Practice of Epidemiology

Regression Calibration When Foods (Measured With Error) Are the Variables of Interest: Markedly Non-Gaussian Data With Many Zeroes

Gary E. Fraser* and Daniel O. Stram

* Correspondence to Dr. Gary E. Fraser, Department of Epidemiology and Biostatistics, School of Public Health, Loma Linda University, 24951 North Circle Drive, Room 2033, Loma Linda, CA 92350 (e-mail: gfraser@llu.edu).

Initially submitted February 18, 2011; accepted for publication August 16, 2011.

Regression calibration has been described as a means of correcting effects of measurement error for normally distributed dietary variables. When foods are the items of interest, true distributions of intake are often positively skewed, may contain many zeroes, and are usually not described by well-known statistical distributions. The authors considered the validity of regression calibration assumptions where data are non-Gaussian. Such data (including many zeroes) were simulated, and use of the regression calibration algorithm was evaluated. An example used data from Adventist Health Study 2 (2002–2008). In this special situation, a linear calibration model does (as usual) at least approximately correct the parameter that captures the exposure-disease association in the “disease” model. Poor fit in the calibration model does not produce biased calibrated estimates when the “disease” model is linear, and it produces little bias in a nonlinear “disease” model if the model is approximately linear. Poor fit will adversely affect statistical power, but more complex linear calibration models can help here. The authors conclude that non-Gaussian data with many zeroes do not invalidate regression calibration. Irrespective of fit, linear regression calibration in this situation at least approximately corrects bias. More complex linear calibration equations that improve fit may increase power over that of uncalibrated regressions.

bias (epidemiology); foods; measurement error; power; regression calibration

Abbreviation: AHS-2, Adventist Health Study 2.

Measurement error is a well-recognized problem in nutritional epidemiology (1, 2). Although it has been described for error correction in dietary analyses, regression calibration has rarely been applied to analyses where foods are the variables of interest. Regression calibration as developed in the epidemiologic literature (3) has usually been limited to data that are approximately normally distributed, although this does not appear to be required by the regression calibration algorithm (4).

Thus, it is unclear whether measurement error correction by regression calibration as has been previously described is a satisfactory procedure where foods are the exposures of interest, and indeed there are some special considerations in this setting. Hypotheses about relations of food/food-group consumption with disease are commonly tested in nutritional epidemiology, and there is a clear need for measurement error correction in such regressions. However, the independent variables may have distributions that are quite irregular and far from normal.

Food variables with many zero values are common in most cohorts because of individual likes and dislikes that result in many subjects’ choosing not to eat some foods. For instance, vegetarians eat no meats, and at least one-sixth of Seventh-day Adventists in the Adventist Health Study 2 (AHS-2) cohort (5) eat virtually no soy, but half of Adventists on average eat soy at Asian levels (6). More than 25% of subjects indicated zero intakes (eaten less than once per month) in questionnaire responses for 21 of 49 foods/food groups evaluated in this study.

Rosner and Gore (7) evaluated the standard approach to regression calibration where data on individual foods (apparently without zero intakes) are transformed to approximate normality. However, such transformations are impossible for some foods or food groups with many zero intakes or irregular distributions.
Other recent work (8–10) dealing with this sort of data has focused on mixture models that use separate models for zero and nonzero data. The method of Kipnis et al. (10) is the most relevant and is described for survey settings in which all study participants have outcome data but also reference dietary data such as data from repeated recalls, which are rare in existing cohort studies. Optional questionnaire dietary data would increase precision, especially for foods consumed less frequently. This motivates joint likelihood-based analysis of the calibration and main studies to estimate disease risk. The models are often quite statistically complex (8–10) and involve stronger assumptions about the distributions of true exposure than are required for the simpler regression calibration methods that we discuss here.

Regression calibration requires that a calibration equation relating the mean of true diet \( T \) to questionnaire estimates \( Q \) be constructed in order to correct bias. When there are many zero values, \( E(T|Q) \) will not have normally distributed homoscedastic residuals and may not have a simple linear form. Errors may be positively skewed at values of \( Q \) close to the origin as negative errors around \( E(T|Q) \) become crowded by the line \( T = 0 \). Does this affect estimation and hypothesis-testing in a calibrated “disease” regression?

Although the calibration relation, \( E(T|Q) \), is usually modeled as a simple linear function, there are other, more complex linear models that can potentially improve statistical power. Transformations such as logarithms may also improve power by reducing heteroscedasticity. With non-Gaussian data, estimates of variances and confidence intervals will usually require nonparametric techniques such as resampling (11, 12). In this article, we simulate a data set containing many zeroes and a positive skew in the nonzero data in order to explore these ideas. An example using real data from the AHS-2 cohort (5) is shown.

**METHODS**

Regression calibration requires that an estimator of \( T \) given \( Q \) be available. Then if \( f \) is the distribution function of a disease indicator \( D \) and if the nondifferential error assumption is satisfied (i.e., \( f(D|T) = f(D|Q, T) \)), then \( E(T|Q) \) is used to perform the variable in the disease regression, the observed coefficient will be very close (except in extreme cases) to an unbiased estimate of that which would have been obtained had \( T \) been available and used instead (13). \( E(T|Q) \) is estimated by means of a calibration equation which as used below is always a linear calibration approximation.

### The linear “disease” regressions

Where the true disease regression is \( g(D) = \alpha_T + \beta_T T + \varepsilon_T \), the calibrated disease regression is \( g(D) = \alpha_{calib} + \beta_{calib} E(T|Q) + \varepsilon_{calib} \), where \( E(\beta_{calib}) = \beta_T \) if assumptions for regression calibration are satisfied. For the moment, it is assumed that \( T \) can in fact be observed in the calibration study. The observed disease regression is \( g(D) = \alpha_S + \beta_S Q + \varepsilon_S \).

### Transformations of the dietary data

Logarithmic transformations are often useful as a variance-stabilizing tool because heteroscedasticity of the calibration equation increases the variance of estimated coefficients (14). Although it is not essential, we transform both \( T \) and \( Q \) but, because of the frequent zeroes, we suggest using \( \log(kX + 1) \), \( X = T, Q \), where \( k \) is chosen so that \( kX \) is close to \( kX + 1 \) for most values of \( X \). Then adding 1 to \( kX \) is a proportionately small increment, and a zero on the original scale remains a zero. \( T' \) and \( Q' \) are defined to indicate nutritional variables that are transformed in this way. In our examples, \( k = 1.0 \) was used, as most nonzero values of \( X \) were relatively large.

### Impact of poor fit of a calibration model on the bias of a linear calibrated disease regression

We show here that poor fit of a calibration equation still results in a consistent calibrated linear disease regression estimator. Consider a function \( h(Q') \) such that \( \text{Cov}(h(Q'), T')/\text{Var}(h(Q')) = 1 \) (i.e., the regression of \( T' \) on \( h(Q') \) has slope 1). By definition, \( \text{Cov}(h(Q'), D)/\text{Var}(h(Q')) \) is the regression slope of \( D \) on \( h(Q') \). This is equal to \( \text{Cov}(h(Q'), \alpha_T + \beta_T T + \varepsilon_T)/\text{Var}(h(Q')) \). So long as \( \varepsilon_T \) is uncorrelated with \( T' \) and \( h(Q') \), the nondifferential error assumption, this regression slope is \( \beta_T \text{Cov}(h(Q'), T'/\text{Var}(h(Q'))) = \beta_T \times 1 = \beta_T \).

The calibration equation, \( h(Q') \), can be constructed as any regression function \( T' = h_c(Q') + \varepsilon_C = \alpha_C + \beta_C h_1(Q') + \beta_C h_2(Q') + \ldots + \beta_C h_q(Q') + \varepsilon_C \), where \( C \) indicates the calibration equation and \( \beta \)'s are the corresponding true multiple regression parameters of \( T' \) on \( h_1(Q') \ldots h_q(Q') \). This implies (proof by contradiction is very simple) that the regression slope of \( T' \) on \( h_c(Q') \) must be 1, and the above proof (see preceding paragraph) applies. Under usual regularity conditions, consistency of the estimate of \( \beta_T \) follows so long as the sizes of both the main and calibration studies are allowed to increase (so that the estimates of parameters in \( h(Q') \) and in the regression given \( h(Q') \) approach their expected values).

Although it was not assumed that the ordinary least squares estimate of \( T' \) given \( h_c(Q') \) is the same as \( E(T'|h_c(Q')) \), a poorly fitting linear calibration equation will still provide an estimate that consistently corrects \( \beta_T \) in order to estimate \( \beta_T \) when the “disease” regression is linear. One caveat is that this requires a calibration study that is a representative sample of the entire study. If the true model for \( E(T'|h_c(Q')) \) is nonlinear in \( Q' \) and if the sample of \( Q' \) is not representative, then the ordinary least squares slope of \( T' \) on \( h_c(Q') \) depends upon which part of the curve is sampled.

### A general regression calibration framework

From consideration of a Taylor’s theorem expansion (see Web Appendix 1, available on the *Journal’s* Web site (http://aje.oxfordjournals.org/)), it can be concluded that the results noted above, which apply when the “disease” model is linear, will also approximately apply to nonlinear “disease” models in many common situations encountered in epidemiologic work. These include 1) approximate deattenuation of effect estimates for measurement error correction (3, 13, 15) and 2) the situation where, among other factors, the fit of the calibration model is a determinant of power of the calibrated result (see below).
Effect of the fit of a linear calibration model on the power of a calibrated linear disease model

Define $R_c^2$ and $R_C^2$ as the multiple correlation coefficients for the correlations between $D$ and $T$ and between $T$ and $Q$, respectively. Then, by considering noncentrality parameters of chi-squared distributions used to test hypotheses about $\beta_T$ and $[\hat{\beta}_C]$ (see Web Appendix 2), it is seen that for small $R_c^2$, the loss of effective sample size associated with the calibration is approximately inversely related to $R_C^2$ but with comparatively greater losses at higher values of $R_f^2$.

The power of an uncalibrated univariate result is approximately the same as that of a calibrated analysis when the calibration equation is simple linear (nonpolynomial) in form (16). An immediate conclusion to be drawn from the above results is that if a more complex calibration equation significantly improves the fit of the calibration equation, this calibrated result will have power exceeding that of an uncalibrated analysis.

Examining the fit of the calibration model

The influence of the fit of the calibration study on power is mediated through the value of $R_c^2$, which is improved by a better fit. To maximize power, it will be worthwhile to examine calibration equation residuals and possibly add higher-order terms, or use a linear spline or other polynomial models.

One flexible calibration model that may improve fit in regions where many subjects have the same values (e.g., zero) is a partitioned approach:

$$T' = x_C + \delta_C H + \sum_{k=1}^{m} y_{ik} B_k + \beta_C Q'(1 - H) + Z^T \gamma_C + e_C, \quad (1)$$

where $H$ is an indicator variable taking the value 1 when $Q < Q_{cont}$, otherwise 0; where $Q_{cont}$ is the value below which the model includes $m$ nonzero categories (a step function), represented by $k = 1, \ldots, m$; and where $B$ are the corresponding indicator variables. Thus, exact mean values of $T'|Q'$ are predicted by $x_C + \delta_C$ when $Q = 0$ and by $x_C + \delta_C + y_{ik}$ when $Q$ falls into category $k$. Coefficient $\beta_C$ describes the slope between $Q'$ and $T'$ when $Q \geq Q_{cont}$.

With the model described by equation 1, by definition the mean function fit is exact when $Q < Q_{cont}$. Thus, in this respect, by potentially providing a “separate” model when $Q = 0$, it is similar to the 2-part model of Kipnis et al. (10). At higher values of $Q$, the skewness of residuals is usually small, and we have used the cumres function in $R$ software (R Foundation for Statistical Computing, Vienna, Austria) to evaluate the fit of the model there (17, 18).

Implications of non-Gaussian (and zero) data for the assumption that $E(R'|T') = T'$

In practice, when developing the calibration equation, $T$ is unobservable, and a surrogate reference measure ($R$), such as the average of repeated dietary recalls or diaries, is employed. This brings the complication of two further assumptions. The first is that the errors in $R'$ and $Q'$ about $T'$ are independent. This assumption is not addressed further, since nothing different results when dealing with non-Gaussian distributions. The second assumption is that $R' = T' + e$, $E(e) = 0$, that is, $E(R'|T') = T'$. Then the random errors, $e$, will not bias estimates of $\beta_C$ (19).

The association between $R'$ and $T'$ has 3 regions of interest: 1) $T = 0$; 2) an intermediate low-intake region of $T$; and 3) $T$ is larger. The $T = 0$ assumption can only be satisfied when both $T = 0$ and $R = 0$, since negative values are not possible. Where subjects in fact consume none of a particular food ($T = 0$), they are most unlikely to then claim that they have eaten that food in a recall or food diary ($R$). Nevertheless, such claims are possible, and sensitivity analyses may be informative. With an intermediate low-intake region of $T$, $R$ is usually a discrete measure even if it is the average from a number of days. A subject’s daily intakes will fluctuate around some underlying mean. If $E(R|T) = T$ for a particular recall day, this will also be true for a (possibly weighted) sum across days, to form (for example) an estimate of weekly intake. When $T$ is small, then for some subjects all days of $R$ may have zero values, and estimated weekly intakes will erroneously also appear to be zero. However, these are counterbalanced by values of $R$ from other subjects with the same value of $T$, where $R$ values are greater than expected. When $T$ is larger, the discrete nature of $R$ is largely masked and the assumption that $E(R|T) = T$ needs no further clarification. In AHS-2, this was when the average from the 6 recalls obtained was sufficiently high that the probability of all recalls being zero was low.

A value of $R$ cannot be an unbiased estimate of $T$ on both the transformed and untransformed scales. However, the approximation may often be sufficiently close. A second-order approximation is $E[\log(kT + 1)] \approx \log(kT + 1) - \sigma^{-2}(kT + 1)^2$, where $\sigma^{-2}$ will probably depend on $T$ because of heteroscedasticity. So long as $\sigma^{-2}$ is much smaller than $(kT + 1)^2$, the assumption that $E(R'|T') = T'$ will be approximately satisfied. An $R$ from many replicates will minimize $\sigma^{-2}$.

Simulated data where $T$ and $Q$ contain many zeroes and where a $\log(X + 1)$ (i.e., $X$) transformation is used

We simulated a distribution with many zeroes and a positive skew. Details of this simulation can be found in Web Appendix 3. Briefly, this was modeled approximately on soy protein intake in the AHS-2 population, where there were many zero intakes. We also investigated the effect of varying the proportion of zeroes between 25% and 60%. Disease events ($D$) were generated using a logistic function, conditional on $T'$, such that the odds ratio for disease comparing $T = 7.84$ g/day with $T = 0$ g/day was 0.6.

Populations of size 13,500 were simulated. Evaluation of the mean values and standard errors of calibrated regression coefficients used the results from 1,000 such populations of $Q$ and $D$ for each set of conditions, all conditional on a fixed set of $T$. A new calibration study of size 4,500 subjects was randomly selected from each new population, large enough so that the variance due to the calibration was small and did not confuse the main results.
Practical estimation of standard errors and confidence intervals for calibrated "disease" regression estimates

When errors are non-Gaussian, a method that bootstraps only within the calibration study has been described by Ferrari et al. (20). This uses the total variance formula

\[ \text{Var}(\beta_{\text{calib}}) = E\{\text{Var}(\text{Bootstrap-estimated } \beta_{\text{calib}})\} + \text{Var}\{E[\text{Bootstrap-estimated } \beta_{\text{calib}}]\}, \quad (2) \]

where the mean and variance on the right side of equation 2 are calculated using the J calibrated "disease" regression results produced by the J bootstrap calibration samples. It was suggested by Ferrari et al. (20) that J ≈ 300 bootstrap samples provides a stable result. A more computation-intensive algorithm (12) has the advantage of not requiring that calibrated beta coefficients are normally distributed when producing confidence intervals.

The large calibration studies that were used in the simulations resulted in only a trivial contribution from the last term of equation 2. Thus, effects of uncertainty in the calibration equation could be ignored for practical purposes.

Adventist Health Study 2

Real data for further illustration come from AHS-2 (5), a cohort study of 96,000 Seventh-day Adventists living in the United States and Canada (2002–2008). A 130-item food frequency questionnaire was completed by study members. For cereals and meat analogs, there were several pages of commonly consumed commercial products and space for write-ins. A sample (n = 1,011) of these subjects comprise a calibration study, which represents the cohort very closely (21). These subjects completed a second food frequency questionnaire and six 24-hour telephone dietary recalls. For each subject, recalls were collected in 2 blocks of 3 days, each 5–6 months apart. Each block provided a synthetic week, created by appropriate weighting of a Saturday, a Sunday, and a weekday—days on which data were always collected. The recalls were obtained interactively by telephone using Nutrition Data System software (22), which also produced nutrient values. Validity correlations between food frequency questionnaire estimates and dietary recall estimates are generally relatively high (21). Energy adjustment was not included in this illustration, since using the density method did not improve the validity of data for soy (or other foods/food groups that were tested). Further details about the cohort are provided elsewhere (5).

The estimate of soy intake in the food frequency questionnaire was gathered from questions that included intake of canned soybeans, fermented soy products, and tofu and a 2-page list of commercial soy products and soy milks often used by this population. In addition, there was space for write-ins, which were coded separately. In the AHS-2 data, standard errors of calibrated coefficients were estimated using the method referred to above (20) with 300 bootstrap samples of the calibration data.

RESULTS

Regression calibration in simulated data with many zero values

These simulated data showed the expected means, variances, and correlations. There were approximately 232, 240, and 270 disease events when zeroes represented 25%, 40%, and 60% of the exposure data, respectively, thus corresponding to disease risks of 0.0172, 0.0178, and 0.0200. Disease risk is higher with a greater proportion of zero intakes, consistent with the negative association between risk and exposure.

The simulated calibration equation was markedly nonlinear near the origin (Figure 1), but for most of the range of Q', it was approximately linear. Thus, a simple linear calibration model applied to these data showed clear evidence of lack of fit, as indicated by cumres plots with highly significant P values.

Several calibration models were explored with each sample: 1) a simple linear model; 2) a linear model with both quadratic and cubic terms in Q'; 3) a linear spline model with a node close to zero; and 4) a partitioned model (equation 1) with m = 0 (i.e., linear except at the zero point in this case), this being the only one of these 4 models where the cumulative residual test indicated a good fit.

When 25% of T values were set to zero (Table 1), the mean of the 1,000 "true" beta coefficients (βT) estimated from simulated data was −0.237. The mean values of the calibrated logistic "disease" model coefficients (βcalib) were relatively close to the true value of −0.234 for all calibration models (Table 1). These values were easily within 2 standard deviations of the estimator for all except the naive (crude) regression, which was seriously biased toward the null. This is consistent with the expectation that lack of fit in a linear calibration model will not create more than trivial biases in typical logistic regressions.

However, the calibration models with better fit produced importantly improved precision and statistical power. This was particularly evident when comparing the z score for the calibrated β (mean bootstrapped β divided by bootstrapped
Table 1. Bias and Precision of Calibrated “Disease” Model Beta Coefficients With Different Calibration Models, Adventist Health Study 2, 2002–2008

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean (β)</th>
<th>Mean (SE(β))</th>
<th>z</th>
<th>Mean(Var(Est[β]))</th>
<th>Var(Est[β])</th>
</tr>
</thead>
<tbody>
<tr>
<td>True regression</td>
<td>-0.237</td>
<td>0.076</td>
<td>3.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naive (crude) regression</td>
<td>-0.101</td>
<td>0.060</td>
<td>1.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibrated coefficients</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form of calibration model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple linear</td>
<td>-0.255</td>
<td>0.151</td>
<td>1.69</td>
<td>0.0228</td>
<td>0.000065</td>
</tr>
<tr>
<td>Quadratic/cubic</td>
<td>-0.238</td>
<td>0.123</td>
<td>1.93</td>
<td>0.0150</td>
<td>0.000020</td>
</tr>
<tr>
<td>Spline</td>
<td>-0.231</td>
<td>0.109</td>
<td>2.12</td>
<td>0.012</td>
<td>0.000018</td>
</tr>
<tr>
<td>Partitioned</td>
<td>-0.231</td>
<td>0.109</td>
<td>2.12</td>
<td>0.012</td>
<td>0.000018</td>
</tr>
</tbody>
</table>

Abbreviation: SE, standard error.

a Simulated data (1,000 replicates) with 25% true zeroes and positive skew (see Web Appendix 3).
b Estimated SEs for the calibration models are calculated as \(\sqrt{\text{Mean}(\text{Var(Est[β])}) + \text{Var(Est[β])}}\).

We also performed simulations that included the reference method, \(R\). Then analyses were based on \(Q\) and \(R\) rather than \(Q\) and \(T\). The results were essentially unchanged, as expected, and details are omitted for the sake of brevity.

AHS-2 data: soy intake

Table 3 shows the results of calibration using real data from AHS-2 and assuming for this purpose that the recall data are true intakes. The distribution of soy protein consumption (16% of subjects consumed <1 g/day, treated as zero) had a strong positive skew and was related to all-cause mortality in 2 separately calibrated proportional hazards regressions. Covariates of both the disease model and the calibration equations were age, race, and gender. The first calibration equation was initially simple linear in form, but after evaluating the fit of the model, adding a squared term in \(\log(\text{soy protein} + 1)\) from the questionnaire slightly improved \(R_c^2\) from 0.336 to 0.339, and was the model finally used in these analyses. For soya, a partitioned calibration model had a slightly lower \(R_c^2\) and was not used here.

The difference between corrected and uncorrected beta coefficients in the “disease” model is substantial, corresponding to odds ratios of 0.60 and 0.75, respectively. Although the standard error of the beta coefficient increases with calibration, it does so in approximate proportion to the beta coefficient values. This is because the calibration study is large, and also because the errors in dietary variables are not strongly related to covariates measured without error. The \(t\) value for this soy model was a little higher after calibration, possibly reflecting the small improvement of fit in the calibration model due to inclusion of the squared term.

DISCUSSION

These results show that non-Gaussian data that include many zeroes need not greatly complicate regression calibration. If the relation between \(T^*\) and \(Q^*\) is not well-described by a straight line, a simple linear calibration will still provide appropriate estimates when the “disease” regression is linear, or under the common condition of a relatively linear region of...
Table 3. Calibration Equation and Corrected and Uncorrected Proportional Hazards Regressions of All-Cause Mortality on Intake of Soy Protein, Adventist Health Study 2, 2002–2008a

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Calibration Model (log[Recall Data + 1]) is Dependent ($R_c^2 = 0.339$)</th>
<th>Proportional Hazards Regressions Predicting All-cause Mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unit or Comparison</td>
<td>$\beta$ (SE)</td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
<td>0.134 (0.165)</td>
</tr>
<tr>
<td>Log(FFQ soy + 1)</td>
<td>g/day</td>
<td>0.775 (0.094)</td>
</tr>
<tr>
<td>[Log(FFQ soy + 1)]$^2$</td>
<td>(g/day)$^2$</td>
<td>-0.051 (0.027)</td>
</tr>
<tr>
<td>Age</td>
<td>10 years</td>
<td>-0.010 (0.021)</td>
</tr>
<tr>
<td>Gender</td>
<td>$F = 0, M = 1$</td>
<td>0.224 (0.060)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>$W = 0, B = 1$</td>
<td>-0.088 (0.058)</td>
</tr>
</tbody>
</table>

|                      | Unit or Comparison | $\beta$ (SE) | $t$ | $\beta$ (SE) | $t$ |
|                      |                    |              |    |              |    |
| Log(soy protein + 1) | 0 g/day vs. 30 g/day | -0.299 (0.085) | 3.620 | -0.514 (0.134) | 3.836 |
| Age                  | 10 years           | 0.961 (0.020) | 47.534 | 0.959 (0.020) | 47.950 |
| Gender               | $F = 0, M = 1$     | 0.472 (0.046) | 10.255 | 0.505 (0.047) | 10.745 |
| Race                 | $W = 0, B = 1$     | 0.233 (0.064) | 3.620 | 0.217 (0.064) | 3.391 |

Abbreviations: B, black; F, female; FFQ, food frequency questionnaire; M, male; SE, standard error; W, white.

aData were obtained from Adventist Health Study 2 (5).

A logistic relation. However, a poorly fitting calibration equation will cause unnecessary loss of power. For instance, if $R_c^2 = 0.25$ and $R_T^2 = 0.3$, there is marked loss of power due to the calibration (measured in relative sample sizes for equal power). A ratio of sample sizes equal to 5.29 is necessary to obtain the same power as if $T$ were observed. If $R_c^2 = 0.40$, this ratio improves to 3.14.

Thus, if there is substantial nonlinearity, an important improvement in power may be produced by the relatively small additional effort necessary to find a calibration model with better fit. In the AHS-2 data, we hardly ever found evidence of strong departures from linearity in the calibration equation, and the soy example (Table 3) is typical. However, in other data sets or with other variables this may not be the case.

A 2-part method described by Kipnis et al. (10) was developed for dietary surveys (with associated outcome data). It gives a combined analysis of calibration and main study data but requires that repeated reference measures (i.e., 24-hour recalls) be available for all subjects as the primary dietary measure. Logistic regression is used to model the probability that true exposure is greater than 0, and then a Box-Cox transformation and linear regression to model exposures greater than zero. It is possible that investigators in future cohort studies may be able to collect repeated recall data from all subjects, but such data are not available in most existing cohort studies.

Was the calibration worthwhile in the AHS-2 results of Table 3? As expected, power is not greatly changed compared with an uncorrected regression, but the magnitude of effect was markedly changed. This is an important benefit, giving greater interest to the effects of the dietary variable on disease risk. In multivariate regression calibration, we have previously shown that effect size, the sizes of statistical tests (especially in large studies), and power may change markedly depending on the correlations between errors (16), providing an even greater motivation for corrected analyses.

Although the variance in the calibration equation coefficients was effectively removed in analyses of the simulated data by using a large calibration study, the main results presented still hold when the calibration coefficients are less precisely identified. Var($\beta_{calib}$) can be divided into the 2 parts identified in equation 2. The first part results when the calibration coefficients are known precisely, and the second results from random variation in these estimates.

The log transformation is warranted when values of $Q$ exhibit significant heteroscedasticity or skewness. Thoresen (23) has demonstrated in logistic regression calibration that although heteroscedasticity is not a severe problem, markedly skewed variables (rather than residuals) can produce biased results. However, as noted, transformations can disturb the required approximation that $E(R' | T') = T'$, even if it does hold exactly for $R$ and $T$.

We also demonstrate that transformations can markedly affect the relative bias of uncorrected analyses.

In summary, where the distribution of intake of a food is markedly non-Gaussian, perhaps containing many zeroes, linear regression calibration can safely be applied, with the usual assumptions. If the calibration relation is nonlinear, the calibration population must properly represent the cohort. In this situation, it could also be worth considerable effort to optimize the fit of the calibration equation, since this will improve the power of a calibrated analysis. As usual, when compared with uncorrected analyses, appropriate regression calibration will improve bias (usually markedly), and where there is a nonlinear calibration relation, it will potentially also improve power.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology and Biostatistics, School of Public Health, Loma Linda University, Loma Linda, California (Gary E. Fraser); and Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California (Daniel O. Stram).

This study was partly supported by National Institute of Health grant R01 CA094594.

Conflict of interest: none declared.

REFERENCES


