EVALUATION TECHNICAL ASSISTANCE BRIEF
for OAH & ACYF Teenage Pregnancy Prevention Grantees

December 2014 • Brief 6

Using the Linear Probability Model to Estimate Impacts on Binary Outcomes in Randomized Controlled Trials

Many researchers are unsure of whether the linear probability model (LPM) – that is, using the same linear regression methodology for a binary outcome that is used for a continuous outcome – is appropriate in the context of calculating impacts on binary outcomes in a randomized controlled trial (RCT). ***The purpose of this brief is to provide those researchers with a technical explanation for why the LPM is appropriate in that context. For less technical readers, we hope that the findings presented at the end of the brief are helpful in selecting an impact estimation method.***

In this brief we examine methodological criticisms of the LPM in general and conclude that these criticisms are not relevant to experimental impact analysis. We also point out that the LPM has advantages in terms of implementation and interpretation that make it an appealing option for researchers conducting experimental impact analysis. An important caveat on these conclusions is that outside of the context of impact analysis, there can be good reasons to avoid using the LPM for binary outcomes.[1]

A. The Linear Probability Model (LPM)

The LPM is simply the application of ordinary least squares (OLS) to binary outcomes instead of continuous outcomes. Equation 1 provides an example of the LPM in the context of experimental impact estimation, where Y is the outcome, T is a binary indicator of treatment status, X is a covariate, β_T is the impact on Y of being assigned to the treatment group and β_X is the mean marginal effect of X on Y.[2]

(1) $$Y = \beta_0 + \beta_T T + \beta_X X$$

By way of comparison, equation 2 provides an example of logistic regression involving the same variables. In equation 2 the interpretation of β_T and β_X are different than in equation 1. Whereas in equation 1 these parameters represent mean marginal effects, in equation 2 they represent a much more difficult to interpret parameter called the "log odds ratio".[3] To calculate marginal effects for a logistic regression, we must take the derivative of equation 2 with respect to the variable of interest (either T or X). Then, to calculate *mean* marginal effects, we must calculate that derivative for every data point and then calculate the mean of those derivatives.[4]

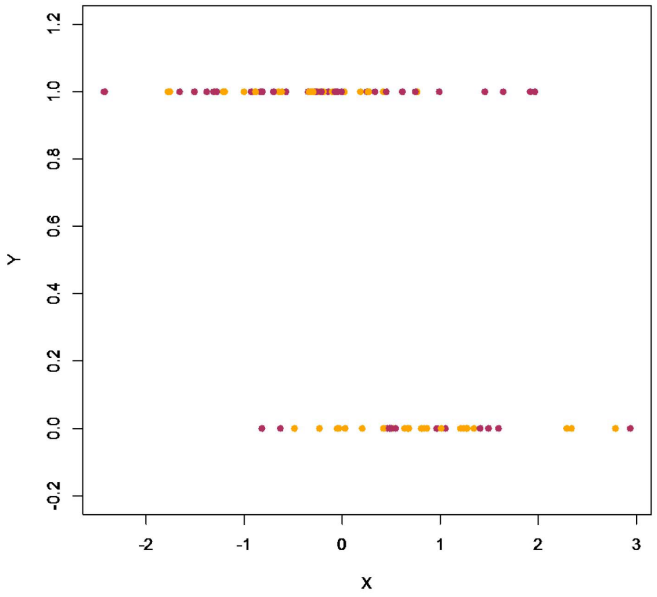
(2) $$\text{Prob}(Y = 1) = \frac{e^{\beta_0 + \beta_T T + \beta_X X}}{1 + e^{\beta_0 + \beta_T T + \beta_X X}}$$

We illustrate how these two different approaches model the relationship between a binary outcome and covariates using artificially generated data (described in the appendix). In Figure 1 we present a scatter-plot of the artificial data (X is on the horizontal axis and represents knowledge about sexually transmitted infections (STIs); Y is on the vertical axis and represents whether or not a youth had unprotected sex or not; T is illustrated via color coding – orange denotes T=1, purple denotes T=0). In Figure 2, we add to that scatter plot predicted probabilities from a logistic regression (predicted probabilities are denoted by the symbol ^). In figure 3, we show predicted probabilities from the LPM.

In both Figures 2 and 3 the impact of treatment (that is, the marginal effect of T) for a particular value of X is the vertical distance between the purple and orange prediction curves/lines. An obvious difference between Figures 2 and 3 is that in Figure 2 the impact (marginal effect of T) varies with respect to X, while in Figure 3 the impact is constant with respect to X. What may not be obvious from simply looking at the figures, however, is that the average vertical distance between these curves (that is, the *mean* marginal effect) is the same in both Figures 2 and 3. That is, both logistic regression and the LPM yield the same expected *average* impact estimate. We will examine this in more detail in section C.

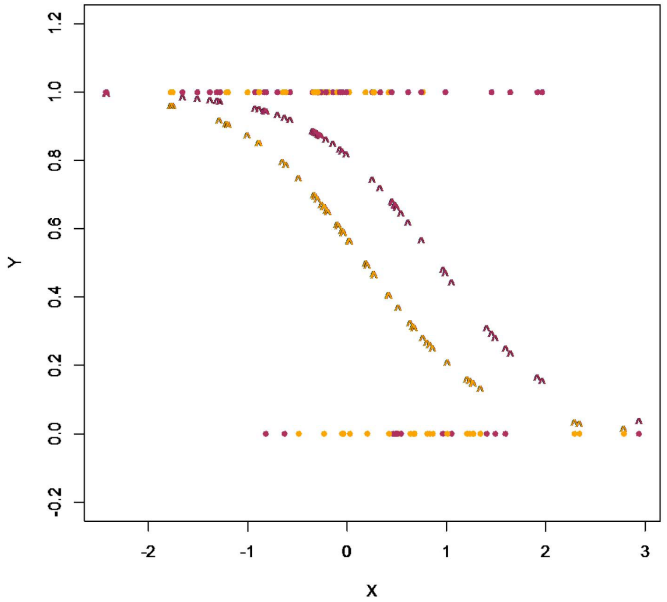
Office of Adolescent Health

Figure 1. Scatterplot of Artificial Data



Note: The data points in the figure are color coded by treatment status. The treatment group is represented by the color orange and the control group is represented by the color purple.

Figure 2. Predicted Probabilities Using Logistic Regression

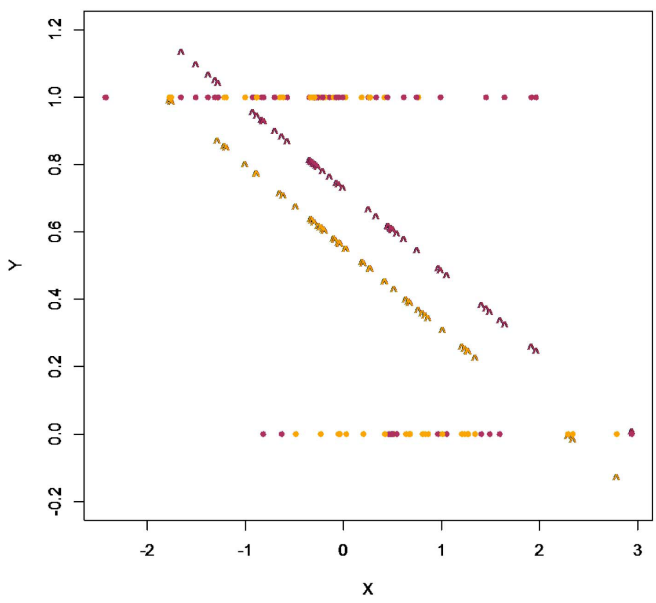


Note: The data points in the figure are color coded by treatment status. The treatment group is represented by the color orange and the control group is represented by the color purple. Predicted probabilities are denoted by the symbol ^ and use the same color-coding.

B. Textbook Advantages and Disadvantages of the LPM for Binary Outcomes

Textbooks (for example, Greene 1993) describe the advantages and disadvantages of the LPM for general use. The main advantage of

Figure 3. Predicted Probabilities Using the LPM



Note: The data points in the figure are color coded by treatment status. The treatment group is represented by the color orange and the control group is represented by the color purple. Predicted probabilities are denoted by the symbol ^ and use the same color-coding.

the LPM is that the parameter estimates can be directly interpreted as the "mean marginal effect" of covariates on the outcome. For example, β_T from Equation 1 is the difference between the treatment and control groups in the prevalence rate of the outcome (the mean vertical distance between the purple and orange prediction ^ lines in Figure 3). This result can be used in statements of the impact of an intervention that can be easily understood by a broad audience. For example: "Youth in the treatment group reported a sexual initiation rate 7 percentage points lower than youth in the control group." By contrast, β_T from equation 2 is a log odds ratio, *not* the mean marginal effect. Consequently β_T from equation 2 does *not* correspond to the mean vertical distance between the purple and orange curves in Figure 2.

The main disadvantage of the LPM that is described in textbooks is that the true relationship between a binary outcome and a continuous explanatory variable is inherently nonlinear.[5] This means that the functional form of the LPM is generally not correctly specified, which can lead to biased estimates of some parameters of interest. For example, estimates of the marginal effect of X for a specific value of X are often biased. This is because the LPM assumes a constant marginal effect of X for all values of X, but the marginal effect of X almost always varies with respect to X. In extreme cases this misspecification of the functional form can even lead to predicted probabilities that are less than 0 or greater than 1 (see Figure 3), or in percentage terms, less than 0 or greater than 100 percent.

C. Performance of LPM In Estimating Experimental Impacts

It turns out that the textbook disadvantages of the LPM described above do not apply to the context of estimating experimental impacts – the LPM yields estimates of experimental impacts that are just as accurate as those estimated by logistic regression. That is, in the context of an impact evaluation, the parameter of interest is β_T, not β_X, and the LPM is an appropriate analytic procedure for estimating β_T. Below we explain why this is the case and use simulations to illustrate the point.

The main reason that the LPM works so well to estimate experimental impacts is that treatment status is a binary variable (not a continuous variable, which would be subject to the potential bias described above). This means that the functional form concerns about LPM do not apply to estimating impacts, since all that is required is to estimate two prevalence rates—one for the treatment group and one for the control group (as opposed to estimating a different prevalence rate for every unique value of a continuous variable).

A second reason that the LPM provides accurate estimates of experimental impacts is that any other covariates included in the impact model are uncorrelated with treatment status (thanks to random assignment), which means that the impact estimate is unbiased regardless of whether the correct functional form is used to adjust for other (possibly continuous) covariates.[6]

We illustrate the strong performance of the LPM relative to logistic regression using Monte Carlo simulations. In each simulation, we create a binary outcome (Y), a continuous baseline covariate (X), and a treatment indicator variable (T). The prevalence rates of Y and the impact of T vary across simulations; the sample size for each of the 10,000 replications is set at 100. For expository purposes, to align with the previous presentations in Figures 1-3, Y can represent whether or not a sample member has had unprotected sex, and X can represent knowledge about STIs. We examine prevalence rates of 50 percent and 80 percent. When the prevalence rate of Y is 50 percent, we examine impacts of 0, 10, and 25 percentage points. When the prevalence rate of Y is 80 percent, we examine impacts of 0, 5, and 15 percentage points. For all simulations Y is negatively correlated with X (as in Figures 2 and 3). Additional details of these simulations are included in the appendix.

Simulation findings are reported in Table 1. There are four key findings in this table:

1. ***Logistic regression cannot always estimate an impact in cases where the LPM can.*** Logistic regression will fail to estimate an impact if treatment status perfectly predicts the outcome (for example, the outcome is equal to 1 for everyone in the treatment group). This can happen when the sample size is small and/or when the prevalence rate of the outcome is very high or low.[7] For example, when the prevalence rate is 80 percent and the true impact is 15 percentage points, an impact cannot be estimated using logistic regression for about 8 percent of our Monte Carlo replications. An implication of this failure is that in situations where the sample size is small and/or the prevalence rate of the outcomes is very high or low *yet it is possible* to estimate an impact using logistic regression, those impacts (and accompanying standard errors) will be biased (see below).

2. ***The LPM yields unbiased impact estimates in all scenarios examined; logistic regression is biased in the scenario where it sometimes fails to estimate an impact.*** For every combination of outcome prevalence rate and impact magnitude, the expected value of the impacts estimated by the LPM equals the true impact (see the two columns under the heading "Mean of Monte Carlo Impact Estimates"). Logistic regression generally yields an unbiased impact, but in the situation where logistic regression is susceptible to failure, the cases where it does not fail yield an average impact that is not equal to the true impact (an impact of 14.5 instead of 15 in the last row of Table 1).

3. ***Impacts estimated using logistic regression are slightly more precise.*** The "true" standard error of impacts estimated by each method is reported under the heading "Standard Deviation of Monte Carlo Impact Estimates." The standard errors for impacts estimated using logistic regression are slightly smaller than those estimated using the LPM.[8]

4. ***Standard errors estimated using the LPM are correct, standard errors estimated for logistic regression are sometimes too small.*** Correctly estimating standard errors is necessary for constructing accurate confidence intervals and for statistical hypothesis testing. We find that the estimated standard errors for impacts estimated by the LPM ("Mean of Monte Carlo Standard Error Estimates") are essentially the same as the true standard errors ("Standard Deviation of Monte Carlo Impact Estimates"). For logistic regression, the estimated standard errors are a little too small for all six scenarios reported in Table 1.[9]

We report additional, more technically nuanced findings from these simulations in the appendix.

Conclusions

Relative to estimating impacts for RCTs using logistic regression, the LPM has two main advantages and no disadvantages. Its main advantages are ease of implementation and interpretation. Specifically, the LPM can estimate impacts in cases where logistic regression cannot and, unlike logistic regression, the parameter estimates from the LPM can be directly interpreted as the impact of the intervention on the prevalence rate of the outcome. The

Table 1. Monte Carlo Comparison of LPM and Logistic Regression

Outcome Prevalence Rate	True Percentage Point Impact	Mean of Monte Carlo Impact Estimates		Standard Deviation of Monte Carlo Impact Estimates		Mean of Monte Carlo Standard Error Estimates		Percentage of Monte Carlo Estimates That Failed	
		LPM	Logistic	LPM	Logistic	LPM	Logistic	LPM	Logistic
50 percent	0	0.0	0.0	8.3	8.2	8.3	8.0	0	0
50 percent	10	10.0	9.9	8.6	8.5	8.5	8.3	0	0
50 percent	25	25.0	24.9	8.4	8.4	8.4	8.1	0	0
80 percent	0	0.0	0.0	7.0	6.8	7.0	6.6	0	0
80 percent	5	4.9	4.9	6.8	6.7	6.7	6.4	0	0
80 percent	15	14.9	14.5	6.0	5.8	5.9	5.6	0	8.1

Source: Monte Carlo experiments, 10,000 replications, each with sample size of 100. See appendix for detailed description.

textbook concern about functional form misspecification for the LPM does not apply to impact estimation since treatment status is a binary variable (meaning that functional form is irrelevant).

Endnotes

[1] This brief focuses on the use of LPM for binary outcomes for RCTs but the conclusions also apply to studies that use a quasi-experimental design (QED) to estimate the impact of a binary treatment variable on a binary outcome.

[2] The "marginal effect" of X on Y is the effect on Y of a small change in X (the derivative of Y with respect to X) and can be calculated for every individual in the data. The "mean marginal effect" is the average of all the individual marginal effects. With a linear regression specification the marginal effect is estimated to be the same for every individual. With a non-linear regression specification, every individual can have a different marginal effect. For example, if the regression specification is $Y = b*X$, then the marginal effect is just b (which is the same for every individual). But if the regression specification is $Y = b*X^2$, then (by calculus) the marginal effect is $2*b*X$, which means that the marginal effect depends on X and therefore varies across individuals.

[3] For example, in the case of equation (2), the log odds ratio β_T is equal to log((Prob(Y=1 | T=1)/Prob(Y=0 | T=1)) / (prob(Y=1 | T=0)/ prob(Y=0 | T=0))). This footnote highlights the complexity involved in interpreting logistic regression, which is why we suggest using the LPM.

[4] Some statistical software will calculate mean marginal effects from a logistic regression "automatically," but it is important to carefully read the manual to ensure that the software is actually calculating the mean marginal effect rather than the marginal effect at the mean and to make sure that the standard error of the mean marginal effect is calculated correctly. It is also important to verify that the software is taking into account whether the mean marginal effect is being calculated for a continuous variable or a discrete variable. This footnote highlights the complexity involved in interpreting logistic regression, which is why we suggest using the LPM.

[5] Another issue with the LPM is conditional heteroskedasticity. If standard error estimates are not adjusted for conditional heteroskedasticity they can be too large in cases where the sample is not evenly split between the treatment and control groups. Standard errors can be adjusted for conditional heteroskedasticity using the Huber (1967)-White (1980) correction.

[6] Functional form also does not matter in the context of a QED with a matched comparison group that is equivalent to the treatment group with respect to observed characteristics.

[7] This issue affects logistic regression because logistic regression is estimated using maximum likelihood, and the maximum likelihood algorithm fails to converge in this situation. It does not affect the LPM because the LPM does not rely on maximum likelihood estimation.

[8] We repeated the Monte Carlo experiment several times to confirm that the difference between the LPM and logistic regression in the variance of the impact is real.

[9] We repeated this experiment without covariate adjustment and found that the estimated standard error for the logistic regression impact was correct in that case, suggesting that logistic regression may overstate the benefit of covariate adjustment. In all scenarios, we estimated standard errors for logistic regression impacts using the delta method, which is also the approach used by the software package STATA. Some readers may be confused that covariate adjustment reduces the standard error of the impact when estimated by logistic regression since adding covariates has been shown to *increase* the standard errors of coefficients in a logistic regression. Note, however, that we are not focusing on the standard error of the *coefficient* on treatment status, but rather the standard error for the *mean marginal effect* of treatment status. See Schochet (2013) for a thorough discussion of the differences between these two parameters and the standard errors of estimates of these parameters.

References

Greene, William H. Econometric Analysis. Englewood Cliffs, NJ: Prentice Hall, 1993.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press, vol. 1, 221–233.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817–830.

Schochet, Peter Z. (2013). Statistical Power for School-Based RCTs With Binary Outcomes. *Journal of Research on Educational Effectiveness*. Volume 6, Issue 3, pages 263-294.

This brief was written by John Deke from Mathematica Policy Research for the HHS Office of Adolescent Health under contract #HHSP233201300416G.

TECHNICAL APPENDIX:

This technical appendix describes the data generating process used to create Figures 1–3 and provides a more detailed description of the simulations that were summarized in Table 1.

A. Data Generating Process Used to Create Figures 1-3

The data used to create Figures 1–3 consist of 100 observations, a normally distributed random variable e, a normally distributed random variable x, a latent dependent variable y^{latent} that is a function of x and e, a dichotomous variable T, and a dichotomous outcome variable y that is a function of T and y^{latent}, such that the proportion of observations where y equals 1 is p when T = 0 and p+impact when T = 1 (for figure 1 p = 0.5 and impact = -0.2). These variables and the relationships between them are shown in Equations 1–4 (F is the cumulative density function for the normal distribution and z is the outcome variable before the impact is added).

$$e \sim N(0,1)$$

(1) $$x \sim N(0,1)$$

$$y_i^{latent} = -x_i + e_i$$

(2) $$T \sim Bernouli(0.5)$$

(3) $$z_i = \begin{cases} 1 \text{ if } y_i^{latent} < F^{-1}_{y^{latent}}(p) \\ 0 \text{ if } y_i^{latent} \geq F^{-1}_{y^{latent}}(p) \end{cases}$$

(4) $$y_i = \begin{cases} z_i \text{ if } T_i = 0 \\ 1 \text{ if } T_i = 1,\ impact > 0, \text{ and } z_i = 1 \\ Bernouli\left(impact/(1-p)\right) \text{ if } T_i = 1,\ impact > 0, \text{ and } z_i = 0 \\ 0 \text{ if } T_i = 1,\ impact < 0, \text{ and } z_i = 0 \\ Bernouli\left((p+impact)/p\right) \text{ if } T_i = 1,\ impact < 0, \text{ and } z_i = 1 \end{cases}$$

B. Simulations Used to Create Table 1

The findings shown in Table 1 come from a Monte Carlo simulation in which impacts were calculated using the LPM and logistic regression on 10,000 data sets. Each data set had 100 observations and was generated using the same process that was used to generate figures 1-3 (described above), but with different values (see Table 1) of the outcome prevalence rate (p) and the true percentage point impact (impact). For logistic regression, the impact was calculated as the mean marginal effect of T on y from a logistic regression of y on T and x. For the LPM, the impact was calculated as the coefficient estimate on T in an ordinary least squares regression of y on T and x.

To assess the sensitivity of our main finding – that the LPM generates unbiased impact and standard error estimates – we ran additional simulations in which we varied different aspects of the data generating process described above. Specifically, we varied the distribution of x (normal, Student's t with 3 degrees of freedom, or dichotomous), we varied sample size (10, 30 or 100), and we varied the prevalence rate (50 percent, 70 percent, 80 percent, 90 percent, and 95 percent) and impact magnitude (a range from 0 to 25 percentage points). Our main findings are robust to these sensitivity analyses – the LPM generates unbiased impact and standard error estimates in all scenarios examined.