

ROUGH AND PARTIAL DRAFT

The Perils of Multicollinearity: A Reassessment

Christopher Winship

Harvard University

September 1999

This research was partially supported by NSF grant SBR-94-11875. I want to thank Ken Land for encouraging me to work on this problem and for helpful comments. Participants in the American Sociological Associations' 1998 Winter Methodology Conference and participants in an Boston American Statistical Associations' seminar provided many helpful comments.

Abstract

Multicollinearity is typically thought of as a problem of large standard errors resulting from the near linear dependence of independent variables. One solution is to have more informative data, possibly in the form of a larger sample.

In this paper I argue that this understanding of multicollinearity is only partially correct. The near collinearity of independent variables results in regression estimates being potentially quite sensitive to small degrees of model misspecification. I examine the assumption that the independent variables and error are uncorrelated and show that when there is multicollinearity, small deviations from this assumption can lead large changes in estimates.

I then present a new estimator based on Bayesian methods that extends the classical OLS estimator by specifying a prior on the correlation between X and e . I show how this estimator can be used to calculate confidence intervals that reflect both sampling error and uncertainty with respect to the model specification.

Introduction

In a classic, but now seldom studied article published in The American Journal of Sociology entitled “Issues in Multiple Regression,” Robert Gordon (1964) analyzed the problems involved in model specification in regression analysis when variables are nearly collinear. He presents a particularly troubling example where two independent variables, each standardized to have variance one (along with other variables), are correlated .8 with each other. One variable has a .6 correlation with the dependent variable and the other variable having a .55 correlation. They have identical correlations with the other variables in the model. He reports that the ordinary least squares estimates for the regression slopes for these two variables are, respectively, .38 and .13. Not too surprisingly, he is disturbed that such a small difference in the two independent variables correlation with Y, .60 versus .55, should result in such a considerable difference in their estimated effect on Y, .38 versus .13.

How are we to understand this example? Gordon considers two possibilities. First, he notes that the X variables’ true correlations with Y might be equal and that their observed difference might simply due to sampling error. In this case, he points out the effects of sampling error should be reflected in the standard errors for the two coefficients. In principle they should be sufficiently large to indicate that the two regression coefficients might be equal. Alternatively, if we were to carry out a formal test of their equivalence, the hypothesis that the two coefficients are equal should not be rejected.

Second, however, he notes that even if .60 and .55 were the true correlations of the two independent variables with the dependent variable, changes in the inclusion of other variables in

the model could substantially effect the regression estimates. That is, the results could be quite sensitive to the particular model specification. This is reason for great concern given the frequent uncertainty that social scientists have about what model specification is correct.

This paper formally explores the issue of model misspecification in the linear regression model in the presence of multicollinearity. Our interest is in situations where the research focus is on the relative and/or absolute size of specific regression coefficients. This is typically the situation when 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 education's effect on earnings.

Traditionally, multicollinearity has been thought about simply as a problem of having weak data or too little data. For example, Kmenta (1971, p. 391; 1997, p. 442) states "that a high degree of multicollinearity is simply a feature of the sample that contributes to the unreliability of the estimated coefficients, but has no relevance for the conclusions drawn as a result of this unreliability." (emphases added). Goldberger (1991) describes at length how the problem of multicollinearity is directly analogous to the problem of having a small sample and has the same consequences. Specifically, the problem of multicollinearity is that of having weak data with the result being that one's standard errors are so large that it is impossible to precisely estimate regression coefficients of interest. Many other citations could be added. Thus the problem induced by multicollinearity is equivalent to that when one has too little data --- the data provides only very imprecise estimates of the regression parameters. From this traditional perspective, if one can collect enough data then it should be possible to solve the multicollinearity problem (Kmenta 1997).

When independent variables are nearly multicollinear and one's sample size is less than enormous, estimates of some or all coefficients will be imprecise and it is quite possible that some estimates will 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 unreasonable estimates are included. From a Bayesian perspective the issue is that the researcher has failed to include sufficient prior information into the analyses about what constitutes realistic estimates for the coefficients (Leamer 1994).

The traditional understanding of multicollinearity as simply weak or insufficiently informative data ignores Gordon's second observation -- that in the presence of multicollinearity, results may be highly dependent on model specification, even with very large samples. This sensitivity is not captured by the traditional standard errors of an estimate and is independent of how large or informative the sample is.

The goals of this paper are two fold. At a general level I present in a nontechnical manner an approach to estimating the precision of estimates that incorporates model specification uncertainty as well as sampling error. This is important. The standard errors associated with conventional methods only reflect potential error due to sampling variability. Given, however, that model specification, (e.g., the question of which variables should be included in a model), is often an open question, it is important to include uncertainty about the specification. In the current paper I show how this can be done for the linear regression model. This approach, however, is applicable to a wide variety of situations including instrumental variables, simultaneous equations, analysis of covariance models, and latent class models. A companion paper will provide a more technical presentation and explore these additional applications.

My approach relies on Bayesian methods, but uses them in atypical way. In a traditional Bayesian approach, the goal is to add additional information into the analysis by specifying a prior distribution for the parameters. In this paper, however, Bayesian methods are used to relax the strict constraints on parameters and/or moments in the classical/frequentist approach. This allows me to incorporate uncertainty about model specification into the analysis. In particular, the resulting standard errors for the parameters estimates reflect both sampling error and model specification uncertainty.

The second goal of the paper is to use my approach to reanalyze the problem of multicollinearity in the single equation multiple regression model. I am able to show that standard thinking that multicollinearity simply amounts to having too little data or data that is weak is only partially correct. I examine the classical regression model's constraint that the correlation between \mathbf{X} and error term, \mathbf{e} , is zero. This is the critical assumption in regression. It implies that there is no measurement error in the independent variables, that there are no omitted independent variables that are correlated with the independent variables included in the model, and that there are no selection or simultaneity problems. In almost all applications, these conditions only hold approximately.

Below I show how Bayesian methods can be used to relax this assumption. The specific methods I develop allow the researcher to incorporate model specification error, as represented by the possible correlation between \mathbf{X} and \mathbf{e} , into the classical regression model. I then show how Bayesian methods can be used to create standard errors and confidence intervals that incorporate the effects of model misspecification. When there is multicollinearity, this can have a large effect on the size of standard errors and associated confidence intervals.

The next section of the paper presents a basic analysis of model misspecification in the classical regression model. The following section demonstrates the sensitivity of estimates within the classical regression model to model misspecification when there is multicollinearity. The subsequent section develops a Bayesian approach to the problem. I then consider the use of these methods in practice. Following this I present an empirical example. I conclude by arguing that it is important that we have methods that incorporate specification uncertainty as well as sampling error into our analysis. This essential if we are to have confidence intervals that accurately reflect the range of estimates that are consistent with the data.

The “Classical” Linear Regression Model

In this paper I will deal simply with the standard linear regression model:

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

where \mathbf{Y} is a \mathbf{n} by 1 column vector consisting of the respondent’s values on the dependent variable, \mathbf{X} is a \mathbf{n} by \mathbf{k} matrix of the \mathbf{k} independent variables, \mathbf{b} is a 1 by \mathbf{k} row vector of “true” regression parameters to be estimated, and \mathbf{e} is a \mathbf{n} by 1 column vector of unobserved errors that are assumed to be independent, and identically distributed normal variables. Without loss of generality, I will assume that both \mathbf{Y} and \mathbf{X} have been centered to have mean zero. Similarly I will assume that the data are the result of a simple random sampling procedure.

The interest here is in estimating the regression coefficients, \mathbf{b} , in equation (1).

Multiplying each side of equation (1) by \mathbf{X}' and rearranging terms gives us:

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

Obviously $\mathbf{X}'\mathbf{e}$ can not be estimated in the sample since \mathbf{e} is unobserved. Note that equation (2) consists of \mathbf{k} equations with $2\mathbf{k}$ unknowns -- the \mathbf{k} regression parameters, \mathbf{b} , and the \mathbf{k} cross-products of the error term with each of the \mathbf{X} 's, $\mathbf{X}'\mathbf{e}$ (which are equal to \mathbf{n} , the sample size, times the covariance of \mathbf{e} with each \mathbf{X}). Rearranging terms to solve for \mathbf{b} we get:

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

\mathbf{b} can not be solved for explicitly in equation (3) without further assumptions. Ordinary least squares (OLS) assumes that the covariances of \mathbf{e} with each of the \mathbf{k} \mathbf{X} 's equals zero, i.e. $\mathbf{X}'\mathbf{e} = \mathbf{0}$.

Replacing $\mathbf{X}'\mathbf{e}$ by 0 reduces equation (3) to:

$$(4) \quad \hat{\mathbf{b}}_{\text{ols}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

where $\hat{\mathbf{b}}_{\text{ols}}$ is the standard OLS estimator of \mathbf{b} . What we have done here is solved for \mathbf{b} by

assuming that $\mathbf{X}'\mathbf{e} = \mathbf{0}$. To see how the OLS estimate of \mathbf{b} critically depends on this

assumption, substitute for \mathbf{Y} in equation 4 the right hand side of equation (1). This gives us:

$$(5) \quad \hat{\mathbf{b}}_{\text{ols}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\mathbf{X}\mathbf{b} + \mathbf{e}) = \mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{e}$$

This analysis is traditional and can be found in many textbooks. It shows that a sufficient condition for the OLS estimator to a “good” estimate of \mathbf{b} is that $\mathbf{X}'\mathbf{e}$ approximately equal zero, that is, \mathbf{X} and \mathbf{e} be approximately uncorrelated. If this is not the case, then $\hat{\mathbf{b}}_{\text{ols}}$ will be a poor estimate of \mathbf{b} .

\mathbf{X} and \mathbf{e} might be correlated for two different reasons. First, it might be the case that \mathbf{X} and \mathbf{e} are uncorrelated in the population, but because we had a relatively small sample, the correlation between \mathbf{X} and \mathbf{e} in the sample 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, $\hat{\mathbf{b}}_{\text{ols}}$ will be a highly biased and thus poor estimate of \mathbf{b} .

A Generalized OLS Estimator

The traditional OLS estimator can easily be extended to deal with the situation where the correlation between \mathbf{X} and \mathbf{e} is some other fixed vector of values besides zero. Assume that $\mathbf{X}'\mathbf{e}$ equaled some value other than zero, call it \mathbf{nc} , where \mathbf{n} is the sample size and \mathbf{c} is a \mathbf{k} by 1 vector.

Here \mathbf{c} here represents our assumption or “guess” about what the covariance between \mathbf{X} and \mathbf{e} is, that is what $\mathbf{X}'\mathbf{e}/\mathbf{n}$, equals. Then substituting for $\mathbf{X}'\mathbf{e}$ in equation (3), the analogue to equation (4) would be:

$$(6) \quad \hat{\mathbf{b}}_{\text{gols}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y} - \mathbf{nc})$$

I refer to this as a generalized ordinary least squares estimator. Note that it is equivalent to the standard OLS estimator, except that we correct $\mathbf{X}'\mathbf{Y}$, which is equal to \mathbf{n} times the covariance of \mathbf{X} with \mathbf{Y} , by subtracting out the “correction” factor \mathbf{nc} , where \mathbf{c} is the assumed covariance between \mathbf{X} with \mathbf{e} . Equivalently we could simply subtract \mathbf{nc} from $\mathbf{X}'\mathbf{Y}/\mathbf{n}$ and then use the standard OLS formula to obtain an estimate of \mathbf{b} . Obviously, equation (6) reduces to equation (4) if $\mathbf{c} = \mathbf{0}$.

To understand the properties of (6) as an estimator, similar to the derivation of equation (5), substitute for \mathbf{Y} in equation (6) using the right-hand side equation (1) and simplify:

$$(7) \quad \hat{\mathbf{b}}_{\text{gols}} = \mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{e} - \mathbf{nc})$$

What equation (6) demonstrates is that $\hat{\mathbf{b}}_{\text{gols}}$ will be a “good” estimate of the true regression parameter, \mathbf{b} , only if $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{e} - \mathbf{nc})$ is approximately equal to zero. A sufficient condition for

this to be true is that $(\mathbf{X}'\mathbf{e} - \mathbf{nc})$ be approximately equal to zero in the sample.

We are interested in why $\mathbf{X}'\mathbf{e}$ and \mathbf{nc} might differ from each other. $\mathbf{X}'\mathbf{e}$ and \mathbf{nc} might differ from each other for two reasons. To see this, let \mathbf{c}^* equal the true covariance between \mathbf{X} and \mathbf{e} in the population (that is, the probability limit of $\mathbf{X}'\mathbf{e}/\mathbf{n}$ as \mathbf{n} goes to infinity) as opposed to \mathbf{c} which is the assumed value of the covariance between \mathbf{X} and \mathbf{e} , that is our “guess”. Rewrite equation (7) as follows:

$$(8) \quad \hat{\mathbf{b}}_{\text{OLS}} = \mathbf{b} + \mathbf{n} (\mathbf{X}'\mathbf{X})^{-1} [(\mathbf{X}'\mathbf{e}/\mathbf{n} - \mathbf{c}^*) + (\mathbf{c}^* - \mathbf{c})]$$

S **M**

There are two reasons why $[(\mathbf{X}'\mathbf{e}/\mathbf{n} - \mathbf{c}^*) + (\mathbf{c}^* - \mathbf{c})]$ might not approximate zero in the sample. First the question is whether with a large enough sample does the **S** component in equation (7), $(\mathbf{X}'\mathbf{e}/\mathbf{n} - \mathbf{c}^*) \approx \mathbf{0}$ or equivalently $(\mathbf{X}'\mathbf{e}/\mathbf{n} \approx \mathbf{c}^*)$. This is the error due to sampling. Under suitable regularity conditions, the law of large numbers guarantees that the error will be small in the sense of consistency, that is the law of large numbers states that as our sample grows the probability that the sample value $\mathbf{X}'\mathbf{e}$ will differ from its true value by an arbitrarily small amount goes to zero.¹ Any differences between $\mathbf{X}'\mathbf{e}/\mathbf{n}$ from \mathbf{c}^* will be due to sampling error and it is these differences that induces sampling variance in $\hat{\mathbf{b}}_{\text{OLS}}$.

Second is the question of whether the **M** component in equation (8), $(\mathbf{c}^* - \mathbf{c}) \approx \mathbf{0}$ or equivalently whether $\mathbf{c}^* \approx \mathbf{c}$. This is a problem of whether the model is correctly specified. That

is, have we made the correct assumption (“guess”) about the probability limit for $\mathbf{X}'\mathbf{e}/n$, the covariance between \mathbf{X} and \mathbf{e} . When our assumption is wrong, we have misspecification error in the sense that we have made an incorrect assumption about the true value of the covariance of \mathbf{X} and \mathbf{e} .

OLS assumes a model specification in which $\mathbf{X}'\mathbf{e}/n = \mathbf{c} = \mathbf{0}$, that is that the \mathbf{X} 's and error term are uncorrelated. This is the key assumption in OLS and its potential Achilles heel. There are a variety of reasons that it may be violated: there are \mathbf{X} 's that have been omitted that effect \mathbf{Y} and are correlated with the \mathbf{X} 's in the model; some or all of the \mathbf{X} 's contain measurement error; \mathbf{Y} and some of the \mathbf{X} 's simultaneously determine each other; or there is selection on the dependent variable. These are all forms of misspecification error. The implication of each is that they will cause \mathbf{X} and \mathbf{e} to be correlated, leading OLS to be inconsistent. Thus, by focusing on the correlation between \mathbf{X} and \mathbf{e} , we are potentially dealing with a wide set of possible reasons for misspecification.

The validity of any set of OLS estimates rests with the plausibility of our assumptions that $\mathbf{X}'\mathbf{e}/n = \mathbf{c} = \mathbf{0}$. The researcher's job is to argue that any deviations are likely to be small, and it is the critics 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 equal to zero. There are always some \mathbf{X} 's that are omitted that potentially have some, though hopefully small, effect on \mathbf{Y} and are correlated with the \mathbf{X} 's in the model. In social science data, most variables have some amount of measurement error. In some cases simultaneity or selection bias is an issue. All of these problems lead to situations where \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 that $\mathbf{X}'\mathbf{e}/n$ will reasonably close enough

to zero that $\hat{\mathbf{b}}_{ols}$, to within sampling error, will be a good estimate of the true regression parameter, \mathbf{b} .

If we knew what \mathbf{c}^* equaled or at least a good guess, then we could use the generalized OLS estimator in equation (5) 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 what the adjustment factor should be. Below I will show how Bayesian methods can be used to deal with uncertainty about \mathbf{c}^* .

Model Misspecification and Multicollinearity

Multicollinearity occurs when one's \mathbf{X} 's or a subset of \mathbf{X} 's are nearly linear dependent on each other. This may occur because two variables are highly correlated or because one variable is well approximated by a linear function of the other independent variables. Multicollinearity is a property of the $\mathbf{X}'\mathbf{X}$ matrix which is equal to n times the covariance matrix of the independent variables \mathbf{X} . A variety of measures of the degree of multicollinearity have been proposed. See Belsley (1991) for a comprehensive review. These will not concern us here.

It is frequently noted that when multicollinearity is a problem, coefficient estimates can be highly sensitive to the particular model specification that is used. The addition or subtraction of specific "control" variables may substantially alter results. Similarly, transforming the scale of one or more variables (e.g. a variable in logged as opposed to linear form) may have a large effect on the coefficient estimates of other variables. In most, if not all cases, however, statistics and econometric textbook authors argue that having sufficiently more informative data, for example a

much larger sample, will solve this problem. As Gordon (1964) understood, and I will now more formally demonstrate, this is incorrect.

Consider equation (8) again:

$$(8) \quad \hat{\mathbf{b}}_{\text{gols}} = \mathbf{b} + \mathbf{n} (\mathbf{X}'\mathbf{X})^{-1} [(\mathbf{X}'\mathbf{e}/\mathbf{n} - \mathbf{c}^*) + (\mathbf{c}^* - \mathbf{c})]$$

S **M**

What we see here is that $(\mathbf{X}'\mathbf{X})^{-1}$ matrix acts on the term $[(\mathbf{X}'\mathbf{e}/\mathbf{n} - \mathbf{c}^*) + (\mathbf{c}^* - \mathbf{c})]$. When multicollinearity is a problem, it is analogous to a column of the $\mathbf{X}'\mathbf{X}$ matrix being nearly equal to zero. The matrix inversion relation is analogous to division. 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 increase the size of the quantity being operated on tremendously. So although the difference between $\mathbf{X}'\mathbf{e}/\mathbf{n}$ and \mathbf{c}^* due to sampling error may be quite small, the quantity $(\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{e}/\mathbf{n} - \mathbf{c}^*)$ may be quite large due to the effect of multicollinearity in $\mathbf{X}'\mathbf{X}$.

Note that the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix operates on two terms: $(\mathbf{X}'\mathbf{e}/\mathbf{n} - \mathbf{c}^*)$, the sampling error, and $(\mathbf{c}^* - \mathbf{c})$, the model misspecification error. Partially expanding equation (8) we get:

$$(9) \quad \hat{\mathbf{b}}_{\text{gols}} = \mathbf{b} + \mathbf{n} (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{e}/\mathbf{n} - \mathbf{c}^*) + \mathbf{n} (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{c}^* - \mathbf{c})$$

S*

M*

As far as I am aware, this decomposition (and equation (8)) are nonstandard, though the derivation is straightforward. This decomposition is key to understanding the effects of collinearity. The \mathbf{S}^* component of equation (8) represents the error in $\hat{\mathbf{b}}_{\text{goals}}$ due to sampling error. This is the traditional concern about multicollinearity. The \mathbf{M}^* component represents the error in $\hat{\mathbf{b}}_{\text{goals}}$ due to model misspecification. In both cases if there is multicollinearity in $\mathbf{X}'\mathbf{X}$, then small departures in either $(\mathbf{X}'\mathbf{e}/\mathbf{n} - \mathbf{c}^*)$ or $(\mathbf{c}^* - \mathbf{c})$ from zero can lead to large departures in $\hat{\mathbf{b}}_{\text{goals}}$ from \mathbf{b} due to the effects of multiplying by $(\mathbf{X}'\mathbf{X})^{-1}$.

As noted above the traditional statistics and econometrics literature is only concerned with the effects of multicollinearity on sampling error. Why hasn't this literature been concerned with misspecification error? For the simple reason that the traditional literature assumes that $\mathbf{X}'\mathbf{e}/\mathbf{n} = \mathbf{c} = \mathbf{0}$ 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.

To understand the problem of multicollinearity and model misspecification more intuitively, consider three hypothetical examples. Below is the covariance/correlation matrix for the first case:

Table 1
 First Hypothetical Example
 Covariance/Correlation Matrix

				Correlation	Covariance
	Y	X ₁	X ₂	e	e
Y	1	.25	.20	.968	.938
X ₁		1	.8	.026	.025
X ₂			1	-.026	-.025

Note that 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 to indicate first its correlation and then its covariance with each of the other variables. As described above in our discussion of equation (6), n times the covariance of \mathbf{X} and \mathbf{e} is the “correction” factor needed to adjust $\mathbf{X}'\mathbf{Y}$ in order for OLS to give consistent estimates. The OLS estimates are: $\hat{\mathbf{b}}_1 = .25$ and

$\hat{\mathbf{b}}_2 = 0$. These estimates are independent of sample size. The standard errors for these

regression estimates can simply be made as small as desired by choosing n to be large enough. In the extreme we could assume in this example 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 \mathbf{X}_2 has

no effect on \mathbf{Y} whereas the partial effect of \mathbf{X}_1 is equal to its zero order effect, i.e. its correlation. This is remarkable given that the correlations of the two \mathbf{X} 's with \mathbf{Y} only differ by .05.

Now we use the generalized ordinary least squares estimator and correct for the fact that the \mathbf{X} 's and \mathbf{e} are correlated. This amounts to adjusting the correlation of \mathbf{X}_1 and \mathbf{X}_2 with \mathbf{Y} , .25 and .20, respectively, by subtracting out their respective covariances with \mathbf{e} , .025 and -.025, giving .225 and .225 and then carrying OLS. Our new estimates are then $\hat{\mathbf{b}}_1 = \hat{\mathbf{b}}_2 = .125$.

What is surprising is that this considerable discrepancy between the OLS estimates and the "true" estimates has been induced by a very modest degree of model misspecification. \mathbf{X}_1 and \mathbf{X}_2 both have +/- .026 correlations with the error. These are small correlations. In most empirical applications it would untenable to argue that the actual correlations of \mathbf{X} 's with the error due to possible model misspecification would be smaller than this.

Assume instead that in the example above that \mathbf{X}_1 and \mathbf{X}_2 are uncorrelated. In this case the corrected regression parameters would be $\hat{\mathbf{b}}_1 = \hat{\mathbf{b}}_2 = .225$ and the OLS estimates would be $\hat{\mathbf{b}}_1 = .25$ and $\hat{\mathbf{b}}_2 = .20$. Each OLS estimate is off by .025, exactly the degree of model misspecification as represented by the covariance of \mathbf{X} and \mathbf{e} . This will generally be true when our \mathbf{X} 's are perfectly orthogonal to each other.

Now consider an example which in some respects produces even more extreme results.

Table 2

Second Hypothetical Example

Covariance/Correlation Matrix

		Correlation			Covariance	
		Y	X ₁	X ₂	e	e
Y	1	1	.22	.20		
X ₁			1	.8	.011	.01
X ₂				1	-.011	-.01

In this case the OLS estimates are $\hat{\mathbf{b}}_1 = .167$ and $\hat{\mathbf{b}}_2 = .067$. Here the effect of X₁ is more than twice the effect of X₂ despite the fact that the correlation of each X with Y differs by only .02. Again, if we corrected the OLS estimates or equivalently did generalized OLS, the “true” estimates would be $\hat{\mathbf{b}}_1 = \hat{\mathbf{b}}_2 = .117$. Here our estimates and the conclusions that we draw from them shift dramatically, even though our X’s are only correlated with the error +/- .011.

Consider one more hypothetical example where the degree of multicollinearity is smaller. I assume that the two X’s are correlated .5, a modest degree of collinearity.

Table 3

Third Hypothetical Example

Covariance/Correlation Matrix

				Correlation	Covariance
	Y	X ₁	X ₂	e	e
Y	1	.25	.20		
X ₁		1	.5	-.026	-.025
X ₂			1	.026	.025

Here the OLS estimates are $\hat{\mathbf{b}}_1 = .20$ and $\hat{\mathbf{b}}_2 = .10$ where, as in the first example, the

correlation of the two \mathbf{X} 's with \mathbf{Y} only differs by .05. Doing generalized OLS, the "true"

estimates are $\hat{\mathbf{b}}_1 = \hat{\mathbf{b}}_2 = .15$.

What these examples show is that multicollinearity can very 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 these examples. The sample size can be infinity and we will still have the same problem. What the reader should find disturbing in these examples is that in the presence of multicollinearity, even if we had a sample

of infinite size, small errors in the model specification can lead to large deviations in one's estimates. Multicollinearity produces potentially extreme sensitivity to model specification errors independent of sample size. Of course, this sensitivity is not reflected in the traditional standard errors computed by statistical packages. I now develop a Bayesian approach that accomplishes this.

A Bayesian Approach

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. In the Bayesian approach we increase the information in the analysis by incorporating information about our prior beliefs about the parameter estimates as opposed to adding new data points (Leamer 1994, Birkes and Dodge 1993, Judge et al. 1985). ²

Researchers often see the use of Bayesian methods to deal with multicollinearity, or even more generally, as unattractive, because they involve stronger assumptions than OLS. The approach here, however, uses Bayesian methods in order to make weaker assumptions than those underlying OLS. Specifically, I pursue a Bayesian approach as a means of incorporating uncertainty about the model specification. In spirit, my approach is similar to Raftery's (1995) work on model uncertainty, though the technical specifics are quite different. (See recent work by Western (1996) for various applications of Raftery's method.) Like Raftery's approach, the methods here involve making weaker assumptions about "the truth" and as result increasing the

amount of uncertainty associated with estimates. As result, these methods may well uncover situations where multicollinearity is a problem, but appears not to be, i.e., where traditional standard errors are small. This is particularly likely to be the case when the data matrix is highly informative, for example when sample sizes are quite large.

The problem with the classic regression model and ordinary least squares is the standard errors associated with parameters estimates only reflect error due to sampling. There is no way to incorporate uncertainty associated with the model specification. In this section I utilize Bayesian methods to provide for the possibility of uncertainty in the model specification. In doing so, I am able to produce confidence intervals for parameters that reflect both sampling error and potential misspecification error.³

Consider equation (6), the generalized ordinary least squares estimator, again:

$$(6) \quad \hat{\mathbf{b}}_{\text{gols}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y} - \mathbf{nc})$$

This model can be thought of as having two sets of parameters, the \mathbf{b} 's or regression parameters, and the \mathbf{c} 's representing a correction factor for $\mathbf{X}'\mathbf{Y}/\mathbf{n}$, equal to the covariance between \mathbf{X} and \mathbf{e} , that is used to account for the fact that the \mathbf{X} 's and \mathbf{e} may be correlated. As specified this model is unidentified without further assumptions about either \mathbf{b} and/or \mathbf{c} . Now in the classic approach we make an assumption about \mathbf{c} in order to solve out for \mathbf{b} . That is, we assume that the vector \mathbf{c} equals a specific set of values. This is a very strong assumption. If we make some specific assumption about the value of \mathbf{c} , then we can use equation (6) to estimate \mathbf{b} . From a Bayesian

perspective, the classic approach seems most peculiar. In a Bayesian context, the classic approach 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 can assume that the value of \mathbf{b} and \mathbf{c} are unknown, but that our beliefs about their true values can be captured in the form of a prior distribution. If our beliefs are represented by a proper distribution (and in some cases where the prior is improper), then the posterior estimates of our parameters, \mathbf{b} , will be well defined. What this means is 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).

In the context of our problem, the intuition as to why the posterior distribution of parameters in a Bayesian model has a proper posterior distribution can be easily understood. For the moment, assume that we have no prior information regarding \mathbf{b} . Using the generalized OLS estimator (equation 5) we can calculate the generalized OLS estimate \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} tells us the likelihood of different values of \mathbf{c} . From this we can then derive the likelihood of different values of \mathbf{b} . The likelihood of different values of \mathbf{b} will be a function of potential sampling error and the likelihood of different values of \mathbf{c} .⁴

Let $L(\cdot)$ be a generic symbol for a likelihood function. The general Bayesian formulation

of our model is from Bayes rule:

$$(9) \quad \mathbf{L}(\boldsymbol{\theta} | \mathbf{Y}) \propto \mathbf{L}(\mathbf{Y} | \boldsymbol{\theta}) \mathbf{L}(\boldsymbol{\theta})$$

where $\mathbf{L}(\boldsymbol{\theta} | \mathbf{Y})$ is the posterior distribution of $\boldsymbol{\theta}$, $\mathbf{L}(\mathbf{Y} | \boldsymbol{\theta})$ is the data likelihood and $\mathbf{L}(\boldsymbol{\theta})$ is the prior distribution for $\boldsymbol{\theta}$. $\boldsymbol{\theta}$ consists of \mathbf{b} , \mathbf{c} and σ^2 , the \mathbf{Y} conditional variance of \mathbf{Y} . I treat the \mathbf{X} as fixed known values and leave \mathbf{X} implicit in order to ease the notation. In order to simplify the math, I also assume for the purposes of this draft that σ^2 is known.⁵

Now since \mathbf{X} and \mathbf{e} are potentially correlated, the expected value of \mathbf{e} for an individual with a particular \mathbf{X} is not necessarily zero. Writing out the regression of \mathbf{e} on \mathbf{X} we get that $\mathbf{E}[\mathbf{e} | \mathbf{X}] = \mathbf{X}\mathbf{a}$ where $\mathbf{a} = \sum_{\mathbf{X}}^{-1} \mathbf{c}$. As a result, the conditional mean of \mathbf{Y} on \mathbf{X} , will depend on two vectors of parameters, \mathbf{b} and $\mathbf{a} = \sum_{\mathbf{X}}^{-1} \mathbf{c}$. Specifically, $\mathbf{E}[\mathbf{Y} | \mathbf{X}] = \mathbf{X}\mathbf{b} + \mathbf{E}[\mathbf{e} | \mathbf{X}] = \mathbf{X}\mathbf{b} + \mathbf{X}\mathbf{a} = \mathbf{X}\mathbf{s}$ where $\mathbf{s} = \mathbf{b} + \mathbf{a} = \mathbf{b} + \sum_{\mathbf{X}}^{-1} \mathbf{c}$. Thus the conditional mean of \mathbf{Y} is a linear function of \mathbf{b} and \mathbf{c} .

I will make the stronger assumption that the only way that \mathbf{Y} , not just its conditional mean, depends on \mathbf{b} and \mathbf{c} (or its linear transformation \mathbf{a}) is through $\mathbf{s} = \mathbf{b} + \sum_{\mathbf{X}}^{-1} \mathbf{c}$. As a result, $\mathbf{L}(\mathbf{Y} | \mathbf{b}, \mathbf{s}) = \mathbf{L}(\mathbf{Y} | \mathbf{b} + \sum_{\mathbf{X}}^{-1} \mathbf{c}) = \mathbf{L}(\mathbf{Y} | \mathbf{s})$. This assumption would hold, for example, if \mathbf{Y} were conditionally multivariate normal and its mean, but not its conditional covariance matrix, were a linear function of \mathbf{X} . The important implication of this is that it implies that the data only provides us with information on $\mathbf{s} = \mathbf{b} + \sum_{\mathbf{X}}^{-1} \mathbf{c}$. A key implication of this that we will use below is that in

terms of estimation we have a standard problem of estimating a regression model where some prior has been assumed for the regression parameters, \mathbf{s} . This means that standard Bayesian methods and programs can be used to do the empirical analysis.

Specifying a Prior

Researchers always object to using Bayesian methods because they involve specifying a prior. Bayesians typically respond that classic methods all implicitly assume some prior. We discussed this above, where I pointed out that classical regression assumes a noninformation prior for \mathbf{b} , but full knowledge of \mathbf{c} , that is, it assumes $\mathbf{c} = 0$.

This section discusses only one very simple approach to specifying a prior for \mathbf{c} . In terms of my discussion of multicollinearity, the major goal of the Bayesian approach developed here is to provide a way of estimating the sensitivity of estimates to small or modest changes in the model specification. In a future companion paper, I will examine more elaborate priors. Specifically, I will focus on ways in which different priors can be used to model specific ways in which we believe that a model may be misspecified. A model could be misspecified due to an omitted variable or variables, measurement error, selection, or simultaneity. Each of these problems potentially suggests a different prior. Here my goal is more modest. I am only interested in using a prior that allows us to move to a slightly more robust assumption than the extraordinarily strong assumption made in the classical regression model that $\mathbf{X}'\mathbf{e} = 0$ exactly.

Here I specifically want to examine situations where our best guess is that \mathbf{X} and \mathbf{e} are uncorrelated, but we are not fully confident that this is *exactly* true. The purpose in using the Bayesian methods presented in the last section is not to get precise results, but rather to get a

rough sense of how much estimates might be changed by a degree of uncertainty in the model specification.

It is difficult to think in terms of covariances because the unit of measurement changes depending on the two variables involved. Correlations, however, are much easier. In a particular piece of analysis the question is how much confidence we have in our model specification. Here, I think of confidence in one's model specification as being represented by a particular assumed value for the correlation between \mathbf{X} and \mathbf{e} . Somewhat arbitrarily I would suggest that a belief that the correlation probably falls between $-.02$ and $.02$ as very strong confidence in one's model specification, $-.05$ and $.05$ represents strong confidence, between $-.10$ and $.10$ moderate confidence, and between $-.20$ and $.20$ weak confidence. If we think about these intervals as approximately 95% Bayesian confidence intervals, then, assuming a normal prior, we are specifying priors for each parameter in the very strong confidence case as $n(0,0.01)$, in the strong confidence case as $n(0,.025)$, in the moderate confidence case as $n(0,.05)$, and in the weak confidence case as $n(0,.10)$. If we are simply interested in studying the sensitivity of estimates to small departures from the assumed model specification, then a first step would be to specify that the prior distribution for \mathbf{e} consists of a series independent normals with standard deviations appropriate to our level of confidence.⁶

The classic regression model is equivalent to a Bayesian model where we have assumed that \mathbf{b} has a diffuse prior, i.e. $\mathbf{f}(\mathbf{b}) \propto \text{constant}$. Here it will be easier, however, to work with a weak prior that will give us nearly the same result. We will assume that \mathbf{b} has a multivariate normal prior with $\boldsymbol{\mu}_b = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_b = \gamma^2 \boldsymbol{\Sigma}_X^{-1}$ where γ is a constant chosen

sufficiently large to represent a nearly diffuse prior for \mathbf{b} and so that the contribution of \mathbf{b} to the formula's below and in the appendix is negligible. This means that for moderate to large n , the contribution of the data will totally dominate the effect of the prior. As a result, our formulas and estimates will closely approximate the case where it is assumed that \mathbf{b} has a constant diffuse prior.

Also we will assume that in the prior \mathbf{b} and \mathbf{c} are independent. In the posterior, however, they will typically be quite highly correlated. Specifically, the \mathbf{b} associated with a specific \mathbf{X} will most cases be highly negatively correlated with the \mathbf{c} associated with that \mathbf{X} .⁷

Posterior for \mathbf{b}

In general our interest will not be in \mathbf{s} , but in \mathbf{b} , the “structural” regression coefficients in our model. Appendix A provides the derivation for the posterior likelihood of \mathbf{b} , $\mathbf{L}(\mathbf{b}|\mathbf{Y})$, as functions of our priors for (\mathbf{b}, \mathbf{c}) and the data, \mathbf{Y} . I use the “ \sim ” to indicate the posterior form of a parameter. There I work exclusively within the multivariate normal family. Above I defined $\mathbf{s} = \mathbf{b} + \mathbf{a}$ and $\mathbf{t} = \mathbf{b} - \mathbf{a} = \mathbf{b} - \sum_{\mathbf{X}}^{-1} \mathbf{c}$. This implies that $\mathbf{b} = (\mathbf{s} + \mathbf{t}) / (1 + \alpha)$ and $\mathbf{a} = (\mathbf{s} - \mathbf{t}) / (1 + \alpha)$. The result derived in Appendix A is that the posterior distribution for \mathbf{a} and \mathbf{b} equals:

$$\tilde{\boldsymbol{\mu}}_{\mathbf{a}} = (\hat{\mathbf{s}}_{\text{ols}} + \alpha \hat{\boldsymbol{\mu}}_{\mathbf{a}}) / (1 + \alpha)$$

$$\tilde{\boldsymbol{\mu}}_{\mathbf{b}} = \hat{\mathbf{s}}_{\text{ols}} - \tilde{\boldsymbol{\mu}}_{\mathbf{a}} = \hat{\mathbf{s}}_{\text{ols}} - \sum_{\mathbf{X}}^{-1} \tilde{\boldsymbol{\mu}}_{\mathbf{c}}$$

where $\tilde{\boldsymbol{\mu}}_{\mathbf{c}} = \sum_{\mathbf{x}} \tilde{\boldsymbol{\mu}}_{\mathbf{a}}$.

Note that the data is informative with respect to \mathbf{a} or equivalently \mathbf{c} . As a result, the posterior mean for \mathbf{b} consists of the posterior for \mathbf{s} which is approximately its OLS estimated adjusted by the posterior estimate of $\tilde{\boldsymbol{\mu}}_{\mathbf{a}} = \sum_{\mathbf{x}}^{-1} \tilde{\boldsymbol{\mu}}_{\mathbf{c}}$, not the prior means for \mathbf{a} and \mathbf{c} .

The components of the posterior covariance matrices for \mathbf{s} and \mathbf{t} are respectively:

$$\tilde{\Sigma}_{\mathbf{a}} = (\mathbf{1} / (\mathbf{1} + \alpha)^2) (\tilde{\Sigma}_{\mathbf{s}} + \tilde{\Sigma}_{\mathbf{t}})$$

$$\tilde{\Sigma}_{\mathbf{b}} = (\mathbf{1} / (\mathbf{1} + \alpha)^2) (\alpha \tilde{\Sigma}_{\mathbf{s}} + \tilde{\Sigma}_{\mathbf{t}})$$

Now since the data is noninformative with respect to \mathbf{t} , the posterior for \mathbf{t} is equivalent to its prior. If we assume that the prior mean of \mathbf{b} makes a negligible contribution to the prior for \mathbf{t} , then $\tilde{\mathbf{t}} = \tilde{\mathbf{a}}$ and we have:

$$\tilde{\Sigma}_{\mathbf{a}} = (\mathbf{1} / (\mathbf{1} + \alpha)^2) (\tilde{\Sigma}_{\mathbf{s}} + \alpha^2 \Sigma_{\mathbf{a}})$$

$$\tilde{\Sigma}_{\mathbf{b}} = (\alpha / (\mathbf{1} + \alpha))^2 (\tilde{\Sigma}_{\mathbf{s}} + \Sigma_{\mathbf{a}})$$

The key point to note here is that the posterior covariance matrix for \mathbf{b} , $\tilde{\Sigma}_{\mathbf{b}}$, is a function of two separate covariances. $\tilde{\Sigma}_{\mathbf{b}} \approx \sigma_e^2 (\mathbf{X}' \mathbf{X})^{-1}$ which is a function of the data and will go to zero as n goes to infinity. This term is the sampling variance of \mathbf{b} due to sampling error. The second component is \mathbf{a} . This is the variance of \mathbf{b} that is due to uncertainty in the model misspecification. It is constant and thus unaffected by sample size.

Empirical Example

As an empirical example I consider the familiar question of the effect of education and earnings on earnings. The data are taken from the National Longitudinal Survey of Youth. I have restricted the sample to white males with positive earnings in 1989. The ability measure is the Armed Forces Vocational Aptitude Battery (AFQT). In order to compare the effects of education and ability both have been standardized on the subpopulation. I have done this so that it is clear how the correction for model specification is done. If the goal were the substantive comparison of the effects of Ability and Education we would most likely want to standardize them relative to the full population or deal with unstandardized variables. The following table reports the correlations between these three variables.

Table 5

Correlations*

	Ln Wages	Ability	Education
Ln Wages	1.000	0.346	0.307
Ability		1.000	0.700
Education			1.000

*Sample restricted to white males with positive earnings in 1989.

As can be seen Ability and Education are highly collinear with a correlation of .70. Ability is somewhat more strongly correlated with log wages than is Education, though in absolute terms the difference is small, approximately .04. Table 6 reports the regression estimates:

Table 6

Regression Estimates

Effects of Ability and Education on Log Wages

	Standardized Coefficient	Standard Error
Ability	.258	.027
Education	.126	.027
	$R^2 = .128$	$N = 2274$

What we see here is that the effect of Ability is twice as large as the effect of Education. This is the case despite the fact that their correlations with Log wages only differ by .04. Notice, however, that the standard errors for the two variables are both quite small and it is quite evident that the two effects are not equal. Thus, what we see here is quite similar to the hypothetical examples above. There are a host of reasons that one might be concerned with error in the model specification here. First, both Ability and Education might be subject to measurement error. Second, there are certainly other variables that might be included in the model. For example many sociologists would argue, though economists would disagree, that measures of family background ought to be included in the model.

Assume that we have weak confidence in our model specification. For both X 's, we assume that the correlation between X and the error is distributed as a normal with mean zero and standard deviation .05. Since we are assuming that these correlations have mean zero, our coefficient estimates in Table 6 are unchanged. However, our standard errors which now incorporate uncertainty in the model specification will be larger. Specifically, we have:

Table 7

Regression Estimates

Effects of Ability and Education on Log Wages

	Standardized Coefficient	Classical Standard Error	Bayesian Standard Error
Ability	.258	.027	.054
Education	.126	.027	.054

$$R^2 = .128 \quad N = 2274$$

Now education is barely significant and the two sigma confidence intervals overlap.

(Further development of this example in the next draft. I will be redoing the calculation of the standard errors here. The direction is right, but I the formulas I used are not correct. Correct formulas next draft.)

Conclusion

Often when empirical researchers carry out a regression analysis they focus on the point estimates they have obtained from their analysis. This gives the impression that the data has given us one answer as to the effect a particular variable has. Of course, data are typically consistent

with a range of effects and the point estimate is only the estimate that is most likely. 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 reasonably accurately reflects the uncertainty about an estimate after the analysis has been carried out. The purpose of this paper has been to develop method for incorporating uncertainty about the model specification into one's standard errors and thus into an estimate's confidence interval. I have focused on the problem of multicollinearity and argued that multicollinearity not only introduces sensitivity in estimates to sampling error, but also model misspecification. I have developed the basic components of a Bayesian approach that allows the analyst to incorporate model specification uncertainty as well as possible sampling error into their analysis. 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 co-authors (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.

My hope is that this paper 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. This paper 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.

Bibliography

- Angrist, J.D., and Krueger, A.B. 1991. "Does compulsory School Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics*, 106: 979-1014.
- _____. 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.
- 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.
- 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*, vol. 90: 443-450.
- 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.
- 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. 1964. "Issues in Multiple Regression." *American Journal of Sociology*, : 592-616.
- 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. 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, Ser. 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.
- Neath, A. A. and Samaniego, F. J. 1996. "On Bayesian Estimation of the Multiple Decrement Function in the Competing Risks Problem." *Statistics and Probability Letters*.
- _____ 1997. "On the efficacy of Bayesian Inference for nonidentifiable models." *The American Statistician*, 51: 225 - 232.
- Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." in *Sociological Methodology 1995*, edited by Peter V. Marsden, 111-63. Cambridge, MA.: Blackwell Publishers.
- Western, Bruce. 1996. "Vague Theory and Model Uncertainty in Macrosciology." in *Sociological Methodology 1996*, edited by Arian E. Raftery, 165-193. Cambridge, MA.: Blackwell Publishers.
- Zellner, Arnold. 1971. *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons.

Appendix A

Derivation of Posterior Distribution for \mathbf{b}

In the text we defined $\mathbf{s} = (\mathbf{b} + \mathbf{a}) = (\mathbf{b} + \sum_{\mathbf{x}}^{-1} \mathbf{c})$. We also noted that the only way that \mathbf{Y} depends on \mathbf{b} and \mathbf{c} is through \mathbf{s} , that is, $\mathbf{L}(\mathbf{Y} | \mathbf{b}, \mathbf{c}) = \mathbf{L}(\mathbf{Y} | \mathbf{b}, \mathbf{s}) = \mathbf{L}(\mathbf{Y} | \mathbf{s})$. In terms of estimating our model, it is convenient to reparameterize in terms of \mathbf{s} and some auxiliary variable \mathbf{t} , such that the function $\mathbf{h}(\cdot)$ relating $(\mathbf{s}, \mathbf{t}) = \mathbf{h}(\mathbf{b}, \mathbf{c})$ is one to one onto. As a result, $\mathbf{h}(\cdot)$ will have a well-defined inverse. It will be convenient to define \mathbf{t} so that \mathbf{s} and \mathbf{t} have independent prior distributions, i.e. $\mathbf{L}(\mathbf{s}, \mathbf{t}) = \mathbf{L}(\mathbf{s}) \mathbf{L}(\mathbf{t})$.

Having reparameterized our model in \mathbf{s} and \mathbf{t} , our problem is to derive their posterior distribution:

$$\mathbf{L}(\mathbf{s}, \mathbf{t} | \mathbf{Y}) = \mathbf{L}(\mathbf{Y} | \mathbf{s}, \mathbf{t}) \mathbf{L}(\mathbf{s}, \mathbf{t})$$

Since \mathbf{s} and \mathbf{t} have independent priors, however we can rewrite this as:

$$\mathbf{L}(\mathbf{s}, \mathbf{t} | \mathbf{Y}) = \mathbf{L}(\mathbf{Y} | \mathbf{s}, \mathbf{t}) \mathbf{L}(\mathbf{s}) \mathbf{L}(\mathbf{t})$$

In the text we assumed that the only way that \mathbf{Y} depended on \mathbf{b} and \mathbf{c} (or \mathbf{a}) was through \mathbf{s} . As a result, $\mathbf{L}(\mathbf{Y} | \mathbf{s}, \mathbf{t}) = \mathbf{L}(\mathbf{Y} | \mathbf{s})$. Substituting this into equation (11) we get:

$$\mathbf{L}(\mathbf{s}, \mathbf{t} | \mathbf{Y}) = \mathbf{L}(\mathbf{Y} | \mathbf{s}) \mathbf{L}(\mathbf{s}) \mathbf{L}(\mathbf{t})$$

This implies that the posteriors of \mathbf{s} and \mathbf{t} are independent and that \mathbf{t} 's posterior distribution is its prior distribution, that is:

$$\mathbf{L}(\mathbf{s}, \mathbf{t} | \mathbf{Y}) = \mathbf{L}(\mathbf{s} | \mathbf{Y}) \mathbf{L}(\mathbf{t} | \mathbf{Y}) = \mathbf{L}(\mathbf{Y} | \mathbf{s}) \mathbf{L}(\mathbf{s}) \mathbf{L}(\mathbf{t})$$

Because the posterior likelihood for \mathbf{s} and \mathbf{t} separates into distinct components, they can be estimated separately, though, in fact, the data is only informative for \mathbf{s} . This means that the parameters of the posterior distribution of \mathbf{s} can be gotten using standard Bayesian methods and readily available computer programs. The parameters for the posterior distribution of \mathbf{t} are simply the parameters of its prior distribution.

Having used the data to estimate the posterior distribution for \mathbf{s} , and knowing the posterior distribution for \mathbf{t} , we can then derive the posterior distribution for (\mathbf{b}, \mathbf{c}) from the posterior distribution of (\mathbf{s}, \mathbf{t}) using standard methods for deriving the distribution of the transformation of variables. A sufficient condition for this to be possible is that our one to one, onto function $(\mathbf{s}, \mathbf{t}) = \mathbf{h}(\mathbf{b}, \mathbf{c})$ be differentiable. In almost all cases it should be possible to define \mathbf{t} with respect to (\mathbf{b}, \mathbf{c}) in such a way that this is the case. Typically, the object of final interest is the posterior distribution of \mathbf{b} , $\mathbf{Y}(\mathbf{b} | \mathbf{Y})$. This can be gotten by integrating the posterior distribution of (\mathbf{b}, \mathbf{c}) , $\mathbf{L}(\mathbf{b}, \mathbf{c} | \mathbf{Y})$ with respect to \mathbf{c} . I now provide a specific derivation based on the multivariate normal.

The derivation of specific formulas consists of three steps. We first derive the prior for (\mathbf{s}, \mathbf{t}) .

Define $\mathbf{t} = \mathbf{b} - \mathbf{a}$ where γ is a scalar chosen so that \mathbf{s} and \mathbf{t} are uncorrelated. We have $(\mathbf{s}, \mathbf{t}) = (\mathbf{b} + \mathbf{a}, \mathbf{b} - \mathbf{a}) = (\mathbf{b} + \sum_X^{-1} \mathbf{c}, \mathbf{b} - \sum_X^{-1} \mathbf{c})$. Second we state the posterior for \mathbf{s} . Because of the way we have structured the problem, this will be a standard problem in Bayesian analysis. As a result, formulas can be found in a multitude of text books on the Bayesian analysis. Finally, we derive the posterior distribution for \mathbf{b} from the posterior distribution for \mathbf{s} .

Prior for (s,t). We assume that the prior for \mathbf{b} is multivariate normal with mean, $\boldsymbol{\mu}_b = \mathbf{0}$ and covariance matrix $\Sigma_b = \gamma^2 \sum_X^{-1}$ where γ , as discussed in the text, is chosen large enough to represent a nearly diffuse prior and so that the contribution of its prior to various formulas is negligible. Further, as discussed in the text, we assume that the priors for \mathbf{b} and \mathbf{c} are independent. We assume that the prior for \mathbf{c} is multivariate normal with mean $\boldsymbol{\mu}_c$ and covariance matrix Σ_c . For the purposes of this paper, we further assume that $\boldsymbol{\mu}_c = \mathbf{0}$ and the covariance matrix $\Sigma_c = \sigma_c^2 \mathbf{I}$ where as discussed in the text σ_c^2 is chosen to represent our uncertainty in the model specification.

In this case, $\boldsymbol{\mu}_a = \sum_X^{-1} \boldsymbol{\mu}_c$ and $\Sigma_a = \sigma_c^2 \sum_X^{-2}$. We have that $(\mathbf{s}, \mathbf{t}) = (\mathbf{b} + \mathbf{a}, \mathbf{b} - \mathbf{a})$. As a result, the prior means, $(\boldsymbol{\mu}_s, \boldsymbol{\mu}_t) = (\boldsymbol{\mu}_b + \boldsymbol{\mu}_a, \boldsymbol{\mu}_b - \boldsymbol{\mu}_a) = (\mathbf{0}, -\boldsymbol{\mu}_a)$ and $\Sigma_s = \Sigma_b + \Sigma_a$ and $\Sigma_t = \Sigma_b + \Sigma_a$.

Posterior for s. Having now derived the prior for \mathbf{s} , standard formulas for the posterior of \mathbf{s} , $L(\mathbf{s}|\mathbf{Y})$ can be found in a variety of text books (Greene 1990,page 203). We use the “~” to represent the posterior parameter. The posterior mean for \mathbf{s} will equal:

$$\tilde{\boldsymbol{\mu}}_s = \mathbf{M}(\boldsymbol{\Sigma}_s^{-1} \boldsymbol{\mu}_s + [\sigma_e^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1} \hat{\mathbf{s}}_{ols})$$

where

$$\mathbf{M} = \tilde{\boldsymbol{\Sigma}}_s = (\boldsymbol{\Sigma}_s^{-1} + [\sigma_e^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1})^{-1}$$

The posterior mean of \mathbf{s} , $\tilde{\boldsymbol{\mu}}_s$, can be interpreted as a weighted average of its prior mean $\boldsymbol{\mu}_s$ and

the standard OLS estimate of \mathbf{s} , $\hat{\mathbf{s}}_{ols}$. In our case because we assume that the prior mean of \mathbf{s} is

zero, the posterior mean will equal $\tilde{\boldsymbol{\mu}}_s = \mathbf{M}([\sigma_e^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1} \hat{\mathbf{s}}_{ols})$. For large n , the first two

terms approximately cancel and thus $\tilde{\boldsymbol{\mu}}_s \approx \hat{\mathbf{s}}_{ols}$. Similarly, for large n , the covariance matrix for

the posterior distribution of \mathbf{s} will approximately equal, $\tilde{\boldsymbol{\Sigma}}_s \approx \sigma_e^2(\mathbf{X}'\mathbf{X})^{-1}$. Below it will be

convenient to use these approximations instead of the exact formulas.

Posteriors for \mathbf{b} and \mathbf{a} (and \mathbf{c}). The posterior of \mathbf{t} is equal to its prior since \mathbf{Y} is noninformative with respect to \mathbf{t} . Given that we have the posterior distributions for \mathbf{s} and \mathbf{t} , we can calculate the posterior for \mathbf{b} and \mathbf{a} and thus \mathbf{b} and \mathbf{c} . Specifically, we have that $\mathbf{b} = (\mathbf{s} + \mathbf{t})/(1 + \lambda)$ and $\mathbf{a} = (\mathbf{s} - \mathbf{t})/(1 + \lambda)$. As a result:

$$\tilde{\boldsymbol{\mu}}_{\mathbf{a}} = (\hat{\mathbf{s}}_{\text{ols}} + \alpha \boldsymbol{\mu}_{\mathbf{a}}) / (1 + \alpha)$$

$$\tilde{\boldsymbol{\mu}}_{\mathbf{b}} = \hat{\mathbf{s}}_{\text{ols}} - \tilde{\boldsymbol{\mu}}_{\mathbf{a}} = \hat{\mathbf{s}}_{\text{ols}} - \sum_{\mathbf{x}}^{-1} \tilde{\boldsymbol{\mu}}_{\mathbf{c}} \quad \text{where } \tilde{\boldsymbol{\mu}}_{\mathbf{c}} = \sum_{\mathbf{x}} \tilde{\boldsymbol{\mu}}_{\mathbf{a}}.$$

Note that the data is informative with respect to \mathbf{a} or equivalently \mathbf{c} . As a result, the posterior mean for \mathbf{b} consists of the posterior for \mathbf{s} which is approximately its OLS estimate adjusted by the posterior estimate of $\tilde{\boldsymbol{\mu}}_{\mathbf{a}} = \sum_{\mathbf{x}}^{-1} \tilde{\boldsymbol{\mu}}_{\mathbf{c}}$, not the prior means for \mathbf{a} and \mathbf{c} .

The components of the posterior covariance matrices for \mathbf{s} and \mathbf{t} are respectively:

$$\tilde{\boldsymbol{\Sigma}}_{\mathbf{a}} = (\mathbf{1} / (\mathbf{1} + \alpha)^2) (\tilde{\boldsymbol{\Sigma}}_{\mathbf{s}} + \tilde{\boldsymbol{\Sigma}}_{\mathbf{t}})$$

$$\tilde{\boldsymbol{\Sigma}}_{\mathbf{b}} = (\mathbf{1} / (\mathbf{1} + \alpha)^2) (\alpha \tilde{\boldsymbol{\Sigma}}_{\mathbf{s}} + \tilde{\boldsymbol{\Sigma}}_{\mathbf{t}})$$

[There is a mistake in the derivation here that I can't figure out. I will straighten it out in the next draft.]. Now since the data is noninformative with respect to \mathbf{t} , the posterior for \mathbf{t} is equivalent to its prior. If we assume that the prior mean of \mathbf{b} makes a negligible contribution to the prior for \mathbf{t} ,

$\mathbf{t} \approx \mathbf{a}$ and we have:

$$\tilde{\Sigma}_a = (\mathbf{1}/(\mathbf{1} + \alpha)^2)(\tilde{\Sigma}_s + \alpha^2 \Sigma_a)$$

$$\tilde{\Sigma}_b = (\alpha / (\mathbf{1} + \alpha))^2(\tilde{\Sigma}_s + \Sigma_a)$$

The key point to note here is that the posterior covariance matrix for \mathbf{b} , $\tilde{\Sigma}_b$, is a function of two separate covariances. $\tilde{\Sigma}_s \approx \sigma_e^2(\mathbf{X}'\mathbf{X})^{-1}$ which is a function of the data and which goes to zero as n goes to infinity. This term is the sampling variance of \mathbf{b} due to sampling error. The second component is Σ_a . This is the variance of \mathbf{b} that is due to uncertainty in the model misspecification. It is constant and thus unaffected by sample size.

[Formula for Σ_a in next draft].

Endnotes

1. One could also worry about whether the value of $(X'X)^{-1} X'e$ 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.
2. The ridge estimator can be justified in part as a special example of a Bayesian estimator (Birkes and Dodge 1993, Judge et al. 1985).
3. In spirit, the approach here is also similar to that of Manski's (1995) work on bounds. Manski's work involves bounding the 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.
4. The thinking here is similar to Raftery's (1985) for pooling across models. In Raftery's approach a "pooled" parameter estimate is obtained by using a weight combination of estimates across different models where the weights are proportional to the posterior likelihood of each model. In Raftery the number of models is finite. Here it is infinite as parameterized by \mathbf{c} . Here the likelihood of different models is determined by the prior distribution for \mathbf{c} as opposed to the posterior likelihoods.
5. Below I will work within the multivariate normal family. The only consequence of not assuming that σ^2 is known, is that our posterior for \mathbf{b} would be a multivariate t instead of a multivariate normal distribution. This could be important in small samples, but will be of negligible importance in samples of greater than a few hundred.
6. In the case with two X 's assuming that their two \mathbf{c} 's were negatively correlated would produce larger standard errors than if we assumed that they were uncorrelated or positively correlated. This is demonstrated in the hypothetical examples presented earlier. The assumption of zero correlation seems like a reasonable middle ground.
7. A problem that needs attention is that I have specified uncertainty about model specification in terms of the correlation between each \mathbf{X} and \mathbf{e} , but \mathbf{c} is a measure of the covariance between \mathbf{X} and \mathbf{e} . To transform the correlation between \mathbf{X} and \mathbf{e} into a covariance we need to know the variance of each \mathbf{X} and the variance of \mathbf{e} . Since we are assuming that the \mathbf{X} 's are fixed, the variance of each \mathbf{X} can simply be calculated from the data. The variance of \mathbf{e} is unobserved and is strictly speaking an unknown parameter of the analysis that should be given a prior of its own. Given that our interest is obtaining a rough sense of the sensitivity of results to model specification, a bit of "fudge" seems appropriate. In particular, we can carry out a standard OLS analysis and calculate the variance of the residual. This, however, is not the variance of \mathbf{e} since \mathbf{e}

depends on \mathbf{X} . The variance of the residual is only the variance of that portion of the error that does not depend on \mathbf{X} .

We can, however, easily calculate what \mathbf{r}^2 is for the regression of \mathbf{e} on \mathbf{X} . This is, $\mathbf{r}^2 = \mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}$ where \mathbf{R} is a \mathbf{k} by 1 vector of assumed correlations between \mathbf{X} and \mathbf{e} . Dividing the residual variance by $1 - \mathbf{r}^2$ then gives us an estimate of the variance of the error. When the sample sizes are relatively large, which is when the methods here are most useful, the “fudge” in creating this estimate for the error variance should have little effect on the analysis.