Problems with using odds ratios as effect sizes in binary logistic regression and alternative approaches

James O. Uanhoro[1], Yixi Wang[1], & Ann A. O'Connell[1]

[1] Ohio State University

Abstract

The standard regression technique for modeling binary response variables in education research is logistic regression. The odds ratios from these models are used to quantify and communicate variable effects. These effects are sometimes pooled together as in a meta-analysis. We argue that this process is problematic as odds ratios calculated from different models and studies are not directly comparable. As an alternative, we recommend the linear probability model for computing risk differences and a Poisson working model for computing risk ratios. These effect sizes are comparable across models and studies. However, standard approaches for estimating these models have their problems, hence, we motivate and present modified estimation techniques for estimating these models that mitigate these problems.

Keywords: effect size, odds ratio, risk difference, risk ratio, linear probability model, Poisson working model

When applied researchers intend to model binary response variables, they often use the logistic regression model. Although there are a variety of approaches for modeling binary response variables, logistic regression is flexible in that it allows for binary, categorical and continuous predictor types. Researchers have applied logistic regression to study a wide variety of questions; examples include degree completion (e.g. Carpenter II, Kaka, Tygret, & Cathcart, 2018; Cox, Reason, Nix, & Gillman, 2016) and college attendance (e.g. Klevan,

The Version of Record of this manuscript has been published and is available in Journal of Experimental Education (25 November, 2019) http://www.tandfonline.com/10.1080/00220973.2019.1693328.

Correspondence concerning this article should be addressed to James O. Uanhoro, Ramseyer Hall, 29 W Woodruff Ave, Columbus, OH 43210. E-mail: uanhoro.1@osu.edu

Weinberg, & Middleton, 2016; Eccles, Vida, & Barber, 2004). The coefficients one obtains from the logistic regression model are logits or log odds, and there are challenges interpreting them substantively. To ease interpretation of the coefficients, it is standard practice to exponentiate them to obtain odds ratios. Our argument is that odds ratios are flawed for quantifying and documenting the magnitude of the effect of predictor variables across studies. And documenting the effect of predictor variables is important for a cumulative science. Hence, our study has impacts for meta-analysis of odds ratios.

In the remainder of this paper, we will:

- Make the case that the process of interpreting logits and odds ratios as variable effects is flawed for establishing a record of variable effects.
- Recommend regression models that estimate alternative effect sizes that do not suffer the same flaws.
- Identify problems with these models, recommend methods to mitigate these problems, and evaluate these methods.
- And finally, demonstrate these alternative models using a real data example.

The problem with odds ratios obtained from the logistic regression model

Before we make the case that the odds ratio obtained from the logistic regression model is problematic for quantifying predictor variable effects, we emphasize one feature of ordinary least squares (OLS) estimation of the linear regression model. It is an established result that when a predictor variable in a regression model is uncorrelated with other determinants of an outcome, its coefficient is an unbiased estimate of the predictor's effect on the outcome (Wooldridge, 2010, p. 49). The easiest way to find such a predictor variable is a variable indicating random assignment of individuals to treatment and control groups. It is for this reason that OLS or its specialized forms such as the Student's t-test can be used to obtain causal effects of a treatment when individuals are randomly assigned to treatment and control groups. Our argument is not about causal analysis but this property of OLS explains why the regression coefficient of such a variable continues to be the simple mean difference in the presence of additional predictors, control variables or covariates, with slight fluctuations.[1] This feature is highly desirable because it allows us to estimate the unconfounded effect of manipulated variables on outcomes of interest with simple study designs like two-group randomized experiments.

Most applied statistical analyses do not involve such random variables as previously

[1] There are some exemptions to this rule; including an interaction created from the variable or a mediator in the model will cause the coefficient of the variable to change.

described. This is why in practice, researchers attempt to control for the effects of other predictors to arrive at better estimates of the unconfounded relation between predictor variables of interest and the outcome. Since most predictors are correlated to some degree, it is common for regression coefficients to change from one model to the next. In fact, it is common for researchers to focus on the change in the regression coefficients of predictor variables of interest from model to model, interpreting changes as a function of which predictor variables are included in the regression model. A simple example is the examination of the effect of elite college attendance on education and career achievement. Brand and Halaby (2006) found that "introducing controls for pre-college academic and family background dramatically reduces the magnitude of the coefficients (the elite college effect), although some are still statistically significant." This change in the magnitude of coefficients only happened because elite college attendance was related with some of these pre-college academic and family background variables, and these background variables were predictive of the education and career outcomes under study. Were elite college attendance independent of these background characteristics, the estimated elite college effect in OLS regression would remain the same with or without controlling for the background variables.

We now return to the logistic regression model and its behavior in the case of a predictor that is unrelated to other determinants of the outcome. The major flaw we intend to point out is that this desirable property of OLS does not apply to the logistic regression model. In fact, as an investigator includes more variables in a model that are predictive of an outcome, the absolute magnitude of the coefficient of a variable unrelated to other determinants of the outcome will keep increasing. The problem goes under different names in different literatures: *neglected heterogeneity* (econometrics; Wooldridge, 2010, p. 470), *unobserved heterogeneity* (sociology; Mood, 2010); *non-collapsibility* (biostatistics; Greenland, Robins, & Pearl, 1999) and *omitted covariates* (biostatistics; Gail, Wieand, & Piantadosi, 1984).

The exposition of the problem by Mood (2010) is easy to follow and we adopt it here. First, we begin with the *latent variable formulation* (Amemiya, 1981; Long, 1997, p. 40) for the logistic regression model. Consider the following data generation process for a binary response variable, y: $y = \mathbf{1}(y^* > c)$, where $\mathbf{1}()$ is the indicator function, y^* is a continuous variable, and c is a cut-off. All values of y^* above the c are a 1 on y, and the others are 0 on y. A simple example would be y^* as test scores, and all participants with scores above a cutoff, c, pass (a 1 on y), and other participants fail (a 0 on y). We make some simplifying assumptions: y^* has a range of negative infinity to infinity, and c is 0. Additionally, the *true* equation for y^* or the model on the logit scale is:

$$y^* = b_0 + b_1 \times x_1 + b_2 \times x_2 + u \tag{1}$$

where b_0 is the intercept, x_1 and x_2 are predictor variables, and b_1 and b_2 are their respective coefficients. u is the error term assumed to be logistic with mean 0 and a variance that is $s^2 \times \pi^2/3$, where s is any positive number[2] − if s is small, the error variance is very small, and vice-versa. We assume x_1, x_2 and u are uncorrelated. Finally, y^* is latent, only y is available. Then logistic regression is the correct model for y, and logistic regression is an attempt to recover the model on the logit scale in equation (1).

During estimation of logistic regression, the error term is always assumed to have a mean of 0, but a variance of $\pi^2/3$. Hence, we re-write u as a function of υ, $u = s \times \upsilon$, where υ is logistic with mean 0 and variance $\pi^2/3$. Thus the logistic regression model that is estimated is:

$$\frac{y^*}{s} = \frac{b_0}{s} + \frac{b_1}{s}x_1 + \frac{b_2}{s}x_2 + \frac{s}{s}\upsilon \tag{2}$$

For now, we assume $u = \upsilon$ implying $s = 1$. In that situation, s vanishes from equation (2). However, should we exclude x_2 from the model, x_2 gets absorbed into the error such that the error becomes $b_2x_2 + \upsilon$. s then becomes:

$$s = \sqrt{\frac{\text{Var}(b_2x_2) + \text{Var}(\upsilon)}{\pi^2/3}} = \sqrt{\frac{b_2^2 \times \text{Var}(x_2) + \pi^2/3}{\pi^2/3}} \tag{3}$$

Hence, when x_2 is excluded from the equation, the coefficient of x_1 then equals:

$$\frac{b_1}{s} = b_1 \times \sqrt{\frac{\pi^2/3}{b_2^2 \times \text{Var}(x_2) + \pi^2/3}} \approx b_1 \times \sqrt{\frac{3.29}{b_2^2 \times \text{Var}(x_2) + 3.29}} \tag{4}$$

In equation (4), the value by which we multiply b_1 will be less than one as long as $b_2 \neq 0$, hence, the coefficient of x_1 will be shrunken towards zero. Recall that x_1 is unrelated to x_2 or u, yet the exclusion of x_2 from the model deflates the coefficient of x_1. So this phenomenon cannot be explained by confounding or suppression effects (MacKinnon, Krull, & Lockwood, 2000). The implication of this for data analysis is that unless one includes all the determinants of the outcome in a logistic regression model, the coefficient of a predictor variable that is unrelated to other determinants (such as x_1) of the outcomes will be shrunken towards zero. The magnitude of shrinkage will depend on the variance of

[2] It is standard to express the variance of the logistic distribution as a factor of $\pi^2/3$.

the omitted variables, and how strongly the omitted variables relate with the outcome. By including more determinants of the outcome in additional models, the absolute magnitude of the coefficient of a variable, such as x_1, will increase.

One additional implication of this result is that the change in odds ratios in logistic regression as additional predictors are included in the model cannot simply be interpreted as a consequence of "controlling for other predictors". The source of this problem is that the latent variable that is modeled in logistic regression, y^*/s, changes from one model to the next model, since s varies as a function of the predictors in the model. If the latent variable changes from one model to the next, then in principle, quantities that rely directly on the latent variable such as logits and odds ratios are not comparable across models. Moving beyond comparison across models is comparison across studies, such as in meta-analysis. The regression estimates for a meta-analysis will come from models with such varying degrees in omitted variables, that the coefficients will not be comparable. This is regardless of whether the constituent studies are randomized trials.

How will the coefficient of a variable whose relationship with the outcome is confounded by other variables respond to the inclusion of these confounders in the model? The coefficient will change for two reasons: (1) its relatedness to other predictors (see equation 6 in Mood, 2010), and (2) the fact that the outcome is better explained (reduction in s). Hence, the source of the change in odds ratios in logistic regression as additional predictors are included in the model cannot simply be interpreted as a consequence of "controlling for other predictors".

Alternative effect sizes that do not suffer the same flaws as the odds ratio

One method by which researchers can address this problem is to compute and report alternative effect size measures for binary response variables. We introduce the *risk ratio* and the *risk difference* as alternatives using the simple example of a two-by-two contingency table. Additionally, we connect these effect sizes to alternative regression models that permit covariate adjustment for effects of interest. The data are hypothetical data obtained from Peng, Lee, and Ingersoll (2002). The sample comprises 189 inner city school children of which 59 were recommended for remedial reading and 130 were not. In Table 1, we break down assignment to remedial reading by gender.

Given the data in Table 1, the odds of recommending boys to remedial reading was 0.63 (36/57). The odds of recommending girls was 0.32 (23/73). Hence, the odds ratio for boys to girls was 2.00 (0.63/0.32). If one performs a logistic regression regressing the indicator for recommendation to the remedial program on an indicator for *boy*, the odds

Table 1
Two-by-two contingency table for 189 children

	Boys	Girls
Yes	36	23
No	57	73

Reproduced from Table 2 in Peng, Lee, and Ingersoll (2002)

ratio matches the exponentiated coefficient for *boy*, $\exp(0.6954) = 2.00$. And it is standard to apply logistic regression even when there are multiple predictors of the binary outcome.

The probability of recommending boys to remedial reading was 0.39 $\left(36/(36 + 57)\right)$. The probability of recommending girls was 0.24 $\left(23/(23 + 73)\right)$. Hence, the probability ratio for boys to girls was 1.62 (0.39/0.24). We can interpret the finding as: boys were on average 62% $\left((1.62 - 1) \times 100\%\right)$ more likely to be recommended to remedial reading than girls. This effect is known as the risk ratio (Valentine, Aloe, & Lau, 2015) or relative risk (Agresti, 2013, p. 43; Jewell, 2004, p. 31). We will call it the risk ratio (RR) going forward. Zou (2004) studied the use of Poisson regression to estimate the log risk ratio under limited simulation conditions and found that Poisson regression produced estimates of the log risk ratio with low bias. In our example, if one performs Poisson regression regressing the indicator for recommendation to the remedial program on an indicator for *boy*, the risk ratio matches the exponentiated coefficient for *boy*, $\exp(0.4798) = 1.62$.

Differently from calculating the ratio ratio, we can calculate the difference in the probability of recommendation between boys and girls; the difference was 0.15 $(0.39 - 0.24)$. We can interpret the finding as: boys were on average 15% points $(0.15 \times 100\%)$ more likely to be recommended to remedial reading than girls. This effect is known as the (absolute) risk difference (Valentine et al., 2015), the excess risk (Jewell, 2004, p. 37) or the difference of proportions (Agresti, 2013, p. 43). We will call it the risk difference (RD). Cheung (2007) studied the use of linear regression to estimate the risk difference under limited simulation conditions and found that linear regression produced estimates of the risk difference with low bias. In our example, if one performs linear regression regressing the indicator for recommendation to the remedial program on an indicator for *boy*, the risk difference matches the coefficient for *boy* (0.15).

We have linked both alternative effect sizes to the regression-based approaches because the simple formulas we presented for obtaining the effect sizes do not apply when one has multiple predictors of the binary response variable. And regression provides a natural approach for modeling an outcome with multiple predictors. It is also the case that the neglected heterogeneity problem with logistic regression coefficients and odds ratios does not affect linear regression and Poisson regression (Gail et al., 1984).[3]

For applied researchers, one benefit of computing these alternative effect sizes is their ease of interpretation. The risk ratio and risk differences are a multiplicative and additive comparison of probabilities respectively; and their interpretations are intuitive. This is in contrast to odds ratio which applied researchers sometimes struggle to interpret (Altman, Deeks, & Sackett, 1998; Greenland, 1987; Osborne, 2006; Schwartz, Woloshin, & Welch, 1999).

Both these modeling alternatives (linear and Poisson models on binary outcomes) are unconventional in the education literature, hence, we devote the next sections to discussing these models, informing the reader of drawbacks, and justifying their use given certain adjustments to the models.

OLS on binary response variables to obtain the risk difference. This model is known as the *linear probability model* (LPM). It goes by this name because the model assumes the relationship between the predictors and the outcome is linear on the probability scale for the outcome. This assumption reveals obvious problems. If a model is linear on the probability scale, then the predicted values (which are supposed to be probabilities) are unbounded such that one can predict probabilities less than zero or greater than one – probabilities outside the unit interval. This undesirable feature of the LPM is a simple rationale for logistic regression where the application of the (inverse-)logit function ensures all the probabilities are bounded between zero and one. Additionally, two assumptions for conducting inference with OLS using classical techniques are that the errors are normally distributed and have constant variance. It is inconceivable that such assumptions are met for binary response variables.

For certain practitioners, these concerns invalidate the LPM as an alternative when the outcome is binary. Our rebuttal begins with the truism that "all models are wrong" (Box, 1976). The logistic regression model for binary response variables is not correct, but is justified by its desirable properties. However, we have shown that it also possesses the undesirable feature that the regression coefficients are not comparable from model to model or from study to study. OLS fit to the data does not suffer this problem (Gail et al., 1984).

The natural question that follows is: how reliable are OLS coefficients for binary

[3] For the reader familiar with generalized linear models (Fox, 2015; McCullagh & Nelder, 1989), it is precisely the log and identity link functions that do not suffer the problem of neglected heterogeneity; while the logit link function does. Experienced modelers may wonder about using the binomial distribution together with a log link or identity link function to obtain the risk ratio and risk difference respectively. However, using such unorthodox link functions with the binomial distribution may result in convergence problems in commonplace statistical packages (Cheung, 2007; Wacholder, 1986).

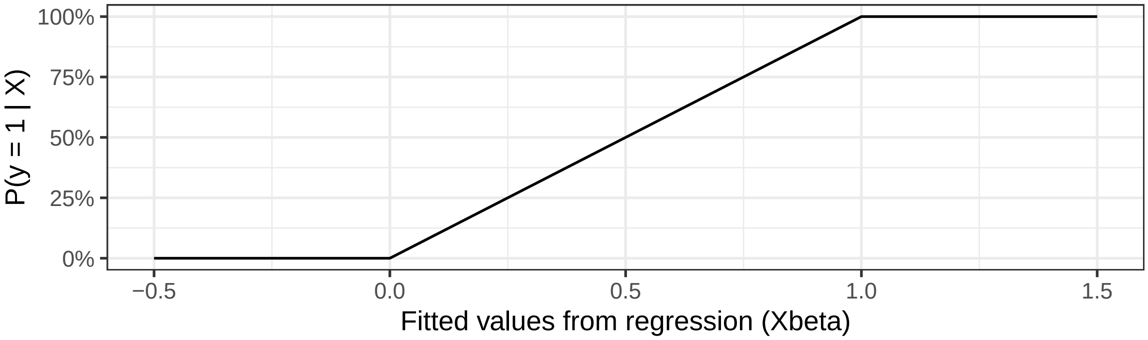

Figure 1. Relationship between fitted values from regression and probabilities implied by linear probability model

response variables? In the latent variable model we used to develop logistic regression, if we assume the error term has a uniform distribution (see Appendix A for derivation; also Olsen, 1980; Amemiya, 1981), we arrive at a linear probability model of the form:

$$P(y=1|X)=\begin{cases}1, & \text{if } X\beta>1\\ X\beta, & \text{if } X\beta\in(0,1]\\ 0, & \text{if } X\beta\le 0\end{cases} \tag{5}$$

where y is the binary response, X is the table of predictors for n cases and k predictors including the constant, β are the k regression coefficients, and $P(y=1|X)$ is the probability of success (1 on y) given the predictors. Figure 1 corresponds to equation (5); when the fitted values from the regression lie between 0 and 1 inclusive, they are the probabilities of success. When these fitted values are less than 0, we believe them to be zero; when these fitted values exceed 1, we believe them to be one.

OLS cannot estimate the non-linear relationship presented in equation (5) and depicted in Figure 1. And the *zero conditional mean* assumption for unbiasedness of OLS coefficients (Wooldridge, 2010, p. 50) is only met when all the fitted values are between 0 and 1 (Horrace & Oaxaca, 2006).

In practice, the investigator never has access to $X\beta$, since β is a population parameter. Horrace and Oaxaca (2003) suggested that one may use $X\hat{\beta}$, the predicted values from any regression application, in place of $X\beta$. Having predicted values outside the $0-1$ or unit interval can easily result in biased coefficients; in our current application, these will be biased risk differences. However, if the predicted probabilities fall on the unit interval, then the OLS estimates may be trustworthy (Horrace & Oaxaca, 2003).

A scenario in which OLS applied to a binary variable is guaranteed to predict probabilities that lie on the unit interval is when the only predictor in the model is a single categorical variable. Consider the standard two-group mean difference question that we usually analyze using *t*-tests, or one-way ANOVA when there are multiple groups, but instead, the outcome is binary. OLS applied to a binary outcome with a single categorical predictor (i.e., *t*-tests or ANOVA) is simply a model for the average probability of success for the different groups or the cell means. This is also the case for multiple categorical predictors as long as all possible interactions are in the model, such as in factorial designs. In such situations, probit, logistic, linear and Poisson regression will yield identical fit to the mean of the data i.e., the probabilities predicted by these approaches will be equivalent (Wacholder, 1986). However, there will be differences in statistical inference as the methods imply different assumptions about the error variance.

The inclusion of continuous predictors, such as in analysis of covariance or typical multiple regression contexts, may result in predicted probabilities that extend beyond the unit interval producing biased OLS estimates. If there are large enough values of the continuous predictor, the predicted values can easily fall outside the unit interval. To remedy the problem of predicted probabilities falling outside the unit interval, Horrace and Oaxaca (2003) recommended the following sequence of steps:

1. Estimate the regression model and compute the predicted values.
2. If all the predicted values fall between 0 and 1, stop, this is the final model.
3. Delete all the cases with predictions less than 0 or greater than 1, return to step 1.

This approach is based on the fact that if the predicted values in a given application are a stand-in for the true fitted values, then all that is needed to estimate the coefficients are the data points with predictions between zero and one. Implementing this approach only requires that a practitioner be able to run analysis on a subset of their data. The practitioner will run the regression on increasingly smaller subsets of their data until all the predicted probabilities fall between 0 and 1. Horrace and Oaxaca (2003) named this approach *sequential least squares* (SLS) and showed using a limited simulation study that this approach reduces the bias of OLS estimation applied to binary response variables.

An additional concern about OLS is the misspecified model for the errors; the residuals obtained are neither normal nor homoskedastic. This misspecification has consequences for statistical inference on model coefficients. To remedy this, one may compute heteroskedasticity-consistent (HC) standard errors (Cribari-Neto, Souza, & Vasconcellos, 2007), which can account for non-constant error variance when the outcome is binary (Cheung, 2007). Finally, we recommend that investigators conduct outlier detection should they choose to apply OLS estimation to binary data. The same concerns that arise in standard linear regression apply here, and diagnostic methods such as dfbetas can be helpful for identifying influential cases.

Poisson regression on binary response variables to obtain the risk ratio. Poisson regression with the log link results in a multiplicative comparison of probabilities, hence exponentiated coefficients are risk ratios. We can also exponentiate the regression prediction to obtain the predicted probability for each case. However, the range of the exponential function is 0 to ∞. The implication of this is that such a model can result in predicted probabilities that exceed one − this is clearly problematic. Additionally, an assumption for conducting inference with Poisson regression is that the probability of success at a given level of the predictor variables equals the variance of the outcome at the same level of the predictor variables − the conditional mean equals the conditional variance assumption. This assumption is problematic for binary response variables, where the variance is a quadratic function of the mean.[4]

Lumley, Kronmal, and Ma (2006) described Poisson regression applied to binary outcomes as a *Poisson working model*, and we borrow this term for the remainder of this paper. So as with OLS applied to binary outcomes, the natural question that follows is: how reliable are the regression coefficients from the Poisson working model? In the latent variable model we used to develop logistic regression, if we assume the error term has an exponential distribution (see Appendix A for derivation), we arrive at a Poisson working model of the form:

$$P(y=1|X) = \begin{cases} 1, & \text{if } X\beta > 0 \\ e^{X\beta}, & \text{if } X\beta \in (-\infty, 0] \end{cases} \tag{6}$$

using the same notation as previously. Here, $X\beta$ are the fitted values on the log scale. Figure 2 corresponds to equation (6); when the fitted values from the Poisson working model are less than 0, if we exponentiate them, we obtain the probabilities of success. When these fitted values exceed 0, we believe the probability of success to be one.

Poisson regression estimated using maximum likelihood does not estimate the non-linearity on the log scale in equation (6) and depicted on the left panel of Figure 2. And the coefficients from the Poisson working model will only be unbiased when all fitted values on the log scale are less than zero (see Appendix A).

[4] The conditional variance of a Bernoulli outcome is $\mu(1-\mu)$ where μ is the conditional mean. Poisson regression assumes the conditional variance is the same as the conditional mean i.e., μ (Krishnamoorthy, 2006, p. 134). However, as μ tends to 0, μ better approximates $\mu(1-\mu)$. This means that the Poisson distribution provides a reasonable approximation to the Bernoulli distribution when the probability of success on the outcome across all levels of the predictor variables is low.

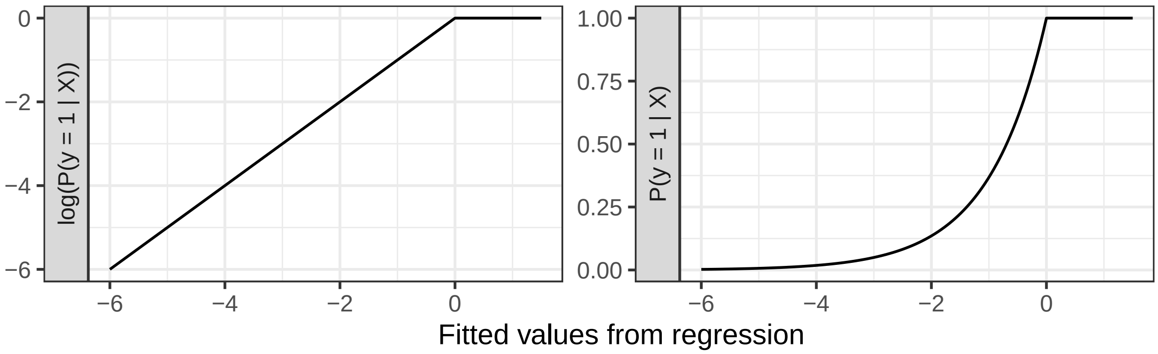

Figure 2. Relationship between fitted values from regression and probabilities implied by Poisson working model

As with all statistical models, the investigator never has access to $X\beta$; all an investigator has is $X\hat{\beta}$. We borrow from the work on Horrace and Oaxaca (2003) on the linear probability model and suggest that one may use $X\hat{\beta}$, the predicted values from any regression application in place of $X\beta$. Having predicted values on the log scale greater than 0 or predicted probabilities exceeding 1 can easily result in biased coefficients; in our current application, the exponentiated coefficients will be biased risk ratios. However, if the predicted probabilities fall on the unit interval, then the Poisson working model estimates may be trustworthy. Similar to OLS, all predicted probabilities are guaranteed to be under one when the predictors are either a single categorical predictor or all-way interactions between multiple categorical predictors (Wacholder, 1986).

When there are predicted probabilities greater than one, we hypothesize that one can extend the sequential least squares approach of Horrace and Oaxaca (2003) to the Poisson working model, following the same series of steps. As with SLS, the programming barrier to this approach is relatively low. An additional problem with the Poisson working model is the problem of variance misspecification. Zou (2004) showed using simulation that applying heteroskedasticity-consistent standard errors lessens the impact of this problem. Hence after taking a sequential approach to coefficient estimation, we hypothesize that applying heteroskedasticity-consistent standard errors should mitigate problems with variance misspecification.

A brief review of approaches derived from the logistic regression model. We note that there are approaches for deriving effect size measures from the logistic regression model that satisfy the criteria of comparability across both models and studies. Mood (2010) focused on additive comparisons of probabilities and lists two approaches in addition to linear regression that satisfy both criteria: average marginal effect (AME) and average partial effect (APE).

Mood (2010) identifies the AME as the expected change in the probability of the outcome for a unit change in the predictor of interest. This value is calculated using information from all cases and variables in the data. Hence, the AME is a single numeric summary of the effect of a predictor. According to Mood (2010), that the AME is a single numeric effect of a predictor differentiates the AME from the APE. For the APE, only cases with values that match or fall within a specified range on the predictor of interest are used in the calculation. This distinction permits the APE to vary as a function of the specified values on the predictor of interest, thus capturing nuanced relations in the data.[5]

In this paper, we focus only on single numeric summaries that communicate the effects of a predictor. We find that graphs are better at communicating nuanced summaries of relations in data, as we demonstrate in the Data Demonstration section. Hence, we eliminate the APE from consideration. On the other hand, the AME returns a value that is highly similar to OLS coefficients, prompting Mood (2010) to state, "deriving AME from logistic regression is just a complicated detour". Moreover, computing the AME may be tedious, hence the availability of special-purpose routines, e.g. Baum (2010) in Stata, and Leeper (2018) in R. Hence, we explore this procedure no further given its similarity to OLS results and the need for special-purpose routines to no notable advantage.

Methods: Simulation studies

In two simulation studies, we evaluated the LPM (study 1) and Poisson working model (study 2) when both models are true for the data. This allowed us to test parameter recovery of the risk difference and risk ratio.

In both simulations, we used the following evaluation criteria:

1. bias: estimated coefficient − population parameter; desirable result is for average bias across replications to be as close to zero as possible.
2. relative bias: bias divided by population parameter. We are only able to calculate relative bias when the population parameter is non-zero. We adopted the proposal in Flora and Curran (2004) where absolute relative bias under 5% is considered trivial, between 5% and 10% is moderate, and greater than 10% is substantial.
3. empirical coverage rate (ECR) of 95% confidence interval (CI): proportion of times across replications that the 95% CI included the population parameter. We adopted the liberal proposal in Bradley (1978) where ECR between 92.5% and 97.5% is considered

[5] There is variation in the use of these terms, as seen in examples from leading econometrics texts. For example, Cameron and Trivedi (2009, sec. 10.6) and Wooldridge (2013, p. 592) use both terms interchangeably to describe what Mood (2010) calls the AME.

adequate.

For both simulations, we have the following expectations:

- When the true fitted values result in plausible probabilities (all $X\beta \in [0, 1]$ for the LPM, all $X\beta \in (-\infty, 0]$ for the Poisson working model), the standard linear and Poisson regression models will have acceptable (relative) bias in parameter recovery for the risk difference and (log) risk ratio respectively. Additionally, applying heteroskedasticity-consistent standard errors will ensure that the empirical coverage rates of the confidence intervals are adequate.
- When the true fitted values result in implausible probabilities (some $X\beta < 0 \vee X\beta > 1$ for the LPM, some $X\beta \in (0, \infty)$ for the Poisson working model), the standard linear and Poisson regression models will have high (relative) bias in parameter recovery for the risk difference and (log) risk ratio respectively. However, applying the sequential approach of Horrace and Oaxaca (2003) will reduce the bias to acceptable levels. Additionally, applying heteroskedasticity-consistent standard errors to the estimates based on the sequential approach will ensure that the empirical coverage rates of the confidence intervals are adequate.

The specific variant of heteroskedasticity-consistent standard errors we applied was HC4. In addition to accounting for heteroskedasticity, it also accounts for leverage (Cribari-Neto et al., 2007).

Study 1: Evaluating OLS and SLS when the linear probability model is true

We simulated data according to the linear probability model to test the performance of OLS and SLS. We controlled the proportion of cases with predicted probabilities that exceeded the unit interval. We generated 4,999 datasets (each $n = 500$) according to the following equation:

$$y = \begin{cases} 1, & \text{if } 0.5 + \beta_b \times x_b + \beta_c \times x_c + v > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where x_b was binary with equal proportion of cases at 0 and 1, x_c was continuous uniform with minimum of -1 and maximum of 1, $\text{Cor}(x_b, x_c) = .3$ suggesting a moderate correlation between both predictors, and v was continuous uniform with minimum of -0.5 and maximum of 0.5. When we generate data according to this process, we would expect that on average, the OLS coefficients for x_b and x_c would be β_b and β_c respectively.

Table 2

Design conditions to test least squares estimation methods for linear probability model

Conditions	Binary coef.	Continuous coef.	Min P	Max P	Expected behaviour
1	0.00	0.00	0.50	0.50	Acceptable
2	0.10	0.00	0.50	0.60	Acceptable
3	0.00	0.25	0.25	0.75	Acceptable
4	0.10	0.25	0.25	0.85	Acceptable
5	0.00	0.50	0.00	1.00	Borderline
6	0.10	0.50	0.00	1.10	Problematic
7	0.00	0.75	-0.25	1.25	Problematic
8	0.10	0.75	-0.25	1.35	Problematic

Note. Min P and Max P are the minimum and maximum prediction respectively given the intercept of 0.5 and coefficients of the binary and continuous predictors. We expect OLS to perform adequately when all the predictions fall between 0 and 1, hence the expectations in the final column.

In equation (7), the regression prediction for each case is $0.5 + \beta_b \times x_b + \beta_c \times x_c$. If there are cases in the data for which this value is less than 0 or greater than 1, we should expect OLS estimates to be biased. This is what we set out to test. Making the continuous variable uniform ensures we can generate a regression prediction that will remain within a pre-specified range. To control the proportion of cases with predictions outside the unit interval, we varied the regression coefficients (see Table 2). Table 2 also includes the values for the minimum and maximum predictions. These values are found through the following formula:

$$\text{For condition } j \begin{cases} \text{Min P} & = 0.5 + \beta_b \times \min(x_b) + \beta_c \times \min(x_c) \\ \text{Max P} & = 0.5 + \beta_b \times \max(x_b) + \beta_c \times \max(x_c) \end{cases}, \ j = \{1, 2, \ldots, 8\} \tag{8}$$

Using condition 8 in the Table as an example, the minimum probability was $0.5+(0.1\times 0) + (0.75 \times -1) = -0.25$, and the maximum probability was $0.5 + (0.1 \times 1) + (0.75 \times 1) = 1.35$. Consequently, we would expect OLS to return biased coefficients in condition 8, hence the expected *problematic* behaviour. Finally, for both OLS and SLS, we applied heteroskedasticity-consistent standard errors to account for the heteroskedasticity implied by the linear probability model.

Table 3

Design conditions to test Poisson regression applied to binary response variables

Conditions	Binary coef.	Continuous coef.	Min P	Max P	Expected behaviour
1	0.00	0.00	0.37	0.37	Acceptable
2	0.50	0.00	0.37	0.61	Acceptable
3	0.00	0.50	0.22	0.61	Acceptable
4	0.50	0.50	0.22	1.00	Borderline
5	0.00	1.00	0.14	1.00	Borderline
6	0.50	1.00	0.14	1.65	Problematic

Note. Min P and Max P are the minimum and maximum probabilities respectively given the intercept of -1 and coefficients of the binary and continuous predictors. We exponentiated the regression prediction to obtain the minimum and maximum predicted probabilities. We expect Poisson regression to perform adequately when all the predicted probabilities are less than 1, hence the expectations in the final column.

Study 2: Evaluating the performance of the Poisson working model when the Poisson working model is true

We simulated data according to the Poisson working model to test the performance of Poisson regression and a sequential Poisson regression approach. We generated 4,999 datasets (each $n = 500$) according to the following equation:

$$y = \begin{cases} 1, & \text{if } -1 + \beta_b \times x_b + \beta_c \times x_c + \upsilon > -1 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where x_b was binary with equal proportion of cases at 0 and 1, x_c was continuous uniform with minimum of -1 and maximum of 1, $\mathrm{Cor}(x_b, x_c) = .3$ suggesting a moderate correlation between both predictors, and υ was exponential with scale parameter of $1 - 1$, $\upsilon \sim \exp(1) - 1$. When we generate data according to this process, we would expect that on average, the Poisson regression coefficients for x_b and x_c would be β_b and β_c respectively.

In equation (9), the regression prediction for each case is $-1 + \beta_b \times x_b + \beta_c \times x_c$. However, if there are cases in the data for which this value is greater than 0, we should expect the model estimates to be biased. This is what we set out to test. To control the proportion of cases with predictions over zero, we varied the regression coefficients (see Table 3). Table 3 also includes the values for the minimum and maximum predicted probabilities. These values are found through the following formula:

$$\text{For condition } j \begin{cases} \text{Min P} & = \exp(-1 + \beta_b \times \min(x_b) + \beta_c \times \min(x_c)) \\ \text{Max P} & = \exp(-1 + \beta_b \times \max(x_b) + \beta_c \times \max(x_c)) \end{cases}, \; j = \{1, 2, \ldots, 6\} \tag{10}$$

Using condition 6 in the Table as an example, the minimum probability was $e^{[-1+(0.5\times 0)+(1\times -1)]} = e^{-2} = 0.14$, and the maximum probability was $e^{[-1+(0.5\times 1)+(1\times 1)]} = e^{0.5} = 1.65$. Consequently, we would expect the Poisson working model to return biased coefficients in condition 6, indicated as *problematic* behaviour. For both Poisson and sequential Poisson approaches, we applied HC4 standard errors.

Results

Study 1: Evaluating OLS and SLS when the linear probability model is true

We report the complete set of results in Appendix B. For the first four conditions in Table 2, there were no predicted probabilities outside the unit interval, so OLS and SLS were equivalent. The range of the coefficient bias was -0.0006 to 0.001. The range of the empirical coverage rate of the 95% CI was 94.5% to 95.2%. Both of these results are adequate suggesting that when there are no predicted probabilities exceeding the unit interval, OLS estimation adequately recovers coefficients and we may proceed with statistical inference using heteroskedasticity-consistent standard errors.

We report results for the binary variable when there were predicted probabilities outside the unit interval, but the true coefficient of the binary variable was zero i.e. conditions 5 and 7. The range of bias under OLS and SLS was -0.005 to 0.0001 respectively. The range of the empirical coverage rate of the 95% CI was 94.1% to 94.9%. These are small values of bias, and acceptable empirical coverage rates.

We report the remaining results in Figure 3. SLS had acceptable relative bias and empirical coverage rates across all (remaining) conditions. However, as the percentage of fitted values outside the unit interval increased, the performance of OLS worsened. As an extreme example, for conditions 7 and 8, the empirical coverage rate of OLS for the continuous coefficient was 0%.

With these results, we find that the more extreme the proportion of cases with predicted probabilities outside the unit interval, the greater the potential for inadequate effect estimation and inference. In such situations, using sequential least squares to obtain

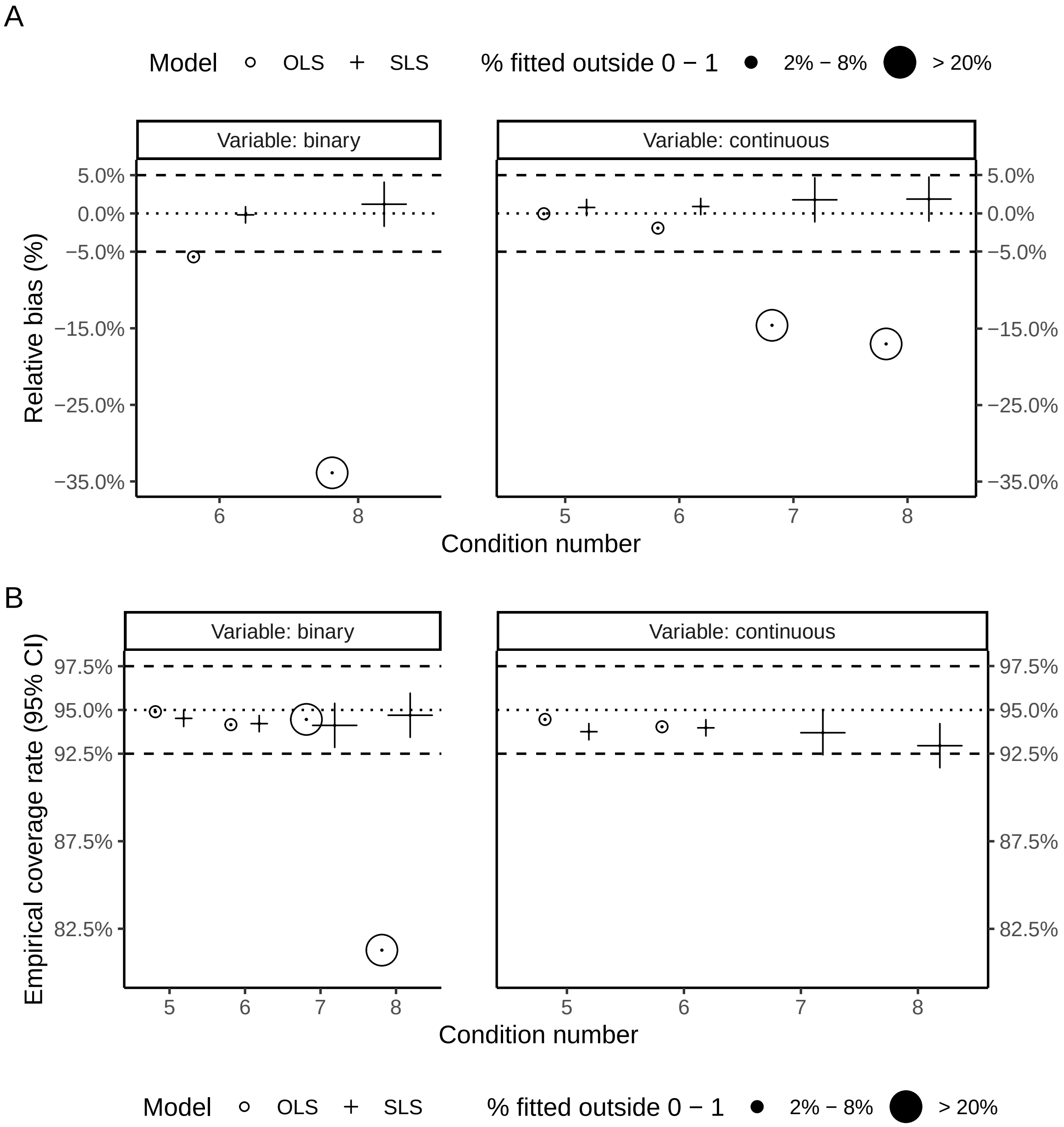

Figure 3. OLS and SLS were equivalent and performed adequately in conditions 1 to 4, see Appendix B. Panel A: Average relative bias of coefficients of two estimation approaches for the linear probability model. SLS coefficients always demonstrated acceptable levels of relative bias, -5% < relative bias < 5%. Conditions 5 and 7 are missing for the binary coefficient because the population parameters were zero, hence, it was impossible to calculate relative bias. The bias of SLS under these conditions was negligible, see Appendix B. Panel B: Empirical coverage rate of 95% confidence interval of two estimation approaches for the linear probability model. SLS always maintained a reasonably good coverage rate, 92.5% < ECR < 97.5%. The ECR of OLS in conditions 7 and 8 was 0%, see Appendix B.

the risk difference can reduce the problems with using OLS. These results apply when the LPM is true for the data.

Study 2: Evaluating the performance of the Poisson working model when the Poisson working model is true

We report the complete set of results in Appendix B. For the first three conditions in Table 3, there were no predicted probabilities above one, so Poisson and sequential Poisson approaches were equivalent. The range of the coefficient bias was -0.004 to 0.003. The range of the empirical coverage rate of the 95% CI was 94.9% to 95.3%. Both of these results are adequate suggesting that when there are no predicted probabilities above one, Poisson regression estimated using maximum likelihood adequately recovers coefficients and we may proceed with statistical inference using heteroskedasticity-consistent standard errors.

We report results for the binary variable when there were predicted probabilities outside the unit interval, but the true coefficient of the binary variable was zero i.e. condition 5. The bias under Poisson and sequential Poisson approaches was 0.001 and -0.0001 respectively. The empirical coverage rate of the 95% CI under Poisson and sequential Poisson approaches was 95.0% and 94.9% respectively. These are small values of bias, and acceptable empirical coverage rates.

Same as the contrast between OLS and SLS, the sequential Poisson approach performed adequately in situations Poisson regression failed. We see this when we compare the methods on the relative bias (see top panel of Figure 4) and the empirical coverage rate of the 95% confidence interval (see bottom panel of Figure 4). And we reach the same conclusions: the more extreme the proportion of cases with predicted probabilities greater than one, the greater the potential for inadequate effect estimation and inference. In such situations, estimating the (log-)risk ratio using a sequential Poisson approach can reduce some of the problems with Poisson regression.

Data Demonstration

We use data from the Early Childhood Longitudinal Study − Kindergarten (ECLS-K). The data are publicly available from the National Center for Educational Statistics (https://nces.ed.gov/ecls). The ECLS-K data provide descriptive information on children's status at their entry into school and their transition through school. For our demonstration, we focus on a subsample of 702 children whose proficiency levels have been previously analyzed by O'Connell (2006). The children in the sample fell into three levels of reading

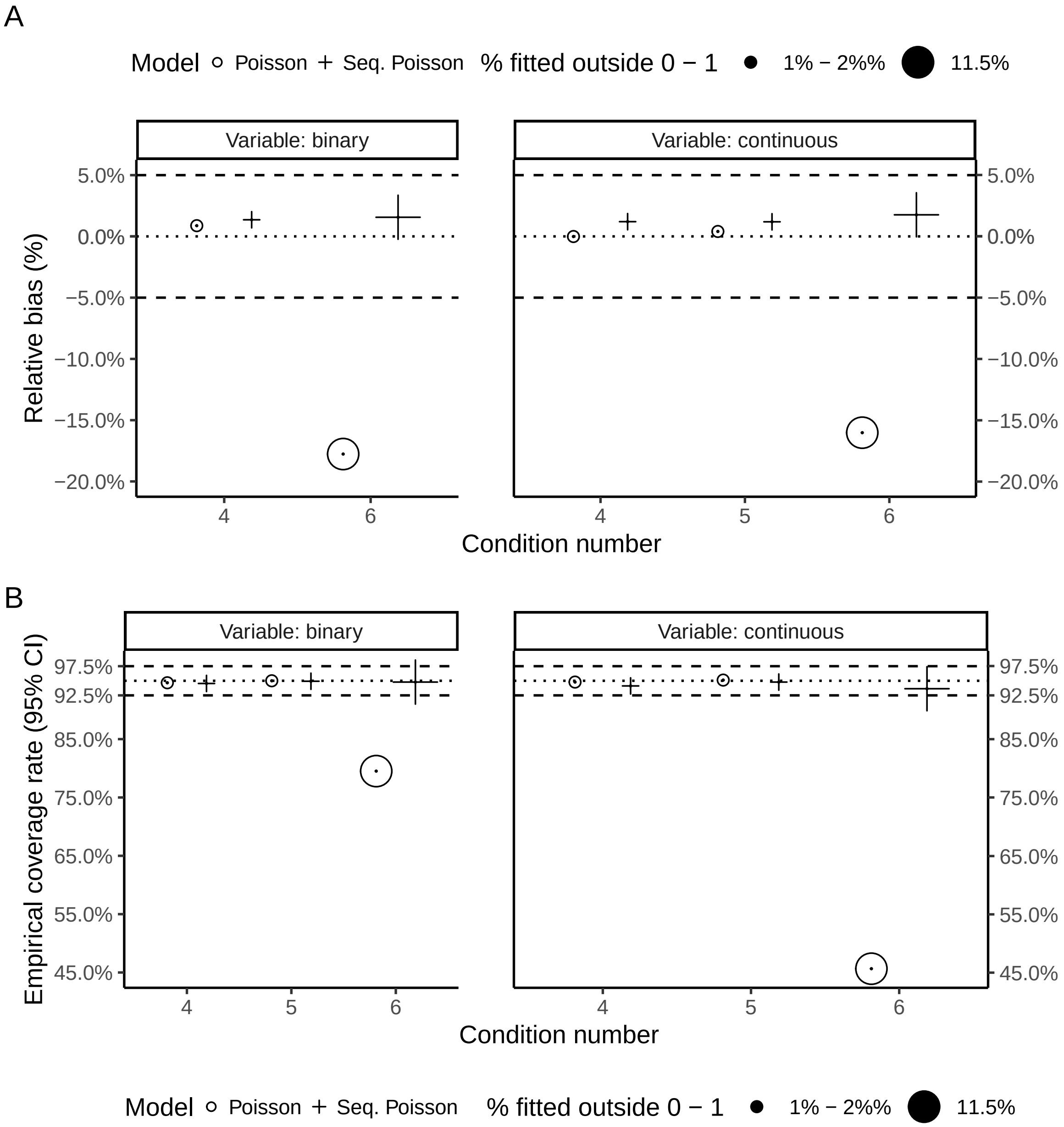

Figure 4. Poisson and sequential Poisson were equivalent and performed adequately in conditions 1 to 3, see Appendix B. Panel A: Average relative bias of coefficients of two estimation approaches for the Poisson working model. Sequential Poisson coefficients always demonstrated acceptable levels of relative bias, -5% < relative bias < 5%. Condition 5 is missing for the binary coefficient because its population parameter was zero, hence, it was impossible to calculate relative bias. The bias of sequential Poisson under this condition was negligible, see Appendix B. Panel B: Empirical coverage rate of 95% confidence interval of two estimation approaches for the linear probability model. Sequential Poisson always maintained a reasonably good coverage rate, 92.5% < ECR < 97.5%.

proficiency: can read words in context (level 5), can identify upper/lowercase letters (level 1), and children who did not pass level 1 proficiency (level 0) as of first grade. To create a binary response variable, we grouped children in the last two categories together i.e., we modeled the outcome with success as level 5 proficiency ($n = 357$) against failure as level 0 or 1 proficiency, $n = 345$.

As predictors of the outcome, we used the same variables as O'Connell (2006). According to O'Connell (2006), the selection of these predictors was based on the literature on reading proficiency. The predictors are sex (0 = male, 1 = *female*), *famrisk* (risk factors, 0 = none, 1 = one or more), *noreadbo* (parent reads books to child at least three times a week = 0, less than three times a week = 1), *halfdayK* (0 = child attended full-day kindergarten, 1 = child attended half-day kindergarten), *center* (whether or not child ever received center-based day care prior to attending kindergarten; 0 = no, 1 = yes), *minority* (0 = white/Caucasian background, 1 = other background), *wksesl* (family SES assessed prior to kindergarten, range of -4.18 to 2.64), and *p1ageent_66* (age of child in months at kindergarten entry centered at the median of 66 months, range of -9 to 13).

To model the outcome, we used logistic regression, OLS and the Poisson working model. As there were predicted probabilities outside the unit interval for both OLS and the Poisson working model, we also applied the sequential approaches. For all models except the logistic regression model, we used heteroskedasticity-consistent standard errors. We report model results in Table 3.

First, we note that the pattern of statistical significance was relatively similar across models. We expect this finding to hold in most applications at large enough sample sizes. Next, sequential least squares and Poisson working models had sample sizes that were 21% and 24% lower than the total sample size. This might be of concern to applied researchers. Based on our simulation results, this is the price we have to pay to obtain less biased estimates of the RD and RR assuming the linear probability or Poisson working model is true. Additionally, for *female* and *wksesl*, the effects from the sequential approaches are substantially larger than the effects from the standard OLS and Poisson approaches.

To better understand these models in this empirical example, we produce a counterfactual plot (Gelman & Hill, 2007, Chapter 9; McElreath, 2015, Section 5.1.3.2) of the model-implied effect of *wksesl* on the probability of success for each of these models. To do this, we generate a new dataset of the same sample size as the original data, with all predictors except *wksesl* set to their median value, and we retained the original values of *wksesl*. This allowed us to explore the relationship between *wksesl* and the probability of success implied by these models for a child with the following characteristics: *female* =

Table 4

Analysis of dichotomized reading proficiency

	Outcome variable: Reading proficiency				
	Logistic	OLS	SLS	Poisson	Seq-Poisson
female	0.83	0.13	0.18	0.29	0.44
	(0.20)***	(0.03)***	(0.04)***	(0.06)***	(0.10)***
famrisk	−0.61	−0.11	−0.11	−0.35	−0.37
	(0.22)**	(0.04)**	(0.04)**	(0.10)***	(0.12)**
center	0.21	0.04	0.04	0.10	0.12
	(0.24)	(0.04)	(0.05)	(0.10)	(0.12)
noreadbo	−0.74	−0.13	−0.14	−0.48	−0.46
	(0.28)**	(0.04)**	(0.05)**	(0.17)**	(0.17)**
minority	−0.20	−0.04	−0.04	−0.12	−0.15
	(0.21)	(0.03)	(0.04)	(0.07)	(0.11)
halfdayK	0.04	0.01	0.02	0.04	0.08
	(0.20)	(0.03)	(0.04)	(0.06)	(0.10)
wksesl	1.72	0.25	0.35	0.48	0.91
	(0.17)***	(0.02)***	(0.03)***	(0.04)***	(0.10)***
p1ageent_66	0.09	0.01	0.02	0.03	0.04
	(0.02)***	(0.004)***	(0.005)***	(0.01)***	(0.01)***
Constant	−0.34	0.46	0.42	−0.93	−1.02
	(0.29)	(0.05)***	(0.05)***	(0.11)***	(0.15)***
Observations	702	702	556	702	533

Note: *p<0.05; **p<0.01; ***p<0.001
Standard errors for the linear and Poisson models are HC4 robust standard errors.

0, *famrisk* = 0, *center* = 1, *noreadbo* = 0, *minority* = 0, *halfdayK* = 0, *p1ageent_66* = 0. We used the regression equations together with the mean functions to calculate the predicted probabilities, and we left and right censored the predicted probabilities at 0 and 1 respectively.

From the plot in Figure 5, we find that the sequential approaches better approximate the logistic curve compared to the standard approaches. Moreover, the RD from SLS approximates the linear effect of *wksesl* on the probability of success in the region where *wksesl* influences the probability of success. One can make the same comment about the sequential Poisson working model; the implied RR conveys the multiplicative effect of *wksesl* on the probability of success in the region where *wksesl* influences the probability of success. This examination alongside the simulation results suggest that for the RD and RR, we interpret the results from the sequential approaches despite the drop in sample size.

As examples, we interpret the coefficients for *female* and *wksesl*. If they were similar on all other predictors, female children had an average odds of success that was 2.29 ($e^{0.83}$) times the odds of success for male children. In terms of the RD, the female children were on

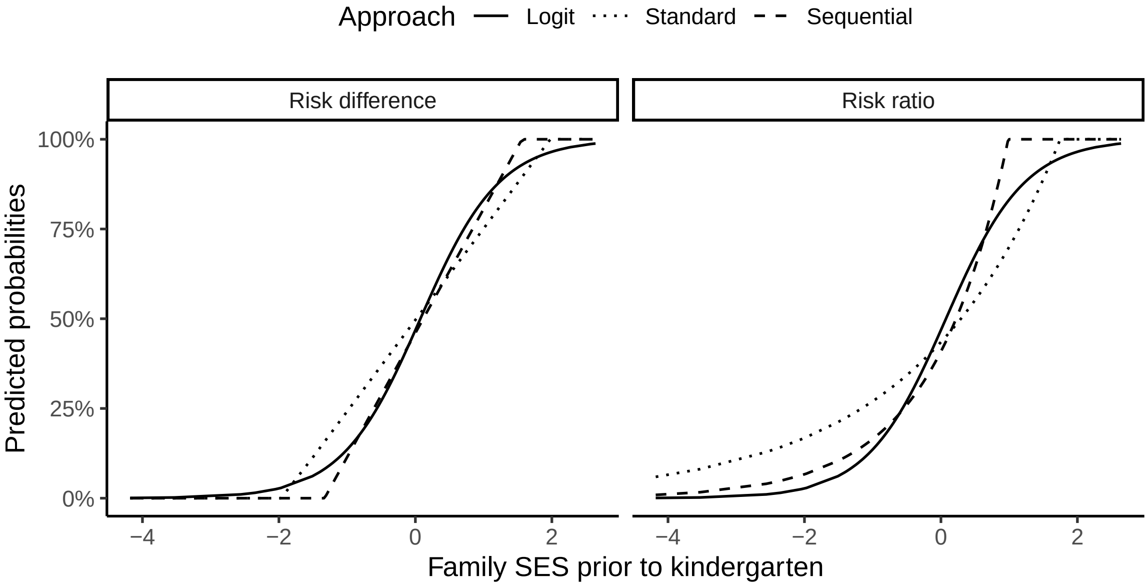

Figure 5. Relationship between family SES and predicted probabilities under different approaches holding all other predictors at their median values. The standard approach for the RD and RR are OLS and the Poisson working model respectively. The sequential approaches are the modifications to these standard approaches.

average 18% points more likely to succeed than their male counterparts. In terms of the RR, the female children were on average 55% $((e^{0.44} - 1) \times 100\%)$ more likely to succeed than their male counterparts. If children were similar on all other predictors, but they differed in family SES by one point, then the higher SES children had an average odds of success that was 5.58 $(e^{1.72})$ times the odds of success of the lower SES children. In terms of the RD, the higher SES children were on average 35% points more likely to succeed than the lower SES children. In terms of the RR, the higher SES children were on average 148% $((e^{0.91} - 1) \times 100\%)$ more likely to succeed than the lower SES children.

Discussion

We have made the case that the odds ratio is not comparable across models and in particular, across studies. A question that follows is: are meta-analyses of odds ratios of value? Although we have presented methods for calculating effect sizes that are comparable across studies, we think that there remains value to meta-analysis of odds ratios. It is important to note that the log odds ratio of a variable that is unconfounded with other determinants of the outcome is deflated towards zero. So a meta-analysis of the odds ratios from well-executed randomized trials will be conservative. Moreover, the lack of comparability does not affect the sign of odds ratios. If a meta-analysis provides evidence for an average odds ratio above 1 (or a positive log odds ratio), our arguments about the comparability of odds ratios do not provide reasons to doubt the findings from the meta-analysis.

In relation to the sequential least squares and Poisson approaches we have presented in this paper, we recommend them as additional tools for the data analyst during the analysis of binary response variables. Our work here does not discourage the use of logistic regression. The alternative regression approaches provide effect sizes that are comparable across models and are useful for comparing the effect of variables of interest across studies. Hence, alongside the odds ratios that are presented with standard logistic regression results, we recommend the inclusion of at least one other effect size measure that can be included in a comparision of effects from different studies. We also advise that researchers produce the counterfactual plots implied by the different modeling approaches for predictors of interest, as demonstrated in the real data example. We expect that in many applications, this practice will facilitate the understanding and interpretation of the effects communicated by the coefficients from different models. For example, practitioners can check how well an SLS estimate matches the linear approximation to the relationship implied by the logistic regression.

Some practitioners may have reservations about having to delete portions of their data to obtain effect size estimates. We note that such data modifications have a long history in statistical practice. For example, popular forms of robust (to outliers) regression use estimators that weight cases depending on how "extreme" the cases are (Faraway, 2002, Chapter 13). If a case is extreme enough, its weight can become zero effectively deleting the case. Additionally, data modification in the form of *winsorizing* and discarding data in the form of *trimming* to compute effect sizes are common methodological recommendations that go as far back as Tukey (1962), with a more recent example in Algina, Keselman, and Penfield (2005).[6] There is also the practice of data augmentation where practitioners add cases to their data to improve the statistical properties of estimators, e.g. Greenland and Mansournia (2015). These are some examples that demonstrate the history of modifying the data during analysis, when the practice is justifiable. In our particular situation, we repeatedly delete cases with predictions outside the unit interval because we assume that their predicted values are known. If their predicted probabilities are under zero/above one, then the predicted probabilities are zero or one. And given the estimation methods of least squares and maximum likelihood, using such cases to estimate the regression coefficients will bias estimation, justifying their exclusion.

The major design consideration for the simulation we conducted was the degree to which

[6] A very common example is the median; it is the trimmed mean with $50\% - 1$ of data points discarded from each tail (Velleman & Wilkinson, 1993).

the true fitted values resulted in probabilities on the unit interval. Within the simulation that followed from this design consideration, we found that the sequential approaches had good parameter recovery and empirical coverage rates when compared to the standard approaches. We welcome more thorough methodological examination via simulation, and mention a few relevant design considerations for such a simulation study. The sequential approaches we recommend rely on case deletion, and case deletion can lead to consequential drops in power and estimation efficiency especially at smaller sample sizes. A similar challenge occurs when group membership is highly unbalanced. Case deletion might lead to the loss of a consequential number of cases within a group with a small sample size even when the total sample size is barely affected. This loss of cases would compromise the power and efficiency of interactions. Given the aforementioned, a more comprehensive study of the recommended sequential methods should include sample size and degree of group balance as deisgn considerations. Additionally, the rate of the outcome impacts case deletion. For very low rate outcomes, SLS may suffer significant loss of cases. While for very high rate outcomes, both SLS and sequential Poisson approaches may suffer significant loss of cases. Hence, the rate of the outcome should be an additional design consideration. Our data generation processes are flexible enough to accommodate these design considerations.[7]

Finally, hierarchical data structures are commonplace in educational data. These hierarchies and grouping structures can lead to violation of the independence assumption of generalized linear models. This fact partially accounts for the popularity of multilevel regression methods in educational research (McNeish, Stapleton, & Silverman, 2017). The reader should note that the problems we have documented about comparability of odds ratios across models and studies also apply to multilevel logistic regression for binary outcomes. In the future, we intend to investigate possible approaches to fitting multilevel linear and Poisson models that may help with the estimation of the risk difference and risk ratio when the data are multilevel.

[7] The binary variable, x_b, was generated as Bernoulli with a probability of success equal to 0.5. Varying this probability will result in unbalanced group sizes. Also, one can manipulate the intercept in both data generation processes to arrive at different outcome rates.

References

Agresti, A. (2013). *Categorical data analysis*. Hoboken, NJ: Wiley-Interscience.

Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, *10*(3), 317–328. doi:10.1037/1082-989X.10.3.317

Altman, D. G., Deeks, J. J., & Sackett, D. L. (1998). Odds ratios should be avoided when events are common. *BMJ: British Medical Journal*, *317*(7168), 1318.

Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature*, *19*(4), 1483–1536. Retrieved from http://www.jstor.org/stable/2724565

Baum, C. F. (2010). Stata tip 88: Efficiently evaluating elasticities with the margins command. *Stata Journal*, *10*(2), 309–312.

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. doi:10.1080/01621459.1976.10480949

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. doi:10.1111/j.2044-8317.1978.tb00581.x

Brand, J. E., & Halaby, C. N. (2006). Regression and matching estimates of the effects of elite college attendance on educational and career achievement. *Social Science Research*, *35*(3), 749–770. doi:10.1016/j.ssresearch.2005.06.006

Cameron, A. C., & Trivedi, K. P. (2009). *Microeconometrics Using Stata* (p. 692). College Station, TX: Stata Press.

Carpenter II, D. M., Kaka, S. J., Tygret, J. A., & Cathcart, K. (2018). Testing the efficacy of a scholarship program for single parent, post-freshmen, full time undergraduates. *Research in Higher Education*, *59*(1), 108–131. doi:10.1007/s11162-017-9456-0

Cheung, Y. B. (2007). A modified least-squares regression approach to the estimation of risk difference. *American Journal of Epidemiology*, *166*(11), 1337–1344. doi:10.1093/aje/kwm223

Cox, B. E., Reason, R. D., Nix, S., & Gillman, M. (2016). Life happens (outside of college): Non-college life-events and students' likelihood of graduation. *Research in Higher Education*, *57*(7), 823–844. doi:10.1007/s11162-016-9409-z

Cribari-Neto, F., Souza, T. C., & Vasconcellos, K. L. P. (2007). Inference under heteroskedasticity and leveraged data. *Communications in Statistics - Theory and Methods*, *36*(10), 1877–1888. doi:10.1080/03610920601126589

Eccles, J. S., Vida, M. N., & Barber, B. (2004). The relation of early adolescents' college plans and both academic ability and task-value beliefs to subsequent college enrollment. *The Journal of Early Adolescence*, *24*(1), 63–77. doi:10.1177/0272431603260919

Faraway, J. J. (2002). *Practical regression and Anova using R* (p. 212). Retrieved from https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. doi:10.1037/1082-989X.9.4.466

Fox, J. (2015). *Applied regression analysis and generalized linear models.* Thousand Oaks: SAGE Publications.

Gail, M. H., Wieand, S., & Piantadosi, S. (1984). Biased estimates of treatment effects in randomized experiments with nonlinear regression and omitted covariates. *Biometrika*, *71*(3), 431–444. doi:10.2307/2336553

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* (p. 625). Cambridge, UK: Cambridge University Press.

Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology*, *125*(5), 761–768. doi:10.1093/oxfordjournals.aje.a114593

Greenland, S., & Mansournia, M. A. (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine*, *34*(23), 3133–3143. doi:10.1002/sim.6537

Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, *14*(1), 29–46. doi:10.1214/ss/1009211805

Horrace, W. C., & Oaxaca, R. L. (2003, January). *New wine in old bottles: A sequential estimation technique for the LPM.* Retrieved from https://ssrn.com/abstract=383102

Horrace, W. C., & Oaxaca, R. L. (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, *90*(3), 321–327. doi:10.1016/j.econlet.2005.08.024

Jewell, N. P. (2004). *Statistics for epidemiology.* New York: Chapman & Hall/CRC.

Klevan, S., Weinberg, S. L., & Middleton, J. A. (2016). Why the boys are missing: Using social capital to explain gender differences in college enrollment for public high school students. *Research in Higher Education*, *57*(2), 223–257. doi:10.1007/s11162-015-9384-9

Krishnamoorthy, K. (2006). *Handbook of statistical distributions with applications*. New York: Chapman & Hall/CRC.

Leeper, T. J. (2018). *margins: Marginal effects for model objects*. Retrieved from https://cran.r-project.org/package=margins

Long, J. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks: SAGE Publications.

Lumley, T., Kronmal, R., & Ma, S. (2006). *Relative risk regression in medical research: Models, contrasts, estimators, and algorithms*. UW Biostatistics. Retrieved from http://biostats.bepress.com/uwbiostat/paper293

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, *1*(4), 173–181. doi:10.1023/A:1026595011371

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman & Hall/CRC.

McElreath, R. (2015). *Statistical rethinking : A Bayesian course with examples in R and Stan* (p. 469). Boca Raton, FL: Chapman; Hall/CRC. Retrieved from https://xcelab.net/rm/statistical-rethinking/

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114–140. doi:10.1037/met0000078

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, *26*(1), 67–82. doi:10.1093/esr/jcp006

O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks: SAGE Publications.

Olsen, R. J. (1980). A least squares correction for selectivity bias. *Econometrica*, *48*(7), 1815–1820. doi:10.2307/1911938

Osborne, J. W. (2006). Bringing balance and technical accuracy to reporting odds ratios and the results of logistic regression analyses. *Practical Assessment, Research and Evaluation*, *11*(7), 1–6. Retrieved from http://pareonline.net/getvn.asp?v=11&n=7

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, *96*(1), 3–14. doi:10.1080/00220670209598786

Schwartz, L. M., Woloshin, S., & Welch, H. G. (1999). Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. *The New England Journal of Medicine*, *341*(4), 279–283; discussion 286–287. doi:10.1056/NEJM199907223410411

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67. doi:10.1214/aoms/1177704711

Valentine, J. C., Aloe, A. M., & Lau, T. S. (2015). Life after NHST: How to describe your data without "p-ing" everywhere. *Basic and Applied Social Psychology*, *3533*(5), 1–14. doi:10.1080/01973533.2015.1060240

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, *47*(1), 65–72. doi:10.1080/00031305.1993.10475938

Wacholder, S. (1986). Binomial regression in glim: Estimating risk ratios and risk differences. *American Journal of Epidemiology*, *123*(1), 174–184. doi:10.1093/oxfordjournals.aje.a114212

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* Cambridge, MA: MIT press.

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (p. 881). Mason, OH: South-Western Cengage Learning.

Zou, G. (2004). A modified Poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology*, *159*(7), 702–706. doi:10.1093/aje/kwh090

Appendix A
Derivations for linear probability and Poisson working models

All that is required to follow the derivations are familiarility with common probability distributions, and basic algebra. For formulas for the distribution functions, we use Krishnamoorthy (2006).

Derivation for linear probability model

We note that the derivation presented here is an old finding, see Olsen (1980), Amemiya (1981), Horrace & Oaxaca (2006). The key question of interest in regression models for binary response variables is: what are the probabilities of success underlying a binary response variable, $\mathbf{y}$, given k predictors in $\mathbf{X}$, i.e., $P(\mathbf{y}=1|\mathbf{X})$? $\mathbf{y} \in \{0,1\}$ for n cases, and $\mathbf{X}$ is an $n \times k$ matrix.

$$\begin{aligned} P(\mathbf{y}=1|\mathbf{X}) &= P(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\upsilon} > \tau|\mathbf{X}) \quad \text{using a latent variable formulation for } \mathbf{y} \\ &= P(\boldsymbol{\upsilon} > \tau - \mathbf{X}\boldsymbol{\beta}|\mathbf{X}) = 1 - P(\boldsymbol{\upsilon} < \tau - \mathbf{X}\boldsymbol{\beta}|\mathbf{X}) \end{aligned} \tag{11}$$

where $\boldsymbol{\beta}$ are the k regression coefficients, τ is the dichotomization threshold, and $\boldsymbol{\upsilon}$ is the error term. For model identification, we set $\tau = 0.5$, such that the last line of equation (11) becomes $1 - P(\boldsymbol{\upsilon} < 0.5 - \mathbf{X}\boldsymbol{\beta}|\mathbf{X})$. Finally, we assume $\boldsymbol{\upsilon} \sim \text{unif}(-0.5, 0.5)$.

Selecting unif$(-0.5, 0.5)$ as the distribution for $\boldsymbol{\upsilon}$ means that $P(\boldsymbol{\upsilon} < 0.5 - \mathbf{X}\boldsymbol{\beta})$ is the distribution function of unif$(-0.5, 0.5)$ evaluated at $0.5 - \mathbf{X}\boldsymbol{\beta}$. Hence, based on the uniform distribution function:

$$\begin{aligned} P(\boldsymbol{\upsilon} < 0.5 - \mathbf{X}\boldsymbol{\beta}) &= \begin{cases} 0, & \text{if } 0.5 - \mathbf{X}\boldsymbol{\beta} < -0.5 \\ \frac{(0.5 - \mathbf{X}\boldsymbol{\beta}) - (-0.5)}{0.5 - (-0.5)}, & \text{if } 0.5 - \mathbf{X}\boldsymbol{\beta} \in [-0.5, 0.5) \\ 1, & \text{if } 0.5 - \mathbf{X}\boldsymbol{\beta} \geq 0.5 \end{cases} \\ &= \begin{cases} 0, & \text{if } \mathbf{X}\boldsymbol{\beta} > 1 \\ 1 - \mathbf{X}\boldsymbol{\beta}, & \text{if } \mathbf{X}\boldsymbol{\beta} \in (0, 1] \\ 1, & \text{if } \mathbf{X}\boldsymbol{\beta} \leq 0 \end{cases} \end{aligned} \tag{12}$$

Since $P(\mathbf{y}=1|\mathbf{X}) = 1 - P(\boldsymbol{\upsilon} < 0.5 - \mathbf{X}\boldsymbol{\beta})$, then:

$$P(\mathbf{y}=1|\mathbf{X}) = \begin{cases} 1, & \text{if } \mathbf{X}\boldsymbol{\beta} > 1 \\ \mathbf{X}\boldsymbol{\beta}, & \text{if } \mathbf{X}\boldsymbol{\beta} \in (0, 1] \\ 0, & \text{if } \mathbf{X}\boldsymbol{\beta} \leq 0 \end{cases} \tag{13}$$

OLS estimates $\mathbf{X}\boldsymbol{\beta}$ without respect for the bounds, hence the expectation for the regression error term, $\boldsymbol{\epsilon}$, is:

$$\mathbb{E}(\boldsymbol{\epsilon}|\mathbf{X}) = P(\mathbf{y}=1|\mathbf{X}) - \mathbf{X}\boldsymbol{\beta} = \begin{cases} 1-\mathbf{X}\boldsymbol{\beta}, & \text{if } \mathbf{X}\boldsymbol{\beta} > 1 \\ 0, & \text{if } \mathbf{X}\boldsymbol{\beta} \in (0,1] \\ 0-\mathbf{X}\boldsymbol{\beta}, & \text{if } \mathbf{X}\boldsymbol{\beta} \leq 0 \end{cases} \tag{14}$$

$\mathbb{E}(\boldsymbol{\epsilon}|\mathbf{X}) = 0$, the *zero conditional mean* assumption for unbiased estimation with OLS (Wooldridge, 2010, p. 50), is only met when all values in $\mathbf{X}\boldsymbol{\beta}$ are in $[0,1]$.

Derivation for Poisson working model

We have yet to see this derivation presented elsewhere in the literature on binary response models. The key question of interest is: $P(\mathbf{y}=1|\mathbf{X})$? From equation (11), we know that: $P(\mathbf{y}=1|\mathbf{X}) = 1 - P(\boldsymbol{\upsilon} < \tau - \mathbf{X}\boldsymbol{\beta}|\mathbf{X})$. For model identification, we set threshold, $\tau = -1$, such that the right hand side of the equation becomes $1 - P(\boldsymbol{\upsilon} < -1 - \mathbf{X}\boldsymbol{\beta}|\mathbf{X})$. We assume the error term plus 1 is exponential with a scale parameter of 1, $\boldsymbol{\upsilon} + 1 \sim \exp(1)$. Adding 1 to the error term ensures $\text{mean}(\boldsymbol{\upsilon}) = 0$ since $\text{mean}(\exp(1)) = 1$. If we increment $\boldsymbol{\upsilon}$ by 1, then we must increment the RHS of the equation:

$$P(\mathbf{y}=1|\mathbf{X}) = 1 - P(\boldsymbol{\upsilon}+1 < -1 - \mathbf{X}\boldsymbol{\beta} + 1|\mathbf{X}) = 1 - P(\boldsymbol{\upsilon}+1 < -\mathbf{X}\boldsymbol{\beta}|\mathbf{X}) \tag{15}$$

Based on the exponential distribution function with scale parameter of 1:

$$\begin{aligned} P(\boldsymbol{\upsilon}+1 < -\mathbf{X}\boldsymbol{\beta}|\mathbf{X}) &= \begin{cases} 0, & \text{if } -\mathbf{X}\boldsymbol{\beta} < 0 \\ 1-e^{-(-\mathbf{X}\boldsymbol{\beta})}, & \text{if } -\mathbf{X}\boldsymbol{\beta} \in [0,\infty) \end{cases} \\ &= \begin{cases} 0, & \text{if } \mathbf{X}\boldsymbol{\beta} > 0 \\ 1-e^{\mathbf{X}\boldsymbol{\beta}}, & \text{if } \mathbf{X}\boldsymbol{\beta} \in (-\infty,0] \end{cases} \end{aligned} \tag{16}$$

From equation (15), $P(\mathbf{y}=1|\mathbf{X}) = 1 - P(\boldsymbol{\upsilon}+1 < -\mathbf{X}\boldsymbol{\beta}|\mathbf{X})$, then:

$$P(\mathbf{y}=1|\mathbf{X}) = \begin{cases} 1, & \text{if } \mathbf{X}\boldsymbol{\beta} > 0 \\ e^{\mathbf{X}\boldsymbol{\beta}}, & \text{if } \mathbf{X}\boldsymbol{\beta} \in (-\infty,0] \end{cases} \tag{17}$$

Poisson regression estimates $e^{\mathbf{X}\boldsymbol{\beta}}$ without respect for the zero upper bound, hence the expectation for the regression error term, $\boldsymbol{\epsilon}$, is:

$$\mathbb{E}(\boldsymbol{\epsilon}|\mathbf{X}) = P(\mathbf{y}=1|\mathbf{X}) - e^{\mathbf{X}\boldsymbol{\beta}} = \begin{cases} 1-e^{\mathbf{X}\boldsymbol{\beta}}, & \text{if } \mathbf{X}\boldsymbol{\beta} > 0 \\ 0, & \text{if } \mathbf{X}\boldsymbol{\beta} \in (-\infty,0] \end{cases} \tag{18}$$

$\mathbb{E}(\boldsymbol{\epsilon}|\mathbf{X}) = 0$, the *zero conditional mean* assumption for unbiased estimation with Poisson regression, is only met when all values in $\mathbf{X}\boldsymbol{\beta}$ are in $(-\infty, 0]$.

Appendix B

Complete simulation results for linear probability model and Poisson working model

Table B1

Complete results for simulation comparing OLS and SLS when both predictors have a correlation of 0.3

	True coefficients				Binary		Continuous	
Condition	Binary	Continuous	% outside 0-1	Model	Bias	ECR	Bias	ECR
1	0	0	0.0%	Both	0.000	94.5%	0.001	94.8%
2	0.1	0	0.0%	Both	0.000	94.9%	0.001	95.0%
3	0	0.25	0.0%	Both	0.000	95.2%	0.000	94.8%
4	0.1	0.25	0.0%	Both	0.000	95.0%	-0.001	94.8%
5	0	0.5	2.6%	OLS	0.000	94.9%	0.000	94.5%
5				SLS	0.000	94.5%	0.004	93.8%
6	0.1	0.5	7.4%	OLS	-0.006	94.2%	-0.010	94.0%
6				SLS	0.000	94.2%	0.005	94.0%
7	0	0.75	21.7%	OLS	-0.005	94.5%	-0.110	0.0%
7				SLS	-0.001	94.1%	0.013	93.7%
8	0.1	0.75	21.8%	OLS	-0.034	81.3%	-0.128	0.0%
8				SLS	0.001	94.7%	0.014	93.0%

Table B2

Complete results for simulation comparing Poisson and sequential Poisson when both predictors have a correlation of 0.3

	True coefficients				Binary		Continuous	
Condition	Binary	Continuous	% outside 0-1	Model	Bias	ECR	Bias	ECR
1	0	0	0.0%	Both	-0.004	95.3%	0.002	94.9%
2	0.5	0	0.0%	Both	0.003	95.0%	0.001	95.0%
3	0	0.5	0.0%	Both	0.002	95.0%	0.000	95.1%
4	0.5	0.5	1.5%	Poisson	0.004	94.7%	0.000	94.8%
4				Seq. Poisson	0.007	94.5%	0.006	94.1%
5	0	1	1.3%	Poisson	0.001	95.0%	0.004	95.1%
5				Seq. Poisson	0.000	94.9%	0.012	94.8%
6	0.5	1	11.5%	Poisson	-0.089	79.5%	-0.160	45.6%
6				Seq. Poisson	0.008	94.8%	0.018	93.6%