International Statistical Review (1991), **58**, 2, pp. 227-240. Printed in Great Britain
© International Statistical Institute

Some Surprising Results About Covariate Adjustment in Logistic Regression Models

Laurence D. Robinson and Nicholas P. Jewell

Program in Biostatistics and Department of Statistics, University of California, Berkeley, CA 94720, USA

Summary

Results from classic linear regression regarding the effect of adjusting for covariates upon the precision of an estimator of exposure effect are often assumed to apply more generally to other types of regression models. In this paper we show that such an assumption is not justified in the case of logistic regression, where the effect of adjusting for covariates upon precision is quite different. For example, in classic linear regression the adjustment for a non-confounding predictive covariate results in improved precision, whereas such adjustment in logistic regression results in a loss of precision. However, when testing for a treatment effect in randomized studies, it is always more efficient to adjust for predictive covariates when logistic models are used, and thus in this regard the behavior of logistic regression is the same as that of classic linear regression.

Key words: Adjustment for covariates; Asymptotic relative efficiency; Classic linear regression; Logistic regression; Omitted covariate; Precision.

1 Introduction

The ability of covariance adjustment to improve the precision of estimates is a long-standing idea in statistics that originated with R.A. Fisher (1932). In particular, in a randomized experiment, when the assumptions of 'classic' linear regression apply, adjustment for covariates that are associated with the response variable is not required to obtain a valid estimate of the treatment effect, but nonetheless is desirable, as it will improve the precision of the treatment effect estimate. This improvement in precision can be explained in terms of a reduction of residual variance, an intuitive notion so persuasive that it has become the conventional wisdom to assume that similar gains in precision will be achieved with respect to regression models other than the classic, such as logistic regression (Mantel & Haenszel, 1959; Mantel, 1989).

Recently, however, some authors have recognized that in some situations the conventional wisdom regarding covariate adjustment does not apply. Wickramaratne & Holford (1989) give a specific $2\times2\times2$ contingency table example (for which a logistic regression analysis is appropriate) in which the pooled and stratum specific (log) odds ratio estimates are equal, and where the variance of the pooled (log) odds ratio estimate is less than that of the stratified estimate. This point was also addressed by Breslow & Day (1987).

In this paper it will be proven that adjustment for covariates always leads to a loss (or at best no gain) of precision with respect to logistic regression models. Section 2 outlines the details of the classic linear regression model, which are then used for comparison with logistic model results obtained in later sections. In § 3 logistic regression models which parametrize the $2\times2\times2$ contingency table situation are introduced, and asymptotic variance formulae for the pooled and adjusted estimates of exposure effect are stated. In § 4 it is demonstrated that the variance of the pooled estimate is always less than or equal to the variance of the adjusted estimate, and this result is then extended to the more general case of several strata. In § 5 it is demonstrated that the result of § 4 also applies to common finite sample estimates of the asymptotic variances. In § 6 a simple argument, involving the symmetric nature of logistic regression, is given which demonstrates that the conventional wisdom cannot apply to logistic regression. In § 7 the effects of certain key factors which influence precision are examined. In § 8, it is shown that it is always as or more efficient to adjust for covariates when testing for the presence of a treatment effect in randomized studies, in the context of a logistic regression model, despite the associated loss of precision demonstrated in § 4.

2 The 'Conventional (Classic Linear Regression) Wisdom'

Suppose the following two classic linear regression models provide valid descriptions of the structure of a population:

$$E(Y \mid X_1) = a^* + b_1^* X_1, \quad \text{var}\,(Y \mid X_1) = \sigma^2_{Y.1} \quad \text{(a constant);} \tag{1}$$

$$E(Y \mid X_1, X_2) = a + b_1 X_1 + b_2 X_2, \quad \text{var}\,(Y \mid X_1, X_2) = \sigma^2_{Y.12} \quad \text{(a constant).} \tag{2}$$

Suppose now that a large simple random sample is obtained from the population, and both models are fitted to the data via the method of least squares, resulting in estimators $\hat{b}_1^*$ and $\hat{b}_1$ (of b_1^* and b_1, respectively). We will denote the asymptotic relative precision of the estimator $\hat{b}_1$ (relative) to the estimator $\hat{b}_1^*$ by ARP ($\hat{b}_1$ to $\hat{b}_1^*$), and define it as follows:

$$\text{ARP}\,(\hat{b}_1 \text{ to } \hat{b}_1^*) = \frac{[\text{var}\,(\hat{b}_1)]^{-1}}{[\text{var}\,(\hat{b}_1^*)]^{-1}} = \frac{\text{var}\,(\hat{b}_1^*)}{\text{var}\,(\hat{b}_1)}.$$

Thus, our measure of the precision of an estimator is the inverse of its asymptotic variance.

For the estimators $\hat{b}_1^*$ and $\hat{b}_1$ associated with models (1) and (2), the following formula for ARP ($\hat{b}_1$ to $\hat{b}_1^*$) can be obtained:

$$\text{ARP}\,(\hat{b}_1 \text{ to } \hat{b}_1^*) = \frac{1 - \rho_{12}^2}{1 - \rho_{Y2.1}^2}.$$

Here ρ_{12} is the simple correlation between the variables X_1 and X_2, and $\rho_{Y2.1}$ is the partial correlation between the variables Y and X_2 conditional on fixed X_1.

The comparison of asymptotic variances is of particular interest when there is no confounding, i.e. when $b_1^* = b_1$, and hence both estimators $\hat{b}_1^*$ and $\hat{b}_1$ are estimating the same unknown population parameter. For the classic linear regression models described above, there will be no confounding if one or both of the following two conditions holds.

Condition 1. $\rho_{12} = 0$.

Condition 2. $\rho_{Y2.1} = 0$ (note this is equivalent to $b_2 = 0$).

When Condition 1 alone holds, ARP ($\hat{b}_1$ to $\hat{b}_1^*$) $= (1 - \rho_{Y2.1}^2)^{-1} > 1$. It is this result that explains the desirability of adjusting for a predictive covariate in randomized studies, even though a valid estimate of treatment effect can be obtained without adjustment. When Condition 2 alone holds, ARP ($\hat{b}_1$ to $\hat{b}_1^*$) $= 1 - \rho_{12}^2 < 1$. This result explains why it is undesirable to adjust for a non-predictive covariate which is correlated with the risk factor of interest. When both Conditions 1 and 2 hold, ARP ($\hat{b}_1$ to $\hat{b}_1^*$) $= 1$.

We now consider the behavior of the ARP ($\hat{b}_1$ to $\hat{b}_1^*$) more generally, not restricting ourselves to conditions of no confounding. This provides important insight into key factors which influence the precision of the estimator $\hat{b}_1$. In general, the value of ARP ($\hat{b}_1$ to $\hat{b}_1^*$) is seen to be (i) less than, (ii) equal to, or (iii) greater than 1 depending on whether $\rho_{Y2.1}^2$ is (i) less than, (ii) equal to, or (iii) greater than ρ_{12}^2. Thus a strong association between Y and X_2 has a beneficial effect upon the precision of $\hat{b}_1$, whereas a strong association between X_1 and X_2 has a detrimental effect, and hence the precision of $\hat{b}_1$ reflects the competing effects of these $Y-X_2$ and X_1-X_2 relationships.

It is the above behavior of the ARP ($\hat{b}_1$ to $\hat{b}_1^*$), and more generally of the precision of $\hat{b}_1$, that we loosely refer to as the conventional wisdom. The purpose of this paper is to demonstrate that the conventional wisdom breaks down with respect to the logistic regression model.

3 The Logistic Regression Model

Let Y, X_1, and X_2 each be a dichotomous variable taking on the values 0 and 1. The variable Y will be considered the response variable, and the variables X_1 and X_2 potential risk factors. The variable X_1 will be considered the risk factor of primary interest. Individuals for whom the value of X_1 equals 1 will be referred to as 'exposed', and those for whom the value of X_1 equals 0 as 'unexposed'. Individuals for whom the value of Y equals 1 will be referred to as 'diseased', and those for whom the value of Y equals 0 as 'non-diseased'. Of course, all of the following results and comments apply whatever the variables refer to in specific applications. Let us now assume the following two logistic regression models both provide a valid description of the population structure:

$$\log\left[\frac{\operatorname{pr}(Y=1 \mid X_1)}{1-\operatorname{pr}(Y=1 \mid X_1)}\right]=a^*+b_1^*X_1, \tag{3}$$

$$\log\left[\frac{\operatorname{pr}(Y=1 \mid X_1, X_2)}{1-\operatorname{pr}(Y=1 \mid X_1, X_2)}\right]=a+b_1X_1+b_2X_2. \tag{4}$$

Model (3) always provides a valid description of the relationship between the dichotomous variables Y and X_1, whereas model (4) imposes an assumption of no interaction (i.e. the variables X_1 and X_2 are assumed to have additive effects with respect to the log odds).

Suppose now that simple random samples of N_1 exposed and N_0 unexposed individuals are obtained, and that both logistic regression models are fit via the method of maximum likelihood, resulting in respective estimators $\hat{b}_1^*$ and $\hat{b}_1$. Standard likelihood theory techniques result in the following asymptotic variance formulae (Gart, 1962):

$$\operatorname{var}(\hat{b}_1^* \mid X_1)=\frac{1}{N_1p_1}+\frac{1}{N_1q_1}+\frac{1}{N_0p_0}+\frac{1}{N_0q_0}, \tag{5}$$

where $p_i=\operatorname{pr}(Y=1 \mid X_1=i)$ for $i=0, 1$ and $q_i=1-p_i$;

$$\operatorname{var}(\hat{b}_1 \mid X_1, X_2)=\left\{\left[\frac{1}{N_{10}p_{10}}+\frac{1}{N_{10}q_{10}}+\frac{1}{N_{00}p_{00}}+\frac{1}{N_{00}q_{00}}\right]^{-1}\right.$$
$$\left.+\left[\frac{1}{N_{11}p_{11}}+\frac{1}{N_{11}q_{11}}+\frac{1}{N_{01}p_{01}}+\frac{1}{N_{01}q_{01}}\right]^{-1}\right\}^{-1}. \tag{6}$$

Here $p_{ij}=\operatorname{pr}(Y=1 \mid X_1=i, X_2=j)$ for $i, j=0, 1$, $q_{ij}=1-p_{ij}$, and N_{ij} equals the number of individuals sampled for whom $X_1=i$ and $X_2=j$. Note that $N_{11}+N_{10}=N_1$ and

$N_{01} + N_{00} = N_0$. Also, here and throughout, the term asymptotic refers to both N_0 and N_1 tending to infinity.

The second variance formula given above is conditional on both X_1 and X_2. However, in accordance with the sampling scheme, in which prespecified numbers of exposed and unexposed individuals are sampled, but where the distribution of X_2 is allowed to vary, for the purpose of defining the asymptotic relative efficiency we shall require the variance conditional only on X_1, that is $\operatorname{var}(\hat{b}_1 \mid X_1)$, and thus must take the expectation of $\operatorname{var}(\hat{b}_1 \mid X_1, X_2)$ with respect to X_2. This results in the following formula:

$$\operatorname{var}(\hat{b}_1 \mid X_1) = \left\{ \left[\frac{1}{N_1 c_{11} p_{11}} + \frac{1}{N_1 c_{11} q_{11}} + \frac{1}{N_0 c_{01} p_{01}} + \frac{1}{N_0 c_{01} q_{01}} \right]^{-1} + \left[\frac{1}{N_1 c_{10} p_{10}} + \frac{1}{N_1 c_{10} q_{10}} + \frac{1}{N_0 c_{00} p_{00}} + \frac{1}{N_0 c_{00} q_{00}} \right]^{-1} \right\}^{-1}. \qquad (7)$$

In the above formula $c_{ij} = \operatorname{pr}(X_2 = j \mid X_1 = i)$ for $i, j = 0, 1$.

Table 1 gives the set of tables which represent the outcomes expected to result from the sampling scheme described above. In all tables in this paper D will denote 'diseased', D̄ 'non-diseased', E 'exposed', and Ē 'unexposed'. To avoid technical difficulties, we will assume that the population contains no 'structural zeroes' (McCullagh & Nelder, 1983, p. 61), so that none of the expected cell entries are 0.

The entries in the pooled table equal the sum of the corresponding entries in the two sub-tables, for example

$$N_1 c_{10} p_{10} + N_1 c_{11} p_{11} = N_1[c_{10} p_{10} + c_{11} p_{11}] = N_1 p_1.$$

Furthermore, we see that $\operatorname{var}(\hat{b}_1^* \mid X_1)$ simply equals the sum of the inverses of the expected cell entries of the pooled table, and that $\operatorname{var}(\hat{b}_1 \mid X_1)$ can be expressed as $[V_1^{-1} + V_0^{-1}]^{-1}$, where V_j equals the sum of the inverses of the expected cell entries of the sub-table $X_2 = j$, for $j = 0, 1$.

The estimator $\hat{b}_1$ referred to above is the maximum likelihood estimator. Another estimator commonly used is the 'inverse variance weighted, stratified estimator' (Weinberg, 1985), in which the parameter b_1 is estimated separately from the two sub-tables (using observed proportions), and then a weighted average of the two estimates is obtained, the weights being inversely proportional to their respective estimated variances. This estimator, also referred to as the 'Woolf estimator' (Woolf, 1955), can easily be computed by hand, whereas the maximum likelihood estimator generally requires an iterative scheme, and hence the use of a computer. It can be shown that the variance formulae given for the maximum likelihood estimator $\hat{b}_1$ also apply to the Woolf estimator (Gart, 1962).

4 The Asymptotic Relative Precision of $\hat{b}_1$ versus $\hat{b}_1^*$

As in § 2, we will denote the asymptotic relative precision of the estimator $\hat{b}_1$ (relative) to the estimator $\hat{b}_1^*$ by ARP ($\hat{b}_1$ to $\hat{b}_1^*$), and define it in terms of the ratio of the inverse of

Table 1

Expected cell frequencies for pooled and sub-tables from cohort design.

Pooled table				Sub-table $X_2 = 0$				Sub-table $X_2 = 1$			
	D	D̄	Total		D	D̄	Total		D	D̄	Total
E	$N_1 p_1$	$N_1 q_1$	N_1	E	$N_1 c_{10} p_{10}$	$N_1 c_{10} q_{10}$	$N_1 c_{10}$	E	$N_1 c_{11} p_{11}$	$N_1 c_{11} q_{11}$	$N_1 c_{11}$
Ē	$N_0 p_0$	$N_0 q_0$	N_0	Ē	$N_0 c_{00} p_{00}$	$N_0 c_{00} q_{00}$	$N_0 c_{00}$	Ē	$N_0 c_{01} p_{01}$	$N_0 c_{01} q_{01}$	$N_0 c_{01}$

asymptotic variances:

$$\text{ARP}\,(\hat{b}_1 \text{ to } \hat{b}_1^*) = \frac{[\text{var}\,(\hat{b}_1 \mid X_1)]^{-1}}{[\text{var}\,(\hat{b}_1^* \mid X_1)]^{-1}} = \frac{\text{var}\,(\hat{b}_1^* \mid X_1)}{\text{var}\,(\hat{b}_1 \mid X_1)}.$$

Here the asymptotic variances are those stated in § 3. Note again that both of these asymptotic variances are conditional on X_1, in accordance with the sampling scheme.

The main result of this paper is that ARP $(\hat{b}_1$ to $\hat{b}_1^*) \leqslant 1$, with equality occurring if and only if the variable X_2 is independent of (Y, X_1). Since ARP $(\hat{b}_1$ to $\hat{b}_1^*) \leqslant 1$ is equivalent to

$$[\text{var}\,(\hat{b}_1^* \mid X_1)]^{-1} \geqslant [\text{var}\,(\hat{b}_1 \mid X_1)]^{-1},$$

we must show that

$$\left[\frac{1}{N_1 p_1} + \frac{1}{N_1 q_1} + \frac{1}{N_0 p_0} + \frac{1}{N_0 q_0}\right]^{-1} \geqslant \left[\frac{1}{N_1 c_{10} p_{10}} + \frac{1}{N_1 c_{10} q_{10}} + \frac{1}{N_0 c_{00} p_{00}} + \frac{1}{N_0 c_{00} q_{00}}\right]^{-1} + \left[\frac{1}{N_1 c_{11} p_{11}} + \frac{1}{N_1 c_{11} q_{11}} + \frac{1}{N_0 c_{01} p_{01}} + \frac{1}{N_0 c_{01} q_{01}}\right]^{-1}.$$

This result follows readily as an application of Minkowski's inequality (Hardy, Littlewood & Polya, 1952, pp. 30–31), which for our purposes may be stated as follows: assume all a_{ij} positive, for $i = 1, \ldots, I$ and $j = 0, \ldots, J-1$. For finite $r < 1$, but not equal to 0, we have the following:

$$\left\{\sum_{i=1}^{I}\left[\sum_{j=0}^{J-1} a_{ij}\right]^r\right\}^{1/r} \geqslant \sum_{j=0}^{J-1}\left[\sum_{i=1}^{I} a_{ij}^r\right]^{1/r},$$

with equality occurring if and only if $a_{ij} = k\, a_{ij'}$ for all i and all choices of $j \neq j'$ and for some finite $k > 0$, where the value of k depends on the specific choice of j and j'.

For our particular application, we restate the above theorem for the specific case of $I = 4$, $J = 2$, and $r = -1$. This yields

$$\left[\frac{1}{a_{10} + a_{11}} + \frac{1}{a_{20} + a_{21}} + \frac{1}{a_{30} + a_{31}} + \frac{1}{a_{40} + a_{41}}\right]^{-1} \geqslant \left[\frac{1}{a_{10}} + \frac{1}{a_{20}} + \frac{1}{a_{30}} + \frac{1}{a_{40}}\right]^{-1} + \left[\frac{1}{a_{11}} + \frac{1}{a_{21}} + \frac{1}{a_{31}} + \frac{1}{a_{41}}\right]^{-1},$$

with equality occurring if and only if $a_{i1} = k\, a_{i0}$ for $i = 1, \ldots, 4$ and for some finite $k > 0$.

From this we can immediately conclude that

$$[\text{var}\,(\hat{b}_1^* \mid X_1)]^{-1} \geqslant [\text{var}\,(\hat{b}_1 \mid X_1)]^{-1},$$

and hence that ARP $(\hat{b}_1$ to $\hat{b}_1^*) \leqslant 1$, as an application of Minkowski's inequality, where

$$a_{10} = N_1 c_{10} p_{10}, \quad a_{11} = N_1 c_{11} p_{11}, \quad a_{20} = N_1 c_{10} q_{10}, \quad a_{21} = N_1 c_{11} q_{11},$$
$$a_{30} = N_0 c_{00} p_{00}, \quad a_{31} = N_0 c_{01} p_{01}, \quad a_{40} = N_0 c_{00} q_{00}, \quad a_{41} = N_0 c_{01} q_{01}.$$

In the above statement of Minkowski's inequality it is stated that ARP $(\hat{b}_1$ to $\hat{b}_1^*) = 1$ if and only if $a_{i1} = k\, a_{i0}$ for $i = 1, \ldots, 4$, a condition referred to as 'proportionality' by Hardy et al (1952). It can be shown (Bishop, Fienberg & Holland, 1975, p. 47) that, for our application, such proportionality is equivalent to the variable X_2 being independent of (Y, X_1).

We now examine the behavior of the ARP $(\hat{b}_1$ to $\hat{b}_1^*)$ for logistic regression, and compare it with the behavior seen for classic linear regression in § 2. As with regard to classic linear regression, of particular interest are those situations where there is no confounding.

For the logistic regression models (3) and (4) stated in § 3 there will be no confounding, that is $b_1^* = b_1$, if one or both of the following two conditions holds (Gail, 1986).

Condition 1′. X_1 and X_2 are independent given Y.

Condition 2′. Y and X_2 are independent given X_1 (note this is equivalent to $b_2 = 0$).

Condition 2′ is very much analogous to the no confounding Condition 2 $\rho_{Y2.1} = 0$ of § 2. In particular, for classic linear regression the condition 'Y and X_2 independent given X_1' does in fact imply that $\rho_{Y2.1} = 0$. When Condition 2′ alone holds, ARP $(\hat{b}_1$ to $\hat{b}_1^*) < 1$, which is the same result as was obtained for classic linear regression when the analogous no confounding Condition 2 holds. Thus we see that, for both logistic and classic linear regression, adjustment for a non-predictive covariate X_2 which is associated with the predictor variable X_1 results in a loss of precision.

Condition 1′ may also be regarded as analogous to the no confounding Condition 1 $\rho_{12} = 0$ of § 2, in that both conditions refer to a lack of an association between the variables X_1 and X_2. However, for logistic regression the absence of association is conditional on Y, which is not the case for classic linear regression. In particular, it is not Condition 1′ which implies Condition 1 $\rho_{12} = 0$ with respect to classic linear regression, but rather the condition 'X_1 and X_2 independent'. When Condition 1′ alone holds, ARP $(\hat{b}_1$ to $\hat{b}_1^*) < 1$, which is not consistent with the analogous result from classic linear regression, where we saw that ARP $(\hat{b}_1$ to $\hat{b}_1^*) > 1$ when the no confounding Condition 1 holds. Thus, whereas adjusting for a non-confounding covariate X_2 which is associated with the dependent variable Y (conditional on X_1) results in a gain in precision with respect to classic linear regression, it results in a loss of precision with respect to logistic regression.

When both Conditions 1′ and 2′ hold, the variable X_2 is independent of (Y, X_1), and thus ARP $(\hat{b}_1$ to $\hat{b}_1^*) = 1$. Furthermore, note that when X_2 is independent of (Y, X_1), with respect to classic linear regression, both $\rho_{12} = 0$ and $\rho_{Y2.1} = 0$. Thus, for both logistic and classic linear regression, ARP $(\hat{b}_1$ to $\hat{b}_1^*) = 1$ when the variable X_2 is independent of (Y, X_1).

More generally, for classic linear regression we saw that the value of ARP $(\hat{b}_1$ to $\hat{b}_1^*)$ can be less than, equal to, or greater than 1 depending on the relative strengths of the $Y - X_2$ and $X_1 - X_2$ relationships, whereas for logistic regression the value of ARP $(\hat{b}_1$ to $\hat{b}_1^*)$ is always less than or equal to 1 (again with equality occurring if and only if X_2 is independent of (Y, X_1)). This suggests that, unlike classic linear regression, where the $Y - X_2$ and $X_1 - X_2$ relationships have opposing effects which compete with each other to determine the relative precision of $\hat{b}_1$, with respect to logistic regression these two relationships have similar effects which combine to cause an automatic loss of precision. Sections 6 and 7 give additional insight into the behavior of the ARP $(\hat{b}_1$ to $\hat{b}_1^*)$ with respect to logistic regression.

The result we have obtained for the case of two strata, i.e. two levels of the variable X_2, can be extended to the more general case of $J \geqslant 2$ strata in a straightforward manner. In particular, the asymptotic variance of the maximum likelihood estimator of b_1 can be shown to equal (Gart, 1962)

$$\operatorname{var}(\hat{b}_1 \mid X_1) = \left\{ \sum_{j=0}^{J-1} \left[\frac{1}{N_1 c_{1j} p_{1j}} + \frac{1}{N_1 c_{1j} q_{1j}} + \frac{1}{N_0 c_{0j} p_{0j}} + \frac{1}{N_0 c_{0j} q_{0j}} \right]^{-1} \right\}^{-1}$$

where

$$c_{ij} = \operatorname{pr}(X_2 = j \mid X_1 = i), \quad p_{ij} = \operatorname{pr}(Y = 1 \mid X_1 = i, X_2 = j), \quad (i = 0, 1; j = 0, \ldots, J-1).$$

This asymptotic variance, which also pertains to the Woolf estimator of b_1, is a simple extension of the formula given previously for the $J = 2$ strata case. The desired result,

$[\text{var}(\hat{b}_1^* \mid X_1)]^{-1} \geq [\text{var}(\hat{b}_1 \mid X_1)]^{-1}$, then follows as an application of Minkowski's inequality with $I = 4$, $r = -1$, and $J =$ the number of strata, in a manner completely analogous with the two strata case. This also allows extension of the asymptotic relative precision result to the case of adjustment for a set of discrete covariates.

5 The Relationship Between Estimated Variances

In § 4 it was proven that $\text{var}(\hat{b}_1^* \mid X_1) \leq \text{var}(\hat{b}_1 \mid X_1)$, a result which pertains to the asymptotic variances. Suppose now that an actual set of data is obtained, and from that data set estimates of b_1^* and b_1 are computed. In this section we will consider both the maximum likelihood estimator and the Woolf estimator of b_1. Typically an investigator will also obtain estimates of the variances $\text{var}(\hat{b}_1^* \mid X_1)$ and $\text{var}(\hat{b}_1 \mid X_1)$. Although the maximum likelihood estimator and the Woolf estimator of b_1 have the same asymptotic variance, the method by which this variance is estimated is generally different. In this section we will examine the question of whether the result of § 4, which pertains to asymptotic variances, extends to their common finite sample estimates.

The estimation of $\text{var}(\hat{b}_1^* \mid X_1)$ given by (5) is very straightforward. There is only one commonly used method for estimating $\text{var}(\hat{b}_1^* \mid X_1)$, namely to substitute the maximum likelihood estimates $\hat{p}_1$ and $\hat{p}_0$ for p_1 and p_0, respectively. These maximum likelihood estimates are the observed proportions of diseased individuals (that is $Y = 1$) among exposed (that is $X_1 = 1$) and unexposed (that is $X_1 = 0$) individuals. Suppose now that the data is as given in Table 2. From this data set we obtain the estimated variance of $\hat{b}_1^*$, denoted by $\widehat{\text{var}}(\hat{b}_1^*)$, as

$$\widehat{\text{var}}(\hat{b}_1^*) = \frac{1}{a_{10} + a_{11}} + \frac{1}{a_{20} + a_{21}} + \frac{1}{a_{30} + a_{31}} + \frac{1}{a_{40} + a_{41}}.$$

Let us now examine the issue of estimating the variance of $\hat{b}_1$. Usually, we further condition on X_2 in calculating an estimated variance of $\hat{b}_1$. Therefore, regardless of whether the maximum likelihood estimator or the Woolf estimator has been used for $\hat{b}_1$, the estimate of $\text{var}(\hat{b}_1 \mid X_1, X_2)$, which we shall denote by $\widehat{\text{var}}(\hat{b}_1)$, is obtained by substituting estimates $\hat{p}_{ij}$ (for the unknown probabilities p_{ij}, for $i, j = 0, 1$) into the asymptotic variance formula (6).

When the method of maximum likelihood is used to obtain an estimate $\hat{b}_1$, typically the estimates $\hat{p}_{ij}$ which are substituted into the asymptotic variance formula are the maximum likelihood estimates based on the regression model. When the Woolf estimator is used, typically the estimates $\hat{p}_{ij}$ which are substituted are the observed proportions, which, in this case, are not the same as the maximum likelihood estimates. Note, however, that it is actually valid to substitute either set of estimates $\hat{p}_{ij}$ into the asymptotic variance formula regardless of which estimator $\hat{b}_1$ has been obtained (although the maximum likelihood estimates $\hat{p}_{ij}$ will generally not be available when the Woolf estimator has been obtained).

First consider the $\widehat{\text{var}}(\hat{b}_1)$ obtained by substituting the observed proportions in (6);

Table 2

Data for pooled and sub-tables arising from cohort studies.

Pooled table			Sub-table $X_2 = 0$			Sub-table $X_2 = 1$		
	D	$\bar{\text{D}}$		D	$\bar{\text{D}}$		D	$\bar{\text{D}}$
E	$a_{10} + a_{11}$	$a_{20} + a_{21}$	E	a_{10}	a_{20}	E	a_{11}	a_{21}
$\bar{\text{E}}$	$a_{30} + a_{31}$	$a_{40} + a_{41}$	$\bar{\text{E}}$	a_{30}	a_{40}	$\bar{\text{E}}$	a_{31}	a_{41}

then

$$\hat{\text{var}}(\hat{b}_1) = \left\{ \left[\frac{1}{a_{10}} + \frac{1}{a_{20}} + \frac{1}{a_{30}} + \frac{1}{a_{40}} \right]^{-1} + \left[\frac{1}{a_{11}} + \frac{1}{a_{21}} + \frac{1}{a_{31}} + \frac{1}{a_{41}} \right]^{-1} \right\}^{-1}.$$

Thus it follows immediately from Minkowski's inequality that when the observed proportions $\hat{p}_{ij}$ are used, $\hat{\text{var}}(\hat{b}_1^*) \leq \hat{\text{var}}(\hat{b}_1)$, and hence we see that, in this case, the relationship between the asymptotic variances extends to these estimated variances.

Now consider the case where the maximum likelihood estimates $\hat{p}_{ij}$ are used. It is a well known property of maximum likelihood estimation that the fitted sub-table cell frequencies must sum to the pooled table (Breslow & Day, 1980). Thus we have, for example, $N_{10}\hat{p}_{10} + N_{11}\hat{p}_{11}$ equalling the total number of diseased, exposed (that is $Y = 1, X_1 = 1$) individuals, which in the above data set is $a_{10} + a_{11}$. However, this total number equals $N_1\hat{p}_1$, and thus we have $N_{10}\hat{p}_{10} + N_{11}\hat{p}_{11} = N_1\hat{p}_1$. Similarly we have $N_{10}\hat{q}_{10} + N_{11}\hat{q}_{11} = N_1\hat{q}_1$, etc. Thus, Minkowski's inequality also applies when the maximum likelihood estimates $\hat{p}_{ij}$ are used, and once again we have the result $\hat{\text{var}}(\hat{b}_1^*) \leq \hat{\text{var}}(\hat{b}_1)$.

6 The Symmetric Nature of the Logistic Regression Model

As in § 3, we will assume that the logistic regression models (3) and (4) provide a valid description of the structure of a three dichotomous variable system. It is a well known property of logistic regression that when models (3) and (4) are valid, models (8) and (9) are also valid (Breslow & Powers, 1978):

$$\log\left[\frac{\text{pr}(X_1 = 1 \mid Y)}{1 - \text{pr}(X_1 = 1 \mid Y)}\right] = c^* + d_1^* Y, \tag{8}$$

$$\log\left[\frac{\text{pr}(X_1 = 1 \mid Y, X_2)}{1 - \text{pr}(X_1 = 1 \mid Y, X_2)}\right] = c + d_1 Y + d_2 X_2. \tag{9}$$

Another well known property of logistic regression is that $b_1^* = d_1^*$ and $b_1 = d_1$. Thus, we see that, for the purpose of estimating the parameters b_1^* and b_1, we can either treat Y as the response variable and X_1 as a predictor variable, or we can treat X_1 as the response variable and Y as the predictor variable.

Suppose now that the variables X_1 and X_2 are independent given Y, and that there exists a non-null association between X_2 and Y given X_1 (and thus $b_2 \neq 0$). Considering models (3) and (4), in which Y is the response variable and X_1 a predictor variable, from the conventional wisdom we would expect that $\text{var}(\hat{b}_1 \mid X_1) < \text{var}(\hat{b}_1^* \mid X_1)$, that is ARP $(\hat{b}_1$ to $\hat{b}_1^*) > 1$. Thus, we would expect adjustment for the covariate X_2 to result in an increase in precision.

But now suppose the variable X_1 is treated as the response variable and the variable Y as a predictor variable, as in models (8) and (9). In this situation, the covariate X_2 and predictor variable Y are not independent given response variable X_1, whereas X_2 is independent of the response variable X_1 given Y (and hence $d_2 = 0$). Now, from the conventional wisdom we would expect that $\text{var}(\hat{d}_1 \mid Y) > \text{var}(\hat{d}_1^* \mid Y)$, which implies ARP $(\hat{b}_1$ to $\hat{b}_1^*) < 1$. Thus, from this point of view, the conventional wisdom suggests that adjustment for the covariate X_2 would result in a loss of precision.

Because of the asymptotic equivalence of the estimators $\hat{b}_1^*$ and $\hat{d}_1^*$, and also of $\hat{b}_1$ and $\hat{d}_1$, we see that application of the conventional wisdom leads to a contradiction. Thus, from the 'symmetric nature' of the logistic regression model alone we can conclude that the conventional wisdom must break down with respect to logistic regression. Furthermore, the symmetric nature immediately suggests that the $Y-X_2$ and X_1-X_2 relationships have similar effects which combine to influence precision, in contrast to the situation observed with respect to classic linear regression, as was discussed previously in § 4.

The previous argument was stated specifically with respect to logistic regression models (3), (4), (8), and (9), for which the covariate X_2 is dichotomous, and where it was assumed that the variables X_1 and X_2 are independent given Y. However, the validity of the argument applies more generally to situations where there is confounding, and to situations involving adjustment for a set of covariates, some of which may be continuous. Thus, we strongly suspect that for logistic regression, when the risk factor of primary interest is dichotomous, adjustment for any set of covariates will result in a loss (or at best no gain) of precision.

7 The Effects of Key Factors which Influence Precision

In this section we look at the effects of certain key factors upon precision. First we examine how the strength of the $Y-X_2$ association, as measured by b_2, affects precision. Subsequently we consider the influence of the marginal distribution of Y.

The results of previous sections suggest that, for logistic regression, the stronger the association between the variables Y and X_2, conditional on X_1, that is the larger the magnitude of b_2, the poorer the precision of the estimator $\hat{b}_1$. Furthermore, we might suspect that as the magnitude of b_2 goes to infinity, the variance of $\hat{b}_1$ might also go to infinity. We will address these issues by examining the behavior of the ARP ($\hat{b}_1$ to $\hat{b}_1^*$) as the value of b_2 varies, while the values of other parameters are held fixed.

In a three dichotomous variable system there are $2^3-1=7$ parameters which are free to vary. Let us now assume that the variables X_1 and X_2 are independent given Y; that is we will focus our attention on a situation where there is no confounding. This assumption actually imposes two restrictions upon the three dichotomous variable system, i.e. independence of X_1 and X_2 at both levels of Y (also note that this assumption of conditional independence at both levels of Y implies that there is no interaction). Given these two restrictions, the three dichotomous variable system can now be parametrized by $7-2=5$ parameters. We will parametrize this system with the following five parameters:

$$p_1=\operatorname{pr}(Y=1\mid X_1=1),\quad p_0=\operatorname{pr}(Y=1\mid X_1=0),\quad \operatorname{pr}(X_1=1),$$
$$m=\operatorname{pr}(X_2=1\mid X_1=1,\,Y=1)=\operatorname{pr}(X_2=1\mid X_1=0,\,Y=1)=\operatorname{pr}(X_2=1\mid Y=1),$$
$$k=\operatorname{pr}(X_2=1\mid X_1=1,\,Y=0)=\operatorname{pr}(X_2=1\mid X_1=0,\,Y=0)=\operatorname{pr}(X_2=1\mid Y=0).$$

The first three probabilities in the above list parametrize the pooled table (note that in a 'cohort' sampling scheme in which N_1 exposed and N_0 unexposed individuals are sampled, the parameter $\operatorname{pr}(X_1=1)$ is fixed at $N_1/(N_1+N_0)$ by the investigator). The last two parameters, m and k, determine how the pooled table gets distributed into the two sub-tables (corresponding to $X_2=0$ and $X_2=1$). Given this parametrization, we have

$$b_2=\log\left[\frac{m/(1-m)}{k/(1-k)}\right].$$

Suppose now that we vary the parameter m while holding the remaining four parameters fixed. In fixing the first three parameters, we have fixed the pooled table, and hence $\operatorname{var}(\hat{b}_1^*\mid X_1)$ also. Since b_2 is a function of m, it varies with m. Since the distribution of the pooled table into the sub-tables varies with m, we see that $\operatorname{var}(\hat{b}_1\mid X_1)$, and hence ARP ($\hat{b}_1$ to $\hat{b}_1^*$), also vary as m varies. Consider now the

Table 3

Population probabilities that form the basis of Fig. 1.

Pooled table				Sub-table $X_2 = 0$			Sub-table $X_2 = 1$		
	D	D̄	Total		D	D̄		D	D̄
E	0·250	0·250	0·500	E	$0{\cdot}250(1-m)$	0·150	E	$0{\cdot}250m$	0·100
Ē	0·125	0·375	0·500	Ē	$0{\cdot}125(1-m)$	0·225	Ē	$0{\cdot}125m$	0·150

population probabilities summarized in Table 3. In Table 3 the four parameters have been fixed as follows: $p_1 = 0{\cdot}50$, $p_0 = 0{\cdot}25$, $\mathrm{pr}\,(X_1 = 1) = 0{\cdot}50$, $k = 0{\cdot}40$. Suppose that we also arbitrarily fix the total sample size at 200. From this information we may immediately obtain $\mathrm{var}\,(\hat{b}_1^* \mid X_1) = 0{\cdot}09333$. By varying the parameter m, a graph of ARP ($\hat{b}_1$ to $\hat{b}_1^*$) versus b_2 is obtained, shown in Fig. 1.

From Fig. 1 we see that ARP ($\hat{b}_1$ to $\hat{b}_1^*$) achieves a maximum value of 1 at $b_2 = 0$, and that it decreases monotonically as b_2 moves away from 0 in either direction. Furthermore, it is clear from this graph that ARP ($\hat{b}_1$ to $\hat{b}_1^*$) reaches asymptotes as b_2 goes to plus and minus infinity. The values of these asymptotes can be easily computed. As $m \rightarrow 1$, $b_2 \rightarrow \infty$, and the $X_2 = 0$ sub-table provides increasingly less information regarding the parameter b_1. Thus $\mathrm{var}\,(\hat{b}_1 \mid X_1)$ can be computed solely from the $X_2 = 1$ sub-table, where m has been set equal to 1. This results in

$$\mathrm{var}\,(\hat{b}_1 \mid X_1) = 50^{-1} + 20^{-1} + 25^{-1} + 30^{-1} = 0{\cdot}14333.$$

From this we obtain the asymptote as $b_2 \rightarrow \infty$ as $0{\cdot}09333/0{\cdot}14333 = 0{\cdot}65116$. In a similar manner we obtain the asymptote as $b_2 \rightarrow -\infty$ as $0{\cdot}09333/0{\cdot}11556 = 0{\cdot}80769$.

In this example we see that the loss of precision induced by adjustment for X_2 increases with the magnitude of b_2 in both directions. However, the variance of $\hat{b}_1$ does not go off to infinity as b_2 goes to plus and minus infinity, but rather approaches asymptotes in both directions. This reflects the fact that the variance depends not only on b_2 but also on other factors, particularly the marginal distributions of X_2 and Y, and that these other factors have been fixed at levels (by our specific choices of values for the four fixed parameters) which limit the potential for loss of precision due to a strong $Y - X_2$ association.

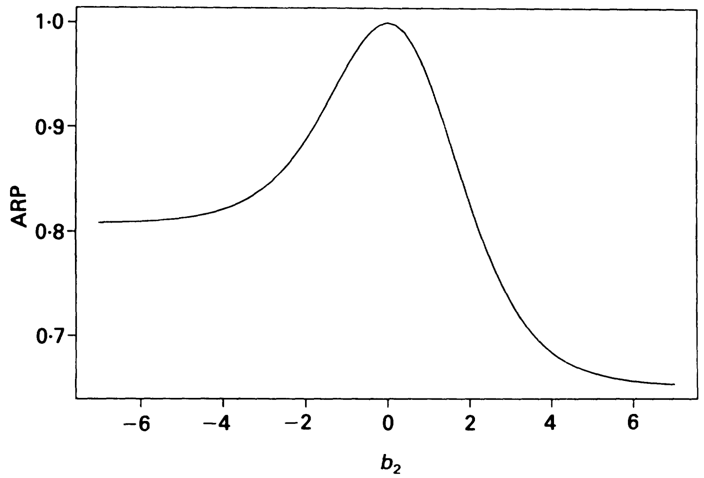

Figure 1. *Plot of asymptotic relative precision* (ARP) *of the adjusted estimator* $\hat{b}_1$ *to the pooled estimator* $\hat{b}_1^*$, *against* b_2, *holding all other parameters fixed at values shown in Table* 3.

Table 4

Population probabilities that form the basis of Fig. 2.

Pooled table				Sub-table $X_2 = 0$			Sub-table $X_2 = 1$	
	D	$\bar{D}$	Total		D	$\bar{D}$	D	$\bar{D}$
E	$0{\cdot}500p_1$	$0{\cdot}500(1-p_1)$	0·500	E	$0{\cdot}200p_1$	$0{\cdot}400(1-p_1)$	E $0{\cdot}300p_1$	$0{\cdot}100(1-p_1)$
$\bar{E}$	$0{\cdot}500p_0$	$0{\cdot}500(1-p_0)$	0·500	$\bar{E}$	$0{\cdot}200p_0$	$0{\cdot}400(1-p_0)$	$\bar{E}$ $0{\cdot}300p_0$	$0{\cdot}100(1-p_0)$

We now examine the effect of the marginal distribution of Y upon the asymptotic relative precision by varying $\text{pr}(Y = 1)$, while the values of other parameters are held fixed. Again we will assume that the variables X_1 and X_2 are independent given Y, so that the parametrization of the previous example also applies here. In this case we will fix the values of the parameters $\text{pr}(X_1 = 1)$, m, and k, and then vary both p_1 and p_0 in such a way as to hold

$$b_1^* = \log\left[\frac{p_1/(1-p_1)}{p_0/(1-p_0)}\right]$$

fixed. As p_1 and p_0 vary, so does the overall incidence of disease $\text{pr}(Y = 1)$. Table 4 gives an example of certain population probabilities.

In Table 4 we have set $\text{pr}(X_1 = 1) = 0{\cdot}50$, $k = 0{\cdot}20$, and $m = 0{\cdot}60$ (and thus $b_2 = \log 6$). Again we consider a total sample size of 200. We now vary both p_1 and p_0 so as to hold b_1^* fixed at log 3, to obtain a graph of ARP ($\hat{b}_1$ to $\hat{b}_1^*$) versus $\text{pr}(Y = 1)$, shown in Fig. 2.

From Fig. 2 we see that for both small and large values of $\text{pr}(Y = 1)$, the ARP ($\hat{b}_1$ to $\hat{b}_1^*$) is relatively close to the maximum value of 1, while for values of $\text{pr}(Y = 1)$ closer to 0·5 the ARP ($\hat{b}_1$ to $\hat{b}_1^*$) is further from 1. Thus, this particular example suggests that the potential for loss of precision due to adjustment for covariates will tend to be greater in cohort studies where the disease is relatively common. It must be noted, however, that the minimum ARP ($\hat{b}_1$ to $\hat{b}_1^*$) value does not occur exactly at $\text{pr}(Y = 1) = 0{\cdot}5$, and that in fact the minimum can occur at values of $\text{pr}(Y = 1)$ quite far from 0·5 when the marginal distribution of X_2 is very skewed. Nonethelesss, in most studies the marginal distribution

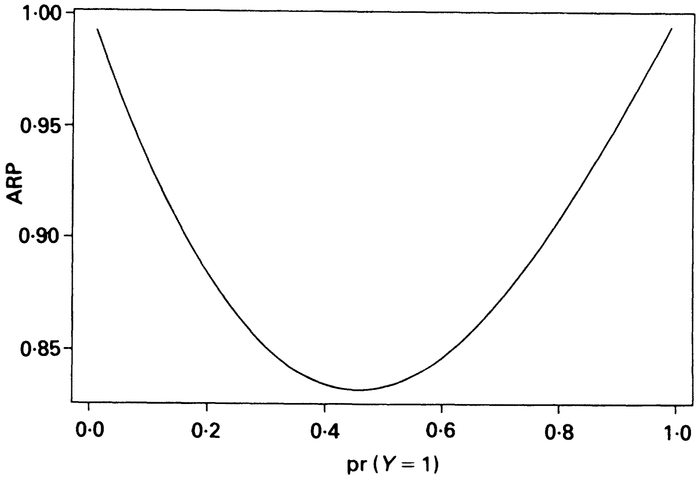

Figure 2. *Plot of asymptotic relative precision* (ARP) *of the adjusted estimator $\hat{b}_1$ to the pooled estimator $\hat{b}_1^*$, against the probability of disease,* $\text{pr}(Y = 1)$, *holding all other parameters fixed at values shown in Table* 4.

of X_2 will not be highly skewed, so that the conclusion reached above remains valid. The above example also suggests that the potential for loss of precision will tend to be particularly great for case-control studies, where the oversampling of cases ensures a relatively high frequency of disease in the sample.

8 Testing for No Treatment Effect in Randomized Studies

In this section we examine the use of the estimators $\hat{b}_1$ and $\hat{b}_1^*$ for testing the hypothesis of no treatment effect in randomized studies (or other situations where X_1 and X_2 are known to be independent), for both classic linear regression and logistic regression. As in previous sections, we will assume that the variables X_1 and X_2 are dichotomous. The variable X_1 will now indicate whether a particular individual received treatment or not. The random allocation of treatment and control to study subjects ensures that X_1 and X_2 are independent.

Let us now suppose that the response variable Y is such that the classic linear regression models (1) and (2) of § 2 are valid. From the results of § 2 we know that $b_1^* = b_1$ and that $\hat{b}_1$ is a more precise estimator of the treatment effect than is $\hat{b}_1^*$. It follows immediately that if we wish to test the null hypothesis of no treatment effect, which can be expressed as $H_0: b_1 = b_1^* = 0$, then test statistics based on the more precise estimator $\hat{b}_1$ will generally have greater power than tests based on $\hat{b}_1^*$.

Now suppose that the outcome variable Y is such that the logistic regression models (3) and (4) of § 3 are valid. Independence of X_1 and X_2 does not in this case ensure no confounding, so that generally b_1 will not equal b_1^*. However, when b_1 equals 0, b_1^* also equals 0, and thus the null hypothesis of no treatment effect can be expressed as $H_0: b_1 = b_1^* = 0$ in this case as well. Because for the logistic case $\hat{b}_1^*$ is a more precise estimator than $\hat{b}_1$, at first glance we might suspect that tests of $H_0: b_1 = b_1^* = 0$ based on $\hat{b}_1^*$ would give greater power. However, from the well known result that when X_1 and X_2 are independent, the value of the parameter b_1^* falls between 0 and b_1 (Gail, 1986), we also see that the point estimate of treatment effect $\hat{b}_1^*$ will tend to be smaller than the estimate $\hat{b}_1$. Hence, upon closer examination it is not clear which type of hypothesis test, that based on $\hat{b}_1^*$ or $\hat{b}_1$, will give greater power in testing H_0.

Pitman (Cox & Hinkley, 1974, p. 338) developed a general definition of the asymptotic relative efficiency of two hypothesis tests which may be applied to determine which of the two types of test statistics gives greater power. In particular, we have

$$\text{ARE}\,(\hat{b}_1 \text{ to } \hat{b}_1^* \text{ at } b_1 = 0) = \left[\lim_{b_1 \to 0} \left\{ \left(\frac{d}{db_1} b_1\right) \Big/ \left(\frac{d}{db_1} b_1^*\right) \right\}\right]^2 \left[\lim_{b_1 \to 0} \frac{\operatorname{var}(\hat{b}_1^* \mid X_1)}{\operatorname{var}(\hat{b}_1 \mid X_1)}\right].$$

Here we are considering the parameter b_1^* as a function of b_1. Notationally, let $E(A_j) = \Sigma \operatorname{pr}(X_2 = j) A_j$, that is expectation over the distribution of X_2. According to this notation, $p_1 = E(p_{1j})$ and $p_0 = E(p_{0j})$. Consequently, b_1^* can be expressed as

$$\log [E(p_{1j}) E(q_{0j}) / E(p_{0j}) E(q_{1j})]$$

and $\operatorname{var}(\hat{b}_1^* \mid X_1)$ can be expressed as

$$[N_1 E(p_{1j}) E(q_{1j})]^{-1} + [N_0 E(p_{0j}) E(q_{0j})]^{-1}.$$

Also, using the independence of X_1 and X_2, the formula for $\operatorname{var}(\hat{b}_1 \mid X_1)$ given by (7) can be expressed as

$$\operatorname{var}(\hat{b}_1 \mid X_1) = \left\{ E\left[\frac{1}{N_1 p_{1j} q_{1j}} + \frac{1}{N_0 p_{0j} q_{0j}}\right]^{-1} \right\}^{-1}.$$

Now, using the fact that

$$p_{ij} = [\exp(a + b_1 i + b_2 j)]/[1 + \exp(a + b_1 i + b_2 j)]$$

for $i, j = 0, 1$, we take the derivative of b_1^* with respect to b_1 to obtain

$$\frac{d}{db_1} b_1^* = \frac{E(p_{1j}q_{1j})}{E(p_{1j})E(q_{1j})}.$$

As $b_1 \rightarrow 0$, we also have $p_{10} \rightarrow p_{00}$, $p_{11} \rightarrow p_{01}$, and $p_1 \rightarrow p_0$, and thus

$$\lim_{b_1 \rightarrow 0} \frac{d}{db_1} b_1^* = \frac{E(p_{0j}q_{0j})}{E(p_{0j})E(q_{0j})}.$$

Similarly,

$$\lim_{b_1 \rightarrow 0} \frac{\mathrm{var}(\hat{b}_1^* \mid X_1)}{\mathrm{var}(\hat{b}_1 \mid X_1)} = \lim_{b_1 \rightarrow 0} \frac{[\mathrm{var}(\hat{b}_1 \mid X_1)]^{-1}}{[\mathrm{var}(\hat{b}_1^* \mid X_1)]^{-1}} = \frac{E(p_{0j}q_{0j})}{E(p_{0j})E(q_{0j})}.$$

Finally, we obtain the result

$$\text{ARE}\,(\hat{b}_1 \text{ to } \hat{b}_1^* \text{ at } b_1 = 0) = \left[\frac{E(p_{0j})E(q_{0j})}{E(p_{0j}q_{0j})}\right]^2 \left[\frac{E(p_{0j}q_{0j})}{E(p_{0j})E(q_{0j})}\right] = \frac{E(p_{0j})E(q_{0j})}{E(p_{0j}q_{0j})}.$$

We immediately conclude that ARE $(\hat{b}_1$ to $\hat{b}_1^*$ at $b_1 = 0) \geqslant 1$, with equality occurring if and only if X_2 is independent of (Y, X_1). Thus, to test the null hypothesis of no treatment effect in a randomized study, it is always as or more efficient to adjust for the covariate X_2 when logistic models are used. Thus, in this regard the logistic regression model behaves similarly to the classic linear regression model. This result is essentially a special case of a result of Gail, Tan & Piantadosi (1988), although these authors work with the score test rather than the asymptotically equivalent Wald test. We note that it is straightforward to extend the above derivation to allow for a discrete multivariate X_2.

9 Discussion

For classic linear regression models, the precision of the estimator $\hat{b}_1$ depends upon the relative strengths of the $Y - X_2$ and $X_1 - X_2$ associations. In particular, a strong $Y - X_2$ association has a beneficial effect upon precision, whereas a strong $X_1 - X_2$ association has a detrimental effect. It has been, heretofore, conventional wisdom to assume that the above behavior of classic linear regression with respect to precision applies more generally to other types of regression models. In this paper, however, we have shown that the behavior of logistic regression with respect to precision is quite different from that of classic linear regression. In particular, while a strong $X_1 - X_2$ association again has a detrimental effect upon precision for logistic regression, a strong $Y - X_2$ association also has a detrimental effect. Consequently, whereas in classic linear regression adjustment for predictive covariates can result in either increased or decreased precision, adjustment for predictive covariates will always result in a loss of precision for logistic regression. However, we have seen that for logistic regression, as for classic regression, adjustment for predictive covariates results in greater efficiency when testing for a treatment effect in randomized studies.

In any particular investigation, one may be interested in estimating b_1, b_1^*, or both. Given the behavior of logistic regression with respect to precision, when the parameter of interest is b_1 it seems plausible that in some situations it might be preferable to use the biased but more precise $\hat{b}_1^*$ to estimate b_1, rather than the unbiased but less precise $\hat{b}_1$, as the estimator $\hat{b}_1^*$ may result in greater accuracy, as measured by mean square error. Work is currently in progress to determine guidelines for when one might use $\hat{b}_1^*$ as an estimator of b_1. It should be noted that for large sample sizes the mean square error will be dominated by its bias rather than variance component, so that for sufficiently large samples the adjusted estimator will always be preferable.

Although we have largely focused on the cohort design, it is straightforward to show that the results and arguments extend to other standard designs, including cross-sectional studies. In further work we shall consider other common generalized linear models used with dichotomous dependent variables and standard regression models used in survival analysis applications.

Acknowledgements

We thank W.J. Redfearn for several illuminating discussions, and S. Selvin for showing us an early proof of a special case of the result of § 4, as well as numerous invaluable discussions. We are also indebted to P. Armitage for comments with regard to the material in § 8. Support for this research was provided in part by Grant A129162 from the National Institute of Allergy and Infectious Diseases. The manuscript was completed while the second author visited the Department of Statistics, University of Oxford, with support from a travel grant from the Burroughs Wellcome Fund.

References

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Mass: MIT Press.

Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research,* **1**: *The Analysis of Case-Control Studies.* Lyon, France: IARC Scientific Publications.

Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research,* **2**: *The Design and Analysis of Cohort Studies.* Lyon, France: IARC Scientific Publications.

Breslow, N.E. & Powers, W. (1978). Are there two logistic regressions for retrospective studies? *Biometrics* **34,** 100–105.

Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

Fisher, R.A. (1932). *Statistical Methods For Research Workers.* Edinburgh: Oliver and Boyd (13th ed., 1958).

Gail, M.H. (1986). Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In *Modern Statistical Methods in Chronic Disease Epidemiology,* Ed. S.H. Moolgavkar and R.L. Prentice, pp. 3–18. New York: Wiley.

Gail, M.H., Tan, W.Y. & Piantadosi, S. (1988). Tests for no treatment effects in randomized clinical trials. *Biometrika* **75,** 57–64.

Gart, J.J. (1962). On the combination of relative risks. *Biometrics* **18,** 601–610.

Hardy, G.H., Littlewood, J.E. & Polya, G. (1952). *Inequalities.* London: Cambridge University Press.

Mantel, N. (1989). Confounding in epidemiologic studies. *Biometrics* **45,** 1317–18.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.* **22,** 719–48.

McCullagh, P. & Nelder, J.A. (1983). *Generalized Linear Models.* London: Chapman and Hall.

Weinberg, C.R. (1985). On pooling across strata when frequency matching has been followed in a cohort study. *Biometrics* **41,** 117–27.

Wickramaratne, P.J. & Holford, T.R. (1989). Confounding in epidemiologic studies. Response. *Biometrics* **45,** 1319–22.

Woolf, B. (1955). On estimating the relationship between blood group and disease. *Ann. Human Genetics* **19,** 251–53.

Résumé

Les résultats de l'analyse de régression linéaire classique concernant l'effet d'ajustement pour des variables concomitantes sur la précision d'un estimateur d'exposition, sont souvent supposés s'appliquer de façon plus générale à d'autres types de modèles de régression. Dans cet article, on montre qu'une telle supposition n'est pas jutifiée dans le cas d'une régression logistique, où l'effet d'ajustement de variables concomitantes sur la précision est tout à fait different. Par exemple, en régression linéaire classique, l'ajustement pour une variable concomitante de prévision non confondante se traduit en une précision ameliorée. Par contre, le même ajustement en régression linéaire logistique, se traduit en une perte de précision. Quoiqu'il en soit, quand l'effet d'un traitement est testé dans une étude randomisée il est toujours plus efficace d'ajuster pour des variables concomitantes prévisionnelles quand un modèle logistique est utilisé et ainsi, le comportement en régression logistique est identique à celiu en régression linéaire classique.

[*Received May* 1990, *accepted October* 1990]