

# Evaluating Main Effects in the Generalized Linear Model

Max A. Halvorson  
University of Washington

Connor J. McCabe  
University of Washington

Xiaolin Cao  
University of Washington

Dale S. Kim  
University of California, Los Angeles

Kevin M. King  
University of Washington

*Keywords:* count models, logistic regression, generalized linear models, data visualization

## Introduction

*The odds of having drank any alcohol during the past weekend are 1.8 times higher ( $p < .001$ ) amongst those who act rashly when experiencing positive emotion (i.e., high on positive urgency), controlling for the effect of gender.*

Given the above information, we can say that there is a direct relation between positive urgency and drinking. However, there are further research questions we likely set out to answer with the analysis we conducted. What is the probability of drinking alcohol on a given day for a person who is low on positive urgency? How about for someone who is high on positive urgency? Is this effect stronger for men? Stronger for women? Or equivalent across men and women?

If these questions are difficult to answer, it's not because the analysis can't answer them, and it's not because you've forgotten what you learned in the first couple years of graduate school - it's because these questions are unanswerable, given the information presented above. When modelling outcomes that are binary (e.g., drinking or no drinking) or counts (e.g., number of drinks in an evening), communication and interpretation of model parameters beyond simple significance tests can become difficult. Whereas regression weights, or unstandardized beta values, express a straightforward relation between a predictor and an outcome (for a 1-unit increase in X, Y increases by beta units), binary and count model parameters have less ready interpretations. Additional care needs to be given to communicating and interpreting model results. In the current work, we sought to 1) highlight common misconceptions about binary and count model interpretation, 2) advocate for the use of quantities of scientific interest (QSI) in the presentation of results, and 3) provide guidelines for effective communication of binary and count model findings.

## Binary and Count Outcomes Should be Modeled using Generalized Linear Models

Recent work has called for more widespread use of extensions of the general linear model which fit implicit assumptions about the data generation process. To facilitate this work, tutorials, software examples, and widely-available software packages make fitting these models more and more accessible (Atkins et al., 2007; Atkins & Gallop, 2013).

Despite widespread use of count and binary outcome models, reporting results of these models is typically done in a cursory fashion. Most commonly, researchers provide odds ratios (ORs) or risk ratios (RRs). Almost no researchers in the field of psychology (lit search results) report model results in terms of the quantities of scientific interest (QSI) they set out to understand. Moreover, graphics of these effects are rarely found in the literature. We found only two manuscripts out of 55 using binary or count outcomes in the *Journal of Abnormal Psychology* and *Journal of Consulting and Clinical Psychology* between 2007 and 2017 that reported XXXX. (FOOTNOTE)

Literature search was conducted using the same criteria as Norton (2004) and/or Brambor, Clark, & Golder (2006).

To make results easier to understand, methodologists recommend that GzLM parameters are best represented 1) in units that are most substantively meaningful and 2) that covariate values are chosen thoughtfully and reported clearly. In count models, this entails reporting predicted counts, and in binary outcome models, this entails reporting predicted probabilities of an event occurring. We advocate for presenting quantities of interest directly, as fitted models are readily able to output direct predictions of these quantities. Moreover, we strongly believe that graphical presentation of data can facilitate understanding in a way that is difficult or impossible for tables or single values to convey. Thoughtful graphical and tabular presentation of data can facilitate intuition even when models are complicated, and present a richer

source of information than single parameters. (GIVE SOME CITES HERE)

### The Case for Reporting Quantities of Scientific Interest

IS THERE COGNITIVE SCIENCE ON UNDERSTANDING STATISTICAL MODELS? EVIDENCE OF BIAS IN SCIENTIFIC INTERPRETATION (BRAIN PHOTOS) GARY KING AND CHRIS ADOLPH WORK DATA VISUALIZATION FROM CONNOR'S WORK

#### Interpretation of Single Coefficients does not Characterize Generalized Linear Models

In addition to the conceptual case for presenting QSIs, there is also a practical case to be made for not using transformed beta weights to interpret results from GzLMs. In examining predicted counts and probabilities, it becomes apparent that the single parameters reported in GzLMs do not map onto relations between predictors and outcomes as readily as they do in ordinary least squares (OLS) regression. The reason for this complexity is that typically-reported transformed versions of parameters, such as odds ratios and rate ratios, A) do not represent constant first differences in QOI, and B) do not account for what we call "inherent interactivity" of effects. By "inherent interactivity," we mean that main effects are not conditionally independent from one another. Thus, even when interaction terms are not specified, the effect of predictor 1 is partially dependent on the level of predictor 2. In the sections that follow, we work through several common GzLMs with several goals: to make the case for presenting QSIs, to explicate properties A and B for each class of models, and to review and build intuition on the models themselves..

In this paper, we demonstrate that these two properties hold for binary outcome and count models via a simulated-data example with two predictors for each type of model.

#### The Generalized Linear Model

Generalized linear models are those for which a link function is used to describe the relationship between predictors and the outcome they predict, when the relationship is not well described by a straight line. The mathematical formulation for any GLM can be shown as:

$$g(E[Y|X]) = X\beta, \quad (1)$$

where  $g(\cdot)$  is the link function,  $X\beta$  are linear predictors, and  $E[Y|X]$  is taken with respect to the probability distribution. In essence, the link function  $g(\cdot)$  transforms the left side of the equation so that the relation between  $Y$  and  $X$  is not simply a direct relationship with slope  $\beta$ ; rather, the entire left side of the equation varies with  $X\beta$ . Herein lies some of the ambiguity associated with results of GzLMs. In this form,

we lose sight of the idea that our QOI is the argument to the link function rather than the entire left side of the equation.

In the case of linear regression, there is an implicit identity function serving as the link function. However, this link function can take many forms, all of which imply a particular data generative process underlying the data themselves. Link functions take the shape of a relationship between a predictor and an outcome (in linear regression, the slope of a line) and allow that relationship to take on whatever shape is defined by the link function.

The link function we choose embodies our assumptions about the process that generates our data. In the case of OLS regression, we assume that our outcome is generated by a process with normally distributed errors. In the case of logistic regression, we assume that there is a latent variable representing underlying levels of some trait. For example, a person's diagnostic status as depressed or non-depressed represents an underlying level of continuous depression. Using a logit function, we can translate this latent level into a predicted probability of an outcome occurring. The link function allows predictor variables to influence this latent variable continuously, and describe the effect in terms of changes in probability of occurrence.

#### Binary Outcome Models

Binary outcome models involve predicting the probability of an outcome occurring (vs. not occurring) based on a set of predictors. The QOI researchers tend to be most interested in is the predicted probability of an occurrence (e.g., of a disease) for individuals with a characteristic or set of characteristics. More specifically, research questions tend to focus on how strongly a change in a predictor affects the probability of an outcome occurring. Motivated by parsimony and led by software defaults, scientists typically report effects of predictors in terms of a single odds ratio. This practice not only leads to coefficients with unintuitive interpretations, but also implies an independence of covariates and a constancy of the reported effect that are inconveniently untrue.

#### The Logistic Model

The logistic model is formulated as follows:

$$\pi_i = \frac{\exp(X\beta)}{1 + \exp(X\beta)}, \quad (2)$$

Since the logit link function is log-linear - that is, predictors relate to predicted probabilities through a log function - the extent of change in  $Y$  for a 1-unit increase in  $X$  is not linear. An odds ratio describes the relation between the odds of an outcome occurring and the odds of an outcome not occurring, and is a reasonable proxy for the percentage increase in probability, at least when proportions are small. Unfortunately, an odds ratio provides no information regarding the magnitude of predicted probabilities.

Model	Link Function	Quantity of Scientific Interest (QSI)	Typical parameter reported	Equation
Logistic	Logit	Probability of event occurring	Odds Ratio ()	$Y \sim X$
Poisson	Exponential	Number of events occurring	Odds Ratio ()	$Y \sim X$
s Logistic	Logit	Probability of event occurring	Odds Ratio ()	$Y \sim X$

Table 1

Parameter reporting

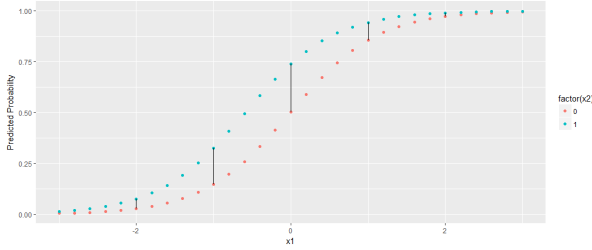


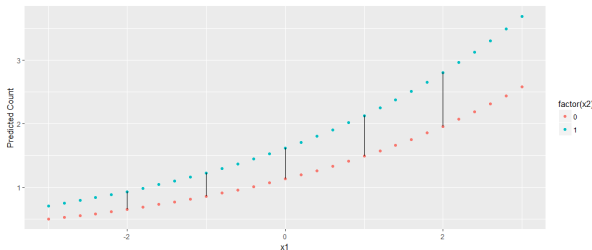
Figure X depicts the relation between a continuous predictor  $X_1$ , a categorical predictor  $X_2$ , and the model-predicted probability of a binary outcome  $Y$  occurring. The curve of each line makes it visually clear that a unit increase in  $X_1$  does not lead to a constant increase in  $P(Y)$ . Similarly, the black lines between the curves for  $X_2 = 0$  and  $X_2 = 1$  are different lengths, demonstrating visually that the effect of  $X_2$  (the amount that the blue line is above the red line) varies based on the level of  $X_1$ . Thus, even without interaction terms in a model, effects in logistic regression are not independent from one another.

### Count Models

#### The Poisson Model

$$\mathbb{P}(Y_i = y_i | x_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \quad \text{for } y_i = 0, 1, \dots \quad (3)$$

$$\mathbb{E}[Y_i | x_i] = \lambda_i = \exp(x_i^T \beta) \quad (4)$$



EXPLAIN INCIDENT RATE RATIOS

#### The Negative Binomial Model

negative binomial formulation:

$$\mathbb{P}(Y_i = y_i | x_i) = \frac{\Gamma(y_i + \theta)}{\Gamma(\theta) y_i!} \cdot \frac{\lambda_i^{y_i} \theta^\theta}{(\lambda_i + \theta)^{(y_i + \theta)}}, \quad (5)$$

$$\mathbb{E}[Y_i | x_i] = \lambda_i = \exp(x_i^T \beta), \quad (6)$$

where  $\Gamma(\cdot)$  is the gamma function and  $\theta$  is a constant shape parameter.

The negative binomial model is a generalization of the poisson model with an extra parameter, allowing for non-constant exposure. As such, the properties of non-constant first differences and non-independence of regression coefficients apply.

#### The Hurdle Model

hurdle model formulation:

$$\mathbb{P}(Y_i = y_i | x_i) = \begin{cases} \pi_i, & \text{if } y_i = 0 \\ (1 - \pi_i) \mathbb{P}_c(Y_i = y_i | x_i), & \text{if } y_i = 1, 2, \dots \end{cases} \quad (7)$$

$$\pi_i = \frac{\exp(x_i^T \beta_\pi)}{1 + \exp(x_i^T \beta_\pi)}, \quad (8)$$

$$\mathbb{E}[Y_i | x_i] = (1 - \pi_i) \mathbb{E}_c[Y_i | x_i], \quad (9)$$

where  $\pi_i$  is once again assumed to be a Bernoulli parameter with logit link.

The hurdle model is a piecewise model involving 1) a logistic regression portion which accounts for zeros and 2) a poisson (or other count) model to account for non-zero values. Hurdle models assume a two-step data generating processes: one that explains whether an observation was a 0 or a positive count (portion 1), and one that predicts that positive count (2). As these models are composed of the aforementioned models, the properties of non-constant first differences and non-independence of regression coefficients applies.

#### The Zero-Inflated Model

zero-inflated model formulation:

$$\mathbb{P}(Y_i = y_i | x_i) = \begin{cases} \pi_i + (1 - \pi_i) \mathbb{P}_c(Y_i = 0 | x_i), & \text{if } y_i = 0 \\ (1 - \pi_i) \mathbb{P}_c(Y_i = y_i | x_i), & \text{if } y_i = 1, 2, \dots \end{cases} \quad (10)$$

$$\pi_i = \frac{\exp(x_i^T \beta_\pi)}{1 + \exp(x_i^T \beta_\pi)}, \quad (11)$$

$$\mathbb{E}[Y_i|x_i] = (1 - \pi_i)\mathbb{E}_c[Y_i|x_i], \quad (12)$$

where  $\pi_i$  refers to the probability of an excess zero and  $\mathbb{E}_c[Y_i|x_i]$  is the expectation with respect to  $\mathbb{P}_c(Y_i|x_i)$ .

The zero-inflated model is a piecewise mixture model involving the poisson model (or another count model), which accounts for two processes of generating zeros. Structural zeros, represented by the first term in part 1 of the piecewise function, occur when zeros are thought to be the result of a lack of exposure. Sampling zeros, represented by the second term in part 1 of the piecewise function, are zeros that were sampled from the full count distribution. Positive counts are modeled by part 2 of the piecewise function. As such, the properties of non-constant first differences and non-independence of regression coefficients applies.

$$\mathbb{P}(Y_i = y_i|x_i) = \begin{cases} \pi_i + (1 - \pi_i)e^{-\lambda_i}, & \text{if } y_i = 0 \\ (1 - \pi_i)\frac{\lambda_i^{y_i}e^{-\lambda_i}}{y_i!}, & \text{if } y_i = 1, 2, \dots \end{cases} \quad (13)$$

$$\pi_i = \frac{\exp(x_i^T \beta_\pi)}{1 + \exp(x_i^T \beta_\pi)}, \quad (14)$$

$$\lambda_i = \exp(x_i^T \beta_\lambda), \quad (15)$$

$$\mathbb{E}[Y_i|x_i] = (1 - \pi_i)\lambda_i, \quad (16)$$

where  $\pi_i$  refers to the probability of an excess zero as before.

When presenting results, it is important to recall that the choice of covariate levels chosen as 0 values can influence the interpretation of results to the majority of readers, who likely do not have the available time or information to probe a model fully.

We propose that research producers should choose their covariate values thoughtfully.

In order to characterize an effect accurately, researchers may have to probe an effect at multiple covariate levels, even when no interactions are included in the model.

For example, an interpretation of the above example might be written as follows: There was a main effect of positive urgency on negative alcohol consequences (OR = XX, SE = XX,  $p = XX$ ). For a 21-year-old male, the predicted number of negative alcohol consequences was 3 for someone low in positive urgency and 8 for someone high on positive urgency. In contrast, for a 21-year-old female, the predicted number of consequences for 5 for a low-urgency individual and 12 for a high-urgency individual.

Make a comment about cumulative psychological science. This effort is in the direction of quantifying effect sizes, not just reporting p-values. Graphical methods are one means of giving more information about nonlinear models. Another

means is to give additional interpretation in the text of results and discussion sections. Especially at the beginning of a discussion section, this type of information can be very useful for readers to contextualize findings. This is also helpful in providing more information and intuition about base rates. But it generally contributes

## Recommendations for Model Reporting: Tables and Graphics

### Binary Outcome Data

Table of first differences with covariate values made explicit

Graphic of first differences for some X1 and X2 of interest  
Real data example???

Show graphs from InterActive? With uncertainty.

### Count Data

// Count Model

We used data from a study of XXXX (ABMRF) to construct a zero-inflated Poisson model predicting the number of negative consequences in the past year (negpysum). By inspecting a histogram of the consequences variable, we can see that it has excessive zeros and the distribution is not overdispersed. Therefore, we expected a zero-inflated Poisson model to describe the data well in this scenario. We used gender, positive urgency, premeditation and sensation seeking as predictor variables. In the process, we treated sensation seeking as the source of structural zeros and other predictors governed by Poisson distribution. The graph () shows that even in the absence of interaction terms, the first difference of predicted consequences is not constant across different genders. (Is this statement correct?) This variance comes from the inherent interactions featured by the model we chose.

//Rate Ratio, Odds Ratio part

//RR for Poisson, OR for zero-inflated

(Most researchers do report the transformed RR and OR and some of them translate the RR/OR in English, but is this the focus of this paper?) In practice, the differentiation between the rate ratio for Poisson part and odds ratio for zero-inflated part is not always illustrated.

**Risk Ratio:** Risk ratio is the ratio of the probability of an outcome in an exposed group to the probability of an outcome in an unexposed group.

$$RR = \frac{\frac{Event}{Event + No - event_{inexperimentalgroup}}}{\frac{Event}{Event + No - event_{incontrolgroup}}} \quad (17)$$

**Odds Ratio:** Odds ratio approaches to risk ratio if the probability of events is small. (EE and CE is smaller than

EN and CN)

$$OR = \frac{\frac{Event}{No - eventinexperimentalgroup}}{\frac{Event}{No - eventincontrolgroup}} \quad (18)$$

Rate Ratio: Rate ratio is closely related to risk ratio. It is computed by the incidence rate in an exposed group divided by the incidence rate in an unexposed group.

$$IRR = \frac{\frac{Event}{No - eventinexperimentalgroup}}{\frac{Event}{No - eventincontrolgroup}} \quad (19)$$

// Binary Model

(By binary model, do we mean the variable of interest is binary (yes/no) or the binary part of the zero-inflated model? If the latter, then it is included in the above content. If not, we need to find another variable of interest in the ABMRF dataset.)

The binary model is constructed from simulated data. The graph shows that the first difference of predicted binary outcome is not consistent across different x1's.