STATISTICS IN MEDICINE
*Statist. Med.* 2003; **22**:1069–1082 (DOI: 10.1002/sim.1388)

# Power and sample size calculations for generalized regression models with covariate measurement error

Tor D. Tosteson[1,*,†], Jeffrey S. Buzas[2], Eugene Demidenko[1] and Margaret Karagas[1]

[1]*Dartmouth Medical School, Lebanon NH 03756, U.S.A.*
[2]*University of Vermont, Burlington, VT 05401-1455, U.S.A.*

## SUMMARY

Covariate measurement error is often a feature of scientific data used for regression modelling. The consequences of such errors include a loss of power of tests of significance for the regression parameters corresponding to the true covariates. Power and sample size calculations that ignore covariate measurement error tend to overestimate power and underestimate the actual sample size required to achieve a desired power. In this paper we derive a novel measurement error corrected power function for generalized linear models using a generalized score test based on quasi-likelihood methods. Our power function is flexible in that it is adaptable to designs with a discrete or continuous scalar covariate (exposure) that can be measured with or without error, allows for additional confounding variables and applies to a broad class of generalized regression and measurement error models. A program is described that provides sample size or power for a continuous exposure with a normal measurement error model and a single normal confounder variable in logistic regression. We demonstrate the improved properties of our power calculations with simulations and numerical studies. An example is given from an ongoing study of cancer and exposure to arsenic as measured by toenail concentrations and tap water samples. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: measurement error; sample size; generalized linear models; logistic regression

## 1. INTRODUCTION

Measurement error in exposure risk factors has frequently been identified as an important feature of epidemiologic and other scientific studies. A consequence of measurement error for linear regression, logistic regression and other generalized linear models is the loss of power of tests of significance for the relative risk regression parameters corresponding to the true exposures. This loss of power needs to be included in careful assessments of sample size requirements for studies with large exposure measurement errors.

Sample size calculations are often done based on simple planned comparisons of means or proportions. With the increasing use of sophisticated measurement techniques, environmental risk factors are more likely to be continuous variables, so that power calculations cannot be

* Correspondence to: Tor D. Tosteson, Dartmouth Medical School, 7927 Rubin Building DHMC, One Medical Center Drive, Lebanon, NH 03756, U.S.A.
† E-mail: tor.tosteson@dartmouth.edu

Contract/grant sponsor: National Institutes of Health; contract/grant numbers: CA50597, ES07373, CA57494.

*Received November 2001*
*Accepted August 2002*

based solely on the comparison of a small number of group means or proportions. Furthermore, known confounding risk factors need to be included in any assessment of a potential causal relationship between the exposure of interest and a health outcome. To accommodate these features of the data, the final planned analysis for an epidemiologic study typically involves a multiple logistic regression model or other generalized linear model for binary, continuous or counted data.

In this paper we provide a simple and general method for performing sample size and power calculations for regression analyses when the exposure risk factors may be subject to substantial measurement error. Our work is related to previous work on the properties of score tests in generalized linear models in the presence of measurement error [1, 2]. The paper by Tosteson and Tsiatis [1] described the form of the optimal score test in generalized linear models with exposure measurement error and derived expressions for the asymptotic relative efficiency under local alternatives, an approximation appropriate for small alternative relative risk regression parameters. Stefanski and Carroll [2] explored a more general form of the score test incorporating a non-parametric estimate of the measurement error distribution. The current paper extends the results of earlier work on the score test in the presence of exposure measurement error to relax the assumption of local alternatives and to provide a practical implementation for use in performing routine sample size calculations.

Other authors have studied modified sample size calculations for epidemiologic regression models in the presence of measurement error using an assortment of simplifying assumptions and restricted models. McKeown-Eyssen and Tibshirani [3] and Devine and Smith [4] studied regression models with log-linear risk functions and normal distributions for exposures and measurement error, and White *et al.* [5] considered case-control studies with conditionally normal exposures within the case and control groups. Our methods are applicable to a broad class of generalized regression models, including logistic regression, and are not based on small relative risk assumptions or restricted models. The asymptotic theory is based on the properties of the generalized score test under measurement error [6] with fixed rather than local alternatives.

Self and Mauritsen [7] investigated the power function for score tests in generalized linear models, the same fundamental problem considered in this paper but without consideration of exposure measurement errors. They employ local alternative approximations, restricting some of their results to small alternative hypotheses for the exposure regression parameter, and do not consider continuous covariates. A follow-up paper by this group [8] considered improved power functions based on likelihood ratio tests rather than the score test, but with a similar reliance on local alternatives and categorical covariates. Shieh [9] generalized the results for the likelihood ratio tests to include continuous covariates. The problems we consider are similar to those encountered for generalized linear models without measurement error, but our approach differs in that we consider an asymptotic theory appropriate for fixed (larger) alternatives.

## 2. GENERALIZED REGRESSION MODELS WITH MEASUREMENT ERROR

Suppose that $y_i$ $(i=1,\ldots,n)$ is related to a $(p+q)\times 1$ vector of covariates $(\mathbf{x}_i',\mathbf{z}_i')'$ via the regression model

$$E(y_i\,|\,\mathbf{x}_i,\mathbf{z}_i)=f(\beta_0+\boldsymbol{\beta}_z'\mathbf{z}_i+\boldsymbol{\beta}_x'\mathbf{x}_i)$$

 

with variance function

$$\mathrm{var}(y_i|\mathbf{x}_i,\mathbf{z}_i)=\sigma^2 g^2(\beta_0+\boldsymbol{\beta}_z'\mathbf{z}_i+\boldsymbol{\beta}_x'\mathbf{x}_i;\theta)$$

where $\theta$ is a variance function parameter. The covariates $\mathbf{x}_i$ are possibly contaminated with measurement error, whereas the covariates $\mathbf{z}_i$ are observed without error. In typical epidemiologic applications, $y_i$ is an indicator of disease, $\mathbf{x}_i$ contains one or more risk factors subject to non-differential misclassification or continuous measurement error, and $\mathbf{z}_i$ represents an additional confounder or controlling variables to be observed without error.

The measurement error problem arises because we cannot directly observe the 'real' exposure $\mathbf{x}_i$ and must instead make do with the 'surrogate' covariate $\mathbf{w}_i$. For scalar $x$ and $w$, the classical additive measurement error model specifies $w_i=x_i+\sqrt{v}u_i$, where $\mathbf{u}_i$ has a mean of zero, variance of one, and is independent of both $x_i$ and $y_i$. The measurement error is represented by $v$. A book by Fuller [10] discusses this case thoroughly in the context of linear models, and the non-linear case is discussed in the book by Carroll *et al.* [6]. In a simple linear regression, the effect of classical measurement error is a bias towards zero in the regression coefficient obtained when $\mathbf{w}_i$ is substituted for $\mathbf{x}_i$. Thus the effect of measurement error is to underestimate the associated regression coefficient in an absolute sense.

A useful generalization of the additive measurement error model can be obtained by using the conditional independence assumption

$$P_{y,\mathbf{w}|\mathbf{z},\mathbf{x}}(\mathbf{y}_i,\mathbf{w}_i|\mathbf{z}_i,\mathbf{x}_i)=P_{y|\mathbf{z},\mathbf{x}}(\mathbf{y}_i|\mathbf{z}_i,\mathbf{x}_i)P_{w|\mathbf{z},\mathbf{x}}(\mathbf{w}_i|\mathbf{z}_i,\mathbf{x}_i) \tag{1}$$

here $P(\cdot|\cdot)$ is a conditional probability density or mass function. All that is required in this specification is that the surrogate exposure be independent of the outcome given the true exposure. It is therefore possible to consider situations where $\mathbf{w}$ is related to $\mathbf{x}$ in a more flexible way. For example, non-differential misclassification in contingency tables will satisfy this condition. We assume that the conditional independence assumption holds throughout the paper.

## 3. THE POWER OF THE GENERALIZED SCORE TEST

In this section we define the generalized score and score test and discuss the power function of this test for a fixed alternative under measurement error assumptions. Our development assumes that $p=1$, that is, only one covariate is subject to measurement error, and thus $\beta_x$ is a scalar.

Generalized score statistics for testing hypotheses concerning the $(\beta_0,\boldsymbol{\beta}_z',\beta_x)$ are based on the quasi-likelihood score equations. In the presence of measurement error, the generalized score [6] for testing $H_0\colon \beta_x=0$ based on the observed covariates and the surrogate exposure is

$$L(\beta_0,\boldsymbol{\beta}_z)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} d_i(\beta_0,\boldsymbol{\beta}_z,\theta)[y_i-f(\beta_0+\boldsymbol{\beta}_z'\mathbf{z}_i)]E[x_i|\mathbf{z}_i,w_i;\tau]$$

where $d_i(\beta_0,\boldsymbol{\beta}_z,\theta)=f^{(1)}(\beta_0+\boldsymbol{\beta}_z'\mathbf{z}_i)/g^2(\beta_0+\boldsymbol{\beta}_z'\mathbf{z}_i,\theta)$ and $f^{(1)}(x)=(d/\mathrm{d}x)f(x)$.

Our presentation assumes that $\theta$, the variance function parameter, and $\tau$, the parameters of the conditional expectation of the true exposure given the surrogate and other covariates, are known, but the results hold when $\sqrt{n}$ consistent estimators are substituted for these unknown values. The generalized score thus accommodates designs in which a measurement error correction is to be made using replicates or a large external validation study.

The score test statistic is

$$\frac{L^2(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_z)}{\hat{\sigma}^2 \hat{\sigma}_0^2} \tag{2}$$

where $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_z)$ are consistent regression parameter estimates satisfying the $q+1$ non-linear equations

$$\sum_{i=1}^{n} d_i(\beta_0, \boldsymbol{\beta}_z, \theta)[y_i - f(\beta_0 + \boldsymbol{\beta}_z' \mathbf{z}_i)] \begin{pmatrix} 1 \\ \mathbf{z}_i \end{pmatrix} = \mathbf{0}$$

Expressions for the normalizing scalars $\hat{\sigma}^2$ and $\hat{\sigma}_0^2$ are given in the Appendix.

The asymptotic distribution of the score test statistic (2) under the null hypothesis is a central chi-square on one degree of freedom, and a level $\alpha$ test rejects the null hypothesis when the test statistic is greater than $\chi_\alpha^2$, the $(1-\alpha)$ percentile of a chi-square distribution.

For a fixed alternative, a derivation given in the Appendix shows that for large $n$

$$\frac{L^2(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_z)}{\hat{\sigma}^2 \hat{\sigma}_0^2} \sim k(\beta_0, \boldsymbol{\beta}_z', \beta_x)^{-1} \chi^2\{\phi(\beta_0, \boldsymbol{\beta}_z', \beta_x)\} \tag{3}$$

where $\chi^2(\phi)$ represents a chi-square random variable on one degree of freedom with non-centrality parameter $\phi$. The asymptotic power function for the generalized score test under the fixed alternative is given by

$$\Pr(\chi^2(\phi) > k\chi_\alpha^2) \tag{4}$$

The functions $k$ and $\phi$ are defined in the Appendix and can be computed upon specification of the joint distribution of $(\mathbf{z}, x, w)$ and $(\beta_0, \boldsymbol{\beta}_z', \beta_x)$.

Calculations for the asymptotic relative efficiency of score tests in generalized linear models have employed the local alternative condition $\beta_x = O(1/\sqrt{n})$ [1, 11]. This assumption, suitable for small alternatives $\beta_x$, implies that for large $n$

$$\frac{L^2(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_z)}{\hat{\sigma}^2 \hat{\sigma}_0^2} \sim \chi^2(\lambda) \tag{5}$$

where $\chi^2(\lambda)$ represents a non-central chi-square on one degree of freedom and non-centrality parameter

$$\lambda = n\beta_x^2 \sigma_0^2 \sigma^2 / 2$$

The associated power function is

$$\Pr(\chi^2(\lambda) > \chi_\alpha^2) \tag{6}$$

An interesting consequence of (5) given in previous papers on the properties of score tests in the presence of measurement error [1, 11] is that, for simple regression models with linear measurement error, for example, normal exposure and measurement error distributions, the asymptotic relative efficiency (ARE) of the score test using the surrogate $w$ to the score test using the true exposure $x$ under local alternatives is

$$\text{ARE} = n_x/n_w = \rho_{xw}^2 \tag{7}$$

where $n_w$ is the sample size required to achieve a given power using $w$, $n_x$ is the required sample size using $x$, and $\rho_{xw}$ is the correlation between $x$ and $w$. This would suggest that, at least for small alternatives, the increase in the required sample size due to using $w$ instead of $x$ is given by $1/\rho_{xw}^2$.

It turns out that the same inflation factor obtained under local alternatives is suggested by other authors [3–5] studying sample size corrections based upon consideration of log-linear risk functions and conditionally normal distributions for exposures and measurement error. Our new method improves on both the local alternatives and the other approximations for examples of designs and generalized regression models not fitting these assumptions.

## 4. AN IMPLEMENTATION FOR LOGISTIC REGRESSION

Computing power or sample size with the power function defined above requires calculating $\phi$ and $k$ which in turn requires multi-dimensional integral evaluations (expectations) of the intermediate expressions given in the Appendix. These expectations can be evaluated upon specification of the joint distribution of $(\mathbf{z}, x, w)$ and values for $(\beta_0, \boldsymbol{\beta}_z, \beta_x)$.

We have implemented a version of the calculation for continuous predictors in logistic regression where $(\mathbf{z}, x, w)$ are jointly normal and $\mathbf{z}$ is scalar. This program can be used for any measurement error structure satisfying the conditional independence assumption (1), including both the 'classical' and 'Berkson case' measurement error assumptions. For classical measurement error, it is assumed that $w = x + \varepsilon$, where $\varepsilon$ is independent of $x$. In Berkson case error, it is assumed that $x = w + \delta$, where $\delta$ is independent of $w$. For general surrogate measurement error, no assumptions are imposed on the correlations among $z, x$ and $w$.

The researcher must specify the correlation between the true and the surrogate, $\rho_{xw}$, as well as the correlation between $z$ and $x$. For classical measurement error, $w$ and $z$ are conditionally independent given $x$, and the correlation between $w$ and $z$ is given by $\rho_{zw} = \rho_{zx}\rho_{xw}$. For Berkson case measurement error, $x$ and $z$ are conditionally independent given $w$, and the correlation between $w$ and $z$ is given by $\rho_{zw} = \rho_{zx}/\rho_{xw}$. For more general surrogate measurement error models corresponding to neither the classical or Berkson models, $\rho_{zw}$ and $\rho_{xw}$ must be specified independently.

To make the program as flexible as possible, the variables are considered to be standardized to have a mean of zero and a variance of one. $\beta_x$ and $\beta_z$ are specified by providing of the odds ratio for a one standard deviation increase in the true exposure and covariate, respectively. $\beta_0$ is specified as the prevalence of the condition at the means for the covariates, that is, $\beta_0 = \log\{\Pr(y=1|x=0, z=0)/\Pr(y=0|x=0, z=0)\}$. Programs have been written for calculating either sample size required for a given power and significance level or the power achieved for a given sample size and significance level. Figure 1 shows examples of the output generated by a web-based demonstration program implementing the classical case. The current URL is http://biostat.hitchcock.org/MeasurementError/Analytics/SampleSizeCalculationsforLogisticRegression.asp.

 

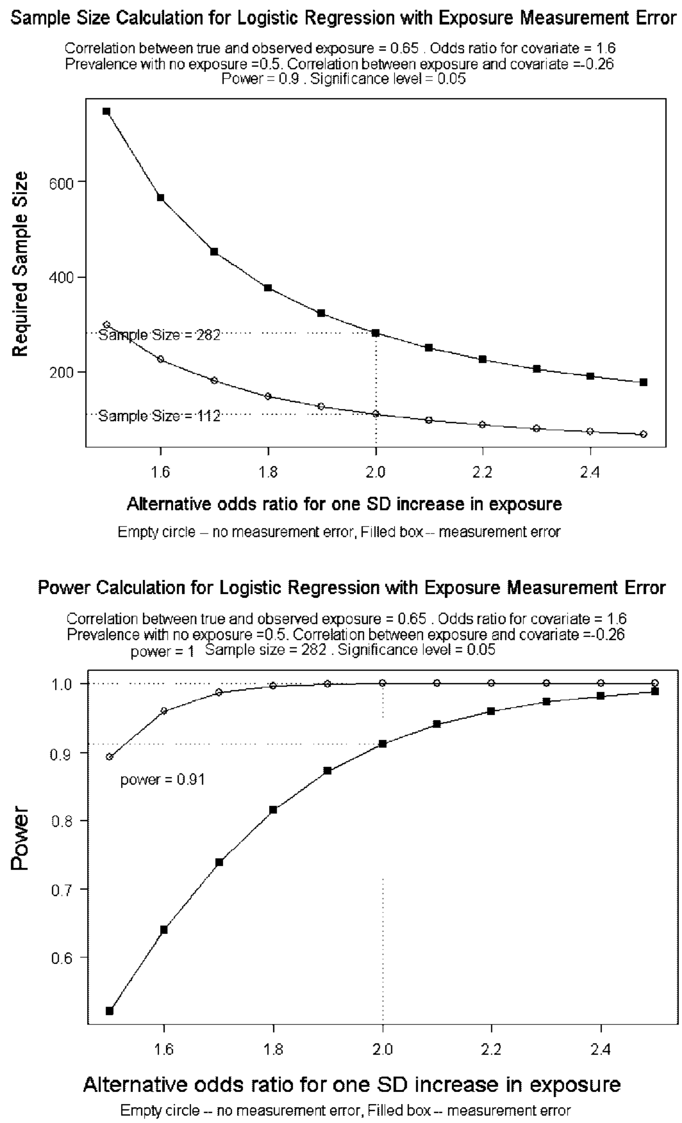


Figure 1. Output from implemented programs for sample size and power calculations for logistic regression with normal measurement error.

## 5. EXAMPLE: ARSENIC EXPOSURE AND SKIN CANCER

Karagas *et al.* [12, 13] describe the design of a case-control study for assessing the risk of bladder and two forms of non-melanoma skin cancer due to arsenic exposure. Controls were frequency matched by age to the combined age distributions of basal and squamous cell cases. Participants in this study are contributing both drinking water samples and toenails for

 

Table I. Total sample size required for a case-control studies of skin cancer, assuming a 1:1 case to control ratio, a two-sided significance level of 0.05, a power of 0.9 and an alternative odds ratio of 1.5 for a one standard deviation increase in toenail concentration.

| Cancer type | All water concentrations $\rho_{xw} = 0.45, \rho_{xz} = 0.20$ | Water $\geqslant 1$ $\mu$g/l $\rho_{xw} = 0.69, \rho_{xz} = 0.20$ |
|---|---|---|
| *Classical measurement error model* | | |
| Basal cell ($\beta_z = \ln 0.85$) | 1479 | 620 |
| Squamous cell ($\beta_z = \ln 1.62$) | 1510 | 633 |
| *Berkson measurement error model* | | |
| Basal cell ($\beta_z = \ln 0.85$) | 1713 | 636 |
| Squamous cell ($\beta_z = \ln 1.62$) | 1749 | 649 |
| *General surrogate measurement error model*, $\rho_{zw} = -0.05$ | | |
| Basal cell ($\beta_z = \ln 0.85$) | 1433 | 598 |
| Squamous cell ($\beta_z = \ln 1.62$) | 1464 | 611 |

measurement of arsenic concentrations. The toenail concentrations are considered to be the best available measure of 'true' exposure. We have based this example on the two forms of skin cancer for which preliminary results have been published and can be used to define the parameters necessary to compute sample size requirements for future studies under measurement error assumptions.

An evaluation of the age-adjusted effects of toenail arsenic on squamous and basal cell skin cancer using the same case-control design has demonstrated possible evidence of an elevation in cancer risk [14, 15]. Based on data for 587 basal cell cancer cases, 284 squamous cell cancer cases and 524 controls, the odds ratio for a one standard deviation increase in age ($\exp \beta_z$) is 0.85 for basal cell cancer and 1.62 for squamous cell cancer. Among 588 subjects with both toenail and water measurements, the correlation ($\rho_{xw}$) between log-transformed toenail and water arsenic is 0.45. For 113 of these subjects with water arsenic concentrations of 1 $\mu$g/l or greater, the correlation is 0.69. The correlation between age and toenail arsenic ($\rho_{zx}$) is $-0.20$ and the correlation between age and water arsenic ($\rho_{zw}$) is $-0.05$.

Table I shows the results from our sample size program treating this example as a classical measurement error, a Berkson case measurement error, or a general surrogate measurement error model. Sample size requirements necessary to detect an odds ratio of 1.5 ($\exp \beta_x$) for an increase of one standard deviation in toenail arsenic are provided for case control studies of either basal or squamous cell cancer using water arsenic as a surrogate exposure. The correlations between water and toenail concentrations are varied between $\rho_{xw} = 0.45$ and $\rho_{xw} = 0.69$, corresponding to studies using water concentrations in the overall population and studies using water concentrations from a more highly exposed population. For comparison, calculations for studies using direct toenail measurements ($\rho_{xw} = 1$) give a sample size requirement of 287 for basal cell and 293 for squamous cell carcinoma.

Large increases in required sample sizes are indicated for studies based on water alone, and the somewhat more confounding effect of age slightly increases the requirements for squamous cell cancer as compared to basal cell cancer. For this example, the assumption of a Berkson case measurement error structure requires the largest sample size. The general surrogate measurement error sample size requirements more closely resemble those for classical measurement error.

 

## 6. SIMULATIONS

Computer simulations have been conducted to verify the properties of the power function described in the previous sections and to evaluate its utility for finite sample sizes. The logistic regression model is used in the simulation so that $y$ is a binary random variable such that

$$E(y|z,x) = F(\beta_0 + \beta_z z + \beta_x x)$$

where $F(\cdot)$ is the logistic cumulative distribution function.

We used the power function to calculate the sample size required to detect a fixed alternative with power 0.9 using a size 0.05 significance test. Sample sizes were computed for a range of values for $(\beta_0, \beta_z, \beta_x, \rho, v)$ as described below. For each set of values, 5000 data sets were generated and the empirical power, defined as the proportion of data sets in which the generalized score test rejected the null hypothesis, was calculated.

The scalar covariates $z$ and $x$ were jointly normal with mean zero, unit variances and correlation $\rho$ which was varied at 0, 0.6 and $-0.6$. The value for $\beta_z$ was determined by setting the odds ratio at 2 for the 90th to 10th percentiles of $z$. Values for $\beta_x$ were determined by setting the odds ratio at 3 or 6 for the 90th to 10th percentiles of $x$. Values for $\beta_0$ were determined by setting $E(y) = 0.1$ or 0.5. The measurement error model was $w = x + \sqrt{v}u$ where $u \sim \mathrm{N}(0, 1)$ independently of all other random variables. The measurement error variance, $v$, was varied at 0, 0.25 and 0.75, representing no measurement error, moderate and large amounts of measurement error. There were a total of $2 \times 1 \times 2 \times 3 \times 3 = 36$ configurations for the parameters $(\beta_0, \beta_z, \beta_x, \rho, v)$ and the sample size. Empirical power was computed for each configuration.

As evident from Table II, the empirical power is close to the nominal power of 0.9 in each case. Note also that sample size requirement can increase dramatically as measurement error increases. Additional simulations were run to evaluate the power function for powers and sizes other than 0.9 and 0.05. In all cases, the empirical power was a close match to the nominal power.

The sensitivity of the calculations to the normality assumptions on $(z, x, u)$ was examined through additional simulations. Using the sample sizes in Table I, computed assuming $(z, x, u)$ are normal, we generated data sets where the marginal distributions of $z$ and $x$ where either both skewed, symmetric (but non-normal) or one skewed and the other symmetric. Additionally, the distribution of $u$ was varied among skew and symmetric distributions. A description of a simulation where $x$ was symmetric and $z$ skewed is as follows. We generated $x$ as uniform random variable, centred and scaled to have mean zero and variance one. We then generated $\tilde{z}$ as an independent log-normal, centred and scaled to have mean zero and variance one. We defined $z = \rho x + \sqrt{(1-\rho^2)}\tilde{z}$, implying that $E(z) = 0$, $\mathrm{var}(z) = 1$ and $\mathrm{corr}(x, z) = \rho$. For the simulations where $x$ and $z$ where both skewed, we generated $x$ as a centred and scaled log-normal and $\tilde{z}$ and $z$ were then generated as above. Centred and scaled uniform or log-normal distributions were examined for the measurement error $u$.

The results of the simulations where the distribution of $z$ is skewed, the distribution of $x$ is symmetric or skewed and the distribution of $u$ is normal are shown in Table III. The power function is little affected by the distribution of $u$ regardless of the distributions for $x$ and $z$, so that only results for $u$ normal are given in Table III. The results can be summarized as follows. When the distribution of $x$ is skewed and $z$ is skewed or

 

Table II. Empirical power as estimated from 5000 simulated data sets with sample sizes calculated using the fixed alternative assumptions to achieve a power of 0.9 for a test with a nominal significance level of 0.05. See text for details of the logistic regression and measurement error models.

| $E(y)$ | Odds ratio | Corr$(x,z)$ | Measurement error variance | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | $v=0$ | | $v=0.25$ | | $v=0.75$ | |
| | | | Sample size | Empirical power | Sample size | Empirical power | Sample size | Empirical power |
| 0.1 | 3 | 0 | 686 | 0.902 | 852 | 0.904 | 1181 | 0.908 |
| | | 0.6 | 1069 | 0.909 | 1478 | 0.914 | 2294 | 0.903 |
| | | −0.6 | 1041 | 0.909 | 1438 | 0.902 | 2231 | 0.903 |
| | 6 | 0 | 288 | 0.922 | 354 | 0.918 | 485 | 0.914 |
| | | 0.6 | 440 | 0.916 | 603 | 0.910 | 929 | 0.902 |
| | | −0.6 | 421 | 0.909 | 577 | 0.908 | 887 | 0.911 |
| 0.5 | 3 | 0 | 249 | 0.907 | 312 | 0.901 | 438 | 0.900 |
| | | 0.6 | 398 | 0.899 | 555 | 0.912 | 867 | 0.895 |
| | | −0.6 | 375 | 0.901 | 523 | 0.902 | 819 | 0.903 |
| | 6 | 0 | 103 | 0.912 | 130 | 0.906 | 183 | 0.901 |
| | | 0.6 | 165 | 0.905 | 231 | 0.900 | 362 | 0.907 |
| | | −0.6 | 152 | 0.908 | 212 | 0.914 | 334 | 0.908 |

Table III. Robustness simulations. Table entries are empirical powers for the score test based on 5000 data sets. The sample sizes for each configuration are given in Table II. The distribution of $z$ is skewed and the distribution for $u$ is normal.

| $E(y)$ | Odds ratio | Corr$(x,z)$ | Measurement error variance | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | $v=0$ Distribution for $x$ | | $v=0.25$ Distribution for $x$ | | $v=0.75$ Distribution for $x$ | |
| | | | Symmetric | Skewed | Symmetric | Skewed | Symmetric | Skewed |
| 0.1 | 3 | 0 | 0.903 | 0.978 | 0.903 | 0.979 | 0.897 | 0.982 |
| | | 0.6 | 0.944 | 0.984 | 0.942 | 0.989 | 0.955 | 0.996 |
| | | −0.6 | 0.922 | 0.956 | 0.938 | 0.965 | 0.942 | 0.975 |
| | 6 | 0 | 0.915 | 0.978 | 0.908 | 0.980 | 0.899 | 0.980 |
| | | 0.6 | 0.940 | 0.978 | 0.941 | 0.985 | 0.952 | 0.994 |
| | | −0.6 | 0.902 | 0.935 | 0.899 | 0.937 | 0.896 | 0.933 |
| 0.5 | 3 | 0 | 0.906 | 0.780 | 0.908 | 0.753 | 0.908 | 0.738 |
| | | 0.6 | 0.851 | 0.703 | 0.819 | 0.606 | 0.803 | 0.562 |
| | | −0.6 | 0.880 | 0.828 | 0.884 | 0.806 | 0.873 | 0.786 |
| | 6 | 0 | 0.924 | 0.689 | 0.922 | 0.652 | 0.929 | 0.636 |
| | | 0.6 | 0.836 | 0.649 | 0.807 | 0.552 | 0.776 | 0.493 |
| | | −0.6 | 0.882 | 0.773 | 0.872 | 0.725 | 0.875 | 0.713 |

symmetric, the power function computed assuming $z$ and $x$ are jointly normal generally underestimates the actual power. When $x$ is symmetric and $z$ is skewed, the power is moderately affected.

 

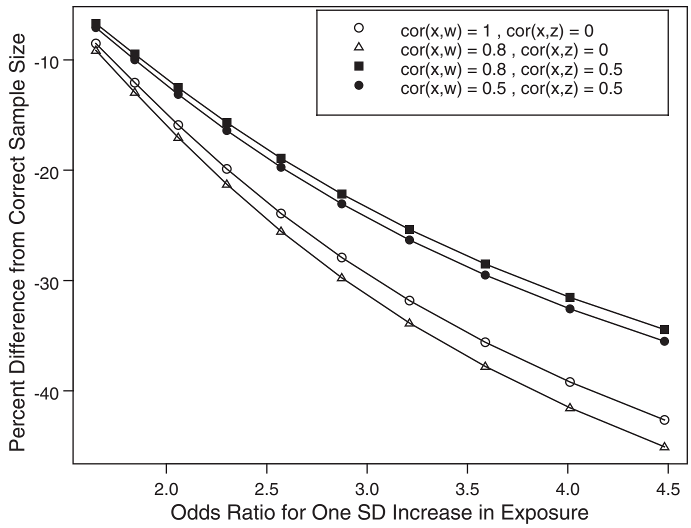


Figure 2. Sample size requirements for logistic regression with normal measurement error calculated using local alternative or exponential risk assumptions, shown as a percentage difference from sample size as calculated using the fixed alternative assumptions.

## 7. COMPARISON OF FIXED AND LOCAL ALTERNATIVES

An advantage of our method is that it provides an accurate accounting of power and required sample sizes without relying on 'local' alternatives or other modelling assumptions leading to the simplified power function (6) and inflation factor $1/\rho^2_{xw}$. To illustrate the degree to which this is important, Figure 2 shows a graph of the required sample size for logistic regression computed using the local approximation (6), plotted as a percentage of the correct sample size computed using (4). The calculations assume that $\exp \beta_z = 1.5$ and $\beta_0 = 0$, with normal covariates and measurement error. This graph indicates that for larger relative risk parameters the local approximation seriously underestimates the required sample size. For this example, greater levels of correlation between the true exposure and the covariates (confounding) apparently improve the quality of the local alternative approximation.

## 8. DISCUSSION

Measurement error is an important consideration in designing an epidemiologic study. Validation data providing estimates of the measurement error variances are necessary for this process and are just as critical as estimates for other parameters needed for conventional power calculations. There have been a variety of proposals for power and sample size calculations in epidemiologic regression models. The developments that we have presented are more general in that they apply to a large class of generalized regression models and measurement error assumptions. They are based on a generalized score test modified for measurement error corrections and do not require further assumptions about the size of the relative risk regression coefficients or other modelling restrictions. Previous methods based on more restrictive assumptions could result in inaccurate sample size calculations for some common models, particularly logistic regression with normal exposures and measurement error.

The numerical and simulation studies illustrate the impact of both covariate measurement error and the correlation between exposure ($x$) and confounder ($z$) on sample size calculations. Increases in sample size are required to maintain power if either the magnitude of the measurement error or the correlation with the confounder are increased. The simulations also indicate that the power function yields accurate sample sizes for detecting both small and large alternatives.

Specific power calculations require specification of the joint distribution of $(\mathbf{z}, x, w)$. Our implemented program specifies joint normality, but the same methods can be applied using other forms for the joint distribution which satisfy the basic conditional independence assumption. The robustness simulation studies using the normality-based power functions have demonstrated a moderate to high degree of sensitivity to the introduction of skewed exposure or measurement error distributions, suggesting that joint distributions need to be correctly specified for the power functions to be accurate.

## APPENDIX

Below we provide the definition of necessary quantities and derivation of the asymptotic distribution of the score statistic under fixed alternative (3).

### A.1. Definition of $k, \phi$, $\hat{\sigma}^2$ and $\hat{\sigma}_0^2$

Define the scalar, $1 \times (q+1)$ vector and $(q+1) \times (q+1)$ matrix

$$c_1(\beta_0, \boldsymbol{\beta}_z) = n^{-1} \sum_{i=1}^{n} H_i^2 m_i(\beta_0, \boldsymbol{\beta}_z, \theta)$$

$$\mathbf{c}_2(\beta_0, \boldsymbol{\beta}_z) = n^{-1} \sum_{i=1}^{n} H_i(1, \mathbf{z}_i') m_i(\beta_0, \boldsymbol{\beta}_z, \theta)$$

$$\mathbf{C}_3(\beta_0, \boldsymbol{\beta}_z) = n^{-1} \sum_{i=1}^{n} (1, \mathbf{z}_i')'(1, \mathbf{z}_i') m_i(\beta_0, \boldsymbol{\beta}_z, \theta)$$

where $m_i(\beta_0, \boldsymbol{\beta}_z, \theta) = d_i(\beta_0, \boldsymbol{\beta}_z, \theta) f_i^{(1)}(\beta_0 + \boldsymbol{\beta}_z' \mathbf{z}_i)$ and $H_i = E[x_i \mid \mathbf{z}_i, w_i; \tau]$. Then $\sigma_0^2$ as a function of $\beta_0, \boldsymbol{\beta}_z$ is defined as

$$\sigma_0^2(\beta_0, \boldsymbol{\beta}_z) = c_1(\beta_0, \boldsymbol{\beta}_z) - \mathbf{c}_2(\beta_0, \boldsymbol{\beta}_z) \mathbf{C}_3^{-1}(\beta_0, \boldsymbol{\beta}_z) \mathbf{c}_2'(\beta_0, \boldsymbol{\beta}_z)$$

The generalized score statistic for testing $H_0\colon \beta_x = 0$ is

$$\frac{L^2(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_z)}{\hat{\sigma}^2 \hat{\sigma}_0^2}$$

where

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\frac{(y_i - f(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_z'\mathbf{z}_i))^2}{g^2(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_z'\mathbf{z}_i, \theta)}$$

and $\hat{\sigma}_0^2 = \sigma_0^2(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_z)$.

It is important in the following to distinguish between the values of the regression coefficients under the alternative, represented by $(\beta_0, \boldsymbol{\beta}_z', \beta_x)$, and the values that the estimators $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_z')$ converge to under the alternative, represented by $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_z')$. Note that $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_z')$ are defined by the identity

$$E(d(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_z, \theta)[y - f(\tilde{\beta}_0 + \tilde{\boldsymbol{\beta}}_z'\mathbf{z})](1, \mathbf{z}')') =$$

$$E(d(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_z, \theta)[f(\beta_0 + \beta_x x + \boldsymbol{\beta}_z'\mathbf{z}) - f(\tilde{\beta}_0 + \tilde{\boldsymbol{\beta}}_z'\mathbf{z})(1, \mathbf{z}')') = \mathbf{0}$$

Under the null hypothesis, $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_z') = (\beta_0, \boldsymbol{\beta}_z)$. Note that $(\tilde{\beta}_0, \tilde{\beta}_z)$ are a function of $(\beta_0, \boldsymbol{\beta}_z, \beta_x)$. Define

$$\begin{aligned}\tilde{\mathbf{c}}_2 = E[\{&d^{(1)}(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_z, \theta)[f(\beta_0 + \boldsymbol{\beta}_z'\mathbf{z} + \beta_x x) - f(\tilde{\beta}_0 + \tilde{\boldsymbol{\beta}}_z'\mathbf{z})]\\ &- d(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_z, \theta) f^{(1)}(\tilde{\beta}_0 + \tilde{\boldsymbol{\beta}}_z'\mathbf{z})\}(1, \mathbf{z}')H]\end{aligned} \tag{A1}$$

$$\begin{aligned}\tilde{\mathbf{C}}_3 = E[\{&d^{(1)}(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_z, \theta)[f(\beta_0 + \boldsymbol{\beta}_z\mathbf{z} + \beta_x x) - f(\tilde{\beta}_0 + \tilde{\boldsymbol{\beta}}_z'\mathbf{z})]\\ &- d(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_z, \theta) f^{(1)}(\tilde{\beta}_0 + \tilde{\boldsymbol{\beta}}_z'\mathbf{z})\}(1, \mathbf{z}')'(1, \mathbf{z}')]\end{aligned} \tag{A2}$$

and independently and identically distributed random variables

$$V_i = d(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_z, \theta)[y_i - f(\tilde{\beta}_0 + \tilde{\boldsymbol{\beta}}_z'\mathbf{z}_i)]\{H_i - \tilde{\mathbf{c}}_2\tilde{\mathbf{C}}_3^{-1}(1, \mathbf{z}_i')'\} \tag{A3}$$

Let

$$\begin{aligned}\sigma_0^2 &= \operatorname{plim}_{n\to\infty}\sigma_0^2(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_z) = E[c_1(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_z)]\\ &\quad - E[\mathbf{c}_2(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_z)]\{E[\mathbf{C}_3(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_z)]\}^{-1}E[\mathbf{c}_2'(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_z)]\end{aligned}$$

$$\begin{aligned}\tilde{\sigma}^2 &= \operatorname{plim}_{n\to\infty}\hat{\sigma}^2\\ &= E\left[\frac{\sigma^2 g^2(\beta_0 + \boldsymbol{\beta}_z'\mathbf{z} + \beta_x x, \theta) + \{f(\beta_0 + \boldsymbol{\beta}_z'\mathbf{z} + \beta_x x) - f(\tilde{\beta}_0 + \tilde{\boldsymbol{\beta}}_z'\mathbf{z})\}^2}{g^2(\tilde{\beta}_0 + \tilde{\boldsymbol{\beta}}_z'\mathbf{z}, \theta)}\right]\end{aligned}$$

Finally, let $\mu = E(V_i)$, $\sigma_1^2 = E(V_i^2) - \mu^2$. Then $\phi$ and $k$ used to calculate the power function (4) are

$$\phi = \frac{n\mu^2}{2\sigma_1^2} \quad \text{and} \quad k = \frac{\tilde{\sigma}^2\sigma_0^2}{\sigma_1^2}$$

 

### *A.2. Asymptotic distribution of the score statistic*

In this section the asymptotic behaviour of the generalized score statistic under a fixed alternative is derived for the case $\dim(x)=1$.

Recall that we assume $\theta$ and $\tau$ are known but that the results hold when $\sqrt{n}$ consistent estimators are substituted. Using standard results from the theory of M-estimators [6], we have that

$$\sqrt{n}\begin{pmatrix}\hat{\beta}_0-\tilde{\beta}_0\\ \hat{\boldsymbol{\beta}}_z-\tilde{\boldsymbol{\beta}}_z\end{pmatrix}=\frac{\tilde{\mathbf{C}}_3^{-1}(\beta_0,\boldsymbol{\beta}_z,\beta_x)}{\sqrt{n}}\sum_{i=1}^{n}d_i(\beta_0,\boldsymbol{\beta}_z,\theta)[y_i-f(\tilde{\beta}_0+\tilde{\boldsymbol{\beta}}_z'\mathbf{z}_i)](1,\mathbf{z}_i')'+o_p(1)$$

where $\tilde{\mathbf{C}}_3(\beta_0,\boldsymbol{\beta}_z,\beta_x)$ is defined by (A2). Next, with $\tilde{\mathbf{c}}_2$ and $V_i$ defined by (A1) and (A3), we obtain

$$\begin{aligned}
L(\hat{\beta}_0,\hat{\boldsymbol{\beta}}_z) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}d_i(\hat{\beta}_0,\hat{\boldsymbol{\beta}}_z,\theta)[y_i-f(\hat{\beta}_0+\hat{\boldsymbol{\beta}}_z'\mathbf{z}_i)]H_i\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}|\,d(\tilde{\beta}_0,\tilde{\boldsymbol{\beta}}_z,\theta)[y_i-f(\tilde{\beta}_0+\tilde{\boldsymbol{\beta}}_z'\mathbf{z}_i)]H_i+\tilde{\mathbf{c}}_2\sqrt{n}\begin{pmatrix}\hat{\beta}_0-\tilde{\beta}_0\\ \hat{\boldsymbol{\beta}}_z-\tilde{\boldsymbol{\beta}}_z\end{pmatrix}+o_p(1)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}d(\tilde{\beta}_0,\tilde{\boldsymbol{\beta}}_z,\theta)[y_i-f(\tilde{\beta}_0+\tilde{\boldsymbol{\beta}}_z'\mathbf{z}_i)]\{H_i-\tilde{\mathbf{c}}_2\tilde{\mathbf{C}}_3^{-1}(1,\mathbf{z}_i')'\}+o_p(1)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_i+o_p(1)
\end{aligned}$$

Recalling that $\mu=EV_i$ and $\sigma_1^2=EV_i^2-(\mu)^2$, it follows that under the alternative hypothesis $L(\hat{\beta}_0,\hat{\boldsymbol{\beta}}_z)\approx \mathrm{N}(\sqrt{n}\mu,\sigma_1^2)$. It follows that the distribution of the generalized score statistic for large $n$ is

$$\frac{L^2(\hat{\beta}_0,\hat{\boldsymbol{\beta}}_z)}{\hat{\sigma}^2\sigma_0^2(\hat{\beta}_0,\hat{\boldsymbol{\beta}}_z)}\sim k^{-1}\chi^2(\phi)$$

ACKNOWLEDGEMENTS

This work was supported in part by grants CA50597, ES07373 and CA57494 from the National Institutes of Health.

REFERENCES

1. Tosteson TD, Tsiatis AA. The asymptotic relative efficiency of score tests in the generalized linear model with surrogate covariates. *Biometrika* 1988; **75**:507–514.
2. Stefanski, LA and Carroll, RJ. Score tests in generalized linear measurement error models. *Journal of the Royal Statistical Society*, *Series B* 1990; **152**:345–359.
3. McKeown-Eyssen GE, Tibshirani R. Implications of measurement error in exposure for the sample sizes of case-control studies. *American Journal of Epidemiology* 1994; **139**:415–421.
4. Devine OJ, Smith JM. Estimating sample size for epidemiologic studies: the impact of ignoring exposure measurement uncertainty. *Statistics in Medicine* 1998; **12**:1375–1389.

 

5. White E, Kushi LH, Pepe MS. The effect of exposure variance and exposure measurement error on study sample size. Implications for design of epidemiologic studies. *Journal of Clinical Epidemiology* 1994; **47**:873–880.
6. Carroll RJ, Ruppert D, Stefanski L. *Measurement Error in Nonlinear Models*. Chapman and Hall: London, 1995.
7. Self SG, Mauritsen RH. Power/sample size calculations for generalized linear models. *Biometrics* 1988; **44**: 79–86.
8. Self SG, Mauritsen RH, O'Hara J. Power calculations for likelihood ratio tests in generalized linear models. *Biometrics* 1992; **48**:31–39.
9. Shieh G. On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics* 2000; **56**:1192–1196.
10. Fuller WA. *Measurement Error Models*. Wiley: New York, 1987.
11. Lagakos S. Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Statistics in Medicine* 1988; **7**:257–274.
12. Karagas MR, Tosteson TD, Blum J, Morris SJ, Baron JA, Klaue B. Design of an epidemiologic study of drinking water arsenic and skin and bladder cancer risk in a U.S. population. *Environmental Health Perspectives* 1998; **106**:1047–1050.
13. Karagas MR, Tosteson TD, Blum J, Klaue B, Weiss JE, Stannard V, Spate V, Morris JS. Measurement of low levels of arsenic exposure: a comparison of water and toenail concentrations. *American Journal of Epidemiology* 2000; **152**:84–90.
14. Karagas MR, Stukel TA, Morris JS, Tosteson TD, Weiss JE, Spencer SK, Greenberg ER. Skin cancer risk in relation to toenail arsenic concentrations in a US population-based case-control study. *American Journal of Epidemiology* 2001; **153**:559–565.
15. Karagas MR, Stukel TA, Tosteson TD. Assessment of cancer risk and environmental levels of arsenic in New Hampshire. *International Journal of Hygiene and Environmental Health* 2002; **205**:85–94.

 