

Measurement error in the explanatory variable of a binary regression : regression calibration and integrated conditional likelihood in studies of residential radon and lung cancer

Short title: Measurement error in binary regression

T. Fearn¹, D.C. Hill² and S.C. Darby²

¹Department of Statistical Science, University College London, London, U.K.

²Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU), University of Oxford, Oxford, U.K.

Correspondence to: Prof. T. Fearn, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, U.K. E-mail: tom@stats.ucl.ac.uk, Tel: 020 7679 1873, Fax 020 7383 4703

Contract/grant sponsors: This study was carried out with funding from Cancer Research UK and the European Commission (contract no: 516483 (FI6R); project: alpha-risk). These financial sponsors had no role in study design, study implementation, or writing of the report.

Summary

In epidemiology the dependence of disease risk on an explanatory variable in the presence of several confounding variables is commonly investigated by fitting a binary regression using a conditional likelihood, thus eliminating the nuisance parameters. When the explanatory variable is measured with error the estimated regression coefficient is biased, usually towards zero. Motivated by the need to correct for this bias in analyses that combine data from a number of case-control studies of lung cancer risk associated with exposure to residential radon, two approaches are investigated. Both employ the conditional distribution of the true explanatory variable given the measured one. The method of regression calibration uses the expected value of the true given the measured variable as the covariate. The second approach integrates the conditional likelihood numerically by sampling from the distribution of the true given the measured explanatory variable. The two approaches give very similar point estimates and confidence intervals, not only for the motivating example but also for an artificial data set. These results and some further simulations that demonstrate correct coverage for the confidence intervals suggest that for studies of residential radon and lung cancer the regression calibration approach will perform very well, so that nothing more sophisticated is needed to correct for measurement error.

Keywords: binary regression; case-control study; conditional likelihood; measurement error; radon; regression calibration

1 Introduction

In most countries, the natural radioactive gas radon is the largest source of exposure to ionizing radiation in the general population [1]. Around 20 case-control studies investigating the risk of lung cancer associated with radon exposure in the home have been carried out in various different countries. In most of these studies the lung cancer risk tends to increase with increasing exposure, but no individual study has been large enough to provide an estimate of the risk that is sufficiently precise for use in formulating policy to control radon-associated risks. The motivation for the investigation reported here is the need to combine information from these studies, in order to obtain a more precise estimate of the risk. The statistical analysis, described in more detail below, is essentially a binary regression of case-control status on the estimated radon exposure of each individual included in the analysis. Estimating the exposure involves attempting to measure the radon concentration in both current and previous homes of each of the individuals. Where measurements are obtained they are subject to substantial error, with a coefficient of variation of around 50% [2, 3]. Furthermore, some measurements are missing, because the home cannot be accessed. It is known that in both linear [4] and nonlinear [5] regression measurement error in the explanatory variable will affect the regression coefficient, usually shrinking it towards zero compared with that from a regression on the same variable measured without error. A combined analysis correcting for this attenuation has recently been carried out for data from 13 European studies [2], and a further analysis combining data from over 20 studies worldwide is underway [2, 6, 7].

It is sometimes argued that such correction is inappropriate, because any predictions or actions would in any case be based on measured, not true, radon exposures so that risk estimates should therefore be based on the measured exposures. In the case of an exercise combining several studies there is a problem with this argument. Exactly for simple linear regression, and approximately for nonlinear regression, the bias in the regression coefficient estimated using measured values is determined by the ratio of the measurement error variance to the variance of the explanatory variable. In the radon studies these ratios vary from study to study [2, 8] and so in the absence of any correction the different studies will estimate regression parameters that are different both numerically and in their precise meaning. If the data from several studies are combined with no correction for measurement error, the extent of the attenuation in the resulting estimate of the regression coefficient will depend on the variance ratios in all the component studies and will not have a clear interpretation in the context of any one of them. Implicit in any exercise combining studies is the desire to estimate some globally meaningful quantity that is comparable across studies. The obvious such quantity here is the dependence of risk on true radon exposure.

Sections 2 and 3 briefly describe the European collaborative analysis, and review some of the possible approaches to correcting for measurement error in this situation. The two approaches selected for further investigation are regression calibration [5] and a Monte Carlo integration of the unobserved true radon exposure from a likelihood involving both the unobserved true and the measured exposure. These approaches are described in Sections 4 and 5 and the results of applying them

in the European collaborative analysis are reported in Section 6. In Section 7 the same methods are applied to an artificial data set and to repeated simulations of this artificial data set, where they give very accurate results. Section 8 contains a brief discussion of the results and some conclusions.

2 Collaborative analysis of 13 radon studies

Data from 13 case-control studies of residential radon and lung cancer, carried out in nine European countries, have recently been combined [2, 9]. In total there were 7148 cases of lung cancer and 14 208 controls in the combined data set. The explanatory variable for each individual was his or her average exposure to residential radon concentration over a period of 30 years. There were, on average, 2.7 addresses per individual. Attempts were made to measure the radon concentrations in as many of these as possible using long-term α -track detectors, and measurements were available for around 80% of the 30-year period on average. They were, however, subject to considerable measurement error. Investigations have been carried out in several of the countries to estimate at least some of the components of this error by making repeated measurements in the same home on different occasions [2]. All these investigations indicated normally distributed additive measurement error on a log scale, with a typical standard deviation being 0.5. This corresponds to a coefficient of variation of 50% in the untransformed radon measurements. It is the magnitude of this variation that makes correction for measurement error so important.

All of the correction methods considered here utilise the population distributions of residential radon concentrations in each of the study areas. As the disease stud-

ied has low incidence these distributions may be estimated from the large number of measurements made on control individuals in the studies. In every study the distribution of log radon was found to be approximately normal.

To adjust for confounding variables, the data were cross-classified by study, age, sex, region of residence, and smoking history. This cross classification produced around 1700 strata. For an individual in stratum s the probability of disease given exposure was modelled as

$$p(y = 1|r, s) = \frac{\alpha_s(1 + \beta r)}{1 + \alpha_s(1 + \beta r)}, \quad (1)$$

where $y = 1$ for cases and 0 for controls, r is radon exposure for the individual and α_s is a baseline odds for stratum s . The use of the linear form $1 + \beta r$ rather than the more usual $\exp(\beta r)$ is common in modelling radiation risks [10]. However, with the exception of the result in the appendix, which is specific to this linear odds model, the methodology described here would apply equally well to the logistic model.

In situations such as this one, where there are very many strata S , it is usual to base the analysis on the conditional likelihood

$$\ell_c(\beta, y, r) = \prod_{s=1}^S \frac{\prod_{i \in C_s} (1 + \beta r_{is})}{\sum_{p \in P} \prod_{i \in C_s^p} (1 + \beta r_{is})}. \quad (2)$$

Here r_{is} is the exposure for individual i in stratum s and C_s is the set of cases in stratum s . Suppose there are n_{0s} controls and n_{1s} cases in stratum s . Then the sum in the denominator is over the $(n_{0s} + n_{1s})!/(n_{0s}!n_{1s}!)$ possible selections of n_{1s} cases from $n_{0s} + n_{1s}$ individuals, and C_s^p indexes the cases for selection p . This likelihood, which is appropriate for use in the analysis of a case-control study, is the result of an argument that conditions on the set of exposures in each stratum [11,

12]. The conditioning eliminates the baseline parameters α_s , saving computation and avoiding the bias that can arise when maximum likelihood is used with very many nuisance parameters.

3 Correcting for measurement error

Carroll *et al.* [5] provide a good review of the general methodology for measurement error in nonlinear regression, including binary logistic regression. One of the methods described there, regression calibration, is attractive because of its simplicity, and appears to work well in many situations.

The basic idea of regression calibration is to replace the unobserved true explanatory variable in the regression by its expectation given its measured value. This requires some modelling of both the explanatory variable and the measurement process. Here, the log-normal population distribution of residential radon concentrations and the log-normal multiplicative measurement errors are used to derive the expectation of true radon exposure given measured exposure. In the case of a simple linear regression, regression calibration would reproduce the correct regression function. For a nonlinear one the regression function is only approximately correct, but the approximation is often a good one. Rosner *et al.* [13] use this approach for logistic regression in an epidemiological context, and both Lagarde *et al.* [14] and Wang *et al.* [15, 3] use it to analyse studies of residential radon and lung cancer, employing the model in (1).

In the case of logistic regression, Reeves *et al.* [16] exploit the similarity between the logistic and probit functions to derive an approximation to the true regression on

the measured covariate that is better than the regression calibration approximation. Their correction to the exposure variable depends on the unknown parameters of the regression, but the fitting can be implemented using a simple iterative approach. This method was used in the original analysis of one of the 13 European radon studies [17]. One limitation of the method is that it does not apply, in general, to the conditional likelihood. Reeves *et al.* do not give any results for the linear odds model (1), but one of their results is extended this model in the appendix to this paper.

4 Integrating the likelihood

Consider a population in which three variables, x, y, z , are associated with each individual. These correspond here to true radon exposure, disease status and measured radon exposure, though the argument below is general. In the population

x has distribution $p(x)$,

y given x has distribution $p(y|x, \theta)$ depending on an unknown parameter θ but not on z , and

z given x has distribution $p(z|x)$ independently of y and not involving θ .

Suppose a random sample is taken from the population and y and z are observed but x is not. Then the likelihood for θ may be obtained by integrating x from the joint probability of observing all three variables, thus

$$p(y, z|\theta) = \int p(y, z, x|\theta) dx$$

$$\begin{aligned}
&= \int p(y|z, x, \theta)p(x|z, \theta)p(z|\theta)dx \\
&= \int p(y|x, \theta)p(x|z)p(z)dx.
\end{aligned}$$

In passing from the second to the third line the conditioning on z in $p(y|z, x, \theta)$ may be omitted because y is independent of the measured variable z given the true value x . Similarly, the conditioning on θ is dropped from the other two terms because neither the measurement process nor the marginal distribution of x depends on θ . Finally $p(z)$, which can be taken outside the integral and does not involve θ , is dropped, leaving

$$p(y, z|\theta) \propto \int p(y|x, \theta)p(x|z)dx. \quad (3)$$

The right hand side of (3) is the integral with respect to the distribution of the unobserved true explanatory variable x given the measured one z of the likelihood for the regression of y on x .

The method of regression calibration may be seen as a single-point approximation to this integral, $p(y|x_{rc}, \theta)$, with $x_{rc} = E(x|z)$. The alternative approach studied here involves direct evaluation of (3) by Monte Carlo integration.

The derivation above of (3) assumed a random sample from a population. Here (3) will be used in the analysis of case-control studies, with $p(y|x, \theta)$ replaced by the conditional likelihood (2). This needs some justification.

The use in the analysis of case-control studies of the logistic likelihood that would be appropriate for a cohort study is common practice. It has been shown [11, 18, 19] that this leads to correct inferences for slope parameters, though the intercepts will have different interpretations. In particular, Carroll *et al.* [19] consider the effect of measurement error as well as other complicating factors in this context and

conclude that even in the presence of these complications it will not be misleading to analyse as though the data were from a cohort study. Farewell [11] gives a specific justification of the use of the conditional likelihood for data from a case-control study. These arguments carry through equally well for the linear odds model (1) and its conditional likelihood (2).

An examination of Farewell's conditioning argument shows that it cannot provide a formal justification for the use of the conditional likelihood in (3). The conditional likelihood is a ratio of two probabilities, which should be integrated separately and then divided, whereas it is the ratio that will be integrated in (3). An informal justification might be made along the lines that the conditional likelihood captures the marginal information about β . If it actually was a marginal likelihood, obtainable by integrating out the intercept parameters over some suitable prior distribution, one could justify the procedure here formally. It seems unlikely that there is any prior that gives exactly the conditional likelihood, but it may still be reasonable to treat it as though there was.

Thus to analyse the case-control studies (3) will be replaced by

$$\ell(\beta, y, z) = \int \ell_c(\beta, y, x)p(x|z)dx \quad (4)$$

where $\ell_c(\beta, y, x)$ is the conditional likelihood in (2). Then the regression calibration approximation becomes $\ell_c(\beta, y, x_{rc})$ and the challenge for the Monte Carlo integration is to evaluate (4). This is made easier by the fact that the resulting likelihood is a function of a single parameter β , allowing it to be tabulated. However, if β is one dimensional, x is certainly not. The single integral sign in (4) conceals the fact that the integral is over as many dimensions as there are individuals in the data.

The next section discusses the estimation of this integral.

5 Implementation of Monte Carlo Integration

The approach to estimating β is to tabulate its likelihood on a one-dimensional grid over a suitably chosen range. For each value of β the likelihood $\ell(\beta, y, z)$ is calculated by Monte Carlo evaluation of the integral in (4). A point estimate and confidence intervals for β may then be derived directly from the likelihood.

The most straightforward way to carry out the Monte Carlo integration would be to generate repeated samples from the distribution $p(x|z)$, evaluate the conditional likelihood $\ell_c(\beta, y, x)$ for each sample, and average these conditional likelihoods. However there are two ways in which it is possible to improve on this approach.

5.1 Factorisation by strata

The conditional likelihood (2) factorises into a product of independent contributions, one from each stratum. When $p(x|z)$ also factorises, which it usually will do, the multiple integral in (4) may be written as a product over strata of an integral for each stratum. For a given number of Monte Carlo samples, a much more accurate result will be obtained by averaging for each stratum separately the contributions to the conditional likelihood and then multiplying these averages than it would by multiplying and then averaging. This is because treating the strata separately reduces the dimensionality of the integrals being estimated from the size of the entire study to the size of a single stratum. The conditional likelihood approach is typically used for data sets that have many strata, each containing a modest number

of individuals. The dimension reduction is crucial to the feasibility of the Monte Carlo integration in this situation.

5.2 More efficient sampling

For Monte Carlo integration to be efficient it is preferable that the samples generated should be concentrated in the region where the integrand is non-negligible. To achieve that here it may be desirable to sample x from a density $q(x)$ different to $p(x|z)$. Rewriting the integral in (4) as

$$\ell(\beta, y, z) = \int \left\{ \ell_c(\beta, y, x) \frac{p(x|z)}{q(x)} \right\} q(x) dx, \quad (5)$$

shows that one then needs to compute for each sample from $q(x)$ the term in $\{ \}$ and average this over samples. Though the dependence is not explicitly shown, $q(x)$ may depend on z or even on y . In particular it may be helpful to sample from different distributions for cases and controls, because $\ell_c(\beta, y, x)$ as a function of x for fixed $\beta > 0$ and y attains its largest values when the cases have greater x than the controls, and will be relatively small for most samples when $p(x|z)$ is used for both cases and controls. This idea is used in the example described below and is discussed further in that context.

5.3 Computation

The computations for the example below were carried out partly using Epicure [20] and partly using MATLAB. The original intention had been to use Epicure for all the computations with the conditional likelihood, but the program was never intended to be embedded in a Monte Carlo simulation, and proved to be unacceptably slow for

such use. Therefore the algorithm of Gail *et al.* [21] was programmed in MATLAB and checked against results from Epicure. One point of note in the computations is that it is the likelihoods, not the log likelihoods, that are averaged, and care is needed to avoid overflows or underflows in these.

5.4 Some comments

The idea of integrating the likelihood is not new. Carroll et al. [5] (Section 7.9.1) describe an approach due to McFadden [22] in which the likelihood is integrated numerically at each step in an iterative maximum likelihood procedure for multiple parameters. What is different here is that because there is only one parameter, β , it is possible to tabulate the likelihood, simplifying the whole procedure considerably.

An alternative procedure that might appear tempting would be to compute the maximum likelihood estimate $\hat{\beta}_{ml}$ for each Monte Carlo sample from $p(x|z)$ and use the mean and standard deviation of the distribution of these values to estimate β and the extra variability due to the measurement error. However, the average of the locations of the maxima of the individual likelihoods, which is what this procedure uses to estimate β , can be very far from the location of the maximum of the averaged likelihoods. Theoretical calculations in the case of simple linear regression suggest that the average of the $\hat{\beta}_{ml}$ obtained by this alternative procedure will be almost the same as if the measured values z had been used in the regression. Theory also suggests that a weighted average, using the values of the likelihood at the maximum as weights, might give a reasonable point estimate for β . Some computations with the linear odds model confirm that these results hold for this model also. The

approach of tabulating the integrated likelihood seems preferable though, because it provides a simple way of quantifying the uncertainty.

6 Some results from the European collaborative analysis

One analysis of the European data was carried out using measured radon concentrations with no adjustment for measurement error. In this analysis the value used for a home for which no measurement was available was an estimate of the mean measured radon concentration in the population of homes in the region, or in some cases a subregion, of the study. The regression coefficient β was estimated as 0.00084, with a 95% likelihood-based confidence interval of (0.00030, 0.00158) [2, 9]. This value of β corresponds to an increase of 8.4% in the risk of lung cancer per 100 Bq/m³ increase in radon concentration. Correcting for measurement error using the integrated likelihood approach nearly doubled the point estimate to 0.0016, and changed the confidence interval to (0.0005, 0.0031). Correcting using regression calibration gave virtually identical results. The implementation of these corrections is described in detail elsewhere [2]. Essentially the same procedures are followed in the artificial example of the next section.

7 An artificial example

A possible conclusion from the results described above is that since both approaches agree they are both working well. In order to increase the confidence in this conclu-

sion an artificial example was constructed to mimic at least the main features of the radon studies. For this artificial data set the true radon exposures and the true value of β are known, making it possible to judge the performance of correction methods.

7.1 Simulation of the data

Populations were generated, separately for each of two areas, as follows. Each individual was assumed to have lived in two homes, with equal lengths of time in each home. A true radon value for each home was generated as $\exp(x)$, with x sampled independently from $N(4.25, 0.64)$ in area 1 and $N(5.25, 0.64)$ in area 2. Given the two radon values $\exp(x_1)$ and $\exp(x_2)$ for an individual, a binary disease status y was generated with

$$p(y = 1) = \frac{\alpha(1 + \beta r)}{1 + \alpha(1 + \beta r)},$$

where the exposure variable $r = \{\exp(x_1) + \exp(x_2)\}/2$ and the two constants are $\alpha = 0.0023$ and $\beta = 0.008$.

This process generates populations with incidence rates of 0.4% in area 1 and 0.7% in area 2. The populations were generated to be large enough so that data for two case-control studies, one for each area, could be constructed by random sampling of 500 diseased cases and 1000 non-diseased controls. Within each study the 1500 observations were randomly split into 60 strata of 25 observations each, the resulting numbers of cases per stratum varying from 3 to 14. The combined sample of 3000 observations in 120 strata forms the artificial data set. Of course it would be possible to get 500 cases without generating such large populations by increasing α very substantially, but this would lead to a quite different distribution of radon

values for the cases in the resulting data.

The final step was to add measurement errors sampled independently from $N(0, 0.25)$ to the log-radon values x for each house to give a measured log-radon z , and to flag some of these measurements as missing. Each measurement was flagged independently with probability 0.2, resulting in 20% missing measurements overall and, as expected, 4% of individuals with missing measurements for both their houses.

The main differences between the artificial example and the collaborative analysis are that the artificial example is smaller, with 3000 rather than 21 000 individuals, and the effect is larger, with $\beta = 0.008$ compared to the estimate of 0.0016. The net effect of these two differences is that β is estimated with roughly the same relative precision in the two data sets. The radon distributions and the overall structure are very similar. The result derived in the appendix (see also the discussion section) suggests that the fact that the effect is larger in the artificial example should make it more of a challenge for the regression calibration method.

7.2 Analysis with true and measured values

The artificial data were first analysed using the conditional likelihood (2), first with $r = \{\exp(x_1) + \exp(x_2)\}/2$, the true radon, and then with $m = \{\exp(z_1) + \exp(z_2)\}/2$, the measured radon, as the explanatory variable, in each case with no missing observations. Figure 1 shows the two log likelihoods, each with its maximum value subtracted. Maximum likelihood estimates and 95% likelihood confidence intervals, the latter derived from the intersection of the log likelihoods with the horizontal line

at 1.92 in Figure 1, are given in the first two rows of Table I.

***** Insert Figure 1 and Table I around here *****

The discrepancy between the maximum likelihood estimate of 0.0062 based on the true radon concentrations and the value of 0.0080 used to generate the data set is well within sampling variability. The attenuation due to measurement error is considerable: both estimate and confidence interval are shrunk by a factor of around 0.5.

7.3 Regression calibration

Implementing the regression calibration approach requires the conditional expectation of the true exposure variable given the measured one. Here the true log radon concentrations x in area A are drawn from $N(\mu_A, \sigma_A^2)$ for $A = 1, 2$, and the corresponding measurement z has distribution $N(x, \sigma_m^2)$. Combining these two leads to the distribution of $x|z$ as $N(\mu_{x|z}, \sigma_{x|z}^2)$ with

$$\mu_{x|z} = \left(\frac{1}{\sigma_m^2} + \frac{1}{\sigma_A^2} \right)^{-1} \left(\frac{z}{\sigma_m^2} + \frac{\mu_A}{\sigma_A^2} \right) \quad (6)$$

and

$$\sigma_{x|z}^2 = \left(\frac{1}{\sigma_m^2} + \frac{1}{\sigma_A^2} \right)^{-1}.$$

Then the true radon $\exp(x)$ for a particular home, given the measurement z on that home, has expectation $\exp(\mu_{x|z} + 0.5\sigma_{x|z}^2)$, and to apply the regression calibration method we use these values to construct the explanatory variable. Homes for which there is no measurement are covered by the same formulae if we let $\sigma_m^2 \rightarrow \infty$, so that $\mu_{x|z} = \mu_A$ and $\sigma_{x|z}^2 = \sigma_A^2$. This corresponds, though now using true rather than

measured radon, to the way such homes were treated in the collaborative analysis without correction for measurement error, since $\exp(\mu_A + 0.5\sigma_A^2)$ estimates the mean of the distribution of true radon concentrations in area A.

Implementing the above requires estimates of the measurement error variance σ_m^2 and the means μ_A and variances σ_A^2 of the distributions of true log radons in each area. For σ_m^2 we used the correct value 0.25. In the collaborative analysis measurement error variances were estimated from the replicated measurements made in several of the studies. To estimate the other parameters we used, as in the collaborative analysis, the measured log radons in the data set. This estimation is complicated by the fact that we have separate random samples of cases and controls, but μ_A and σ_A^2 correspond to the overall populations in the two areas. The first two rows of Table II show the observed means and variances of the non-missing measured log radons, separately for cases and controls, in areas 1 and 2. The variances in the columns headed ‘Corrected’ have had the measurement error variance $\sigma_m^2 = 0.25$ subtracted from them so that they, like the means, are unbiased estimates of the corresponding population parameters for true log radon values.

***** Insert table II around here *****

The third row of Table II combines the separate case and control estimates to produce estimates for the populations from which the case and control samples were drawn. This requires independent estimates of the incidence rates ϕ in the populations, taken to be 0.04 for area 1 and 0.07 for area 2 here. Then

$$\mu_{pop} = \phi\mu_{case} + (1 - \phi)\mu_{control}$$

and

$$\sigma_{pop}^2 = \phi^2 \sigma_{case}^2 + (1 - \phi)^2 \sigma_{control}^2 + \phi(1 - \phi)(\mu_{case} - \mu_{control})^2.$$

The effect of the low incidence rates and the modest differences between radon concentrations in case and control groups is that the result is the same, to the accuracy used, as simply using the mean and corrected variance from the control sample to estimate μ_A and σ_A^2 . This is what was done in the collaborative analysis.

Using these data-based estimates for the parameters involved in $p(x|z)$ introduces an extra source of variability into the procedure, and induces various dependencies between quantities treated as independent in some of the arguments above. Because the estimates are based on large samples, both the extra variability and the dependencies are small, and they will be neglected.

The result of applying the regression calibration method to the artificial data set is the log likelihood shown as a dashed line in Figure 2. For comparison, the solid line in Figure 2 is the log likelihood using true radon values, already seen in Figure 1. The maximum likelihood estimate and likelihood-based 95% confidence interval using regression calibration are given in the third row of Table I.

***** Insert Figure 2 around here *****

The correction recovers almost exactly the point estimate obtained with the true explanatory variable, as it is designed to do. The spread of the log likelihood, and in consequence the width of the confidence interval, are increased compared with that derived from the true values. This is appropriate, because there is now extra uncertainty due to the measurement error. In Section 7.5 the coverage of the interval is investigated in some further simulations.

7.4 Integrating the likelihood

To integrate the likelihood the same distributions of x given z were employed as in the regression calibration approach, using the same estimates for population means and variances. A naive approach, sampling from $p(x|z)$, and a more efficient sampling strategy, described in detail below, were both investigated. In both cases the likelihoods were averaged separately for each of the 120 strata and the resulting averages multiplied.

To make the sampling more efficient the idea described in Section 5.2 was used. Controls were still sampled from $p(x|z)$. For cases, the case means, m_{case} , with values 4.30 and 5.52 for areas 1 and 2 (Table II, row 1), were used instead of the population means, m_{pop} , with values 4.26 and 5.26 (Table II, row 3), as substitutes for μ_A in calculating $\mu_{x|z}$ via (6). The variance estimates, v_{pop} , of 0.58 and 0.68 were not changed. Radon concentrations for homes associated with cases were then sampled from this modified form of $p(x|z)$. With this choice for the density q in (5), the factor $p(x|z)/q(x)$ in (5) is 1 for controls but can be shown after some algebra to receive a multiplicative contribution of

$$\exp \left\{ \left(\frac{m_{pop} - m_{case}}{v_{pop}} \right) x \right\} \quad (7)$$

for each x sampled for a home associated with a case. It is worth emphasizing that no change has been made to the integral being computed, which is still (4) with $p(x|z)$ based on the overall population radon distribution. The distribution of radon for cases is introduced only to obtain samples that give higher likelihoods, and the effect of the change is exactly compensated by the factors in (7). Some preliminary comparisons showed that this approach gave a worthwhile reduction in

the variance of the Monte Carlo integration results, and the results reported below employ it.

The dotted curve in Figure 2 is the log of the integrated likelihood averaged over ten thousand samples. The resulting maximum likelihood estimate $\hat{\beta}$ and confidence interval (L, U) are given in the last row of Table I. To give some idea of the extent of convergence, Figure 3 shows ten log likelihoods, each using one thousand of the ten thousand samples, and each with its maximum value subtracted. The ten estimates of $\hat{\beta}$, L and U have relative standard deviations of around 1% in each case. The upper confidence limits U , which are the most variable in absolute terms, range from 0.0131 to 0.0136. This is acceptable accuracy in this context.

***** Insert Figure 3 around here *****

7.5 Repeated sampling behaviour of the regression calibration estimate and confidence interval

A further simulation was carried out to investigate the behaviour under repeated sampling of the point estimate and confidence interval given by regression calibration. Fixing the strata, the case-control status, and the missing-value indicators in the artificial data set, 10 000 such sets were generated by repeated simulation of the true and measured radon values from the appropriate distributions. Point estimates and likelihood confidence intervals were computed for each data set using the true radon values and by applying the regression calibration method to the measured ones. Including the integrated likelihood approach in this simulation would have been impractical in terms of computer time.

Using the true radon values, the 10 000 point estimates of β , whose value is known here to be 0.008, had sample mean 0.0085, median 0.0080 and coefficient of variation 0.31 and were distributed lognormally. The 10 000 point estimates derived from the measured radons using regression calibration had sample mean 0.0088, median 0.0080 and coefficient of variation 0.42 and were also distributed lognormally.

On a log scale, both samples of estimates of β had mean and median $\log(0.0080)$ making them unbiased estimators of $\log(\beta)$. The biases on the untransformed scale are due to the skewed sampling distributions, and the difference in the biases is a consequence of the difference in the coefficients of variation. The increase in the coefficient of variation (or the standard deviation on the log scale) of the samples of point estimates from 0.31 when the true radon values are used to 0.42 in the case of regression calibration is an indication of the impact of the measurement errors and missing values on the inference.

The initial artificial data set, for which the results are given in Table I, was not specially selected, but it is clear that such perfect agreement between the regression calibration $\hat{\beta}$ and that using true radon values was fortuitous. With measurement errors as large as they are here there is substantial variability in both estimators of β and in the differences between them. However the simulations confirm that there is no systematic error in the regression calibration estimate.

The results for the coverage of the 95% likelihood confidence intervals also confirm the correct performance of the regression calibration approach. In 10 000 repetitions the true $\beta = 0.008$ should be excluded 250 times at each end of the interval,

with the observed numbers of exclusions having a standard deviation of 16, derived from the binomial distribution with $N = 10\,000$ and $p = 0.025$. The observed numbers of exclusions at the lower and upper ends respectively were 257 and 243 using the true radons, an overall coverage of exactly 95%, and 258 and 267 using the measured radons and regression calibration, giving an overall coverage of 94.75%. All these results are consistent with correct coverage properties.

8 Discussion

The initial motivation for implementing the integrated likelihood approach was to improve on regression calibration for the collaborative analysis of the 13 radon studies. When, after a great deal of computation, the two approaches gave virtually identical results, the further investigations described here were carried out. For the artificial example, which copies the key features of the radon studies and for which the two approaches also give very similar results, it seems that there is little scope for improving on regression calibration. Not only does it give good point estimates but, perhaps more surprisingly, the coverage properties of the resulting confidence interval are also correct. We conclude that there is no need for any more sophisticated approach in the collaborative analysis.

Should it have been obvious a priori that regression calibration was adequate? One way of assessing this for the linear odds model is to examine the values of the factor f in equation (11) of the appendix. If these are close to 1 for all observations, it suggests that regression calibration may perform well. For the artificial example the values of f are mostly close to 1, with the most extreme values reaching only

1.1. Thus the fact that regression calibration produces such good point estimates of β might have been predicted. However the remarkable accuracy of the confidence intervals is still somewhat surprising.

Though the integrated likelihood approach is not needed for the radon studies, regression calibration will not always work so well. For example, some further limited simulations with larger values of β suggest that the integrated likelihood and regression calibration begin to diverge as β increases, as one would expect from examination of (11). When it is necessary, integrating the likelihood is a feasible approach, even with quite large datasets.

Appendix

When the exposure variable in the linear-odds model (1) is $r = \exp(x)$, and the distribution of the true x given the observed z is normal, it is possible to derive a good approximation to the form of $p(y = 1|z)$, thus extending the results of Reeves *et al.* [16] to this case.

Dropping the subscript s and expressing the probability in terms of x gives

$$p(y = 1|x) = \frac{\alpha(1 + \beta e^x)}{1 + \alpha(1 + \beta e^x)} , \quad (8)$$

which can be written in the form

$$p(y = 1|x) = \frac{\alpha}{1 + \alpha} + \frac{1}{1 + \alpha} \Lambda(\theta + x) \approx \frac{\alpha}{1 + \alpha} + \frac{1}{1 + \alpha} \Phi(k\{\theta + x\}) , \quad (9)$$

where $\Lambda(t) = e^t/(1 + e^t)$ is the logistic function, $\theta = \ln\{\beta\alpha/(1 + \alpha)\}$, $\Phi(\cdot)$ is the cdf of the standard normal distribution, and $k = 0.588 = 1/1.70$ is the constant in the well-known [23] approximation $\Lambda(t) \approx \Phi(kt)$.

If $x|z$ has a normal distribution with mean z^* and variance σ^2 it is possible to integrate the approximate form (9) over this distribution to give

$$p(y = 1|z) \approx \frac{\alpha}{1 + \alpha} + \frac{1}{1 + \alpha} \Phi \left(\frac{k\{\theta + z^*\}}{\sqrt{1 + k^2\sigma^2}} \right) .$$

Reversing the initial manipulations, this becomes

$$p(y = 1|x) \approx \frac{\alpha(1 + f\beta e^{z^*})}{1 + \alpha(1 + f\beta e^{z^*})} , \quad (10)$$

where the factor f is given by

$$f = \left(\frac{\alpha}{1 + \alpha} \beta e^{z^*} \right)^{-\delta} , \quad (11)$$

with $\delta = 1 - (1 + k^2\sigma^2)^{-0.5}$. The only approximation in this argument is the use of $\Lambda(t) \approx \Phi(kt)$, and this is known to be good except in the extreme tails [23].

The expression in (10) is the equivalent for this model of equation (20) of Reeves *et al.* for the logistic model. It shares with the logistic case the problem that the intercept parameters do not cancel if this form is used in the conditional likelihood, making it difficult to exploit this result in the analysis here. However, when f is close to 1, (10) is close to (8) with x replaced by $z^* = E(x|z)$. Thus, examination of the size of f may suggest when regression calibration will be adequate for the linear odds model.

References

1. United Nations Scientific Committee on the Effects of Atomic Radiation.
Sources and Effects of Ionizing Radiation. *UNSCEAR 2000 Report to the*

General Assembly, with Scientific Annexes. Vol. I: Sources, United Nations: New York, 2000.

2. Darby S, Hill D, Deo H , Auvinen A, Barros-Dios JM, Baysson H, Bochicchio F, Falk R, Farchi S, Figueiras A, Hakama M, Heid I, Hunter N, Kreienbrock L, Kreuzer M, Lagarde F, Mäkeläinen I, Muirhead C, Oberaigner W, Pershagen G, Ruosteenoja E, Schaffrath Rosario A, Tirmarche M, Tomášek L, Whitley E, Wichmann HE, Doll R. Residential radon and lung cancer: detailed results of a collaborative analysis of individual data on 7,148 subjects with lung cancer and 14,208 subjects without lung cancer from 13 epidemiological studies in Europe. *Scandinavian Journal of Work, Environment and Health*, in press.
3. Lubin JH, Wang ZY, Wang LD, Boice JD Jr, Cui HX, Zhang SR, Conrath S, Xia Y, Shang B, Cao JS, Kleinerman RA. Adjusting lung cancer risks for temporal and spatial variations in radon concentration in dwellings in Gansu Province, China. *Radiation Research* 2005; **163**:571–579.
4. Fuller WA. *Measurement Error Models*, Wiley: New York, 1987.
5. Carroll RJ, Ruppert D, Stefanski LA. *Measurement Error in Nonlinear Models*, Chapman and Hall: London, 1995.
6. Krewski D, Lubin JH, Zielinski JM, Alavanja M, Catalan VS, Field RW, Klotz JB, Létourneau EG, Lynch CF, Lyon JI, Sandler DP, Schoenberg JB, Steck DJ, Stolwijk JA, Weinberg C, Wilcox HB. Residential radon and risk of lung cancer. A combined analysis of 7 North American case control studies. *Epidemiology* 2005; **16**:137–145.

7. Lubin JH, Wang ZY, Boice JD Jr, Xu ZY, Blot WJ, Wang LD, Kleinerman RA. Risk of lung cancer and residential radon in China: pooled results of two studies. *International Journal of Cancer* 2004; **109**:132–137.
8. Heid IM, Küchenhoff H, Wellmann J, Gerken M, Kreienbrock L, Wichmann HE. On the potential of measurement error to induce differential bias on odds ratio estimates: an example from radon epidemiology. *Statistics in Medicine* 2002; **21**:3261–3278.
9. Darby S, Hill D, Auvinen A., Barros-Dios, JM, Baysson H, Bochicchio F, Deo H, Falk R, Forastiere F, Hakama M, Heid I, Kreienbrock L, Kreuzer M, Lagarde F, Mäkeläinen I, Muirhead C, Oberaigner W, Pershagen G, Ruano-Ravina A, Ruosteenoja E, Schaffrath Rosario A, Tirmarche M, Tomášek L, Whitley E, Wichmann HE, Doll R. Radon in homes and the risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. *British Medical Journal* 2005; **330**:223–226.
10. Preston DL, Shimizu Y, Pierce DA, Suyama A, Mabuchi K. Studies of mortality of atomic bomb survivors. Report 13: Solid cancer and non-cancer disease mortality, 1950-1997. *Radiation Research* 2003; **160**:381-407.
11. Farewell, VT. Some results on the estimation of logistic models based on retrospective data. *Biometrika* 1979; **66**:27–32..
12. Prentice RL, Pyke, R. Logistic disease incidence models and case-control studies. *Biometrika* 1979; **66**:403–411.

13. Rosner B, Willett WC, Spiegelman, D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine* 1989; **8**:1051–1069.
14. Lagarde F, Pershagen G, Åkerblom G, Axelson O, Båverfält U, Damberg L, Enflo A, Svartengren M, Swedjemark GA. Residential radon and lung cancer in Sweden: risk analysis accounting for random error in the exposure assessment. *Health Physics* 1997; **72**:269–276.
15. Wang Z, Lubin JH, Wang L, Zhang S, Boice JD Jr, Cui H, Zhang S, Conrath S, Xia Y, Shang B, Brenner A, Lei S, Metayer C, Cao J, Chen K, Lei S, Kleinerman RA. Residential radon and lung cancer risk in a high-exposure area of Gansu Province, China. *American Journal of Epidemiology* 2002; **155**:554–564.
16. Reeves GK, Cox DR, Darby SC, Whitley E. Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statistics in Medicine* 1998; **17**:2157–2177.
17. Darby SC, Whitley E, Silcocks P, Thakrar B, Green M, Lomas P, Miles J, Reeves G, Fearn T, Doll, R. Risk of lung cancer associated with residential radon exposure in south-west England: a case-control study. *British Journal of Cancer* 1998; **78**:394–408.
18. Breslow NE, Day NE. *Statistical Methods in Cancer Research. Vol. 1 - The Analysis of Case-Control Studies*. IARC: Lyon, 1980.

19. Carroll RJ, Wang S, Wang CY. Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association* 1995 **90**:157–169.
20. Preston DL, Lubin JH, Pierce DA, McConney ME. *Epicure: Release 2.10*. HiroSoft International Corporation: Seattle, WA, 1998.
21. Gail MH, Lubin JH, Rubinstein LV. Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika* 1981; **68**:703–707.
22. McFadden D. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* 1989; **57**:239–265.
23. Johnson NL, Kotz S. *Continuous Univariate Distributions. Vol. 2*. Wiley: New York, 1970.

Table I. Point estimates, $\hat{\beta}$ and the lower, L , and upper, U , limits of 95% likelihood-based confidence intervals for the regression parameter in the artificial example of Section 7.1.

Analysis	$\hat{\beta}$	L	U
True radon values	0.0062	0.0035	0.0109
Measured radon, no correction	0.0034	0.0019	0.0057
Regression calibration	0.0059	0.0027	0.0127
Integration	0.0061	0.0028	0.0134

Table II. Means and variances of non-missing measured log radons in each area, separately for cases and controls and combined to produce estimates for the true log radons in the overall population.

	Area 1			Area 2		
	Mean	Variance		Mean	Variance	
		Raw	Corrected		Raw	Corrected
Cases	4.30	0.97	0.72	5.52	0.91	0.66
Controls	4.26	0.83	0.58	5.26	0.93	0.68
Overall	4.26		0.58	5.26		0.68

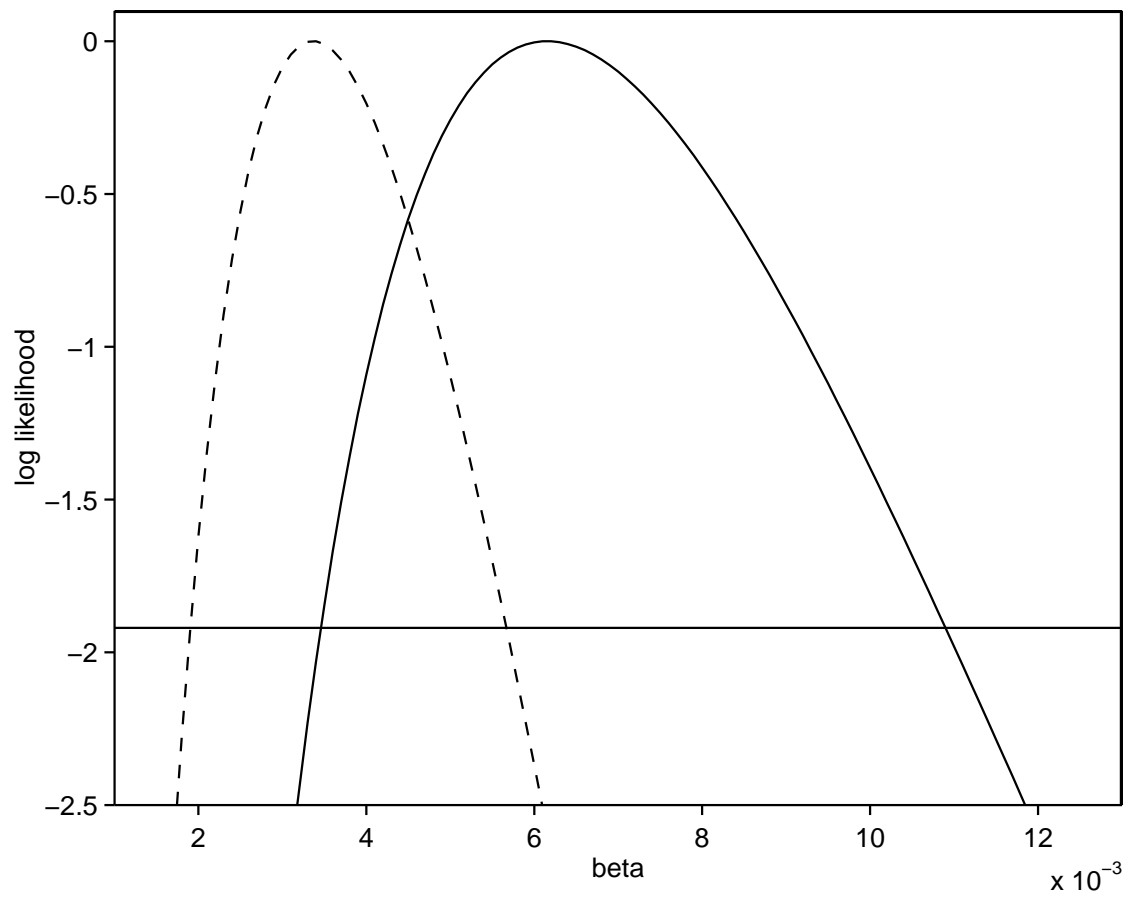


Figure 1. Log likelihoods for β for the artificial example of Section 7.1 using the true (solid line) and the measured (dashed line) radon exposures.

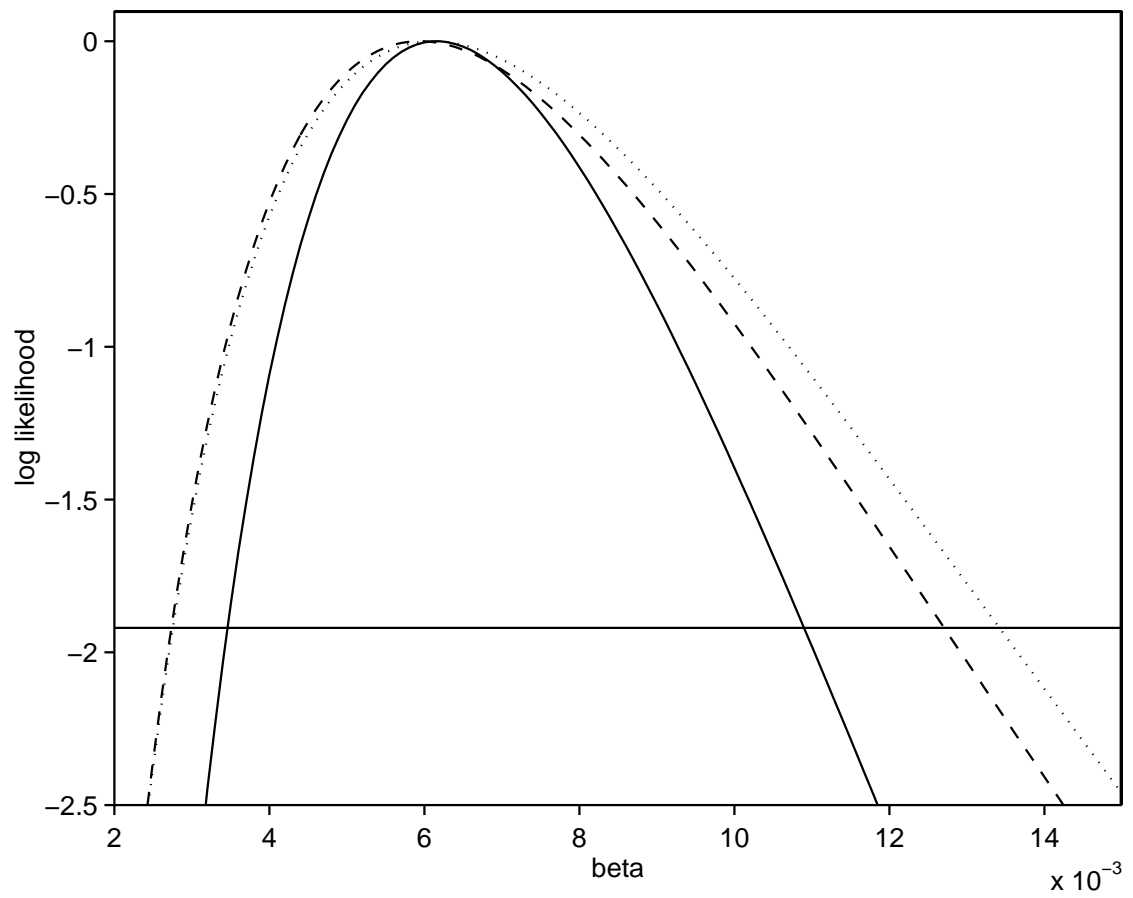


Figure 2. Log likelihoods for β for the artificial example of Section 7.1 using the true radon exposures (solid line) and using two approaches to correct for the effect of measurement error, regression calibration (dashed line) and integrated likelihood (dotted line).

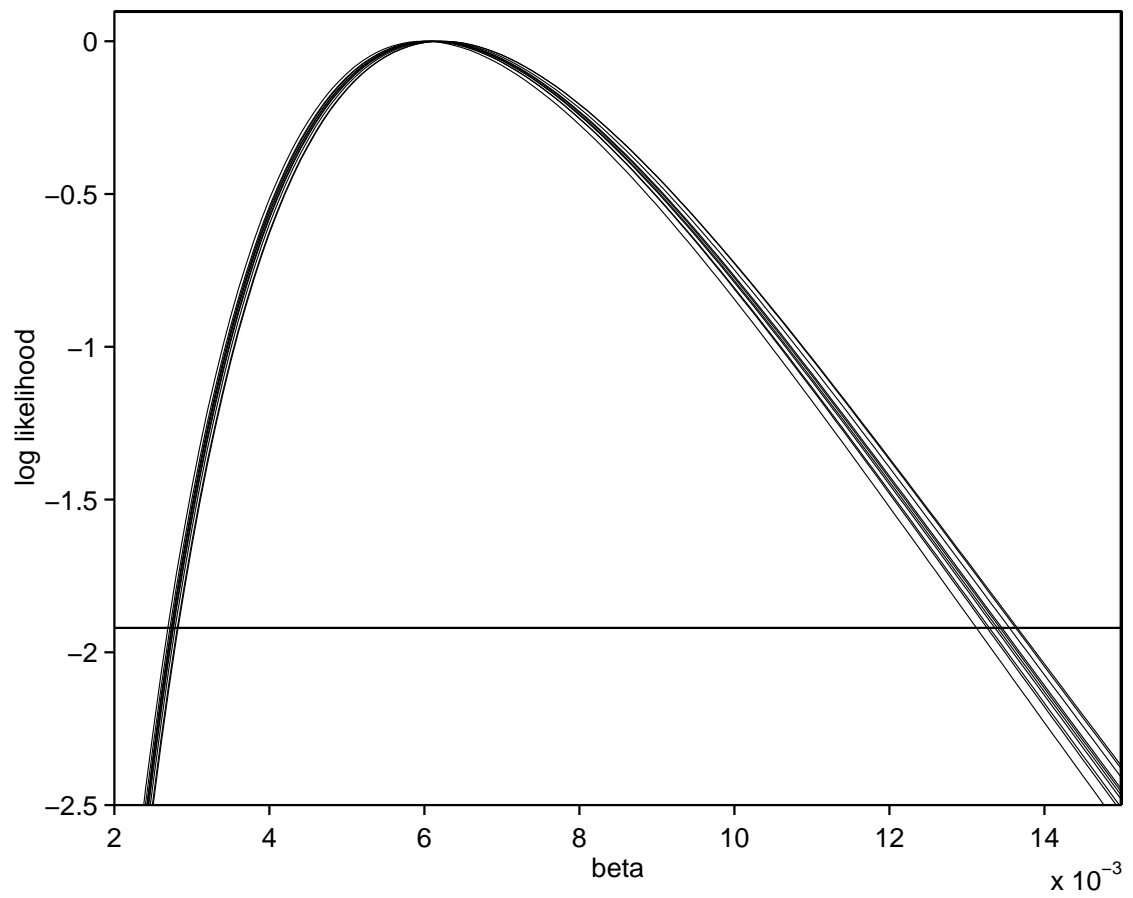


Figure 3. Ten log integrated likelihoods for β for the artificial example of Section 7.1. Each is the result of a separate Monte Carlo integration.