Four Reasons To Use Logit*

Christopher Zorn
Department of Political Science
Pennsylvania State University
Contact: zorn@psu.edu
Version 1.1

September 22, 2016

Abstract

In choosing among regression models for binary outcomes, there are a number of little-known but ultimately compelling reasons for selecting logit regression over its alternatives.

*I thank Brian Habing and Cyrus Samii for useful conversations. All errors are my own.

Overview: Logit and Probit

Despite increased recent attention to alternative approaches, regression-based models remain the dominant approach to adjusting for observable confounders in quantitative analyses. This is particularly true among analysts working with observational data and in other contexts where experimental manipulation and other design-based approaches are infeasible. When the outcome of interest is binary, the workhorse regression models are the familiar logit and probit variants of generalized linear models (McCullagh and Nelder 1989).[1] For a binary response Y observed for $i = \{1, 2, ...N\}$ observations, and a corresponding $N \times k$ matrix of predictors $\mathbf{X}$, logistic regression for a binary response ("logit") is typically written as:

$$\Pr(Y_i = 1) = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})} \tag{1}$$

and probit as:

$$\Pr(Y_i = 1) = \Phi(\mathbf{X}_i\boldsymbol{\beta}). \tag{2}$$

where Φ denotes the cumulative standard normal function. These models can be motivated in a number of ways, but are often presented as reflecting an underlying latent (continuous) phenomenon that is measured discretely (e.g. Long 1997, 40-47).

The similarity of logit and probit is widely known and discussed (e.g., Chambers and Cox 1967).[2] Most expositions note that estimates $\hat{\boldsymbol{\beta}}$ from logit and probit models are related by a scale

[1] Despite its recent rehabilitation in some quarters, I do not address the linear probability model ("LPM") here.

[2] Mathematically, logit and probit are related in a particular way; Andrews and Mallows (1974) demonstrate that mixtures of independent normal distributions, where the mixing distribution is the Kolmogorov-Smirnov distribution, yield the logistic distribution (see also Poirier 1978; West 1987; Stefanski 1991).

factor,[3] and that inferences about effect sizes, predicted probabilities, and other quantities of interest are typically very similar,[4] with the result that "either model will give identical substantive conclusions in most applications" (Liao 1994, 24). Because of these similarities, nearly all textbooks present the choice between these two regression models as a low-stakes affair. Gelman and Hill call the choice "a matter of taste or convenience" (2007, 199), while Aldrich and Nelson state that "there is little to guide the choice between the two" (1984, 35). Long provides a better-than-average summary of the "textbook" perspective. He writes:

> "(T)he choice between the logit and probit models is largely one of convenience and convention, since the substantive results are generally indistinguishable... Often the choice is a matter of convention. Some research areas tend to use logit, while others favor probit. For some users the simple interpretation of logit coefficients as odds ratios is the deciding factor. In other cases, the need to generalize a model may be an issue. For example, multiple-equation systems involving qualitative dependent variables are based on the probit model, as discussed in Chapter 9. Or, if an analysis also includes equations with a nominal dependent variable, the logit model may be preferred since the probit model for nominal dependent variables is computationally too demanding. Or, in case-control studies where sampling is stratified by the binary outcome, the logit model is required" (Long 1997, 83).

An (admittedly impressionistic) survey of published research in political science suggests that probit and logit models are both common, but that the latter has overtaken the former in popularity during the past two decades. We find similar balance in sociology, social psychology, communi-

[3] Camilli (1994) retraces the origins of the widely-used scaling constant $d = 1.702$ for translating between the two CDFs.

[4] E.g., Aldrich and Nelson (1984, 65), who note that "(O)nly if there are a lot of observations at extreme probability values will the two estimation techniques differ noticeably, for the probit and logit functional bases are essentially identical at all but the tails of their respective distributions (and even there the differences are but slight)."

cation, and most other social sciences. The lone exception is in economics, where probit models dominate; one possible explanation for that dominance is the common econometric motivation of binary-response models as models of individual choice (e.g. Judge et al. 1985, 761-762), with its corresponding affinity for Normally-distributed errors. Outside of the social sciences, the logit model sees far greater use, especially in biological and medical fields.

For most regression applications with observational data, then, the choice between logit and probit *seems* of little consequence. In fact, there remain a number of ultimately compelling reasons to prefer logit to probit when fitting regression models for binary outcomes.

Reason #1: Maximum Entropy

It is typically the case that, at least with non-simulated data, we generally neither know nor have strong theoretical reasons to suspect a particular distribution characterizes the stochastic component of the latent continuous response. Given this fact, it is generally the case that "theory" alone cannot arbitrate among probit, logit, and other alternatives. Yet, perhaps ironically, it is exactly the *lack* of a strong theoretical justification for any particular choice of distribution that provides the first argument in favor of logit.

In an influential article, Jaynes (1957) set forth the principle of "maximum entropy" as a method of guiding the choice of probability distribution in the absence of compelling theoretical bases. Building on LaPlace's "Principle of Insufficient Reason," and on early work by Gibbs, Shannon, and others in information theory, the principle states that "in making inferences on the basis of partial information we must use that probability distribution which has maximum entropy subject to whatever is known" (1957, 623). A number of subsequent authors have shown that Jaynes' maximum entropy distribution corresponds to the multinomial logistic model for nominal outcomes, of which binary regression is a special case (see e.g. Good 1963; Collins et al. 2002; Donoso and de Grange 2010). In fact, models based on the logistic distribution (including both bi-

nary and multinomial logit) are commonly referred to as "MaxEnt" models in the machine learning and natural language processing communities.

Leblanc and Shapiro summarize the substantive significance of this fact, noting that logit "is the optimal model available under the maximum entropy principle. It is the most conservative approach for the analysis that can be taken in the given context, that is considering the available information... Logistic analysis in that sense is not resting or relying on a very sophisticated mathematical tool for analyzing discrete data, but rather simply on the best model to use when one is driven by a data-oriented approach to describe the situation observed, given the information" (1999, 61). Jaynes' fundamental insight, that "the fact that a probability distribution maximizes the entropy subject to certain constraints becomes the essential fact which justifies use of that distribution for inference" (1957, 621), is therefore especially likely to be relevant when considering models for data which are weakly theorized and/or not subject to manipulation by the researcher; it is no particular exaggeration to note that such circumstances are very common in social scientific applications.

Reason #2: Invariance Under Sampling Schemes

A second benefit of logistic regression is its ability to consistently estimate model parameters under different sampling schemes. In particular, it is well-understood that estimates of $\hat{\beta}$s (other than the intercept) from logistic regression are valid under either prospective or retrospective ("case-control," or "choice-based") sampling plans (see e.g. Prentice and Pyke 1979; Cosslett 1981; Kagan 2001). McCullagh and Nelder (1989, 111-114) and Agresti (2002, 170-171) provide lucid summaries of this characteristic of the logit model, which is rooted in the central role of odds ratios in the model's exposition (Cornfield 1951). Kagan (2001) demonstrates that the logistic model is the *only* link function for generalized linear models of binary responses that posesses this characteristic; more recently, Osius establishes a class of log-bilinear association models which are

also invariant to sampling on X and Y, while noting that "the logistic regression model is the only one among generalized linear models for binary Y which is equivalent to an association model" (2009, 468).

While this property of logistic regression is not unknown in political science,[5] its significance is not widely appreciated. For example, one implication of this property is that logit preserves the marginal probabilities in the sample data. This means that – like OLS,[6] but unlike probit – predicted probabilities generated from logit estimates of $\hat{\beta}$ are consistent estimators of treatment effects (Freedman 2008, 242-246). More generally, the increased use of retrospective / case-control sampling in political science (including in field experiments and other quasi-experimental contexts where the interest is in causal inference) counsel wider use of the logistic regression model.

Reason #3: It's the Canonical Link

For generalized linear regression models with a binary response, the canonical link function is the logit (McCullagh and Nelder 1989, 30-32). Canonical-link GLMs have a number of advantages that are not widely appreciated. Statistically, use of the canonical link ensures that the resulting estimator can be expressed in terms of sufficient statistics (McCullagh and Nelder 1989, 115-16). In addition, use of the canonical link in GLMs means that the expected and actual values of the Hessian matrix are identical, and that Newton-Raphson and Fisher scoring algorithms are the same (McCullagh and Nelder 1989, 43). This means that logit-based estimates of standard errors will be the same whether one uses the observed or expected information matrix, while the same is not true for probit or other models (Agresti 2002, 247).

More recent work underscores the practical value of canonical-link GLMs. Firth and Bennett derive a class of "internally bias-calibrated" models that are asymptotically design-consistent; that

[5]For example, it is noted in passing in King and Zeng (2001, 160).

[6]Importantly, Firth and Bennett (1998, 19) find that, among design-consistent estimators for binary responses, the logit model is consistently more efficient than the LPM.

is, they are "approximately unbiased for (the population quantity) regardless of whether the corresponding linear model approximately represents the population regression" (1998, 4). They note that a number of canonical-link GLMs belong to this class, including the linear-Gaussian model with an identity link and the binomial model with the logit link; this means that "the maximum likelihood fit of a suitably specified logistic regression can be used directly to yield a design-consistent estimator of T, but the same is not true of, for example, probit or complementary log-log-regression" (Firth and Bennett 1998, 5). As in the discussion of maximum entropy above, this design consistency characteristic is a particularly useful trait in the theory-poor contexts in which much observational data analysis in political science is conducted.

Reason #4: Extensibility and Ease of Interpretation

"The advantage of the logit is that it is easier to interpret, since effects on the logistic scale can be expressed as odds ratios" (Francis and Payne 1977, 244). The value of odds ratios as a straightforward, easily-understood means of understanding the substantive influence of covariates in non-linear binary-response regression models has been a theme in political science for nearly four decades, and in statistics even longer. For a binary covariate X, the exponentiated logit estimate $\exp(\hat{\beta})$ reflects the expected change in the odds of $Y = 1$ associated with a one-unit change in X; "(N)o such simple interpretation exists for other links such as the probit" (Faraway 2006, 32). While debate about the interpretive value of absolute versus relative risk estimates remains open, the utility of odds ratios as a summary of the substantive marginal association between a response and a regressor cannot be understated.

Beyond ease of interpretation, logit models also offer at least as great a range of extensions as their normal-based counterparts. For ordinal responses, there is little to distinguish logistic models from their Normal-based analogues. As described above, multinomial logit alone satisfies the maximum-entropy criterion, while logit-based alternatives to multinomial logit (e.g. Zeng 2000)

have the dual advantage of computational simplicity and relaxation of the independence of irrelevant alternative assumption. On the nonparametric front, ther is no probit-like analogue to exact logistic regression (Mehta and Patel 1995), which allows for combinatoric-based inference about regression parameters in models with binary outcomes and provides finite estimates of parameter values even in the presence of perfect separation.

Summary Thoughts

In a spirited defense of the logistic model, the eminent statistician Joseph Berkson (who coined the term "logit" in 1944) noted that "(I)t is not that the logistic function is necessarily the physical law of all sigmoidally represented phenomena... It is rather that the logistic function refers to a wide range of phenomena, the intimate physical mechanisms of which are different in different cases" (1951, 334). His statement is reminiscent of George Box's famous comment about the verity and value of models in general, and presages a number of the more technical benefits of the logit model described above.

Of course, as with any statistical model, there remain reasons for caution. For example, unlike in the Gaussian-linear case, likelihood-ratio and Wald tests are not identical in the logit case (Hauck and Donner 1977), and Freedman (2008) demonstrates that both linear regression and binary-response GLMs can be misleading when used to estimate treatment effects. But these are also issues with other binary-response GLMs; in the main, logit regression models offer a wider range of potential benefits – and none of the costs – of their more commonly-used alternatives, and do so while being computationally straightforward and simpler to interpret.

References

Agresti, Alan. 2002. *Categorical Data Analysis*, 2nd Ed. New York: Wiley.

Aldrich, John H., and Forrest D. Nelson. 1984. *Linear Probability, Logit, and Probit Models*. Newbury Park, CA: Sage.

Berkson, Joseph. 1951. "Why I Prefer Logits To Probits." *Biometrics* 7(4):327-339.

Byrne, Simon P. J., and A. Philip Dawid. 2014. "Retrospective-Prospective Symmetry in the Likelihood and Bayesian Analysis of Case-Control Studies." *Biometrika* 101(1):189-204.

Camilli, Gregory. 1994. "Origin of the Scaling Constant $d = 1.7$ in Item Response Theory." *Journal of Educational and Behavioral Statistics* 19(3):293-295.

Chambers, Elizabeth A., and David R. Cox. 1967. "Discrimination Between Alternative Binary Response Models." *Biometrika* 54(3/4):573-578.

Collins, Michael, Robert E. Schapire, and Yoram Singer. 2002. "Logistic Regression, AdaBoost and Bregman Distances." *Machine Learning* 48:253-285.

Cornfield, J. 1951. "A Method of Estimating Comparative Rates from Clinical Data; Applications to Cancer of the Lung, Breast, and Cervix." *Journal of the National Cancer Institute* 11:1269-1275.

Cosslett, Stephen R. 1981. "Maximum Likelihood Estimator for Choice-Based Samples." *Econometrica* 49(5):1289-1316.

Donoso, Peter, and Louis de Grange. 2010. "A Microeconomic Interpretation of the Maximum Entropy Estimator of Multinomial Logit Models and Its Equivalence to the Maximum Likelihood Estimator." *Entropy* 12:2077-2084.

Faraway, Julian J. 2006. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. New York: Chapman and Hall.

Francis, John G., and Clive Payne. 1977. "The Use of the Logistic Model In Political Science: British Elections, 1964-1970." *Political Methodology* 4(3):233-270.

Freedman, David. 2008. "Randomization Does Not Justify Logistic Regression." *Statistical Science* 23(2):237-249.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Good, I. J. 1963. "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables." *The Annals of Mathematical Statistics* 34(3):911-934.

Hauck, Walter W. Jr., and Allan Donner. 1977. "Wald's Test as Applied to Hypotheses in Logit Analysis." *Journal of the American Statistical Association* 72(360):851-853.

Jaynes, E. T. 1957. "Information Theory and Statistical Mechanics." *The Physical Review* 106(4):620-630.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T.-C. Lee. 1985. *The Theory and Practice of Econometrics*, 2nd ed. New York: Wiley.

Kagan, Abram. 2001. "A Note On The Logistic Link Function." *Biometrika* 88(2):559-601.

King, Gary, and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis* 9(2):137-163.

Liao, Tim Futing. 1994. *Interpreting Probability Models*. Thousand Oaks, CA: Sage Publications.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.

McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd Ed. London: Chapman & Hall.

Mehta, C. R., and N. R. Patel. 1995. "Exact Logistic Regression: Theory and Examples." *Statistics in Medicine* 14(19):2143-2160.

Osius, Gerhard. 2009. "Asymptotic Inference for Semiparametric Association Models." *The Annals of Statistics* 37(1):459-489.

Poirier, Dale J. 1978. "A Curious Relationship between Probit and Logit Models." *Southern Economic Journal* 44(3):640-641.

Prentice, R. L., and R. Pyke. 1979. "Logistic Disease Incidence Models and Case-Control Studies." *Biometrika* 66(3):403-411.

Stefanski, Leonard A. 1991. "A Normal Scale Mixture Representation of the Logistic Distribution." *Statistics and Probability Letters* 11(1):69-70.

West, Mike. 1987. "On Scale Mixtures of Normal Distributions." *Biometrika* 74(3):646-648.

Zeng, Langche. 2000. "A Heteroscedastic Generalized Extreme Value Discrete Choice Model." *Sociological Methods and Research* 29(1):118-144.