Power analyses for Southern Resident killer whale demographic modeling

Eric Ward, Will Satterthwaite

7/13/2020

Background

This markdown file presents a set of relatively simple power analysis for logistic regression models used to link external covariates with SRKW fecundity or survival. Using fecundity as an example, the general form of these models is

$$logit(Pr(birth)) = B_0 + B_1 * X + B_2 * age + B_3 * age^2$$

where B_0 is the intercept, X is an optional covariate (such as prey) linked to birth rates through the parameter B_1, and B_2 and B_3 are optional covariates to account for a quadratic age effect. We assume that X is z-transformed, i.e. standardized to have mean 0 and standard deviation of 1. Note that the true functional form of the relationship is likely different, measurements are uncertain (detection of births is due in part to the whales' presence in inland waters, which is variable by pod and year), and other unmeasured covariates have effects as well. These complications are not accounted for in these power analyses.

Following off work by the 2011-2012 bilateral science panel, much of the recent regression work has asked to what degree external perturbations to covariates affecting prey (hatchery production, fishing closures) might affect vital rates. This document attempts to link the magnitude of these effects (% increases over status quo) to the idea of statistical power, or the probability of detecting a change when it exists. We present simulations based around the question: What is the power to detect effects of interventions whose effects are not mediated through directly observed covariates, through a before-after comparison? We also offer some qualitative guidance about what this means with respect to our ability to detect effects of measured covariates, but a full power analysis of this question turns out to be very computationally expensive, and very sensitive to assumptions and modeling choices.

Simulations

Simulation overview

To simulate before/after analysis of some intervention, we performed simulations, using fecundity of SRKWs as the response variable. For this first set of simulations, prey was not included as a covariate (i.e., there was no X and no B_1 term). These simulations were done as follows:

1. First, we fit several alternative models to the entire SRKW dataset. These models largely follow from those published previously, including quadratic effects of age for fecundity and stages for survival (Hilborn et al. 2012, Ward et al. 2013, PFMC 2020). We fit a simple logistic regression model (GLM, binomial family, logit link) without time varying effects, and a second generalized additive model (GAM, binomial family, logit link) with a smooth term estimated over the year effects. The smooth term can be thought of as the collection of drivers affecting fecundity or that are unrelated to age or sex and not directly measured (noise, toxins, disease, etc., also prey if it is not modeled as a covariate).

2. Given these model fits, we performed simulations that generated new datasets based on the estimated model parameters, but varying the intercept parameter of each model, in logit space, for the second half of the simulated time series (after the year 2000). In the absence of covariates, the intercept in logistic regression models of SRKW fecundity can be viewed as the mean birth rate (on the logit scale). We varied each intercept from 0.00 to 1, in steps of 0.01. For each proposed intercept, we simulated 1000 new datasets from the model fit to the data in Step 1. Each dataset was simulated to have the

1

same intervention, in the middle of the time series (for these simulated data, data after the year 2000 were generated the new intercept, while data before 2000 were generated using the original intercept).

3. For each of the 1000 simulated datasets, we re-fit GLMs and GAMs, as in Step 1, but for each included the effect of an indicator variable (0 before the year 2000, 1 after). As a last binary fork in the simulation, we simulated datasets to either have parameters shared across individuals and time steps, or random parameters for each individual and time step.

4. We calculated statistical power in Step 3 as the proportion of simulated datasets where the effect of the indicator was "statistically significant" (in the typical sense of $p <= 0.05$).

With two modeling approaches (GLMs with constant rates through time, GAMs with time varying rates) and two assumptions about parameter variability (constant, random by individual-year), this approach will produce four power curves. We expect that simpler models (time-invariant GLMs) and less parameter variability (constant values shared across individuals and years) will yield higher power for a given effect.

Interpreting results

Though our simulations included a range of effect sizes (i.e., multipliers applied to the base fecundity rate – such that an effect size of 1.0 means no effect), we truncated the upper limit of these to be 1.3. This number was chosen because an increase in fecundity rates of 30% would result in fecundity rates comparable to Northern Resident killer whales (Ward et al. 2013, Ward et al. 2009), which is presumably close to the biological maximum for the species.

First, we can plot the effect sizes versus statistical power for each combination of model and parameter uncertainty (Figure 1). Statistical power here is defined as the probability of detecting a statistically significant effect ($p <= 0.05$). This plot illustrates two important concepts. First, simpler models without time-varying parameters (GLMs) can produce higher power than those that have time-varying fecundity rates (GAMs). Second, when simulated datasets are generated based on constant parameters across individuals and time (sampling = "constant") the power is higher than with full (and more realistic) variability (sampling = "random"). Combined, these results confirm our hypothesis that simpler models will yield higher statistical power. The simple models used in this simulation (GLMs or GAMs with constant parameters) should be seen as oversimplifications, both in assumptions made about time-varying trends, and propagating parameter uncertainty.

An important result from this plot is that for most cases, statistical power is low - even for the largest effect sizes. For example, when parameters are held constant in simulating new datasets, the GAM model's highest power is 0.722. When parameter uncertainty is not held constant, power is 0.368 and 0.126 for the temporally constant GLM and time-varying GAM, respectively. Effect sizes smaller than about 1.1 are unlikely to be detected even under the most optimistic modeling assumptions. Designed experiments often target a power of at least 80%, which can only be achieved for these models under the most optimistic modeling assumptions and for fairly strong effects.

Note also that for an effect size of 1.0 (i.e., no effect) the power reflects the probability of a false positive, which should be close to the target p-value of 0.05. If the critical p-value is raised to 0.10, power goes up, but so does the risk of a false positive (Figure 2).

Interpreting effect sizes

One obvious question that arises from these simulations is how they relate to proposed adjustments to prey levels (separate from the question of estimating the effect of known changes in prey abundance, which we address later). We first ask, in the absence of measures of prey abundance, but assuming a logit-scale relationship between prey abundance and SRKW demography, how much power do we have to detect the demographic response to an increase in mean prey abundance? That is, given prey effect term B_1, what is our power to detect a demographic response to increase prey abundance? We can address this by calculating the changes in $B1 * X$ that are equivalent to a specified change in the intercept. Note that the logit link function makes the relationship between the effect and response non-linear. Note also that this is a separate question
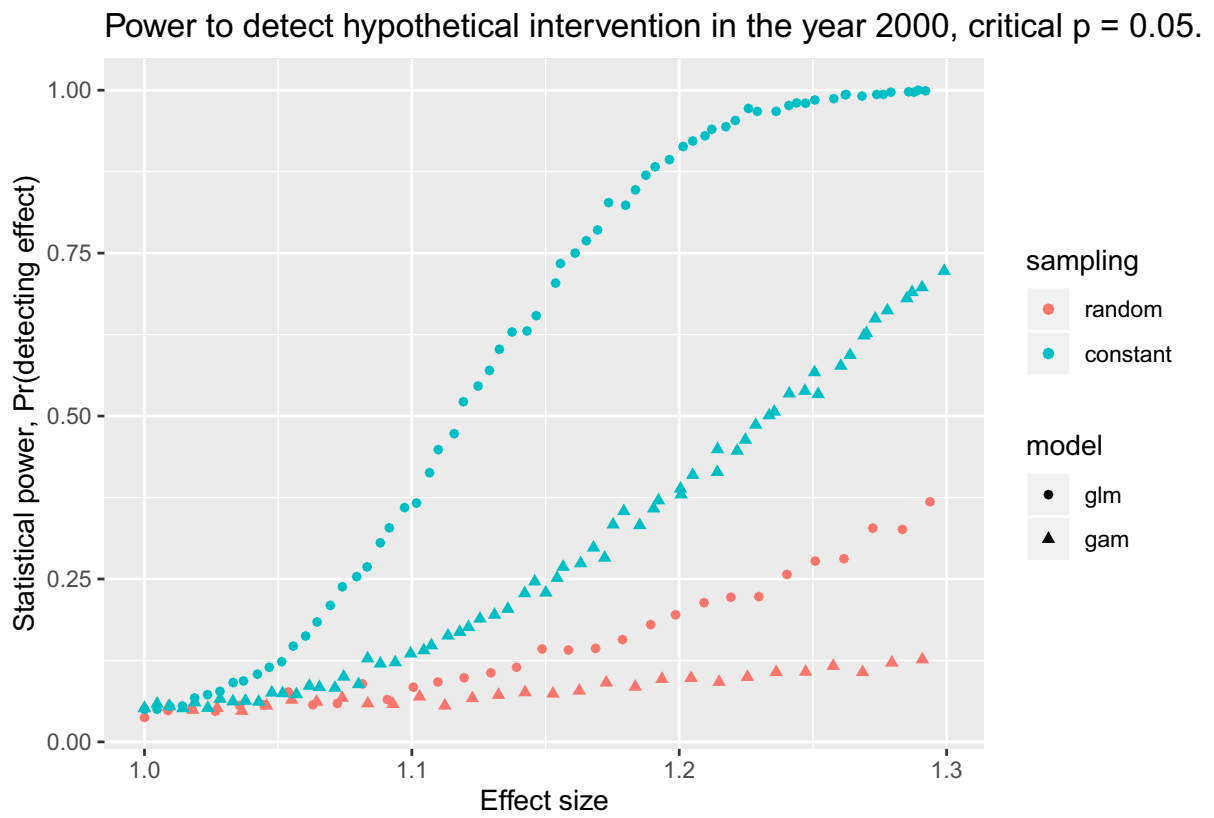
2

Figure 1: Statistical power as a function of changes the effect size in normal-space (1.0 = no effect, 1.3 = 30% increase), using critical p value of 0.05.
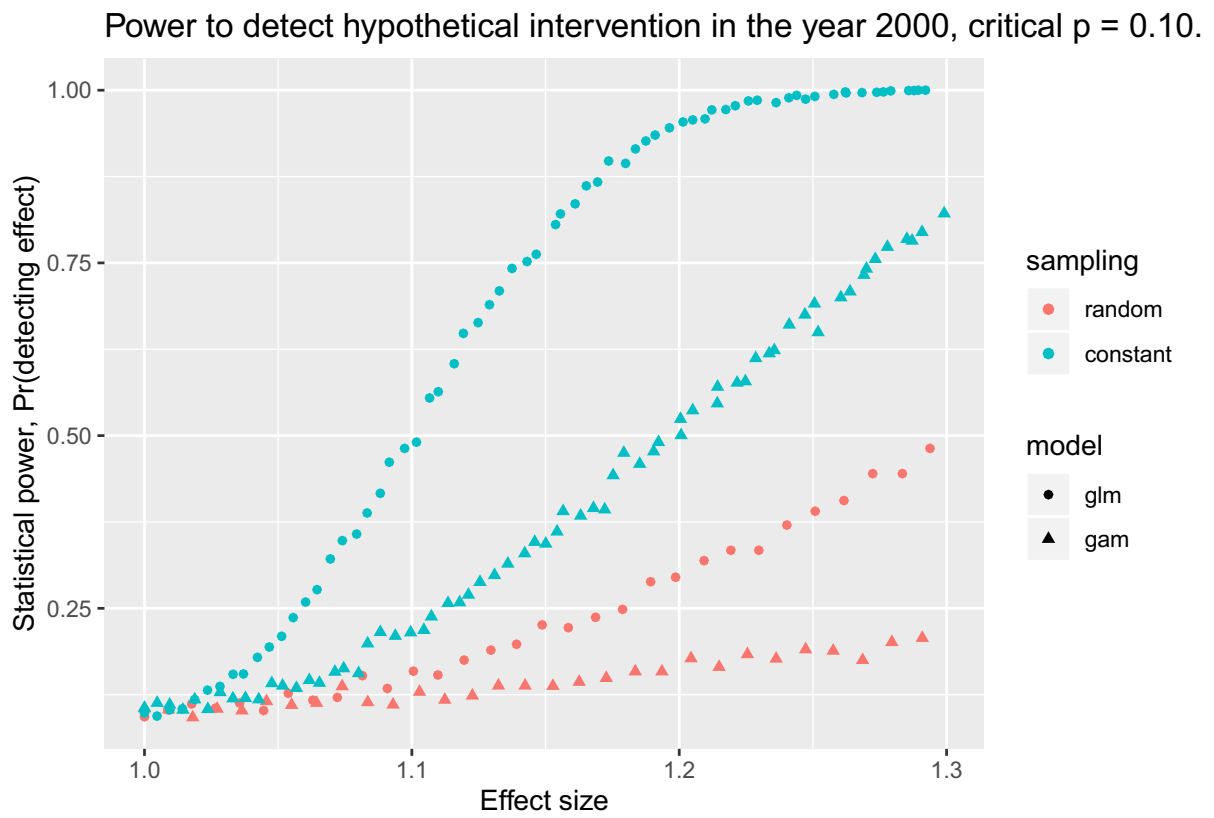
Figure 2: Statistical power as a function of changes the effect size in normal-space (1.0 = no effect, 1.3 = 30% increase), using critical p value of 0.1.
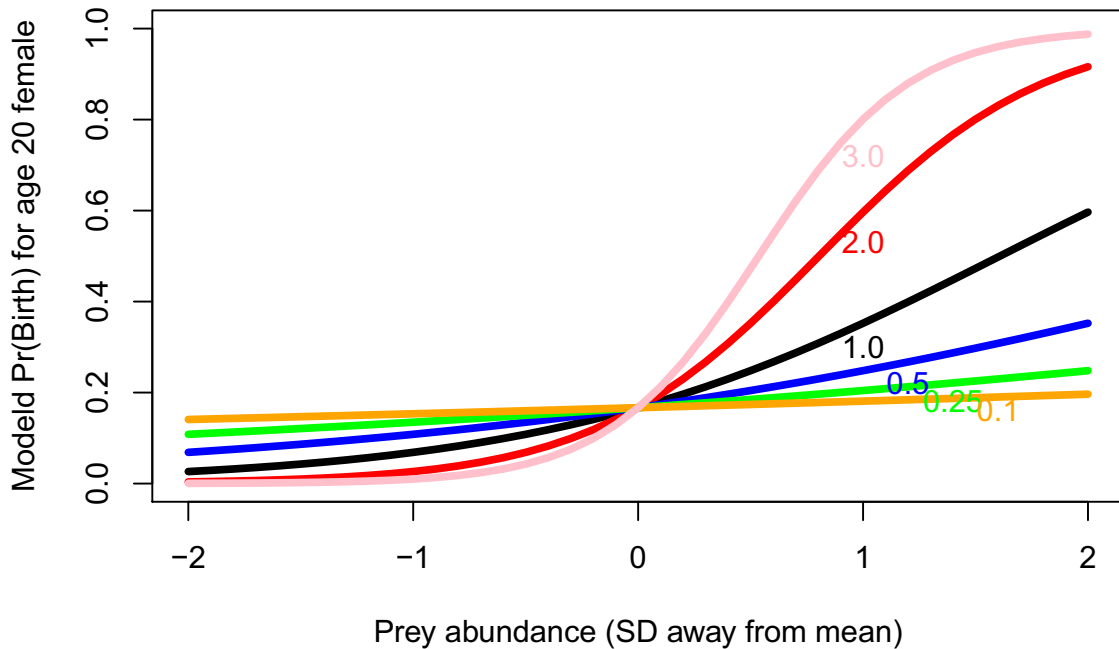
Figure 3: Interpreting modeled effects of prey abundance on the linear scale

from our power to detect whether B_1 is significantly different from zero, i.e. it is a separate question from our power to detect effects of changing prey abundance on demography, given measurements of prey abundance.

Due to the way our model is formulated, the modeled response of fecundity to prey abundance is not linear, and it also depends on age. Figure 3 shows how modeled fecundity changes as a function of prey abundance (in units of standard deviations away from the mean) for a range of values of B_1. $B_1 = 1.0$ implies that the probability of an age-20 female giving birth increases from 16.7% to 35.2% when prey increases by one standard deviation above its mean. Note that the response is approximately linear for small changes in prey abundance, but deviates from linearity in responses to large changes. Nonlinearity is also more pronounced for larger values of B_1.

As an example, we'll look at the coefficients from either the GLM or GAM model - which are identical, except for the time-varying components in the GAM. The intercept and age effects are identical, e.g.

$$logit(Pr(birth)) = -3.708357 + B_1 * X + 0.190689 * age - 0.004285 * age^2$$

Instead of presenting Figure 1 with the effect size on the x-axis, we can think about the values of the prey effect B_1 times the covariate X that would be needed to result in the same statistical power. Mapping these covariate values onto effect sizes yields the following plot, which shows effect sizes are just a function of the sampling scheme, and not the modeling approach (GLM vs GAM). An important point is that though these lines might look linear, they're actually much better predicted with non-linear quadratic relationships.

5
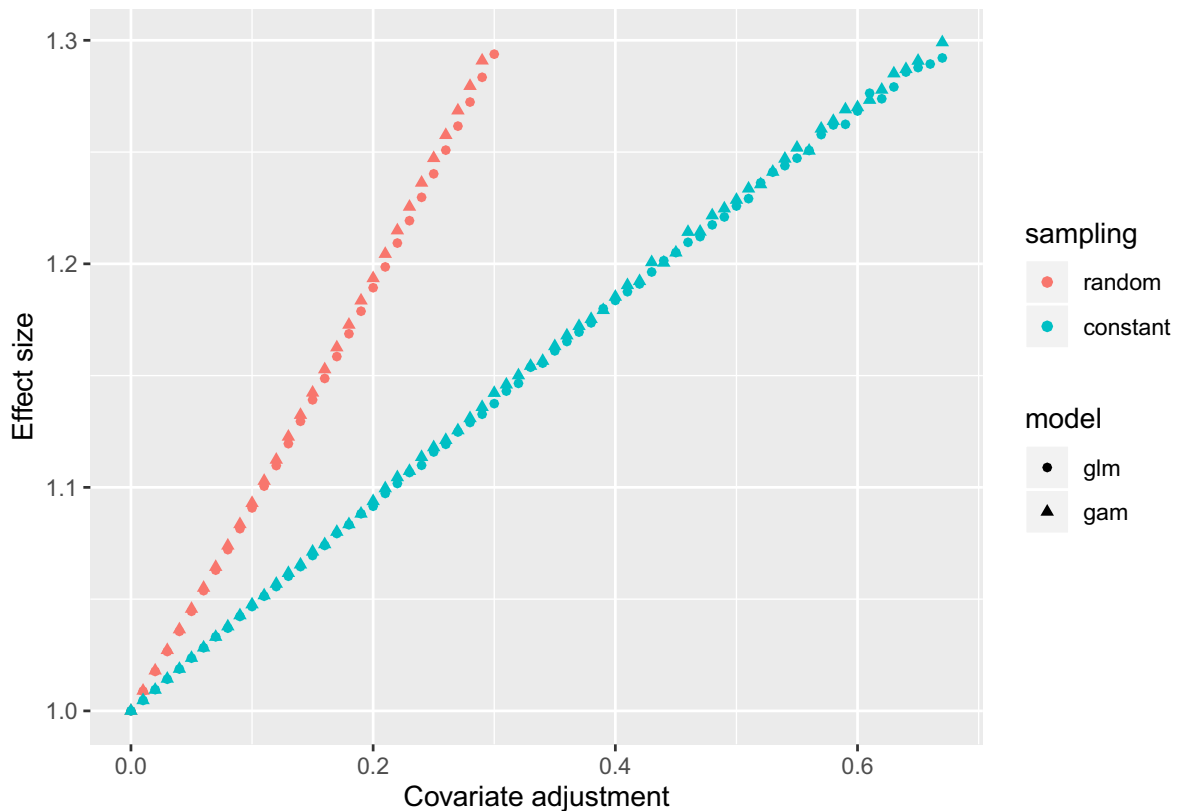
Examples

Two simple examples are included here to demonstrate how the above relationships can be used.

As a first example, suppose we were using a GLM model of SRKW fecundity, assuming parameters not to be constant across animals and time in our simulations, and were interested in finding what level of change in the covariate would be needed to produce an effect size of 1.2 (raising fecundity rates by 20%). The steps would be

1. From the estimated slopes in Figure 2, our predicted equation would be $1.2 = 1 + 0.886 * B_1 * X + 0.350 * X * X$.

2. Solving for $B_1 * X$, we get a solution at $B_1 * X = 0.209$. So our standardized covariate time series would need to change from the mean (0) to $0.209/B_1$ to yield this desired effect. In other words, mean prey abundance would need to increase by about 21 percent of its standard deviation if B_1 is equal to 1.0, and by correspondingly larger amounts as B_1 becomes smaller, or by smaller amounts as B_1 becomes larger. Note that this is a separate question from our power to detect whether B_1 is significantly different from zero.

3. Given B_1 and a mean and standard deviation of the raw data used in scaling the time series, we can back-calculate what this value corresponds to in normal space (e.g. translate 0.209 to raw indices of prey abundance).

As a second example, suppose we wanted to examine effects of fishing restrictions and ask what level of statistical power would we have to detect the intervention (effect of closures). Suppose that external experts have estimated that such closures would increase the mean of our standardized prey time series from 0 to 0.1, i.e. increase the mean by 10 percent of a standard deviation.

1. Assuming we're using a GAM and parameters are drawn randomly, and assuming $B_1 = 1$, we can use the relationships between covariate values and effect sizes to estimate an effect size of $1.09 =$

6

$1 + 0.886 * 0.1 + 0.350 * 0.1 * 0.1.$

2. Examining the curves between power and effect sizes shown in Figure 1, an effect size for this model (GAM, random) would produce a statistical power of ~ 0.06 if B_1 was 1.0. Power would be higher if B_1 were larger than 1.0, and lower if B_1 were smaller. This could be evaluated by replacing $0.886 * 0.1$ with $0.886 * B_1 * 0.1$ in the equation above, calculating a new value in place of 1.09, and determining the power of that new effect size based on Figure 2.

Performing a power analysis with respect to detecting whether B_1 is significantly different from zero would be very computationally expensive. The power associated with estimates of B_1 are also dependent on the nature of prey abundance – power will be higher if there is a large amount of data from years at all different abundance levels, and lower if only a limited range of prey abundances are well represented in the data. Power will also vary based on the number of females of prime reproductive age each year who did not give birth the previous year. Finally, power will depend critically on the appropriateness of model assumptions, and a misleadingly optimistic assessment of power will result if factors beyond those included in the data-generating model affect the individual- and/or year-specific probabilities of demographic events. Observation error in the covariate would also need to be considered in a reliable power analysis.