# Effect Sizes for Categorical Data

# Seamus Donnelly1 & Jay Verkuilen 2

# Max Planck Institute for Psycholinguistics1

# Graduate Center of the City University of New York2

# Author Note

#### This research was funded by a grant from Australian Research Council.

#### Correspondence should be sent to Seamus Donnelly, Max Planck Institute for Psycholinguistics.

###### Abstract

#### Here is an abstract

###### Effect Sizes for Categorical Data

Categorical data are widespread in language acquisition research: response accuracy in computerized tasks, production data in corpus and priming studies, and choices in the forced point paradigm. Because categorical data do not meet the assumptions of linear models, researchers have recently turned to generalized linear models–such as logistic regression, multinomial regression and the cumulative logit model–to analyze their data. While such models are better suited to analysing categorical data, they quantify relationships between variables in logits, which are notoriously difficult to interpret, even for statisticians. Here, we compare the logit to two alternative effect size metrics for binary data (percentage point differences, risk ratios). We discuss the different interpretations of the three metrics and show how to estimate them using Bayesian methods. We emphasize, throughout, that there is no generally correct choice of effect size measure; rather, we encourage researchers to carefully consider all three in reaching statistical inferences.

## Example

We begin with a brief review of logistic regression, emphasizing the problems it solves as well as some of its interpretational challenges. Consider a hypothetical syntactic priming experiment, where 100 participants of various ages are assigned to one of two conditions (an active or passive prime) and complete one test trial. In such a situation, we have three variables (response, prime type and age) for 100 participants. One might fit a linear regression to this model with the form:

An advantage of linear regression, such as the model in Equation 1, is its easy interpretation. First, it assumes that prime and age combine additively and relate linearly to the dependent variable. In particular, this model assumes that the effect of prime is the same for all participants (regardless of their age) and that the effect of age is linear (that is, the difference in passives produced between three and four year olds is the same as the difference in passives produced between four and five year olds). Second, the model assumes error variance () is independent of the other parameters. In other words, the discrepancies between the model’s predicted values and observed values are assumed to be the same across all the model’s predictions [[1]](#footnote-29).

We could relax the assumptions of this model by allowing Age and Prime structure to interact (as in Equation 2):

However, doing so complicates the interpretation of the above parameters. For example, now refers to the difference between participants primed with actives or passives at whichever value of age is set to 0. In other words, the effect of Prime is now assumed to be conditional on the value of age. This is why researchers must be cautious in their choice of coding scheme for categorical variables with interactions (Citations).

While Equations 1 and 2 are reasonably easy to interpret, they are often inappropriate for binary data. Figure 1 shows that when applied to binary data, linear regression yields impermissible predictions. In particular, this model predicts that older participants in the passive prime condition will produce passives more than 100% of the time. This is a necessary consequence of the assumptions of linear regression. By assuming predictors relate linearly to outcomes, we assume that scores on the outcome increase without bound. Nothing in linear regression restricts to be between 0 and 1. This is one reason why fitting linear regression to binary data can result in biased parameter estimates (Barr, 2008, Jaeger, 2010, Mirman, Dixon & Magnusson, 2008), and even create spurious significant effects (Donnelly & Verkuilen, 2017).

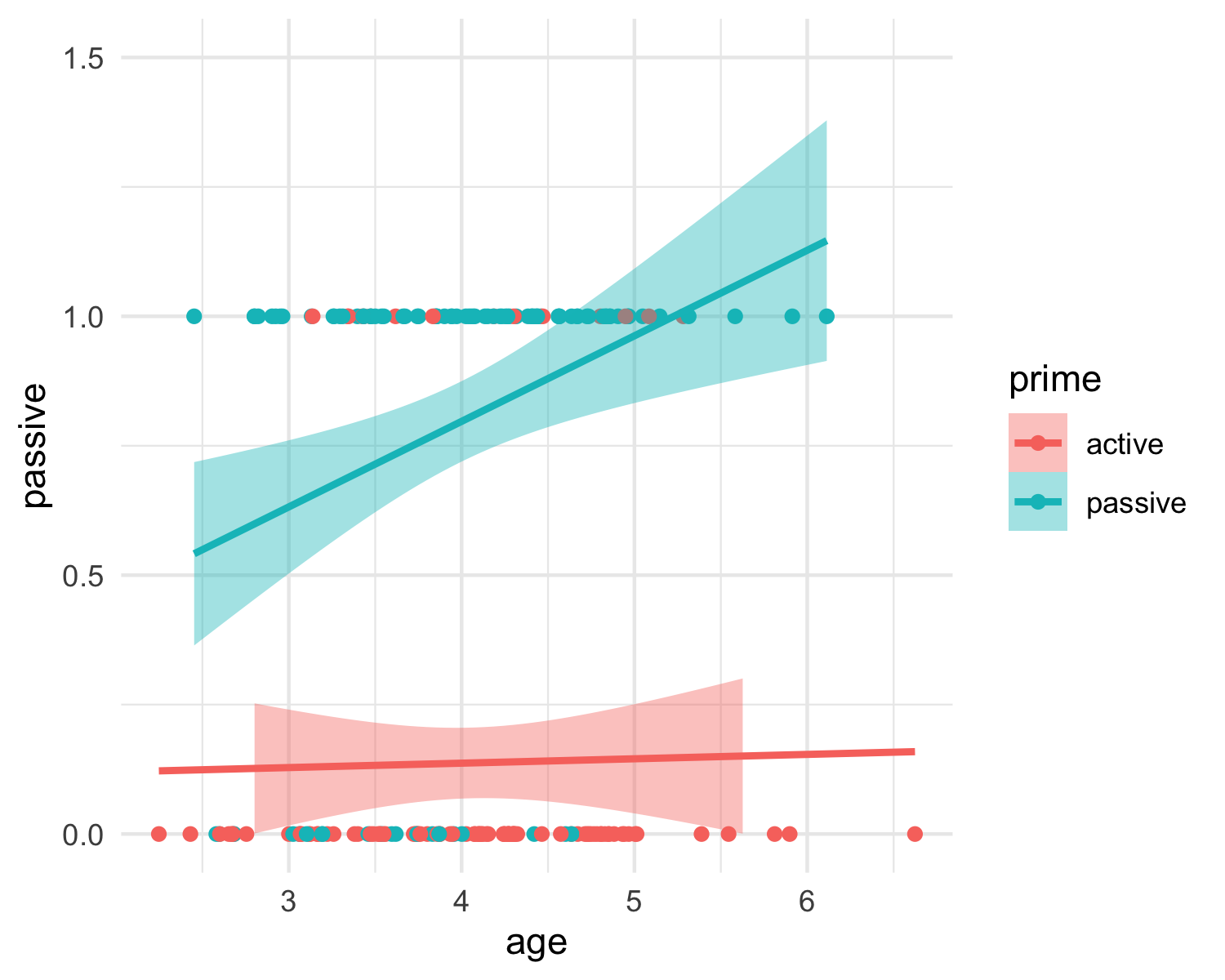


Figure 1. Data with regression lines from Equation 1

Logistic regression handles this problem by diverging from linear regression in two ways (as seen in Equation 3). First linear regression assumes that data are conditionally normally distributed, governed by a mean and a standard deviation . Logistic regression, on the other hand, assumes that data are conditionally binomially distributed and governed by a probability (), which can be seen in the first line of Equation 3. The lack of a separate variance parameter in the binomial likelihood function means that logistic regression assumes that a variable’s variance changes with its mean. If this sounds counter-intuitive, imagine flipping (a) a fair coin and (b) a coin weighted to produce 99% heads 100 times each. One would expect far more variability across trials for coin a than for coin b.

Second, in order to keep its predictions between 0 and 1, logistic regression employs a logistic link function, which we break into two parts below in Equation 4. The first line is called the **linear predictor**, which is an additive function of the predictor variables which can vary from to , much like the linear regression model in Equation 2. The linear predictor is then transformed by the logistic function (also called the inverse logit function), which is constrained to be between 0 and 1. Figure 2 shows how the linear predictor maps to by way of the logistic link function.

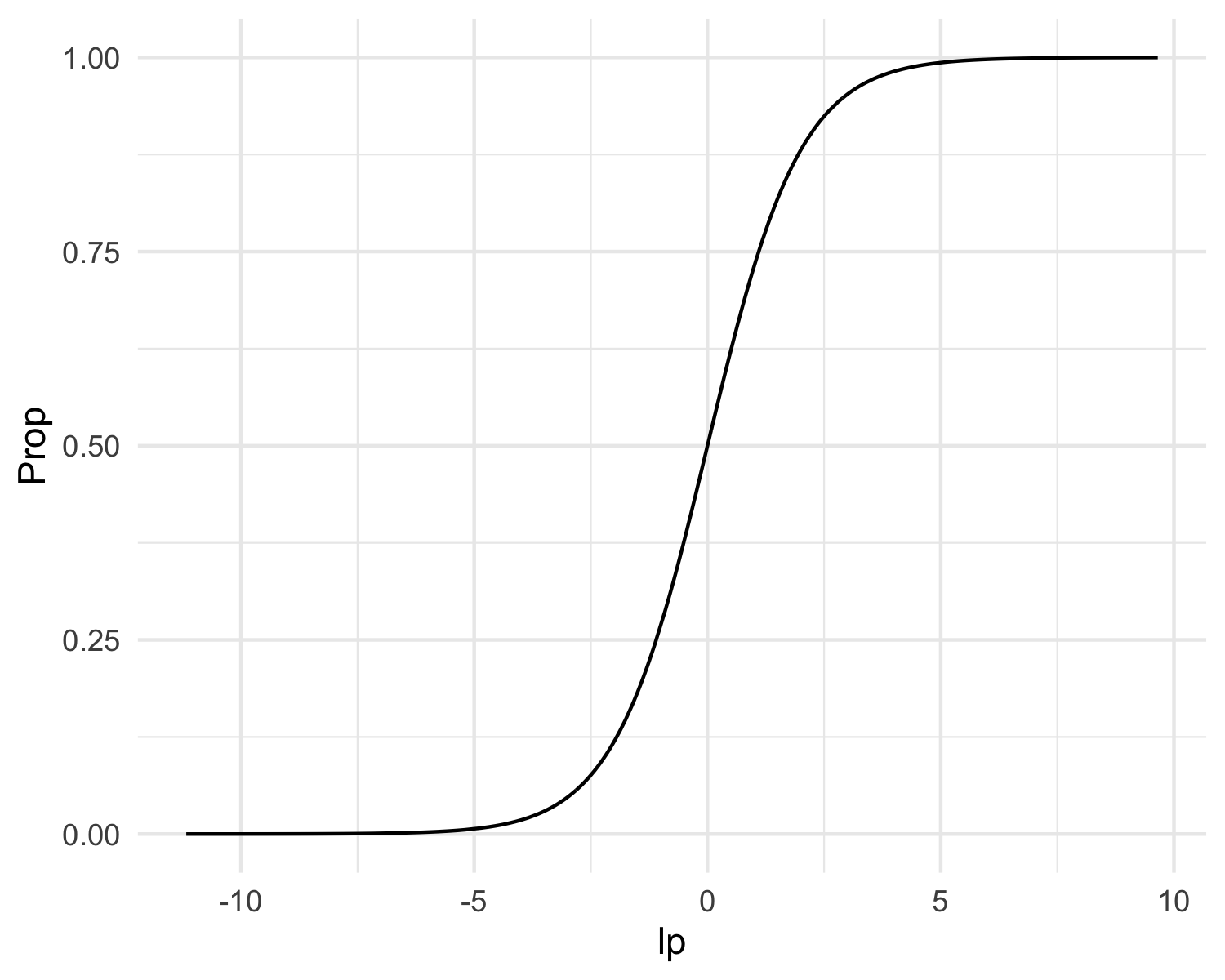


Figure 2. Relationship Between Linear Predictor and Y in Logistic Regression

Equation 4 is useful for illustrating why the predictions from logistic regression are constrained to be between 0 and 1. However, it masks one of the challenges in logistic regression: the interpretation of of the regression parameters. The first line of Equation 4 indicates that the parameters combine linearly and additively on the scale of the *linear predictor*. In other words, the model is linear inside the link function. However, because the link function is non-linear, the regression parameters do not combine additively or linearly to on the scale of . This is emphasized by Equation 5, which is an alternate expression of Equation 4.

If the reader is no longer sure how to interpret the parameters above, they are not alone. As an example, consider a model where = -2, = 1.1 and =0.5. Assume that Prime is dummy coded, with 0 for an active prime and 1 for a passive prime. The predicted probability that a 3 year old participant produces a passive after an active prime is 0.38 and the probability that a participant produces a passive after a passive is 0.65, for a difference of 0.27 percentage points. However, if the same child were 7 years old, the predicted probabilities would be 0.82 and 0.93 respectively, for a difference of 0.11 percentage points. In other words, even though prime and age are additive on the scale of the linear predictor, they interact on the scale of probability. This is induced by the non-linear link function–precisely the thing that constrains our predictions to be between 0 and 1.

In general, in logistic regression, the magnitude of the relationship between any parameter and y depends on the values of all the other parameters in the model (including ). Thus, while the logit link function is an excellent tool for estimating a logistic regression equation, its standard parameter estimates are often not well suited to describing the magnitude of the relationship between predictors and the outcomes. Fortunately, there is a long literature in epidemiology considering varying effect sizes for categorical data analysis, including the logit described above. We describe these measures in turn.

## One Table, 3 Effect Sizes.

Consider a simpler version of our example from above. Assume that, rather than treating age as a continuous variable, we treat it as a binary variable (children vs adults). As before, our independent variable is the prime type and our dependent variable is the proportion of passives produced on test trial (See Table 1). How can we quantify the magnitude of the priming effect for each group?

Adults

Children

Active

0.30

0.05

Passive

0.35

0.10

The most obvious method is to subtract the proportion of passives produced after active primes from the proportion of passives produced after passive primes, **percentage point changes**. In this case, the magnitude of the priming effect for both groups is 5 percentage points. This is a very similar approach to calculating Cohen’s d, which divides this difference score by the pooled standard deviation, and has been commonly used in the developmental syntactic priming literature (Rowland et al. 2012; Peter et al. 2015).

However, an alternative approach is to divide the proportion of passives produced after passive primes by the proportion of active primes, so called **risk ratios**. This shows us that children’s production of passives has increased two-fold, whereas adults’ productions of passives has increased by 1.16 or 16%. Risk ratios tell us how the experimental variable (in this case, prime structure) changed the frequency of the event. We suspect this quantity is what most researchers in syntactic priming research are interested in. For example, Peter et al. 2015 interpreted.

A third method of quantifying the strength of the priming effects from the data in this table is to calculate the odd’s ratio, which is the basis of logistic regression (logits are logarithm-transformed odds ratios). The odds of an event is the probability that the event occurs divided by the probability that the event does not occur, so the odds that children produce a passive after a passive prime is 0.1 divided by 0.9, or 0.111. We can divide this by the odds of producing a passive after an active prime, 0.053, to get the odds ratio 2.222. If we follow the same procedure for adults, we get an odds ratio of 1.252. These odds ratios tell us how the experimental variable affected the odds of producing a passive. However as odds are notoriously difficult to interpret, the widespread popularity of the odds ratio comes from its mathematical properties, rather than its intepretational properties.

## 3 Effect Sizes, 3 Models

The three effect sizes correspond to three different statistical models:

The linear regression model in Equation 2 produces parameter estimates on the percentage-point scale. This model assumes that percentage point differences vary as a linear function of the predictor variable. While this model seems counter-intuitive, if mean proportions are between .25 and .75, and percentage point changes are the desired method, linear regression (perhaps with robust standard errors accounting for heteroskedasticity), can be a reasonable choice of model. For example, XXXX strongly advocates this approach. An alternative, quite similar approach, is to retain the binomial likelihood function but do away with the logistic link function. Many statistical packages may fail to estimate this model but it can be estimated in the Bayesian context with suitable priors.

The logistic regression model defined in Equation 3 will produce coefficients defined on the scale of the odd’s ratio. In particular, logistic regression tells us how the (log transformed) odds ratio changes with the model’s predictor variables. The advantage of this model is that, by forcing all predictions to be between 0 and 1, it can handle

If we change the link function in equation 3 from the logistic to log link function, we have the model in Equation 4, whose coefficients are risk ratios. This model tells us how the risk ratio varies linearly as a function of the predictor variables. In practice, this model is difficult to estimate, as it does not restrict predictions to be between 0 and 1. When this link function is applied with the binomial likelihood function, traditional statistical software may fail to converge. However, such a model is estimable in a Bayesian context with an appropriate choice of priors.

1. Some readers may be more familiar with the form , which is different notation for the same model. This notation is very clear for linear regression; however, it does not easily extend to generalized linear models. In particular, for most GLMS a model’s variance is determined by its mean so s do not exist [↑](#footnote-ref-29)