To Offset or Not: Using Offsets in Count Models

Stephen Parry

1 Introduction

When collecting count data, larger or smaller counts may be observed based on the size of observational units or the amount of time spent measuring a particular unit. For example, suppose that the count variable of interest is the number of ants that arrive at a food source. Ideally, each food source would be observed for the same amount of time. However, if the food sources were observed for varying amounts of time, then we would expect to observe larger numbers of ants at the food sources observed for longer periods of time. In this case, it may be more appropriate to model the *rate* of ants observed per unit of time rather than the *number* of ants observed at each food source.

To model a count variable as a rate we use an *offset* variable.

2 Offsets in count regression models

Poisson and negative binomial regression models are frequently used to model count data. The Poisson model can be written as

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

where μ is the mean of the response variable and $x_1, \ldots, x_p$ are the predictor variables. The exponentiated coefficient of a predictor variable from a Poisson model tells us how much the expected count changes multiplicatively for a one unit increase in the predictor variable.

An offset variable represents the size, exposure or measurement time, or population size of each observational unit. The regression coefficient for an offset variable is constrained to be 1, thus allowing our model to represent rates rather than counts. In the count regression model given above, the offset variable is equal to the log of the measurement time (population size, unit size, etc.). For the ant arrival example, the offset variable would be the log of the amount of time spent observing each food source.

Suppose that A is the amount of measurement time. Then the Poisson regression model, including the offset, is given by

$$\log(\mu) = 1 \times \log(A) + \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

which is equivalent to

$$\log\left(\frac{\mu}{A}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

When you fit a model with an offset, the exponentiated regression coefficient of a predictor variable tells us how much the expected *rate* changes multiplicatively for a one unit increase in the predictor variable. For the ant example, the rate is expressed as “number of ants per unit of time”.

When using an offset, the assumption is made that doubling the unit size (measurement time, etc.) will lead to a doubling of the count outcome. If this assumption is not appropriate, controlling for the unit size as a covariate instead of an offset may be more appropriate. A likelihood ratio test could be used to determine if the fit of these two models is significantly different.

By not controlling for exposure or effort, the model may have over-dispersion due to a lack of fit.

For studies that involve running a series of trials where each trial has a binary outcome (success or failure), it may be tempting to use a count model where the response is the count of the number of successes and the number of trials is considered the unit size. However, since the number of successes and the number of trials have the same units, and the number of successes cannot exceed the number of trials, this type of data would be better analyzed using a binomial (logistic regression) model.

Created March 2018. Last updated April 2022.