistic regression model is the true generative model, estimates of π and the intercept α are highly correlated. In practice the logistic regression model is both unlikely to be the true model and unverifiable. Although logistic regression models are a useful approximation, to estimate π we are placing a greater reliance on that model being the truth. It is safer to have a good

estimate of π from another source, and run sensitivity analyses on the results.

A.3 Proof of penalized likelihood

We wish to maximize the log-likelihood

$$\log L(\eta|X) = \sum_{i\,;\,z_i=1} \log f_1(\mathbf{x}_i)$$

subject to the constraint that the density integrate to 1: $\int f_1(\mathbf{x})d\mathbf{x} = 1$. We take a specific case of the proof of Silverman (1982), to show this is equivalent to unconstrained maximization of the penalized log-likelihood

$$\log L^*(\eta|X) = \sum_{i\,;\,z_i=1} \log f_1(\mathbf{x}_i) - n_p \int_{\mathbf{x}} f_1(\mathbf{x})d\mathbf{x}. \tag{A.1}$$

To do this, we define $g(\mathbf{x}) = \log f_1(\mathbf{x})$ and the two loss functions:

$$\mathbf{L}(g) = -\frac{1}{n_p} \sum_{i\,;\,z_i=1} g(\mathbf{x}_i)$$

$$\mathbf{L}^*(g) = -\frac{1}{n_p} \sum_{i\,;\,z_i=1} g(\mathbf{x}_i) + \int_{\mathbf{x}} e^{\,g(\mathbf{x})} d\mathbf{x},$$

so that $\mathbf{L}(g) = -\frac{1}{n_p} \log L(\eta|X)$ and $\mathbf{L}^*(g) = -\frac{1}{n_p} \log L^*(\eta|X)$. Now define

$$g^*(\mathbf{x}) = g(\mathbf{x}) - \log \int_{\mathbf{x}} e^{\,g(\mathbf{x})} d\mathbf{x},$$

and note that

$$\int_{\mathbf{x}} e^{\,g^*(\mathbf{x})} = \int_{\mathbf{x}} \frac{e^{\,g(\mathbf{x})}}{\int_{\mathbf{x}} e^{\,g(\mathbf{x})} d\mathbf{x}} d\mathbf{x} = 1.$$

Thus g^* is the standardized version of g, in that it integrates to 1, and so is a proper density. We show that minimizing the penalized loss $\mathbf{L}^*$ with respect to g^* is equivalent to minimizing it with respect to g:

$$\mathbf{L}^*(g^*) = -\frac{1}{n_p} \sum_{i\,;\, z_i=1} g(\mathbf{x}_i) + \log \int_{\mathbf{x}} e^{\,g(\mathbf{x})} d\mathbf{x} + \int_x e^{\,g^*(\mathbf{x})} d\mathbf{x}$$

$$= \mathbf{L}^*(g) - \int_{\mathbf{x}} e^{\,g(\mathbf{x})} d\mathbf{x} + \log \int_{\mathbf{x}} e^{\,g(\mathbf{x})} d\mathbf{x} + 1.$$

Noting that $t - \log t \geq 1$ for $t > 0$, with equality only when $t = 1$, we have $\mathbf{L}^*(g^*) < \mathbf{L}^*(g)$ with equality only if g integrates to 1. Thus unconstrained maximization of the penalized likelihood is equivalent to constrained maximization of the original likelihood.

Appendix B

Implementation of the presence-only models

B.1 The `ecogbm` package

Models 2, 3 and 4 described in Chapter 7 were implemented by adapting the `gbm` package in R. The new package, named `ecogbm` retains much of the functionality of `gbm`, but is implemented for the three new presence-only models.

The original `gbm` package is set up so that the different distributions are all programmed in C++ and must be directly referenced in another C++ source file. Although this provides for very fast model fitting, it is not easy to add another distribution, unlike the relative ease of adding another family for the generalized linear model functions. In addition, two of the three new models require an extra parameter π to be specified in the function call. Thus I have copied version 1.5-7 of `gbm`, stripped out the existing distributions, and

replaced them with the three distributions `ponlysample`, `ponlypopn` and `ponlylog`. These distributions implement the sample logistic, population logistic and population log models respectively. The default distribution `ponly` defers to the `ponlypopn` distribution. These new distributions in `ecogbm` work the same as existing distributions in `gbm`, with the same functions for performance analysis, fitted values, variable importance and marginal effects.

Bibliography

Araújo, M. B., Thuiller, W., Williams, P. H., and Reginster, I. (2005). Downscaling european species atlas distributions to a finer resolution: implications for conservation planning. *Global Ecology and Biogeography*, 14:17–30.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159.

Buston, P. (2003). Social hierarchies: Size and growth modification in clownfish. *Nature*, 424:145–146.

De'ath, G. (2007). Boosted trees for ecological modeling and prediction. *Ecology*, 88:243–251.

Della Pietra, S., Della Pietra, V., and Lafferty, J. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:1–13.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38.

Dudík, M., Schapire, R., and Phillips, S. (2006). Correcting sample selection bias in maximum entropy density estimation. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*, pages 323–330. MIT Press, Cambridge, MA.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression (with discussion). *Annals of Statistics*, 32:407–499.

Elith, J., Graham, C. H., Anderson, R. P., Dudk, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Soberón, J., Williams, S., Wisz, M. S., and Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29:129–151.

Engler, R., Guisan, A., and Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, 41:263–274.

Frair, J. L., Nielsen, S. E., Merrill, E. H., Lele, S. R., Boyce, M. S., andG B Stenhouse, R. H. M. M., and Beyer, H. L. (2004). Removing GPS collar bias in habitat selection studies. *Journal of Applied Ecology*, 41:210–212.

Freedman, B. (1994). *Environmental Ecology: The Ecological Effects of Pollution, Disturbance, and Other Stresses*. Academic Press, San Diego, California, Second edition.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Pathwise coordinate optimization.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 19:1–141.

Friedman, J. H. (2001). Greedy function approximation: the gradient boosting machine. *Annals of Statistics*, 29:1189–1232.

Friedman, J. H. and Popescu, B. E. (2004). Gradient directed regularization for linear regression and classification. Technical report, Stanford University, Department of Statistics.

Gotelli, N. J. (1995). *A primer of ecology*. Sinauer Associates, Inc, Sunderland, MA, U.S.A.

Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8:993–1009.

Guisan, A. and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135:147–186.

Guo, Q., Kelly, M., and Graham, C. H. (2005). Support vector machines for predicting distribution of sudden oak death in california. *Ecological Modelling*, 182:75–90.

Hastie, T. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, London.

Hastie, T., Tibshirani, R. J., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, First edition.

Johnson, C. J., Nielsen, S. E., Merrill, E. H., McDonald, T. L., and Boyce, M. (2006). Resource selection functions based on use-availability data: theoretical motivation and evaluation methods. *Journal of Wildlife Management*, 70:347–357.

Keating, K. A. and Cherry, S. (2004). Use and interpretation of logistic regression in habitat-selection studies. *Journal of Wildlife Management*, 68:774–789.

Kerfoot, W. C. and Sih, A. (1987). *Predation: Direct and Indirect Impacts on Aquatic Communities*. University Press of New England.

Lancaster, T. and Imbens, G. (1996). Case-control studies with contaminated controls. *Journal of Econometrics*, 71:145–160.

Leathwick, J. R., Elith, J., Francis, M. P., Hastie, T., and Taylor, P. (2006). Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series.*

Leathwick, J. R., Rowe, D., Richardson, J., Elith, J., and Hastie, T. (2005). Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology*, 50:2034–2051.

Lillesand, T. M., Kiefer, R. W., and Chipman, J. W. (2004). *Remote sensing and image interpretation*. John Wiley & Sons Ltd.

Liu, B., Dai, Y., Li, X., Lee, W. S., and Yu, P. S. (2003). Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining*, pages 179–186.

Manly, B. F. J., McDonald, L. L., Thomas, D. L., McDonald, T. L., and Erickson, W. (2002). *Resource selection by animals: statistical design and analysis for field studies*. Chapman & Hall, New York, Second edition.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London, Second edition.

McDowall, R. M. and Richardson, J. (1983). The new zealand freshwater fish survey, a guide to input and output. *New Zealand Ministry of Agriculture and Fisheries, Fisheries Research Division Information Leaflet*, 12:1–15.

Milsom, T. P., Langton, S. D., Parkin, W. K., Peel, S., Bishop, J. D., Hart, J. D., and Moore, N. P. (2000). Habitat models of bird species' distribution: an aid to the management of coastal grazing marshes. *Journal of Applied Ecology*, 37:706–727.

New Zealand Statistics (2005). *Fish Species Articles: Freshwater Eels*, `http://www.stats.govt.nz/environment/environmental-statistics/publications/default.htm`.

Owen, A. (2006). Infinitely unbalanced logistic regression.

Park, M. Y. and Hastie, T. (2006). L_1 regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society B*, (to appear).

Parker, J. M. and Anderson, S. H. (2003). Habitat use and movements of repatriated wyoming toads. *Journal of Wildlife Management*, 67:439–446.

Pearce, J. L. and Boyce, M. S. (2006). Modeling distribution and abundance with presence-only data. *Journal of Applied Ecology*, 43:405–412.

Peterson, A. T. and Vieglais, D. A. (2001). Predicting species invasions using ecological niche modeling. *BioScience*, 51:363–371.

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

Reese, G. C., Wilson, K. R., Hoeting, J. A., and Flather, C. H. (2005). Factors affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications*, 15:554–564.

Ridgeway, G. (2004). gbm: Generalized boosted regression models. R package, version 1.3-5. `http://www.i-pensieri.com/gregr/gbm.shtml`.

Rosset, S., Zhu, J., and Hastie, T. (2004). Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973.

Roughgarden, J. (1983). Competition and theory in community ecology. *The American Naturalist*, 122:583–601.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, 10:795–810.

Stein, B. R. and Wieczorek, J. (2004). Mammals of the world: Manis as an example of data integration in a distributed network environment. *Biodiversity Informatics*, 1:14–22.

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., Erasmus, B. F. N., de Siqueira, M. F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., van Jaarsveld, A. S., Midgley, G. F., Miles, L., Ortega–Huerta, M. A., Peterson, A. T., Phillips, O. L., and Williams, S. E. (2004). Extinction risk from climate change. *Nature*, 427:145–148.

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

Walker, P. A. and Cocks, K. D. (1991). HABITAT: a procedure for modeling a disjoint environmental envelope for a plant or animal species. *Global Ecology and Biogeography Letters*, 1:108–118.

Weimerskirch, H., Bonadonna, F., Bailleul, F., Mabille, G., DellOmo, G., and Lipp, H. P. (2002). Gps tracking of foraging albatrosses. *Science*, 295:1259.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the 21st International Conference on Machine Learning*, pages 903–910.

Zammit, O., Descombes, X., and Zerubia, J. (2006). Burnt area mapping using support vector machines. *Forest Ecology and Management*, 234:S240.

Zaniewski, A. E., Lehmann, A., and Overton, J. M. (2002). Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, 157:261–280.