Chapter 5
Simple Ways to Interpret Effects in Modeling Binary Data

Alan Agresti, Claudia Tarantola, and Roberta Varriale

5.1 Introduction

Suppose you are consulting with a non-statistician colleague in academia, government, or industry, for a study that has a binary response variable. If you use a standard binary regression model such as logistic or probit regression, is your colleague able to understand its natural effect measures, such as odds ratios or probit differences? In our consulting experiences as well as teaching such methods to students in various disciplines, interpretation can be challenging.

Models for binary responses that apply link functions to the probability of "success," such as logistic regression models, are generalized linear models that employ non-linear link functions. With such models, effect parameters are not as simple to interpret as slopes and correlations for ordinary linear regression. This article surveys simple measures that can supplement the ordinary model-based measures, being easier to interpret. Our intention is not to present new methodology but rather to show ways of using existing approaches to supplement the most popular model-based analyses as well as more complex models for binary data.

We consider a binary response variable y taking values 0 and 1 and a set of explanatory variables $(x_1, \ldots, x_p)$, which may be a mixture of quantitative

A. Agresti (✉)
Department of Statistics, University of Florida, Gainesville, FL, USA
e-mail: aa@stat.ufl.edu

C. Tarantola
Department of Economics and Management, University of Pavia, Pavia, Italy
e-mail: claudia.tarantola@unipv.it

R. Varriale
Istat and La Sapienza University, Rome, Italy
e-mail: varriale@istat.it

M. Kateri, I. Moustaki (eds.), *Trends and Challenges in Categorical Data Analysis*, Statistics for Social and Behavioral Sciences,
https://doi.org/10.1007/978-3-031-31186-4_5

and categorical. In describing effect summaries for comparing two groups of a categorical explanatory variable, we sometimes use a separate indicator variable z to distinguish between the groups. The logistic regression model, defined by

$$\log\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = \alpha + \beta z + \beta_1 x_1 + \cdots + \beta_p x_p,$$

is a generalized linear model (GLM) with link function $\text{logit}[P(Y=1)] = \log[P(Y=1)/(1-P(Y=1))]$. This model has effects most naturally interpreted using odds ratios. For example, adjusting for the other explanatory variables, the odds that $y=1$ for the group having $z=1$ divided by the odds that $y=1$ for the group having $z=0$ are

$$\frac{P(Y=1 \mid z=1, x_1, \ldots, x_p)/P(Y=0 \mid z=1, x_1, \ldots, x_p)}{P(Y=1 \mid z=0, x_1, \ldots, x_p)/P(Y=0 \mid z=0, x_1, \ldots, x_p)} = \exp(\beta).$$

The coefficient β_k of x_k is the change in the log odds per each 1-unit increase in x_k, adjusting for the other explanatory variables, so $\exp(\beta_k)$ is a multiplicative effect of each 1-unit increase in x_k on the odds of response $y=1$ versus response $y=0$.

To compare two levels of an explanatory variable such as two groups, however, it is easier for methodologists or practitioners to understand a difference or a ratio of *probabilities* than a ratio of *odds*. In our experience, many (even some statisticians) misinterpret the odds ratio as if it were a ratio of probabilities. When two groups have probabilities close to 0, the ratio of odds is similar to the ratio of probabilities, but this is not true otherwise. In fact, [22] noted that when the probabilities exceed 0.2, the odds ratio is better approximated by the *square* of the ratio of probabilities. For example, if an odds ratio is 9, one group may have success probability merely about 3 times the success probability for the other group.

Other aspects of logistic regression that are due to its nonlinear link function are not as well known to users. For instance, suppose explanatory variables x_1 and x_2 are uncorrelated, such as in many experimental designs. In ordinary linear models, the estimated effect of x_1 is the same when x_1 is the sole predictor as when x_1 and x_2 are joint predictors. For logistic regression, this is not the case with model-based odds effect measures. For instance, the effect β_1^* when x_1 is the sole predictor relates approximately to the effect β_1 when x_2 is also in the model by $\beta_1^* \approx \beta_1\sqrt{3.29/[3.29+\beta_2^2\text{var}(x_2)]}$, where $3.29 = \pi^2/3$ is the variance of the standard logistic distribution [18]. For the model with probit link, $\beta_1^* = \beta_1\sqrt{1/[1+\beta_2^2\text{var}(x_2)]}$. Equality of the effects in the two cases is, however, approximately true for the simpler measures discussed in this article.

The structure of this paper is as follows. In Sect. 5.2, we show that generalized linear models using the identity link function and the log link function, although not as natural for binary data, have simpler summaries and can sometimes supplement logistic and probit models. We illustrate these summary measures with an Italian study to model an employment response variable. In Sect. 5.3, we focus on probit and logit models and we present alternative probability-based summaries that can be used to study the effect of an explanatory variable, while adjusting for other explanatory variables in the model. For group comparisons, these include average differences and average log-ratios of probabilities and comparisons that result directly from corresponding latent variable models. In this section we also show the correspondence between these effect measures obtained for logistic and probit models and the model-based effect measures obtained with the identity and log link functions. We conclude this section illustrating the proposed measures with the Italian study. Section 5.4 uses the measures of Sect. 5.3 to aid in interpreting effects for more complex models, such as generalized additive models. We illustrate with an example about horseshoe crab mating, generalizing existing results for a logistic model.

5.2 Alternative Models for Binary Data

Standard models for binary response variables are special cases of the GLM

$$\text{link}[P(Y = 1)] = \alpha + \beta z + \beta_1 x_1 + \cdots + \beta_p x_p, \tag{5.1}$$

for link functions such as the logit and probit. For describing effects, we find it useful to refer to the model expressed as

$$F^{-1}[P(Y = 1)] = \alpha + \beta z + \beta_1 x_1 + \cdots + \beta_p x_p, \tag{5.2}$$

where the link function F^{-1} is the inverse of a standard cumulative distribution function (cdf). For logistic regression, $F(z) = \exp(z)/[1 + \exp(z)]$ is the standard logistic cdf. For probit regression, F is the standard normal cdf, which we denote by Φ. The nonlinear link function naturally produces effects on the link scale. For example, with the probit link, β is the difference between $F^{-1}[P(Y = 1)]$ when $z = 1$ and when $z = 0$, and β_k is the change in $F^{-1}[P(Y = 1)]$ per each 1-unit increase in x_k, adjusting for the other explanatory variables. Such effect measures are not easy to interpret by those who need to understand the effects in more real-world terms. Although the probit model was the first model for binary data to receive much attention (pre-dating logistic regression by nearly 10 years), its use by methodologists has undoubtedly been hampered by the difficulty of interpretation unless one uses a corresponding latent variable model. The same applies to other link functions that are potentially very useful, such as those with log-log and complementary log-log link functions.

In addition, effects often behave in a way that is counterintuitive to those mainly familiar with ordinary linear models. For example, as mentioned in the introductory section, if an explanatory variable uncorrelated with x_1 is added to a logistic regression model, the partial effect of x_1 is typically different than in the model without the other explanatory variable; it would be identical in an ordinary linear model. For contingency tables, this relates to standard collapsibility results [e.g., 1, pp. 53–54]. For example, consider several 2×2 tables relating binary y to binary x_1 at different levels for x_2. If the difference or the ratio of proportions is the same in each table, then when x_1 and x_2 are marginally independent, the marginal table collapsing over x_2 has the same value for that measure. For the odds ratio, however, collapsibility occurs when x_1 and x_2 are *conditionally* independent, given y, rather than marginally independent. Because of this, regardless of correlation structure among explanatory variables, it can be challenging to compare the effect of an explanatory variable to its effect when other variables are added to the model. Generally, the relation between conditional and marginal effect measures depends on the model and measure considered. For related literature, particularly for logistic regression, see [4, 7, 8, 10, 19], and [20]. Related remarks also occur in comparing effects in marginal models for multivariate responses with effects in corresponding models that add a random effect to the model [1, pp. 495–497].

5.2.1 Identity and Log Link Models for Binary Data

For comparing groups, simple difference and ratio measures on the proportion scale result from alternative link functions in model (5.1). For the identity link function, the coefficient β of an indicator variable in that model is the difference between $P(Y = 1)$ for two groups, adjusting for other variables. The corresponding model is called the *linear probability model*. For the log link function, β is the log ratio of probabilities.

Generalized linear models with identity and log link functions are relatively rarely used for binary data. The link values for the linear probability model are restricted to the [0, 1] range, rather than the entire real line that is the range of linear predictor values in the model. The log-link values are restricted to negative values. Because of these restrictions, ordinary maximum likelihood (ML) fitting of such models, assuming a binomial distribution for the response, may fail. One can always fit the linear probability model using least squares, as in fitting ordinary linear models, but the fitted values may be outside [0, 1] for some values of explanatory variables. When ML works for such a model and it fits the data decently, however, one obtains the advantage of simpler interpretation of effects than in the logistic model.

The appearance of the linear probability model is similar to the logistic and probit models for probabilities between about 0.2 and 0.8. To illustrate, the first panel in Fig. 5.1 shows 500 observations in which X was uniformly distributed over (0, 100), and conditional on $X = x$, $P(Y = 1)$ follows a logistic model with $P(Y = 1)$ increasing from 0.2 to 0.8 over the range of x values. (For clarity of showing the data, the binary observations are jittered slightly.) The figure also shows the ML fits of the logistic and linear probability models. The appearance of the log-link model is similar to the logistic and probit models when probabilities are uniformly less than about 0.25 over the ranges of explanatory-variable values and similar to those models with link applied to $P(Y = 0)$ when probabilities are uniformly above about 0.75. To illustrate, the second panel of Fig. 5.1 shows 500 observations in which X was uniformly distributed over (0, 100), and conditional on $X = x$, $P(Y = 1)$ follows a logistic model with $P(Y = 1)$ increasing from 0.01 to 0.25 over the range of x values. The figure also shows the ML fits of the logistic model and the model with log link.

When we have reason to expect probabilities to fall in the previously specified ranges, we believe that it can be helpful in summarizing the size of an effect to use the models with identity and log link functions, even if only to supplement ordinary logistic and probit models. In addition, the binary models with identity and log links share the property with ordinary linear models that effects remain stable when explanatory variables are added to the model that are uncorrelated with ones already in the model.

5.2.2 Example: Models for Italian Survey Data

In this section, we fit generalized linear models with logit, log and indentity link functions to some data from a simple random sample of about 100,000 Italians from the Toscana region in December 2015. The information comes from administrative sources collected and organized by Istituto Nazionale di Statistica (Istat). Administrative data relevant for the labor statistics derive mainly from social security and fiscal authority and are organized in an information system having a linked employer-employees structure. From this data structure it is possible to obtain information on the statistical unit of interest, i.e., the worker. The response variable y indicates whether the subject is present in any administrative source (1 = yes, 0 = no). Assuming there are no measurement errors, a person not present in an administrative labor source either is not working or is doing so illegally, so in the following we refer to y as whether employed (1 = yes, 0 = no). The examined explanatory variables are x_1 = gender (1 = female, 0 = male), x_2 = Italian (1 if the individual is an Italian citizen, 0 otherwise), and x_3 = pension (1 if the individual is receiving a pension, 0 otherwise). For Istat confidentiality reasons, we cannot report the exact data, but we provide in tables the approximate cross-classified sampled proportions.

We first restrict attention to the 27,775 subjects having age over 65. The sample proportions that were employed ($y = 1$) in the eight cases that cross classify the three explanatory variables were small, so we fitted models both with logit and log links, as shown in Table A1 of the Appendix. The main-effects model fits are

$$\text{logit}[\hat{P}(Y = 1)] = -1.8686 - 1.3236x_1 - 0.4295x_2 + 0.2162x_3$$

and

$$\log[\hat{P}(Y = 1)] = -2.0374 - 1.2388x_1 - 0.3619x_2 + 0.2003x_3$$

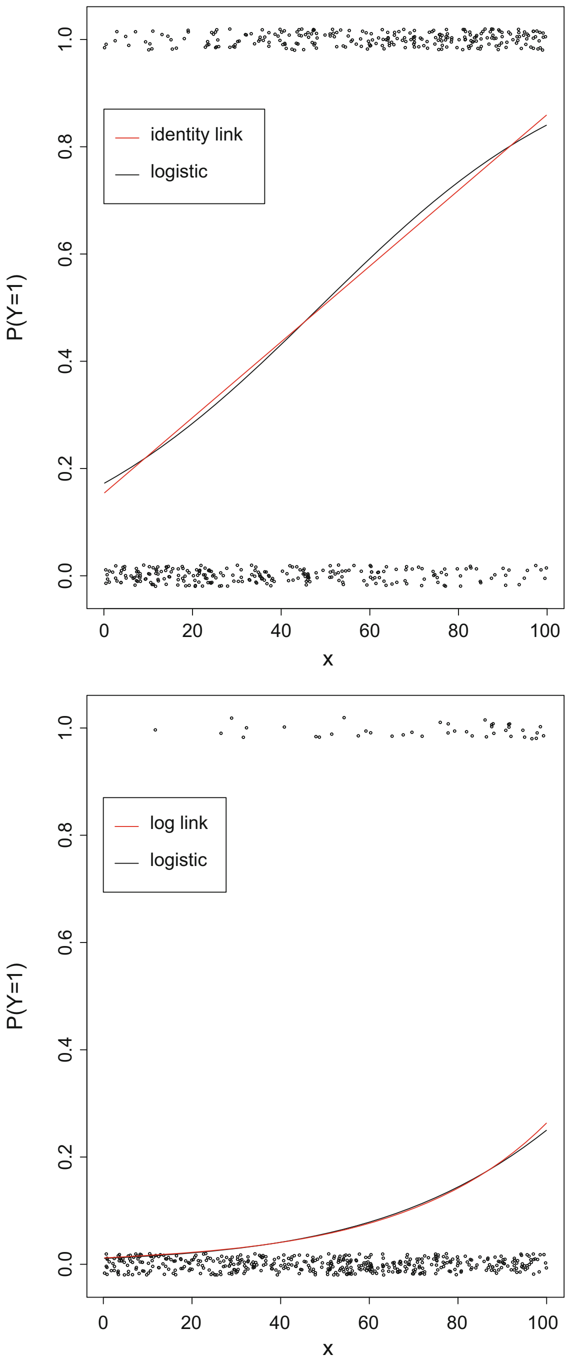

Fig. 5.1 Data sets showing jittered binary data and fits of logistic regression model and (1) linear probability model when $P(Y = 1)$ varies from 0.2 to 0.8, (2) log-link model when $P(Y = 1)$ is less than 0.25

Table 5.1 Fitted values for Istat sample of older subjects, for models with logit and log links for predicting employment using gender (G), Italian (I), and pension (P)

			Main effects		Gender/Italian interaction		
			Logit link	Log link	Logit link	Log link	Sample proportion
G	I	P	$\hat{P}(Y=1)$	$\hat{P}(Y=1)$	$\hat{P}(Y=1)$	$\hat{P}(Y=1)$	(Sample size)
1	1	1	0.0321	0.0321	0.0314	0.0314	0.0316 (13,300)
1	1	0	0.0260	0.0263	0.0255	0.0257	0.0244 (2300)
1	0	1	0.0485	0.0461	0.0865	0.0864	0.0690 (100)
1	0	0	0.0395	0.0378	0.0709	0.0708	0.0812 (200)
0	1	1	0.1109	0.1109	0.1118	0.1118	0.1116 (11,000)
0	1	0	0.0913	0.0908	0.0920	0.0915	0.0949 (600)
0	0	1	0.1608	0.1593	0.1106	0.1111	0.1238 (100)
0	0	0	0.1337	0.1304	0.0911	0.0909	0.0800 (100)

Note: The sample sizes for the sample proportions were not the actual ones used but are rounded to the nearest hundredth, for Istat confidentiality reasons

Table 5.1 shows the fitted values for the two models. They are uniformly very close, with the absolute difference averaged over the 27,775 cases being only 0.000097. The residual deviances (for the grouped data files) are 13.17 and 13.85, with df = 4.

The log-link model has the advantage of simplicity of interpretation, the exponentiated coefficients estimating ratios of probabilities instead of ratios of odds. For instance, adjusting for whether an Italian citizen and whether receiving a pension, the probability that a woman is employed is estimated to be $\exp(-1.2388) = 0.2897$ times the probability that a man is employed.

Table 5.1 also shows sample proportions for the eight cases. Both models show clear lack of fit for the non-Italians, although the sample sizes for those cases are relatively small. In fact, for non-Italians, fitted and sampled values are quite different. Improved fits result from adding an interaction term between gender and whether an Italian citizen to reflect that the gender effect seems to be larger for Italian citizens than for non-citizens. Table 5.1 also shows fitted values for the model with this interaction term, with logit and log links. For this model, fitted values are again uniformly very close, with the absolute difference averaged over the 27,775 cases being only 0.000062. The residual deviances are 1.35 and 1.42 with $df = 3$.

We next consider the 72,225 subjects having age under 65. The sample proportions that were employed ($y = 1$) in the eight cases that cross classify the three explanatory variables fell between 0.20 and 0.75, so we fitted models both with logit and identity links, as shown in Appendix Table A2. The main-effects model fits are

$$\text{logit}[\hat{P}(Y=1)] = 0.3502 - 0.6440x_1 + 0.7017x_2 - 1.8737x_3$$

and

$$\hat{P}(Y = 1) = 0.5876 - 0.1386x_1 + 0.1513x_2 - 0.4078x_3$$

Again, the identity-link model has the advantage of simplicity of interpretation. For instance, adjusting for whether an Italian citizen and whether receiving a pension, the probability that a woman is employed is estimated to be 0.1386 lower than the probability that a man is employed. (Interestingly, the estimated effects of x_2 and x_3 have reverse sign from the estimated effects for the older sample, and the gender effect in the logit model is about half the size.)

Table 5.2 shows the fitted values for the two models and sample proportions for the eight cases. The fits are quite close, with the absolute difference averaged over the 72,225 cases being only 0.00430. These two models show lack of fit for the non-Italians with a pension, although these are only 195 of the 72,225 cases. Improved fits result from adding an interaction term between the Italian citizen and pension variables. The gender main-effect estimate in the identity-link model changes only from −0.1386 to −0.1397. Table 5.2 also shows fitted values for the interaction models with logit and identity links. They are quite close, with the average absolute difference being only 0.00286. The residual deviances are 15.80 and 30.32 with $df = 3$, not particularly large for this enormous sample size.

We do not wish to suggest by these examples that one should *not* use logistic regression. Indeed, an obvious advantage of it compared to the models with log and identity links is that it is relevant regardless of the range of values for $P(Y = 1)$. However, we believe that the log-link model and identity-link model can sometimes supplement the logit-link model, in particular by providing effect interpretations that are simpler for many to understand.

Table 5.2 Fitted values for Istat sample for younger subjects, for models with logit and identity links for predicting employment using gender (G), Italian (I), and pension (P)

			Main effects		Italian/Pension interaction		
			Logit	Identity	Logit	Identity	Sample proportion
G	I	P	$\hat{P}(Y = 1)$	$\hat{P}(Y = 1)$	$\hat{P}(Y = 1)$	$\hat{P}(Y = 1)$	(Sample size)
1	1	1	0.1876	0.1924	0.1845	0.1775	0.1991 (3400)
1	1	0	0.6006	0.6002	0.6011	0.6020	0.5974 (27,700)
1	0	1	0.1027	0.0410	0.2153	0.2119	0.2202 (100)
1	0	0	0.4271	0.4489	0.4243	0.4334	0.4339 (5200)
0	1	1	0.3054	0.3310	0.3012	0.3171	0.2879 (3800)
0	1	0	0.7411	0.7389	0.7416	0.7416	0.7453 (27,500)
0	0	1	0.1789	0.1797	0.3433	0.3516	0.3372 (100)
0	0	0	0.5867	0.5875	0.5840	0.5731	0.5725 (4400)

Note: The sample sizes for the sample proportions were not the actual ones but are rounded to the nearest hundred, for Istat confidentiality reasons

5.3 Alternative Effect Measures for Explanatory Variables

Because of the range restrictions for probabilities, the identity and log links are often not appropriate. But even in fitting a model such as logistic or probit regression, one can construct summary measures based on differences and ratios of probabilities to help others understand the size of the effects. In this section, we describe two types of interpretation that supplement estimated model-parameter effects with simpler effects reported on the probability scale rather than on the scale of the link function. Such effects also exhibit greater stability in terms of the impact of uncorrelated explanatory variables.

When a binary regression model of generalized linear model form contains solely main effects, $P(Y = 1)$ changes monotonically as a quantitative explanatory variables increases, with others at fixed values. This is the situation that we assume in forming these supplementary summary measures.

5.3.1 Probability Effect Measures

A simple summary for the effect of an explanatory variable x_k averages the rate of change in $P(Y = 1)$, as a function of x_k. For this, we consider the expression (5.2) of the model, namely $F^{-1}[P(Y = 1)] = \alpha + \beta z + \beta_1 x_1 + \cdots + \beta_p x_p$. Let $f(y) = \partial F(y)/\partial y$ denote the corresponding probability density function. For a quantitative explanatory variable x_k, the rate of change in $P(Y = 1)$ when other explanatory variables are fixed at certain values $\boldsymbol{x}^*$ is

$$\partial P(Y = 1|\boldsymbol{x} = \boldsymbol{x}^*)/\partial x_k = f(\alpha + \beta z^* + \beta_1 x_1^* + \cdots + \beta_p x_p^*)\beta_k.$$

These measures are denoted in different ways depending on the context; for example, the econometric literature [6] uses the term *elasticity*, while the statistics literature calls them either *marginal effects* or *partial effects*. Long and Mustillo [16] and many others refer to such an instantaneous effect as a *marginal effect*. This terminology is a bit misleading, as this partial derivative refers to a *conditional* effect of x_k rather than its *marginal* effect as the term *marginal* is commonly used (i.e., for a sole predictor, collapsing over the other explanatory variables). Some authors, e.g. [14], instead use the term *partial effect*, which we use in this paper.

For the logit link, the partial effect for x_k on $P(y = 1)$ has the expression

$$\partial P(Y = 1|\boldsymbol{x} = \boldsymbol{x}^*)/\partial x_k = \beta_k P(y = 1|\boldsymbol{x} = \boldsymbol{x}^*)[1 - P(y = 1|\boldsymbol{x} = \boldsymbol{x}^*)].$$

This takes values bounded above by its highest value of $\beta_k/4$ that occurs when $P(Y = 1|\boldsymbol{x} = \boldsymbol{x}^*) = 1/2$. For probit models, the highest value of this instantaneous change is $\beta_k/\sqrt{2\pi}$, also when $P(Y = 1|\boldsymbol{x} = \boldsymbol{x}^*) = 1/2$. These maximum values need not be relevant, as $P(Y = 1)$ need not be near 1/2 for most or all the data.

Any particular way of fixing values of the explanatory variables has its corresponding partial effect value for x_k. Long and Mustillo [16] summarize various versions. Here, we mainly consider the *average partial effect*, which estimates the partial effect of x_k at each of the n sample values of the explanatory variables, and then averages them. We could instead estimate the partial effect with every explanatory variable, including x_k, set at its mean, which is the *partial effect at the mean*. Or, we could set all explanatory variables at values considered to be of particular interest. This might be more appropriate if the sample is not random or not representative of the population of interest, in which case it is sometimes referred to as a *partial effect at a representative value*. For each version, the summary value obtained still reflects the effect of x_k adjusting for the explanatory variables, unlike an averaging of the effect over values of a random effect in a generalized linear mixed model, in which case the effect changes nature to being population-averaged and can have quite different magnitude.

For a categorical explanatory variable, for each version one would instead use a *discrete change*, estimating the change in $P(Y = 1)$ for a change in an indicator variable. To compare two groups, for instance, for the n sample observations, we could find the difference between estimates of $P(Y = 1)$ when $z = 1$ and when $z = 0$ at the sample values for the other predictors and average the obtained values. When the number of possible values of the categorical explanatory variable is greater than two, the discrete change is computed as the difference in the predicted probabilities for cases in one category relative to the reference level.

Discrete changes are also relevant for quantitative explanatory variables, to summarize estimated changes in $P(Y = 1)$ over a particular range of x_k values. For example, to summarize the effect of a quantitative variable x_k on y, it can be useful to report the difference between the model-fitted estimate of $P(Y = 1)$ at the maximum and minimum values of x_k, when other explanatory variables are set at particular values such as their means. A caveat for such measures is that their relevance depends on the plausibility of x_k taking extreme values when all other explanatory variables fall at their means. Also, this summary can be misleading when outliers exist on x_k, in which case one can instead report the estimated probabilities at more resistant quantiles. Reporting them at the upper and lower quartiles of x_k summarizes the estimated change in $P(Y = 1)$ over the range of the middle half of the observations on x_k, with other explanatory variables fixed. Such a measure has greater scope for reflecting reality.

A useful and easy-to-obtain measure that we've not seen proposed for the two-group comparison focuses on *ratios* of estimated probabilities for the two groups. For example, we could average the n log-ratios of probability estimates, to obtain a measure comparable to the effect in the log-link GLM, and then exponentiate that average for interpretive purposes. Again, other versions are possible, such as finding the ratio at the mean of the other explanatory variables. Such measures would seem to be especially useful when fitted probabilities are near 0 for the groups being compared.

Greene [9, pp. 775–785] showed how to obtain standard errors for the maximum likelihood estimators of some effect measures based on instantaneous rates of change and differences of probabilities. We have used the bootstrap to obtain a standard error (SE) for the log-ratio measure just proposed. Mood [18] pointed out that the average partial effect has behavior reminiscent of effects in ordinary linear models, in the sense that it is roughly stable when we add an explanatory variable to the model that is uncorrelated with the variable for which we are describing the effect. Such behavior is expected, as such an average partial effect typically takes similar value as the effect using the linear probability model discussed in Sect. 5.2. The effect measures are available in software, such as presented by Leeper [13], Long and Freese [15, pp. 341–351], and Sun [21, pp. 527–531]. Agresti and Tarantola [3] and Iannario and Tarantola [12] proposed analogous measures for modeling ordinal data.

5.3.2 A Probability Summary for Ordered Comparison of Groups

It is sometimes realistic to regard a categorical variable as crude measurement of an underlying continuous latent variable y^* that, if we could observe it, would be the response variable for an ordinary linear model. In fact, model (5.2) is implied by a model in which a latent response has conditional distribution with standard cdf given by the inverse of the link function [1, p. 252]. We next use this connection to suggest an alternative way to summarize an effect, in the context of comparing two groups ($z = 0$ and $z = 1$). Let y_1^* and y_2^* denote independent underlying latent variables for the binary response, representing the underlying distributions when $z = 1$ and when $z = 0$ respectively. At a particular setting $\boldsymbol{x}$ for other explanatory variables, $P(Y_1^* > Y_2^*; \boldsymbol{x})$ is a summary measure of relative size, suggested by Agresti and Kateri [2] for ordinal response variables.

The normal latent variable model with $y^* \sim N(\beta z + \beta_1 x_1 + \cdots + \beta_p x_p, 1)$ implies the probit model

$$\Phi^{-1}[P(Y = 1)] = \alpha + \beta z + \beta_1 x_1 + \cdots + \beta_p x_p,$$

with α the cutpoint on the underlying scale between y^* values for which $y = 1$ and for which $y = 0$. For this model,

$$P(Y_1^* > Y_2^*; \boldsymbol{x}) = P\left[\frac{(y_1^* - y_2^*) - \beta}{\sqrt{2}} > \frac{-\beta}{\sqrt{2}}\right] = \Phi\left(\frac{\beta}{\sqrt{2}}\right). \tag{5.3}$$

This is true regardless of the $\boldsymbol{x}$ value, so we denote it by $P(Y_1^* > Y_2^*)$. For the logit link,

$$P(Y_1^* > Y_2^*) \approx \frac{\exp(\beta/\sqrt{2})}{[1 + \exp(\beta/\sqrt{2})]}, \tag{5.4}$$

for the β coefficient of z in the logistic model.

When the latent variable model for binary data is realistic, this type of probability comparison of the groups supplements the ordinary interpretation of the effect coefficient β. As β increases from 0, the probability increases from 0.5 toward 1. In addition, a natural way to construct a summary measure of predictive power is to estimate R^2 for the linear model that is specified for the underlying latent response variable. McKelvey and Zavoina [17] suggested this measure for a probit model for ordinal responses, for which the underlying latent variable model is the ordinary normal linear model, but it applies also for binary data and for other link functions for ordinal data [3].

5.3.3 Example: Measures for Italian Survey Data

We illustrate these measures for the example from earlier in this article of modeling Italian employment status. For simplicity, here we consider only the main effects models.

The average partial effect for a logistic model approximates the corresponding effect from the binary model with identity link. For the younger age group, we obtained the gender effect of -0.1386 ($SE = 0.0035$) with the identity-link model. The estimated average partial effect for the model with logit link is -0.1409 ($SE = 0.0035$). This can be easily found with an existing package in `R` applied to the ungrouped data file, with code such shown in Table A2 in the Appendix.

The average partial effect for log-ratios that we suggested for a logistic model approximates the corresponding effect from the binary model with log link. For the older age group the gender effect estimate is equal to -1.2388 ($SE = 0.0516$) with the log-link model. The estimated average log-ratio partial effect for the model with logit link is -1.2398 ($SE = 0.0517$). Table A3 in the Appendix presents edited `R` code for obtaining the estimated average log-ratio partial effect and for using the bootstrap with 1000 resamplings of the data to obtain its SE. As one can do with the log-link model parameter estimate, one could utilize the asymptotic normality of the sample measure to obtain a correponding confidence interval for the population value, such as the 95% confidence interval $-1.2398 \pm 1.96(0.0517)$, which is $(-1.341, -1.138)$. The exponentiated endpoints of the interval, that is (0.26, 0.32), are a confidence interval on the probability-ratio scale. (Recall that the log-link model provided ML estimate 0.2897.) Alternatively, one can find a bootstrap confidence interval, such as shown with the percentile method in Table A3.

Whether a latent variable is sensible for measuring propensity toward employment is debatable. But if so, from Eq. (5.4) with the estimated gender effect $\hat{\beta} = -1.3236$ for the older sample, the estimated probability that a randomly selected female would be higher on the latent variable than a randomly selected male is $\exp(-1.3236/\sqrt{2})/[1 + \exp(-1.3236/\sqrt{2})] = 0.282$. For the younger sample, the effect is $\exp(-0.6440/\sqrt{2})/[1 + \exp(-0.6440/\sqrt{2})] = 0.388$.

5.4 Generalized Additive Model for Binary Data

A generalized additive model (GAM) replaces the linear predictor in a binary generalized linear model (GLM) by additive unspecified smooth functions. Its basic version has the form

$$\text{link}[P(y = 1)] = s_1(x_1) + \cdots + s_j(x_j) + \cdots + s_p(x_p),$$

where the smooth function s_j is typically based on cubic splines [11] and more generally uses basis expansions of low rank with complexity controlled by ridge penalties on regression coefficients [e.g. 23]. The name *additive* derives from the additive structure of the predictor. GAMs have the advantage over GLMs of greater flexibility, with an ordinary GLM with $s_k(x_k)$ replaced by $\beta_k x_k$. In practical application, it is often helpful to use both smooth and linear terms in a model. Using a graphical portrayal of a GAM fit, we may discover patterns that we would miss with ordinary GLMs, and we may obtain potentially better estimates of mean responses. A disadvantage of GAMs and other smoothing methods, compared with GLMs, is that interpretability is even more difficult. It can be more difficult to summarize an effect and judge when it has substantive importance.

Fasiolo et al. [5] described an efficient visual method for interpreting GAMs, using the `mcgViz` package in `R`. The proposed methods include ones to bin the data and summarize them in a form that can be displayed effectively, interactive Q-Q plots, portrayals of conditional residuals, and visualizations of the uncertainty of the fitted smooth effects. To supplement these with simple numerical summaries, we believe that measures that aid in interpreting binary GLMs can also be useful for GAMs. When an effect of a quantitative explanatory variable seems to be monotonic and not highly variable in the degree of non-linearity, useful measures include measures of average partial rates of change of probabilities and comparisons of the fitted probability at extreme values or quartiles or other quantiles of the explanatory variable.

How does one describe quantitatively the effect of an explanatory variable or obtain a confidence intervals for the true effect? Here, we suggest a way to construct an estimated average partial effect using the fit of a GAM. For explanatory variable k, let $\boldsymbol{x}_{i(k)}$ denote the values of the other explanatory variables for observation i. The fitted rate of change for explanatory variable x_k for observation i can be approximated by

$$[\hat{P}(y = 1 \mid \boldsymbol{x}_{i(k)}, x_{ik} + \epsilon) - \hat{P}(y = 1 \mid \boldsymbol{x}_{i(k)}, x_{ik} - \epsilon)]/2\epsilon$$

for a very small ϵ. Finding the mean of these values for the n observations yields an approximate average partial effect for that predictor. We suggest starting with a trial value such as $\epsilon = 0.000001$ and then using a smaller value yet to ensure that results are stable to several decimal places. For comparing two groups, one could find an average difference or average ratio of estimated probabilities at the n values of the explanatory variables.

5.4.1 Example: GAM for Horseshoe Crab Study

We illustrate the use of the average partial effect in the context of GAMs with a data set analyzed extensively with logistic regression in [1], from a study of nesting horseshoe crabs. During spawning season, a female migrates to the shore to breed. With a male attached to her spine, she lays clusters of eggs, which are fertilized externally. During spawning, other male crabs, called *satellites*, may cluster around the pair and fertilize the eggs. The response outcome for each female crab is whether she had any satellites (1 = yes, 0 = no). Explanatory variables associated with this response were the female crab's carapace (shell) width, which is a summary of her size, and her color (four categories from light to dark), which is a surrogate for the crab's age, older crabs being darker. In the sample, width had a mean of 26.3 cm and a standard deviation of 2.1 cm. Logistic modeling showed that width had a positive effect on the presence of a satellite, and color being dark (category 4) had a negative effect.

The logistic ML fit with predictors width and an indicator for color that is 1 for dark-colored crabs and 0 for others is

$$\text{logit}[\hat{P}(Y = 1)] = -11.6790 + 0.4782(\text{width}) - 1.3005(\text{color}),$$

with standard errors of 0.104 for width and 0.526 for color. Table A4 in the appendix shows edited results for the logistic regression and a GAM fit with these data, which is

$$\text{logit}[\hat{P}(Y = 1)] = -11.2470 + s(\text{width}) - 1.2805(\text{color}),$$

s(width) being a smoothing spline. Figure 5.2 shows the GAM fit, with jittered observations. Adding an interaction term does not provide a significantly improved fit.

Table A5 in the Appendix shows edited R code for finding the average partial effects for width and for color for this GAM as well as for the corresponding logistic model. Interpretation is relatively simple. For the logistic fit at the $n = 173$ observed width values, the average rate of change is 0.087 in the estimated probability of a satellite per 1 cm increase in width, adjusting for color. At those width values, the estimated probability of a satellite averages 0.261 lower if the crab has dark color than if it has a lighter color. For the GAM, the corresponding values are 0.085 (standard error = 0.015) and 0.254 (standard error = 0.112), quite similar because the logistic model fits relatively well.

Table A6 in the Appendix shows edited R code for using the bootstrap with 1000 resamplings of the data to obtain standard errors and confidence intervals for the

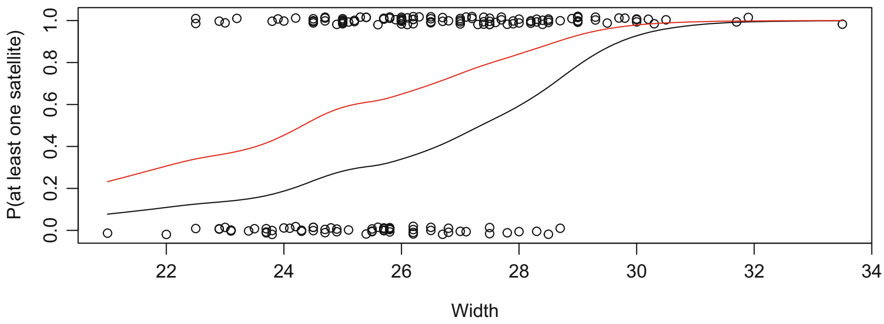

Fig. 5.2 Portrayal of GAM fit for the effects of width and color (black for dark, red for other colors) on jittered responses for whether a female horseshoe crab has at least one satellite

average partial effects in the GAM. For example, the bias-corrected and accelerated (BC_a) confidence interval of $(-0.466, -0.028)$ for the average partial effect for color indicates that at the sampled width values, the probability of a satellite is estimated to average between 0.028 and 0.466 lower if the crab has dark color than if it has a lighter color. The relatively wide interval reflects partly that the sample had only 22 dark-colored crabs.

5.5 Discussion and Future Research

Future research could apply the methods of this paper to other models for binary responses. In particular, using alternative link functions to aid in interpretation would be useful for marginal models, whether fitted by GEE methods or maximum likelihood. The binary and log links are more challenging for random effects models, as the usual assumption of normally-distributed random effects adds another restriction to models with bounded range values. Effect measures such as average partial effects are also relevant for models for multi-category responses. See [3] for their use with cumulative link models for ordinal responses.

Perhaps more challenging for future research is the development of effect measures for generalized additive models. The average partial effect measure presented in this article is of use when relationships are monotone, but often that is not the case. Even when it is the case, difference or ratio effects are sometimes highly variable across the range of an explanatory variable, and a single summary may be too simplistic. Also for the binary generalized linear models considered here, we assumed that $P(Y = 1)$ is monotone in quantitative explanatory variables, and alternative measures are needed when this is not the case.

In summary, in these days in which statistical science is ever more visible, partly because of the emergence of data science and methods for “big data,” it is increasingly important for statisticians to develop ways to present relatively simple summaries of complex methods that will be understandable by a relatively wide audience. We hope that this paper is a step in that direction.

Acknowledgments The authors appreciate helpful comments from two referees and from Pablo Inchausti and Maria Kateri.

Appendix

This appendix provides the source code for the R analyses described in the text.

Table A1 R code for fitting logistic and log-link models to the older-age Istat sample

```
--------------------------------------------------------------------------------
>Italian1 <- read.csv("http://www.stat.ufl.edu/~aa/cat/data/Italian_older.csv",
+                      header=TRUE)
>mod.logit <- glm(empl ~ female + italian + pension, family=binomial,
+                 data=Italian1)
>summary(mod.logit) # fit of logistic model; default link is logit
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.8686     0.1631  -11.46   <2e-16
female       -1.3236     0.0546  -24.26   <2e-16
italian      -0.4295     0.1632   -2.63   0.0085
pension       0.2162     0.0948    2.28   0.0225
---
>mod.log <- glm(empl ~ female + italian + pension, family=binomial(link=log),
+               data=Italian1)
>summary(mod.log) # fit of model with log link function
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0374     0.1465  -13.91   <2e-16
female       -1.2388     0.0516  -24.00   <2e-16
italian      -0.3619     0.1460   -2.48    0.013
pension       0.2003     0.0885    2.26    0.024
--------------------------------------------------------------------------------
```

Table A2 R code for fitting logistic and linear probability models to the younger-age Istat sample and finding the average partial effect for the logistic regression model

```
--------------------------------------------------------------------------------
>Italian2 <- read.csv("http://www.stat.ufl.edu/~aa/cat/data/Italian_younger.csv",
+                      header=TRUE)
>mod.logit <- glm(empl ~ female + italian + pension, family=binomial,
+                 data=Italian2)
>summary(mod.logit)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3502     0.0224    15.6   <2e-16
female       -0.6440     0.0161   -39.9   <2e-16
italian       0.7017     0.0225    31.2   <2e-16
pension      -1.8737     0.0288   -65.1   <2e-16
---
> mod.linprob <- glm(empl ~ female + italian + pension,
               family=quasi(link=identity, variance="mu(1-mu)"), data=Italian2)
>summary(mod.linprob) # fit of linear probability model
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5876    0.0052   112.4   <2e-16
female      -0.1386    0.0035   -40.1   <2e-16
italian      0.1513    0.0052    29.1   <2e-16
pension     -0.4078    0.0052   -78.4   <2e-16
---
>library(mfx)
>logitmfx(mod.logit, atmean=FALSE, data=Italian2)
Marginal Effects:
           dF/dx Std. Err.     z  P>|z|
female  -0.14062   0.00346 -40.6 <2e-16
italian  0.15820   0.00512  30.9 <2e-16
pension -0.41602   0.00508 -81.9 <2e-16
--------------------------------------------------------------------------------
```

Table A3 R code for finding average log-ratio partial effect and bootstrap SE and bootstrap CI for the logistic regression model applied to the older-age Istat sample

```
--------------------------------------------------------------------------------
>library(plyr)
>library(boot)
>attach(Italian1)
---
 APER.log<-function(formula, data, indices, fam, var_exp)
{
dat<-data[indices,]
mod <- glm(formula, family=fam, data=dat)
pred.prob <- (predict(mod,type="response"))
var_exp_ind<-var_exp[indices,]
r_new<- as.data.frame(cbind(var_exp_ind, pred.prob))
r1_new<-count(r_new, vars = c(names(r_new)))
row_r1<-nrow(r1_new)
pred.prob.Male_new<-r1_new$pred.prob[1:(row_r1/2)]
pred.prob.Female_new<-r1_new$pred.prob[((row_r1/2)+1):row_r1]
r2_new <- count(var_exp_ind[,-1],vars = c(names(var_exp_ind)[-1]))
APER.log_new <- ((log(pred.prob.Female_new/pred.prob.Male_new)%*%r2_new$freq)
+                 /sum(r2_new$freq))
return(APER.log_new)
}
APER.log(formula=empl ~ female + italian + pension, data=Italian1, indices=
+     c(1:nrow(Italian1)), fam=binomial,var_exp=cbind(female, italian, pension))
        [,1]
[1,] -1.2398
---
APER.log_boot <- boot(data=Italian1, statistic=APER.log, R=1000,
    formula=empl ~ female + italian + pension, fam=binomial,
    var_exp=cbind(female, italian, pension) )
> APER.log_boot
Bootstrap Statistics :
    original      bias    std. error
t1*  -1.2398 -0.00039268    0.051689
---
> boot.ci(APER.log_boot,type="perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates
Intervals :
Level     Percentile
95%   (-1.344, -1.141 )
--------------------------------------------------------------------------------
```

Table A4 `R` code for GAM fit for using width and color as predictors of whether a female horseshoe crab has any satellites

```
-------------------------------------------------------------------------------
>Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",
+                      header=TRUE)
>Crabs$c4 <- ifelse(Crabs$color == 4, 1, 0) # indicator for color cat. 4
>fit.glm <- glm(y ~ width + c4, family=binomial, data=Crabs)
>summary(fit.glm)
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.6790     2.6925  -4.338 1.44e-05
width         0.4782     0.1041   4.592 4.39e-06
c4           -1.3005     0.5259  -2.473   0.0134
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.96  on 170  degrees of freedom
---
>library(gam)
>fit.gam <- gam(y ~ s(width) + c4, family=binomial, data=Crabs)
>summary(fit.gam)
    Null Deviance: 225.7585 on 172 degrees of freedom
Residual Deviance: 185.4678 on 167.0001 degrees of freedom
Anova for Parametric Effects
           Df  Sum Sq Mean Sq F value   Pr(>F)
s(width)    1  17.774 17.7736 18.3127 3.15e-05
c4          1   5.928  5.9278  6.1076  0.01446
Residuals 167 162.084  0.9706
>fit.gam$coefficients
(Intercept)      c4
-11.2470    -1.2805
-------------------------------------------------------------------------------
```

Table A5 `R` code for finding the average partial effects for width and for color for the logistic regression model and for the generalized additive model for the presence of horseshoe crab satellites

```
-------------------------------------------------------------------------------
> Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",
                       header=TRUE)
> Crabs$c4 <- ifelse(Crabs$color == 4, 1, 0) # indicator for dark color
> fit <- glm(y ~ width + c4, family=binomial, data=Crabs)
> library(mfx)
> logitmfx(fit, atmean=FALSE, data=Crabs) # with atmean=TRUE, finds
Marginal Effects:                          # effect only at the mean
          dF/dx  Std. Err.         z    P>|z|
width   0.08748    0.02447   3.5748  0.00035
c4     -0.26142    0.10569  -2.4735  0.01338
---
dF/dx is for discrete change for the following variables: "c4"

#Function to obtain Average Partial Effects in a GAM model
APE_GAM<-function(formula,data,indices, pvar1,pvar2, fam,epsilon){
   d <- data[indices,]
   fit <- gam(formula,family=fam, data=d)
   data_plus <- data
   data_minus <- data
   data_plus[,pvar1] <- data[,pvar1] + epsilon
   data_minus[,pvar1] <- data[,pvar1] - epsilon
   data1_plus <- data_plus[,c(pvar1, pvar2)]
   data1_minus <- data_minus[,c(pvar1, pvar2)]
   tvec <- (predict(fit, data1_plus, type="response")
              - predict(fit,data1_minus, type="response"))/(2*epsilon)
   APE <- mean(tvec)
   return(APE)
}

# APE for width
> APE_GAM(formula = y ~ s(width) + c4, data=Crabs, indices=c(1:nrow(Crabs)),
+         pvar1=5,  pvar2=8, fam=binomial, epsilon=0.000001)
[1] 0.0850665

# Function to obtain a discrete change in a GAM model
dchange_GAM <-function(formula, data, indices, pvar1, pvar2, fam,epsilon){
   d <- data[indices,]
   fit <- gam(formula,family=fam, data=d)
   data_plus <- data
   data_minus <- data
   data_plus[,pvar2] <- 1
   data_minus[,pvar2] <- 0
   data1_plus <- data_plus[,c(pvar1, pvar2)]
   data1_minus <- data_minus[,c(pvar1, pvar2)]
   tvec <- (predict(fit, data1_plus, type="response")
                 - predict(fit, data1_minus, type="response"))
   APE <- mean(tvec)
   return(APE)
   }

dchange_GAM (formula = y ~ s(width) + c4,data=Crabs,indices=c(1:173),
             pvar1=5,  pvar2=8, fam=binomial,epsilon=0.000001)
[1] -0.2539021
-------------------------------------------------------------------------------
```

Table A6 R code for using a bootstrap to find confidence intervals for the average partial effects for width and for color for the generalized additive model for the presence of horseshoe crab satellites

```
-------------------------------------------------------------------------------
#width variable
>  APE_boot <- boot(data=Crabs, statistic=APE_GAM, R=1000, formula =
+          y ~ s(width) + c4, pvar1=5, pvar2=8, fam=binomial, epsilon=0.000001)
> APE_boot
Bootstrap Statistics :
     original        bias    std. error
t1* 0.0850665 -0.0007855886  0.01513634
>  boot.ci(APE_boot, type="bca")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates
Level       BCa
95%   ( 0.0536,  0.1122 )

#color variable
> disc_boot_1 <- boot(data=Crabs, statistic=dchange_GAM, R=1000, formula =
+        y ~ s(width) + c4, pvar1=5, pvar2=8, fam=binomial, epsilon=0.000001)
> disc_boot_1
Bootstrap Statistics :
      original       bias    std. error
t1* -0.2539021 -0.002005479   0.1118384

 boot.ci(disc_boot_1, type="bca")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates
Level       BCa
95%   (-0.4658, -0.0285 )
-------------------------------------------------------------------------------
```

References

1. Agresti, A.: Categorical Data Analysis, 3rd edn. Wiley, Hoboken (2013)
2. Agresti, A., Kateri, M.: Ordinal probability effect measures for group comparisons in multinomial cumulative link models. Biometrics **73**, 214–219 (2017)
3. Agresti, A., Tarantola, C.: Simple ways to interpret effects in modeling ordinal categorical data. Statistica Neerlandica **72**, 210–223 (2018)
4. Brumback, B., Berg, A.: On effect-measure modification: relationships among changes in the relative risk, odds ratio, and risk difference. Stat. Med. **27**, 3453–3465 (2008)
5. Fasiolo, M., Nedellec, R., Goude, Y., Wood, S.N.: Scalable visualisation methods for modern generalized additive models. J. Comput. Graph. Stat. **29**, 78–86 (2020)
6. Franses P.H., Paap, R.: Quantitative Models in Marketing Research. Cambridge University Press, Cambridge (2004)
7. Gail, M.H.: Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In: Moolgavkar, S.H., Prentice, R.L. (eds.) Modern Statistical Methods in Chronic Disease Epidemiology, pp. 3–18. Wiley, New York (1986)
8. Gail, M.H., Wieand, S., Piantadosi, S.: Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika **71**, 431–444 (1984)

9. Greene, W.H.: Econometric Analysis, 6th edn. Pearson Prentice Hall, Upper Saddle River (2008)
10. Greenland, S., Robins, J.M., Pearl, J. Confounding and collapsibility in causal inference. Stat. Sci. **14**, 29–46 (1999)
11. Hastie, T., Tibshirani, R.: Generalized Additive Models. Chapman and Hall, London (1990)
12. Iannario, M., Tarantola, C.: How to interpret the effect of covariates on the extreme categories in ordinal data models. Sociol. Methods Res. (2021). https://doi.org/10.1177/0049124120986179
13. Leeper, T.J.: Margins: An R Port of Stata's Margins Command. R package version 0.3.23 (2018)
14. Long, J.S.: Regression Models for Categorical and Limited Dependent Variables. Sage, Thousand Oaks (1997)
15. Long, J.S., Freese, J.: Regression Models for Categorical Dependent Variables Using Stata, 3rd edn. Stata Press, College Station (2014)
16. Long, J.S., Mustillo, S.A.: Using predictions and marginal effects to compare groups in regression models for binary outcomes. Sociol. Methods Res. **50**, 1284–1320 (2021)
17. McKelvey, R.D., Zavoina, W.: A statistical model for the analysis of ordinal level dependent variables. J. Math. Sociol. **4**, 103–120 (1975)
18. Mood, C.: Logistic regression: Why we cannot do what we think we can do, and what we can do about it. Eur. Sociol. Rev. **26**, 67–82 (2010)
19. Robinson, L., Jewell, N.P.: Some surprising results about covariate adjustment in logistic regression models. Int. Stat. Rev. **59**, 227–240 (1991)
20. Stanghellini, E., Doretti, M.: On marginal and conditional parameters in logistic regression models. Biometrika **106**, 732–739 (2019)
21. Sun, C.: Empirical Research in Economics: Growing up with R. Pine Square LLC, Baton Rouge (2015)
22. VanderWeele, T.J.: Optimal approximate conversions of odds ratios and hazard ratios to risk ratios. Biometrics **76**, 746–752 (2020)
23. Wood, S.N.: Generalized Additive Models: An Introduction with R, 2nd edn. CRC Press, Boca Raton (2017)