# Using Poisson regression and OLS to obtain effect sizes when analyzing binary outcomes

James Uanhoro <uanhoro.1@osu.edu>

Educational Studies (Quantitative Research, Evaluation and Measurement), The Ohio State University

## Why not logistic regression?

The goto statistical method for analyzing binary outcomes in education is logistic regression, and the most commonly reported effect size from this model is the odds ratio (OR). There are two reasons to re-think this practice:

1. Based on a reading of odds ratio interpretations in research reports, odds ratios may not be the effect size of interests for substantive researchers.
2. When analyzing the same outcome, the OR for a predictor variable in one model is not comparable to the OR for the same variable in another model with different predictor variables.

## Are odds ratio the effect of interest?

As noted by Osborne (2006), education researchers commonly interpret odds ratios – a ratio of odds – using "likely" language. Consider the following example:

| | Pass | Fail |
|---|---|---|
| Treatment | 200 | 100 |
| Control | 160 | 140 |

Table 1:Data from an education intervention, values are student counts

$$\text{OR} = \frac{Pass_{treat}}{Fail_{treat}} \div \frac{Pass_{control}}{Fail_{control}} = \frac{200}{100} \div \frac{160}{140} = 1.75 \quad (1)$$

The correct interpretation is: "the odds of success of students in the treatment group was on average 1.75 times the odds of success of students in the control group." To use "likely" language, researchers would have to calculate the *risk ratio* (RR) – a ratio of probabilities:

$$\text{RR} = \frac{Pass_{treat}}{All_{treat}} \div \frac{Pass_{control}}{All_{control}} = \frac{200}{300} \div \frac{160}{300} = 1.25 \quad (2)$$

with the interpretation, "students in the treatment group were on average 25% $\left((1.25 - 1) \times 100\%\right)$ more likely to succeed than their counterparts in the control group." In this example, misinterpreting 1.75 as the RR would have resulted in a 40% inflation of the treatment effect. If $p$ is the probability of success on an outcome, then the odds of success on the outcome are $\frac{p}{1-p}$. Hence, $p \approx \frac{p}{1-p}$ as $p$ tends to zero. It follows that the misinterpretation of OR using "likely" language is an acceptable approximation for low rate events.

Another effect size that may be of interest to applied researchers is the *risk difference* (RD):

$$\text{RD} = \frac{Pass_{treat}}{All_{treat}} - \frac{Pass_{control}}{All_{control}} = \frac{200}{300} - \frac{160}{300} = 0.1\overline{3} \quad (3)$$

## ORs are not comparable across models

Consider the following data generation process (DGP) for the probability, $p$, parameter of a Bernoulli distribution:

$$p = (1 + e^{-(0.5 \times x_b + 1.5 \times x_c)})^{-1}, \quad \text{where} \quad x_b \sim B(1, .5) \quad \text{and} \quad x_c \sim U(-\sqrt{12}/2, \sqrt{12}/2) \quad (4)$$

I simulated data according to this DGP with $n = 350$ and 4999 replications. Assuming $x_b$ as the primary variable of interest, I performed two logistic regression models within each replication, varying the predictors: $x_b$ only; and both $x_b$ and $x_c$. See Figure 1 for the outcome of this process.

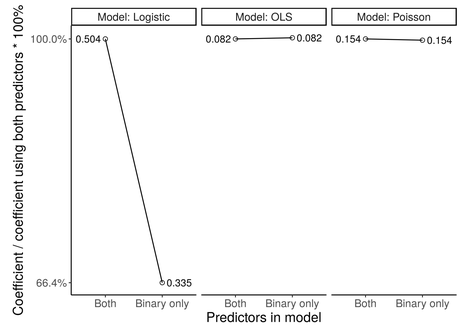


Figure 1:All models include a constant predictor. Point labels are average regression coefficients across replications. Log odds, hence ORs, are not comparable across models.

It is easy to see the reason for the result in Figure 1 using the latent variable formulation of logistic regression:

$$y^* = \beta_0 + \beta_b x_b + \beta_c x_c + \sigma\epsilon \quad \text{and} \quad y = \begin{cases} 1, & \text{if } y^* > 0 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where $\epsilon \sim \text{Logistic}(0, 1)$ and $\sigma^2$ is a ratio of the actual error variance to the error variance of Logistic$(0, 1)$, $\pi^2/3$. In practice, the error variance is always fixed to $\pi^2/3$ for model identification, hence, the actual model that is estimated is:

$$y^*/\sigma = \frac{\beta_0}{\sigma} + \frac{\beta_b}{\sigma} x_b + \frac{\beta_c}{\sigma} x_c + \epsilon \quad (6)$$

When we correctly specify the model using $x_b$ and $x_c$, $\sigma = 1$ and the estimated coefficients, $b = \beta$. When we omit $x_c$:

$$\sigma = \sqrt{\frac{\beta_c^2 \text{var}(x_c) + \pi^2/3}{\pi^2/3}} \quad \text{and} \quad b_b = \beta_b \sqrt{\frac{\pi^2/3}{\beta_c^2 \text{var}(x_c) + \pi^2/3}} \quad (7)$$

The more predictors of an outcome are missing from a model, the more shrunken are the regression coefficients of the variables in the model. (7) assumes as in (4) that $x_b$ and $x_c$ are independent. This problem is described in the econometrics literature as *neglected heterogeneity* (Wooldridge, 2002) and in the sociology literature as *unobserved heterogeneity* (Mood, 2010).

## Poisson regression

As seen in Figure 1, Poisson regression coefficients are comparable across models. The Poisson model for binary data is of the form:

$$log(p) = \beta_0 + \beta_b x_b + \beta_c x_c \quad \text{and} \quad \text{var}(Y) = p \quad (8)$$

Since the model is additive on the log-probability scale, when we exponentiate the coefficients, we obtain a multiplicative comparison of probabilities or the RR. Hence, the results from Figure 1 suggest that on average, cases with $x_b = 1$ were about 17% $\left((e^{.154} - 1) \times 100\%\right)$ more likely to succeed on the outcome than cases with $x_b = 0$.

The model for the variance is misspecified, which affects statistical inference. The variance of a binary variable is $p(1 - p)$ which is always less than $p$. However, same as OR $\approx$ RR as $p$ tends to 0, $p(1 - p) \approx p$ as $p$ tends to 0. Approaches to coping with variance overestimation include heteroskedasticity-correct standard errors (HCSE; Zou (2004)), a quasi-Poisson approach, and likelihood based approaches that estimate dispersion parameters such as generalized Poisson and Conway-Maxwell Poisson. In any one application, simulation-based model-diagnostics (Gelman & Hill, 2007) can help reveal the extent to which competing models capture relevant features of the data.

## OLS estimation

As seen in Figure 1, OLS estimation coefficients are comparable across models. The model underlying OLS estimation alongside requirements for statistical inference are:

$$Y = \beta_0 + \beta_b x_b + \beta_c x_c + \epsilon \quad (9) \qquad \epsilon \sim N(0, \sigma^2) \quad (10)$$

Since the model is additive on the probability scale, it is also known as the *linear probability model*, and the coefficients are an additive comparison of probabilities or the RD. Hence, the results from Figure 1 suggest that on average, cases with $x_b = 1$ had an 8.2% higher probability of success on the outcome than cases with $x_b = 0$. Horrace and Oaxaca (2006) provided proof of the bias and inconsistency of the RD obtained from the LPM. Practically, the utility of the RD obtained in this manner depends on the extent to which the predicted probabilities from the model exceed the unit interval.

Additionally, the model for the variance is misspecified, which affects statistical inference. The variance of a binary variable varies as a function of $p$, however, OLS estimates the variance as a constant. As a general remedy, Cheung (2007) recommended HCSEs.

The presence of high leverage cases can result in RDs that misrepresent the data, hence, outlier detection is important.

## Statistical power of different methods

Using the sample size, number of replications and DGP as in Figure 1, I performed model comparisons, see Figure 2.

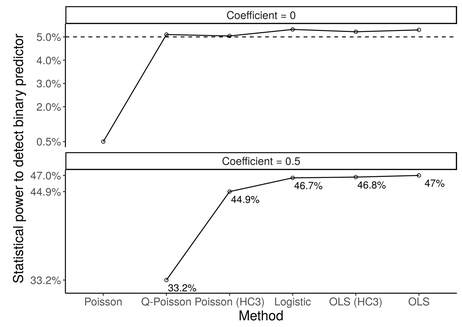


Figure 2:Statistical power of different methods. Tests were based on models with both predictors. Q-Poisson = quasi-Poisson, HC3 = HCSE Type III. Poisson was not able to maintain the nominal error rate under a nil effect, hence, I have not reported its power when the effect was non-nil. All methods considered except for Poisson and quasi-Poisson had comparable power to logistic regression when the effect was non-nil.

## Conclusion

"Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise." (Tukey, 1962). Poisson regression and OLS though clearly wrong may provide informative measures of effect size during model building that are not only comparable across models and studies and but may ease the communication of results. Simulation & chart script (R): ghostbin.com/paste/odqo7

## References

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

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

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

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

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`

Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics, 33*(1), 1–67. Retrieved from `https://www.jstor.org/stable/2237638`

Wooldridge, J. M. (2002). Econometric Analysis of Cross Section and Panel Data. *The MIT Press*, 735. doi: 10.1515/humr.2003.021

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

## Acknowledgements

I thank Professor Ann O'Connell for her support of my graduate career.