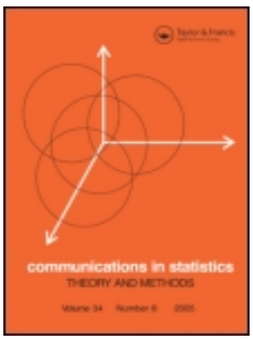

Taylor & Francis
Taylor & Francis Group

Communications in Statistics—Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

Quantifying the Impact of Unobserved Heterogeneity on Inference from the Logistic Model

Salma Ayis

To cite this article: Salma Ayis (2009) Quantifying the Impact of Unobserved Heterogeneity on Inference from the Logistic Model, Communications in Statistics—Theory and Methods, 38:13, 2164-2177, DOI: 10.1080/03610920802491782

To link to this article: https://doi.org/10.1080/03610920802491782

Communications in Statistics—Theory and Methods, 38: 2164–2177, 2009

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920802491782

Taylor & Francis
Taylor & Francis Group

Quantifying the Impact of Unobserved Heterogeneity on Inference from the Logistic Model

SALMA AYIS

Department of Social Medicine, University of Bristol, Bristol, UK

While consequences of unobserved heterogeneity such as biased estimates of binary response regression models are generally known; quantifying these and awareness of situations with more serious impact on inference is however, remarkably lacking. This study examines the effect of unobserved heterogeneity on estimates of the standard logistic model. An estimate of bias was derived for the maximum likelihood estimator $\hat{\beta}$, and simulated data was used to investigate a range of situations that influence size of bias due to unobserved heterogeneity. It was found that the position of the probabilities, along the logistic curve, and the variance of the unobserved heterogeneity, were important determinants of size of bias.

Keywords Biased estimate; Logistic model; Unobserved heterogeneity.

Mathematics Subject Classification Primary 62J12; Secondary 62P10.

1. Introduction

Theoretical models, such as health modes, generally conceptualize outcomes as a result of interaction among a complex set of components including biological, genetic, behavioral, and socio-economics (Mosley and Chen, 2003; World Health Organization, 2001). In practical situations of data analysis, however, it is not possible to account for all variables that result in an outcome by including these as explanatory variables in a statistical model. Even in the richest model specification, several factors would be unobserved, immeasurable or unknown, and some of these would be of high importance to the resulting outcome (Lee and Lee, 2003; Zohoori and Savitz, 1997). Nonetheless, it is not uncommon in many specialized journals to find some conclusions that were reached on the basis of inference from observed variables, assuming unobserved heterogeneity is of little relevance. Economists were puzzled for nearly two decades by the spurious positive association between drinking alcohol (medically known as drug with depressant properties and is unlikely to positively affect productivity) and high wage, where drinkers were

Received May 30, 2008; Accepted September 19, 2008
Address correspondence to Salma Ayis, Department of Social Medicine, Canynge Hall, Whiteladies Road, University of Bristol, BS8 2PR, UK; E-mail: s.ayis@bristol.ac.uk

persistently found to earn more than alcohol abstainers, the association was reversed by the introduction of individual specific fixed effect, which rid results of bias due to time invariant unobserved heterogeneity (Peters, 2004).

Unobserved factors whether environmental or personal may have a large effect on outcomes of relevance (for example, health, economics, social). Ignoring such effects may lead to the identification of incorrect risk factors, in addition the magnitude of the association of these may be so seriously biased and as a result conclusions may be misleading.

The objectives of this study were to quantify bias of the maximum likelihood (ML) estimate, of a standard logistic model due to un-modeled unobserved heterogeneity, and to highlight situations where the impact of such bias was more serious. Using Taylor series theory, an approximate estimate of bias was derived then simulation was used to investigate situations that were thought to affect the size of bias. The importance of higher-order terms of the derived approximation was also examined under various situations, including different variance of unobserved heterogeneity and differing positions for the probabilities of outcome within the logistic curve. The estimation described was confined to the case of a single binary explanatory variable x, a binary response y, and unobserved heterogeneity that was linked to each individual.

It was found that in most situations the first-order approximation defined as δ_1, provides an adequate approximation of bias due to unobserved heterogeneity. At special situations with large variance and large difference between the two probabilities, however, the first-order approximation becomes inadequate, and does worse as the difference between the two probabilities increases.

2. Methods

2.1. *Assumptions and Models*

We consider a hypothetical simple example, i.e., lung cancer, as an outcome, and smoking as the only explanatory variable. It was also assumed that there are other factors that were likely to cause lung cancer but these were either unobserved, difficult to measure, or were totally unknown. These may include some genetic factors, personal differences in diet, childhood exposures, lifestyle, or other individual specific factors. In a statistical model, the relationship may be expressed as:

$$y = f(x, \varepsilon), \tag{2.0}$$

where y was the outcome, x was the observed binary explanatory variable, and ε was the unobserved variable or variables. The conditional expectation of y given x and ε, may be written as:

$$E(y \mid x, \varepsilon) = h(\beta_0 + \beta_1 x + \varepsilon) = h(\eta). \tag{2.1}$$

Under the linear model assumptions, the estimates of β_1 will be unbiased whether ε, was considered by the model or not. At other situations, where nonlinear models such as the logistic regression model (Agresti, 2002) were used, the effect of the omission of ε on estimates was, however, different. The situation with a missing

binary variable was described before (Gail et al., 1984). Here, we shall consider the situation, where the omitted variable was continuous which is quite common in many research areas, such as health. The probability of positive outcome (response) in our example with one explanatory variable x, may be written as:

$$p_1(x) = p(Y = 1 \mid X = x) \tag{2.2}$$

and the logit as:

$$\text{logit}[p_1(x)] = \log\left(\frac{p_1(x)}{1 - p_1(x)}\right) = x'\beta. \tag{2.3}$$

The maximum likelihood (ML) estimate equation for such a model may be written as:

$$\hat{\beta} = \log n_1\hat{p}_1 - \log n_1(1 - \hat{p}_1) - \log n_0\hat{p}_0 + \log n_0(1 - \hat{p}_0), \tag{2.4}$$

where n_1 and n_0 were the numbers of subjects exposed ($X = 1$) and unexposed ($X = 0$), respectively, and the conditional probabilities, $\hat{p}_1 = p(Y = 1 \mid X = 1)$ and $\hat{p}_0 = p(Y = 1 \mid X = 0)$, are the probabilities of positive response, among the exposed and unexposed subjects, respectively.

2.2. *Estimation of Bias, Using Taylor Series Expansion: The Uncorrelated Case*

We consider the situation where sampled observations were identically independently distributed (IID), expected value of y, $E(Y)$ exists, X was a binary variable (0/1), and ε was a random variable that was unobserved but has an influence on the response y. The unobserved term was assumed to follow a normal distribution, with mean zero and a variance σ_ε^2. The maximum likelihood estimator, $\hat{\beta}$ for the logistic model, with one explanatory variable x, was first obtained ignoring unobserved heterogeneity; the effect of unobserved heterogeneity, ε on the response was then brought in, and the expected value of the (ML) estimator was evaluated again. Taylor series expansion was used and the estimates were compared (ignoring unobserved heterogeneity and including the influence of the term on the response y). The derivation showed that the (ML) estimator is biased due to un-modeled unobserved heterogeneity and the bias may be approximated by a first order term δ_1, of Taylor series as follows:

$$\delta_1 = \left\{\frac{g_1(p_1) - p_1}{p_1(1 - p_1)}\right\} - \left\{\frac{g_1(p_0) - p_0}{p_0(1 - p_0)}\right\}. \tag{2.5}$$

The bias may also be approximated by the first- and second-order terms, $\delta_1 + \delta_2$, and/or by first-, second-, and third-order terms, $\delta_1 + \delta_2 + \delta_3$, where δ_i for $i = 1, 2$, and 3 together with details of the terms $g_1(p_1)$ and $g_1(p_0)$ were fully described in Appendix A, was the derivation steps.

2.3. *Simulation*

Simulation was used to examine situations that were thought to affect the severity of bias. Data was simulated from a logistic model with a response y, an explanatory variable x, and an extra term representing unobserved heterogeneity. A Fortran program was written to generate data from a logit function of the form:

$$\log\left\{\frac{p_j}{1-p_j}\right\} = \beta_0 + x_j'\beta_1 + \varepsilon_j, \tag{2.6}$$

where x_j is a vector of a binary variable for subjects $j = 1, 2, \ldots, n$, $p_j = p(y_j = 1 \mid x_j)$ is the conditional probability, β_0 and β_1 are the logistic regression parameters, and ε_j an extra term representing unobserved heterogeneity, assumed to be normally distributed random variable with zero mean and a variance σ_ε^2. Several parameter values were considered, so that they cover a range of probabilities, including situations where one of the probabilities was 0.5, the two probabilities were lying on one side of 0.5, and where they were lying further apart on the two sides of 0.5. We also considered a range of values for the variance of the unobserved heterogeneity term. The standard logistic model was then fitted to predict the response y, using x as the only explanatory variable, in order to detect the effect of ignoring unobserved heterogeneity on the estimated parameter, $\hat{\beta}$. Estimates were calculated from 100 simulations, each based on a sample of 1,000 of identically independently distributed (IID) individuals.

3. Results

3.1. *Theoretical Bias: The Effect of Position of the Two Probabilities*

We examined the size of bias with focus on the importance of first-, second-, and third-order terms in the overall bias approximation. Table 1 illustrates the contribution of each term under varying situations. Different positions of the two

Table 1
Theoretical contribution of first-, second-, and third-order terms to the approximation of bias due to unobserved heterogeneity, with variance $\sigma_\varepsilon^2 = 1.0$

p_0, p_1	δ_1	δ_2	δ_3
0.73, 0.82	−0.08	0.02	0.0
0.88, 0.99	−0.28	0.13	−0.03
0.73, 0.88	−0.15	0.04	−0.01
0.50, 0.38	−0.08	0.0	0.0
0.50, 0.73	−0.18	0.0	0.0
0.50, 0.82	−0.26	0.02	−0.01
0.50, 0.92	−0.39	0.09	−0.02
0.38, 0.62	−0.20	0.0	0.0
0.27, 0.73	−0.36	0.01	0.0
0.12, 0.88	−0.67	0.10	−0.03

probabilities were examined, while the variance of unobserved heterogeneity was kept fixed at 1.0. The positions considered include: (i) the two probabilities lie on one side of 0.5 of the logistic curve; (ii) one of the probabilities is 0.5; and (iii) the two probabilities lie on the opposite side of 0.5.

For all three situations, the first term δ_1 was dominant, the second-order term is of less importance, and the third-order term has a very small contribution that may be safely ignored at almost all situations. There are, however, special situations where the second-order term becomes more important, for example, as one probability approaches 0.9.

Table 2

A comparison between theoretical bias and numerical (simulation) bias, for the (MLH) estimator $\hat{\beta}$, of the logistic model, for various probabilities, and for a range of variance (0.01–1.0) for the unobserved heterogeneity term ε

Variance σ_ε^2	Theoretical bias				Simulation bias			
	δ_1	δ_2	δ_3	Total	Bias $\hat{\beta}-\beta$	95% C.I	Rej%	(p_0, p_1)
1.0	−0.12	0.0	0.0	−0.12	−0.12	−0.15, −0.09	13	0.62, 0.73
0.75	−0.07	0.01	0.0	−0.06	−0.07	−0.10, −0.04	12	
0.5	−0.04	0.0	0.0	−0.05	−0.05	−0.06, 0.0	7	
0.1	−0.01	0.0	0.0	−0.01	−0.01	−0.04, 0.02	6	
0.01	0.0	0.0	0.0	0.0	0.01	−0.03, 0.03	3	
1.0	−0.39	0.11	−0.03	−0.31	−0.33	−0.36, −0.30	36	0.62, 0.95
0.75	−0.29	0.06	−0.01	−0.24	−0.23	−0.27, −0.19	17	
0.5	−0.19	0.02	0.0	−0.17	−0.15	−0.19, −0.11	15	
0.10	−0.03	0.0	0.0	−0.03	−0.04	−0.09, 0.01	6	
0.01	0.0	0.0	0.0	0.0	0.05	0.0, 0.10	4	
1.0	−0.09	0	0	−0.10	−0.08	−0.10, −0.06	12	0.5, 0.62
0.75	−0.07	0	0	−0.07	−0.06	−0.08, −0.04	8	
0.5	−0.05	0	0	−0.05	−0.05	−0.07, −0.03	5	
0.1	−0.01	0	0	−0.01	0	−0.03, 0.03	4	
0.01	0.0	0	0	0.0	0.01	−0.02, 0.02	2	
1.0	−0.34	0.05	−0.01	−0.30	−0.29	−0.32, −0.26	49	0.5, 0.88
0.75	−0.27	0.03	−0.01	−0.24	−0.23	−0.27, −0.21	27	
0.5	−0.18	0.01	0	−0.17	−0.16	−0.19, −0.13	14	
0.1	−0.04	0	0	−0.04	−0.03	−0.06, 0.0	4	
0.01	0	0	0	0.0	−0.01	−0.02, 0.02	3	
1.0	−0.48	0.11	−0.03	−0.40	−0.39	−0.43, −0.35	49	0.5, 0.95
0.75	−0.36	0.04	−0.01	−0.32	−0.31	−0.35, −0.27	32	
0.5	−0.25	0.04	0	−0.21	−0.20	−0.24, −0.16	21	
0.1	−0.05	0	0	−0.04	−0.03	-0.08, 0.02	7	
0.01	0.0	0	0	0	0.06	0.01, 0.11	5	
1.0	−0.20	0	0	−0.20	−0.17	−0.19, −0.15	20	0.38, 0.62
0.75	−0.15	0	0	−0.15	−0.14	−0.17, −0.11	18	
0.5	−0.10	0	0	−0.10	−0.10	−0.13, −0.07	13	
0.10	−0.02	0	0	−0.02	−0.03	−0.06, 0.0	10	
0.01	0	0	0	0	.03	−0.01, 0.04	3	

Note: 95% C.I : upper and lower 95% confidence intervals for the simulated bias.

Rej%: the number of times, in percentage that the true parameter β, lied outside the 95% confidence intervals of the simulated estimator $\hat{\beta}$.

3.2. *Simulation Results: Theoretical and Numerical Bias*

Table 2, reports the average estimates of 100 simulations, each based on a sample of size 1,000. The table examined various positions for the two probabilities, and a range of variance for unobserved heterogeneity (1.0–0.01) including the special case where unobserved heterogeneity, was almost constant ($\sigma_\varepsilon^2 = 0.01$). Three situations regarding the position of the two probabilities were covered as described earlier. A comparison was then drawn between the theoretical bias, and the numerical bias, which was calculated from the simulation. The 95% confidence intervals for the simulated bias were reported and the number of times the confidence intervals of the estimate $\hat{\beta}$, fail to cover the true parameter β was reported as percentage rejected (Rej%). The main findings may summarized as:

1. The bias was well approximated by the first-order term at most of the situations considered.
2. The bias becomes more serious as the difference between the two probabilities gets larger, especially as one probability gets closer to 0.9 (the situation with one probability approaching 0.1 is identical) or where the two probabilities lie further apart at the two sides of the logistic curve.
3. The bias was more serious for relatively large variance of unobserved heterogeneity, 1.0 and 0.75, from the range of values considered.
4. Contributions from the second-order term to the bias approximation become of some importance at special situation where the difference between the two probabilities was large, and the variance of unobserved heterogeneity was relatively large.
5. For small variance of unobserved heterogeneity, the bias was small and the percentage of rejections was modest.
6. For the special case of unobserved heterogeneity with variance $= 0.01$, the bias either disappeared or became negligible, and the percentage of estimates outside the coverage property of $\hat{\beta}$ was particularly small, less than 5% in most cases.

4. Discussion

Much attention in the 1980's, and after, was given to the asymptotic bias of the maximum likelihood (ML) estimators of binary response regression models, that are widely used to describe associations between binary outcome and explanatory variables in trials and surveys. In general, ML estimators may not hold in small and finite samples, as shown by Anderson and Richardson (1979) where a simulation was used to investigate bias of the logistic model estimates, the study found that bias can be substantial if the sample size is small, a formulae for correction was developed. Another similar study (Griffiths et al., 1987) examined the bias and other sample properties such as mean square error based on three alternative covariance matrix estimators for the Probit model, also reached the same conclusion with regard to the bias. A simpler formula using Taylor series expansion for correction of bias in logistic ML estimate was also developed by Copas (1988). For exponential family such as the logistic model, the bias was of order $O(1/n)$ suggesting that for large samples it was negligible relative to the standard errors of the estimates, the bias was treated by Jeffery's priors' as reported in McCullagh and Nelder (1989) and Firth (1992). A set of GLIM macros was developed to reduce bias (Firth, 1993; Steyerberg and Eijkemans, 2004), but the reduction achieved was reported to be small.

Another issue of importance to the (ML) estimation was the deviation of data from the assumed identical independent distribution, that was addressed in survey methodology and procedures for correction of estimates were developed (Skinner and Smith, 1989).

Of no less importance was the problem of unobserved heterogeneity, although awareness of the problem has recently increased (Aprahamian et al., 2007; Arana and Leon, 2006; Cramer, 2007), but still many influential articles continue to report important findings ignoring the possibility of any impact of unobserved heterogeneity on these findings. In practice, in almost any biological investigation, there are factors (exogenous or endogenous, independent of a biological process, or part of it, time varying or time invariant, personal or contextual) that would be unobserved. Using nonlinear models under such situations lead to biased estimates of population parameters.

Here we present the case with unobserved heterogeneity that was linked to individuals (for example, taste, charisma, emotions), another scenario is where unobserved heterogeneity was correlated, that may occur with repeated measurements within individual, or due to clusters such as household (family related gene, for example), more details on the correlated case were reported in an earlier study (Ayis, 1995) where it was shown that the leading term of bias approximation for the correlated case was the same as that for the non correlated. An extra term, due to replications, however, becomes important if the number of clusters was small, and where the two probabilities lie further apart within the logistic curve.

While there are situations where estimates from the logistic model may be fairly robust to unobserved heterogeneity, there are others where the problem deserves more attention. For situations with outcomes such as fertility or incidence of disease, where all of the probabilities were on the same side of 0.5, the potential for bias was there, but perhaps not as bad as where extreme probabilities occur in both tails, for example ($p_0 \leq 0.2$, $p_1 \geq 0.8$). The work by Copas (1988) is also relevant to the latter situation of extreme probabilities, although the assumption was that extreme values occur due to mis-recording, that is where the values of the response "y" was being transposed in error between 0 and 1, rather than due to the nature of the association between the response and the explanatory variable we present. Mont Carlo simulation was used to examine the sensitivity of different binary response models to such extreme values of probabilities, a model was proposed to allow for robust estimation where a small number of outcomes was being mis-recorded, and techniques for diagnostics where developed. For situations like the one we present in this study, extreme values will be more common, but detection of such values may help in assessing the quality of the estimates and possibly in whether to use alternative methods at situations were the bias is serious.

The misspecification bias of the logistic model, due to unobserved heterogeneity described in this study, is similar to the misspecification bias due to incorrectly assuming the error term was logistically distributed when it was not, as described in Horowitz (1993), where the effect on estimates was measured using simulation. The bias was found to be small as long as the assumed error distribution has the same qualitative shape as the true distribution (unimodal for the logistic case considered) and more serious when the true distribution of the error term was bimodal or heteroskedastic. These findings suggest the need to explore the effect of other forms of distribution on the bias derived in this study. Similar findings were shown by Arana and Leon (2005) where a Monte-Carlo simulation was used to test

the performance of a Bayesian mixture normal distribution (semi parametric model), with other parametric models (including the logit) and nonparametric models, using alternative assumptions for the error distribution and using different sample sizes, when data exhibit unobserved heterogeneity. The mean square error (MSE) in all models and for all sample sizes was found to be considerably large reflecting the difficulty in modeling this type of data. The Bayesian specification, however, performed better than the competing models for small as well as for large samples, and substantially reduces the bias and the MSE, the improvements in bias was larger for small samples.

Misspecification bias due to unobserved heterogeneity can seriously affect estimates from the logistic model as well as other binary regression models. Using panel data and suitable random effect models that allow for individual fixed effect adjustment is a solution, but obtaining such data may not always be easy. Attempting other semiparametric models such as Bayesian normal mixture model seems to be an appealing approach, especially as suitable software are rapidly developing. Further work is needed to examine the impact on estimates in real situations where there are several explanatory variables often correlated. Methods of detection may also be developed to assess the need for alternative more flexible models, at situations where unobserved heterogeneity is likely to have more serious impact on the estimates due to data structure, before drawing inference that might be misleading.

Appendix A

Consider a MLH estimator, $\hat{\beta}$ for a logistic model with binary response y and a single binary explanatory variable x,

$$\hat{\beta} = \log n_1\hat{p}_1 - \log n_1(1-\hat{p}_1) - \log n_0\hat{p}_0 + \log n_0(1-\hat{p}_0). \tag{A.1}$$

We may rewrite the first term of the right-hand side (R.H.S) of Eq. (A.1) as follows:

$$\log n_1\hat{p}_1 = \log n_1 p_1\left\{1 + \frac{\hat{p}_1 - p_1}{p_1}\right\}. \tag{A.2}$$

Using Taylor series expansion, we rewrite Eq. (A.2) as follows:

$$\log n_1\hat{p}_1 = \log n_1 p_1 + \underbrace{\left\{\frac{\hat{p}_1 - p_1}{p_1}\right\}}_{\text{I}} - \underbrace{\frac{1}{2}\left\{\frac{\hat{p}_1 - p_1}{p_1}\right\}^2}_{\text{II}} + \underbrace{\frac{1}{3}\left\{\frac{\hat{p}_1 - p_1}{p_1}\right\}^3}_{\text{III}} + \cdots H.O.T. \tag{A.3}$$

The symbols, I, II, and III were introduced to make referral to the original, more complex terms simpler, each term will be manipulated by algebraic procedures separately, (H.O.T. stands for higher-order terms). We first consider term I; we rewrite it in a form that involves the response y;

$$\left\{\frac{\hat{p}_1 - p_1}{p_1}\right\} = \frac{1}{n_1}\sum\left\{\frac{y_{1j} - p_1}{p_1}\right\} = \frac{1}{n_1 p_1}\sum y_{1j} - 1. \tag{A.4}$$

To appreciate the effect of unobserved heterogeneity defined by term ε, we consider the conditional expectation of the response y_{1j} given ε, we evaluate the expectation of each of the three terms, I, II, and III, and then collecting terms with common powers of σ_ε. Some preliminary results are needed for the expectation and these will be described in this section. We first consider the expectation of the response y, given unobserved heterogeneity ε,

$$E(y_{1j}) = E_1\{E_2(y_{1j} \mid \varepsilon)\}, \tag{A.5}$$

where E_1, is the expectation over ε, and E_2 is the expectation given ε

$$E_2(y_{1j} \mid \varepsilon) = \left\{\frac{p_1}{1-p_1}e^{\varepsilon}\right\} \Big/ \left\{1 + \frac{p_1}{1-p_1}e^{\varepsilon}\right\}. \tag{A.6}$$

We introduce a function $f(\varepsilon)$, such that

$$f(\varepsilon) = \left\{\frac{p_1}{1-p_1}e^{\varepsilon}\right\} \Big/ \left\{1 + \frac{p_1}{1-p_1}e^{\varepsilon}\right\}. \tag{A.7}$$

Taylor, series theorem, was then used to expand the function about zero, the expansion may be written as:

$$f(\varepsilon) = f(0) + f^{(1)}(0)\varepsilon + f^{(2)}(0)\frac{\varepsilon^2}{2!} + f^{(3)}(0)\frac{\varepsilon^3}{3!} + \cdots + f^{(n-1)}(0)\frac{\varepsilon^{n-1}}{(n-1)!} + \text{H.O.T}, \tag{A.8}$$

where $f^{(1)}, f^{(2)}, \ldots, f^{(n-1)}$ are the first, second, $\ldots, (n-1)$th, derivatives of the function, and the derivatives for $f(\varepsilon)$, are as follows;

$$f^{(1)}(\varepsilon) = \left\{\frac{p_1}{1-p_1}e^{\varepsilon}\right\} \Big/ \left\{1 + \frac{p_1}{1-p_1}e^{\varepsilon}\right\}^2 \tag{A.9}$$

$$f^{(2)}(\varepsilon) = \frac{\left\{1 + \frac{p_1}{1-p_1}e^{\varepsilon}\right\}^2\left\{\frac{p_1}{1-p_1}e^{\varepsilon}\right\} - 2\left\{\frac{p_1}{1-p_1}e^{\varepsilon}\right\}^2\left\{1 + \frac{p_1}{1-p_1}e^{\varepsilon}\right\}}{\left\{1 + \frac{p_1}{1-p_1}e^{\varepsilon}\right\}^4}$$

$$= \frac{\left\{\frac{p_1}{1-p_1}e^{\varepsilon}\right\}\left\{1 - \frac{p_1}{1-p_1}e^{\varepsilon}\right\}}{\left\{1 + \frac{p_1}{1-p_1}e^{\varepsilon}\right\}^3} \tag{A.10}$$

$$f^{(3)}(\varepsilon) = \frac{\left\{\frac{p_1}{1-p_1}e^{\varepsilon}\right\}[1 - 4\left\{\frac{p_1}{1-p_1}e^{\varepsilon}\right\} + \left\{\frac{p_1}{1-p_1}e^{\varepsilon}\right\}^2]}{\left\{1 + \frac{p_1}{1-p_1}e^{\varepsilon}\right\}^4}. \tag{A.11}$$

To simplify notation, define $q = \left\{\frac{p_1}{1-p_1}e^{\varepsilon}\right\}$; accordingly, $f^{(3)}(\varepsilon)$ may be written as:

$$f^{(3)}(\varepsilon) = \frac{q[1 - 4q + q^2]}{(1+q)^4}, \tag{A.12}$$

and, the fourth, fifth, and sixth's derivatives as:

$$f^{(4)}(\varepsilon) = \frac{q\{(1+q)^4(1-8q+3q^2) - 4q(1-4q+q^2)(1+q)^3\}}{(1+q)^8}$$
$$= \frac{q(1-11q+11q^2-q^3)}{(1+q)^5} \tag{A.13}$$
$$f^{(5)}(\varepsilon) = \frac{q\{(1+q)^5(1-22q+33q^2-4q^3) - 5q(1-11q+11q^2-q^3)(1+q)^4\}}{(1+q)^{10}}$$
$$= \frac{q\{(1-26q+66q^2-26q^3+q^4)\}}{(1+q)^6} \tag{A.14}$$
$$f^{(6)}(\varepsilon) = \frac{\begin{array}{c} q\{(1+q)^6(1-52q+198q^2-104q^3+5q^4) \\ -6q(1-26q+66q^2-26q^3+q^4)(1+q^5)\} \end{array}}{(1+q)^{12}}$$
$$= \frac{q\{(1-57q+302q^2-302q^3+57q^4-q^5)\}}{(1+q)^7}. \tag{A.15}$$

Accordingly, the functions about zero are:

$$f(0) = p_1 \tag{A.16}$$
$$f^{(1)}(0) = p_1(1-p_1) \tag{A.17}$$
$$f^{(2)}(0) = p_1(1-p_1)(1-2p_1) \tag{A.18}$$
$$f^{(3)}(0) = p_1(1-p_1)(1-6p_1+6p_1^2) \tag{A.19}$$
$$f^{(4)}(0) = p_1(1-p_1)(1-14p_1+36p_1^2-24p_1^3) \tag{A.20}$$
$$f^{(5)}(0) = p_1(1-p_1)(1-30p_1+150p_1^2-240p_1^3+120p_1^4) \tag{A.21}$$
$$f^{(6)}(0) = p_1(1-p_1)(1-62p_1+540p_1^2-1560p_1^3+1800p_1^4-720p_1^5) \tag{A.22}$$

Based on the normality assumption of ε, the odd moments would be equal to zero, the even moments were: $E(\varepsilon^2) = \sigma_\varepsilon^2$, $E(\varepsilon^4) = 3\sigma_\varepsilon^4$, $E(\varepsilon^6) = 15\sigma_\varepsilon^6 \ldots$ etc. Hence, if we consider the first six derivatives of $f(\varepsilon)$, and if terms up to and including σ_ε^6 were involved, we may rewrite the expectations of Term I as:

$$E_1[E_2(y_{1j} \mid \varepsilon)]$$
$$= p_1 + \frac{\sigma^2}{2}p_1(1-p_1)(1-2p_1) + \frac{3\sigma^4}{4!}p_1(1-p_1)(1-14p_1+36p_1^2-24p_1^3)$$
$$+ \frac{15\sigma_\varepsilon^6}{6!}p_1(1-p_1)(1-62p_1+540p_1^2-1560p_1^3+1800p_1^4-720p_1^5). \tag{A.23}$$

For the convenience of notations, we define Eq. (A.23) as:

$$E_1[E_2(y_{1j}|\varepsilon)] = g_1(p_1);$$

hence,

$$E_1\left\{E_2\left(\frac{\hat{p}_1 - p_1}{p_1}\right)\right\}$$
$$= \left\{\frac{g_1(p_1) - p_1}{p_1}\right\} \tag{A.24}$$
$$= (1-p_1)\left\{\begin{array}{l}\frac{\sigma_\varepsilon^2}{2}(1-2p_1) + \frac{3\sigma_\varepsilon^4}{4!}p_1(1-14p_1+36p_1^2-24p_1^3) \\ \quad + \frac{15\sigma_\varepsilon^6}{6!}(1-62p_1+540p_1^2-1560p_1^3+1800p_1^4-720p_1^5)\end{array}\right\}. \tag{A.25}$$

Corollary A.1. *Consider the expectations of corresponding terms of* $\log n_1(1-\hat{p}_1)$ *of* Eq. (A.1)*. These may be written as*:

$$E_1\left\{E_2\left(\frac{1-\hat{p}_1-(1-p_1)}{(1-p_1)}\right)\right\}$$
$$= -\left[\begin{array}{l}\frac{\sigma_\varepsilon^2}{2}p_1(1-2p_1) + \frac{3\sigma_\varepsilon^4}{4!}p_1(1-14p_1+36p_1^2-24p_1^3) \\ \quad + \frac{15\sigma_\varepsilon^6}{6!}p_1(1-62p_1+540p_1^2-1560p_1^3+1800p_1^4-720p_1^5)\end{array}\right] \tag{A.26}$$
$$= -\left\{\frac{g_1(p_1) - p_1}{(1-p_1)}\right\}. \tag{A.27}$$

We apply the same procedures for terms associated with $\hat{p}_0$ and $1-\hat{p}_0$, from Eq. (A.1) and define similar functions of p_0 at this stage.

Now consider Term II,

$$\left\{\frac{\hat{p}_1 - p_1}{p_1}\right\}^2 = \frac{1}{(n_1p_1)^2}\sum (y_{1j})^2 - \frac{2}{n_1p_1}\sum (y_{1j}) + 1 \tag{A.28}$$
$$= \frac{1}{(n_1p_1)^2}\sum y_{1j}^2 + \sum_j\sum_{j'}\frac{y_{1j}y_{1j'}}{(n_1p_1)^2} + 1 - \frac{2}{n_1p_1}\sum y_{1j}. \tag{A.29}$$

Since y_{1j}, is binary, the conditional expectation is:

$$E_1\{E_2(y_{1j}^2 \mid \varepsilon)\} = E_1\{E_2(y_{1j} \mid \varepsilon)\} \tag{A.30}$$
$$E_1\{E_2(y_{1j}y_{1j'} \mid \varepsilon_j\varepsilon_{j'})\} = \{E_1E_2(y_{1j}|\varepsilon)\}^2. \tag{A.31}$$

Assuming conditional independence of y_{1j}, we may then write:

$$E_1\left\{E_2\left(\frac{\hat{p}_1 - p_1}{p_1}\right)^2\right\} = o\left(\frac{1}{n_1}\right) + \frac{n_1(n_1-1)}{(n_1p_1)^2}[g_1(p_1)]^2 + 1 - \frac{2}{p_1}g_1(p_1), \tag{A.32}$$

where $o\left(\frac{1}{n_1}\right)$ includes all terms of order $(n_1)^{-1}$, which are zero based on the asymptotic theory assumption, hence the R.H.S of Eq. (A.32) may be written as:

$$E_1\left\{E_2\left(\frac{\hat{p}_1 - p_1}{p_1}\right)^2\right\} = \left\{\frac{g_1(p_1) - p_1}{p_1}\right\}^2 + o\left(\frac{1}{n_1}\right), \tag{A.33}$$

by including corresponding terms of $\log n_1(1-\hat{p}_1)$, from Eq. (A.1) these give:

$$E_1\left\{E_2\left(\frac{\hat{p}_1-p_1}{1-p_1}\right)^2\right\}=\left\{\frac{g_1(p_1)-p_1}{(1-p_1)}\right\}^2. \tag{A.34}$$

We also consider Term III, and apply similar procedures; we write:

$$\left\{\frac{\hat{p}_1-p_1}{p_1}\right\}^3=\left\{\sum_j\frac{y_{1j}}{(n_1p_1)}-1\right\}^3 \tag{A.35}$$

$$=\frac{1}{(n_1p_1)^3}\{\sum y_{1j}\}^3-\frac{3}{(n_1p_1)^2}\{\sum y_{1j}\}^2+\frac{3}{n_1p_1}\{\sum y_{1j}\}-1 \tag{A.36}$$

$$=\frac{1}{(n_1p_1)^3}\left\{\sum y_{1j}^3+3\sum_{j\pm j'}\sum y_{1j}^2y_{1j'}+\sum_{j\neq j'\neq j''}\sum\sum y_{1j}y_{1j'}y_{1j''}\right\}$$

$$-\frac{3}{(n_1p_1)^2}\left\{\sum y_{1j}^2+\sum_{j\pm j'}\sum y_{1j}y_{1j'}\right\}+\frac{3}{(n_1p_1)}\{\sum y_{1j}\}-1; \tag{A.37}$$

therefore,

$$E_1\left\{E_2\left(\frac{\hat{p}_1-p_1}{p_1}\right)^3\right\}=\frac{n_1(n_1-1)(n_1-2)}{(n_1p_1)^3}\{g_1(p_1)\}^3-\frac{3n_1(n_1-1)}{(n_1p_1)^2}\{g_1(p_1)\}^2$$

$$+\frac{3}{p_1}g_1(p_1)-1+o\left(\frac{1}{n_1}\right) \tag{A.38}$$

$$=\left\{\frac{g_1(p_1)-p_1}{p_1}\right\}^3+o\left(\frac{1}{n_1}\right). \tag{A.39}$$

If we similarly consider the corresponding terms of $\log n_1(1-\hat{p}_1)$ of Eq. (A.1), that gives:

$$-E_1\left\{E_2\left(\frac{\hat{p}_1-p_1}{1-p_1}\right)^3\right\}=-\left\{\frac{g_1(p_1)-p_1}{(1-p_1)}\right\}^3 \tag{A.40}$$

Proof. We first restrict the bias to include only the first-order terms. The proof comes directly by substituting in Eq. (A.1), terms from I, from Eqs. (A.24), and complementary terms from (A.27), plus other similar terms for, $\log(n_1\hat{p}_0)$ and $\log(n_1(1-\hat{p}_0))$, the later terms are identical to the functions of $\hat{p}_1$ and $(1-\hat{p}_1)$, but they are functions of $\hat{p}_0$ and $(1-\hat{p}_0)$ and they take different signs as Eq. (A.1) showed. The expectation of $\hat{\beta}$ may then be written as:

$$E_1\{E_2(\hat{\beta})\}=\beta+\left\{\frac{g_1(p_1)-p_1}{p_1(1-p_1)}\right\}-\left\{\frac{g_1(p_0)-p_0}{p_0(1-p_0)}\right\}; \tag{A.41}$$

therefore,

$$\text{bias}\,(\hat{\beta})=\left\{\frac{g_1(p_1)-p_1}{p_1(1-p_1)}\right\}-\left\{\frac{g_1(p_0)-p_0}{p_0(1-p_0)}\right\}=\delta_1 \tag{A.42}$$

if the second-order term II of Eq. (A.3) was also considered by bringing its relevant components from Eqs. (A.33) and (A.34), and the corresponding terms related to $\hat{p}_0$ and $(1-\hat{p}_0)$, and substitute all in Eq. (A.1), then we may rewrite the expectations of $\hat{\beta}$ as:

$$E_1\{E_2(\hat{\beta})\} = \beta + \delta_1 - \frac{1}{2}\left[\left\{\frac{g_1(p_1)-p_1}{p_1(1-p_1)}\right\}^2(1-2p_1) - \left\{\frac{g_1(p_0)-p_0}{p_0(1-p_0)}\right\}^2(1-2p_0)\right]. \tag{A.43}$$

To simplify we write:

$$E_1\{E_2(\hat{\beta})\} = \beta + \delta_1 + \delta_2, \tag{A.44}$$

where

$$\delta_2 = -\frac{1}{2}\left[\left\{\frac{g_1(p_1)-p_1}{p_1(1-p_1)}\right\}^2(1-2p_1) - \left\{\frac{g_1(p_0)-p_0}{p_0(1-p_0)}\right\}^2(1-2p_0)\right]. \tag{A.45}$$

If term III was also considered, components from Eqs. (A.39) and (A.40), and similar terms relevant to $\hat{p}_0$ and $(1-\hat{p}_0)$, were also brought and substituted in Eq. (A.1), by procedures similar to those used in manipulating and including I and II, the expected value $E_1\{E_2(\hat{\beta})\}$ may be written as:

$$\begin{aligned} E_1\{E_2(\hat{\beta})\} = \beta + \delta_1 &- \frac{1}{2}\left[\left\{\frac{g_1(p_1)-p_1}{p_1(1-p_1)}\right\}^2(1-2p_1) - \left\{\frac{g_1(p_0)-p_0}{p_0(1-p_0)}\right\}^2(1-2p_0)\right] \\ &+ \frac{1}{3}\left[\left\{\frac{g_1(p_1)-p_1}{p_1(1-p_1)}\right\}^3(1-3p_1+3p_1^2) - \left\{\frac{g_1(p_0)-p_0}{p_0(1-p_0)}\right\}^3(1-3p_0+3p_0^2)\right], \end{aligned} \tag{A.46}$$

where

$$\delta_3 = \frac{1}{3}\left[\left\{\frac{g_1(p_1)-p_1}{p_1(1-p_1)}\right\}^3(1-3p_1+3p_1^2) - \left\{\frac{g_1(p_0)-p_0}{p_0(1-p_0)}\right\}^3(1-3p_0+3p_0^2)\right].$$

To simplify, we may write:

$$E_1\{E_2(\hat{\beta})\} = \beta + \delta_1 + \delta_2 + \delta_3 \tag{A.47}$$

and

$$\text{bias}\,(\hat{\beta}) = \delta_1 + \delta_2 + \delta_3. \tag{A.48}$$

By substituting, the actual components of $g_1(p_1)$ and $g_1(p_0)$, in Eqs. (A.41), (A.43), and (A.46), we may evaluate bias, including first-, second-, and third-order terms.

Acknowledgments

I am grateful to professor D. Holt for the initiatives, advice, and suggestions he gave to the theoretical and simulation investigations. Much thanks to Dr. Marie South and Dr. Peter Egger for considerable help with the FORTRAN programs used.

References

Agresti, A. (2002). *Categorical Data Analysis*. Probability and Statistics. New York: Wiley-Interscience.

Anderson, J. A., Richardson, S. C. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics* 21:71–78.

Aprahamian, F., Chanel, O., Luchini, S. (2007). Modeling starting point bias as unobserved heterogeneity in contingent valuation surveys: an application to air pollution. *Amer. J. Agri. Econ.* 89:533–547.

Arana, J. E., Leon, C. J. (2005). Flexible mixture distribution modeling of dichotomous choice contingent valuation with heterogenity 250541. *J. Environm. Econ. Manage.* 50:170–188.

Arana, J. E., Leon, C. J. (2006). Modelling unobserved heterogeneity in contingent valuation of health risks. *Appl. Econ.* 38:2315–2325.

Ayis, S. A. M. (1995). *Modelling Unobserved Heterogeneity: Theoretical and Practical Aspects*. Southampton, UK: University of Southampton.

Copas, J. B. (1988). Binary regression models for contaminated data. *J. Roy. Statist. Soc. Ser. B (Methodological)* 50:225–265.

Cramer, J. S. (2007). Robustness of logit analysis: unobserved heterogeneity and mis-specified disturbances. *Oxford Bull. Econ. Statist.* 69:545–555.

Firth, D. (1992). Bias reduction, the Jeffreys prior and GLIM. In: Fahrmeir, L., Francis, B., Gilchrist, R., eds. *Advances in GLIM and Statistical Modelling*. New York: Springer, pp. 91–100.

Firth, D. A. V. I. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* 80:27–38.

Gail, M. H., Wieand, S., Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 71:431–444.

Griffiths, W. E., Hill, R. C., Pope, P. J. (1987). Small sample properties of probit model estimators. *J. Amer. Statist. Assoc.* 82:929–937.

Horowitz, J. L. (1993). Semiparametric and nonparametric estimation of quantal response models. In: Maddala, G. S., Rao, C. R., Vinod, H. D., eds. *Handbook of Statistics*. Amsterdam: Elsevier.

Lee, S., Lee, S. (2003). Testing heterogeneity for frailty distribution in shared frailty model. *Commun. Statist. Theor. Meth.* 32:2245–2253.

McCullagh, P. A., Nelder, J. A. (1989). *Generalized Linear Models*. Monographs on Statistics and Applied Probability. 2nd ed. London: Chapman Hall.

Mosley, W. H., Chen, L. C. (2003). An analytical framework for the study of child survival in developing countries. 1984. *Bull. World Health Organ.* 81:140–145.

Peters, B. L. (2004). Is there a wage bonus from drinking? Unobserved heterogeneity examined. *Appl. Econ.* 36:2299–2315.

Skinner, C. J., Holt, D., Smith, T. M. F. (eds.). (1989). *Analysis of Complex Surveys*. Chichester: John Wiley Sons, Ltd.

Steyerberg, E. W., Eijkemans, M. J. C. (2004). Heterogeneity bias: the difference between adjusted and unadjusted effects. *Med. Deci. Making* 24:102–104.

World Health Organization (2001). *International Classification of Functioning, Disability and Health: ICF* (Geneva ed.).

Zohoori, N., Savitz, D. A. (1997). Econometric approaches to epidemiologic data: Relating endogeneity and unobserved heterogeneity to confounding. *Ann. Epidemio.* 7:251–257.