

Multicollinearity and Model Misspecification

Christopher Winship, Bruce Western

Harvard University

Abstract: Multicollinearity in linear regression is typically thought of as a problem of large standard errors due to near-linear dependencies among independent variables. This problem can be solved by more informative data, possibly in the form of a larger sample. We argue that this understanding of multicollinearity is only partly correct. The near collinearity of independent variables can also increase the sensitivity of regression estimates to small errors in the model misspecification. We examine the classical assumption that independent variables are uncorrelated with the errors. With collinearity, small deviations from this assumption can lead to large changes in estimates. We present a Bayesian estimator that specifies a prior distribution for the covariance between the independent variables and the error term. This estimator can be used to calculate confidence intervals that reflect sampling error and uncertainty about the model specification. A Monte Carlo experiment indicates that the Bayesian estimator has good frequentist properties in the presence of specification errors. We illustrate the new method by estimating a model of the black–white gap in earnings.

Keywords: bias; multicollinearity; model misspecification

IN a classic but now seldom-studied article published in *The American Journal of Sociology* in 1968 entitled “Issues in Multiple Regression,” Robert Gordon analyzed the problems involved in model specification in regression analysis when variables are nearly collinear. He presents a particularly troubling example in which two independent variables correlate at 0.80. These predictors correlate with the dependent variable at 0.60 and 0.55. They have identical correlations with the other predictors in the model. He reports that the ordinary least squares (OLS) estimates for the standardized regression slopes for these two variables are, respectively, 0.38 and 0.13. Not surprisingly, he is disturbed that a small difference in the two predictors’ correlation with the dependent variable, 0.60 versus 0.55, results in a large difference in their estimated effects.

How are we to understand this example? Gordon considers two possibilities. First, he notes that the predictors’ true correlations with Y might be equal and that their observed difference might be due to sampling error. In this case, he observes that sampling error should be reflected in the standard errors for the two coefficients. In principle, the standard errors should be sufficiently large to indicate that the two regression coefficients might be equal. More formally, a test of their equality could be constructed. As numerous textbook authors have argued, however, when predictors are collinear, one’s sample size may not be large enough to provide a test of sufficient power to reject the null hypothesis of equality. The obvious (though often infeasible) solution is to increase sample size.

Second, Gordon notes that even if 0.60 and 0.55 were the true correlations of the two independent variables with the dependent variable, changing other predictors that had different correlations with the two variables could substantially affect the

Citation: Winship, Christopher, and Bruce Western. 2016. “Multicollinearity and Model Misspecification.” *Sociological Science* 3: 627–649.


Received: February 5, 2016

Accepted: March 5, 2016

Published: July 26, 2016

Editor(s): Jesper Sørensen, Olav Sorenson

DOI: 10.15195/v3.a27

Copyright: © 2016 The Author(s). This open-access article has been published under a Creative Commons Attribution License, which allows unrestricted use, distribution and reproduction, in any form, as long as the original author and source have been credited. 

estimates. The results may be sensitive to the model specification. This sensitivity is due to collinearity, not sample size.

Specification sensitivity is worrying for two reasons. First, the sensitivity may not be obvious because collinearity can be difficult to detect (Belsley 1991). Detection can be difficult in the presence of strong multiple—rather than bivariate—correlations among predictors. Second, social scientists are usually uncertain about the correct model specification. In practice, the true model is virtually never known with certainty. A variety of models are usually consistent with a particular theory.

Within the traditional frequentist perspective, collinearity has been analyzed as a problem of imprecise estimates or equivalently large standard errors resulting from having weak or too little data.¹ Goldberger's widely used textbook (1991) describes at length how the problem of multicollinearity is analogous to the problem of small sample size, or what he jokingly calls micronumerosity. Specifically, with multicollinearity, sample data may be too uninformative to estimate regression coefficients precisely. Consequently, some estimates may be nonsensical. From a frequentist perspective, this is simply due to the fact that the confidence interval associated with an estimate is so large that it includes unreasonable estimates. For Bayesians, the researcher may have failed to include sufficient prior information to restrict estimates to a realistic range (Leamer 1994). The cure for harmful multicollinearity is either additional information in the form of more sample data or stronger priors (nonsample) information for the coefficients of interest (Belsley 1991).

The standard analysis of multicollinearity as simply weak or uninformative data ignores Gordon's second observation—that estimates from collinear data may be highly sensitive to the model specification, even in large samples. This sensitivity is not captured by traditional standard errors and is independent of the amount of sample information. This article formally explores the consequence of model misspecification in the linear regression model in the presence of multicollinearity. Specifically, we develop Gordon's insight that multicollinearity can create problems not only when one has too little data but also when the model is misspecified. As far as we are aware, with the exception of Mela and Kopalle (2002), this is not an issue that has been formally analyzed in the research literature.² Our interest is in situations where the research focus is on the relative and/or absolute size of specific regression coefficients and the intent is to interpret regression coefficients as estimates of the causal effects of variables. For example, we might be interested in the effect of family background on occupational attainment or race's effect on earnings.

Model specification uncertainty is not easily analyzed from a traditional frequentist perspective. For frequentists, models are either true or false. In the frequentist framework, there is no possibility of specifying a model that is true with some probability (cf. Leamer 1978; Draper 1994). Model misspecification, however, can be analyzed in a frequentist perspective. Essentially, one assumes that there is a true model and asks what the consequences are of estimating the wrong model. A common example of this is found in textbooks on regression where the effects of omitting an independent variable are analyzed. More general analysis is also possible (e.g., White 1994).

This article has two goals. First, from a frequentist perspective, we carry out a simple analysis demonstrating the consequences of model misspecification in the presence of multicollinearity. The critical insight is that multicollinearity can enormously magnify the effects of model misspecification. A key implication, counter to the traditional perspective, is that multicollinearity may be particularly dangerous with large sample sizes. Here, one's sampling standard errors may be quite small, giving the researcher confidence in their estimates. In fact, however, small changes in the specification of the model may radically change estimates.

Second, from a Bayesian perspective, we describe a method that allows for model uncertainty. We almost never know the true model exactly: misspecification is the norm, not the exception. How then should we deal with the potential effects of misspecification in the presence of multicollinearity? In our approach, a Bayesian prior distribution is used to propagate uncertainty about the correct model specification through its standard errors. The Bayesian standard errors reflect both uncertainty because of sampling error and uncertainty about the model. As a result, the Bayesian standard errors are larger than the classical standard errors that reflect only sampling variability.

In important respects, our approach challenges much of the traditional literature on multicollinearity.³ Given that multicollinearity can produce imprecise estimates, the traditional approach proposes methods to reduce standard errors. In fact, however, beyond adding information either by increasing the sample size or introducing prior information, little can be done. In our view, traditional standard errors that fail to reflect model uncertainty are too small, not too large. Such standard errors may lead researchers to be unjustifiably confident about the precision of estimates. This overconfidence may be most acute with large samples, exactly the situation in which the traditional literature argues multicollinearity is less likely to be a problem (Goldberger 1991).⁴ Our goal is to construct standard errors that incorporate model uncertainty and thus are necessarily larger than traditional sampling standard errors.

The next section of the article presents a basic analysis of model misspecification in the classical regression model. We demonstrate the sensitivity of classical regression estimates to model misspecification when there is multicollinearity. The following section develops a Bayesian solution to the problem of incorporating specification uncertainty into a model. Next, we report results from a Monte Carlo experiment that shows that the Bayesian regression can perform much better than OLS in the presence of specification errors. Following this, we present an empirical example. We conclude by arguing that it is important that we have methods that incorporate specification uncertainty as well as sampling error into our analysis. This is essential if we are to have confidence intervals that accurately reflect the range of estimates that are consistent with the data.

Model Misspecification and Multicollinearity

The “Classical” Linear Regression Model

In this article, we will only deal with the standard linear regression model:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \quad (1)$$

where \mathbf{y} is an $n \times 1$ vector consisting of observations on the dependent variable, \mathbf{X} is an $n \times k$ matrix of the k independent variables, \mathbf{b} is a $k \times 1$ vector of regression coefficients to be estimated, and \mathbf{e} is an $n \times 1$ vector of unobserved errors that are assumed to be independent, identically distributed, normal variables. Without loss of generality, we will assume that both \mathbf{y} and \mathbf{X} have been centered to have mean zero. Similarly, we will assume that the data are generated by simple random sampling. Interest focuses on estimating the regression coefficients, \mathbf{b} , in Equation (1). Pre-multiplying each side of Equation (1) by \mathbf{X}' and rearranging terms yields,

$$\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{e} = \mathbf{X}'\mathbf{X}\mathbf{b} \quad (2)$$

Obviously, $\mathbf{X}'\mathbf{e}$ cannot be estimated since \mathbf{e} is unobserved. Note that Equation (2) consists of k equations with $2k$ unknowns—the k regression coefficients, \mathbf{b} , and the k cross-products of the error term \mathbf{e} with each of the columns of \mathbf{X} . The cross-products vector, $\mathbf{X}'\mathbf{e}$, equals the sample size, n , times the covariance of \mathbf{e} with each column of \mathbf{X} . Rearranging terms to solve for \mathbf{b} yields:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{e}) \quad (3)$$

The coefficient vector, \mathbf{b} , cannot be solved for explicitly in Equation (3) without further assumptions. The least squares estimator starts by assuming that in expectation $\mathbf{X}'\mathbf{e} = \mathbf{0}$. OLS then assumes that in the particular sample being analyzed that in fact the covariances of the error, \mathbf{e} , with each of the k predictors equals zero; that is, $\mathbf{X}'\mathbf{e} = \mathbf{0}$, giving:

$$\begin{aligned} \mathbf{b}_{OLS} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{V}_{XX}^{-1}\mathbf{v}_{Xy} \end{aligned} \quad (4)$$

where \mathbf{V}_{XX} is the sample covariance matrix for \mathbf{X} and \mathbf{v}_{Xy} is a vector of the sample covariances between \mathbf{X} and \mathbf{y} .

We obtain the OLS solution for \mathbf{b} by assuming that $E[\mathbf{X}'\mathbf{e}] = \mathbf{0}$. To see the importance of this assumption, replace \mathbf{y} in Equation (4) with the right hand side of Equation (1). Taking the expectation of both sides, one gets:

$$\begin{aligned} E[\mathbf{b}_{OLS}] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\mathbf{b} + \mathbf{e})] \\ &= \mathbf{b} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}] \end{aligned} \quad (5)$$

This analysis is traditional and can be found in many textbooks. It shows that OLS is unbiased when the correlations between \mathbf{X} and \mathbf{e} in the population is zero, resulting in $E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}] = \mathbf{0}$. If this is not the case, then \mathbf{b}_{OLS} will be a biased estimate of \mathbf{b} .

\mathbf{X} and \mathbf{e} may be correlated in a specific sample for two reasons. First, \mathbf{X} and \mathbf{e} may be uncorrelated in the population, but the sample correlation between \mathbf{X} and \mathbf{e} might differ, perhaps substantially, from zero. This source of error would be reflected in the sampling errors of the parameter estimates. A second possibility is that \mathbf{X} and \mathbf{e} are correlated in the population. In this case, if the correlation is large or, as we will see below, if there is multicollinearity, \mathbf{b}_{OLS} can be highly biased.

A Generalized OLS Estimator

The traditional OLS estimator can be generalized to situations in which the population correlation or covariance between \mathbf{X} and \mathbf{e} is some other fixed vector of values besides zero. Let \mathbf{c}_0 be a possibly nonzero vector representing our assumption based on prior information (or guess) about what the covariance between \mathbf{X} and \mathbf{e} —that is, $\mathbf{X}'\mathbf{e}/n$ —equals in the population. Then substituting for $\mathbf{X}'\mathbf{e}$ in Equation (3), the analogue to Equation (4) would be:

$$\begin{aligned}\mathbf{b}_{GOLS} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y} - n\mathbf{c}_0) \\ &= \mathbf{V}_{XX}^{-1}(\mathbf{v}_{Xy} - \mathbf{c}_0)\end{aligned}\quad (6)$$

We refer to this as the generalized OLS estimator. Note that it is equivalent to the standard OLS estimator, except that we adjust $\mathbf{X}'\mathbf{y}$ by subtracting out the “correction” factor $n\mathbf{c}_0$, n times our assumed covariance between \mathbf{X} with \mathbf{e} . Equation (6) reduces to Equation (4) if $\mathbf{c}_0 = \mathbf{0}$.

To understand the properties of Equation (6), substitute for \mathbf{y} in Equation (6) using the right-hand side Equation (1), giving:

$$\mathbf{b}_{GOLS} = \mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{e} - n\mathbf{c}_0)\quad (7)$$

Taking expectations of both sides gives one:

$$E[\mathbf{b}_{GOLS}] = \mathbf{b} + [(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{e} - n\mathbf{c}_0)]\quad (8)$$

Equation (8) shows that \mathbf{b}_{GOLS} will be an unbiased estimate of \mathbf{b} if $E[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{e} - n\mathbf{c}_0)] = \mathbf{0}$. A sufficient condition for this to be true is that $E(\mathbf{X}'\mathbf{e} - n\mathbf{c}_0) = \mathbf{0}$.

Why might $(\mathbf{X}'\mathbf{e} - n\mathbf{c}_0) \neq \mathbf{0}$ in a particular sample (that is, $\mathbf{X}'\mathbf{e}$ and $n\mathbf{c}_0$ not be equal)? To answer this question, let \mathbf{c} equal the true covariance between \mathbf{X} and \mathbf{e} in the population as opposed to \mathbf{c}_0 , which is our assumed value of the covariance between \mathbf{X} and \mathbf{e} . Rewrite Equation (7) as follows:

$$\mathbf{b}_{GOLS} = \mathbf{b} + n(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{s} + \mathbf{m})\quad (9)$$

where the sampling error

$$\mathbf{s} = \mathbf{X}'\mathbf{e}/n - \mathbf{c}$$

and the specification error

$$\mathbf{m} = \mathbf{c} - \mathbf{c}_0$$

There are two reasons why $\mathbf{s} + \mathbf{m}$ might not approximate zero in the sample. First, \mathbf{s} may not be approximately zero—or equivalently, $\mathbf{X}'\mathbf{e}/n \not\approx \mathbf{c}$ —because of sampling

error. Under suitable regularity conditions, the law of large numbers guarantees that the error will be small in the sense of consistency. This means that as our sample grows, the probability that the sample value of $\mathbf{X}'\mathbf{e}$ will differ from its true value by an arbitrarily small amount goes to zero.⁵ In finite samples, $\mathbf{X}'\mathbf{e}/n$ may differ significantly from \mathbf{c} . This difference induces sampling variance in \mathbf{b}_{GOLS} .

Second, \mathbf{m} may not be approximately zero; that is, $\mathbf{c} \neq \mathbf{c}_0$. A large difference between \mathbf{c} and \mathbf{c}_0 indicates we have incorrectly specified the limiting behavior of the covariance between \mathbf{X} and \mathbf{e} . For instance, our regression model will be misspecified if we assume that $\mathbf{X}'\mathbf{e}/n$ is zero when in fact it is not. It is critical to note that the difference $\mathbf{c} - \mathbf{c}_0$ is not a function n . Model misspecification error (\mathbf{M}) is not affected by the sample size. This is in sharp contrast to the sample error (\mathbf{S}), which will by the law of large numbers go to zero in probability as the sample size goes to infinity.

Traditional OLS assumes a model in which the independent variables and error term are uncorrelated, $\mathbf{X}'\mathbf{e}/n = \mathbf{c}_0 = \mathbf{0}$. This is the key assumption in OLS and its potential Achilles heel. There are a variety of reasons why it may be violated: there are X s that have been omitted that affect y and are correlated with the X s in the model, some or all of the X s contain measurement error, y and some of the X s simultaneously determine each other, or there is selection on the dependent variable. Indeed, simultaneity and selection bias can be thought of as inducing special kinds of omitted variable bias. All these specification errors produce correlations between \mathbf{X} and \mathbf{e} . In these cases, OLS is inconsistent. Consistent estimates can be obtained by correctly assuming the covariance between \mathbf{X} and \mathbf{e} . If we know, approximately at least, the covariance between \mathbf{X} and \mathbf{e} , we can correct misspecification from a variety of sources with the generalized OLS estimator. Thus, a potential strength of the generalized ordinary least squares estimator is its ability to deal with a wide set of specification problems in the standard linear regression model.

The validity of any set of OLS estimates rests with the plausibility of our assumptions that $\mathbf{X}'\mathbf{e}/n = \mathbf{c}_0 = \mathbf{0}$. The researcher's job is to argue that any deviations are likely to be small, and it is the critic's responsibility to demonstrate why they are likely to be large. In almost all cases, it is implausible to assume that $\mathbf{X}'\mathbf{e}/n$ is exactly zero. There are always some omitted independent variables that may have some, hopefully small, effect on y and are correlated with the included predictors. In social science data, most variables are measured with error. Sometimes simultaneity or selection bias is an issue. All of these problems lead to situations in which \mathbf{X} and \mathbf{e} are likely to be at least weakly correlated. The hope (and prayer) in using OLS is that these problems are sufficiently minimal so that $\mathbf{X}'\mathbf{e}/n$ will be close enough to zero that \mathbf{b}_{OLS} will estimate the true regression parameter, \mathbf{b} , reasonably well to within sampling error.

If \mathbf{c} was known or accurately approximated, then we could use the generalized OLS estimator in Equation (6) and perform OLS on the adjusted covariance of \mathbf{X} and \mathbf{y} , $(\mathbf{X}'\mathbf{y}/n - \mathbf{c})$. Unfortunately, we almost never have precise knowledge about the adjustment factor. Below, we show how Bayesian methods can be used to deal with uncertainty about the true value of \mathbf{c} .

The Consequences of Multicollinearity

Multicollinearity occurs when predictors in a linear regression are nearly linearly dependent. This may occur because two variables are highly correlated or because one variable is well approximated by a linear function of other independent variables. Multicollinearity is a property of the $\mathbf{X}'\mathbf{X}$ matrix. With centered data, $\mathbf{X}'\mathbf{X}$ equals n times the covariance matrix of the independent variables, \mathbf{X} . A variety of multicollinearity diagnostics have been proposed. These will not concern us here (Belsley 1991 provides a comprehensive review).

Consider Equation (9) again:

$$\mathbf{b}_{GOLS} = \mathbf{b} + n(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{s} + \mathbf{m}) \quad (9)$$

In this equation, $(\mathbf{X}'\mathbf{X})^{-1}$ multiplies the vector $(\mathbf{s} + \mathbf{m}) = [(\mathbf{X}'\mathbf{e}/n - \mathbf{c}) + (\mathbf{c} - \mathbf{c}_0)]$. When multicollinearity is a problem, it is analogous to a column of $\mathbf{X}'\mathbf{X}$ matrix being nearly equal to zero. Matrix inversion is similar to the division by a scalar. Like division when we divide by a number close to zero, when there is multicollinearity and we multiply by $(\mathbf{X}'\mathbf{X})^{-1}$, we may tremendously increase the size of the quantity being multiplied.⁶

So, although the difference between $\mathbf{X}'\mathbf{e}/n$ and \mathbf{c} due to sampling error may be quite small, $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{e}/n - \mathbf{c})$ may be large because of the effect of multicollinearity in $\mathbf{X}'\mathbf{X}$.

Note that $(\mathbf{X}'\mathbf{X})^{-1}$ multiplies two terms: $(\mathbf{X}'\mathbf{e}/n - \mathbf{c})$, the sampling error, and $(\mathbf{c} - \mathbf{c}_0)$, the specification error. Expanding Equation (9),

$$\mathbf{b}_{GOLS} = \mathbf{b} + \mathbf{s}^* + \mathbf{m}^* \quad (10)$$

where

$$\mathbf{s}^* = n(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{e}/n - \mathbf{c}_0)$$

and

$$\mathbf{m}^* = n(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{c}_0 - \mathbf{c})$$

Although nonstandard, this decomposition is key to understanding the potential effects of collinearity. The vector \mathbf{s}^* represents the error in \mathbf{b}_{GOLS} because of sampling—the traditional concern about multicollinearity. The vector \mathbf{m}^* represents error because of model misspecification. If there is multicollinearity, small departures from zero in either $(\mathbf{X}'\mathbf{e}/n - \mathbf{c}_0)$ or $(\mathbf{c}_0 - \mathbf{c})$ can lead to large departures in \mathbf{b}_{GOLS} from \mathbf{b} because of multiplication by $(\mathbf{X}'\mathbf{X})^{-1}$.⁷

As noted above, the traditional statistics and econometrics literature focuses on the effects of multicollinearity on sampling error. Why hasn't this literature been concerned with misspecification error? The traditional literature assumes that in the population $\mathbf{X}'\mathbf{e}/n = \mathbf{c}_0 = 0$ exactly, thus making the problem of misspecification disappear. Of course, it is never the case that social scientists have perfectly specified models, making this strict assumption implausible.

Hypothetical Examples

The following example helps sharpen intuitions about the problem of multicollinearity and model misspecification. Table 1 shows the covariance/correlation matrix for two predictors, x_1 and x_2 , and a dependent variable y . Because all the variables except the error term have variance one, the covariances are equivalent to correlations. Two columns have been used for the error term, e , to indicate its correlation and covariance with each of the other variables. Because the R^2 in this model is relatively small (0.1825), the covariance between each variable and the error term and their correlation are approximately equal. As described in our discussion of Equation (6), adjusting $X'y$ by subtraction of n times the covariance of X and e yields consistent estimates of the coefficients. Applying Equation (4), the OLS estimates are: $\hat{b}_1 = 0.25$ and $\hat{b}_2 = 0$. These estimates are independent of sample size. The standard errors for these regression estimates can be made arbitrarily small by choosing n to be sufficiently large. In the extreme, we could assume that n is infinity, making the standard errors of the estimates zero. What is astounding in this example is that the OLS estimates suggest that x_2 has no effect on y , whereas the partial effect of x_1 is equal to its bivariate correlation. This is remarkable given that the correlations of the two predictors with y only differ by 0.05.

Now, consider the problem of model misspecification. In this example, our two predictors, x_1 and x_2 , covary at 0.025 and -0.025 with the error term. Covariances of this size might be created by omission of a third predictor, x_3 . For example, these correlations would occur if the effect of x_3 on y was 0.125 and if x_3 correlated with x_1 and x_2 at 0.20 and -0.20 . The omitted predictor might also be a composite variable consisting of all omitted variables. The effect of this omitted variable might be quite large in standardized units—say, 0.70—accounting for approximately half of the variance in y . In this case, the respective correlations would only need to be 0.036 and -0.036 to induce covariances of the above size between the included predictors and e .

We now use the generalized OLS estimator to correct for the correlation between the predictors and e . Applying Equation (6), this is equivalent to adjusting the correlation of x_1 and x_2 with y (0.25 and 0.20) by subtracting out their covariances with e (0.025 and -0.025) and then carrying OLS. Our new estimates are then $\hat{b}_1 = \hat{b}_2 = 0.125$. Notably, a modest degree of model misspecification induces a large difference between the OLS estimates and the true coefficients. The predictors x_1 and x_2 have small correlations with the error of ± 0.028 . In practice, it would often seem untenable to claim that the correlations of independent variables with the error are smaller than this.

In this example, the covariances between the predictors and the error term are equal in magnitude but opposite in sign. If we write this covariance as c and the covariance between the two predictors as r , then the bias in the OLS estimates for the two coefficients is $c/(1-r)$ and $-c/(1-r)$. Figure 1 plots the bias in the OLS estimate of the slope for x_1 as a function of collinearity between the predictors, r , given a fixed level of model misspecification, $c = 0.025$. Bias is expressed as a percentage of the true coefficient value, 0.125. Bias in OLS increases nonlinearly with increasing collinearity. Very high levels of collinearity can produce extremely large biases even when the model misspecification is very modest as in this example.

Table 1: Covariance/Correlation Matrix for Example 1.

		Correlation			Covariance	
	y	x_1	x_2	e	e	
y	1.00	0.25	0.20	0.900	0.818	
x_1		1.00	0.80	0.028	0.0250	
x_2			1.00	-0.028	-0.0250	

If the predictors are uncorrelated, however, the bias simply equals the covariance between the predictor and the error. While bias because of model misspecification in OLS increases nonlinearly with collinearity, for a fixed level of collinearity, bias only increases linearly with misspecification. In this sense, OLS estimates are more sensitive to collinearity than model misspecification.

In sum, multicollinearity can substantially affect the robustness of OLS estimates to model misspecification. Small errors in model misspecification can lead to large biases in parameter estimates. Note that this has nothing to do with sample size or classically calculated standard errors, as in the traditional discussion of multicollinearity. Increasing the sample size has no effect on the biases observed in this example. We can have a sample of infinite size, sampling standard errors of zero, and we would have the same problem. This suggests that multicollinearity may actually be most dangerous with large data sets. In this case, because our sampling standard errors are small, we may well believe that multicollinearity is not a problem when in fact small changes in our model specification might have a large effect on our estimates.

Modeling Specification Uncertainty

Sensitivity Analysis

Collinearity increases the sensitivity of estimates to the model misspecification. What can be done? Sensitivity analysis provides the simplest solution. In this approach, the researcher studies how estimates change with the model specification. Although sensitivity analysis is often valuable, it does not provide a way of calculating standard errors that incorporates model uncertainty or a method for hypothesis testing that allows for the uncertainty (although cf. Leamer 1983). We now discuss a Bayesian approach that allows for the incorporation of model uncertainty into estimates of standard errors.

Bayesian Approach

Bayesian regression is rarely used because typical applications make stronger assumptions than OLS.⁸ Our approach, however, uses Bayesian methods to relax the OLS assumptions. We use a Bayesian prior distribution to allow for uncertainty about the model specification. This approach is similar in spirit to Raftery's (1995) work on model uncertainty, though the technical specifics are different. Like

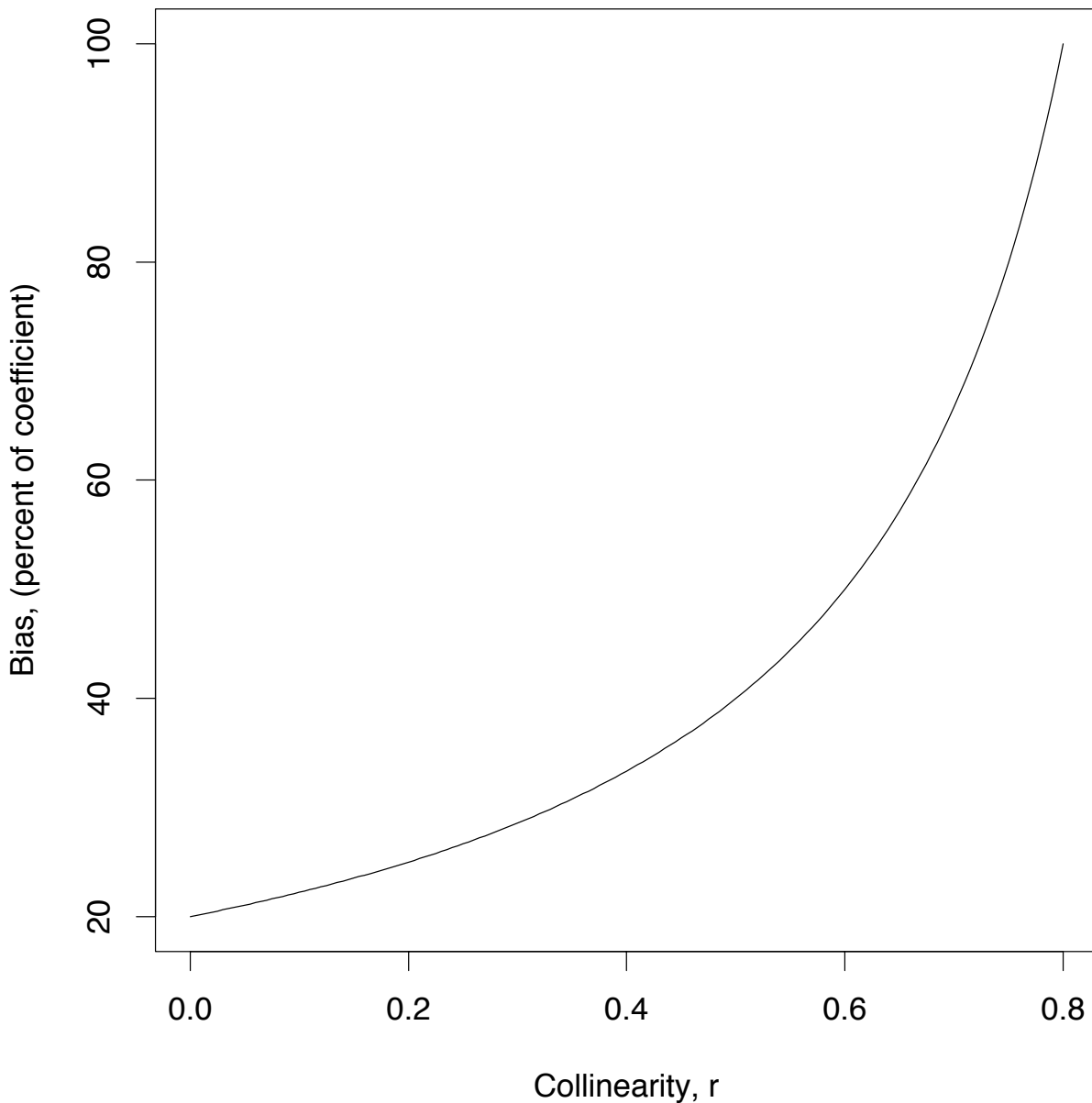


Figure 1: Bias in OLS coefficient estimates as a function of collinearity for a misspecified model. Collinearity (r) is measured by the covariance between the two predictors in the regression. Misspecification is indexed by the covariance between the predictors and the errors ($c = 0.025$).

Raftery's work, our method relaxes the assumption that the true model is known with certainty. This causes standard errors to increase. With these weaker assumptions, our method may well uncover situations where multicollinearity is a problem, but appears not to be—where classical standard errors are small, but estimates are not robust to different model specifications. This is particularly likely to be the case

when the data matrix appears to be highly informative—for example, when sample size is large.

The problem with OLS is that the standard errors of the coefficients only reflect sampling error. There is no way to incorporate uncertainty associated with the model specification. In this section, we utilize Bayesian methods to provide for the possibility of uncertainty in the model specification. In doing so, we are able to produce confidence intervals for parameters that reflect both sampling error and potential misspecification error.⁹

Again, consider Equation (6), the generalized ordinary least squares estimator:

$$\begin{aligned} \mathbf{b}_{GOLS} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y} - n\mathbf{c}_0) \\ &= \mathbf{V}_{XX}^{-1}(\mathbf{v}_{Xy} - \mathbf{c}_0) \end{aligned} \quad (6)$$

This model can be thought of as having two sets of parameters: \mathbf{b} , the regression coefficients, and \mathbf{c} , the correction factor for $\mathbf{X}'\mathbf{y}/n$ —equal to the covariance between \mathbf{X} and \mathbf{e} —that is used to account for the fact that the predictors and the errors may be correlated. As specified, this model is unidentified without further assumptions about either \mathbf{b} or \mathbf{c} . The traditional frequentist approach assumes a particular set of values for \mathbf{c} to solve for \mathbf{b} . This is a very strong assumption. From a Bayesian perspective, the frequentist approach is peculiar: it amounts to assuming that we have no prior information at all about the true value of \mathbf{b} , but that we have absolutely perfect information about \mathbf{c} , typically that $\mathbf{c} = \mathbf{0}$.

In the Bayesian approach, we assume that the values of \mathbf{b} and \mathbf{c} are unknown, but that our beliefs about their values can be described by a prior distribution. If at least some of our beliefs are represented by a proper probability distribution, then the posterior distribution of our parameters, \mathbf{b} , will also be proper.¹⁰ This means that unidentified models can be analyzed by Bayesian methods by imposing distributional assumptions on parameters, whereas in a classical frequentist approach, strict constraints would have to be imposed on these parameters in order to identify the model. This fact is well known but has seen only limited application (Neath and Samaniego 1996; Kadane 1975; Dreze 1975; Lindley and El-Sayyad 1968).

Why do proper priors yield proper posteriors? Assume that we have no prior information regarding \mathbf{b} . Using the generalized OLS estimator (Equation [5]), we can calculate the generalized OLS estimate of \mathbf{b} for any value of \mathbf{c} . Now if we think of \mathbf{c} not as a specific value but instead as a distribution of values, then for any particular value of \mathbf{c} in the distribution we can estimate \mathbf{b} . The distribution of \mathbf{c} describes our degree of belief in different values of \mathbf{c} . With this probability distribution, we can derive the probability of different values of \mathbf{b} . The probability of different values of \mathbf{b} will be a function of potential sampling error and the probability of different values of \mathbf{c} .¹¹

Posterior Mean and Covariances

For the purposes of this article, we assume that we have no prior information about \mathbf{b} . Equivalently, \mathbf{b} is given an improper diffuse prior. A standard result shows that with this prior, assuming $\mathbf{c} = \mathbf{0}$ and normal errors, the Bayesian posterior means for \mathbf{b} equal the OLS estimates (Gelman et al. 1995). To incorporate model uncertainty,

we place a proper informative prior on \mathbf{c} . We specify this prior distribution to be normal and centered at \mathbf{c}_0 , with prior covariance matrix \mathbf{U}_0 . Generalization to the case where \mathbf{b} has a proper prior is straightforward.

In Appendix A of the online supplement, we derive the posterior means and covariance matrix for \mathbf{b} under the above assumptions. The posterior mean of the regression coefficients is given by:

$$\mathbf{b}_{Bayes} = \mathbf{b}_{OLS} - n(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}_0 \quad (11)$$

and the posterior covariance matrix is

$$V(\mathbf{b}_{Bayes}) = \mathbf{R} + n^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{U}_0(\mathbf{X}'\mathbf{X})^{-1} \quad (12)$$

where \mathbf{R} is the OLS estimate of covariance matrix for \mathbf{b} . The posterior mean is simply the OLS estimate of \mathbf{b} adjusted by the prior specification of the bias. The posterior covariance is simply the OLS covariance plus a function of the prior covariance matrix for \mathbf{c} .

Note that the second term in Equation (12), $n^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{U}_0(\mathbf{X}'\mathbf{X})^{-1}$, is equal to $\mathbf{V}_{XX}^{-1}\mathbf{U}_0\mathbf{V}_{XX}$, where \mathbf{V}_{XX} is the covariance matrix of \mathbf{X} . Because the covariance matrix for \mathbf{X} does not depend on n , the Bayesian adjustment to the OLS standard errors is independent of sample size. The adjustment expresses uncertainty in our estimates because of model uncertainty. Thus, as sample size goes to infinity, the Bayesian standard errors will not go to zero.

Specifying a Prior

Choice of a prior is key to dealing with model misspecification of various types. A model could be misspecified for a variety of reasons: an omitted variable or variables, measurement error, selection, or simultaneity. Each of these problems implies a different type of prior. A full discussion of prior specification, however, is an article itself. This is a topic for future research. Here, we provide only a basic discussion.

In terms of our analysis of multicollinearity, our goal is to provide a way of measuring the sensitivity of estimates to modest changes in the model specification. For now, we are only interested in using a prior that allows us to move to moderately weaker assumptions than the extraordinarily stringent assumption made in the classical regression model, that $E[\mathbf{X}'\mathbf{e}] = \mathbf{0}$ exactly. Specifically, we want to examine situations in which we suspect that \mathbf{X} and \mathbf{e} are nearly uncorrelated, but we are not fully confident that this is exactly true.

The purpose in using the Bayesian methods is not to get precise results, but rather to get a rough sense of how much larger standard errors should be in order to reflect uncertainty about the model specification as well as sampling error. We begin by assuming that the prior mean $\mathbf{c}_0 = \mathbf{0}$ and that the covariance matrix, \mathbf{U}_0 , is diagonal. This simplifies the choice of prior, restricting attention to the variances in \mathbf{U}_0 . We further simplify the prior covariance matrix by specifying $\mathbf{U}_0 = u_0\mathbf{I}$, in which \mathbf{I} is the identity matrix. Choice of the prior covariance matrix thus reduces to choosing a single prior parameter, u_0 . We can elicit u_0 by specifying an interval that

represents the likely range of the covariances between \mathbf{X} and the error term. With $\mathbf{c}_0 = \mathbf{0}$, this interval will be centered at zero. The plausible interval for \mathbf{c} will cover four standard deviations of the normally distributed prior. Dividing the length of the interval by four yields the prior standard deviation of \mathbf{c}_0 . Squaring the prior standard deviation yields the prior variance, u_0 , that fills the diagonal elements of the prior covariance matrix, \mathbf{U}_0 .

How large is the covariance between a predictor and the error term? We limit our discussion here to the problem of omitted variables. It is easiest to think about the effects of omitted variables if our predictors and the errors have standard deviations of one. To work in this metric, each predictor, x_i , is scaled by its standard deviation, $s(x_i)$. To ensure the standard deviation of the errors is approximately one, first regress \mathbf{y} on the scaled \mathbf{X} and calculate the standard deviation of the residuals, $s(\hat{\epsilon})$. Scaling \mathbf{y} by $s(\hat{\epsilon})$ ensures that a regression fit to the scaled data will have OLS residuals with a standard deviation of one. After the Bayesian regression is fit, the coefficient standard errors can be transformed back to their original scales by multiplying by $s(\hat{\epsilon})/s(x_i)$.

Once the independent and dependent variables are suitably scaled, the covariance and correlation between \mathbf{X} and \mathbf{e} will be equal. In general, the range of possible correlations between the independent variables and the error should be small; we would suggest less than 0.10 in absolute value. If this is not the case, then one's model is so misspecified that further analysis of it is unlikely to be worthwhile.

Consider the case in which there is only minimal knowledge about the correlations between \mathbf{X} and \mathbf{e} . We suggest that correlations between \mathbf{X} and \mathbf{e} in the interval $[-0.10, 0.10]$ represent a high degree of model specification uncertainty, the interval $[-0.05, 0.05]$ expresses a moderate degree of uncertainty, and $[-0.02, 0.02]$ describes strong confidence in one's specification. Dividing these intervals by four and squaring them yields the prior standard variances, 0.0001, 0.000625, and 0.0025 that represent a reasonable range of values in \mathbf{U}_0 .

The prior specification for \mathbf{U}_0 implies that: (1) each predictor is equally affected by a model misspecification that is summarized by one parameter, u_0 , and (2) because \mathbf{U}_0 is diagonal, our uncertainty about the components of \mathbf{c} are uncorrelated. The prior could be elaborated by relaxing the restriction that the prior variances be equal. Of course, if one has strong intuitions about the nature of the model misspecification, then this should be incorporated into one's choice of both the prior mean, \mathbf{c}_0 , and the prior covariance matrix, \mathbf{U}_0 .

Choosing \mathbf{U}_0 to be diagonal represents a middle position between two extremes. Consider a regression with two predictors. Without loss of generality, assume that both have a positive effect on \mathbf{y} . One possibility is that the covariances of these two X s with the error term have the same sign: either both positive or both negative. Correcting for misspecification of this type principally leads to the absolute size of the variables' two regression slopes changing in the same direction, not their relative sizes. Formally, if $a\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{e} = n\mathbf{c}_0$ for a positive scalar a so the covariances between each \mathbf{X} and \mathbf{e} is positively proportional to the covariance between that \mathbf{X}

and \mathbf{y} , then:

$$\begin{aligned}
 \mathbf{b}_{GOLS} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y} - n\mathbf{c}_0) \\
 &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{e}) \\
 &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y} - a\mathbf{X}'\mathbf{y}) \\
 &= (1 - a)(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y}) \\
 &= (1 - a)\mathbf{b}_{OLS}
 \end{aligned}$$

Here, correcting for misspecification reduces the absolute size of the OLS estimates. Their size relative to each other remains constant. As before, multicollinearity magnifies the effects of model misspecification, but the size of coefficients is only changed absolutely. This situation would occur if, for example, the degree of measurement error in the X s as measured by their reliabilities is similar. To the extent that the covariances between the X s and the error term fail to meet the above proportionality condition, misspecification will change the relative size of coefficients.

The other possible pattern of misspecification is that the covariances of the two X s with the error term are likely to be of the opposite sign. This was illustrated in the three hypothetical examples described above. In this case, both the absolute and relative size of the two coefficients will change. The particular pattern of variation in the covariances will differ across situations. If an error term consists of random measurement error associated with a particular X , then that variable will have a negative covariance with the error, but the other X s will be uncorrelated with the error term. If the nonrandom component of the error term consists of omitted X s, the covariance of each observed X with the error term will depend on its covariance with these omitted variables. In this case, as the hypothetical examples above demonstrated, multicollinearity is likely to severely exacerbate the effects of model misspecification.

Specifying \mathbf{U}_0 to be diagonal represents a compromise between the two cases just described: in one model, uncertainty is likely to change the relative size of effects only modestly, and in the other model, uncertainty is likely to dramatically change the size of relative effects. Substantively, this prior implies that our uncertainty about the components of \mathbf{c}_0 are uncorrelated. In this case, our beliefs about a specification error that would bias one coefficient are unrelated to beliefs about specification errors affecting other coefficients.

A Monte Carlo Experiment

We can compare the frequentist properties of our Bayesian method to OLS with a Monte Carlo experiment. To perform the experiment, we generate the predictors and the dependent variable from a known distribution. We then fit models using OLS and Bayesian regression, varying four experimental conditions: (1) sample size n , (2) collinearity r , (3) the model misspecification C , and (4) the prior variance for the specification error u_0 . Our interest centers on the rate at which nominal confidence intervals cover the true regression coefficients. Ideally, a 90 percent

confidence interval for a coefficient should cover the true regression coefficient 90 percent of the time in repeated sampling. Of course, OLS confidence intervals have this property for correctly specified models. If the true coverage rate of a confidence interval is lower than its nominal level, the interval provides a falsely precise inference. How does OLS perform relative to the Bayesian regression for misspecified models?

This question is studied with the following experimental design. A sample of size n ($n = 100, 1,000$) for each of three variables— x_1 , x_2 , and x_3 —is generated from a trivariate normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\Sigma = \begin{bmatrix} 1 & r & c \\ & 1 & 0 \\ & & 10 \end{bmatrix}, \quad r = 0, 0.60, \quad \text{and} \quad c = 0, 0.15, 0.30$$

We then formed a dependent variable, $y = x_1 + x_2 + x_3$, and fit a misspecified model including only the predictors x_1 and x_2 . The true regression coefficients for x_1 and x_2 are thus known to equal one. The model is misspecified in the sense that there is an omitted variable, x_3 , that is correlated with one of the included predictors. In fitting the misspecified model, r controlled the collinearity among the included predictors, and c controlled the degree of model misspecification. We can express the misspecification as a correlation, rather than a covariance, by dividing by the variance of the omitted variable. The correlation between the omitted variable and x_1 is $C = c/10$. The Bayesian regression uses a prior that assumes that the correlations, C , between predictors and the error is in the interval $[-0.05, 0.05]$ ($u_0 = 0.000625$) or in the interval $[-0.10, 0.10]$ ($u_0 = 0.0025$). When the misspecification correlation $C = 0.03$ and the prior $u_0 = 0.000625$, model misspecification is towards the edge of prior distribution. For $u_0 = 0.0025$, however, the largest specification errors are well within the range entertained by the prior. For each simulation, OLS and Bayesian 90 percent confidence intervals were calculated. For each design point (r, c, n, u_0) in the experiment, we performed 500 simulations. The coverage rate for a coefficient equals the percentage of confidence intervals out of 500 simulations that cover the true coefficient value of one. (Computer code used to perform these and other Bayesian calculations is reported in Appendix B of the online supplement.)

Table 2 compares the OLS and Bayesian estimators for the coefficient of x_1 when the prior, $u_0 = 0.00625$. As we would hope, an OLS 90 percent interval yields the correct coverage rate when the model is correctly specified; that is, when $C = 0$. As the specification error increases from $C = 0$ to $C = 0.03$, the performance of OLS deteriorates substantially. The final column of Table 2 shows that bias in OLS ranges from 15 percent to 50 percent of the true coefficient even with relatively little specification error. Bias also impairs the performance of inference with OLS. In small samples, when omitted variables correlate with observed predictors at 0.03, a nominally significant interval of 90 percent covers the true parameter between 56 percent and 78 percent of the time. With small samples and assuming only moderate model misspecification, the Bayesian regression offers little protection from specification errors. With $n = 100$, a prior of $u_0 = 0.000625$ increases the OLS standard errors by less than 10 percent. Because the Bayesian adjustment for model

Table 2: Monte Carlo Results Comparing Coverage Rates of 90% Intervals and Standard Errors (S.E.) for OLS and Bayes Estimators, with Prior $u_0 = 0.000625$.

Specification	Collinearity	OLS Coverage (%)	Bayes Coverage (%)	OLS Average S.E.	Bayes Average S.E.	Average Coefficient
Error (C)	(r)					
<i>Sample size, n = 100</i>						
0.000	0.0	92	92	0.32	0.33	1.00
0.000	0.6	91	92	0.40	0.43	1.02
0.015	0.0	88	89	0.32	0.33	1.15
0.015	0.6	83	86	0.40	0.43	1.26
0.030	0.0	75	77	0.32	0.33	1.29
0.030	0.6	66	70	0.40	0.42	1.49
<i>Sample size, n = 1,000</i>						
0.000	0.0	91	97	0.10	0.13	1.00
0.000	0.6	89	99	0.13	0.19	1.00
0.015	0.0	57	74	0.10	0.13	1.15
0.015	0.6	43	76	0.13	0.19	1.23
0.030	0.0	08	17	0.10	0.13	1.30
0.030	0.6	02	10	0.12	0.19	1.46

Note: A prior of $u_0 = 0.000625$ assumes that correlations between predictors and the error fall in the interval $[-0.05, 0.05]$.

misspecification is independent of sample size and coefficient standard errors tend to be large in small samples, the adjustment contributes only a small fraction to the Bayesian standard error. Consequently, there is no significant efficiency loss with the Bayes estimator but little gain in inferential accuracy for misspecified models.

The advantage of Bayesian inference is clearer in large samples. With a sample size of 1,000 and a correctly specified model, the Bayesian interval is substantially too conservative (too wide) only at a high level of collinearity. With moderate misspecification ($C = 0.015$), the prior of $u_0 = 0.000625$ increases the OLS standard errors between 30 percent and 50 percent, depending on the level of collinearity. With this model misspecification, OLS performs poorly. With uncorrelated predictors, the OLS confidence interval covers the true parameter just over half the time. When the predictors are moderately collinear ($r = 0.60$), the nominal OLS 90 percent interval includes the true coefficient less than half the time. The Bayes estimator offers a large gain in inferential accuracy in this case, with coverage rates over 70 percent. When the model misspecification is severe, $C = 0.03$, the OLS interval rarely covers the true interval, even with uncorrelated predictors. Although prior information about specification errors inflates the OLS standard errors substantially, confidence intervals with a prior of $u_0 = 0.000625$ are also extremely optimistic.

Table 3 reports results for the prior $u_0 = 0.0025$, which assumes that correlations between observed predictors and the errors are between -0.10 and 0.10 . This prior produces similar results to those obtained with the more modest assumption that

Table 3: Monte Carlo Results Comparing Coverage Rates of 90% Intervals and Standard Errors (S.E.) for OLS and Bayes Estimators, with Prior $u_0 = 0.0025$.

Specification		OLS	Bayes	OLS	Bayes	
Error	Collinearity	Coverage	Coverage	Average	Average	Average
(C)	(r)	(%)	(%)	S.E.	S.E.	Coefficient
<i>Sample size, n = 100</i>						
0.00	0.0	89	92	0.32	0.36	0.99
0.00	0.6	91	96	0.40	0.50	1.00
0.15	0.0	87	93	0.32	0.36	1.14
0.15	0.6	81	89	0.40	0.50	1.26
0.30	0.0	75	81	0.32	0.36	1.29
0.30	0.6	70	80	0.40	0.49	1.47
<i>Sample size n = 1,000</i>						
0.00	0.0	90	100	0.10	0.19	1.00
0.00	0.6	91	100	0.13	0.31	1.00
0.15	0.0	53	94	0.10	0.19	1.16
0.15	0.6	41	98	0.13	0.31	1.23
0.30	0.0	09	54	0.10	0.19	1.30
0.30	0.6	02	66	0.12	0.31	1.46

Note: A prior of $u_0 = 0.0025$ assumes that correlations between predictors and the error fall in the interval $[-0.10, 0.10]$.

$u = 0.000625$. In the analysis of small samples, the Bayesian adjustment is slightly larger than under the more conservative prior. Good results are obtained for modest specification error ($C = 0.015$), with Bayesian coverage rates in the vicinity of their nominal level of 90 percent. In large samples, the Bayesian regression performs very well at this level of specification error. With $C = 0.015$ in large samples, OLS intervals are overly optimistic (too narrow) by a very large margin. The Bayesian interval, on the other hand, provides coverage rates close to their nominal levels. In highly misspecified models, $C = 0.03$, the Bayesian intervals yield a significant improvement over OLS, although the intervals are still very optimistic.

The Monte Carlo experiment suggests that regression analysis with a prior allowing for specification error is a useful diagnostic tool in nearly all circumstances. In small samples, the Bayesian analysis does not perform substantially worse than OLS and can provide small improvements in the presence of specification errors. In large samples, the Bayesian analysis can perform significantly better, especially with moderate levels of specification error and highly collinear predictors.

An Empirical Example

To illustrate the Bayesian regression in a more realistic example, we consider the race gap in male earnings controlling for measures of skill and schooling. This analysis arises in the context of the exchange between Cancio, Evans, and Maume

Table 4: Correlations among Variables for Regression Analysis of Log Hourly Earnings, NLSY Men, 1998 ($N = 3,526$).

	Log Wages	Schooling	AFQT	Married	Age	Black	Hispanic
Log Wages	1.00	0.40	0.44	0.24	0.06	-0.21	-0.05
Schooling		1.00	0.63	0.15	0.01	-0.09	-0.12
AFQT			1.00	0.20	0.14	-0.38	-0.14
Married				1.00	0.06	-0.22	0.01
Age					1.00	0.00	0.00
Black						1.00	-0.31
Hispanic							1.00

(1996) and Farkas and Vicknair (1996). Briefly, Cancio et al. (1996) argue that a large racial gap in wages remains even after controlling for social background and market characteristics of workers, and indeed, the wage gap has grown since 1976. Farkas and Vicknair (1996) reply that skill differences can account for the black–white wage gap among men. We take no particular position on the substantive issues, but the debate does usefully illustrate reliance on statistical inference about coefficients in the presence of collinearity. Like Farkas and Vicknair (1996), we analyze data from the National Longitudinal Survey of Youth (NLSY). Our sample is restricted to males with positive hourly earnings in 1998. The dependent variable is log hourly earnings, and predictors include measures of skill, years of schooling, a dummy variable for married respondents, age, and dummy variables for black and Hispanic respondents. Skill is measured by the respondent’s percentile score for the Armed Forces Qualification Test (AFQT).

Table 4 reports correlations among the the dependent and independent variables. Schooling and AFQT are highly correlated, and black is moderately correlated with AFQT and married. Regressing AFQT on all the other independent variables yields an R^2 of 0.56, indicating moderate collinearity among the predictors.

Interest centers on the residual black–white gap in wages expressed by the coefficient for black. The OLS results indicate that black men in the NLSY earned 10 percent less than white men after controlling for schooling, AFQT, marital status, and age (Table 5). The black coefficient is statistically significant. If the model is only only slightly misspecified so that correlations of omitted variables with the predictors are of the order ± 0.01 , our uncertainty about the black–white wage gap increases, but we can remain quite confident about the earnings difference. If specification errors are larger than this, the Bayesian analysis indicates that we cannot draw a confident conclusion about the earnings gap. Indeed, this result usefully quantifies the state of the debate—*inference about the black–white wage gap is sensitive to the choice of model specification.*

Table 5: Bayesian and OLS Regression Results from Analysis of Log Hourly Earnings, NLSY Men, 1998.

	Coefficient (1)	OLS (2)	Absolute <i>t</i> statistics		
			$u_0 =$ 0.0001 (3)	$u_0 =$ 0.000625 (4)	$u_0 =$ 0.0025 (5)
Schooling	0.06	10.60	7.54	3.98	2.10
AFQT	0.58	11.68	7.89	4.02	2.10
Married	0.20	9.71	8.21	5.19	2.93
Age/10	0.03	0.76	0.64	0.40	0.23
Black	-0.10	3.62	2.70	1.48	0.79
Hispanic	-0.02	0.86	0.69	0.40	0.22

Note: Results for columns 3 to 5 are from the Bayesian regression analysis. $R^2 = 0.25$, $N = 3,526$.

Conclusion

Often when empirical researchers carry out a regression analysis, they focus on the point estimates they obtain. This gives the impression that the data provides one answer about the magnitude of an effect. Of course, data are typically consistent with a range of effects, and the point estimate is only the most likely estimate for a given model. We are much better off to think of the confidence interval around an estimate as representing the values of the effect that are consistent with the data.

A confidence interval, however, is only useful if it accurately reflects the uncertainty about an estimate after the analysis has been carried out. In the first part of this article, we showed that parameter estimates are likely to be highly sensitive to model specification when multicollinearity is present. The following section of the article presented a Bayesian approach to incorporating uncertainty about the model specification into one's standard errors and thus into one's confidence intervals. Monte Carlo results indicated there are few disadvantages and often significant improvements in inference with Bayesian methods. We then demonstrated how this methodology could be used in two empirical situations in which multicollinearity is a potential problem.

The approach here could be applied in other situations. An obvious application is to the problem of "weak" instruments. The weak instruments literature uses instrumental variables that are weakly correlated with the dependent and independent variable of concern, but obtain precise estimates by using extremely large samples. The most discussed example is the use of quarter of birth as an instrument for education in estimating the effect of education on earnings (Angrist and Krueger 1991, 1992). As John Bound and his coauthors (Bound et al 1995; Bound 1996) have argued, these analyses are quite sensitive to model specification error. More generally, the approach developed here could be used to analyze the sensitivity of identification restrictions in simultaneous equation models or in latent class analysis.

Our hope is that this article will encourage others to think more generally about how to incorporate model specification uncertainty into one's analysis. If

anything is true of social science analyses, it is that we typically are uncertain about what the "right" model is. Even if we are individually confident about a particular specification, there are almost certainly others who will argue that we have gotten it wrong. Paradoxically, the frequentist methods that we all commonly use assume that we know the correct model. A Bayesian approach potentially offers a constructive alternative. This article is a first attempt to provide what will hopefully be a useful approach to incorporating uncertainty and disagreement about model specification into one's analysis.

Notes

- 1 We have examined numerous statistics and econometric textbooks as well as class notes posted on the web. In almost all cases, the discussion of multicollinearity has focused solely on its consequence for standard errors, as opposed to the additional problem of bias considered here. Farrar and Glauber (1967) and Mela and Kopalle (2002) are exceptions. We thank the editor for pointing out these two citations.
- 2 A closely related issue that has received considerable attention in the econometric literature is that of weak instruments. The problem with weak instruments is that the $Z'X$ matrix in the IV formula $b_{IV} = (Z'X)^{-1}Z'Y$ may be close to singular, possibly resulting in substantial bias if the exclusion restriction does not hold exactly (Bound et al. 1996). This is directly analogous to the problem we discuss at length in this article, in which multicollinearity results in the $X'X$ being nearly singular, with the consequence that OLS estimates can be severely biased if the assumption that the X 's and error are uncorrelated is not met exactly.
- 3 For an recent overview of that literature, see Neelman 2012. Much of the literature on multicollinearity has been published prior to 2000. Important exceptions are several articles that have analyzed the consequences of multicollinearity in multilevel models (Yu 2015; Shieh and Foulai 2003; Graham 2003).
- 4 The astute reader will wonder why the problem of model misspecification in the presence of multicollinearity cannot simply be solved in the case where one has a very large sample by choosing a subsample in which the X s were uncorrelated. Unfortunately, this increases the absolute size of the covariance between the X s and the error, exacerbating the misspecification problem. This is most easily seen in the case of measurement error. Consider the case of two X s that are positively correlated. We would want to choose cases where individuals were high on one value of X , but low on the other. But these are precisely cases that are likely to have considerable measurement error associated with them.
- 5 One could also worry about whether the value of $X'e$ in the sample is an unbiased estimate of the population value of $X'e$. If X is assumed fixed, then the linearity of this expression guarantees unbiasedness.
- 6 More formally, the effect of multicollinearity can be understood using the adjoint matrix method for calculating a matrix's inverse. The key observation is that the elements of a matrix's inverse are equal to the corresponding element in the matrix's adjoint matrix divided by its determinant. A matrix's determinant will be zero when there is strict linear dependence and near zero when there is strong multicollinearity. The determinant for a k by k matrix is its volume in k dimensional space. The adjoint of a matrix is simply calculated as the transpose of a matrix, replacing each element by its

cofactor. The element M_{ij} 's cofactor is the determinant of the submatrix of \mathbf{M} defined by deleting the i th row and j th column (Cullen, 1972, section 3.3.)

- 7 As discussed in Mela and Kopalle (2002), it is possible in some cases, depending on the exact pattern of correlation, for multicollinearity to lead to more and not less precise estimates.
- 8 A Bayesian approach is commonly recommended as a means of dealing with multicollinearity. The thinking is analogous to the recommendation for dealing with multicollinearity by increasing the sample size. If the analysis is based on more information, then we should be able to estimate the parameters more precisely (Leamer 1994; Birkes and Dodge 1993; Judge et al. 1985). The ridge estimator, another way of dealing with multicollinearity, can be justified in part as a special example of a Bayesian estimator (Birkes and Dodge 1993; Judge et al. 1985) in which the regression coefficients are being shrunk to a prior with zero mean and a diagonal covariance matrix.
- 9 In spirit, the approach here is similar to that of Manski's (1995) work on bounds. Manski's work involves determining the bounded set of estimates that are consistent with a minimal set of assumptions or restrictions. These restrictions are deterministic. In essence, the approach here involves examining the implications of assumptions that are probabilistic.
- 10 A probability distribution (or density) is "proper" if it integrates to one.
- 11 The thinking here is similar to Raftery's (1995) for pooling across models. In Raftery's approach, a pooled parameter estimate is obtained by using a weighted combination of estimates across different models in which the weights are proportional to the posterior probability of each model. Raftery focuses on applications with a finite number of alternative models. Our approach involves an infinite number of alternative models implied by the continuous and unbounded prior on \mathbf{c} . In our approach, the likelihood of different models is determined by the prior distribution for \mathbf{c} as opposed to the posterior model probabilities.

References

- Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics* 106: 979–1014. <http://dx.doi.org/10.2307/2937954>
- Angrist, Joshua D., and Alan B. Krueger. 1992. "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association* 87: 328–336. <http://dx.doi.org/10.1080/01621459.1992.10475212>
- Belsley, David A. 1991. *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. New York: John Wiley & Sons.
- Birkes, David and Yadolah Dodge. 1993. *Alternative Methods of Regression*. New York: John Wiley & Sons. <http://dx.doi.org/10.1002/9781118150238>
- Bound, John. 1996. "On the Validity of Season of Birth as an Instrument in Wage Equations: a Comment on Angrist and Krueger's Does Compulsory School Attendance Affect Schooling and Earnings?" NBER Working Paper 5835.
- Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation when the Correlation Between Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association* 90: 443–450. <http://dx.doi.org/10.1080/01621459.1995.10476536>

- Cancio, A. Silvia, T. David Evans, and David J. Maume. 1996. "Reconsidering the Declining Significance of Race: Racial Differences in Early Career Wages." *American Sociological Review* 61: 541–57. <http://dx.doi.org/10.2307/2096391>
- Cullen, Charles G. 1972. *Matrices and Linear Transformations*. New York, Dover Publications.
- Draper, David. 1995. "Assessment and Propagation of Model Uncertainty." *Journal of the Royal Statistical Society, Series B* 57: 45–97.
- Dreze, J. 1975. "Bayesian Theory of Identification in Simultaneous Equation Models," in *Studies in Bayesian Econometrics and Statistics*, eds. S. Feinberg and A. Zellner, Amsterdam: North Holland: 159–174.
- Farkas, George and Keven Vicknair. 1996. "Appropriate Tests of Racial Wage Discrimination Require Controls for Cognitive Skill: Comment on Cancio, Evans, and Maume." *American Sociological Review* 61: 557–560. <http://dx.doi.org/10.2307/2096392>
- Farrar, Donald E. and Robert R. Glauber. 1967. "Multicollinearity in Regression Analysis: The Problem Revisited." *The Review of Economics and Statistics*. 49(1): 92-107. <http://dx.doi.org/10.2307/1937887>
- Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis*. London: Chapman & Hall.
- Goldberger, Arthur S. 1991. *A Course in Econometrics*. Cambridge: Harvard University Press.
- Gordon, Robert. 1968. "Issues in Multiple Regression." *American Journal of Sociology* 73: 592–616. <http://dx.doi.org/10.1086/224533>
- Graham, M.H., 2003. "Confronting multicollinearity in ecological multiple regression." *Ecology*, 84: 2809–2815. <http://dx.doi.org/10.1890/02-3114>
- Harville, David. 1997. *Matrix Algebra from a Statistician's Perspective*. New York: Springer. <http://dx.doi.org/10.1007/b98818>
- Judge, George, W.E. Griffiths, R. Carter Hill, Helmut Lutkepohl. 1985. *The Theory and Practice of Econometrics*. Second Edition. New York: John Wiley & Sons.
- Kadane, J. B. 1975. "The Role of Identification in Bayesian Theory." in *Studies in Bayesian Econometrics and Statistics*, eds. S. Feinberg and A. Zellner, Amsterdam: North Holland
- Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.
- Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73: 31–43.
- Leamer, Edward E. 1991. "A Bayesian Perspective on Inferences from Macroeconomic Data." *Scandinavian Journal of Econometrics* 93: 225–48. <http://dx.doi.org/10.2307/3440330>
- Leamer, Edward E. 1994. *Sturdy Econometrics*. Hants, England: Edward Elgar.
- Lindley, D. V. and G. M. El-Sayyad. 1968. "The Bayesian Estimation of a Linear Functional Relationship." *Journal of the Royal Statistical Society, Series B* 30: 190–202.
- Kmenta, Jan. 1986. *Elements of Econometrics*. Second Edition. Ann Arbor: The University of Michigan Press.
- Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.
- Mela, Carl F. and Praveen K. Kopalle. 2002. "The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations." *Applied Economics* 34: 667-677. <http://dx.doi.org/10.1080/00036840110058482>

- Neath, Andrew A. and Fancisco J. Samaniego. 1996. "On Bayesian Estimation of the Multiple Decrement Function in the Competing Risks Problem." *Statistics and Probability Letters*. [http://dx.doi.org/10.1016/s0167-7152\(96\)00016-8](http://dx.doi.org/10.1016/s0167-7152(96)00016-8)
- Neath, Andrew A. and Fancisco J. Samaniego. 1997. "On the Efficacy of Bayesian Inference for Nonidentifiable Models." *The American Statistician* 51: 225–232.
- Neeleman, Dirk. 2012. *Multicollinearity in linear economic models* (Vol. 7). Springer Science and Business Media.
- Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." In *Sociological Methodology*, edited by Peter V. Marsden, 111-63. Cambridge, MA.: Blackwell Publishers. <http://dx.doi.org/10.2307/271063>
- Shieh, Yann-Yann and Rachel T. Fouladi. 2003. "The effect of multicollinearity on multi-level modeling parameter estimates and standard errors." *Educational and psychological measurement*, 63: 951-985. <http://dx.doi.org/10.1177/0013164403258402>
- Western, Bruce. 1996. "Vague Theory and Model Uncertainty in Macrosociology." In *Sociological Methodology*, edited by Adrian E. Raftery, 165-193. Cambridge, MA.: Blackwell Publishers.
- Western, Bruce. 1999. "Priors for Regression with Endogeneity and Selection Bias." Unpublished. Princeton University.
- White, Halbert. 1994. *Estimation, Inference and Specification Analysis*. New York: Cambridge.
- Yu, Han, Shanhe Jiang, and Kenneth C. Land. 2015. "Multicollinearity in Hierarchical Linear Models." *Social Science Research* 53: 118-136. <http://dx.doi.org/10.1016/j.ssresearch.2015.04.008>
- Zellner, Arnold. 1971. *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons.

Acknowledgements: We thank Kinga Makovi for help in the preparation of the manuscript. We also appreciate the editor's suggestions for citations that we were unaware of.

Christopher Winship: Department of Sociology, Harvard University.
E-mail: cwinship@wjh.harvard.edu.

Bruce Western: Department of Sociology, Harvard University.
E-mail: western@wjh.harvard.edu.