Inflation of Type I error in multiple regression when independent variables are measured with error

L. Jerome Brunner Peter C. Austin [1]

September 11, 2007

[1]Lawrence J. Brunner is Associate Professor, Department of Statistics, University of Toronto, Toronto M5G 3G3, Canada (email brunner@utstat.toronto.edu), and Peter Austin is Senior Scientist, Institute for Clinical Evaluative Sciences, Toronto M4N 3M5, Canada (email peter.austin@ices.on.ca). This work was partially supported by Natural Sciences and Engineering Research Council of Canada Grant OGPIN 014.

Abstract

When independent variables are measured with error, ordinary least squares regression can yield parameter estimates that are biased and inconsistent. Here, we document the inflation of Type I error that can occur in this situation. In addition to analytic results, we report a large-scale Monte Carlo study showing unacceptably high Type I error rates, under circumstances that could easily be encountered in practice. The problem applies to various types of regression and various types of measurement error. Statistical methods that incorporate measurement error are available, but their use requires multiple indicators of the independent variables. This implies a new tradition of data collection.

Keywords: Errors in variables, Measurement error, Type I error, Structural equation models, Monte Carlo.

Introduction

This is a story about something everyone knows, but few seem to appreciate. Consider the usual univariate multiple regression model with independent normal errors. Everyone knows that even though the independent variables are supposed to be fixed constants, in non-experimental studies they are usually random variables. This is okay, but if the independent variables are measured with error, everyone knows that there is trouble. Expressions of concern go back at least to Stouffer (1936), who observed that estimates of partial correlations can be biased when the variables for which one is controlling are measured with error. By the seventh edition of *Statistical methods for research workers*, Fisher (1938) was warning scientists about the problem, again in the context of partial correlation. For multiple regression proper, earlier discussions are reviewed and clarified by Cochran (1968), who shows that when the independent variables are measured with error, ordinary least squares estimates of the regression coefficients can be inconsistent and biased, even asymptotically.

The misleading quality of measurement error in the independent variables has figured in one important political debate. Initial analyses of data from the Head Start program (Cicirelli *et al.* 1969, Barnow 1973) suggested that even controlling for socioeconomic status, students receiving a shorter (summer-only) version of the program performed worse on an educational test than students who were not exposed to any version of the Head Start program. The conclusion was that Head Start could actually be harmful. This claim was challenged by Campbell and Erlbacher (1970) on the grounds that socioeconomic status was measured with error, and so attempts to control for it using ordinary least squares would not completely correct for differences between the treatment group and the non-randomized comparison group.

In subsequent debate and re-analysis of the data allowing for measurement error (Magidson 1977, Bentler and Woodward 1978, Magidson 1978), harmful effects are

entirely ruled out. It is now accepted that all versions of the Head Start program were helpful for African-American and Mexican-American children, and that the full program was also helpful for White children; disagreement is limited to whether the data also provide adequate evidence of a positive effect (not a negative effect) for White children receiving the summer-only version of the program. What we take from this example is that one can get away with ignoring measurement error, but not when the conclusions of the study make a serious difference.

If measurement error is not to be ignored, it must be included in the statistical model. The modelling of measurement error in the predictor variables has a long history, especially in economics; see Wald (1940) and Madansky (1959) for references to early writings. Today, there is a well-developed literature on regression models that incorporate measurement error; for example, see the discussions and references in Fuller (1987), Cheng and Van Ness (1999) and Wansbeek and Meijer (2000). These measurement error models are special cases of the structural equation and related models that have been long been popular in the social and biomedical sciences: see for example Jöreskog (1978), Bollen (1989), and the generalizations of Skrondal and Rabe-Hesketh (2004), Muthén (2002) and Muthén & Muthén (2006).

So, it is widely recognized that measurement error can present a problem for ordinary least-squares regression, and a class of high-quality solutions is in place. But please glance at the regression text that is closest to hand, provided that it is not an econometrics text. It may or may not contain a warning about measurement error in the independent variables, but look at the examples and sample data sets. You will be reminded that in practice, measurement error is routinely ignored, and that individuals at all levels of statistical sophistication are encouraged to go ahead and carry out ordinary least-squares regression on observational data without worrying too much about it. It is as if people are saying that asymptotic bias and inconsistency do

sound pretty bad, but maybe there is is just a little bit of measurement error. Maybe it does not matter much.

Unfortunately, it can matter a lot. Ignoring measurement error in the independent variables of a regression can drastically inflate Type I error. Simply put, when one tries to "control" for an independent variable that is measured with error, traditional methods do not completely do the job. This holds under circumstances that can easily be encountered in practice, and applies to various types of regression and various types of measurement error.

We focus upon Type I error rather than bias, because significance tests are often used in the biological and social sciences as a kind of filter, to reduce the amount of random noise that gets into the scientific literature. In fact, we view this as the primary function of statistical hypothesis testing in the discourse of science. Essentially, $p < 0.05$ means that it is socially acceptable to speak. Therefore, when a common statistical practice can be shown to inflate Type I error, there is a problem — a problem that will be recognized by a large class of practitioners who are totally unmoved by calculations of asymptotic bias.

Of course there is a connection between asymptotic bias and Type I error. If the asymptotic bias occurs when the null hypothesis is true, *and* the estimated standard deviation of the estimator tends to zero under the incorrect model, then the Type I error rate will necessarily increase to unity. This accounts for passing references (Fuller, 1978, p. 55; Cochran, 1968, p. 653) to incorrect Type I error rates when measurement error is ignored. What we are doing in this paper is documenting the connection and making it explicit for a particular class of examples.

In Section 1, we revisit an example discussed by Cochran (1968), in which ignoring measurement error is shown to produce inconsistent least-squares estimates of regression coefficients when the independent variables as well as the dependent variable are

normal. We observe that the inconsistency applies regardless of distribution, and that the Type I error rate of the usual F or t test tends almost surely to one as the sample size approaches infinity. These analytic results are supported by a large-scale Monte Carlo study showing unacceptably high Type I error rates, even for small amounts of measurement error and moderate sample sizes.

In Section 2, we report a set of smaller-scale simulations. First, we present an example of how ignoring measurement error can result in rejection of the null hypothesis virtually always when the null hypothesis is false — but with the regression coefficient having the wrong sign. Finally, we combine references to the literature and small Monte Carlo studies to show that ignoring measurement error in the independent variables can inflate Type I error for various types of regression (such as logistic regression and Cox proportional hazards regression for survival data), and various types of measurement error, including classification error for categorical independent variables. This calls into question many non-experimental studies which claim to have "controlled" for potential confounding variables or risk factors using the standard tools.

Modelling measurement error is preferable to ignoring it, and good solutions are available. However, models that include measurement error usually require multiple indicators of the independent variables in order to be uniquely identified in the model parameters. For linear regression with measurement error, a simple solution is to measure the independent variables twice. If it can be assumed that errors of measurement on different occasions are uncorrelated, then appropriate methods can applied in a routine manner.

1 Inflation of Type I error rate

To see how badly things can go wrong when measurement error is ignored, consider a multiple regression model in which there are two independent variables, both measured with simple additive error. This situation has been thoroughly studied, notably by Cochran (1968), but the following is a bit more general than usual.

Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} Y_i &= \beta_0 + \beta_1\xi_{i,1} + \beta_2\xi_{i,2} + \epsilon_i \\ X_{i,1} &= \nu_1 + \xi_{i,1} + \delta_{i,1} \\ X_{i,2} &= \nu_2 + \xi_{i,2} + \delta_{i,2}, \end{aligned} \tag{1}$$

where β_0, β_1 and β_2 are unknown constants (regression coefficients), and

$$E\begin{bmatrix} \xi_{i,1} \\ \xi_{i,2} \end{bmatrix} = \begin{bmatrix} \kappa_1 \\ \kappa_2 \end{bmatrix} \qquad Var\begin{bmatrix} \xi_{i,1} \\ \xi_{i,2} \end{bmatrix} = \boldsymbol{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{bmatrix}$$

$$E\begin{bmatrix} \delta_{i,1} \\ \delta_{i,2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad Var\begin{bmatrix} \delta_{i,1} \\ \delta_{i,2} \end{bmatrix} = \boldsymbol{\Theta} = \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{12} & \theta_{22} \end{bmatrix}$$

$$E[\epsilon_i] = 0 \qquad Var[\epsilon_i] = \sigma^2.$$

The true independent variables are $\xi_{i,1}$ and $\xi_{i,2}$, but they are latent variables that cannot be observed directly. They are independent of the error term ϵ_i and of the measurement errors $\delta_{i,1}$ and $\delta_{i,2}$; the error term is also independent of the measurement errors. The constants ν_1 and ν_2 represent measurement bias. For example, if ξ_1 is true average minutes of exercise per day and X_1 is reported average minutes of exercise, then ν_1 is the mean amount by which people exaggerate their exercise times.

Also, it is reasonable to allow the measurement errors to be correlated. Again, suppose that ξ_1 is true amount of exercise and X_1 is reported amount of exercise, while ξ_2 is true age and X_2 is reported age. It is natural to imagine that adults who exaggerate how much they exercise might tend to under-report their ages. Thus, the covariance parameter θ_{12} is quite meaningful.

When a model such as (1) holds, all one can observe are the triples $(X_{i,1}, X_{i,2}, Y_i)$ for $i = 1, \ldots, n$. Suppose the interest is in testing whether ξ_2 is related to Y, conditionally on the value of ξ_1. The natural mistake is to take X_1 as a surrogate for ξ_1 and X_2 as a surrogate for ξ_2, fit the model

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i \tag{2}$$

by ordinary least squares, and (assuming ϵ_i normal) test the null hypothesis $H_0 : \beta_2 = 0$ using the usual t or F-test.

Suppose that in fact $\beta_2 = 0$ in Model (1), so that conditionally upon the value of ξ_1, the dependent variable Y is independent of ξ_2. We now observe that that except under special circumstances, the least squares quantity $\widehat{\beta}_2$ based on Model (2) converges almost surely to a quantity different from $\beta_2 = 0$ as the sample size increases, with the p-value of the standard test going to zero and the Type I error rate going to one.

1.1 Almost sure disaster

The ordinary least-squares estimate $\widehat{\beta}_2$ (based upon the incorrect Model 2) is a function of the sample variance-covariance matrix, which by the Strong Law of Large Numbers, converges almost surely to the true variance-covariance matrix of the observed data. This variance-covariance matrix is in turn a function of the parameters of the true model (1). So by a continuity argument, the ordinary least-squares estimate converges almost surely to the corresponding function of the true model parameters.

Our focus is upon Type I error for the present, so we examine the case where $H_0 : \beta_2 = 0$ is true. Setting $\beta_2 = 0$ and simplifying, we find that as n tends to infinity,

$$\widehat{\beta}_2 \stackrel{a.s.}{\to} \frac{\beta_1(\phi_{12}\theta_{11} - \phi_{11}\theta_{12})}{(\phi_{11} + \theta_{11})(\phi_{22} + \theta_{22}) - (\phi_{12} + \theta_{12})^2} \tag{3}$$

Expression (3) is the asymptotic bias of $\widehat{\beta}_2$ as an estimate of the true regression parameter β_2, in the case where $\beta_2 = 0$. Notice that it does not depend upon the

intercept β_0, the measurement bias terms ν_1 and ν_2, nor upon κ_1 and κ_2, the expected values of the latent independent variables.

Clearly, the bias is zero only if $\beta_1 = 0$ (the dependent variable is unrelated to ξ_1) or if $\phi_{12}\theta_{11} = \phi_{11}\theta_{12}$. Notice the parallel roles played by ϕ_{12}, the covariance between the latent "true" independent variables, and θ_{12}, the covariance between error terms. If they have opposite signs they pull in the same direction, but if they have the same sign they can partially or even completely offset one another. The effect of ϕ_{12} is augmented by the variance of the error in measuring ξ_1, while the effect of θ_{12} is augmented by the variance of ξ_1 itself.

The parameter θ_{11}, the variance of the error term δ_1, represents the amount of noise in the independent variable for which one is trying to control, while θ_{22} is the amount of noise in the independent variable one is trying to test. Clearly, θ_{11} is a greater potential problem, because θ_{22} appears only in the denominator; measurement error in the variable for which one is testing actually *decreases* the asymptotic bias, in this case where $\beta_2 = 0$. Incidentally, the denominator of (3) is the determinant of the covariance matrix of X_1 and X_2; it will be positive provided that at least one of $\mathbf{\Phi}$ and $\mathbf{\Theta}$ are positive definite. This condition is required for convergence.

All these details aside, the main point is that when Y is conditionally independent of ξ_2, the estimator $\widehat{\beta}_2$ converges to a quantity that is not zero in general. Now, $\widehat{\beta}_2$ is the numerator of the t-statistic commonly used to test $H_0 : \beta_2 = 0$ as a substitute for the real null hypothesis $H_0 : \beta_2 = 0$. The denominator, the estimated standard deviation of $\widehat{\beta}_2$, may be written as

$$S_{\widehat{\beta}_2} = \frac{W_n}{\sqrt{n}}.$$

Using the same approach that led to (3), we find that W_n converges almost surely to a positive constant, again provided that at least one of the covariance matrices $\mathbf{\Phi}$ and $\mathbf{\Theta}$ are positive definite. Consequently, the absolute value of the t-statistic blows

up to infinity, and the associated p-value converges almost surely to zero. That is, we almost surely commit a Type I error.

1.2 A Monte Carlo study of Type I error inflation

The preceding result applies as $n \to \infty$. To get an idea of how much Type I might error be inflated in practice, we conducted a large-scale Monte Carlo study in which we simulated data sets from Model (1) using various sample sizes, probability distributions and parameter values.

Since Expression (3) for the asymptotic bias does not depend on any of the expected value terms, we set all expected values to zero for the simulations, except for an arbitrary intercept $\beta_0 = 1$ in the latent regression equation. Also, we let $\theta_{1,2} = 0$, so there is no correlation between the measurement errors.

This is a complete factorial experiment with six factors.

1. *Sample size:* There were 5 values; $n = 50$, 100, 250, 500 and 1000.

2. *Correlation between latent (true) independent variables:* Letting R_1. R_2 and R_3 be independent random variables with mean zero and variance one, we calculated

$$\begin{aligned} \xi_1 &= \sqrt{1-\phi_{1,2}}\, R_1 + \sqrt{\phi_{1,2}}\, R_3 \text{ and} \\ \xi_2 &= \sqrt{1-\phi_{1,2}}\, R_2 + \sqrt{\phi_{1,2}}\, R_3, \end{aligned} \tag{4}$$

   yielding $Var(\xi_1) = Var(\xi_2) = 1$ and a correlation of $\phi_{1,2}$ between ξ_1 and ξ_2. A quiet but important feature of this construction is that when $\phi_{1,2} = 0$, ξ_1 and ξ_2 are independent, even when the distributions are not normal. There were five correlation values: $\phi_{1,2} = 0.00$, 0.25, 0.75, 0.80 and 0.90.

3. *Variance explained by ξ_1:* With $\beta_1 = 1$, $\beta_2 = 0$ and $Var(\xi_1) = \phi_{1,1} = 1$, we have $Var(Y) = 1 + \sigma^2$. So, the proportion of variance in the dependent variable that comes from ξ_1 is $\frac{1}{1+\sigma^2}$. We used this as an index of the strength of relationship between ξ_1 and Y, and adjusted it by changing the value of σ^2. There were three values of explained variance, corresponding to a weak, moderate and strong relationship between ξ_1 and Y: 0.25, 0.50 and 0.75.

4. *Reliability of X_1:* In classical psychometric theory (for example Lord and Novick, 1968) the *reliability* of a test is the squared correlation between the observed score and the true score. It is also the proportion of variance in the observed score that comes from the true score. From Model (1), we have

$$[Corr(\xi_1, X_1)]^2 = \left[\frac{\phi_{1,1}}{\sqrt{\phi_{1,1}}\sqrt{\phi_{1,1} + \theta}_{1,1}}\right]^2 = \frac{1}{1 + \theta_{1,1}}.$$

   Thus we may manipulate the reliability by changing the value of the error variance $\theta_{1,1}$. Five reliability values were employed, ranging from lackluster to stellar: 0.50, 0.75, 0.80, 0.90 and 0.95.

5. *Reliability of X_2:* The same five values were used: 0.50, 0.75, 0.80, 0.90 and 0.95.

6. *Base distribution:* In all the simulations, the distribution of the errors in the latent regression (ϵ_i) are normal; we have no interest in revisiting the consequences of violating the assumption of normal error in multiple regression. But the distributions of the latent independent variables and measurement errors are of interest. We constructed the measurement error terms by multiplying standardized random variables by constants to give them the desired variances. These standardized random variables, and also the standardized variables R_1. R_2 and R_3 used to construct ξ_1 and ξ_2 – see Equations (4) – come from a com-

mon distribution, which we call the "base" distribution. Four base distributions were examined.

- Standard normal
- Student's t with degrees of freedom 4.1, scaled to have unit variance.
- Uniform on the interval $(-\sqrt{3}, \sqrt{3})$, yielding mean zero and variance one.
- Pareto (density $f(x) = \frac{\alpha}{x^{\alpha+1}}$ for $x > 1$) with $\alpha = 4.1$, but standardized.

Distributions and base distributions Because the simulated data values are linear combinations of standardized random variables from the base distribution, the base distribution is the same as the distribution of the simulated data only for the normal case. Otherwise, the independent variables (both latent and observed) are nameless linear combinations that inherit some of the properties of the base distribution. The t base distribution yielded heavy-tailed symmetric distributions, the Pareto yielded heavy-tailed nonsymmetric distributions, and the uniform yielded light-tailed distributions.

Results Again, this is a complete factorial experiment with $5 \times 5 \times 3 \times 5 \times 5 \times 4 = 7,500$ treatment combinations. Within each treatment combination, we independently generated 10,000 random sets of data, for a total of 75 million simulated data sets. For each data set, we ignored measurement error, fit Model (2) and tested $H_0 : \beta_2 = 0$ with the usual "extra sum of squares" F-test. The proportion of simulated data sets for which the null hypothesis was rejected at $\alpha = 0.05$ is a Monte Carlo estimate of the Type I error rate.

Considerations of space do not permit us to reproduce the entire set of results here. Instead, we give an excerpt that tells the main part of the story, referring the reader to `www.utstat.toronto.edu/~brunner/MeasurementError` for the rest.

On the Web, the full set of results is available in the form of a 6-dimensional table with 7,500 cells, and also in the form of a plain text file with 7,500 lines, suitable as input data for further analysis. Complete source code for our special-purpose fortran programs is also available for download, along with other supporting material.

Table 1 shows the results when all the variables are normally distributed and the reliabilities of both independent variables equal 0.90; that is, only 10% of the variance of the independent variables arises from measurement error. In the social and behavioral sciences, a reliability of 0.90 would be considered impressively high, and one might think there was little to worry about.

In Table 1, we see that except when the latent independent variables ξ_1 and ξ_2 are uncorrelated, applying ordinary least squares regression to the corresponding observable variables X_1 and X_2 results in a substantial inflation of the Type I error rate. As one would predict from Expression (3) with $\theta_{1,2} = 0$, the problem becomes more severe as ξ_1 and ξ_2 become more strongly related, as ξ_1 and Y become more strongly related, and as the sample size increases. We view the Type I error rates in Table 1 as shockingly high, even for fairly moderate sample sizes and modest relationships among variables.

This same pattern of results holds for all four base distributions, and for all twenty-five combinations of reliabilities of the independent variables. In addition, the Type I error rates increase with decreasing reliability of X_1, and decrease with decreasing reliability of X_2 (the variable being tested). The distribution of the error terms and independent variables does not matter much, though average Type I error rates are slightly lower when the base distribution is the skewed and heavy-tailed Pareto; the marginal mean estimated Type I error rate was 0.37 for the Pareto, compared to 0.38 for the Normal, t and Uniform.

Table 1: Estimated Type I error rates when independent variables and measurement errors are all normal, and reliability of X_1 and X_2 both equal 0.90

25% of Variance in Y is Explained by ξ_1

	Correlation Between ξ_1 and ξ_2				
N	0.0	0.2	0.4	0.6	0.8
50	0.0476†	0.0505†	0.0636	0.0715	0.0913
100	0.0504†	0.0521†	0.0834	0.0940	0.1294
250	0.0467†	0.0533†	0.1402	0.1624	0.2544
500	0.0468†	0.0595†	0.2300	0.2892	0.4649
1000	0.0505†	0.0734	0.4094	0.5057	0.7431

50% of Variance in Y is Explained by ξ_1

	Correlation Between ξ_1 and ξ_2				
N	0.0	0.2	0.4	0.6	0.8
50	0.0460†	0.0520†	0.0963	0.1106	0.1633
100	0.0535†	0.0569†	0.1461	0.1857	0.2837
250	0.0483†	0.0625	0.3068	0.3731	0.5864
500	0.0515†	0.0780	0.5323	0.6488	0.8837
1000	0.0481†	0.1185	0.8273	0.9088	0.9907

75% of Variance in Y is Explained by ξ_1

	Correlation Between ξ_1 and ξ_2				
N	0.0	0.2	0.4	0.6	0.8
50	0.0485†	0.0579†	0.1727	0.2089	0.3442
100	0.0541†	0.0679	0.3101	0.3785	0.6031
250	0.0479†	0.0856	0.6450	0.7523	0.9434
500	0.0445†	0.1323	0.9109	0.9635	0.9992
1000	0.0522†	0.2179	0.9959	0.9998	1.00000

†Not Significantly different from 0.05, Bonferroni corrected for 7,500 tests.

2 And there's more

2.1 Significance in the wrong direction

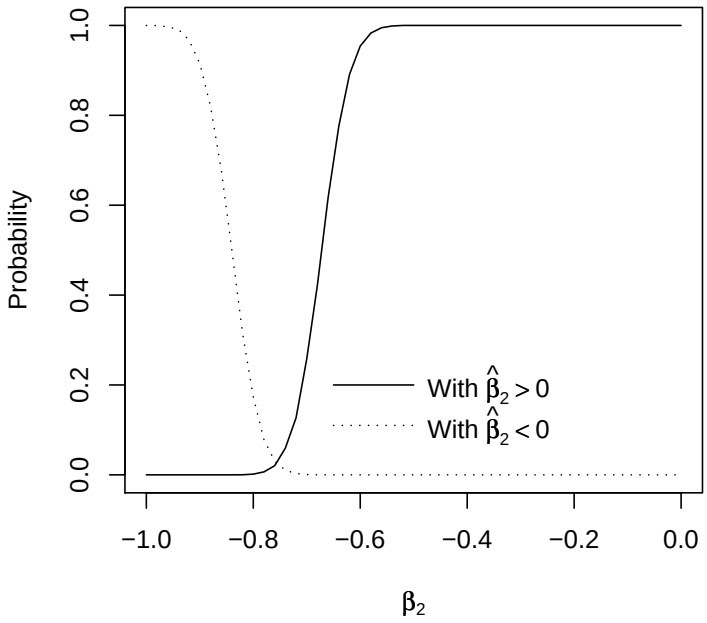
Consider Model (1) again. Let the covariance between ξ_1 and ξ_2 be positive, the partial relationship between ξ_1 and Y be positive, and the partial relationship between ξ_2 and Y be *negative*. That is, $\phi_{1,2} > 0$, $\beta_1 > 0$, and $\beta_2 < 0$. Again, suppose we ignore measurement error and fit Model (2) with ordinary least squares, and test $H_0 : \beta_2 = 0$. We now describe a simulation showing how small negative values of β_2 can be overwhelmed by the positive relationships between ξ_1 and ξ_2, and between ξ_1 and Y, leading to rejection of the null hypothesis at a high rate, accompanied by a *positive* value of $\widehat{\beta}_2$.

This kind of "Type III error" (Kaiser, 1960) is particularly unpleasant from a scientist's perspective, because the reality is that for each value of the first independent variable, the second independent variable is negatively related to the dependent variable. But application of the standard statistical tool leads to the conclusion that the relationship is positive – the direct opposite of the truth. Almost certainly, such a finding will muddy the literature and interfere with the development of any worthwhile scientific theory.

As in the first set of simulations, we set all expected values in Model (1) to zero except for the intercept $\beta_0 = 1$. We also let $\theta_{1,2} = 0$, $\beta_1 = 1$, and $\phi_{1,1} = \phi_{2,2} = 1$. We then employed a standard normal base distribution, together with a sample size and set of parameter values guaranteed to cause problems with Type I error: $n = 500$, $\phi_{1,2} = 0.90$, $\sigma^2 = \frac{1}{3}$ (so that ξ_1 explains 0.75 of the variance in Y), $\theta_{1,1} = 1$ (so that the reliability of X_1 is 0.50), and $\theta_{2,2} = \frac{1}{19}$ (so that the reliability of X_2 is 0.95).

We then varied β_2 from minus one to zero, generating 10,000 data sets for each value of β_2. For each data set, we fit Model (2) and tested $H_0 : \beta_2 = 0$ at $\alpha = 0.05$ with the usual F-test. Each test was classified as significant with $\widehat{\beta}_2 > 0$, significant

Figure 1: Probability of Rejecting $H_0 : \beta_2 = 0$



with $\widehat{\beta}_2 < 0$, or nonsignificant.

Figure 1 shows the results. For substantial negative values of β_2, the null hypothesis $H_0 : \beta_2 = 0$ is rejected at a high rate with $\widehat{\beta}_2 < 0$, leading to the correct conclusion even though the model is wrong. As the value of β_2 increases, the proportion of significant tests decreases to near zero around $\beta_2 = -0.76$ Then for values of β_2 closer to zero (but still negative), the null hypothesis is increasingly rejected again, but this time with $\widehat{\beta}_2 > 0$, leading to the conclusion of a positive relationship, when in fact the relationship is negative. This example shows how ignoring measurement error in the independent variables can lead to firm conclusions that are directly opposite to reality.

2.2 The generality of the problem

We have illustrated inflation of Type I error for the normal linear model with simple additive measurement error, but the problem is much more general. We suggest that *regardless of the type of measurement error and regardless of the statistical method used, ignoring measurement error in the independent variables can seriously inflate Type I error*. We will now support this assertion by references to the literature, supplemented by a collection of quick, small-scale Monte Carlo studies. All the simulations in this section were carried out using *R* Version 2.1.1 (R Development Core Team, 2006). Code is available at `www.utstat.toronto.edu/~brunner/MeasurementError`.

Logistic regression with additive measurement error In this small simulation, we constructed data sets with a pair of latent independent variables ξ_1 and ξ_2, and corresponding manifest variables X_1 and X_2, using a normal base distribution and the troublesome $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ values of Section (2.1). We then constructed a binary dependent variable Y, with the log odds of $Y = 1$ equal to $\beta_0 + \beta_1\xi_1 + \beta_2\xi_2$, where $\beta_0 = \beta_1 = 1$ and $\beta_2 = 0$. Ignoring the measurement error, we fit a standard logistic regression model with the log odds of $Y = 1$ equal to $\beta_0 + \beta_1 X_1 + \beta_2 X_2$, and used a likelihood ratio test of $H_0 : \beta_2 = 0$. The parallel to what we did with ordinary least squares regression should be clear.

In 1,000 simulations with $n = 250$, we incorrectly rejected the null hypothesis 957 times. This shows that the problem described in this paper applies to logistic regression as well as to the normal linear model.

Normal linear regression with censored independent variables Austin and Brunner (2003) describe inflation of Type I error for the case where an independent variable has a "cutoff" – a value that is recorded for the independent variable if it equals or exceeds the cutoff value. The inflation of Type I error occurs when

the one attempts to test another variable that is correlated with the true version of the censored variable, while "controlling" for the censored version with ordinary regression. If one views the censoring as an obscure type of measurement error, this fits neatly into the framework of the present paper.

Normal linear regression and logistic regression with categorized independent variables The most common variant of this data analytic crime arises when independent variables are split at the median and converted to binary variables. The loss of information about the independent variables is a type of measurement error, albeit one that is deliberately introduced by the data analyst. Maxwell and Delaney (1993) show how Type I error can be inflated in this situation. While their argument depends upon a multivariate normal distribution for the data, in fact the inflation of Type I error does not depend upon the distribution (apart from the existence of moments). Median splitting the independent variables has also been shown to inflate Type I error in logistic regression (Austin and Brunner, 2004).

Normal linear regression, ranking the independent variable We have unpublished work showing that in terms of Type I error, median splits are worse than dividing the independent variable into three categories, three categories are worse than four, and so on. The limiting case is when an independent variable is ranked, and one performs a regression controlling for the ranked version, rather than for the independent variable itself. Even here there can be substantial Type I error inflation; we demonstrate this with a quick simulation.

We constructed data sets according to Model (1) again using the $\mathbf{\Phi}$ values of Section (2.1), a reliability of 0.95 for X_2, a normal base distribution, $\beta_0 = \beta_1 = 1$ and $\beta_2 = 0$. However, the observable independent variable X_1 contained the ranks of ξ_1, rather than ξ_1 plus a piece of random noise. As usual, we fit the incorrect regression

model (2) and tested $H_0 : \beta_2 = 0$ with the usual "extra sum of squares F-test. In 1,000 simulated data sets, the null hypothesis was rejected 544 times at the 0.05 level.

Log-linear models with classification error For categorical independent variables, the most natural kind of measurement error is *classification error*, in which the recorded value of a variable is different from the true one. In this case, the structure of measurement error corresponds to a matrix of transition probabilities from the latent variable to the observable variable.

Now we construct an example to show that ignoring measurement error can lead to unacceptable inflation of the Type I error rate in this situation. Again there are two correlated latent variables ξ_1 and ξ_2, only this time they are binary. The corresponding observable variables X_1 and X_2 are also binary. There is a binary dependent variable Y that is dependent upon ξ_1 and conditionally independent of ξ_2.

The components of the measurement error model are two-way tables of the joint probabilities of ξ_1 and ξ_2, ξ_1 with X_1, and ξ_2 with X_2. The values we used are given in Table 2.

Table 2: Joint probabilities for the classification error model

	ξ_1	
ξ_2	0	1
0	0.40	0.10
1	0.10	0.40

	X_1	
ξ_1	0	1
0	0.30	0.20
1	0.20	0.30

	X_2	
ξ_2	0	1
0	0.45	0.05
1	0.05	0.45

The data were constructed by first sampling a (ξ_1, ξ_2) pair from a multinomial distribution, and then simulating X_1 conditionally on ξ_1 and X_2 conditionally on ξ_2. Finally, we generated Y conditionally on ξ_1 using $P(Y = 0|\xi_1 = 0) = P(Y = 1|\xi_1 = 1) = 0.80$. Repeating this process $n = 250$ times yielded a simulated data set of (X_1, X_2, Y) triples. We then tested for conditional independence of X_2 and

Y given X_1, as a surrogate for for the conditional independence of ξ_2 and Y given $\xi_!$. Specifically, we used R's `loglin` function to fit a hierarchical loglinear model with an association between X_1 and X_2, and between X_1 and Y. Comparing this to a saturated model, we calculated a large-sample likelihood ratio test of conditional independence with two degrees of freedom. In 1,000 independent repetitions of this process, the null hypothesis was incorrectly rejected 983 times at the 0.05 level.

Factorial ANOVA with classification error In an unbalanced factorial design with a quantitative dependent variable, a common approach — say using the Type III sums of squares of SAS `proc glm` (SAS Institute Inc., 1999) — is to test each main effect controlling for all the others as well as the interactions. We now report a quick simulation showing that in a two-factor design, if factor level membership is subject to classification error in one of the independent variables, then Type I error may be inflated in testing for a main effect of the other independent variable.

We started with two correlated binary latent independent variables ξ_1 and ξ_2, and their corresponding observable versions X_1 and X_2, constructed according to the same classification error model we used for loglinear models; see Table 2. We then generated the dependent variable as $Y = 1 + \xi_1 + \epsilon$, where ϵ is Normal with mean zero and variance $\frac{1}{4}$. Because ξ_1 is Bernoulli with probability one-half, its variance is also $\frac{1}{4}$, and it accounts for half the variance in Y. Conditionally upon the latent (true) independent variable ξ_1, Y is independent of ξ_2 and there is no interaction.

Repeating this process $n = 200$ times yielded a simulated data set of (X_1, X_2, Y) triples. As usual, we conducted the analysis using the observable variables X_1 and X_2 in place of ξ_1 and ξ_2 respectively, ignoring the measurement error. We fit a regression model with effect coding and a product term for the interaction, and tested for a main effect of X_2 at the 0.05 level with the usual F test. Again, this is equivalent to the test based on Type III sums of squares in SAS `proc glm`. Conducting this test

on 1,000 simulated data sets, we incorrectly rejected the null hypothesis 995 times.

Discarding data to get equal sample sizes in factorial ANOVA In Section 1, we saw that inflation of Type I error arises not just from measurement error in the independent variables, but from the combination of correlated independent variables and measurement error in the one for which one is attempting to "control." Now sometimes, researchers (not statisticians, we hope) randomly discard data from observational studies to obtain balanced factorial designs, and it might be tempting to try this here to eliminate the correlation between independent variables. It doesn't work, though, because it is association between the *latent* independent variables that is the source of the problem.

To verify this, we simulated random sets of data exactly as in the last example, except that when one of the four combinations of X_1, X_2 values reached 50 observations, we discarded all subsequent observations in that cell, continuing until we had 50 data values in each of the four cells. Then we tested for a main effect of X_2 (as a surrogate for ξ_2) exactly as before. The result was that we wrongly rejected the null hypothesis 919 times in 1,000 simulated data sets.

Proportional hazards regression with additive measurement error The last mini-simulation shows that the problem of inflated Type I error extends to survival analysis. Proceeding as in earlier examples, we constructed data sets with a pair of latent independent variables ξ_1 and ξ_2, and also corresponding manifest variables using a normal base distribution and the the $\mathbf{\Phi}$ and $\mathbf{\Theta}$ values of Section (2.1). We then sampled the dependent variable Y from an exponential distribution with mean $\exp \beta_0 + \beta_1\xi_1 + \beta_2\xi_2$, with $\beta_0 = \beta_1 = 1$ and $\beta_2 = 0$. So again, Y is conditionally independent of ξ_2. We then right-censored all the data for which $Y > 5$ (Type I censoring), so that around a quarter of the data in each data set were censored.

Ignoring the measurement error, we fit a proportional hazards model (Cox, 1972) with R's `coxph` function, using X_1 and X_2 as the independent variables, testing the relationship of X_2 to Y controlling for X_1. In 1,000 simulated data sets with $n = 100$, we incorrectly rejected the null hypothesis 994 times, showing that proportional hazards regression, too, is subject to severe inflation of Type I error when measurement error in the independent variables is ignored.

3 Discussion

We are not suggesting that ignoring measurement error *always* inflates Type I error to the degree indicated by our Monte Carlo results. Usually there are more than two independent variables; in this case, ordinary least-squares estimates of regression parameters are still asymptotically biased, but the pattern is complex, with many parameters having the potential to partially cancel or magnify the effects of others when the null hypothesis is true. With estimated standard errors going to zero, Type I error will still approach one as the sample size tends to infinity for most parameter confiurations, but the magnitude of the effect for a given sample size will depend upon the variances and covariances among the independent variables and among the measurement errors.

Still, we cannot escape the conclusion that measurement error in the independent variables will inflate Type I error to some degree. The severity of the problem in practice is unknown, but our Monte Carlo results suggest that it can be very bad. Given this, it seems unduly optimistic to continue applying standard regression and related methods in the presence of obvious measurement error, and hoping that the various sources of trouble will cancel out and everything will be okay. Surely, we can do better.

For linear models with measurement error, we prefer to use classical structural

equation modelling of the kind described by Jöreskog (1978) and Bollen (1989), rather than, for example, the arguably more sophisticated methods of Fuller (1987). This is partly because structural equation models are easier to present to students and clients, and partly because of the availability of high-quality commercial software such as LISREL (Jöreskog and Sörbom, 1996), AMOS (Arbuckle, 2006) and SAS proc calis (SAS Institute, 1999). There is also a structural equation modelling package for R (Fox, 2006). Estimation and testing methods have been developed for categorical variables, both latent and observed (Lee and Xia, 2006; Muthén, 2002; Muthén and Muthén, 2006; Skrondal and Rabe-Hesketh, 2004). Our hope is that tools like these will soon become part of the statistical mainstream.

However, it is not just a matter of applying a new statistical method to the same old data. In many cases, a different kind of data set is required. The reason is that for even the simplest measurement error models, multiple indicators of the independent variables are required for the model to be identified; see for example the discussions by Fuller (1987) and Bollen (1989). A simple solution for linear regression with measurement error is measure each independent variable twice, preferably on two different occasions and using different methods or measuring instruments — perhaps as in Campbell and Fiske's (1959) "multi-trait multi-method matrix." If it can be assumed that the measurement errors on the two occasions are uncorrelated, scientists and undergraduates without much mathematical background should have no trouble using commercially available software to carry out a valid measurement error regression.

References

Arbuckle, J. L. (2006), *AMOS 7.0 User's Guide*. Chicago: SPSS Inc.

Austin, P. C. and Brunner, L. J. (2003), "Type I Error Inflation in the Presence of a Ceiling Effect," *American Statistician*, 57, 97-104.

Austin, P. C. and Brunner, L. J. (2004), "Inflation of the Type I error rate when a continuous confounding variable is categorized in logistic regression analysis," *Statistics in Medicine*, 23, 1159-1178.

Barnow, B. S. (1973), "The effects of Head Start and socioeconomic status on cognitive development of disadvantaged children." Doctoral dissertation, University of Wisconsin, Madison.

Bentler, P. M. and Woodward, J. A. (1978), "A Head Start re-evaluation: Positive effects are not yet demonstrable." *Evaluation Quarterly*, 2, 493-510.

Bollen, K. A. (1989), *Structural equations with latent variables,* New York: Wiley.

Campbell, D. T. and Erlbacher, A. (1970), "How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education programs look harmful," In J. Hellmuth (Ed.) *The disadvantaged child: Vol 3, Compensatory education: A national debate,* (Pp. 185-210) New York: Brunner/Mazel.

Campbell, D. T. and Fiske, D. W. (1959), "Convergent and discriminant validation by the multi-trait multi-method matrix," *Psychological Bulletin*, 56, 81-105.

Cheng, C. L. and Van Ness, J. W. (1999), *Statistical regression with measurement error,* London: Chapman & Hall.

Cicirelli, V. G. *et al.* (1969), *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development,* Athens, Ohio: Ohio University and Westinghouse Learning Corporation.

Cochran, W. G. (1968), "Errors of measurement in statistics." *Technometrics*, 10,

637-666.

Cox, D. R. (1972), "Regression models and life tables (with discussion)," *Journal of the Royal Statistical Society B*, 34, 187-202.

Fisher, R. A. F. (1938), *Statistical methods for research workers (7th ed.),* London: Oliver and Boyd.

Fox, J. (2006), "Structural equation modeling with the `sem` package in R," *Structural equation modelling*, 13, 465-486.

Fuller, W. A. (1987), *Measurement error models,* New York: Wiley.

Jöreskog, K. G. (1978), "Structural analysis of covariance and correlation matrices," *Psychometrika*, 43, 443-477.

Jöreskog, K.G., and Sörbom, D. (1996), *LISREL 8: Structural equation modelling with the SIMPLIS command language*. London: Scientific Software International.

Kaiser, H.F. (1960), "Directional Statistical Decisions," *Psychological Review*, 67, 160-167.

Lee, S. Y. and Xia, Y. M. (2006), "Maximum likelihood methods in treating outliers and symetrically heavy-tailed distributions for nonlinear structural equation models with missing data," *Psychometrika*, 71, 565-595.

Lord, F. M. (1960),"Large-sample covariance analysis when the control variable is fallible," *Journal of the American Statistical Association*, 55, 307-321.

Lord, F. M. and Novick, M. R. (1968), *Statistical theories of mental test scores,* Reading: Addison-Wesley.

Madansky, A. (1959), "The fitting of straight lines when both variables are subject

to error." *J. Am. Statist. Assoc.*, 54, 173-205.

Magidson, J. (1977), "Towards a causal model approach to adjusting for pre-existing differences in the non-equivalent control group situation: A general alternative to ANCOVA." *Evaluation Quarterly*, 1, 511-520.

Magidson, J. (1978), "Reply to Bentler and Woodward: The .05 level is not all-powerful." *Evaluation Quarterly*, 2, 399-420.

Maxwell, S..E. and Delaney, H. D. (1993), "Bivariate median splits and spurious statistical significance," *Psychological Bulletin* 113, 181-190.

McCallum, B. T. (1972), "Relative asymptotic bias from errors of omission and measurement," *Econometrica*, 40, 757-758.

Muthén, B. O. (2002), "Beyond SEM: General latent variable modelling. *Behaviormetrika*, 29, 81-117.

Muthén, L. K. and Muthén, B. O. (2006), *Mplus users guide (4th ed.).* Los Angeles: Muthén and Muthén.

R Development Core Team (2006), "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

SAS Institute, Inc. (1999), "SAS/STAT User's guide, Version 8." Cary, N. C.: SAS Institute, Inc. 3884 pp.

Scheffé, H. (1959), *The analysis of variance,* New York: Wiley.

Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models,* London: Chapman & Hall.

Stouffer, S. A. (1936), "Evaluating the effect of inadequately measured variables in partial correlation analysis." *J. Am. Statist. Assoc.*, 31, 348-360.

Wald A. (1940), "The fitting of straight lines if both variables are subject to error." *Ann. Math. Statist.*, 11, 284-300.

Wansbeek, T. J. and Meijer, E. (2000), *Measurement error and latent variables in econometrics,* New York: Elsevier.

Wickens, M. R. (1972), "A note on the use of proxy variables," *Econometrica*, 40, 759-761.