The sign of the logistic regression coefficient

A. B. Owen
Stanford University

P. A. Roediger
UTRS, Inc.

January 2014

Abstract

Let Y be a binary random variable and X a scalar. Let $\hat{\beta}$ be the maximum likelihood estimate of the slope in a logistic regression of Y on X with intercept. Further let $\bar{x}_0$ and $\bar{x}_1$ be the average of sample x values for cases with $y = 0$ and $y = 1$, respectively. Then under a condition that rules out separable predictors, we show that $\mathrm{sign}(\hat{\beta}) = \mathrm{sign}(\bar{x}_1 - \bar{x}_0)$. More generally, if x_i are vector valued then we show that $\hat{\beta} = 0$ if and only if $\bar{x}_1 = \bar{x}_0$. This holds for logistic regression and also for more general binary regressions with inverse link functions satisfying a log-concavity condition. Finally, when $\bar{x}_1 \neq \bar{x}_0$ then the angle between $\hat{\beta}$ and $\bar{x}_1 - \bar{x}_0$ is less than ninety degrees in binary regressions satisfying the log-concavity condition and the separation condition, when the design matrix has full rank.

1 Introduction

This short note is to introduce and prove an interesting fact about logistic regression with a scalar predictor x and an intercept. The fact is that the sign of the slope coefficient's maximum likelihood estimate (MLE), be it positive, negative or zero, matches the sign of the difference in sample means of the predictors.

This finding about signs was put forth as a conjecture in a discussion (Ray et al., 2013) of the paper by Wu and Tian (2013) on sensitivity experiments. These sensitivity experiments are sequential experimental designs to estimate quantities such as a dose with 50% lethality (LD50) in toxicology. Other applications, such as safety and reliability of explosives, require estimates of x with $\Pr(Y = 1 \mid X = x)$ much closer to 0 or 1. Every once in a while a chance pattern will lead to a negative coefficient for a predictor x when it is known scientifically that $\Pr(Y = 1 \mid X = x)$ can only increase with x. Ray et al. (2013) advocate continued testing in this circumstance and remark that the mean difference is a simple way to detect it. Similarly, the conjecture would allow one to constrain the slope's MLE to be positive, by the simple device of adding an artificial data point $(x_0, 0)$ with a very small x_0, and/or $(x_{n+1}, 1)$ with a very large x_{n+1}. Their conjecture can be proved using elementary methods. The finding is interesting and does not seem to be widely known.

It is intuitively clear that a positive coefficient should be more likely when $\bar{x}_1$, the average x value for $y = 1$, is larger than $\bar{x}_0$, the average x value for $y = 0$. But the pattern is absolute; there can be no exceptions stemming from different variances, skewnesses or outliers among the x values.

Logistic regression with fixed x_i and random y_i is an exponential family model. Given $x_1, \dots, x_n$, the sufficient statistic is the pair of values $n_1 = \sum_{i=1}^n y_i$ and $\sum_i x_i y_i$. (McCullagh and Nelder, 1989, Chapter 2.2.4). As a result, the logistic regression coefficients are determined by $x_1, \dots, x_n$ along with n_1 and $\bar{x}_1$. Given x_i, the sufficient statistics can be used to compute $(n_0, n_1, \bar{x}_0, \bar{x}_1)$ and vice versa, where $n_0 = n - n_1$. But this latter quadruple does not determine the MLE on its own. The sufficiency is only a conditional one and does not quite explain the sign result. Moreover, probit regression models do not have such sufficient statistics, yet we find that their slope also has a sign determined by that of $\bar{x}_1 - \bar{x}_0$, without regard to other features of the x_i sample. The sign result stems from log concavity of the logistic and Gaussian density functions and it extends to many other binary regressions.

Section 2 introduces the notation we need. Section 3 has our elementary proof and Section 4 considers two generalizations for vector-valued predictors $x_i \in \mathbb{R}^d$. Both generalizations make some mild assumptions about the configuration of x_i values. In binary regressions, the inverse of the link function is usually a cumulative distribution function (CDF). When that CDF corresponds to a log concave probability density function, we find that the coefficient of x is zero if and only if the mean x-values for $y = 0$ and $y = 1$ coincide. When the means do not coincide, the regression coefficient makes less than a 90 degree angle with the difference in x means.

We conclude this section by mentioning some similar results. When $X \in \mathbb{R}^d$ are independently $\mathcal{N}(\mu_y, \Sigma)$ distributed conditionally on $Y = y$, for a covariance matrix Σ of full rank, then the population version of the logistic regression coefficient is $\beta = \Sigma^{-1}(\mu_1 - \mu_0)$. In this case, $\beta = 0$ if and only if $\mu_1 = \mu_0$. Furthermore, when $\mu_1 \neq \mu_0$ then $\beta^{\mathsf{T}}(\mu_1 - \mu_0) > 0$. In some infinitely imbalanced limits where $n_0 \to \infty$ while n_1 remains fixed, the MLE of the slope coefficient depends on x_i for $y_i = 1$ only through $\bar{x}_1$ (Owen, 2007). This work grew from a correspondence about using the result in Owen (2007) to address the conjecture in Ray et al. (2013).

2 Notation and basic result

In the scalar predictor case, the data are (x_i, y_i) for $i = 1, \dots, n$ with $x_i \in \mathbb{R}$ and $y_i \in \{0, 1\}$. There are n_0 observations with $y_i = 0$ and n_1 with $y_i = 1$. To avoid trivial complications, we assume that $\min(n_0, n_1) > 0$.

Let $\bar{x}_1 = (1/n_1)\sum_{i=1}^n x_i y_i$ and $\bar{x}_0 = (1/n_0)\sum_{i=1}^n x_i(1 - y_i)$ be the sample averages of x for observations with $y_i = 1$ and $y_i = 0$, respectively. A logistic regression model has

$$\Pr(Y = 1 \mid X = x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \equiv p(x; \alpha, \beta).$$

The likelihood function is

$$L(\alpha, \beta) = \prod_{i=1}^{n} p(x_i; \alpha, \beta)^{y_i} (1 - p(x_i; \alpha, \beta))^{1-y_i}. \tag{1}$$

This model has a well defined maximum likelihood estimate if the x data for $y = 0$ overlap sufficiently with those for $y = 1$. In Section 4 we state the overlap conditions given by Silvapulle (1981).

For scalar x, Silvapulle's conditions simplify. Let $L_0 = \min\{x_i \mid y_i = 0\}$, $U_0 = \max\{x_i \mid y_i = 0\}$, $L_1 = \min\{x_i \mid y_i = 1\}$, and $U_1 = \max\{x_i \mid y_i = 1\}$ be the extreme values of x in each of the two groups. It is sufficient to have

$$L_0 < U_1 \ \& \ L_1 < U_0. \tag{2}$$

If the intervals $[L_0, U_0]$ and $[L_1, U_1]$ overlap in an interval of positive length, then (2) is satisfied. That is the usual case when logistic regression is used, but other corner cases satisfy Silvapulle's condition too. For instance, if all the x's for one y value, say $y = 1$, are identical, then (2) can still hold so long as $L_0 < L_1 = U_1 < U_0$.

We can even weaken (2) to allow the x values for one group to form a zero-length interval tied with an extreme value from the other group:

$$\begin{aligned} &L_0 = U_0 = L_1 < U_1 \quad \text{or} \quad L_1 < U_1 = L_0 = U_0 \quad \text{or} \\ &L_1 = U_1 = L_0 < U_0 \quad \text{or} \quad L_0 < U_0 = L_1 = U_1. \end{aligned} \tag{3}$$

We cannot however have $L_0 = L_1 = U_0 = U_1$. For scalar x with $n_0 > 0$ and $n_1 > 0$, Silvapulle's conditions are equivalent to (2) or (3).

The sign function we use is defined for $z \in \mathbb{R}$ by $\mathrm{sign}(z) = 1$ for $z > 0$, $\mathrm{sign}(z) = -1$ for $z < 0$, and $\mathrm{sign}(0) = 0$. Our first result is the following:

Theorem 1. *Let $x_i \in \mathbb{R}$ and $y_i \in \{0, 1\}$ for $i = 1, \ldots, n$. Assume that both $n_1 = \sum_{i=1}^{n} y_i > 0$ and $n_0 = n - n_1 > 0$ and that x_i and y_i satisfy an overlap condition* (2) *or* (3). *Then the likelihood* (1) *has a unique maximizer $(\hat{\alpha}, \hat{\beta})$ with* $\mathrm{sign}(\hat{\beta}) = \mathrm{sign}(\bar{x}_1 - \bar{x}_0)$.

The conclusion of Theorem 1 still holds in some cases where neither (2) nor (3) hold, though it may require some interpretation. For instance, if $U_0 < L_1$, then $\bar{x}_1 > \bar{x}_0$ and also $\hat{\beta} = +\infty$. Likewise, if $U_1 < L_0$, then $\bar{x}_1 < \bar{x}_0$ with $\hat{\beta} = -\infty$. So these two cases are included if we take $\mathrm{sign}(\pm\infty) = \pm 1$. An exception arises when all x_i have the same value. Then $\bar{x}_1 - \bar{x}_0 = 0$ but the likelihood has no unique maximizer.

3 Proof of Theorem 1

The existence of a unique maximizer for logistic regression here follows from the theorem in Silvapulle (1981). So we only need to consider the sign of $\hat\beta$.

The log likelihood in the logistic regression is

$$\ell(\alpha,\beta) = \sum_{i=1}^n y_i(\alpha+\beta x_i) - \log(1+\exp(\alpha+\beta x_i)).$$

This is a concave function of the parameter (α,β). The maximum likelihood estimates $(\hat\alpha,\hat\beta)$ are attained by setting

$$0 = \frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^n (y_i - p(x_i;\alpha,\beta)) \tag{4}$$

and

$$0 = \frac{\partial \ell}{\partial \beta} = \sum_{i=1}^n (y_i - p(x_i;\alpha,\beta))x_i. \tag{5}$$

We will abbreviate $p(x_i;\hat\alpha,\hat\beta)$ to $\hat p_i$. From equation (4) we find that $\bar p \equiv (1/n)\sum_{i=1}^n \hat p_i = n_1/n$. This is a well-known consequence of including an intercept term in logistic regression.

Subtracting $\bar x$ times equation (4) from equation (5) yields $\sum_{i=1}^n (y_i - \hat p_i)(x_i - \bar x) = 0$. After rearranging the sum we have

$$n_1(\bar x_1 - \bar x) = \sum_{i=1}^n \hat p_i (x_i - \bar x). \tag{6}$$

On the left of (6) we find that $\mathrm{sign}(\bar x_1 - \bar x) = \mathrm{sign}(\bar x_1 - (n_1/n)\bar x_1 - (n_0/n)\bar x_0) = \mathrm{sign}((n_0/n)(\bar x_1 - \bar x_0)) = \mathrm{sign}(\bar x_1 - \bar x_0)$. For the right side, we consider three cases. First, if $\hat\beta = 0$, then $\hat p_i$ is constant and the right side of (6) is zero.

Second, if $\hat\beta > 0$, then $\hat p_i$ is a strictly increasing function of x_i. Then

$$\sum_{i=1}^n \hat p_i (x_i - \bar x) = \sum_{i=1}^n (\hat p_i - \widetilde p)(x_i - \bar x), \tag{7}$$

where $\widetilde p = \exp(\hat\alpha + \hat\beta \bar x)/(1+\exp(\hat\alpha+\hat\beta\bar x))$. Each term on the right of (7) is a product of two positive numbers, two negative numbers or two zeros. Therefore (7) cannot be negative. Because the x_i cannot all be equal under either (2) or (3), at least two of the terms in (7) must be strictly positive. Therefore the right side of (6) is positive when $\hat\beta > 0$.

Similarly if $\hat\beta < 0$ (the third case) then the right side of (6) is negative. In all three cases, the sign of $\hat\beta$ matches the sign of the right side of (6) and hence equals the sign of $\bar x_1 - \bar x_0$.

4 Generalizations

Here we generalize the connection between $\hat{\beta}$ and the difference in group means for $x_i \in \mathbb{R}$ to some other settings. The first setting is to allow $x_i \in \mathbb{R}^d$ for an integer $d \geq 1$. As before, we let $\bar{x}_0$ and $\bar{x}_1$ be the averages of x_i for $y = 0$ and for $y = 1$ respectively. The logistic regression model is now $\Pr(Y = 1 \mid X = x) = \exp(\alpha + x^\mathsf{T}\beta)/(1 + \exp(\alpha + x^\mathsf{T}\beta))$.

The second generalization extends the logistic model to models of the form $\Pr(Y = 1 \mid X = x) = G(\alpha+x^\mathsf{T}\beta)$ where $G(\cdot)$ is a non-decreasing function from $\mathbb{R}$ to $[0, 1]$. The function G^{-1} is called the link function. The link function applied to $\Pr(Y = 1 \mid X = x)$ is an affine function, $\alpha + x^\mathsf{T}\beta$, of x. Besides the logistic model, some important alternatives are the probit model with $G(z) = \Phi(z)$ where Φ is the CDF of the $\mathcal{N}(0, 1)$ distribution, and the complementary log-log link function whose inverse is $G(z) = 1 - \exp(-\exp(z))$. See McCullagh and Nelder (1989).

4.1 Assumptions

Before generalizing to other contexts, we present conditions that we will need. We need assumptions about the data and assumptions about the link function. For the data, we first extend x_i to $\widetilde{x}_i = (1, x_i^\mathsf{T})^\mathsf{T}$ whose first component creates the intercept term.

Assumption 1 (Full rank condition)**.** The matrix $\mathcal{X} \in \mathbb{R}^{n\times(d+1)}$ with i'th row equal to $\widetilde{x}_i^\mathsf{T}$ has full rank $d + 1 \leq n$.

Assumption 1 is commonly made in regression settings. When it fails to hold, then one of the component variables in x_i can be replaced by an affine combination of the others. In that case, such redundant variables are often dropped from the model until Assumption 1 holds.

Next we state an assumption that keeps the maximum likelihood estimates bounded. This assumption imposes some overlap among the x's for which $y = 0$ and the ones for which $y = 1$. Using the extended predictors, let

$$S = \Biggl\{ \sum_{i:y_i=1} k_i\widetilde{x}_i \mid k_i > 0 \Biggr\}, \quad \text{and} \quad F = \Biggl\{ \sum_{i:y_i=0} k_i\widetilde{x}_i \mid k_i > 0 \Biggr\}. \qquad (8)$$

These are open convex cones in $\mathbb{R}^{d+1}$ generated by the extended predictor values for $y = 1$ and $y = 0$ respectively. Silvapulle's overlap condition is that either $S \cap F \neq \varnothing$ or $S = \mathbb{R}^{d+1}$ or $F = \mathbb{R}^{d+1}$. The latter two possibilities cannot hold in our setting with an intercept, so we only need the first condition, which we label Assumption 2.

Assumption 2. [Overlap condition] Let the cones S and F be defined from the data as at (8). Then $S \cap F \neq \varnothing$.

The above are the assumptions we need on the data. Next, we give an assumption for the link function.

Assumption 3 (Silvapulle's (1981) link condition)**.** The inverse link function G is strictly increasing at every value of z with $0 < G(z) < 1$, and both $-\log(G(z))$ and $-\log(1 - G(z))$ are convex functions of z.

Silvapulle's definition of convexity allows functions that take the value $+\infty$. For example if G is the CDF of the $U[0, 1]$ distribution then both $-\log(G(z))$ and $-\log(1-G(z))$ are convex functions, the former equalling ∞ for $z \le 0$ and the latter equalling ∞ for $z \ge 1$. There is a typographical error in part (iii) of the Theorem in Silvapulle (1981): it supposes that $-\log(G)$ and $\log(1-G)$ are convex but the second one should be $-\log(1 - G)$.

Assumption 3 requires both the CDF G and the survivor function $1 - G$ to be log-concave functions of z. A log-concave function is one whose logarithm is concave. Many probability density functions are log-concave. Both the CDF and survivor functions inherit log concavity from the density function.

Lemma 1. *Let $g(z)$ be a probability density function for $z \in \mathbb{R}$. If $-\log(g(z))$ is convex then $G(z) = \int_{-\infty}^{z} g(t)\,\mathrm{d}t$ satisfies Assumption 3.*

Proof. Log concavity of G and $1 - G$ both follow from Lemma 3 of An (1998). An (1998) requires g to be measurable but that holds automatically for log concave g. Let $z \in \mathbb{R}$ satisfy $0 < G(z) < 1$. Then there is a point $a < z$ with $g(a) > 0$ and a point $b > z$ with $g(b) > 0$. From log concavity of g we have $g(t) > \min(g(a), g(b)) > 0$ for all t in the interval (a, b) which contains z. Therefore G is strictly increasing at z. □

Lemma 1 makes it easy to identify a large family of link functions that satisfy Silvapulle's condition. If $g(z)$ is a log concave density on $\mathbb{R}$ then it satisfies Assumption 3 including the requirement to be strictly positive where the corresponding CDF G satisfies $0 < G(z) < 1$. Bagnoli and Bergstrom (2005) list the following log-concave densities among others: uniform, normal, exponential, logistic, extreme value, double exponential, cz^{c-1} on $(0, 1]$ for $c \ge 1$, Weibull with shape parameter at least 1, and the Gamma distribution with shape at least 1.

The inverse link for the complementary log-log model is easily seen to satisfy Assumption 3: It corresponds to the density $g(z) = \exp(z-\exp(z))$ and $\log(g(z))$ has second derivative $-\exp(z) < 0$. The Cauchy CDF, $G(z) = (1/\pi)\arctan(z)+1/2$, has been suggested for binary regression models (Morgan and Smith, 1992). It is somewhat robust to mislabeling among the y_i, but this G is not log concave.

4.2 Equivalence of $\hat{\beta} = 0$ and $\bar{x}_0 = \bar{x}_1$

The log likelihood function for binary regression with inverse link G is

$$\ell(\alpha, \beta) = \sum_{i=1}^{n} y_i \log(G(\alpha + x_i^{\mathsf{T}}\beta)) + (1 - y_i)\log(1 - G(\alpha + x_i^{\mathsf{T}}\beta)). \tag{9}$$

Theorem 2. *Let $x_i \in \mathbb{R}^d$ and $y_i \in \{0,1\}$ for $i = 1,\dots,n$ satisfy the full rank Assumption 1 and the overlap Assumption 2. Let G satisfy the link Assumption 3. Then $\bar{x}_1 = \bar{x}_0$ if and only if the model* (9) *has a unique maximum likelihood estimate $(\hat{\alpha}, \hat{\beta})$ with $\hat{\beta} = 0$.*

Proof. Under these assumptions, the Theorem in Silvapulle (1981) yields that there is a unique maximum likelihood estimate $(\hat{\alpha}, \hat{\beta})$. It solves the equations

$$0 = \frac{\partial}{\partial \alpha}\ell(\alpha, \beta) = \sum_{i=1}^{n} y_i \frac{g(\alpha + x_i^{\mathsf{T}}\beta)}{G(\alpha + x_i^{\mathsf{T}}\beta)} - (1 - y_i)\frac{g(\alpha + x_i^{\mathsf{T}}\beta)}{1 - G(\alpha + x_i^{\mathsf{T}}\beta)} \tag{10}$$

and

$$0 = \frac{\partial}{\partial \beta}\ell(\alpha, \beta) = \sum_{i=1}^{n} y_i \frac{g(\alpha + x_i^{\mathsf{T}}\beta)}{G(\alpha + x_i^{\mathsf{T}}\beta)}x_i - (1 - y_i)\frac{g(\alpha + x_i^{\mathsf{T}}\beta)}{1 - G(\alpha + x_i^{\mathsf{T}}\beta)}x_i \tag{11}$$

where g is the derivative of G.

If the MLE has $\hat{\beta} = 0$, then we can solve equation (10) to find that $G(\hat{\alpha}) = n_1/n$. Then equation (11) yields $n_1\bar{x}_1 g(\hat{\alpha})/G(\hat{\alpha}) = n_0\bar{x}_0 g(\hat{\alpha})/(1 - G(\hat{\alpha}))$ from which we get $\bar{x}_1 = \bar{x}_0$. Conversely, if $\bar{x}_0 = \bar{x}_1$ then $\hat{\alpha} = G^{-1}(n_1/n)$ and $\hat{\beta} = 0$ jointly solve equations (10) and (11) and hence provide the unique maximum likelihood estimate. □

4.3 Angle between $\hat{\beta}$ and $\bar{x}_1 - \bar{x}_0$

For the special case of logistic regression, where $G(z) = \exp(z)/(1 + \exp(z))$, suppose that Assumptions 1 and 2 are satisfied. Then the maximum likelihood estimates are well defined. Using the same argument as in Section 3, but multiplying both sides of (6) by $\hat{\beta}^{\mathsf{T}}$, we find that $\hat{\beta} = 0$ if and only if $\bar{x}_0 = \bar{x}_1$. Otherwise $\hat{\beta}^{\mathsf{T}}(\bar{x}_1 - \bar{x}_0) > 0$. In other words, when $\bar{x}_1 \neq \bar{x}_0$, then $\hat{\beta}$ makes less than a ninety degree angle with $\bar{x}_1 - \bar{x}_0$. The result holds more generally.

Theorem 3. *Let $x_i \in \mathbb{R}^d$ and $y_i \in \{0,1\}$ for $i = 1,\dots,n$ satisfy the full rank Assumption 1 and the overlap Assumption 2. Let G satisfy the link Assumption 3. If $\bar{x}_1 - \bar{x}_0 \neq 0$ and $(\hat{\alpha}, \hat{\beta})$ maximize the log likelihood* (9)*, then $\hat{\beta}^{\mathsf{T}}(\bar{x}_1 - \bar{x}_0) > 0$.*

Proof. We use two data sets. The original and a shifted one with $x_i^* = x_i - y_i\Delta$ where $\Delta \equiv \bar{x}_1 - \bar{x}_0$. We use ℓ_* to denote the log likelihood of the shifted data set. In the shifted data set, $(1/n_1)\sum_{i=1}^{n} y_i x_i^* = (1/n_0)\sum_{i=1}^{n}(1 - y_i)x_i^*$ by construction. The overlap assumption also holds in the shifted data set. The shifted data set has MLE $\beta_* = 0$ and $\alpha_* = G^{-1}(n_1/n)$. Now suppose to the contrary of the theorem that $\hat\beta^\mathsf{T}\Delta \le 0$. Then

$$\begin{aligned}\ell(\hat\alpha,\hat\beta) &= \sum_{i=1}^n y_i\log(G(\hat\alpha+\hat\beta^\mathsf{T}x_i)) + (1-y_i)\log(1-G(\hat\alpha+\hat\beta^\mathsf{T}x_i))\\ &\le \sum_{i=1}^n y_i\log(G(\hat\alpha+\hat\beta^\mathsf{T}(x_i-\Delta))) + (1-y_i)\log(1-G(\hat\alpha+\hat\beta^\mathsf{T}x_i))\\ &= \ell_*(\hat\alpha,\hat\beta) \le \ell_*(\alpha_*,0) = \ell(\alpha_*,0).\end{aligned}$$

As a result, $\hat\beta$ is not the unique MLE of β that it would have been, had it maximized (9) under the given assumptions. The first inequality arises because G is nondecreasing. The second inequality follows because, from Theorem 2, the maximizers of ℓ_* are $(\alpha_*,0)$. The full rank assumption may fail to hold in the shifted data set, but if it does then $(\alpha_*,0)$ is still a maximizer of ℓ_* though not the unique maximizer, and our proof here does not need uniqueness in the shifted data. □

Acknowledgments

We thank Alan Agresti, Brad Efron, Peter McCullagh, Mervyn Silvapulle and Jeff Wu and Danny Wang for helpful comments. ABO was supported by the National Science Foundation under grant DMS-0906056. PAR was contractually supported by the U.S. Army Armament Research, Development and Engineering Center (ARDEC).

References

An, M. Y. (1998). Logconcavity versus logconvexity: A complete characterization. *Journal of Economic Theory*, 80:350–369.

Bagnoli, M. and Bergstrom, T. (2005). Log-concave probability and its applications. *Economic Theory*, 26(2):445–469.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman & Hall, Boca Raton, FL.

Morgan, B. J. T. and Smith, D. M. (1992). A note on Wadley's problem with overdispersion. *Journal of the Royal Statistical Society, Series C*, 41(2):349–354.

Owen, A. B. (2007). Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, 8:761–773.

Ray, D. M., Roediger, P. A., and Neyer, B. T. (2013). Commentary: Three-phase optimal design for sensitivity experiments. *Journal of Statistical Planning and Inference*. In press.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimates for the binomial response models. *Journal of the Royal Statistical Society, Series B*, 43:310–313.

Wu, C. F. J. and Tian, Y. (2013). Three-phase optimal design of sensitivity experiments. *Journal of Statistical Planning and Inference*. In press.