The Canadian Journal of Statistics
Vol. ??, No. ?, 2009, Pages ???-???
La revue canadienne de statistique

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

Jerry BRUNNER and Peter C. AUSTIN

Key words and phrases: Errors in variables; Measurement error; Monte Carlo; Structural equation models; Type I error rate.
MSC 2000: Primary 62J99; secondary 62H15.

Abstract: When independent variables are measured with error, ordinary least squares regression can yield parameter estimates that are biased and inconsistent. This paper documents an inflation of Type I error rate that can also occur. In addition to analytic results, a large-scale Monte Carlo study shows unacceptably high Type I error rates under circumstances that could easily be encountered in practice. A set of smaller-scale simulations indicate that the problem applies to various types of regression and various types of measurement error.

Title in French: we can supply this

Résumé : When independent variables are measured with error, ordinary least squares regression can yield parameter estimates that are biased and inconsistent. This paper documents an inflation of Type I error rate that can also occur. In addition to analytic results, a large-scale Monte Carlo study shows unacceptably high Type I error rates under circumstances that could easily be encountered in practice. A set of smaller-scale simulations indicate that the problem applies to various types of regression and various types of measurement error.

1. INTRODUCTION

This is a story about something everyone knows, but few seem to appreciate. It is well known that when standard regression methods are applied to data in which the independent variables are measured with random error, serious difficulties can arise. 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 & 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 & Woodward 1978; Magidson 1978), harmful effects are entirely ruled out, and disagreement is limited to whether the data provide evidence of a positive effect (not a negative effect) for White children receiving a summer-only version of the program. What we take from this example is that it is more difficult to get away with ignoring measurement error in data that have political significance.

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), Madansky (1959) and Wickens (1972) 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 & Van Ness (1999), Wansbeek & Meijer (2000) and Carroll, Ruppert, Stefanski & C. M. Crainiceanu (2006). Many common measurement error models are special cases of the structural equation models that have have long been popular in the social and biomedical sciences; see for example Jöreskog (1978), Bollen (1989), and the generalizations of Skrondal & Rabe-Hesketh (2004), Muthén (2002) and Muthén & Muthén (2006). Gustafson (2004) discusses the adjustment of more traditional regression methods to account for measurement error.

So, it is widely recognized that measurement error can present a problem for ordinary least-squares regression, and a class of high-quality alternatives is in place. But please glance at the regression text that is closest to hand. It may or may not contain a warning about measurement error, but look at the examples and sample data sets. You will be reminded that in practice, individuals at all levels of statistical sophistication are encouraged to apply ordinary least-squares regression to data where the predictor variables are obviously measured with error.

When we shield our students and clients from technical difficulties in this manner, presumably we are guided by the famous principle "Essentially all models are wrong, but some are useful." (Box & Draper, 1987, p. 424). But the operative term here is *some*; Box's rule was never intended to justify the use of imperfect models in situations where their imperfections almost guarantee false conclusions. This paper is a reminder of how misleading standard regression methods can be when the independent variables are measured with error.

The main message is that if two independent variables in a regression are correlated, measurement error in one of them can drastically inflate the Type I error rate in tests of the other predictor. This artifact is not observed in the case of simple regression, where ignoring measurement error in the single predictor yields an estimated slope that is asymptotically biased toward zero. Familiarity with this case can actually lull the data analyst into a false sense of security, for it can give the impression that the effect of measurement error generally is to weaken the apparent relationships among true (error-free) variables.

But with two or more independent variables, the effects of measurement error are more complex. In particular, traditional methods of "controlling" for risk factors or potential confounding (lurking) variables are only partly successful. The result is that even when a predictor variable of interest is conditionally independent of the response given the risk factor, the usual test may still reject the null hypothesis with high probability. 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 rate rather than bias, because tests of statistical significance 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 the Type I error rate, there is a problem — a problem that may be taken seriously by empirical scientists who are unmoved by calculations of asymptotic bias.

Of course there is a connection between asymptotic bias and Type I error rate. 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 power of the test to detect a non-existent effect goes to one, and the Type I error rate rate necessarily increases to unity. This accounts for passing references — for example by Fuller (1978, p. 55), Cochran (1968, p. 653) Carroll *et al.* (2006, pp. 52-53) — to incorrect Type I error rates and incorrect conclusions 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 2, we revisit a canonical example with two independent variables, discussed by Cochran (1968). Cochran showed that when all the random variables involved are normal, ignoring measurement error in one of the predictors can produce an inconsistent least-squares estimate of the regression coefficient for the other predictor. We show 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.

Section 3 describes a set of smaller-scale simulations. First, we present an example in which ignoring measurement error results in rejection of the null hypothesis virtually always when the null hypothesis is false — but with the sample regression coefficient having the wrong sign. Then, 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 rate 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 standard tools. This issue is particularly troublesome in epidemiologic studies (Fewell et al. 2007).

Modelling measurement error is preferable to ignoring it, and good solutions are available. However, the typical data set has only a single measurement of each predictor variable. In the absence of additional information, this means models that include measurement error will not be uniquely identified in the model parameters, so that consistent estimation of the model parameters is impossible. For linear regression with classical additive measurement error, a simple solution is to measure the independent variables twice. If it can safely be assumed that errors of measurement on different occasions are uncorrelated, appropriate methods can be applied in a routine manner.

2. INFLATION OF TYPE I ERROR RATE IN LINEAR REGRESSION

Consider a multiple regression model in which there are two independent variables, both measured with additive error (a classical measurement error model). 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

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

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

$$E\begin{bmatrix} X_{i,1} \\ X_{i,2} \end{bmatrix} = \begin{bmatrix} \kappa_1 \\ \kappa_2 \end{bmatrix} \qquad Var\begin{bmatrix} X_{i,1} \\ X_{i,2} \end{bmatrix} = \mathbf{\Phi} = \begin{bmatrix} \phi_{1,1} & \phi_{1,2} \\ \phi_{1,2} & \phi_{2,2} \end{bmatrix}$$

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

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

The true independent variables are $X_{i,1}$ and $X_{i,2}$, 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, let $X_{i,1}$ be true average minutes of exercise per day for subject i, and let $W_{i,1}$ be 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 $X_{i,1}$ is true amount of exercise and $W_{i,1}$ is reported amount of exercise, while $X_{i,2}$ is true consumption of snack food and $W_{i,2}$ is reported consumption. If people who exaggerate how much they exercise tend to under-report how much snack food they eat, the covariance parameter $\theta_{1,2}$ would be negative.

When a model such as (**??**) holds, all one can observe are the triples $(W_{i,1}, W_{i,2}, Y_i)$ for $i = 1, \ldots, n$. Even if all the intercepts and expected values were zero or known, it would be impossible to estimate the model parameters uniquely without additional information, and such information is usually unavailable in practice. We view this as a shortcoming of the data, not of the model. The model could well be approximately correct, but the variables that are measured do not allow all the model parameters to be recovered, even from an infinite number of cases.

Suppose the interest is in testing whether $X_{i,2}$ is related to Y_i, conditionally on the value of $X_{i,1}$; that is, the null hypothesis is $H_0 : \beta_2 = 0$. The parameters of Model (**??**) cannot be estimated from the available data, so the analyst takes $W_{i,1}$ as a surrogate for $X_{i,1}$ and $W_{i,2}$ as a surrogate for $X_{i,2}$, fits the incorrect model

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

by ordinary least squares, and (assuming ϵ_i normal) tests the null hypothesis $H_0 : \beta_2 = 0$ using the usual t or F-test. Gustafson (1994) calls this the "naive" approach, and indeed it is.

Notice that the same symbols are deliberately used for the regression coefficients in the correct Model (**??**) and the incorrect Model (**??**). This is because when there is measurement error in the independent variables, estimators and tests based on Model (**??**) are really just *ad hoc* methods for the regression coefficients of the more realistic but elusive Model (**??**).

Suppose that Model (**??**) is correct, and that $H_0 : \beta_2 = 0$ is true. We now observe that except under special circumstances, the least squares quantity $\widehat{\beta}_2$ based on Model (**??**) converges almost surely to a quantity different from zero 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.

2.1 Almost sure disaster

The ordinary least-squares estimate $\widehat{\beta}_2$ 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 (**??**). So by a continuity argument, the ordinary least-squares estimate converges almost surely to the corresponding composite function of the true model parameters.

Our focus is upon Type I error rate 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,

$$\widehat{\beta}_2 \stackrel{a.s.}{\to} \frac{\beta_1(\phi_{1,2}\theta_{1,1} - \phi_{1,1}\theta_{1,2})}{(\phi_{1,1} + \theta_{1,1})(\phi_{2,2} + \theta_{2,2}) - (\phi_{1,2} + \theta_{1,2})^2} = \frac{\beta_1(\phi_{1,2}\theta_{1,1} - \phi_{1,1}\theta_{1,2})}{|\mathbf{\Phi} + \mathbf{\Theta}|} \tag{3}$$

Expression (**??**) is the asymptotic bias of $\widehat{\beta}_2$ as an estimate of the true regression parameter β_2, in the case where $\beta_2 = 0$. It does not depend upon any of the expected value terms in Model (**??**), and it is zero only if $\beta_1 = 0$ or if $\phi_{1,2}\theta_{1,1} = \phi_{1,1}\theta_{1,2}$. The denominator will be positive if at least one of $\mathbf{\Phi}$ and $\mathbf{\Theta}$ are positive definite; this condition is required for convergence.

Note that if $\theta_{1,2} = \phi_{1,2} = 0$, there is no correlation between either the measurement errors or the latent independent variables. In this case there is no asymptotic bias in $\widehat{\beta}_2$ under the null hypothesis, and in Section 2.2 we will see that there is also no inflation of the Type I error rate.

The least-squares quantity $\widehat{\beta}_2$ is the numerator of the t-statistic commonly used to test $H_0 : \beta_2 = 0$ based on the incorrect Model (**??**). The denominator, the standard error of $\widehat{\beta}_2$, is just the square root of the quantity Mean Squared Error multiplied by the (3,3) element of $(\mathbf{X}'\mathbf{X})^{-1}$. Writing this denominator as $U_n/\sqrt{n}$ and using the same approach that led to (**??**), we find that

$$U_n \stackrel{a.s.}{\to} \frac{\sqrt{(\phi_{1,1} + \theta_{1,1})(\beta_1^2(\phi_{1,1}|\mathbf{\Theta}| + \theta_{1,1}|\mathbf{\Phi}|) + |\mathbf{\Phi} + \mathbf{\Theta}|(\sigma^2 + 2(\beta_0^2 + \beta_1\kappa_1)^2)}}{|\mathbf{\Phi} + \mathbf{\Theta}|}, \tag{4}$$

again provided that at least one of $\mathbf{\Phi}$ and $\mathbf{\Theta}$ is positive definite. Consequently, the denominator converges to zero, the absolute value of the t-statistic tends to infinity, and the associated p-value converges almost surely to zero. That is, we almost surely commit a Type I error.

2.2 A Monte Carlo study of Type I error rate inflation

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

Since Expression (**??**) 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, the latent independent variables were generated as follows:

$$\begin{aligned} X_1 &= \sqrt{1 - \phi_{1,2}}\, R_1 + \sqrt{\phi_{1,2}}\, R_3 \text{ and} \\ X_2 &= \sqrt{1 - \phi_{1,2}}\, R_2 + \sqrt{\phi_{1,2}}\, R_3, \end{aligned} \tag{5}$$

   yielding $Var(X_1) = Var(X_2) = 1$ and a correlation of $\phi_{1,2}$ between X_1 and X_2. A quiet but important feature of this construction is that when $\phi_{1,2} = 0$, X_1 and X_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 X_1:* With $\beta_1 = 1$, $\beta_2 = 0$ and $Var(X_1) = \phi_{1,1} = 1$ we have $Var(Y) = 1 + \sigma^2$, so that the proportion of variance in the dependent variable that comes from X_1 is $\frac{1}{1+\sigma^2}$. We used this as an index of the strength of relationship between X_1 and Y, and adjusted it by varying the value of σ^2. There were three values of explained variance: 0.25, 0.50 and 0.75.

4. *Reliability of W_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 (**??**), we have

$$[Corr(X_{i,1}, W_{i,1})]^2 = \left[\frac{\phi_{1,1}}{\sqrt{\phi_{1,1}}\sqrt{\phi_{1,1}+\theta_{1,1}}}\right]^2 = \frac{1}{1+\theta_{1,1}}.$$

   Thus one may manipulate the reliability by varying 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 W_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 X_1 and X_2 – see Equations (**??**) – come from a common 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, yielding 75 million simulated data sets. For each data set, we ignored measurement error, fit Model (**??**) and tested $H_0 : \beta_2 = 0$ with the usual t-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 captures their essential features, referring the reader to

Table 1: Estimated Type I error rates when independent variables and measurement errors are all normal, and reliability of W_1 and W_2 both equal 0.90

25% of Variance in Y is Explained by X_1

	Correlation Between X_1 and X_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 X_1

	Correlation Between X_1 and X_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 X_1

	Correlation Between X_1 and X_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.

`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 **??** 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 **??**, we see that except when the latent independent variables X_1 and X_2 are uncorrelated, applying ordinary least squares regression to the corresponding observable variables W_1 and W_2 results in a substantial inflation of the Type I error rate. As one would predict from Expression (**??**) with $\theta_{1,2} = 0$, the problem becomes more severe as X_1 and X_2 become more strongly related, as X_1 and Y become more strongly related, and as the sample size increases. We view the Type I error rates in Table **??** 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 W_1, and decrease with decreasing reliability of W_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.

3. Further Difficulties

3.1 Significance in the wrong direction

Consider Model (**??**) again. Let the covariance between X_1 and X_2 be positive, the partial relationship between X_1 and Y be positive, and the partial relationship between X_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 (**??**) 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 X_1 and X_2 and between X_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 (**??**) 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 X_1 explains 0.75 of the variance in Y), $\theta_{1,1} = 1$ (so that the reliability of W_1 is 0.50), and $\theta_{2,2} = \frac{1}{19}$ (so that the reliability of W_2 is 0.95).

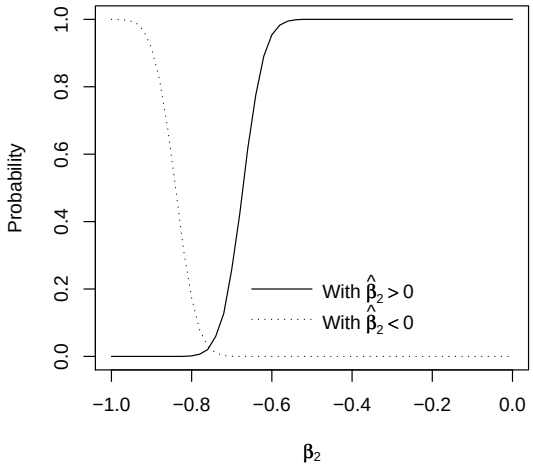
We then varied β_2 from minus one to zero, generating 10,000 data sets for each value of β_2. We fit Model (**??**) to each data set and tested $H_0 : \beta_2 = 0$ at $\alpha = 0.05$ with the usual t-test. Each test was classified as significant with $\widehat{\beta}_2 > 0$, significant with $\widehat{\beta}_2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

Figure **??** 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 it is negative. This example shows how ignoring measurement error in the independent variables can lead to firm conclusions that are directly opposite to reality.

3.2 The generality of the problem

We have illustrated inflation of the Type I error rate 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 the Type I error rate.* 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`.

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



Logistic regression with additive measurement error In this simulation – also see Fewell, Smith and Sterne (2007) – we constructed data sets with a pair of latent independent variables X_1 and X_2, and corresponding manifest variables W_1 and W_2, using a normal base distribution and the troublesome $\mathbf{\Phi}$ and $\mathbf{\Theta}$ values of Section 3.1. We then constructed a binary dependent variable Y, with the log odds of $Y = 1$ equal to $\beta_0 + \beta_1 X_1 + \beta_2 X_2$, where $\beta_0 = \beta_1 = 1$ and $\beta_2 = 0$. Ignoring measurement error, we fit a standard logistic regression model with the log odds of $Y = 1$ equal to $\beta_0 + \beta_1 W_1 + \beta_2 W_2$, and used a likelihood ratio test of $H_0 : \beta_2 = 0$. This is parallel to what we did with ordinary least squares regression.

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 & Brunner (2003) describe inflation of the Type I error rate 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 rate 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 & Delaney (1993) show how Type I error rate 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 rate does not depend upon the distribution (apart from the existence of moments). Median splitting the independent variables has also been shown to inflate the Type I error rate in

logistic regression (Austin & Brunner 2004).

Normal linear regression, ranking the independent variable We have unpublished work showing that in terms of Type I error rate, 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 inflation of the Type I error rate.

We constructed data sets according to Model (**??**) again using the $\mathbf{\Phi}$ values of Section 3.1, a reliability of 0.95 for W_2, a normal base distribution, $\beta_0 = \beta_1 = 1$ and $\beta_2 = 0$. However, the observable independent variable W_1 contained the ranks of X_1, rather than X_1 plus a piece of random noise. As usual, we fit the incorrect regression model (**??**) and tested $H_0 : \beta_2 = 0$ with the usual t-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* (Gustafson 2004), 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 X_1 and X_2, only this time they are binary. The corresponding observable variables W_1 and W_2 are also binary. There is a binary dependent variable Y that is dependent upon X_1 and conditionally independent of X_2.

The components of the measurement error model are two-way tables of the joint probabilities of X_1 and X_2, X_1 with W_1, and X_2 with W_2. The values we used are given in Table **??**.

Table 2: Joint probabilities for the classification error model

	X_1	
X_2	0	1
0	0.40	0.10
1	0.10	0.40

	W_1	
X_1	0	1
0	0.30	0.20
1	0.20	0.30

	W_2	
X_2	0	1
0	0.45	0.05
1	0.05	0.45

The data were constructed by first sampling an (X_1, X_2) pair from a multinomial distribution, and then simulating W_1 conditionally on X_1 and W_2 conditionally on X_2. Finally, Y was generated conditionally on X_1 using $P(Y = 0|X_1 = 0) = P(Y = 1|X_1 = 1) = 0.80$. Repeating this process $n = 250$ times yielded a simulated data set of (W_1, W_2, Y) triples. We then tested for conditional independence of W_2 and Y given W_1, as a surrogate for for the conditional independence of X_2 and Y given $X_!$. Specifically, we used R's `loglin` function to fit a hierarchical loglinear model with an association between W_1 and W_2, and between W_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 experiment, 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 the Type I error rate may be inflated in testing for a main effect of the other independent variable.

We started with two correlated binary latent independent variables X_1 and X_2, and their corresponding observable versions W_1 and W_2, constructed according to the same classification error model used for loglinear models; see Table **??**. We then generated the dependent variable as $Y = 1 + X_1 + \epsilon$, where ϵ is Normal with mean zero and variance $\frac{1}{4}$. Because X_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 X_1, Y is independent of X_2 and there is no interaction.

Repeating this process $n = 200$ times yielded a simulated data set of (W_1, W_2, Y) triples. As usual, we conducted the analysis using the observable variables W_1 and W_2 in place of X_1 and X_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 W_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 2, we saw that inflation of the Type I error rate 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 as a means of eliminating the correlation between independent variables. Unfortunately 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 W_1, W_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 W_2 (as a surrogate for X_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 rate extends to survival analysis. Proceeding as in earlier examples, we constructed data sets with a pair of latent independent variables X_1 and X_2, and also corresponding manifest variables using a normal base distribution and the the $\mathbf{\Phi}$ and $\mathbf{\Theta}$ values of Section 3.1. We then sampled the dependent variable Y from an exponential distribution with mean $\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$, with $\beta_0 = \beta_1 = 1$ and $\beta_2 = 0$. So again, Y is conditionally independent of X_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 W_1 and W_2 as the independent variables, testing the relationship of W_2 to Y controlling for W_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 the Type I error rate when measurement error in the independent variables is ignored.

4. DISCUSSION

We are not suggesting that ignoring measurement error *always* inflates the Type I error rate 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 diminish or magnify the effects of others. Still, one cannot escape the conclusion that measurement error

in the independent variables may inflate the Type I error rate to an unacceptable degree. Given this, it seems unduly optimistic to continue applying standard regression and related methods in the presence of obvious measurement error.

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 & 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 & Xia 2006; Muthén 2002; Muthén & Muthén 2006; Skrondal & 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 new statistical methods 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 measurements of the 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 & 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.

ACKNOWLEDGEMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

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

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

P. C. Austin & L. J. Brunner (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.

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

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

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

G. E. P. Box & N. R. Draper (1987). *Empirical Model-Building and Response Surfaces.* New York: Wiley.

D. T. Campbell & A. Erlbacher (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.

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

R. J. Carroll, D. Ruppert, L. A. Stefanski, & C. M. Crainiceanu (2006). *Measurement error in nonlinear models: a modern perspective. (2nd. ed.)* Boca Raton, FL : Chapman & Hall/CRC.

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

V. G. Cicirelli *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.

W. G. Cochran (1968). Errors of measurement in statistics. *Technometrics*, 10, 637–666.

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

Z. Fewell, G. D. Smith, and J. A.. C. Sterne (2007). The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study. *American Journal of Epidemiology*, 166, 646–655.

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

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

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology - Impacts and Bayesian Adjustments.* Chapman & Hall/CRC, Boca Raton, USA.

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

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

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

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

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

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

A. Madansky (1959). The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, 54, 173–205.

J. Magidson (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.

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

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

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

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

L. K. Muthén & B. O. Muthén (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.

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

S. A. Stouffer (1936). Evaluating the effect of inadequately measured variables in partial correlation analysis. *Journal of the American Statistical Association*, 31, 348–360.

A. Wald (1940). The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics*, 11, 284–300.

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

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

Received ???
Accepted ???

Jerry BRUNNER: `brunner@utstat.toronto.edu`
Department of Statistics, University of Toronto
Toronto. Ontario
Canada, M5S 3G3

Peter C. AUSTIN: `peter.austin@ices.on.ca`
Institute for Clinical Evaluative Sciences
Toronto. Ontario
Canada, M4N 3M5