Regression Models, Fantastic Beasts, and Where to Find Them: A Simple Tutorial for Ecologists Using R

Luca Corlatti

Chair of Wildlife Ecology and Management, University of Freiburg, Freiburg, Germany.

Bioinformatics and Biology Insights
Volume 15: 1–19

DOI: 10.1177/11779322211051522

SAGE

ABSTRACT: Regression modeling is a workhorse of statistical ecology that allows to find relationships between a response variable and a set of explanatory variables. Despite being one of the fundamental statistical ideas in ecological curricula, regression modeling can be complex and subtle. This paper is intended as an applied protocol to help students understand the data, select the most appropriate models, verify assumptions, and interpret the output. Basic ecological questions are tackled using data from a fictional series, "*Fantastic beasts and where to find them*," with the aim to show how statistical thinking can foster curiosity, creativity and imagination in ecology, from the formulation of hypotheses to the interpretation of results.

KEYWORDS: Ecology, fantastic animals, habitat selection, model assumptions, regression modeling

RECEIVED: May 5, 2021. **ACCEPTED:** September 18, 2021.

TYPE: Methods and Protocols

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The article processing charge was funded by the German Research Foundation (DFG) and the University of Freiburg in the funding programme Open Access Publishing.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Luca Corlatti, Chair of Wildlife Ecology and Management, University of Freiburg, Tennenbacher Straße 4, 79106 Freiburg, Germany.
Email: luca.corlatti@wildlife.uni-freiburg.de

Introduction

In 1927, the magizoologist Newt Scamander published a seminal book where he described, with a fair amount of details, all the known magical animals of the world. The Ministry of Magic classified these beasts based on their perceived level of danger, from 1 to 5: "boring," 'harmless "competent wizard should cope," "dangerous," and "impossible to train or domesticate." After reading the book—and notwithstanding the details on the biology and behavior of each animal (Figure 1)—some crucial questions stuck in my head: what drives the distribution of these fantastic beasts, and the geographical variation in their level of danger? To put it in other words, why do beasts live where they live, especially the most dangerous ones? The reader will agree this is a fundamental ecological question! The distribution of animals and their phenotypes can be influenced by a variety of ecological factors such as food or climate, among others. Investigating these relationships may thus provide useful information about the evolutionary forces that shaped magical biological diversity and, ultimately, about adaptation. But how do we tackle this issue?

The scientific method posits that, after observing a phenomenon, we formulate hypotheses to explain its nature, put forward predictions, collect data, analyze them and finally evaluate our results in the light of our expectations. What about our fantastic beasts? We know that magical animals are widespread worldwide,[1] but we have no details about their current distribution, nor we have information about their type-specific abundance (I have deliberately avoided writing "species-specific" abundance: the species concept is fuzzy enough in the standard zoological literature!). All we know is where they originally come from, and what is the mean perceived level of danger of country-specific magical animal communities, which nonetheless can provide useful information about the potential drivers of diversity.

From the map reported in Figure 2, we observe that the native occurrence of magical animals and—even more so—the mean perceived level of danger of native animal communities, are seemingly greater in tropical and sub-tropical regions than in temperate zones. If climate plays a role, we may hypothesize that (1) the native distribution of fantastic beasts and (2) the mean danger level of a native community of beasts are related to temperature. Magizoology is not my field of expertise, and many magical beasts appear to be an odd mixture of "muggle" animals, which makes it quite difficult to hypothesize plausible biological patterns. Perhaps, at least for insect/reptile-like looking beasts, thermal niches with relatively higher environmental temperatures might be required to secure critical physiological processes, thereby shaping the observed uneven distribution. If so, we may predict that (1) the warmer the country, the more likely the native occurrence of magical animals and (2) the warmer the native country, the higher the mean perceived level of danger of the community of beasts that lives in it. To address these questions, we may resort to regression models.

Regression is "*a method that allows researchers to summarize how predictions or average values of an outcome vary across individuals defined by a set of predictors*,"[2] that is, it allows to explore the relationships between a response variable and a set of explanatory variables. Ecological processes often have many drivers or confounding factors that contribute to affect them simultaneously, thus it is unsurprising that multiple regression dominates in the ecological literature.[3] Furthermore, regression can handle different kinds of response variables, which are common in ecological data,[4] and it can be extended to allow for hierarchical sampling designs, which are widespread in ecological studies.[5–7] Importantly, to gain insights into ecological processes, it is crucial that response and explanatory variables are modeled in a biologically sensible manner.[8,9] Indeed, a regression model in itself is not a magical problem-solver: it is

Figure 1. Nifflers are relatively harmless magical animals native to Britain, they resemble a platypus and are attracted to shiny things. Despite they can destroy houses, they are possibly the cutest fantastic beasts. The figures are screenshots from the 2016 movie "*Fantastic beasts and where to find them"* (the adult niffler on the left) and from the 2018 movie "*Fantastic beasts: The crimes of Grindelwald"* (the baby niffler on the right). Screenshots are © & ™ Warner Bros. Entertainment Inc. J.K. ROWLING'S WIZARDING WORLD ™ J.K. Rowling and Warner Bros. Entertainment Inc. Publishing Rights © JKR. (s18), and their use in this article is intended under the Fair Use guidelines.

simply a tool, a "golem" whose "*abundance of power is matched by its lack of wisdom*."[10] Consequently, regression models are just as good as the instructions used to build them, which, despite their apparent simplicity, can be complex and subtle.[2] This, in turn brings up a crucial question: to what extent do we need to master the technicalities of regression modeling to properly analyze our data?

It has been suggested that a solid understanding of calculus, algebra and mathematical statistics is needed to step from "literacy" to "fluency" in statistical modeling (including—but not limited to—regression analysis).[11] This is indubitably true: arguably, however, even moderate mathematical skills may suffice to achieve, if not "fluency," at least the ability to "reason" and "think" statistically (*sensu*[12]), which should allow us to go a long way in regression modeling. As noted by Cohen et al[13] '*to drive a car, one does not need to be a physicist, nor an automotive engineer, nor even a highly skilled auto mechanic, although some of the latter's skills are useful when one is stuck on the highway*.' In order to do that, however, we must know our car sufficiently well: in other words, we may possibly do without a detailed understanding of the model estimation "engine," but we must surely know what the model components are, how inference works, recognize if models are "well-behaved" and understand their outputs.[13,14] This work is a modest contribution toward this aim: I shall present the basic steps of classic ("frequentist") regression modeling in a tutorial style, using just a few chunks of R codes, to help understand the data, select the most appropriate models, verify assumptions and provide useful advices for the interpretation of results. The exposition will broadly follow the protocol for regression-type analysis outlined in Zuur and Ieno.[15] I assume the reader is a student in ecology, with some rudimentary knowledge of R programming and a basic understanding of linear modeling.

Let me end this introduction on a note of caution: Sir Ronald Fisher, one of the fathers of modern statistics, once said that "*to consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.*" This exercise is indeed a 'post mortem examination': as we will see in the next paragraph, the original data (beside being fictional) were not collected with regression analysis in mind. This is *not* what should be done: indeed, decent knowledge of statistical tools is a fundamental prerequisite for proper data collection, and I wholeheartedly recommend the excellent introduction of Sutherland[16] on the matter. Furthermore, I am simply a practitioner, and any competent statistician will likely find my description overly simplistic. For example, I do not touch on issues such as model selection, multilevel modeling or nonlinear regression, which are often encountered in ecological studies, and only hint at alternative estimation methods such as Bayesian analysis. In my defense, this is merely intended as a fairly simple and entertaining technical exercise. Priority was thus given to basic understanding of applied regression over research protocol coherence and analytical complexity, whose importance I do not belittle.

Data Collection

Our predictions can be investigated by assessing the relationship of the country-specific temperature with: (1) the native presence of fantastic beasts and (2) the mean perceived level of danger of the native communities of beasts. To investigate these relationships, we need data on the native occurrence of beasts in each country, on the mean perceived level of danger within each native country and on the temperature of each country.

With the first prediction, we seek to investigate the relationship between temperature and animal distribution. This is essentially a problem of habitat selection, "*the set of rules individuals use to choose among patches that differ in some way*,"[17], that is, rules that ultimately determine the spatial distribution of individuals. Several methods exist to investigate this process. One of the simplest approaches is based on the estimation of

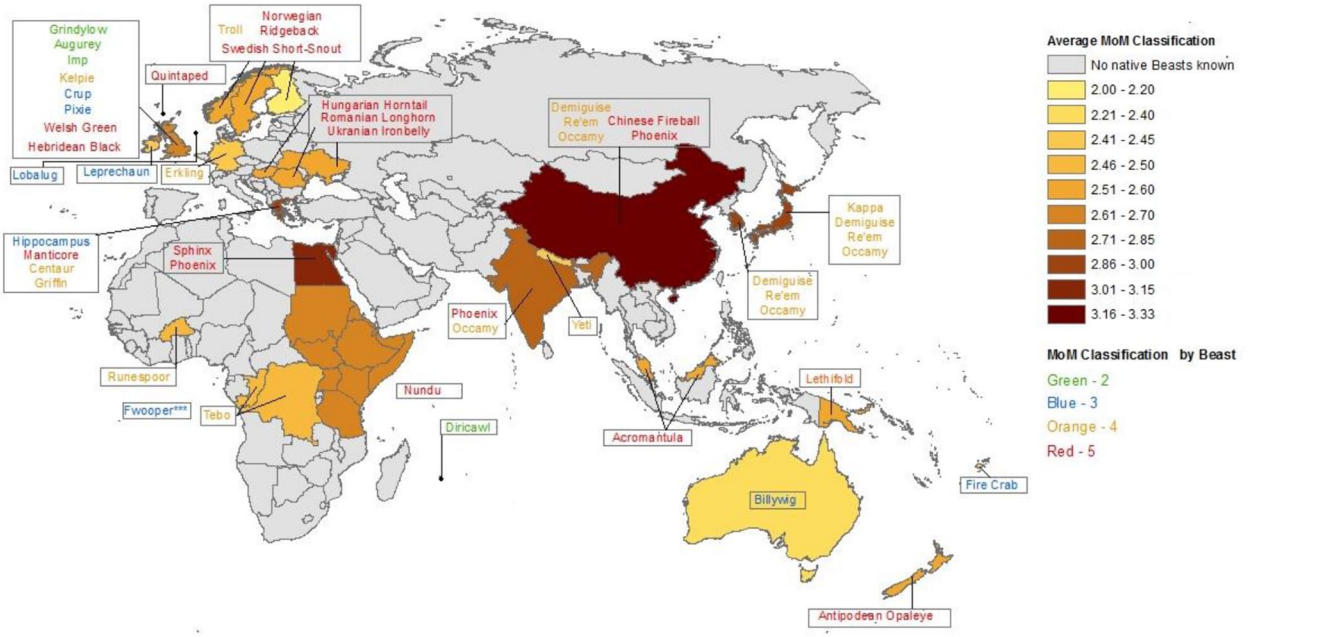


Figure 2. Many fantastic beasts are currently widespread worldwide. This map, modified from www.harrypottercartography.wordpress.com, shows their native distribution. The map also reports the mean perceived level of danger of country-specific magical animal communities, based on the classification of the Ministry of Magic (MoM).

the probability that a sample unit is used by animals, as a function of some features of that unit, for example, elevation, climate or human-disturbance. More generally, this is a problem of "resource selection," as long as the selected units are seen as resources, and the explanatory variables associated with these resource units as covariates of the resources.[18] To this aim, a survey area can be divided into discrete sampling units of the same size, selected randomly or systematically, and the use or non-use of each unit is assessed through appropriate search strategies.[19] This method assumes that presence (use) and absence (non-use) of animals within each sample unit are assessed without error. What do we know about the fantastic beasts? Most information available in the book was collected through observations made over years of intensive traveling.[1] Unfortunately, we do not have information on the occurrence of animals in randomly or systematically distributed sampling units of the same size, nor exact details on animal locations. All we have is information about native presence or absence of any one beast within countries. Assuming that Newt Scamander searched systematically throughout the continents, namely that the sampling effort was evenly distributed across countries, the country itself could be considered as the basic sampling unit. We could thus assign a "0" to countries with no native beasts, and a "1" to countries with at least a native beast. Native presence/absence of fantastic beasts was retrieved from the map in Figure 2 (sample size: n = 159). With the second prediction, we seek to investigate the relationship between temperature and geographical variation of a phenotypic trait, ie, the mean perceived level of danger of native communities of beasts. The map in Figure 2 provides information on the mean perceived level of danger within each native country, varying from 1 to 5, based on the classification of the Ministry of Magic (in n = 36 sampling units). For both predictions, the size of the sampling units must be accounted for: quite intuitively, the larger the country, the higher may be the probability of having magical beasts, and possibly very dangerous ones. Therefore, in order to separate the effect of temperature, we also need data on country size to control for this latter variable.

Wikipedia was used to collect information about each country shown in Figure 2 in terms of temperature (for the sake of simplicity, I used the average yearly temperature: https://en.wikipedia.org/wiki/List_of_countries_by_average_yearly_temperature) and area (https://simple.wikipedia.org/wiki/List_of_countries_by_area). Some data maneuvering was nonetheless necessary. For example, the Island of Drear, home of one of the most dangerous fantastic beasts, the Quintapeds, was not reported by Wikipedia. According to www.hp-lexicon.org, "*the tiny island of Drear is located off the northernmost point of Scotland. Its sole inhabitants are the fierce Quintapeds. The map, of course, shows only the approximate location of Drear, since it is unplottable.*" For Drear, I therefore assumed an island the size of Liechtenstein, with the same average temperature as the United Kingdom. The island might surely have different size or climate, but since it's unplottable, nobody can prove me wrong.

I finally created a data frame with 5 columns (see Supplementary file 1): column 1 reports the name of each country showed in Figure 2 ("Country," categorical variable); column 2, the native presence or absence of beasts within each country ('Native,' binary variable); column 3, the mean perceived level of danger of country-specific communities ('Level_of_danger,'

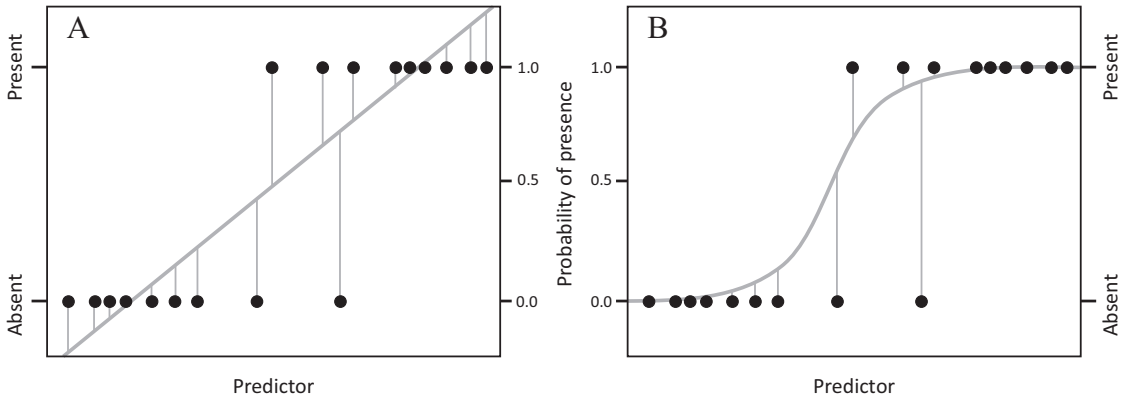


Figure 3. Panel (A) shows how a simple linear model would be inappropriate for modeling the probability of presence of fantastic beasts. The probability would be allowed to go below 0 and above 1, the conditional response distribution would be non-normal and with unequal variance, and the relationship between unscaled response and predictor would be linear. This type of data can be handled correctly by a GLM in the form of a logistic regression (panel (B)), which allows to get probability values between 0 and 1, to model non-linearity, and to handle a conditional distribution with non-normal and unequal variance.

continuous numeric variable from 1 to 5); column 4, the average yearly temperature of each country ('Temperature,' continuous numeric variable, in °C); column 5, the area of each country ('Country_size,' continuous numeric variable, in km^2). All analyses have been conducted with R v. 4.0.4[20] in RStudio v. 1.3.1056[21] for MacOS v. 10.15.6.

Regression Modeling

Native presence of fantastic beasts

Model building. Three elements must be defined when building a (generalized) linear regression model, or (G)LM: the linear predictor, the conditional distribution of the response variable, and the link function.[22] The linear predictor is the systematic part of the model, an additive or interactive combination of products of regression coefficients and explanatory variables; only coefficients must be linear, while continuous variables can take different forms (eg, log, polynomial). The conditional distribution is the random part of the model, and refers to the distribution of the response variable across the regression line, ie, the distribution conditional on a set of explanatory variables; each distribution implies a given variance (for a bestiary of probability distributions, see[23]). The link function, as the name suggests, connects the linear predictor with the mean of the conditional distribution: it is a transformation of the expected response that linearizes the relationship between the fitted value and the predictor. It does so by removing restrictions on the range of the expected response (eg, $\geqslant 0$ or 0-1 in GLMs), and mapping the interval to an unbounded continuous scale (ie, $\pm\infty$), so that the predictions (ie, the fitted values on the scale of the response variable) can be obtained by applying the inverse of the link function. When defining the linear predictor, the response variable should be inspected for the form of its relationship with the explanatory variables (eg, linear, quadratic. . .). When defining the conditional distribution (eg, Gaussian in simple LMs, or binomial, gamma, Poisson in GLMs), it doesn't generally make much sense to look at the distribution of the raw response variable, which, *per se*, might look distinctly non-normal, non-gamma or non-Poisson: rather, the raw response variable should be inspected for its main characteristics, for example, whether it is continuous or discrete, bounded or unbounded.[24] Once the conditional distribution is defined, the choice of the link function (eg, identity, logit, log) is often canonical,[23] though it can be changed based on the desired scale for the response.

In our model, native presence/absence of beasts is assumed to be the response variable, while the linear predictor includes average yearly temperature of the country, plus country size, which is the covariate that accounts for the uneven area of sample units. Presence/absence is a Bernoulli variable expressed as 1/0, thus simple linear models would be unsuitable to fit the data: a Gaussian conditional distribution for the response variable would allow infinite negative or positive fitted values, and it would require symmetric and constant variance across the fitted line; furthermore, the canonical identity link function in LMs does not transform the expected response, thus implying a linear relationship between the mean in the original scale and the predictor (Figure 3A). Instead, binary data are either 0 or 1, their conditional distribution is non-normal and with non-constant variance (the "freedom" of variation of the response variable from the fitted line is asymmetrically reduced toward the extremes), and the original expected response is non-linearly related to the predictor (Figure 3B). We thus need a model with a different conditional distribution and link function to get the likelihood correct, in order to avoid misleading estimates. Logistic regression, a form of generalized linear model, is the natural approach to binary data[18]: it prevents probabilities to go below 0 and above 1, it allows to model asymmetrical and unequal variance, and it captures the non-linearity of the relationship between the unscaled response and the predictor. More generally, a GLM essentially allows to extend simple linear models by capturing a particular kind of variation in the conditional distribution of the response variable, and a

particular kind of non-linearity of the relationship between the predictor and the expected response.[23] We may say that GLMs are "link-linear," so that a "*GLM may be thought of as a linear model for a transformation of the expected response or as a nonlinear regression model for the response*."[25] Excellent, accessible introductions to GLMs are provided by Dunteman and Ho,[26] Buckley,[22] and Pekár and Brabec.[24]

We start by loading the dataset:

```
beasts <- read.csv(file.choose()) # choose the .csv file interactively
```

Before writing the R code for our logistic regression, for the sake of clarity, it is good practice to present the basic structure of the model[15]

$$E\left(Native_i\right) = \pi_i \; Var\left(Native_i\right) = \pi_i \cdot \left(1 - \pi_i\right)$$
$$g\left(\pi_i\right) = \eta_i \text{ and } \eta_i = \beta_0 + \beta_1 \cdot Temperature_i + \beta_2 \cdot log2\left(Country\ size_i\right)$$

In our logistic regression model, the conditional distribution has mean $E(Native_i)$ and variance $Var(Native_i)$. The function $g(\pi_i)$ represents the canonical logit-link function: as π_i is confined between 0 and 1, the logit-link allows to map the interval to the real line, that is, between $\pm\infty$. In the linear predictor η_i, β_i are the regression coefficients to be estimated from the data, while temperature and country size are the explanatory variables. Country size was log-transformed because data were sparse, namely, the gaps between datapoints were large, likely owing to the presence of very small and very large countries: you can easily check that by calling: `plot(Native ~ Country_size, data=beasts)`. A log-transformation helps improving the fit of the model: $\log_2$ also allows an easy interpretation of the results, as the value of country size can be obtained by raising 2 to the power of the log-transformed value, so that increasing of 1 unit the log-transformed value corresponds to doubling the size of the country. Since temperature and $\log_2$-country size have very similar scales, no further transformation is required; if explanatory variables were from widely different scales, or interactions were present, standardizing (ie, centering each continuous explanatory variable to its mean and then dividing by 1 standard deviation) would be an option to make different regression coefficients comparable.[4]

Both temperature and country size can be included in the analysis, as no issue of collinearity was detected in preliminary data exploration between these variables. The scatterplot between temperature and country size, `plot(log2(beasts$Country_size), beasts$Temperature)`, does not suggest any relationship between these variables (Figure 4) and the function `cor(log2(beasts$Country_size), beasts$Temperature)` returned a Pearson correlation value of -0.07, well below the commonly accepted threshold of|0.7.|[27] Collinearity can be also checked by applying the "vif" (Variance Inflation Factor) function in the "car" package[28] to the model fitted in the next section. VIF values < 3 are considered inconsequential.[29]

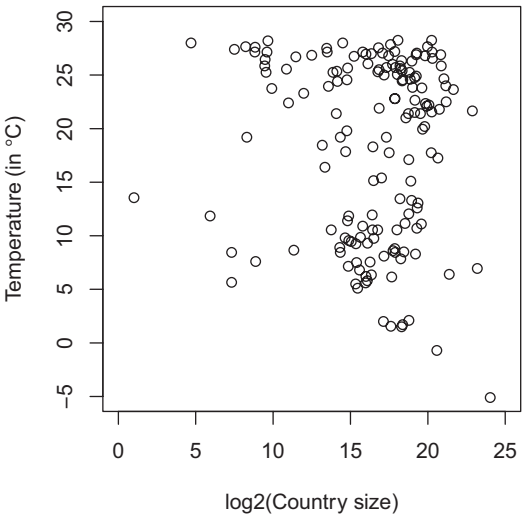


Figure 4. A scatterplot can be used to visually investigate the relationship between numerical explanatory variables: in this case the relationship between country size and temperature does not show any particular pattern.

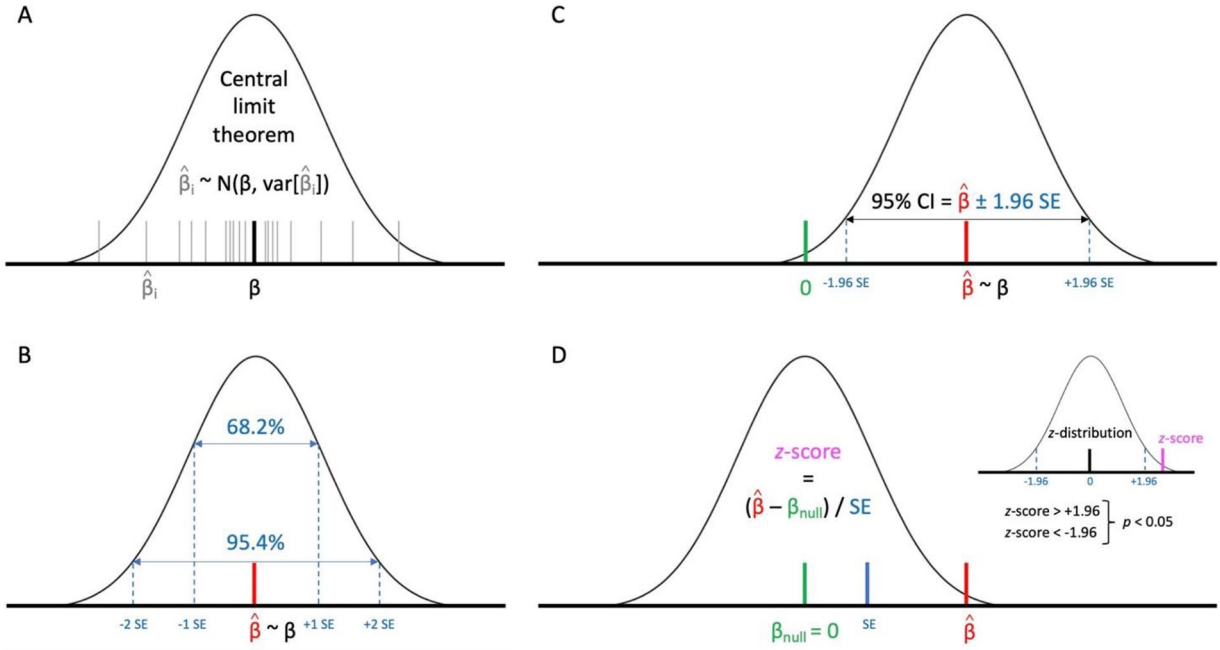
Model fitting and residual diagnostics. The model structure reported above can be coded in R as:

```
glm_native <- glm(Native ~ Temperature + log2(Country_size),      # set explanatory variable and covariate
                  family = binomial(link = "logit"),              # set conditional distribution and link function
                  data = beasts)                                  # select dataset
```

Before inspecting the results, it is essential to make sure the model complies to the underlying assumptions, so that it is possible to make proper inference (*cf.* Box 1). Visual residual diagnostics is commonly employed to this aim.[30] Basically, the idea is this: suppose we have a sample that is representative of the entire population of interest; suppose also that we knew the "true" data generating process, that is, the "true" regression between response and explanatory variables in the whole

Box 1. A bestiary of classic inference for regression models.

Statistical inference is the attempt to draw conclusions about some unknown aspect of a population, based on a sample from that population. Through ordinary least squares (LMs) or maximum likelihood (LMs and GLMs) we estimate $\hat{\beta}$, a sample-based approximation (our 'best guess') of the 'true'—albeit unknown—regression coefficient β that links an explanatory variable to the response in the population. Typically, we are interested to know if this relationship is different from zero ($\beta \neq 0$). To this end, however, $\hat{\beta}$ alone is insufficient: intuitively, using only one sample makes our estimate of β uncertain, as the same model applied to another sample is likely to yield different estimates! The **standard error** (*SE*) returned by the model accounts for this uncertainty due to sampling variability, and allows to make the leap from the known ($\hat{\beta}$) to the unknown (β). How?



Classic (or 'frequentist') inference theoretically assumes that random samples of a given size are drawn repeatedly from the same population, and regression coefficients are estimated for each one of them. The realized coefficients $\hat{\beta}$ will differ because of sampling variability, but the central limit theorem ensures they will distribute themselves normally, for large enough sample sizes (panel A). A **sampling distribution** of regression coefficients is thus defined by a mean β, and by a standard deviation that quantifies the uncertainty in the estimation of β. The mean of the sampling distribution is approximated by $\hat{\beta}$, the 'signal', and the standard deviation is approximated by *SE*, the 'noise' around $\hat{\beta}$, both estimated by our regression model. We also know the % of the area under this curve that lies between ± a given number of standard errors from the mean: for example, 68.2% between ± 1 *SE*, 95.4% between ± 2 *SE* (panel B). This information is particularly relevant to inference, because it allows to account for sampling variability when testing if $\beta \neq 0$.

For example, we can define a **confidence interval** (CI) for β, using the conventional 95% threshold (α-**level** = 0.05). Assuming the model estimator is normally distributed (thus reflecting a theoretical normal sampling distribution) with known variance, the 95% CI bounds around the realized $\hat{\beta}$ are obtained by multiplying the estimated value of *SE* by ± the number of standard errors needed to include 95% of the area under a normal distribution, which corresponds to a *z*-value of 1.96 (panel C). When the variance is unknown (as it is often the case) or sample size is small, the model estimator would follow a slightly different distribution, that is, a *t*-distribution: in this case, the number of standard errors used to define the 95% CI will depend on sample size, but the *t*-value will converge to 1.96 for large samples. The CI can thus be generally defined as $\hat{\beta} \pm z(\text{or } t)_{a/2} \times \text{SE}$, an interval that, if the study is theoretically repeated many times, 100(1 − α)% of the times will include β.

Alternatively, we can set up a **null hypothesis** of no relationship between variables, that is, a sampling distribution of regression coefficients with mean zero ($\beta_{null} = 0$) and standard deviation *SE*. Next, we evaluate how far the realized regression coefficient $\hat{\beta}$ is from β_{null}. To this aim, a **signal-to-noise** ratio ($[\hat{\beta} - \beta_{null}]/SE$) is calculated to obtain a **standardized *z*- or *t*-score**, depending on the distribution of the model estimator (see above). This score, which reflects the distance (measured in standard errors) between $\hat{\beta}$ and zero, will follow a standard normal (*z*-) distribution or a standard *t*-distribution (panel D). The tabulated values of the corresponding distribution allow to associate the realized *z*- or *t*-score with a ***p*-value** for a given α-level, usually 0.05. The *p*-value can thus be defined as the probability of obtaining the *z*- or *t*-score (or more extreme values), if the study is theoretically repeated many times and the null hypothesis is true.

CIs and *p*-values both allow to assess if β is significantly different from zero at a given α-level, while accounting for sampling variability. Namely, for α-level = 0.05, a regression coefficient will be **statistically significant** if its 95% CI excludes zero or the *p*-value < 0.05 (panels C-D). CIs, however, convey more information than plain null hypothesis testing: rather than focusing on the dichotomic alternative 'significant *vs* nonsignificant', CIs focus on effect size, and they tend to be preferred in modern statistics.[34]

population. By definition, this "true" regression will capture the information in the data correctly, that is, it will account for all sources of variation, so that what is left unexplained is just random noise. This random noise is reflected by the "errors," that is, the vertical distances between each datapoint in the sample and the "true" regression line. However, we don't actually know what the "true" regression is, as we hardly have the possibility to investigate the entire population, or include all potential explanatory variables; errors are thus unobservable. We can only approximate the data generating process in a sample by using a model.[31] The question is, can we actually trust this model? To this aim, we can inspect the distances between each of the response datapoint in the sample and the regression line fitted with the model, that is, the "residuals." Though residuals have different properties than errors,[32] if the fitted model complies to its underlying assumptions, the residuals would approximate

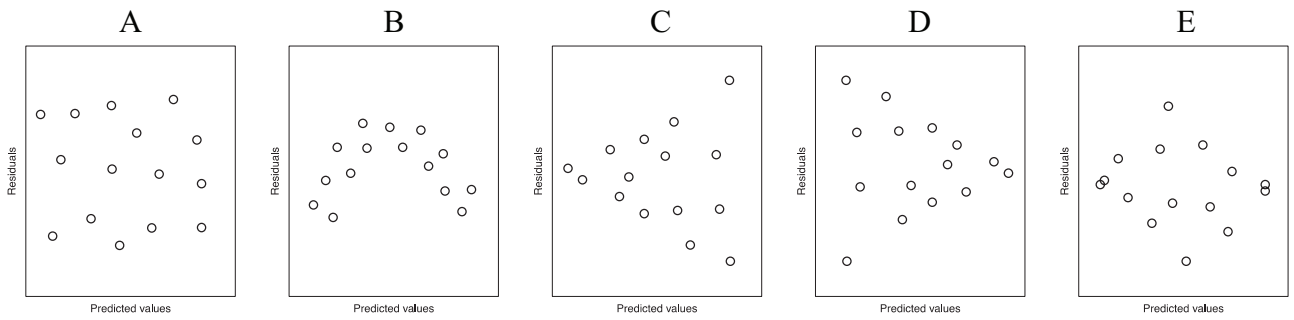


Figure 5. Scatterplots between residuals and predicted values can be used to visually inspect how well a model fits the data. (A) homoscedastic (unsystematic) patterns suggest a good fit, (B) humped (non-linear) patterns suggest that the systematic part of the model might have been misspecified: interactions or non-linear relationships may be considered, (C and D) funnel-shaped or (E) double-bowed scatters indicate that the variance changes with the predicted values (eg, linearly or as a proportion between 0 and 1), hence alternative conditional distributions may be considered.

the errors, that is, they would deviate unsystematically from the fitted regression line. It is important to stress that, because the true data generating process is difficult to approximate, residuals are best seen *not* as tools to assess if the model is exactly correct, but rather as tools to assess if the model is not grossly wrong.[33] Indeed, generic checks such as residuals *vs* fitted values do not ensure that a given model is indeed what we are looking for.[24] For example, we may be missing an important explanatory variable, yet have a decent residual distribution. Despite these limitations, residual plots remain fundamental tools to make sure there are no major violations of model assumptions.

For simple linear models, the main assumptions can be intuitively summarized with the acronym **L.I.N.E.**: the relationship between mean response and predictor must be **L**inear; the responses are **I**ndependent of each other; the conditional distribution of the response variable is **N**ormal, and it shows **E**quality of variance. For GLMs, some of these assumptions can be relaxed: linearity is not required, and the curvature of the relationship between unscaled mean response and predictor can be modeled using the correct link function; the responses are still assumed independent of each other; the conditional distribution of the response variable does not need to be symmetrical (normal) nor equal, and it can be modeled using the correct distribution and variance.[35] Either way, (G)LMs should include all important variables in the predictor and have no multicollinearity issues. When LM assumptions are met, the plot of (standardized) residuals *vs* predicted values would reflect random noise and exhibit no pattern, that is, the residuals would be scattered in an unsystematic way. Severe deviations from near-randomness (eg, non-linear patterns, or "funnel-shaped" variance) would indicate that the model has not captured correctly the information in the data, that is, there are problems with one or more assumptions (eg, important variables are missing, or their form is incorrectly specified; the assumed residual variance is wrong. . .) (Figure 5).

Ideally, the behavior of residuals for GLMs should be similar to the behavior of residuals for LMs. Intuitively, however, in GLMs the plot of raw response residuals *vs* fitted values would show some asymmetry and inequality of variance (*cf.* Figure 3B), thus they would be inadequate for model diagnostics.[35] To handle this issue in logistic regression, a different form of residuals may be used, for example, binned residuals. Arguably, however, quantile residuals are the most appropriate choice for GLM diagnostics.[36] Essentially, the idea is to determine the cumulative probability that an observation is less than or equal to the fitted value of a given distribution; this cumulative probability is then used to find the corresponding value of the standard normal variate, that is, the quantile residual.[35] Technicalities aside, for practitioners it is important to know that quantile residuals have the desirable property of being normally distributed, thus making inspection of GLM residuals as intuitive as residuals for LMs. Quantile residuals can be generated with the package "DHARMa"[37]:

```
library("DHARMa")                                  # load package
sim_glm_native <- simulateResiduals(glm_native)    # generate scaled residuals by simulating from the model
plot(sim_glm_native)                               # plot simulated residuals
```

Figure 6 shows that: the outlier test and the dispersion test on simulated residuals are non-significant; the Kolmogorov-Smirnov test is non-significant, suggesting that the conditional distribution of the response variable conforms to the expectations. The plot of residuals against predicted values does not show any particular pattern, also suggesting that the assumptions of the model have been met. All in all, the model seems to behave well! It is also generally recommendable to inspect the residuals against each explanatory variable, to check if the relationships with the response have been modeled appropriately (results are not shown, but the residuals do not reveal any systematic pattern):

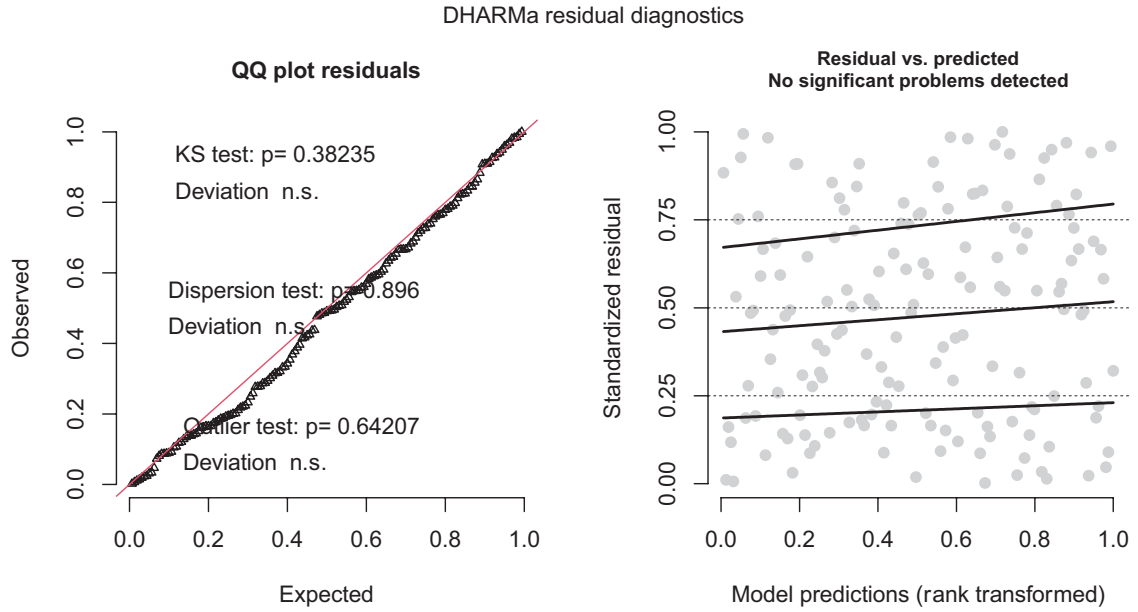


Figure 6. Residual diagnostics for the model fitted to investigate the probability of native presence of fantastic beasts. On the left, "DHARMa" returns a qq-plot to detect overall deviations from the expected distribution, and several goodness of fit tests; on the right, it returns a plot of the simulated residuals against the predicted values. In the panel on the left, QQ stands for "quantile-quantile", and KS for "Kolmogorov-Smirnov".

```
plotResiduals(sim_glm_native, form = beasts$Temperature, xlab = "Temperature")
plotResiduals(sim_glm_native, form = log2(beasts$Country_size), xlab = "log2(Country size)")
```

Importantly, (G)LMs in their basic forms assume that random errors are independent of each other. Lack of independence in the random errors generally arises because of the multilevel nature of the response variable (which can be modeled with the inclusion of a "grouping" or "random" term,[7]), or because of temporal or spatial correlation. In this example observations were assumed to be independent as we have neither grouping nor temporal effect. Some spatial correlation might be present: conditional autoregressive (CAR) models can cope with correlation between neighboring countries (we have no details about the exact location of each fantastic beast), but they are beyond the scope of this tutorial.

Model results and interpretation. Once we accept that our model is well-behaved, we can inspect the estimates with the "parameters" package,[38] which conveniently allows to display essential information:

```
library("parameters")
parameters(glm_native) # show estimate, SE, 95% confidence interval, z-score, p-value
```

The main results, in terms of $\hat{\beta}$, standard errors (*SE*), confidence intervals, *z*-scores and *p*-values are reported in Table 1. This information is used to make inference: if you need a refresher on these concepts, Box 1 should (hopefully) help!

Table 1 shows that the native presence of fantastic beasts is related to country size, but not to temperature. More specifically, the values in Table 1 tell us that, with a temperature of 0 °C and a $\log_2$-country size of zero (the Intercept), the probability of native presence of fantastic beasts would be 0.015 (since we used a logit-link, this value can be obtained by back-transforming the intercept value [-4.16] using the **`inv.logit()`** function, which is simply defined as $e^{(x)}/(1+e^{(x)})$, in the package "boot"[39]). Admittedly, assuming a country size of 1 km^2 (so that the $\log_2$-value is zero) might make poor biological sense, and it may be more appropriate to center the covariate "country size" to its mean, prior to fitting the model, so that the intercept would return the probability of native presence of beasts for a mean $\log_2$-country size (notably, the regression coefficients would not change, try it!). The estimated $\hat{\beta}$ coefficient represents the expected change in the response variable for each unit change in the explanatory variable, while holding the other variables in the model constant at a given level. Unlike in simple linear models,

Table 1. Estimated relationships of temperature and $\log_2$-country size with native presence of fantastic beasts.

COEFFICIENT	$\hat{\beta}$	SE	95% CI	*Z*-SCORE	DF	*P*-VALUE
Native presence						
(Intercept)	−4.16	1.30	(−6.70, −1.61)	−3.20	156	<.001
Temperature	−0.02	0.02	(−0.07, 0.02)	−0.88	156	.381
$\log_2$(Country_size)	**0.19**	**0.07**	**(0.06, 0.33)**	**2.77**	**156**	**.006**

Abbreviations: CI, confidence intervals; SE, standard errors.
The table reports coefficient estimates ($\hat{\beta}$), standard errors (*SE*), 95% confidence intervals (CI), *z*-scores, degrees of freedom (df) and *P*-values. Significant coefficients in bold. The interpretation of these values is detailed in the main text.

where the response is not transformed, in logistic regression the fitted values are in the logit scale, and $\hat{\beta}$ cannot be straightforwardly interpreted. Rather, $\hat{\beta}$ can be exponentiated to obtain the odds-ratio (OR), a metric often used to interpret logistic regression coefficients (if you are not familiar with that, check out Box 2!).

When increasing temperature by 1 °C, we obtain an OR of $e^{(-0.02)} = 0.980$, which can be interpreted as a non-significant change (ie, with low "signal-to-noise ratio," *cf.* Box 1) of -2% in the odds of native presence of beasts. Similarly, when increasing the $\log_2$-size of the country by 1 unit (ie, when doubling country size), we obtain an OR of $e^{(0.19)} = 1.209$, which can be interpreted as a significant change (ie, with high "signal-to-noise ratio") of + 21% in the odds of native presence of beasts. Alternatively, we could look at regression coefficients by using the "divide by 4" rule presented in Box 2. For example, when increasing temperature by 1 °C, we obtain a variation of -0.02 / 4 = -0.005 in the expected value π_i, which can be interpreted as a non-significant decrease of approximately 0.5% in the probability of native presence of beasts. This rule, however, would not work well when increasing the $\log_2$-size of the country by 1 unit (ie, when doubling country size): the expected values are very close to 0 and the relationship is highly non-linear (*cf.* Box 2 and Figure 7).

We can also extract the pseudo-R^2 of the model, which mimics the behavior of the coefficient of determination for ordinary least squares models (ie, the proportionof the variance for a dependent variable that is explainedby the independent variables), with the package "performance"[42]:

Box 2. Logit, odds-ratio and other beasts.

The logit-link function assumed in logistic (and Beta) regression models takes the form:

$logit(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$, which, for a hypothetical linear predictor $\eta_i = \beta_0 + \beta_1 \cdot x_i$ stands for

$$\pi_i = \frac{e^{\beta_0+\beta_1 \cdot x_i}}{1+e^{\beta_0+\beta_1 \cdot x_i}}$$

From these formulas, we see that the logit-transformed mean—not the mean—is modeled as a linear function of the predictor; the logit-link therefore acts as a transformation that linearizes the relationship between the expected response and the predictor, mapping the range of probability (0,1) to the real line.[25]

We also see that the logit function is the logarithm of the odds, $ln\left(\frac{\pi_i}{1-\pi_i}\right)$ or 'log-odds', that is, the log of the ratio between the probability that the event will occur and the probability that it will not occur.

Since log-odds are modeled as a linear function of the predictor, β coefficients represent the increase or decrease in the log-odds of the response for one-unit change in the explanatory variable. Basically, the β of an explanatory variable is evaluated as a subtraction of log-odds, so that β = log-odds(x + 1)—log-odds(x), where log-odds(x + 1) refers to the log-odds when the value of the explanatory variable is just one unit larger than the value of the same variable associated with log-odds(x). This interpretation of β, however, is not very intuitive, and back-transforming the log values to odds is desirable. Because exponentiating a subtraction results in a division, by exponentiating the β coefficients, that is, the difference in log-odds, we obtain the 'odds-ratio' (OR), which can be interpreted as the change in the odds of the response variable with every one-unit increase (indicated with an asterisk) in the explanatory variable. Therefore, $OR = e^{\beta} = \frac{\frac{\pi_i^*}{1-\pi_i^*}}{\frac{\pi_i}{1-\pi_i}}$. In simpler words, the odds-ratio can be defined as '*the odds on something occurring in one situation to the odds of the same event occurring under a second situation. An odds-ratio of 1 implies that the odds on an event occurring (and hence the probability of its occurrence) are unaffected by the change in the situation*'.[40]

Admittedly, odds and odds-ratio sound a bit like obscure beasts. To better understand the relationships between response and explanatory variables is thus recommendable to inspect the mean predicted values graphically.[41] Alternatively, the 'divide by 4' rule may also be used. If the relationship between the response and the explanatory variable does not change much in most of the predictor's range, then this part of the logistic curve can be considered roughly linear. The slope, that is, the first derivative of π_i to the explanatory variable, is $\beta\pi_i(1-\pi_i)$. The slope of the curve is at its maximum when $\pi_i = \frac{1}{2}$, that is, when it equals $\frac{\beta}{4}$.[2] Consequently, if we divide the logit coefficients by 4, we obtain an estimate of the maximum change of the expected response for one-unit increase in the explanatory variable. This rule works well when the probabilities do not flatten, that is, when they do not attain extreme values.[25]

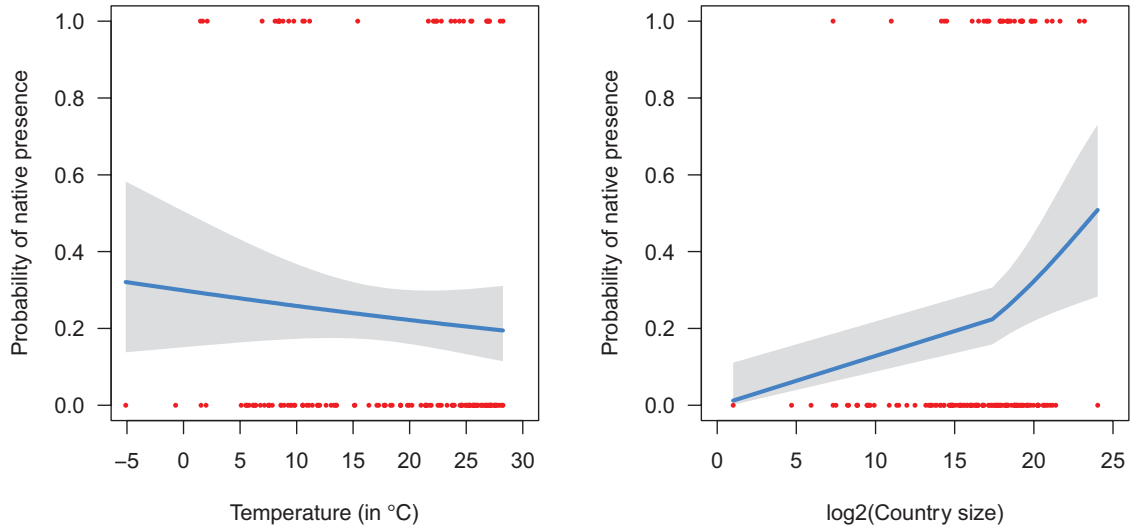


Figure 7. Marginal relationships of temperature and $\log_2$-country size with the probability of native presence of fantastic beasts. When visualizing a marginal relationship in "visreg," the other continuous explanatory variables in the model are set at their median value. Gray shaded areas show 95% confidence intervals. Original datapoints in red.

```
library("performance")
r2(glm_native)              # calculate pseudo-R2 for the fitted model
```

This function returns a fairly low value (about 7%) for the Tjur's coefficient of discrimination.[43] Although the residual plot did not give evidence of gross mis-specification, the model has limited predictive accuracy, that is, much of the variation in the response was not captured by the explanatory variables. We may further explore such accuracy by calculating the area under the receiver operating characteristic (ROC) curve.[44] Simply put, the ROC curve allows to assess the ability of the model to discriminate between presence and absence of beasts in any one country, based on the explanatory variables used. In general, a value of the area under the curve (AUC) of 0.5 suggests random discrimination ability (ie, the model is equally likely to predict either presence or absence), 0.7 to 0.8 is considered an acceptable discrimination ability, 0.8 to 0.9 is considered excellent, and >0.9 is considered outstanding. If the AUC is <0.5, the model does worse than chance in predicting presence/absence. The actual response values and the values predicted by the model can be used to calculate the AUC with the "pROC" package[45]:

```
library("pROC")
roc.glm_native <- roc(beasts$Native, predict(glm_native))  # calculate ROC curve
roc.glm_native$auc                                         # extract area under the ROC curve
plot(roc.glm_native)                                       # plot the ROC curve (not shown)
```

The AUC value is 0.70, suggesting that there is about 70% chance that the model will be able to discriminate between presence and absence of beasts in any one country, given the explanatory variables. This predictive accuracy largely owes to the significant explanatory variable $\log_2$-country size. Should we refit the model with only the non-significant explanatory variable (ie, temperature), we would obtain an AUC value of about 0.45, suggesting that flipping a coin may do better than the temperature-only model in predicting presence or absence of beasts in any one country!

With 36 events of native presence out of 159 datapoints, and only 2 explanatory variables in the model, our estimates should be OK. Rules of thumb suggest that a regression model is likely to be reliable when the number of explanatory variables is less than $m/15$, where m is the "limiting sample size" (which, for logistic regression, would correspond to the number of cases in the less frequent category[46]:). However, to account for potential bias due to small sample sizes, a solution may be to fit a model using the Firth's method,[47] which allows to reduce small-sample bias by placing a penalty term on the maximum likelihood function. Unlike traditional maximum likelihood, Firth's method always allows to generate finite estimates of regression coefficients and associated standard errors. Firth regression can be fitted with the "logistf" package[48]—the estimates (not shown) are consistent with those of the uncorrected logistic regression:

```
library("logistf")
glm_native.firth <- logistf(Native ~ Temperature + log2(Country_size),
                            firth = TRUE,
                            data = beasts)
parameters(glm_native.firth)
```

To further improve the interpretation of results, we may generate graphs of the model, which can be much more effective at imparting information than tables.[15] We can easily plot the marginal relationships with the package "visreg"[49] (Figure 7):

```
library("visreg")

par(mfrow=c(1,2))                                                          # set multi-panel plot

visreg(glm_native, "Temperature" scale = "response" rug = FALSE,           # select model and variable
       xlim=c(-5,30), ylim = c(0,1),                                       # set limits for x & y axes
       xlab = "Temperature (in °C)" ylab = "Probability of native presence")   # define labels
with(beasts, points(Temperature, Native, pch = 16, cex = 0.5, col = "red"))   # plot data points

visreg(glm_native, "Country_size" scale = "response" xtrans = log2, rug = FALSE,
       xlim=c(0,25), ylim = c(0,1),
       xlab = "log2(Country_size)" ylab = "Probability of native presence")
with(beasts, points(log2(Country_size), Native, pch = 16, cex = 0.5, col = "red"))
```

Figure 7 helps visualizing the results reported in Table 1, namely the slightly negative, albeit non-significant, relationship of temperature, and the significant positive relationship of $\log_2$-country size with the probability of native presence of beasts.

Mean perceived level of danger

Model building. Given that fantastic beasts are natively present in some of our sample units, what is the relationship between temperature and the mean perceived level of danger of the animal communities? We focus our attention on countries where beasts occur, because we can't measure a phenotypic trait where animals are not present. We thus need to subset our data:

```
beasts_danger <- subset(beasts, Native==1) # select native countries only
```

As we did for the logistic regression, to model variation in the response variable (mean level of danger) our linear predictor shall include temperature as an explanatory variable, and country size as a covariate to control for uneven area of sampling units. Our next task is choosing the appropriate conditional distribution for the response variable. Since the mean level of danger is a continuous variable, assuming a normal conditional distribution may be tempting (and indeed not wrong, but more on that later), but for didactic purposes let us reflect more carefully on the nature of these data. According to the classification of the Ministry of Magic, we cannot have beasts less than "boring" (<1), nor beasts more than "impossible to train or domesticate" (>5). Consequently, our response variable (mean level of danger) can take any value between 1 and 5. As seen before, in principle Gaussian linear models are not ideal to fit these data, as they allow infinite negative or positive fitted values, and assume that the conditional distribution of the response variable is symmetrically distributed with constant variance[4] (*cf.* Figure 3A). This is clearly not the case with the mean perceived level of danger, whose values are intrinsically constrained between 1 and 5. The conditional distribution of the response variable will be asymmetrical when the fitted values approach 1 or 5, and the variance will tend to be greater at intermediate expected values (intuitively, the "freedom" of variation of the response variable from the fitted line will reduce toward the extremes, *cf.* Figure 3B). We thus need a distribution that can accommodate these issues. An appropriate choice is the Beta distribution.[50] Importantly, the Beta distribution requires data in the open interval between 0 and 1 (that is, not including 0 and 1). We thus need some data manipulation before we fit our model. First, we temporarily scale the raw data y_i, which can vary between 1 and 5, into data between 0 and 1, by taking $y_i' = (y_i — y_{min}) / (y_{max} — y_{min})$:

```
beasts_danger$Level_of_danger.scaled.temp <- (beasts_danger$Level_of_danger-1)/(5-1)
```

(*cf.*[4]). To avoid undefined logits for the endpoints, we may add and subtract a small amount to, respectively, the 0- and the 1-values. Smithson and Verkuilen[51] suggested an alternative, more efficient procedure to avoid zeros and ones, where the scaled data are compressed by taking $y_i'' = [(y_i' \cdot (n - 1) + 0.5) / n]$, where n is the sample size:

```
beasts_danger$Level_of_danger.scaled <- (beasts_danger$Level_of_danger.scaled.temp*
                                          (length(beasts_danger$Level_of_danger.scaled.temp)-1)+0.5)/
                                           length(beasts_danger$Level_of_danger.scaled.temp)
```

This scaling method allows to effectively shrink the y_i'' interval to [.005, .995] (technical details are available in Smithson and Verkuilen[51]). We are now ready to fit our Beta regression model. We start again by presenting the structure of the model:

$$E\left(Level\ of\ danger_i\right) = \mu_i \text{ and}$$
$$Var\left(Level\ of\ danger_i\right) = \frac{\mu_i \cdot \left(1-\mu_i\right)}{1+\phi}$$
$$g\left(\mu_i\right) = \eta_i \text{ and } \eta_i = \beta_0 + \beta_1 \cdot Temperature_i + \beta_2 \cdot log2\left(Country\ size_i\right)$$

where the conditional Beta distribution has mean $E(Level\ of\ danger_i)$ and variance $Var(Level\ of\ danger_i)$. The parameter ϕ is known as the precision parameter since, for given values of μ_i, the larger ϕ the smaller the variance.[50] Intuitively, as in the logistic model, $g(\mu_i)$ represents the canonical logit-link function: as μ_i is confined between 0 and 1, the logit-link allows to map the interval to the real line, that is, between $\pm\infty$. Finally, β_i are the regression coefficients to be estimated from the data.

Model fitting and residual diagnostics. The simplest way to fit a Beta regression model in R is through the "betareg" package[52]:

```
library("betareg")
glm_danger.betareg <- betareg(Level_of_danger.scaled ~ Temperature + log2(Country_size),
                              link = "logit" link.phi = "log"
                              data = beasts_danger)
```

"betareg," however, is not supported by DHARMa, and although residual diagnostics is possible,[52] I prefer to refit the model using the package "glmmTMB."[53] "glmmTMB" uses maximum likelihood estimation via "TMB" (Template Model Builder), but yields (nearly exactly) the same results as "betareg," allows greater flexibility in fitting regression models and can be used in DHARMa.

```
library("glmmTMB")
glm_danger.glmmTMB <- glmmTMB(Level_of_danger.scaled ~ Temperature + log2(Country_size),
                              family = beta_family(link = "logit"),
                              data = beasts_danger)
```

As we did for the logistic regression, we check the fit of the model (Figure 8):

```
sim_glm_danger.glmmTMB <- simulateResiduals(glm_danger.glmmTMB)
plot(sim_glm_danger.glmmTMB)
```

Figure 8A suggests that the model is not a very good fit to the data. Why is that? The graph on the right suggests the presence of a severe outlier (indicated with an asterisk). Looking at the data, we see that the fierce Quintapeds are extremely dangerous (level 5) and they are the sole inhabitants of the tiny island of Drear. So, a tiny island with an extremely high mean level of perceived danger. . . What if we remove this datapoint? (Row number 33 in the dataset).

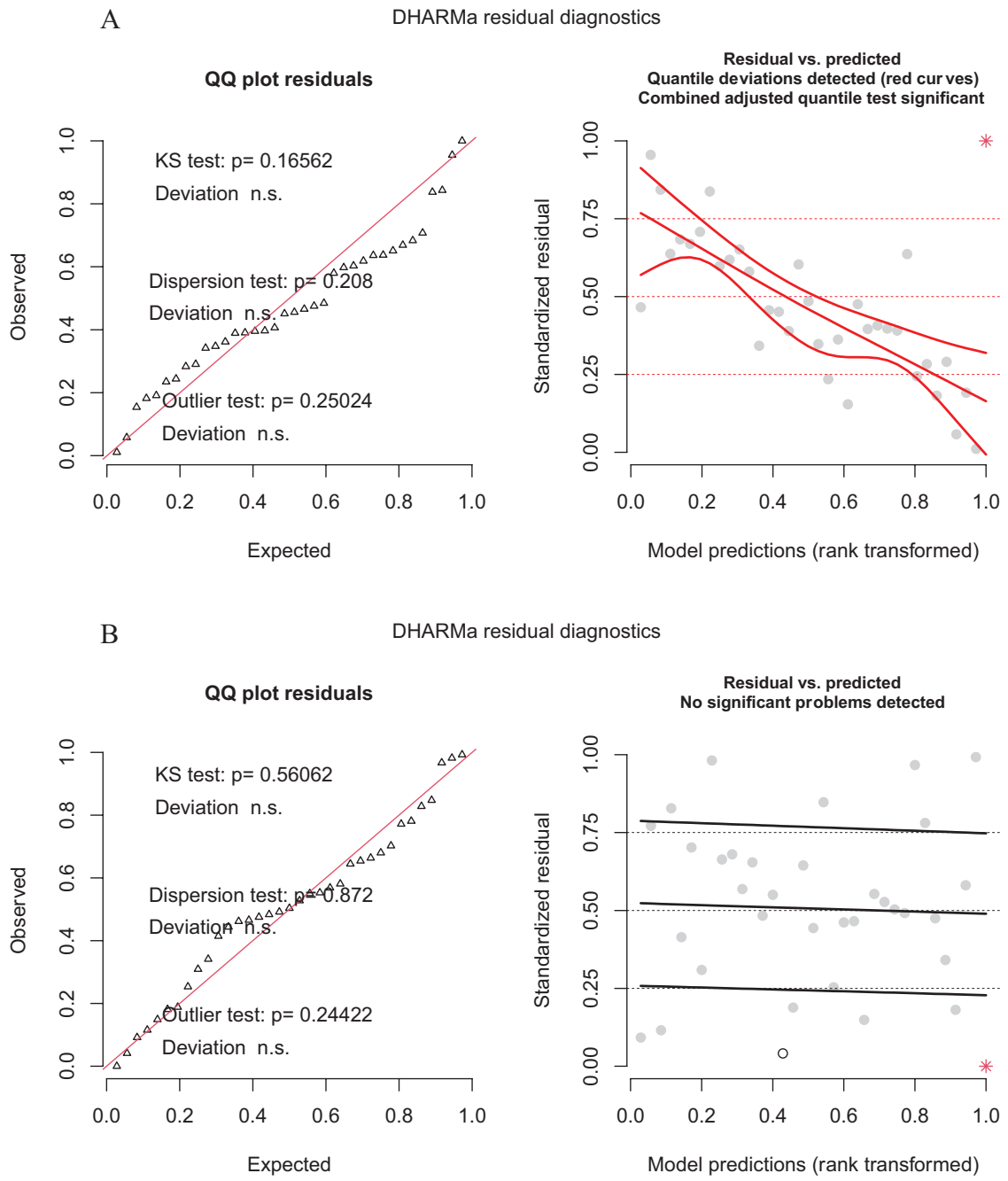


Figure 8. Residual diagnostics for the model fitted to investigate the variation in mean perceived level of danger of country-specific magical animal communities, with (A) and without (B) outlier.

```
glm_danger_no_outlier.glmmTMB <- glmmTMB(Level_of_danger.scaled ~ Temperature + log2(Country_size),
                                                                  family = beta_family(link = "logit"),
                                                                  data = beasts_danger[-c(33),])
sim_glm_danger_no_outlier.glmmTMB <- simulateResiduals(glm_danger_no_outlier.glmmTMB)
plot(sim_glm_danger_no_outlier.glmmTMB)
```

The residuals in Figure 8B look good now!

As we did for the logistic regression, we can also inspect the residuals of the model without outlier against each explanatory variable, to check if the relationships with the response have been modeled appropriately (results are not shown, but the residuals do not reveal any systematic pattern):

Table 2. Estimated relationships of temperature and $\log_2$-country size with mean perceived level of danger of country-specific animal communities.

COEFFICIENT	$\hat{\beta}$	SE	95% CI	*Z*-SCORE	DF	*P*-VALUE
Level of danger (with outlier)						
(Intercept)	1.63	0.51	(0.64, 2.63)	3.21	32	<.001
Temperature	−0.01	0.01	(−0.03, 0.02)	−0.48	32	.632
Log_2(Country_size)	**−0.10**	**0.03**	**(−0.16, −0.04)**	**−3.51**	**32**	**<.001**
Level of danger (without outlier)						
(Intercept)	−1.29	0.30	(−1.88, −0.69)	−4.22	31	<.001
Temperature	0.00	0.00	(−0.01, 0.01)	0.43	31	.665
Log_2(Country_size)	**0.05**	**0.02**	**(0.02, 0.08)**	**2.94**	**31**	**.003**

Abbreviations: CI, confidence intervals; *SE*, standard errors.
The table reports coefficient estimates ($\hat{\beta}$), standard errors (SE), 95% confidence intervals (CI), *z*-scores, degrees of freedom (df) and *P*-values. Significant coefficients in bold.

```
plotResiduals(sim_glm_danger_no_outlier.glmmTMB, form = beasts_danger[-c(33),]$Temperature,
              xlab = "Temperature")

plotResiduals(sim_glm_danger_no_outlier.glmmTMB, form = log2(beasts_danger[-c(33),]$Country_size),
              xlab = "log2(Country size)")
```

Model results and interpretation. Although the model with outlier is not a good fit to the data, for didactic purposes we inspect the results of both models (Table 2):

```
parameters(glm_danger.glmmTMB)
parameters(glm_danger_no_outlier.glmmTMB)
```

The interpretation of results follows the same logic explained above for the logistic regression, as binomial and Beta regressions share the same logit-link function. Therefore, $\hat{\beta}$ coefficients in Beta regressions can be interpreted in terms similar to odds-ratio: for one-unit increase in the explanatory variable, $\hat{\beta}$ would measure a difference in log-ratios of scaled level of danger. With a temperature of 0 °C and a $\log_2$-country size of zero (the Intercept), the inverse-logit danger level of countries would be 0.84 for the model with outlier, and 0.21 for the model without outlier. When increasing yearly average temperature by 1 °C, we obtain $e^{(-0.01)} = 0.990$ for the model with outlier, and $e^{(0.00)} = 1.000$ for the model without outlier, thereby suggesting non-significant changes (respectively, -1% and 0%) in the odds of scaled danger level. When doubling the size of the country, we obtain $e^{(-0.10)} = 0.905$ for the model with outlier, and $e^{(0.05)} = 1.051$ for the model without outlier, thereby suggesting significant opposite changes (respectively, -9.5% and + 5.1%) in the odds of scaled danger level. These results might be difficult to interpret, and the "divide by 4" rule may be a quicker and clearer alternative, at least for the model without outlier, for which the expected values are roughly linear and quite far from the extremes (*cf.* Figure 10). When increasing yearly average temperature by 1 °C, the expected value μ_i varies approximately by 0.002 / 4 = 0.0005, thereby suggesting an increase of some 0.05% in the scaled level of danger. When doubling the size of the country we obtain a variation of 0.05 / 4 = 0.0125, which suggests an increase of approximately 1.25% in the scaled level of danger.

A note of caution: the *z*-test returned by the model requires a normally distributed population with known variance, or large sample size (*cf.* Box 1). With moderate to small samples, the test might produce inaccurate *p*-values and confidence intervals, possibly misleading inference. Based on the *m*/15 rule of thumb for a continuous response variable,[46] the sample size for the level of danger should suffice for our simple model, but we would feel more confident if consistent results were obtained using a different (non-asymptotic) inferential approach. Although outside of the scopes of this paper, Bayesian modeling is an alternative to the classic approach outlined in Box 1.[2] Bayesian regression models allow to—perhaps more intuitively—generate sampling distributions for $\hat{\beta}$. How does this work? Just like for the frequentist model, we start by specifying a linear predictor; we then assign a particular kind of prior distribution to the regression coefficients, for example, a normal distribution

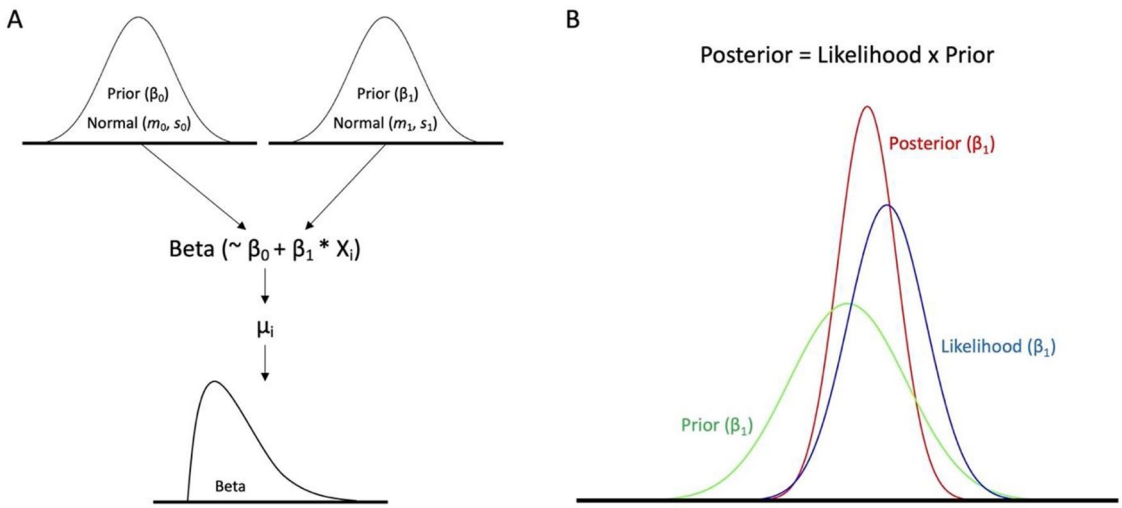


Figure 9. Bayesian regression assumes a given prior distribution for the regression coefficients, and a given linear predictor. This combination is assumed to generate a given conditional distribution for the response (panel (A)). The prior information and the likelihood are then combined to generate a posterior distribution for the regression coefficients through MCMC (panel (B)).

with a given mean *m* and a given standard deviation *s*; this combination is assumed to generate a particular kind of conditional distribution for the response, which, in this case—as in the frequentist counterpart, follows a Beta probability distribution (Figure 9A). As before, a logit-link is used to map the confined interval (0, 1) to the real line. Next, through an iterative procedure known as Markov chain Monte Carlo (MCMC), the prior information and the information derived from the data (likelihood) are combined to generate a posterior distribution for the regression coefficients (Figure 9B).[54]

The estimates for each coefficient β can thus be easily obtained by extracting, eg, the mean and the 2.5 and 97.5 percentiles from the posterior distribution to obtain the 95% credible interval. Unlike the frequentist 95% confidence interval, the Bayesian 95% credible interval has an intuitive interpretation: it is an interval that has 95% probability of including the unobserved β. Our Bayesian regression model (without outlier) can be easily built (using default priors) with the package "brms,"[55,56] which uses a simple glmmTMB-like formula syntax:

```
library("brms")
glm_danger_no_outlier.brm <- brm(Level_of_danger.scaled ~ Temperature + log2(Country_size),
                                 family = Beta(link = "logit" link_phi = "log"),
                                 data = beasts_danger[-c(33),])

parameters(glm_danger_no_outlier.brm, ci = 0.95)
```

The results (not shown) are numerically very similar to those obtained with the classic frequentist approach. For this model, lots of interesting things (estimates, diagnostics . . .) can be interactively inspected with the package "shinystan"[57] (not shown):

```
library("shinystan")
launch_shinystan(glm_danger_no_outlier.brm)
```

Bayesian estimation has the great advantage of allowing to move beyond the need of assuming outcomes that never happened, and it provides a natural way to reflect estimate uncertainty.[23] Bayesian regression models are also generally robust even with small samples.[58] Nonetheless, Bayesian estimation is still relatively rare in ecology, possibly because of prior "subjectivity" and MCMC complexity. Fortunately, the development of user-friendly packages such as "brms" or "rstanarm"[59] has made model specification much more accessible. If this exercise has triggered in you the desire to know more about Bayesian regression, I would recommend the books of Zuur et al,[60] Korner-Nievergelt et al,[14] Kruschke,[54] McElreath[10] or Gelman et al.[2]

Coming back to the frequentist Beta regression model, as we did for the logistic regression, a visualization of the marginal relationships greatly helps the interpretation of model results (Figure 10):

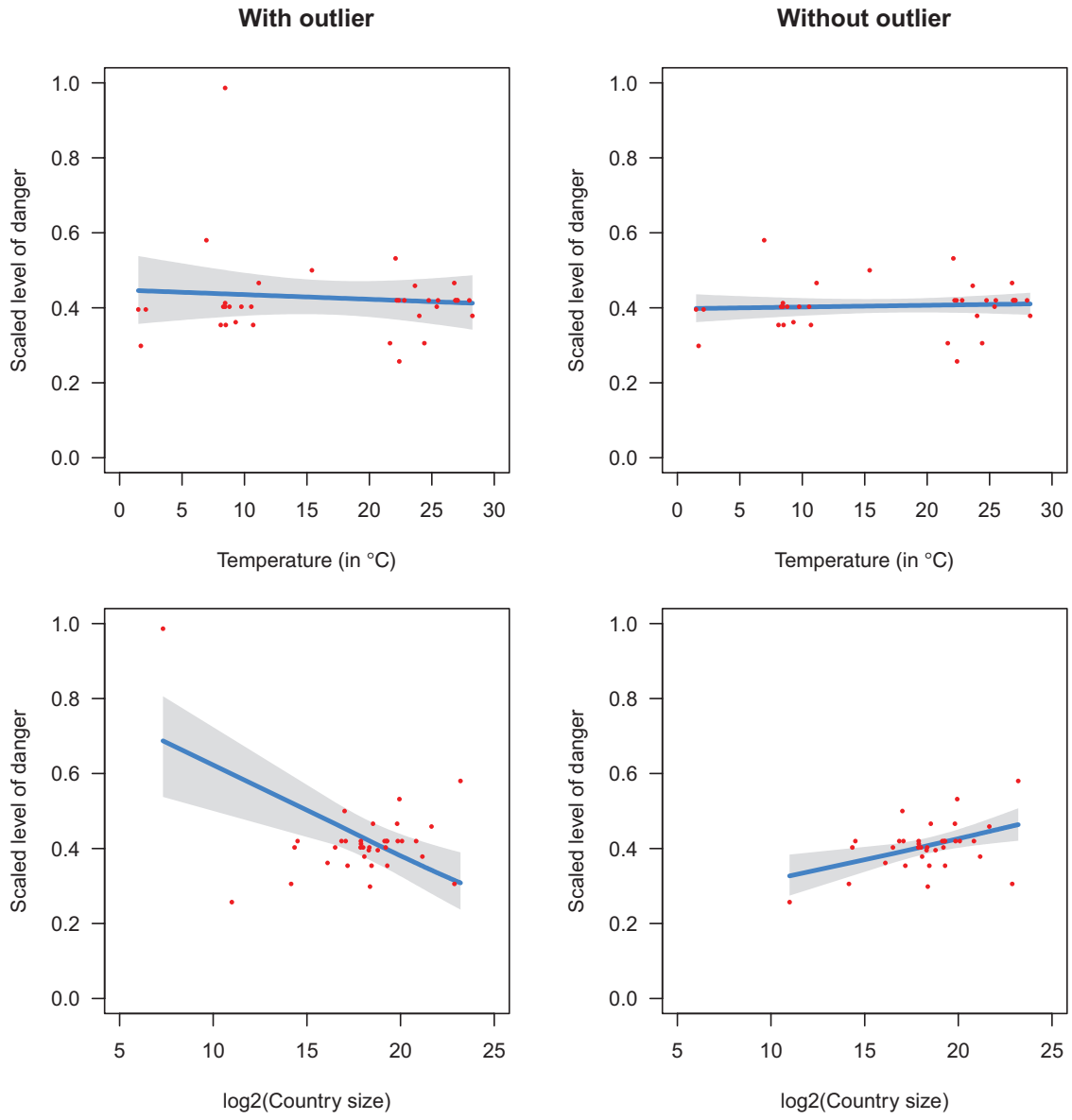


Figure 10. Marginal relationships of temperature and $\log_2$-country size with scaled mean perceived level of danger of country-specific magical animal communities, with (left column) and without (right column) outlier. Gray shaded areas show 95% confidence intervals. Original datapoints in red.

```
par(mfrow=c(2,2))

visreg(glm_danger.glmmTMB, "Temperature" scale = "response" rug = FALSE,
       xlim = c(0,30), ylim = c(0,1),
       main = "With outlier" xlab = "Temperature (in °C)" ylab = "Scaled level of danger")
with(beasts_danger, points(Temperature, Level_of_danger.scaled, pch = 16, cex = 0.5, col = "red"))

visreg(glm_danger_no_outlier.glmmTMB, "Temperature" scale = "response" rug = FALSE,
       xlim = c(0,30), ylim = c(0,1),
       main = "Without outlier" xlab = "Temperature (in °C)" ylab = "Scaled level of danger")
with(beasts_danger[-c(33),], points(Temperature, Level_of_danger.scaled, pch = 16, cex = 0.5, col = "red"))

visreg(glm_danger.glmmTMB, "Country_size" scale = "response" xtrans = log2, rug = FALSE,
       xlim = c(5,25), ylim = c(0,1),
       xlab = "log2(Country_size)" ylab = "Scaled level of danger")
with(beasts_danger, points(log2(Country_size), Level_of_danger.scaled, pch = 16, cex = 0.5, col = "red"))

visreg(glm_danger_no_outlier.glmmTMB, "Country_size" scale = "response" xtrans = log2, rug = FALSE,
       xlim = c(5,25), ylim = c(0,1),
       xlab = "log2(Country_size)" ylab = "Scaled level of danger")
with(beasts_danger[-c(33),], points(log2(Country_size), Level_of_danger.scaled, pch = 16, cex = 0.5,col =
"red"))
```

To obtain the pseudo-R^2, calculated as the squared correlation of linear predictor and link-transformed response,[52] we can refit the models using the "betareg" function:

```
glm_danger.betareg <- betareg(Level_of_danger.scaled ~ Temperature + log2(Country_size),
                              link = "logit" link.phi = "log"
                              data = beasts_danger)
glm_danger.betareg$pseudo.r.squared # the function "summary" would also return a value for pseudo-R2
glm_danger_no_outlier.betareg <- betareg(Level_of_danger.scaled ~ Temperature + log2(Country_size),
                              link = "logit" link.phi = "log"
                              data = beasts_danger[-c(33),])
glm_danger_no_outlier.betareg$pseudo.r.squared
```

Interestingly, the results tell us very different things. In both cases, temperature does not have a significant relationship with danger level; yet, country size has opposite relationships with the perceived mean level of danger, depending whether the outlier is included or not! When including the outlier, the model explained about 22% of the variance, while the explained variance was slightly lower (21%) when excluding it. These issues are discussed in the next section.

Discussion

The results of this study are somewhat disappointing. For example, I hoped that my intuition about the potential relationship between temperature and the country's likelihood of hosting fantastic beasts could be supported by the data. The variation in the response variable, however, was explained by a rather trivial concept: the larger the country, the more likely the presence of a native beast. And if we exclude the extremely dangerous Quintapeds of the tiny island of Drear, the larger the country, the higher the mean perceived level of danger of the magical animal community that lives in it. The interpretation of results, however, always needs to be treated with caution.

First, although the aforementioned predictions hint at a causal relationship between temperature and occurrence/danger level of beasts, it should be pointed out that, strictly speaking, causal interpretations can be ascertained only through controlled experiments. Many, if not most, field studies in ecology are observational, and using regression to infer causal effects requires caution.[2] The regression models presented here were descriptive in nature, ie, they simply aimed to find relationships between variables. Strictly related to this, is the issue of "lurking variables," namely variables that are unknown and not controlled for, which might change the relationships between explanatory and response variables[25]; when available, variables that can have an important effect on the variables of interest should thus be included in the model. In fact, my models did not explain much of the data variation, suggesting that the variation in native presence and in mean perceived level of danger may be better described by the inclusion of further variables. It seems also plausible that the pooling of data might have hampered the amount of available information: beast-specific information, eg, which animals are present in which country, as well as their country-specific abundance, might help to define better models (for example, weighting the level of danger by the number of individuals present in the community, to buffer against the disproportionate effect they might have in our results) and possibly yield more insightful results. However, this was just a preliminary exploration, and other researchers should collect data on further explanatory variables that may clarify the ecological diversity of fantastic beasts, possibly including non-linear relationships.

Perhaps, the most interesting aspect of my investigation was the effect of the outlier on the model that explored the variation in perceived mean level of danger. Since the outlier does not seem to be a mistake in data collection, it is difficult to decide whether to keep this point or not in the dataset. To sort this out, we need some deeper statistical and ecological thinking. Outliers are not uncommon in ecology; natural variation is the rule, not the exception. So, what should we do? First, since we are arguably interested in the "bulk" of the data, it seems fair to inspect two models, one with and one without outlier. Next, it is worth discussing the difference between the two models, and pick either one based on sound theoretical and statistical justifications. From the ecological standpoint, for example, I argue that the outlier may not be very informative about the relationship between level of danger and the associated explanatory variable. Based on Scamander's narration, Quintapeds may have originated from the transmutation of wizards into terrible creatures; if so, I believe they would hardly be representative of the native habitat "choice" of beasts, since this is clearly a case of artificial colonization, and the outlier could be eliminated (also from the logistic regression, which is nonetheless much less affected by the datapoint). Conversely, under a hypothetical scenario where the really dangerous beasts truly evolved on the little island, it would not be justifiable to exclude the point. The model with outlier, however, did not behave well, and should we keep the Quintapeds in the dataset, a robust modeling approach—which downweighs the importance of extreme datapoints—may be more appropriate.

Notably, for the perceived mean level of danger, results similar to those obtained with the Beta models can be obtained by fitting simple linear models with the "lm" function (you can try to do some modeling yourself!). This is reassuring for our

inference, since GLMs are generally data-hungrier than simple linear models: unlike GLMs, LMs estimates are generally obtained through ordinary least squares and the test statistic follows a *t*-distribution, which is equivalent to the normal distribution when the number of cases becomes large, but it is more appropriate when the population standard deviation is not known and the sample size is small (*cf.* Box 1). But how is it possible that similar results are obtained, if our data are not ideal for Gaussian models? It turns out that, when excluding the outlier, the perceived mean level of danger is linearly related to the explanatory variable and most values are far from the extremes (ie, zero and one: *cf.* Figure 10). Therefore, within this realized range of values, the conditional distribution of the response variable is to some extent unconstrained, roughly symmetrically distributed, and with approximately constant variance. It should be noted that perceived danger per country is the mean value of the danger score per beast inhabiting a country, thus the central limit theorem guarantees that this mean value converges to a normal distribution. This explains why simple linear models, though theoretically not ideal to fit these data, worked fairly well in practice. This frequently happens with variables such as biometric measurements (eg, height or body mass) which, despite having biological boundaries, often show a realized data distribution that allows to use simple normal models without major violations of assumptions.[24] The results of LMs are also generally easier to interpret than those of GLMs, since LMs do not employ a transformed response. Notably, for the model without outlier, the $\hat{\beta}$ regression coefficients of LM using scaled data are very similar to the $\frac{\hat{\beta}}{4}$ regression coefficients of the corresponding Beta regression (*cf.* Box 2)! Irrespective of the modeling strategy, however, the results are not really exiting. But that was not the point of the paper anyway.

As a behavioral ecologist, and fan of the Harry Potter saga, I thought it would have been fun to investigate the ecology of the fantastic beasts. While doing that, I realized this was also a challenging and instructive task, from both the ecological and the statistical standpoints. Therefore, I thought to offer students a practical (albeit arguably simplistic) tutorial on regression modeling. Most importantly, my main take home message for them is simple: learning statistics is not only useful; it can also be fun!

Acknowledgements

I thank Niccolò Fattorini (University of Siena), Claudia Hermes (Birdlife International) and Ranjana Pal (Wildlife Institute of India) and an anonymous reviewer for helpful suggestions on earlier drafts of the manuscript.

Author Contributions

Luca Corlatti conceived the idea for this work, did the statistical analyses and wrote all drafts of the manuscript.

ORCID iD

Luca Corlatti https://orcid.org/0000-0002-2706-3875

Data Accessibility

Data are available in the supplementary file 1.

REFERENCES

1. Scamander N. *Fantastic Beasts and Where to Find Them*. Hogwarts, UK: Hogwarts Library Book; 1927.
2. Gelman A, Hill J, Vehtari A. *Regression and Other Stories*. Cambridge, UK: Cambridge University Press; 2020.
3. Mac Nally R. Multiple regression and inference in ecology and conservation biology: further comments on identifying important predictor variables. *Biodivers Conserv*. 2002;11:1397-1401.
4. Zuur AF, Ieno EN, Smith GM. *Analysing Ecological Data*. New York, NY: Springer; 2007.
5. Bolker BM, Brooks ME, Clark CJ, et al. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol Evol*. 2009;24:127-135.
6. Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. *Mixed Effects Models and Extensions in Ecology With R*. New York, NY: Springer; 2009.
7. Harrison XA, Donaldson L, Correa-Cano ME, et al. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*. 2018;6:e4794.
8. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference*. 2nd ed. New York, NY: Springer; 2002.
9. Krausman PR. Important considerations when using models. *J Wildlife Manage*. 2020;84:1221-1223.
10. McElreath R. *Statistical Rethinking: A Bayesian Course With Examples in R and Stan*. 2nd ed. London, England: Chapman and Hall/CRC Press; 2020.
11. Ellison AM, Dennis B. Paths to statistical fluency for ecologists. *Front Ecol Environ*. 2010;8:362-370.
12. delMas RC. Statistical literacy, reasoning, and learning: a commentary. *J Stat Educ*. 2002;10:3.
13. Cohen J, Cohen P, West SG, Alken LS. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates; 2003.
14. Korner-Nievergelt F, Roth T, von Felten S, Guélat J, Almasi B, Korner-Nievergelt P. *Bayesian Data Analysis in Ecology Using Linear Models With R, BUGS, and Stan*. 1st ed. London, England: Academic Press; 2015.
15. Zuur AF, Ieno EN. A protocol for conducting and presenting results of regression-type analyses. *Methods Ecol Evol*. 2016;7:636-645.
16. Sutherland WJ. Planning a research programme. In: Sutherland WJ, ed. *Ecological Census Techniques*. Cambridge, UK: Cambridge University Press; 2006:1-10.
17. Hamilton IM. Habitat selection. In: Breed MD, Moore J, eds. *Encyclopedia of Animal Behavior*. London, England: Academic Press; 2010:38-43.
18. Boyce MS, Vernier PR, Nielsen SE, Schmiegelow FKA. Evaluating resource selection functions. *Ecol Model*. 2002;157:281-300.
19. Murray DL, Sandercock BK. *Population Ecology in Practice*. New York, NY: Wiley; 2020.
20. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2020. https://www.R-project.org/
21. R Studio Team. *RStudio: Integrated Development for R*. Boston, MA: RStudio Inc.; 2020.
22. Buckley YM. Generalized linear models. In: Fox GA, Negrete-Yankelevich S, Sosa VJ, eds. *Ecological Statistics: Contemporary Theory and Application*. Oxford, UK: Oxford University Press; 2015:131-148.
23. Bolker BM. *Ecological Models and Data in R*. Princeton, NJ: Princeton University Press; 2009.
24. Pekár S, Brabec M. *Modern Analysis of Biological Data: Generalized Linear Models in R*. Brno, Czech Republic: Masaryk University Press; 2016.
25. Fox J. *Applied Regression Analysis and Generalized Linear Models*. Los Angeles, CA: SAGE; 2016.
26. Dunteman GH, Ho M-HR. *An Introduction to Generalized Linear Models*. Thousand Oaks, CA: SAGE; 2006.
27. Dormann CF, Elith J, Bacher S, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*. 2013;36:27-46.
28. Fox J, Weisberg S. *An R Companion to Applied Regression*. 3rd ed. Thousand Oaks, CA: SAGE; 2019.
29. Zuur AF, Ieno EN, Elphick CS. A protocol for data exploration to avoid common statistical problems. *Methods Ecol Evol*. 2010;1:3-14.

30. Fox J. *Regression Diagnostics: An Introduction*. 2nd ed. Los Angeles, CA: SAGE; 2020.
31. Westfall PH, Arias AL. *Understanding Regression Analysis: A Conditional Distribution Approach*. Boca Raton, FL: CRC Press; 2020.
32. Faraway JJ. *Linear Models With R*. 2nd ed. Boca Raton, FL: CRC Press; 2014.
33. Faraway JJ. *Extending the Linear Models With R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. 2nd ed. Boca Raton, FL: CRC Press; 2016.
34. Halsey LG. The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biol Lett*. 2019;15:20190174.
35. Dunn PK, Smyth GK. *Generalized Linear Models With Examples in R*. New York, NY: Springer; 2018.
36. Dunn KP, Smyth GK. Randomized quantile residuals. *J Comput Graph Stat*. 1996;5:1-10.
37. Hartig F. DHARMa: residual diagnostics for hierarchical (multi-level / mixed) regression models, R package version 0.3.2.0, 2020. https://CRAN.R-project.org/package=DHARMa
38. Lüdecke D, Ben-Shachar MS, Makowski D. Describe and understand your model's parameters. CRAN, 2020. https://easystats.github.io/parameters
39. Canty A, Ripley B. boot: Bootstrap R (S-Plus) functions, R package version 1.3-25, 2020. https://cran.r-project.org/web/packages/boot
40. Upton G, Cook I. *Oxford Dictionary of Statistics*. Oxford, UK: Oxford University Press; 2002.
41. Best H, Wolf C. Logistic regression. In: Best H, Wolf C, eds. *The Sage Handbook of Regression Analysis and Causal Inference*. Los Angeles, CA: SAGE; 2015:153-172.
42. Lüdecke D, Makowski D, Waggoner P, Patil I. Assessment of regression models performance. CRAN, 2020. https://easystats.github.io/performance
43. Tjur T. Coefficients of determination in logistic regression models—a new proposal: the coefficient of discrimination. *Am Stat*. 2009;63:366-372.
44. Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett*. 2006;27:861-874.
45. Robin X, Turck N, Hainard A, et al. PROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.
46. Harrell FE. *Regression Modeling Strategies. With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. New York, NY: Springer; 2015.
47. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80:27-38.
48. Heinze G, Ploner M. logistf: Firth's bias-reduced logistic regression, R package version 1.23.1, 2020. https://CRAN.R-project.org/package=logistf
49. Breheny P, Burchett W. Visualization of regression models using visreg. *R J*. 2017;9:56-71.
50. Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. *J Appl Stat*. 2004;31:799-815.
51. Smithson M, Verkuilen J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol Methods*. 2006;11:54-71.
52. Cribari-Neto F, Zeileis A. Beta regression in R. *J Stat Softw*. 2010;34:1-24.
53. Brooks ME, Kristensen K, van Benthem KJ, et al. GlmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R J*. 2017;9:378-400.
54. Kruschke JK. *Doing Bayesian Data Analysis*. London, England: Academic Press; 2015.
55. Bürkner P-C. Brms: an R package for Bayesian multilevel models using Stan. *J Stat Softw*. 2017;80:1-28.
56. Bürkner P-C. Advanced Bayesian multilevel modeling with the R package brms. *R J*. 2018;10:395-411.
57. Gabry J. shinystan: interactive visual and numerical diagnostics and posterior analysis for Bayesian models, R package version 2.5.0, 2018. http://CRAN.R-project.org/package=shinystan
58. van de Schoot R, Miočević M. *Small Sample Size Solutions: A Guide for Applied Researchers and Practitioners*. Abingdon, UK: Routledge; 2020.
59. Goodrich B, Gabry J, Ali I, Brilleman S. rstanarm: Bayesian applied regression modeling via Stan, R package version 2.21.1, 2020. https://mc-stan.org/rstanarm
60. Zuur AF, Hilbe JM, Ieno EN. *A Beginner's Guide to GLM and GLMM With R. A Frequentist and Bayesian Perspective for Ecologists*. Newburgh, UK: Highland Statistics Ltd.; 2013.