A Modified Least-Squares Regression Approach to the Estimation of Risk Difference

Yin Bun Cheung

From the MRC Tropical Epidemiology Group, Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom.

Received for publication December 11, 2006; accepted for publication July 6, 2007.

Risk ratio and risk difference are parameters of interest in many medical studies. The risk ratio has a property that the value for the outcome \( Y = 0 \) is not the inverse of the risk ratio for the outcome \( Y = 1 \). This property makes risk ratios inappropriate in some situations. Estimation of risk difference often encounters the problem that the binomial regression model fails to converge. Recently discussed alternatives may have the same problem of nonconvergence or are difficult to implement. Here the author proposes a modified least-squares regression approach—unweighted least-squares regression with a Huber-White robust standard error—for estimation of risk differences. Four versions of the robust standard error are considered. The binomial, ordinary least-squares, and modified least-squares estimators are compared analytically in a simple situation of one exposure variable. Multivariable regression analyses are simulated to demonstrate the usefulness of the approach. For sample sizes of approximately 200 or less, a small-sample version of the robust standard error is recommended. The method is illustrated with data from a patient survey in which the binomial regression fails to converge in the analyses of four out of five outcome variables.

binomial model; least-squares analysis; regression analysis; risk; robustness; variance-covariance matrix; variance heterogeneity

Abbreviations: ECOG, Eastern Cooperative Oncology Group; EQ-5D, EuroQol 5-Dimension; MLE, maximum likelihood estimation; MLS, modified least-squares; OLS, ordinary least-squares.
controlled trial comparing two therapies for malaria (3), a risk difference in clinical failure of less than 3 percent on the part of the investigational product was considered noninferiority. In the elicitation of clinicians’ beliefs about the survival chances of patients under different treatments, the risk difference is usually used for easy understanding (4). The risk ratio also has another interpretation problem in that the risk ratio for the outcome \( Y = 0 \) is not the inverse of the risk ratio for the outcome \( Y = 1 \) (5). Consider a hypothetical example of 250 failures and 750 successes in an unexposed group and 400 failures and 600 successes in an exposed group. Further, consider the equivalence margins \( 0.80 \) and \( 1/0.80 \) for the unexposed group and \( 400 \) failures and \( 600 \) successes in a hypothetical example of 250 failures and 750 successes in an exposed group. Further, consider the equivalence margins \( 0.80 \) and \( 1/0.80 \) for the unexposed group and 400 failures and 600 successes in an exposed group. Further, consider the equivalence margins \( 0.80 \) and \( 1/0.80 \) for the unexposed group and 400 failures and 600 successes in an exposed group. Further, consider the equivalence margins \( 0.80 \) and \( 1/0.80 \) for the unexposed group and 400 failures and 600 successes in an exposed group. Further, consider the equivalence margins \( 0.80 \) and \( 1/0.80 \) for the unexposed group and 400 failures and 600 successes in an exposed group.

Maximum likelihood estimation (MLE) of the log-binomial and binomial regression models often fails to converge (2, 5–8). It has been suggested that Poisson regression with a log link and Huber-White robust variance can be used to estimate risk ratios (8) and that Poisson regression with an identity link and robust variance can be used to estimate risk differences (2). However, the Poisson regression model with an identity link does not always converge either. Although it is possible to derive a standardized risk difference from the outputs of a logistic regression model, the procedure and (in particular) its variance calculation are complex, and quantitative covariates need to be categorized as strata (5, 9).

The ordinary least-squares (OLS) method is commonly used for the analysis of the (conditional) mean of a quantitative outcome. Note that the standard proof of the unbiasedness of OLS estimates does not require assumptions of normality or constant variance (10). These extra assumptions are only relevant to how statistical inference (significance test and confidence interval) is made. When the outcome variable is not normally distributed, the mean may or may not be a good measure of central tendency. However, in the case of a binary response variable coded as 0 or 1, the mean is equal to the proportion or risk. The OLS method can be used to analyze a binary response. Economometricians call this a “linear probability model” (11). The regression coefficients represent risk differences. Nevertheless, the error structure for a binary outcome is different from that assumed in OLS regression. Therefore, the statistical inference based on OLS is usually not valid unless one makes an assumption of a multivariate normal distribution in the exposure variables. This is often a valid assumption in economic studies and is the situation in which economists use the linear probability model (11). The most typical exposure variable in epidemiologic investigations is probably binary. The assumption of multivariate normality in the exposure variables is unlikely to be applicable in this context.

### TABLE 1. Notation for entries in a 2 × 2 table summarizing data for calculation of a risk difference

<table>
<thead>
<tr>
<th>( y = 0 ) (no event)</th>
<th>( y = 1 ) (event)</th>
<th>Risk</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x = 0 ) (unexposed)</td>
<td>( a )</td>
<td>( b )</td>
<td>( \rho_0 = b/(a + b) )</td>
</tr>
<tr>
<td>( x = 1 ) (exposed)</td>
<td>( c )</td>
<td>( d )</td>
<td>( \rho_1 = d/(c + d) )</td>
</tr>
</tbody>
</table>

In this article, I examine the estimation of risk difference by the OLS regression method accompanied by statistical inference based on the Huber-White robust estimate of variance (1, 12, 13). The widely used version of the robust variance may not be suitable if sample size is small (1, 13). Three small-sample versions are examined here. I analytically examine the estimators concerned in a simple situation of one binary exposure variable. I then show simulation results that evaluate the properties of the methods in more complex situations. A real example from a survey of cancer patients is presented. Following the terminology of Zou (8), I refer to this method as modified least-squares (MLS) regression.

### MLE, OLS, AND MLS

A simple case of a 2 × 2 table (table 1) is used for illustration, where \( a, b, c, \) and \( d \) represent cell frequencies. By application of the standard MLE method, it can be shown that the risk difference (RD) estimated by binomial regression is

\[
\text{RD}_{\text{MLE}} = \frac{d}{c + d} - \frac{b}{a + b} = \hat{\rho}_1 - \hat{\rho}_0, \tag{1}
\]

and that its variance, or the square of the standard error, is

\[
\text{Var}(\text{RD}_{\text{MLE}}) = \frac{\hat{\rho}_0(1 - \hat{\rho}_0)}{n_0} + \frac{\hat{\rho}_1(1 - \hat{\rho}_1)}{n_1}. \tag{2}
\]

The regression coefficient in OLS is estimated by the following formula, which yields the same estimate of RD:

\[
\text{RD}_{\text{OLS}} = \frac{\sum_{i=1}^{n} y_i x_i - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{d}{c + d} - \frac{b}{a + b} = \hat{\rho}_1 - \hat{\rho}_0. \tag{3}
\]

The variance of the OLS estimate is given by

\[
\text{Var}(\text{RD}_{\text{OLS}}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \left( \frac{n}{n - 2} \right) \left[ \frac{\hat{\rho}_0(1 - \hat{\rho}_0)}{n_1} + \frac{\hat{\rho}_1(1 - \hat{\rho}_1)}{n_0} \right], \tag{4}
\]

where \( \sigma^2 \) is the variance of the residuals. The residuals are calculated by subtracting the predicted probability of \( y = 1 \) from \( \hat{\mu}(x) \).
from the observed $y$. After some algebraic manipulation, it turns out that the OLS variance looks fairly similar to that of the MLE variance in equation 2. The difference is that, firstly, there is a multiplication factor of $n(n - k)$, where $n$ is the total sample size and $k = 2$ is the number of parameters in the regression model. In most epidemiologic studies, $n$ is large and this multiplicative factor is close to 1 and therefore inconsequential. The second difference is that, when equations 2 and 4 are compared, the denominators $n_0$ and $n_1$ are found to have switched places. With the $n(n - 2)$ factor ignored, the difference between the two variances can be expressed as

$$\text{Var}(\hat{\text{RD}}_{\text{MLE}}) - \text{Var}(\hat{\text{RD}}_{\text{OLS}}) = \frac{(n_1 - n_0)}{n_0 n_1} [\beta_0 (1 - \hat{p}_0) - \beta_1 (1 - \hat{p}_1)].$$

(5)

Several relations between the two variance estimates can be seen here. If $n_1 = n_0$ or $\hat{p}_1 = \hat{p}_0$, the variances based on OLS are identical to those based on binomial regression. The larger the risk difference or $n_1 - n_0$, the larger the difference in the variances. Furthermore, the difference can be positive or negative depending on which exposure group has a larger sample size and the relative magnitude of $\hat{p}_0$ and $\hat{p}_1$. Note that in a binomial distribution with parameter $p$, the quantity $p(1 - p)$ is maximized if $p = 0.5$. Therefore, if the group with $p_i$ further from 0.5 has a larger sample size, the variance of the OLS estimate is smaller than that of the binomial regression.

The Huber-White robust estimate of variance (robust variance) makes no distributional assumption. It is commonly given in matrix form,

$$\text{Var}(\hat{\beta}) = (X'X)^{-1} \sum_{i=1}^{n} e_i^2 x_i x_i' (X'X)^{-1},$$

(6)

where $e_i = y_i - x_i \hat{\beta}$, $x_i$ is the $i$th row of the design matrix $X$, and $\hat{\beta}$ is the vector of OLS estimates. A nontechnical exposition of the derivation of the robust variance in its matrix form is given in Appendix 1, which applies to single or multiple exposure variables and to continuous or binary exposure variables. In the special case of a single binary exposure variable, it can also be shown that this robust variance of the risk difference is identical to the variance estimated by the binomial regression model given in equation 2, that is, $\text{Var}(\text{RD}_{\text{MLE}}) = \text{Var}(\text{RD}_{\text{OLS}})$ (see details in Appendix 2). In sum, in the case of one binary exposure variable, the use of binomial regression and the use of MLES regression produce identical estimates of RD and its variance.

In the case of a continuous exposure variable, the binomial regression model attempts to find the intercept (IntMLE) and risk difference (RDMLE) that satisfy the following relation:

$$\sum_{i=1}^{n} \left[ \frac{y_i x_i}{\text{IntMLE} + \text{RD}_{\text{MLE}}} - \frac{1 - y_i x_i}{1 - \text{IntMLE} - \text{RD}_{\text{MLE}}} \right] = 0.$$

This has to be done by means of an iterative search procedure; there is no analytic solution. Indeed, there is no guarantee that the search will produce any solution at all.

If it does produce a solution, there is no guarantee that its results will exactly agree with those of the MLE in a particular finite sample. The standard errors cannot be obtained analytically either. I will examine situations involving continuous variables by simulations.

**Correction for small sample size**

Equation 6 gives the asymptotic, or large-sample, version of the robust variance. Neither this nor the binomial regression is reliable in small samples (1, 12, 13). Three relatively unknown small-sample versions of the robust variance have been proposed (see Long and Ervin (13) for a review). The literature calls them HC1, HC2, and HC3 and the large-sample version HC0. Briefly, the idea is that the $e_i^2$ in equation 6 tends to be underestimated in small samples, and therefore it is desirable to multiply it by a factor that is larger than 1 when sample size is small but converges to 1 as sample size increases. The correction factors used by HC1, HC2, and HC3 are, respectively, $n(n - k)/n$, $1/(1 - h_i)$, and $1/(1 - h_i)^2$, where $n$ and $k$ are as previously defined and $h_i$ is the $i$th element on the diagonal of the hat matrix $X(X'X)^{-1}X'$. The amount of inflation is smallest in HC1 and largest in HC3. The four versions converge when sample size is large.

**SIMULATIONS**

**Experimental conditions**

Multivariable regression analyses of risk in relation to a binary explanatory variable $X_1$ (1 = exposed and 0 = unexposed) and two continuous variables $X_2$ and $X_3$ were simulated. The probability of $x_1 = 1$ was either 0.5 or 0.7. The continuous variable $X_2$ followed a beta distribution with parameters ($\alpha = 5 + 3x_1$, $\beta = 5$) or ($\alpha = 5 + 5x_1$, $\beta = 5$). These beta distribution parameters make $X_2$ fairly symmetric, though not normal. The correlation between $X_1$ and $X_2$ was approximately 0.37 in the scenario with $\alpha = 5 + 3x_1$ and approximately 0.52 in the latter scenario. These levels of correlation are referred to as moderate and strong collinearity, respectively (14). The continuous variable $X_3$ followed a beta distribution with parameters ($\alpha = 2 + 2x_1$, $\beta = 1$) or ($\alpha = 2 + 5x_1$, $\beta = 1$). These parameters make $X_3$ highly skewed. The correlation between $X_1$ and $X_3$ was approximately 0.32 (moderate) in the scenario with $\alpha = 2 + 2x_1$ and approximately 0.52 (strong) in the latter scenario. The underlying risk was set to equal $\text{RD}_{\text{MLE}} + \text{RD}_{\text{X}1} + \text{RD}_{\text{X}3}$, with $\text{RD}_{\text{X}1} = 0.00, 0.15$, or 0.30 for $i = 1, 2$, and 3. Findings about the estimation of $\text{RD}_{\text{X}1}$ when $\text{RD}_{\text{X}2}$ and $\text{RD}_{\text{X}3}$ are both 0.15 are presented in the Results section. The results regarding the estimation of $\text{RD}_{\text{X}1}$ when $\text{RD}_{\text{X}2}$ and $\text{RD}_{\text{X}3}$ are 0.0 or 0.3 instead. Findings about the estimation of $\text{RD}_{\text{X}2}$ and $\text{RD}_{\text{X}3}$ when the risk differences for their covariates are 0.15 are presented in the Results section. Findings for the case of their covariates having a risk difference of 0.0 or 0.3 were similar and, in the interest of space and brevity, are not presented here. A complementary set of simulations in which $X_2$ and $X_3$ were negatively correlated with $X_1$ by changing the aforementioned $\alpha = a +$
TABLE 2. Mean estimate of risk difference and coverage of the 95% confidence interval for the binary variable \( X_1 \), adjusted for \( X_2 \) and \( X_3 \).

<table>
<thead>
<tr>
<th>Prob((x_1 = 1))</th>
<th>Risk difference</th>
<th>Sample size</th>
<th>Moderate collinearity</th>
<th>Strong collinearity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>OLS*</td>
<td>Coverage of 95% CI*</td>
</tr>
<tr>
<td>0.5</td>
<td>0.00</td>
<td>100</td>
<td>-0.0002</td>
<td>95.1 95.2 95.2 95.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>-0.0009</td>
<td>94.9 95.1 95.2 95.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>0.0004</td>
<td>94.6 94.8 94.8 94.9</td>
</tr>
<tr>
<td>0.15</td>
<td>0.1515</td>
<td>100</td>
<td>94.4 94.3 94.4 94.9</td>
<td>0.1509</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>94.7 94.8 94.8 95.0</td>
<td>0.1505</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>94.4 94.7 94.7 94.8</td>
<td>0.1500</td>
</tr>
<tr>
<td>0.30</td>
<td>0.2996</td>
<td>100</td>
<td>94.4 94.6 94.6 95.1</td>
<td>0.2985</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>94.5 94.6 94.6 94.9</td>
<td>0.2998</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>94.6 95.1 95.1 95.1</td>
<td>0.2998</td>
</tr>
<tr>
<td>0.7</td>
<td>0.00</td>
<td>100</td>
<td>95.8 93.8 94.0 94.5</td>
<td>0.0017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>95.2 94.4 94.5 94.7</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>95.3 94.6 94.6 94.6</td>
<td>0.0001</td>
</tr>
<tr>
<td>0.15</td>
<td>0.1517</td>
<td>100</td>
<td>96.4 93.7 93.9 94.7</td>
<td>0.1505</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>96.4 94.4 94.5 94.8</td>
<td>0.1494</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>96.3 94.9 94.9 95.1</td>
<td>0.1496</td>
</tr>
<tr>
<td>0.30</td>
<td>0.2996</td>
<td>100</td>
<td>96.3 94.2 94.3 94.8</td>
<td>0.2986</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>96.2 94.5 94.5 94.8</td>
<td>0.3007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>96.5 95.2 95.3 95.4</td>
<td>0.3000</td>
</tr>
</tbody>
</table>

* CI, confidence interval; OLS, ordinary least-squares.

\( bx_1 \) beta distribution parameters to \( a = a - bx_1 \) was also conducted. The results are not reported, because they were very similar to the results based on positive collinearity as far as the MLS is concerned.

In each run of the simulation, a multivariable OLS regression was performed to obtain the risk difference and the usual OLS standard error, and then the robust variances were calculated to obtain the MLS standard errors. Total sample sizes considered were 100, 200, and 500. Ten thousand runs were performed for each combination of the simulation parameters. The binomial regression was not included, since it often failed to converge (which is the motivation for the present work). Failure to converge does not happen randomly, and therefore simulation results from the binomial regression are likely to be misleading, except in the unlikely event that the estimation does converge in every run.

RESULTS

Table 2 shows the simulation results concerning the risk difference for \( x_1 \). Two key findings can be easily seen. Firstly, the mean of the risk difference estimates was practically identical to the true parameter in all scenarios. Indeed, if two decimal places had been used, they would all have been identical to the true values. Secondly, the coverage of the 95 percent confidence interval of the MLS method using HC2 and HC3 robust standard errors was very reliable. In most scenarios, their 95 percent confidence intervals covered the true parameter values about 95 percent of the time. Among the 36 situations shown in table 2, the numbers of times OLS, HC1, HC2, and HC3 had coverage of less than 94 percent or more than 96 percent were 10, 5, 3, and 0, respectively. The poorer performance of OLS was concentrated in situations where \( X_1 \) was not equally distributed, that is, when \( \text{Prob}(x_1 = 1) = 0.7 \).

In most cases with \( n = 100 \) or \( n = 200 \), HC1 and HC2 gave coverage that was slightly less than 95 percent, while HC3 gave coverage very close to 95 percent. For the cases with \( n = 500 \), the three analyses gave very similar levels of coverage.

The upper half of table 3 shows the simulation results concerning the risk difference for \( x_2 \). The findings are similar to the findings regarding \( x_1 \). Again, the point estimates were very close to the true values, regardless of the simulation parameters. The coverage based on HC3 was closely scattered around 95.0 percent. HC1 and HC2 tended to provide a level of coverage that was slightly less than the nominal level, especially when \( n = 100 \). However, in all 18 situations shown in table 3 concerning \( x_2 \), which, as aforementioned, was quite symmetrically distributed, the coverage was always between 94 percent and 96 percent no matter which standard error estimate was used.

The lower half of table 3 shows the simulation results concerning the risk difference for \( x_3 \). The findings regarding the point estimate of the risk difference and the coverage using HC1, HC2, and HC3 were similar to the aforementioned findings regarding \( x_2 \). However, for this skewed...
variable, OLS worked poorly. In all 18 situations, the coverage was at least 1 percent different from the intended 95 percent.

The performance of the large-sample version of the robust standard error, HC0, is not reported here because there is no good reason to apply it routinely to finite samples. Since it is smaller than HC1 and since HC1 has a tendency to cover true parameter values less often than it should (tables 2 and 3), HC0 must have even smaller coverage and therefore is less acceptable than HC1.

**ILLUSTRATIVE EXAMPLE**

A survey of cancer patients was conducted in Singapore from September 2004 to July 2006. Details on this institutional review board-approved study will be reported elsewhere (Goh et al., Singapore National Cancer Centre, unpublished manuscript). Briefly, approximately 800 ethnic Chinese patients were interviewed with a questionnaire that included the EuroQol 5-Dimension (EQ-5D) questionnaire (15), after they gave written informed consent. The EQ-5D questionnaire contains five classifiers of health status: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. For each item, respondents may choose one of three levels ("no problem," "some problems," or "severe problems") to rate their health on the day of the interview. Approximately one third of the target population understood Chinese only, one third understood English only, and one third understood both. The patients chose the language of the questionnaire (Chinese or English) according to their preference (409 Chinese and 362 English questionnaires were returned without missing values). Most patients self-administered the questionnaire, but some requested interviewer administration.

One of the objectives of the study was to establish the equivalence of the original English EQ-5D and the Chinese translation. It is common that very few respondents report "severe problems," so the analysis of individual EQ-5D items is often based on the dichotomization of "some or severe problem" ($Y_k = 1$, $k = 1, 2, \ldots, 5$) versus "no problem" ($Y_k = 0$). Lou et al. (16) have suggested defining the minimally clinically important difference of the risk of a positive response to the individual dichotomized EQ-5D item as 0.10. Language preference is often related to age, education, performance status (as measured by the Eastern Cooperative Oncology Group (ECOG) score (17)), and mode of questionnaire administration (18). The ECOG score runs from 0 to 4 in steps of 1, with a larger value indicating worse performance. In order to assess the equivalence of the different versions of the questionnaire, we performed multiple regression analysis of the five dichotomized responses in relation to questionnaire version and covariates, with age and ECOG score treated as quantitative variables.

When binomial regression was used, only the analysis of anxiety/depression converged. The modified Poisson method for estimating risk difference as suggested by Spiegelman and Hertzmark (2) also failed to converge in

**TABLE 3.** Mean estimate of risk difference and coverage of the 95% confidence interval for continuous variables $X_2$ and $X_3$, adjusted for $X_1$, and for each other

<table>
<thead>
<tr>
<th>Variable</th>
<th>Risk difference</th>
<th>Sample size</th>
<th>Moderate collinearity</th>
<th>Strong collinearity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>OLS*</td>
<td>HC1</td>
</tr>
<tr>
<td>$X_2$ (symmetric)</td>
<td>0.00</td>
<td>100</td>
<td>-0.0022</td>
<td>95.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>-0.0005</td>
<td>95.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>0.0005</td>
<td>95.1</td>
</tr>
<tr>
<td>$X_3$ (asymmetric)</td>
<td>0.00</td>
<td>100</td>
<td>0.1522</td>
<td>95.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.1435</td>
<td>95.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>0.1494</td>
<td>95.3</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>100</td>
<td>0.3037</td>
<td>95.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.3007</td>
<td>95.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>0.3039</td>
<td>95.8</td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>100</td>
<td>0.0012</td>
<td>96.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>-0.0007</td>
<td>96.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>0.0002</td>
<td>96.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.1542</td>
<td>96.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>0.1506</td>
<td>96.4</td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>100</td>
<td>0.2984</td>
<td>96.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.3001</td>
<td>96.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>0.3015</td>
<td>96.1</td>
</tr>
</tbody>
</table>

* CI, confidence interval; OLS, ordinary least-squares.
the same four analyses. MLS regression analysis worked for all five responses. Table 4 shows the results of the binomial and MLE regression (HC1) analyses of the anxiety/depression item. They produced very similar estimates of risk differences and standard errors. They also produced very similar predicted risk values (figure 1): The intraclass correlation coefficient was 0.9994. Furthermore, all predicted values were within the range of 0.09 and 0.75 using either method. The conclusion based on the two methods was identical: The English and Chinese versions of the questionnaire showed measurement equivalence. Using HC2 or HC3 gave identical standard errors with three decimal places.

**DISCUSSION**

Many authors have commented on the uncritical use of the odds ratio and have suggested that the risk ratio and risk difference may be of more scientific interest in many investigations (8, 19, 20). A major practical problem is that estimation in binomial and log-binomial regression models often fails to converge. Prespecification of a statistical analysis plan is becoming more common and is often required by scientific and ethical review boards. However, it is difficult for researchers to prepare and justify a plan based on the risk ratio or risk difference without knowing whether the plan will work. Zou (8) made a significant contribution by highlighting the fact that the modified Poisson model—that is, the Poisson model with the Huber-White robust standard error—can be used to consistently estimate risk ratio. The analysis of risk difference has remained difficult.

In this manuscript, I have used analytic methods, simulations, and a real survey data set to illustrate that least-squares regression accompanied by the Huber-White robust standard error can be used to estimate risk differences and make correct inferences. There is no risk of nonconvergence, and the method does not require extensive computing time. This approach can be employed using popular statistical packages such as SAS (2) and Stata (21). Following previous terminology, I call this the MLS regression approach.

When sample size is large (e.g., \( n > 200 \)), different versions of the robust variance provide similar results. For a smaller sample size, HC3 is preferred to HC0, HC1, and HC2. Interestingly, in previous simulation studies of very different situations (continuous outcome variables), it was also found that the variance estimators have similar performances for \( n \geq 250 \) but HC3 performed better for \( n < 250 \) (13). Not all major statistical packages support the three small-sample versions of the robust standard error. As reviewed by Long and Ervin (13), HC0 and HC1 appeared to be the default options of most packages. Stata’s “regress” command uses HC1 as the default and has the options of HC2 and HC3 (21). In other packages, the user may need to program the small-sample versions. HC1 is simply HC0

### Table 4. Results from multivariable regression analyses of difference in the risk of anxiety/depression as measured by the EuroQol 5-Dimension questionnaire among 771 cancer patients in Singapore, 2004–2006

<table>
<thead>
<tr>
<th></th>
<th>Binomial model</th>
<th>Modified least-squares</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RD*</td>
<td>SE*</td>
</tr>
<tr>
<td></td>
<td>RD</td>
<td>SE</td>
</tr>
<tr>
<td>English version of questionnaire</td>
<td>-0.009</td>
<td>0.039</td>
</tr>
<tr>
<td>Performance status†</td>
<td>0.125</td>
<td>0.020</td>
</tr>
<tr>
<td>Age (10 years)</td>
<td>-0.022</td>
<td>0.016</td>
</tr>
<tr>
<td>Female sex</td>
<td>0.092</td>
<td>0.035</td>
</tr>
<tr>
<td>Secondary education‡</td>
<td>0.089</td>
<td>0.049</td>
</tr>
<tr>
<td>Postsecondary education‡</td>
<td>0.064</td>
<td>0.060</td>
</tr>
<tr>
<td>Interviewer administration§</td>
<td>0.055</td>
<td>0.044</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.218</td>
<td>0.121</td>
</tr>
</tbody>
</table>

* RD, risk difference; SE, standard error.
† Eastern Cooperative Oncology Group score (17).
‡ Reference group: persons with a primary-school education or no formal education.
§ Reference group: self-administration.
multiplied by \( n/(n - k) \). HC2 and HC3 use the hat matrix, which usually can be produced by the least-squares regression command of statistical packages. This simplifies the programming task.

The MLE regression approach should prove to be a useful tool for researchers who need to analyze binary outcomes. Nevertheless, this approach may not work for all research purposes. As Spiegelman and Hertzmark have discussed (22), data analysis does not necessarily aim to fit the data. It is usually more important to estimate the parameters of interest (often measures of association). In the development of prognostic indices, the parameter of interest is the predicted risk. As such, for the development of a prognostic index, one may prefer the logistic model over the MLE regression models so that the prediction is bounded within the admissible range of 0 to 1. Furthermore, whether it is reasonable to assume linearity in the modeling of risk in relation to an explanatory variable is also something to be examined in a particular situation.

ACKNOWLEDGMENTS

The author thanks Drs. Neal Alexander, Greg Fegan, Eddy Lam, and Paul Milligan for helpful comments on an earlier draft of this article and Dr. Cynthia Goh for providing the data set used in the illustrative example.

Conflict of interest: none declared.

REFERENCES


APPENDIX 1

Exposition of the Robust Estimate of Variance

In matrix form, the ordinary least-squares (OLS) regression coefficients are given by \( \hat{\beta} = (X'X)^{-1}X'Y \), and the OLS variance-covariance matrix is

\[
\text{Var} (\hat{\beta}) = \text{Var}([X'X]^{-1}(X'Y))
\]

\[
= (X'X)^{-1}\text{Var}(X'Y)(X'X)^{-1}
\]

\[
= (X'X)^{-1}X'(\sigma^2I)X(X'X)^{-1}
\]

\[
= \sigma^2(X'X)^{-1},
\]

where \( \sigma^2I \) is the assumed variance structure that is inappropriate for binary data.

The Huber-White method estimates \( \text{Var}(X'Y) \) by

\[
\hat{\text{Var}}(X'Y) = \sum_{i=1}^{n} \hat{e}_i^2 x_i'x_i,
\]

where \( \hat{e}_i^2 = (y_i - x_i\hat{\beta})^2 \) is an empirical estimate of the variance of \( y_i \). The only assumption required by this procedure is that the \( y_i \)’s are independent. Substituting this into the \( \text{Var}(\hat{\beta}) \) formulae above produces the asymptotic version of the robust estimate of variance in equation 6.
APPENDIX 2
Variance of the Risk Difference Estimated by the Modified Least-Squares Approach for a Single Binary Exposure Variable

Using the notation in table 1, \( \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^2 = (c + d) \),

\[
(X'X)^{-1} = \frac{1}{n \sum_{i=1}^{n} x_i - (\sum_{i=1}^{n} x_i)^2} \begin{pmatrix}
\sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\
-\sum_{i=1}^{n} x_i & n
\end{pmatrix}
\]

\[
= \begin{pmatrix}
\frac{1}{n - (c + d)} & -\frac{1}{n - (c + d)} \\
-\frac{1}{n - (c + d)} & \frac{1}{(c + d) - (c + d)^2/n}
\end{pmatrix},
\]

and

\[
\sum_{i=1}^{n} e_i^2 x_i = \sum_{i=1}^{n} e_i^2 \begin{pmatrix}
1 & x_i \\
1 & x_i^2
\end{pmatrix}
\]

\[
= \begin{bmatrix}
[a(0-b)/(a+b)]^2 + b[1-b/(a+b)]^2 + c[0-d/(c+d)]^2 + d[1-d/(c+d)]^2 & c[0-d/(c+d)]^2 + d[1-d/(c+d)]^2 \\
c[0-d/(c+d)]^2 + d[1-d/(c+d)]^2 & c[0-d/(c+d)]^2 + d[1-d/(c+d)]^2
\end{bmatrix}
\]

\[
= \begin{bmatrix}
ab^2/(a+b)^2 + a^2b/(a+b)^2 + cd^2/(c+d)^2 + c^2d/(c+d) & cd^2/(c+d)^2 + c^2d/(c+d) \\
cd^2/(c+d)^2 + c^2d/(c+d) & cd^2/(c+d)^2 + c^2d/(c+d)
\end{bmatrix},
\]

With the above elements substituted into equation 6 and completing the matrix multiplication, the element in the second row and second column of the robust variance-covariance matrix is

\[
\text{Var}(RD_{MLS}) = ab/(a+b)^3 + cd/(c+d)^3 = \frac{\hat{\rho}_0(1-\hat{\rho}_0)}{n_0} + \frac{\hat{\rho}_1(1-\hat{\rho}_1)}{n_1}.
\]