Published in final edited form as:

Biometrics. 2011 September ; 67(3): 1127–1134. doi:10.1111/j.1541-0420.2010.01536.x.

An Augmented Probit Model for Missing Predictable Covariates in Quantal Bioassay With Small Sample Size

Dean Follmann and Martha Nason¹

Dean Follmann: dfollmann@niaid.nih.gov

¹ Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, 6700B Rockledge Drive MSC 7609, Bethesda, MD 20892

Abstract

Quantal bioassay experiments relate the amount or potency of some compound; e.g. poison, antibody, or drug to a binary outcome such as death or infection in animals. For infectious diseases, probit regression is commonly used for inference and a key measure of potency is given by the ID_P , the amount that results in $P\%$ of the animals being infected. In some experiments, a validation set may be used where both direct and proxy measures of the dose are available on a subset of animals with the proxy being available on all. The proxy variable can be viewed as a messy reflection of the direct variable, leading to an errors-in-variables problem. We develop a model for the validation set and use a constrained seemingly unrelated regression (SUR) model to obtain the distribution of the direct measure conditional on the proxy. We use the conditional distribution to derive a pseudo-likelihood based on probit regression and use the parametric bootstrap for statistical inference. We re-evaluate an old experiment in twenty-one monkeys where neutralizing antibodies to HIV were measured using an old (proxy) assay in all monkeys and with a new (direct) assay in a validation set of eleven who had sufficient stored plasma. Using our methods, we obtain an estimate of the ID_1 for the new assay, an important target for HIV vaccine candidates. In simulations we compare the pseudo-likelihood estimates with regression calibration and a full joint likelihood approach.

Keywords

Assay; Errors-in-variables; HIV Vaccine; ID_{50} ; Measurement Error; Probit Regression; Regression Calibration; Validation Set

1 Introduction

A key issue in HIV vaccine development is identifying the specific immune responses needed to ensure protection from infection. Neutralizing antibodies (nABs) are likely critical and have been the focus of intense investigation. Antibodies are proteins secreted by B-cells that attach to a specific sequence of amino acids, ideally corresponding to part of the outer surface of a virus or pathogen. Each B-cell produces its own specific type of antibody that recognizes its own unique small sequence of amino acids or antigen. Neutralizing antibodies attach to their specific antigen and prevent a pathogen or virus from entering a cell, thus neutralizing an essential part of the pathogen's life cycle and preventing infection.

Nishimura et al (2002) reported the results of an experiment that estimated the amount of antibodies that resulted in 1% of the monkeys being infected or the ID_1 . The ID_1 is of great interest in vaccine development; an ideal vaccine should aim to induce this amount of antibody in vaccinees and thus protect 99% from infection. In Nishimura et al's experiments, different amounts of antibodies from a common source were (passively)

transferred to 21 uninfected monkeys. The effective circulating dose was measured by extracting plasma from the monkey and measured using a laborious assay involving infectable MT4 cells which gives a readout W . Shortly after the plasma extraction, the monkeys “challenged” with virus and infection status recorded.

The MT4 assay is time consuming and new assays have been developed that provide a rapid measure of antibody activity, say X . These TZM-bl assays are widely used to compare different vaccine candidates, but relevant ID_{50} s have not been determined. Of the 21 monkeys, in the original experiment, 11 had sufficient stored plasma to run the TZM-bl assay, while 10 monkeys had insufficient plasma. Thus for 11 monkeys we have (Y, W, X) or a validation set, while for 10 monkeys the data is $(Y, W, ?)$. The goal is to estimate the relationship between Y and X using all available data and to determine the ID_{50} and ID_{10} , for the new assay. Willey, Nason, Nishimura, Follmann & Martin (2009) reported the main results in a virology journal; here we detail the statistical underpinnings of the approach.

Statistically, this can be formulated as a kind of errors-in-variables problem. Assuming a regression relationship between X and W in the validation set of 11 monkeys we can predict X based on W for the remaining 10 monkeys, say $\hat{X}(W)$; this can be viewed as a covariate measured with error. It differs from standard errors-in-variables models in that $\hat{X}$ is not based on an average of replicates. Regression calibration could be directly applied where $\hat{X}(W)$ is used in lieu of X in the probit model (see e.g. Carroll and Stefanski 1990). This concept was unappealing to our collaborators who interpreted the approach as one where “made-up” data was used in place of “real” data. It was also felt that others in the field would have a similar reaction and perhaps be dismissive of our results, especially if the results differed from their expectations. We thus developed a more involved approach where the uncertainty in $\hat{X}(W)$ was incorporated in the estimation procedure. We adopt the approach of pseudo-likelihood (Gong & Samaniego 1981) where estimation of the prediction model is done in the first stage and then these estimates are treated as fixed constants in the second stage where maximum likelihood is used. This approach is similar to that of Follmann, Hunsberger, & Albert (1999), but they had X missing on everyone. Our approach differs from Schaefer (1993) who developed a full likelihood approach for probit errors-in-variables but for large sample size and where each observation had replicates of an error-prone covariate. Carroll & Wand (1991) had our basic data structure of a validation set (Y, W, X) and $(Y, W, ?)$ and used a logistic regression model for Y on X but assumed no parametric relationship between W and X , this is not feasible with small sample size. Other papers developing probit models in the presence of measurement error or errors-in-variables include Carroll, Spiegelman, Lan, Bailey, & Abbott (1984) and Tosteson, Stefanski, & Schafer (1989).

All of these approaches use large-sample theory for inference, which is not viable in our problem. We use profile-likelihood and the parametric bootstrap to obtain accurate tests and confidence intervals for different ID_{50} s. An additional wrinkle was that the TZM-bl assay was used on two different viruses giving a readout of interest $X_1 = X$ and a related virus X_2 . Using theoretical arguments we build a kind of seemingly unrelated regression (SUR; Zellner 1962) model for both readouts. For a general review of methods for measurement error see Carroll Ruppert Stefanski & Crainiceanu (2006); for an discussion of statistical methods for bioassay data, see Morgan (1992).

This paper is organized as follows. In section 2 we describe the experiments and assays in detail. Section 3 develops a probit regression model and associated pseudo-likelihood. We develop profile likelihood based confidence intervals for the ID_{50} in section 4 and use the parametric bootstrap for accurate small sample inference. We apply the methods to our data in section 5. In section 6 we use simulation to evaluate the performance of regression

calibration, pseudo-likelihood, and full likelihood estimates for a variety of settings. We finish with a brief discussion.

2 Experiment & Assays

In the original experiment, twenty-one monkeys were infused with differing amounts of neutralizing antibodies. Twenty four hours after antibody infusion, blood samples were drawn, some stored, and the effective amount of nABs determined using the MT4 assay. The monkeys were also 'challenged' by injection with virus. Following challenge, the infection status of each monkey was recorded.

The MT4 assay is described below. Following an initial 1:6 dilution of the plasma, serial 3-fold dilutions were performed and infectable MT4 cells were mixed with virus and the diluted plasma. This allows the virus to attempt to infect cells and replicate; if there is sufficient antibody from the plasma in the mixture, infection and thus replication cannot occur. Following 14 days, the mixture was examined for any evidence of viral replication. The procedure was performed in quadruplicate and the smallest dilution D with an estimated 50% of the mixtures showing replication was recorded giving $W = \log_{10}(D)$. Larger numbers indicate more circulating antibodies.

The MT4 assay is labor intensive and a scalable, simpler assay is gaining popularity. The new assay uses TZM-bl cells which contain the gene encoding luciferase, a substance from a firefly that causes luminescence. Virions that infect TZM-bl cells will produce a protein "Tat" that will "turn on" the luciferase gene. Thus the intensity of light indicates the abundance of infected cells. If virus entry is blocked, no luciferase will be produced. It is as if the TZM-bl cells contain dark light-bulbs with on-off switches and only a virion can flick the switch on. Plasma from monkeys was serially diluted and incubated with virus in a 96 well plate containing TZM-bl cells. Let W_i be the intensity of the i th dilution and W_0 the average luciferase activity in a control well with no antibody and that glows relatively brightly. The percent neutralization is $P_i = (W_0 - W_i)/W_0$. A four parameter logistic model is then fit to these P_i s and log dilutions (Schofield 2003). The dilution for which the predicted $\hat{P}$ equals .50 is a measure of neutralizing antibody effect, say X . Unfortunately sufficient stored plasma for the TZM-bl assay was only available on a validation set of 11 of the 21 monkeys. For the validation set, the TZM-bl assay was separately run using two different chimeras or combinations of HIV and SIV viruses called the SHIV-293T and the pseudo SHIV-293T viruses. Thus there are actually two readouts from the TZM-bl assay corresponding to the two viruses, say X_1 and X_2 . In this paper we are interested in the effect of $X_1 = X$ on infection status.

3 Models & Likelihoods

To begin, suppose we have n individuals with a binary outcome Y , say infection, an associated covariate X and no missing data. In bioassay, it is common to assume that the relationship between the two is given by a probit regression model

$$P(Y=1|X) = \Phi(\alpha_0 + \alpha_1 X), \quad (1)$$

where $\Phi()$ is the standard normal distribution function, see e.g. Finney (1971). A standard approach is to calculate the maximum likelihood estimates of (α_0, α_1) . Of particular interest in assays is the ID_P or the value of X that results in $P\%$ of the animals being infected. For probit regression the ID_P is given by $\{\Phi^{-1}(P) - \alpha_0\}/\alpha_1$. Feiler's method or profile likelihood are typically used to construct confidence intervals for the ID_P (Morgan 1992).

We are interested in the problem where X is missing on some individuals, W is available on all, and there is a validation set of size n_v containing X and W . With a specified model for X/W one can derive a likelihood for all the data;

$$L(\alpha_0, \alpha_1) = \prod_{i=1}^{n_v} P(Y=1|X_i)^{y_i} \{1 - P(Y=1|X_i)\}^{1-y_i} \prod_{i=n_v+1}^n E_{X|W} \{P(Y=1|X, W_i)\}^{y_i} [1 - E_{X|W} \{P(Y=1|X, W_i)\}]^{1-y_i} \quad (2)$$

where the conditional expectation in $E_{X|W} \{P(Y=1|X, W_i)\}$ is over the missing X . In general we need to assume an ignorable missing data mechanism e.g. $P(X=?|X, W, Y) = h(X)$ for valid inference. In our application, the plasma was collected prior to challenge and only monkeys who had insufficient stored plasma had missing X values, thus ignorable missingness seems reasonable.

To proceed we need to specify a model for X/W , say a simple linear regression:

$$X = \beta_0 + \beta_1 W + U \quad (3)$$

where U is normal $(0, \sigma^2)$. In our setting there is an additional twist to the data structure as we have another outcome variable that is not of direct interest, X_2 , along with the $X = X_1$ of interest. Joint modeling of these two variables leads to a multivariate regression model,

$$X_1 = \beta_{01} + \beta_{11} W + U_1 \quad (4)$$

$$X_2 = \beta_{02} + \beta_{12} W + U_2 \quad (5)$$

where U_1, U_2 are bivariate normal. Such a model with correlated errors is known as a seemingly unrelated regression (SUR) model in the econometrics literature, Zellner (1962). In general SUR estimates can be more efficient than single equation estimates. However, when the covariate W is common to both equations the regression estimates of X_1 on W are the same for SUR as for the single equation (3). Thus to gain efficiency in this setting, additional structure needs to be imposed such as $\beta_{01} = \beta_{02}$ or $\beta_{11} = \beta_{12}$, or additional covariates say $V_1 \neq V_2$ can be brought into (4) & (5).

Under either (3) or (4) & (5), the conditional distribution of X/W is Gaussian with mean $\beta_0 + \beta_1 W$ and variance $(\alpha_1 \sigma)^2$. Under the assumption that $P(Y=1|X, W) = \Phi(\alpha_0 + \alpha_1 X)$ we obtain the following for use in (2):

$$E_{X|W} \{P(Y=1|X, W)\} = E_{X|W} \{\Phi(\alpha_0 + \alpha_1 X)\} = \Phi \left\{ \frac{\alpha_0 + \alpha_1 (\beta_0 + \beta_1 W)}{\sqrt{1 + (\alpha_1 \sigma)^2}} \right\}. \quad (6)$$

This follows from the result that $E_U \{\Phi(a+U)\} = \Phi\{(a+\mu)/\sqrt{1+\omega^2}\}$ for U normal (μ, ω^2) .

If $\beta_0, \beta_1, \sigma^2$ were known, we could substitute them in (6) and maximize (2) directly. Since they have to be estimated, different options are possible. One possibility is the joint

likelihood for Y, X, W . For the model defined by (4) & (5), and the assumption that $f(Y, X/W) = f(Y/X)f(X/W)$ this amounts to forming

$$L_J = \prod_{i=1}^{n_v} P(Y=1|X_i)^{y_i} \{1 - P(Y=1|X_i)\}^{1-y_i} \\ \prod_{i=n_v+1}^n E_{X|W} \{P(Y=1|X, W_i)^{y_i} [1 - E_{X|W} \{P(Y=1|X, W_i)\}]^{1-y_i} \\ \varphi(X_{i1}, X_{i2}|\beta_{01}+\beta_1 W_i, \beta_{02}+\beta_1 W_i, \sigma^2, \rho)\}$$

where $\varphi(X_{i1}, X_{i2}|\mu_1, \mu_2, \sigma^2, \rho)$ is the bivariate normal density function with means μ_1, μ_2 , variance σ^2 and correlation ρ . A REML-type log-likelihood can be defined by adding $-1/2 \log\{\det(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})\}$ to $\log(L_J)$, where $\mathbf{Z}$ is the $2n \times 3$ matrix with rows $(1, 0, W_1), (1, 1, W_1), (1, 0, W_2)$ and $\mathbf{V} = \text{cov}(X_{11}, X_{21}, X_{22}, \dots)$ (see Harville 1977). However, this approach involves a fairly complex likelihood and may be difficult to handle with small n . As an alternative, regression calibration could be considered. Here the missing X is replaced with $\hat{\beta}_0 + \hat{\beta}_1 W$ and the usual probit likelihood used. This approach is very simple but can result in biased estimates if $(\alpha_1 \sigma)^2$ is large.

A kind of middle ground is given by pseudo-likelihood where estimates of the parameters $\beta_0, \beta_1, \sigma^2$ are substituted in (2) (Gong & Samaniego 1981). However, this direct substitution ignores the fact that there is uncertainty in $\hat{\beta}_0$ and $\hat{\beta}_1$ and thus more uncertainty about the missing X than $\text{var}(X/W) = \sigma^2$. With the model given by (4) & (5) one can show that the variance of a future predicted observation, $\hat{\beta}_0 + \hat{\beta}_1 W + U$, is given by $\tau^2(W) = \sigma^2 + \text{var}(\hat{\beta}_0) + W^2 \text{var}(\hat{\beta}_1) + 2W \text{cov}(\hat{\beta}_0, \hat{\beta}_1)$. This is larger than the direct substitution approach which uses σ^2 . Using this larger, W dependent variance, we obtain the following expression for the probability of infection for an individual with only W available in (2)

$$E_{X|W} \{P(Y=1|X, W)\} = E_{X|W} \{\Phi(\alpha_0 + \alpha_1 X)\} = \Phi \left[\frac{\alpha_0 + \alpha_1 \hat{X}(W)}{\sqrt{1 + \{\alpha_1 \hat{\tau}(W)\}^2}} \right], \quad (7)$$

where $\hat{\tau}(W)$ is an estimator of $\tau(W)$, and $\hat{X}(W) = \hat{\beta}_0 + \hat{\beta}_1 W$. We then maximize the following pseudo-log-likelihood to get the mles of α_0, α_1

$$\ell_1(\alpha_0, \alpha_1) = \log \left\{ \prod_{i=1}^{n_v} \Phi(\alpha_0 + \alpha_1 X_i)^{y_i} \{1 - \Phi(\alpha_0 + \alpha_1 X_i)\}^{1-y_i} \times \prod_{i=n_v+1}^n \Phi \left[\frac{\alpha_0 + \alpha_1 \hat{X}(W_i)}{\sqrt{1 + \{\alpha_1 \hat{\tau}(W_i)\}^2}} \right]^{y_i} \left(1 - \Phi \left[\frac{\alpha_0 + \alpha_1 \hat{X}(W_i)}{\sqrt{1 + \{\alpha_1 \hat{\tau}(W_i)\}^2}} \right] \right)^{1-y_i} \right\}. \quad (8)$$

To test whether α_1 is zero, we use a likelihood ratio test. Due to the small sample size, use the parametric bootstrap to identify a critical value (Efron & Tibshirani 1993). To proceed we first obtain the mle of α_0 , based on the pseudo-likelihood (8), with α_1 fixed at 0, say $\hat{\alpha}_0^0$. We then generate new Y data, say Y^* , where $Y_i^* \stackrel{D}{=} \text{Bernoulli}\{\Phi(\hat{\alpha}_0^0)\}$, for $i = 1, \dots, n$. We plug in Y^* for y in (8) and maximize to obtain $\hat{\alpha}_0^*, \hat{\alpha}_1^*$. We also estimate α_0 using Y^* under the null model where $\alpha_1 = 0$; call this estimate $\hat{\alpha}_0^{*0}$. This allows us to form the pseudo-likelihood ratio statistic for the simulated data: $h^* = 2\{\ell_1(\hat{\alpha}_0^*, \hat{\alpha}_1^*) - \ell_1(\hat{\alpha}_0^{*0}, 0)\}$. We generate B values of h^* using this method. An estimated p-value for testing $H_0: \alpha_0 = 0$ is given by $\#(h^* \geq h)/B$

where h is the pseudo-likelihood ratio statistic for the original data

$$h=2\{\ell_1(\widehat{\alpha}_0, \widehat{\alpha}_1) - \ell_1(\widehat{\alpha}_0^0, 0)\}.$$

4 Confidence Intervals for ID_{PS}

In bioassay, the main parameter of interest is $\theta = \text{ID}_P$ or the “dose” of antibodies that results in $p\%$ of the animals being infected. For vaccine development an important target is the ID₁ as an ideal vaccine should protect nearly everyone. Below we illustrate how to construct a confidence interval when $p = .5$, where $\theta = -\alpha_0/\alpha_1$; other ID_{PS} are similar. Methods discussed in the literature include Fieller’s method (Fieller 1944) and profile likelihood see e.g. Williams (1986) or Beyne & Moineddin (2005). Both of the latter papers suggest profile likelihood may be better than Fieller’s method. To proceed, note that $\alpha_0 = -\theta\alpha_1$ and reparameterize the pseudo-log-likelihood as

$$\ell_2(\alpha_1, \theta) = \log\left\{\prod_{i=1}^{n_v} \Phi\{\alpha_1(X_i - \theta)\}^{y_i} [1 - \Phi\{\alpha_1(X_i - \theta)\}]^{1-y_i} \times \prod_{i=n_v+1}^n \Phi\left[\frac{\alpha_1\{\widehat{X}(W_i) - \theta\}}{\sqrt{1+\{\alpha_1\widehat{\tau}(W_i)\}^2}}\right]^{y_i} \left[1 - \Phi\left[\frac{\alpha_1\{\widehat{X}(W_i) - \theta\}}{\sqrt{1+\{\alpha_1\widehat{\tau}(W_i)\}^2}}\right]\right]^{1-y_i}\right\}. \quad (9)$$

We use the notation $\ell_2()$ to emphasize a parameterization different from that in $\ell_1()$.

To obtain a confidence interval for θ we consider one-sided tests of $H_0: \theta \leq \theta_0$ and $H_0: \theta \geq \theta_0$ using the likelihood ratio statistic:

$$h(\theta_0) = 2[\ell_2(\widehat{\alpha}_1, \widehat{\theta}) - \ell_2\{\widehat{\alpha}_1(\theta_0), \theta_0\}],$$

where $\widehat{\alpha}_1(\theta_0)$ is the mle of α_1 from $\ell_2()$ when θ is fixed at θ_0 . We then find $\theta_0 = \theta_L < \widehat{\theta}$ and $\theta_0 = \theta_R > \widehat{\theta}$ that satisfy $h(\theta) = C$ where C is the 95th percentile of the h s under H_0 . The solution (θ_L, θ_R) provides an approximate 95% confidence interval for the ID₅₀.

For large sample sizes, C is approximately 3.84 as h has approximately a chi-squared distribution with 1 degree of freedom on the null. To obtain an accurate C for small samples, we use the parametric bootstrap. We first obtain the mles of θ, α_1 based on the augmented probit pseudo-likelihood (9), $\widehat{\theta}, \widehat{\alpha}_1$. We then generate new Y data, say Y^* using the fixed $X_1, \dots, X_{n_v}, W_{n_v+1}, \dots, W_n$ and the mle: $Y_i^* \stackrel{D}{=} \text{Bernoulli}[\Phi\{\widehat{\alpha}_1(X_i - \widehat{\theta})\}]$ for $i = 1, \dots, n_v$ and

$Y_i^* \stackrel{D}{=} \text{Bernoulli}(\Phi[\{\widehat{\alpha}_1\{\widehat{X}(W_i) - \widehat{\theta}\} / \sqrt{1+\{\widehat{\alpha}_1\widehat{\tau}(W_i)\}^2}])$ for $i = n_v + 1, \dots, n$. For each generated data set, we plug in Y^* for y in (9) and maximize to obtain $\widehat{\alpha}_1^*, \theta^*$. We additionally plug Y^* into (9) but with θ fixed at the mle of the original data $\widehat{\theta}$ and then obtain the mle of α_1 under this constraint, say $\widehat{\alpha}_1^*(\widehat{\theta})$. We then form the log-likelihood ratio statistic based on

$Y^*: h^* = 2[\ell_2(\widehat{\alpha}_1^*, \widehat{\theta}^*) - \ell_2\{\widehat{\alpha}_1^*(\widehat{\theta}), \widehat{\theta}\}]$. We obtain C as the 95th percentile of the distribution of the h^* s.

5 Application

We applied our methods to the dataset of twenty-one monkeys. We begin by developing the prediction model for $X = X_1$ given W . In our setting, immunological arguments regarding the relationship between X_1 and X_2 can be used to derive parameter constraints on the SUR model (4) & (5). Specifically, the regressions of X_1 on W and X_2 on W should have the same slopes but different intercepts.

$$X_{1i} = \beta_{01} + \beta_1 W_i + U_{1i} \quad (10)$$

$$X_{2i} = \beta_{02} + \beta_1 W_i + U_{2i}, \quad (11)$$

where X_{1i} , X_{2i} are the readouts of the TZM-bl assays for the SHIV-293T and the pseudo SHIV-293T viruses, respectively, for monkey $i = 1, \dots, 11$. We assume that the U_1 , U_2 are bivariate normal with mean 0 and common variance σ^2 .

To estimate the parameters we use the method of generalized estimating equations with an exchangeable working correlation matrix (Liang & Zeger 1986) and obtain $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2 = 1.395, .936, .034$. The model assumptions appeared reasonable based on graphical analysis (not shown). Based on these estimated parameters we maximize the pseudo-likelihood (8) to estimate α_0, α_1 . The parameter estimates are $\hat{\alpha}_0 = 4.32$ and $\hat{\alpha}_1 = -1.98$. The likelihood ratio test statistic for testing $\alpha_1 = 0$ is $h = 7.41$. We used the parametric bootstrap with 1,000,000 bootstrap samples to test $\alpha_1 = 0$ and estimated a p-value of .0112.

To check the assumption that $P(Y = 1/X, W) = P(Y = 1/X)$, we fit the simple probit model $P(Y = 1/X, W) = \Phi(\alpha_0 + \alpha_1 X + \alpha_2 W)$ using regression calibration with missing X replaced with $\hat{X}(W)$. The estimates $\hat{\alpha}_1, \hat{\alpha}_2$ had an estimated correlation of $-.93$ and neither were significant ($p > .10$). Results were similar using the 11 monkeys with nonmissing X values. Such behavior is not surprising given the strong relationship between X and W and checking the assumption $P(Y = 1/X, W) = P(Y = 1/X)$ is inherently difficult.

Figure 2 provides the data and estimated probit curve (solid curve) for the TZM-bl assay for the SHIV-293T virus. The 11 observed Y, X pairs are solid squares while the predicted $\hat{X}(W_i)$ are open circles with 80% prediction intervals $\hat{X}(W_i) \pm 1.28 \sqrt{\hat{\tau}^2(W_i)}$ given by lines. Note the 1.28 constant is based on the assumption that U_1 is normal. The dashed curve in Figure 2 provides the “attenuated” probability of infection as a function of $\hat{X}$ at the average value of

$$\bar{\tau}^2; \bar{\tau}^2 = \sum_{i=12}^{21} \hat{\tau}^2(W_i) / 10 \text{ or}$$

$$\Phi \left\{ \frac{\alpha_1 (\hat{X} - \theta)}{\sqrt{1 + (\alpha_1^2 \bar{\tau}^2)}} \right\}.$$

The denominator is about 1.002 and thus the attenuated curve is quite close to the probit curve, suggesting that regression calibration would yield nearly identical results. For comparison, we calculated both the regression calibration and joint REML-likelihood estimates of $\alpha_0, \alpha_1, \text{ID}_{50}$; respectively (4.40, $-2.02, -2.18$) and (4.01, $-1.84, -2.18$). This compares to (4.32, $-1.98, -2.18$) for pseudo-likelihood. Though the estimates slightly differ for the slope and intercept, the ID_{50} is the same for all three methods.

Table 2 presents the pseudo-likelihood estimates of two ID_{Ps} on the original and log scale, along with confidence intervals. The 95th percentile of the bootstrap distribution of h^* was 4.51, based on 1,000,000 samples. We see that the estimated $10^{\text{ID}_{50}}$ is about 150 while the estimated 10^{ID_1} is about 2100 with lower limits of about 29 and 470, respectively. Figure 2 provides the profile log-likelihoods for the ID_1 and ID_{50} along with the confidence intervals. The log-likelihood for ID_{50} is more symmetric while that of ID_1 is skewed.

To interpret the 10^{ID_p} s, we need to know what kind of antibody production is possible either from natural infection or in response to a vaccine. Li et al (2005) reported geometric mean nAB titers from 1:97 to 1:361 based on pools of sera from multiple individuals who were infected with HIV. If a vaccine could mimic natural chronic infection and induce such levels of nABs, perhaps 50% of individuals could be protected. In a more direct experiment, Blish et al (2008) looked at individuals who were already infected with HIV and thus had nABs, but got infected again with a new HIV virus. They reported nAB titers of approximately 1:100, so there is evidence that 1:100 is too low of a titer.

6 Simulations

In this section we evaluate the performance of regression calibration, pseudo-likelihood, and joint REML-likelihood estimates of the ID_p in different settings. We also address the accuracy of the profile confidence intervals for the ID_{50} for our setting with $n = 21$ and smaller and larger sample sizes.

For the first set of simulations, we considered sample sizes of $n = 21$ and $n = 84$ using the original W s (or 4 replicates of each). Bioassay experiments typically have a good distribution of doses ensuring a variety of event probabilities ranging from low to high. Using these W s thus recapitulates an important design feature. Based on these W s we then generate X_1, X_2 under model (4) (5), and then generated outcomes under the probit model (1). We used the joint REML-likelihood estimates of the parameters when generating the data. To create a validation set, we randomly “covered up” a proportion p of the X_1 s. For $n = 21$ we used the original 21 W s and 11 original X_1 s thus the same 10 were always “covered up”.

We varied a number of parameters. We set $(\alpha_0, \alpha_1) = (\hat{\alpha}_0, \hat{\alpha}_1)f$ where $f = 1/2, 1, 2$ and $(\hat{\alpha}_0, \hat{\alpha}_1)$ are the joint REML-likelihood estimates. We also varied the proportion of data in the validation set from $p = .1, .5, .9$, and set $STD(Y) = s\hat{\sigma}$ for $s = 1/4, 1, 4$ i.e. multiples of the joint REML estimate. Table 2 presents the results. Throughout, the first row in a contiguous block of rows provides a reference case, rows below perturb a single parameter e.g. rows 2 and 3 change f from .5 to 2. For the $n = 21$ rows we used a trimmed mean and variance where the largest and smallest .5% of the simulated estimates were discarded. Thus was done to ameliorate the obfuscating effect of outliers. For $n = 21$ we see that regression calibration and pseudo-likelihood have nearly identical performance while the joint REML-likelihood approach can have a smaller or larger variance. All methods have similar averages. Also note that rows 2 and 5 have some very large variances, due to rare occasions where the ID_1 was difficult to estimate.

For the middle block of rows with $n = 84$ a similar story emerges. Regression calibration and pseudo-likelihood have similar performance though regression calibration is slightly biased with larger variance when $s = 4$, but regression calibration has smaller variance with $p = .1$. The performance of the estimates is strongly influenced by f . The joint likelihood estimate is generally worse than the other two.

The bottom block of rows corresponds to a case where data are generated under a quadratic relationship between X, W but an incorrect linear regression is used. The quadratic function was chosen to be a nearly monotone J-shaped curve, broadly similar to the estimated linear relationship. Under this type of model misspecification the pseudo and joint likelihood approaches have less bias than regression calibration, though regression calibration generally has a smaller variance.

To explore the sensitivity of our conclusions to the assumption that $P(Y = 1/X, W)$ was free of W , we conducted a simulation similar to the middle seven rows of Table 2 but where data

were generated under the model $P(Y=1|X, W)=\Phi(\alpha_0^*+\alpha_1^*X+\alpha_2^*W)$ with the α s set at their estimated values from our dataset (results not shown). We regarded the average estimates from the probit likelihood model $P(Y=1|X, W)=\Phi(\alpha_0+\alpha_1X)$ as the “truth” as this “approximate” model is what is of scientific interest and what we would estimate with no missing data. We calculated the median percent bias for the pseudo-likelihood estimates. For $P=.50$ the absolute percent bias was always less than 4% while for $P=.01$ the percent bias was always less than 10% except for $p=.1$ where the bias was 18%. The bias from regression calibration and joint likelihood estimates was very similar, except for $P=.01$, $s=4$ where the pseudo-likelihood estimates had bias 8% while regression calibration and the joint likelihood had biases of 32% and -56% respectively. For these data, bias does not appear to be a substantial problem for pseudo-likelihood under this kind of model misspecification except when the validation set is a small fraction and interest focuses on the $ID_{.01}$.

In other simulations we evaluated the ID_{50} (results not shown). The conclusions were similar to Table 2, except for regression calibration, which was less biased than for the ID_1 . We repeated the simulations using a single regression equation (3) instead of (4) & (5). For pseudo-likelihood and regression calibration, the results were quite similar for all scenarios. Interestingly, the joint likelihood approach generally did better with one equation, possibly related to difficulties in estimating the nuisance parameter ρ . Overall, the extra complication of the two equation approach is probably not worth the benefit. We also evaluated maximum likelihood which was very similar to REML.

All in all, regression calibration and pseudo-likelihood seem to perform more reliably than the joint-REML likelihood estimates, with regression calibration perhaps slightly better. Interestingly, while regression calibration is nearly unbiased for the ID_{50} and slightly biased for the ID_1 , we did observe substantial bias for α_0 , α_1 . Apparently the bias tends to cancel out when forming a ratio of biased estimates; recall that $ID_{50} = -\alpha_0/\alpha_1$.

Our second set of simulations address coverage of our profile confidence intervals. We first generate a “primitive” data set Y using $Y_i \stackrel{D}{=} \text{Bernoulli}[\Phi\{\hat{\alpha}_1(X_i - \hat{\theta})\}]$ for $i = 1, \dots, n_v$ and

$Y_i \stackrel{D}{=} \text{Bernoulli}[\Phi\{\alpha_1(X(W_i) - \theta)/\sqrt{1+\{\alpha_1\tau(W_i)\}^2}\}]$ for $i = n_v + 1, \dots, n$. The constants α_1 , θ , $\tau(W)$, $X(W)$ are set at their estimated values from the previous section.

We calculate the mles $\hat{\alpha}_1$, $\hat{\theta}$ based on this primitive data and then construct a 95% confidence interval using profile likelihood and the parametric bootstrap with 10,000 replications as described at the end of section 4. We repeat this entire endeavor 1,000 times. We also construct confidence intervals using an asymptotic critical value, $C=3.84$. We repeated this exercise by reducing and increasing the sample size. For reduction we ordered the original, covariate data $X_1, \dots, X_{11}$ deleted the even ones and repeated this for $\hat{X}(W_{12}), \dots, \hat{X}(W_{21})$. We chose such a small sample size ($n = 11$) to see how things might break down. For increasing, we doubled and then quadrupled the number of outcomes for each of the 21 covariates. Coverage intervals are reported in Table 3 along with the percentage of times that the confidence interval contained an infinite limit. With $n = 21$, the bootstrap confidence intervals have slightly greater coverage than the chi-square confidence intervals and are closer to the nominal level, though not infrequently the CI limit is infinite. There is not a substantial difference between the bootstrap and asymptotic approaches. With $n = 11$ the confidence intervals have poor coverage and a substantial percentage of CIs have infinite limits. For $n = 42$ and 84, both approaches are similar and accurate.

The occurrence of infinite confidence limits deserves comment. This nearly always happens when we fail to reject a test that the slope is zero. This makes sense as the ID_P is undefined

when the slope, α_1 , is 0. If we have little evidence of a nonzero slope, we should have enormous uncertainty about the ID_P and an infinite CI.

7 Discussion

This paper develops an augmented probit model for use in bioassay for the small sample setting where a validation set is available. Our development was influenced by the small sample size, which is typical for bioassay experiments involving expensive non-human primates. We chose a parametric linear approach for the prediction model with Gaussian errors and incorporated data from a strongly related assay. A pseudo-likelihood approach was used for estimation and the parametric bootstrap was used for inference on both the probit slope and the ID_P . Simulations reflecting our experiment suggest that the asymptotic approach to construction of profile likelihood confidence intervals is reasonably accurate for n as small as about 20. In other simulations we found that regression calibration using a single prediction equation worked very well and was quite similar to the pseudo-likelihood approach. Under model misspecification pseudo-likelihood and regression calibration usually had similar bias though occasionally regression calibration had more bias. A joint likelihood approach had more variable performance, probably related to the difficulty in estimating a nuisance parameter. While regression calibration is very simple and useful, pseudo-likelihood usually gave similar performance and was more appealing to our collaborators as it more clearly propagated the uncertainty associated with missing data.

Acknowledgments

We thank Malcolm Martin, Ron Willey, Yoshiaki Nishimura, Pamela Shaw, and Chiung-Yu Huang for helpful discussions on the assay, experiments, and methods.

References

- Beyene, J.; Moineddin, R. Methods for Confidence Interval Estimation of a Ratio Parameter with Application to Location Quotients; BMC Medical Research Methodology. 2005. p. 1-7. <http://www.biomedcentral.com/1471-2288/5/32>
- Blish CA, Dogan OC, Derby NR, Nguyen MA, Chohan B, Richardson BA, Overbaugh J. Human immunodeficiency virus type 1 superinfection occurs despite relatively robust neutralizing antibody responses. Journal of Virology. 2008; 82:12094–12103. [PubMed: 18842728]
- Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. Measurement Error in Nonlinear Models. Chapman & Hall/CRC; Boca Raton FL: 2006.
- Carroll RJ, Spiegelman CH, Lan KKG, Bailey KT, Abbott RD. On errors-in-variables for binary regression models. 1984; 71:19–25.
- Carroll RJ, Stefanski LA. Approximate quasi-likelihood estimation in models with surrogate predictors. Journal of the American Statistical Association. 1990; 85:652–663.
- Carroll RJ, Wand MP. Semiparametric Estimation in Logistic Measurement Error Models. Journal of the Royal Statistical Society Series B (Methodological). 1991; 53(3):573–585.
- Efron, B.; Tibshirani, RJ. An Introduction to the Bootstrap. Chapman & Hall/CRC; Boca Raton FL: 1993.
- Follmann DA, Hunsberger SA, Albert PS. Repeated Probit Regression When Covariates Are Measured With Error. Biometrics. 1999; 55:403–409. [PubMed: 11318193]
- Finney, DJ. Probit Analysis. Cambridge University Press; Cambridge England: 1971.
- Fieller EC. A Fundamental Formula in the Statistics of Biological Assay, and Some Applications. Quarterly Journal of Pharmacy & Pharmacology. 1944; 17:117–123.
- Gong G, Samaniego FJ. Pseudo Maximum Likelihood Estimation: Theory and Applications. Annals of Statistics. 1981; 9:861–869.
- Harville DA. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. Journal of the American Statistical Association. 1977; 72:320–338.

- Laird NM, Ware JH. Random-Effects Models for Longitudinal Data. *Biometrics*. 1982; 38:963–974. [PubMed: 7168798]
- Liang KY, Zeger S. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrics*. 1986; 38:963–974.
- Li M, Gao F, Mascola JR, Stamatatos L, Polonis VR, Koutsoukos M, Voss G, Goepfert P, Gilbert P, Greene KM, Bilska M, Kothe DL, Salazar-Gonzalez JF, Wei X, Decker JM, Hahn BH, Montefiori DC. Human immunodeficiency virus type 1 env clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. *Journal of Virology*. 2005; 79:10108–10125. [PubMed: 16051804]
- Morgan, BJT. Analysis of Quantal Response Data. Chapman & Hall/CRC; Boca Raton FL: 1992.
- Neter, J.; Wasserman, W. Applied Linear Statistical Models. Richard Irwin; Homewood IL: 1974.
- Nishimura Y, Igarashi T, Haigwood N, Sadjadpour R, Plishka RJ, Buckler-White A, Shibata R, Martin MA. Determination of a statistically valid neutralization titer in plasma that confers protection against simian-human immunodeficiency virus challenge following passive transfer of high-titered neutralizing antibodies. *Journal of Virology*. 2002; 76:2123–2130. [PubMed: 11836389]
- Schafer DW. Likelihood Analysis for Probit Regression with Measurement Errors. *Biometrika*. 1993; 80(4):899–904.
- Schofield, T. Assay Development. In: Chow, Shein-Chung, editor. *Encyclopedia of Biopharmaceutical Statistics*. 2. 2003. p. 55-62.
- Tosteson TD, Stefanski LA, Schafer DW. A Measurement-error model for binary and ordinal regression. *Statistics in Medicine*. 1989; 8(9):1139–1147. [PubMed: 2799134]
- Wiley R, Nason M, Nishimura Y, Follmann DA, Martin MA. A Reassessment of In Vivo Neutralizing Antibody Titers Required to Protect Macaques From a Simian/Human Immunodeficiency Virus Challenge. *AIDS Research and Human Retroviruses*. 2009 submitted.
- Williams DA. Interval Estimation of the Median Lethal Dose. *Biometrics*. 1986; 42:641–646. [PubMed: 3567296]
- Zellner A. An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *Journal of the American Statistical Association*. 1962; 57:348–368.

Figure 1.

Data and probit curve for TZM-bl assay with SHIV-293T virus. The solid curve is the estimated probability of protection as a function of the amount of neutralizing antibody as measured by $\log_{10}(\text{EC}_{50})$ of the TZM-bl assay. Solid squares denote observed (Y, X) data for 11 monkeys, jittered for clarity. Solid lines denote 80% prediction intervals for the $(Y, ?)$ for 10 monkeys based on the the observed MT4 assay. The dotted curve is the attenuation of the probit curve resulting from a prediction of X based on a Gaussian distribution with mean $\hat{\beta}_0 + \hat{\beta}_1 W$ and variance $\overline{\hat{\tau}^2(W)}$, the average over the 10 monkeys.

Figure 2.
Profile log-pseudo-likelihoods for the ID_{50} and ID_1 .

Table 1

Pseudo-likelihood estimates for the ID_{50} and ID_1 . Confidence intervals are obtained by profiling the pseudo-likelihood ratio test statistic.

Parameter	Estimate	95% Profile CI	
		Lower	Upper
ID_{50}	2.17	1.46	2.62
ID_1	3.32	2.67	8.04
$10^{ID_{50}}$	148.57	29.01	412.75
10^{ID_1}	2088.80	466.78	1.10×10^8

Table 2

Performance of different estimates of the ID_1 . Each row summarizes 1,000 simulated data sets. Each scenario is a perturbation of the estimated model of section 5. Perturbations include the sample size (n), a multiplicative factor for α_0 , α_1 (f), the proportion in the validation set (p), and the multiplicative factor for the variance (s). The first five rows use a trimmed sample mean and variance where the most extreme 1% of estimates are excluded. The bottom five rows are when the regression between X and W is truly quadratic, but misspecified as linear.

n	f	p	s	ID_1	Sample Mean			Sample Variance		
					calib	pseudo	joint	calib	pseudo	joint
21	1.	-	1	3.34	3.34	3.29	3.32	0.410	0.453	0.610
21	.5	-	1	4.52	5.04	5.01	5.03	63.400	63.014	60.768
21	2.	-	1	2.76	2.74	2.65	2.66	0.104	0.127	0.148
21	1.	-	.25	3.34	3.31	3.31	3.08	0.422	0.424	11.662
21	1.	-	4	3.34	3.57	3.20	3.10	0.745	0.774	0.653
84	1.	.5	1	3.34	3.37	3.34	3.36	0.094	0.097	0.238
84	.5	.5	1	4.52	4.77	4.75	4.63	2.256	2.266	1.766
84	2.	.5	1	2.76	2.80	2.73	2.75	0.022	0.024	0.062
84	1.	.1	1	3.34	3.41	3.33	3.27	0.130	0.183	0.167
84	1.	.9	1	3.34	3.36	3.35	3.43	0.086	0.084	1.268
84	1.	.5	.25	3.34	3.36	3.36	2.95	0.086	0.085	377.371
84	1.	.5	4	3.34	3.62	3.29	3.27	0.146	0.122	0.117
Quadratic Regression										
84	1.	.5	1	3.34	3.98	3.78	3.73	0.216	0.255	0.277
84	.5	.5	1	4.52	5.03	4.93	4.84	0.790	0.839	0.815
84	2.	.5	1	2.76	3.36	3.00	2.98	0.099	0.126	0.142
84	1.	.5	.25	3.34	4.00	3.85	3.78	0.253	0.287	0.299
84	1.	.5	4	3.34	4.01	3.40	3.36	0.194	0.181	0.165

Table 3

Coverage probabilities for different profile likelihood confidence intervals for the ID_{50} . Both bootstrap and asymptotic critical values are evaluated. Simulation scenarios are perturbations of the motivating example.

n	Intended Coverage	Bootstrap		Asymptotic	
		Coverage	% Inf limit	Coverage	% Inf limit
11	80	.678	18.8	.707	11.6
	95	.871	50.7	.893	36.4
21	80	.809	4.4	.774	2.9
	95	.941	19.4	.929	16.8
42	80	.809	.2	.792	.2
	95	.954	1.8	.944	1.6
84	80	.810	0	.800	0
	95	.953	0	.952	0