Revisiting Multicollinearity:

When Correlated Predictors Exhibit Nonlinear Effects or Contain Measurement Error

Nathan Favero

Department of Public Administration & Policy

School of Public Affairs, American University

favero@american.edu

Prepared for the Association for Public Policy Analysis and Management's Annual Fall Research Conference in Washington, DC, November 3-5, 2016.

Revisiting Multicollinearity:

When Correlated Predictors Exhibit Nonlinear Effects or Contain Measurement Error

Abstract

While multicollinearity weakens statistical power, the presence of correlation among predictors (multicollinearity) violates no assumptions of the standard linear regression model. This fact has led many scholars to conclude that multicollinearity poses no problems to valid statistical inference when significant results are obtained. While this conclusion is correct when all regression assumptions are perfectly met, multicollinearity can exacerbate problems associated with model misspecification or measurement error. In this paper, I use Monte Carlo simulations to demonstrate the effect of multicollinearity on type I errors (false positives) when nonlinearities are incorrectly modeled and when classical measurement error is present in some of the predictors. I conclude by offering a set of practical suggestions to applied researchers who encounter multicollinearity in their data.

Revisiting Multicollinearity:

When Correlated Predictors Exhibit Nonlinear Effects or Contain Measurement Error

Multicollinearity is frequently invoked as a potential explanation for why results are insignificant. Considerable attention has been devoted to discussing a number of approaches—dropping or transforming variables, collecting more data, combining multiple variables into a single index, or ridge regression—that may increase one's ability to find significant coefficient estimates when multicollinearity is present in a dataset. Much less attention has been devoted to the topic of whether multicollinearity can pose (or at the very least signal) problems other than reduced precision of coefficient estimates. The dominant message from econometric literature appears to be that multicollinearity should cause no alarm if statistically significant results are found.[1] After all, under the Gauss-Markov assumptions, coefficient and standard errors estimates are unbiased regardless of how much (imperfect) multicollinearity is present in the data.

While the prevailing advice from econometric treatments of multicollinearity is appropriate for some situations, I argue that multicollinearity should not be ignored—even when results are significant—when researchers are testing for nonlinear effects or examining data that is likely to contain measurement error. When predictors are highly correlated, they can easily serve as proxies for one another, which can cause problems (such as false positives) when

[1] More precisely, the dominant message is that the presence of high levels of multicollinearity does not constitute legitimate grounds for doubting the results of a test of statistical significance where the null hypothesis is rejected. If one cares about obtaining precise estimates of the size of a coefficient (rather than just conducting a hypothesis test), multicollinearity might still pose a problem.

nonlinear relationships are misspecified or variables are imperfectly measured. Applied researchers should be aware of these potential problems so that they can make educated judgements about what to do or what limitations apply to their results when there are high levels of multicollinearity in their data.

In the following pages, I begin by taking a look at how the general topic of multicollinearity is currently discussed in political science and econometric literatures. I then examine how multicollinearity relates to the issue of model misspecification when testing for nonlinear relationships. Monte Carlo simulations are used to demonstrate the performance of competing nonlinear models with regards to the frequency of type I and type II errors under various levels of multicollinearity. Next, I consider the issue of measurement error and how its effect depends on the correlations among predictors. Another set of Monte Carlo simulations are conducted in order to test the extent to which multicollinearity and measurement error interact to produce increasing levels of bias in multiple regression coefficient estimates. I conclude with a discussion of how applied researchers can appropriately respond when they find high levels of multicollinearity in their data.

Multicollinearity in the Literature

Multicollinearity is standard topic for introductory courses on linear regression. In both methods textbooks and journal articles, multicollinearity discussions overwhelmingly focus on three aspects of the topic: (1) how to detect it, (2) what its effects are, and (3) what (if anything) researchers can do to obtain more precise estimates when they encounter multicollinearity (Dormann et al. 2013; Farrar and Glauber 1967; Graham 2003; Greene 2012, 89-94; Grewal, Cote, and Baumgartner 2004; Gujarati and Porter 2009, 320-364; Hill and Adkins 2001; Silvey 1969; Wold et al. 1984; Wooldridge 2013, 95-98). While a number of statistics have been

created to aid in the detection of multicollinearity, the Variance Inflation Factor (VIF) is the most commonly used and discussed (O'Brien 2007; Wooldridge 2013, 98). A separate VIF must be calculated for each independent variable and indicates the extent to which the variance in that variable is shared with other independent variables. It is calculated as $1/(1-R_i^2)$, where R_i^2 is the R^2 value resulting from a regression where variable i is predicted by all of the other independent variables. A VIF of 1 indicates that a variable is completely uncorrelated with the other independent variables while larger values indicate that a greater proportion of the variance can be predicted by the other independent variables. Some scholars have used rules of thumb stating that there is a multicollinearity problem if VIFs exceed 4 or 10, but these rules of thumb have been criticized (O'Brien 2007; Wooldridge 2013, 98).

The effect of multicollinearity that receives the most attention is that it causes large standard errors for coefficients (and thus imprecise coefficient estimates), making it difficult to find statistically significant results (for individual variables). This reflects the fact that it is difficult to precisely determine the independent effects of variables that mostly vary in concert. The square root of a VIF indicates how many times larger the standard error for the given variable is than it would be if there was no multicollinearity present. For example, a variable with a VIF of 9 will have a standard error that is 3 times larger than it would be if it were uncorrelated with the other independent variables. Even if it is impossible to reliably determine independent effects for a set of variables that mostly vary together in a sample, the variables may still exhibit joint significance. In addition to large standard errors, other (related) effects of multicollinearity that are often discussed are that coefficient estimates can change dramatically in response to changes in the sample or model and that coefficients can have implausible directions or magnitudes (Farrar and Glauber 1967; Greene 2012, 89; Hill and Adkins 2001). For nonlinear

models (including maximum likelihood estimation models), multicollinearity may also cause failure to converge (Hill and Adkins 2001).[2] If certain model selection or model averaging approaches are used, multicollinearity can cause bias in the final results because incorrect models are likely to be chosen or given some weight (Freckleton 2011; Graham 2003).

As long as the Gauss-Markov assumptions are met and errors are normally distributed, the inflation in standard errors caused by multicollinearity is appropriately reflected in linear regression results, and all estimates are unbiased. For this reason, some have taken care to not overstate the problems associated with multicollinearity. For example, Achen (1982, 82) writes: "[M]ulticollinearity violates no regression assumptions. Unbiased, consistent estimates will occur, and their standard errors will be correctly estimated. The only effect of multicollinearity is to make it hard to get coefficient estimates with small standard error." Under this viewpoint, multicollinearity is not a problem at all as long as the coefficient estimates provided by the regression analysis are precise enough to fit the researcher's purpose. For example, when hypothesis tests are being conducted, low statistical power due to multicollinearity is not a true problem if the null hypotheses can be rejected. O'Brien (2007, 683) discourages researchers from treating multicollinearity as grounds for doubting the estimates derived from a regression yielding significant results:

> If a regression coefficient is statistically significant even when there is a large amount of multi-collinearity – it is statistically significant in the "face of that

[2] It has also been shown that multilevel models can produce biased parameter estimates when there are high levels of multicollinearity (Can, van de Schoot, and Hox Forthcoming; Shieh and Fouladi 2003).

> collinearity." It is no more appropriate to question its statistical significance because there is multi-collinearity than to question a statistically significant relationship (at a specified level) because the variance explained by the model is low.

Though rarely stated so explicitly, O'Brien's perspective seems to be implicitly accepted among nearly all methods sources writing about multicollinearity. This apparent consensus does not extend to all applied researchers; in fact, the impetus for O'Brien's piece seems to be the prevalence of the belief that multicollinearity is a problem even if results are significant.[3]

A number of suggestions have been offered regarding how one should deal with data exhibiting high levels of multicollinearity. Virtually all of the approaches offered are aimed at increasing the precision of coefficient estimates. Standard suggestions include collecting additional data, altering model specification (dropping a variable, transforming data), reconsidering theoretical arguments, restricting parameters based on nonsample information (theory or prior findings), reducing the number of predictors by combining variables into indexes (e.g., through factor analysis), or using special techniques like ridge regression or principal components regression (Greene 2012, 91; Gujarati and Porter 2009, 342-346; Hill and Adkins 2001). Unfortunately, all of the available statistical solutions risk introducing bias into coefficient estimates; adjustments to specification risk introducing omitted variable bias or

[3] The prevalence of this belief among applied researchers despite an apparent consensus among methods sources that the belief is wrong may be due to the fact that many methods sources never explicitly state that multicollinearity isn't a problem if results are significant. This frequent omission is perhaps itself another impetus for O'Brien's piece.

specification error while ridge regression and principal components regression are known to be biased estimators (Hill and Adkins 2001; O'Brien 2007).

Since the methods literature does not generally view multicollinearity as a problem when one is satisfied with the level of precision in the coefficient estimates (as in the case where one finds a significant relationship in a hypothesis test), there is very little explicit discussion of what should be done in such cases. Gujarati and Porter (2009, 342) explain that one reasonable response to multicollinearity is to simply "do nothing" and learn what one can from the regression results. Similarly, O'Brien (2007, 681) states that, "Even with VIF values that greatly exceed the rules of 4 or 10, one can often confidently draw conclusions from regression analyses. How confident one can be depends upon the t-values and/or confidence intervals, which the variance of the regression coefficients help generate."

Though the existing literature is certainly correct to point out that multicollinearity violates no regression assumption, it is not enough to simply understand the behavior of our models when assumptions are perfectly met. Discussions of multicollinearity have generally neglected to consider whether multicollinearity might act to either mitigate or exacerbate problems caused by common violations of regression assumptions. In the following two sections, I consider the effect of multicollinearity within the context of two violations: model misspecification and measurement error.

Nonlinear Relationships and Model Misspecification

It is widely known that omitting an important independent variable that is correlated with a predictor that is included in the regression will cause omitted variable bias. However, little attention in political science or economics has been paid to the bias that may be introduced in models where one nonlinear term (e.g., an interaction) is included but other potential nonlinear

terms (e.g., squared terms) that are correlated with the included term are omitted from the regression. While largely ignored within political science, this issue has been discussed in some detail by a few studies published in psychology and management journals (Cortina 1993; Ganzach 1997, 1998; MacCallum and Mar 1995). These studies caution researchers of the potential bias that can plague the estimation of an interaction effect between two correlated variables if the variables exhibit unmodeled direct curvilinear effects. This bias results because "when the correlation between X and Z increases so does the correlation between XZ and X^2, which results in an overlap between the variance explained by XZ and the variance explained by X^2" (Ganzach 1997, 236). Just as failing to control for some variable x_2 that may reasonably be expected to affect the dependent variable risks biasing estimates for the main predictor x_1 if x_2 is correlated with x_1, failing to control for other nonlinear effects (squared or interaction terms) that may reasonably be expected to exist risks biasing estimates for a nonlinear relationship of interest (x_1x_2, ${x_1}^2$, or ${x_2}^2$) if relevant predictors (x_1 and x_2) are correlated. Given the complex nature of social phenomena, there is good reason to believe that curvilinear and interactive relationships abound for many variables of interest to political scientists.

Of course, including variables that are not part of the true regression model makes it more difficult to obtain precise estimates, though it does not bias results. When variables that may be unnecessary are highly correlated with other independent variables, including them in the regression risks interfering with one's ability to find significance for variables that are truly related to the outcome. Ganzach (1998, 621) uses simulation results to argue that "in most of the situations which are encountered by researchers in management, adding quadratic terms does not result in a considerable increase in the probability of type II error in detecting interaction if the true regression equation does not include quadratic terms." More specifically, he finds that

"[o]nly when multicollinearity is very high (i.e., [correlation] above .7), does the addition of quadratic terms have a substantial impact on the probability of type II error" (619).

Table 1 reports the results of a set of Monte Carlo simulations that demonstrate the biasing effect that misspecification of nonlinear effects can have when predictors are correlated. 5,000 samples were randomly generated for each of 20 sets of parameters. The 20 conditions varied in terms of the true regression equation (ether a curvilinear quadratic or an interactive model), the correlation between the two predictors (0, .3, .6,. .9, .95), and the sample size (50, 500).[4] When the true regression equation contains a curvilinear effect for one out of two predictors ($y = x_1 + x_2 + x_1{}^2 + \varepsilon$), running an interactive model with no squared terms ($y = \boldsymbol{\beta_1} x_1 + \boldsymbol{\beta_2} x_2 + \boldsymbol{\beta_5} x_1 x_2 + \varepsilon$) can produce a substantial number of false positives. When the two predictors are uncorrelated, false positives at the .05 level (two-tailed tests) for the interactive term occur 24.8% of the time in small samples (50 observations) and 29.2% of the time in large samples (500 observations). If the predictors are correlated at .3, the false positive rates increase to 55.6% and 99.9%. At correlations of .6 and higher, false positive rates are at 94.5% or higher.

Just as an unmodeled curvilinear relationship can bias estimates for an interactive effect, an unmolded interaction can bias estimates of a curvilinear effect. Estimating a curvilinear model without any interaction term ($\boldsymbol{y = x_1 + x_2 + x_1{}^2 + \varepsilon}$) when the true model is interactive ($\boldsymbol{y =}$

[4] $\boldsymbol{x_1}$ and $\boldsymbol{\varepsilon}$ were independently drawn from a standard normal distribution. $\boldsymbol{x_2}$ was computed as $\boldsymbol{\rho * x_1 + \sqrt{1 - r^2} * v}$, where $\boldsymbol{\rho}$ is the correlation between $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ selected for that set of samples and $\boldsymbol{v}$ is drawn from an independent normal distribution with a mean of zero and a variance of one. Thus, $\boldsymbol{x_2}$ has a mean of zero, a variance of one, and an expected correlation of $\boldsymbol{r}$ with $\boldsymbol{x_1}$.

$\boldsymbol{\beta_1 x_1 + \beta_2 x_2 + \beta_5 x_1 x_2 + \varepsilon}$) produces false positives at rate that is alarming but still slightly lower than in the reverse situation. 19.2% of small samples and 26.1% of large samples produced false positives for the squared term when the predictors were uncorrelated. Correlation at the .3 level causes false positive rates of 49.9% for small samples and 99.7% for large samples. False positive rates exceed 90% whenever predictors are correlated at .6 or higher.

The problem of false positives can be eliminated by running an overspecified model that includes both curvilinear quadratic terms and an interaction term ($y = \boldsymbol{\beta_1} x_1 + \boldsymbol{\beta_2} x_2 + \boldsymbol{\beta_3} {x_1}^{\boldsymbol{2}} + \boldsymbol{\beta_4} {x_2}^{\boldsymbol{2}} + \boldsymbol{\beta_5} x_1 x_2 + \varepsilon$). The rate of false positives drops down to 5% regardless of the parameters chosen. The overspecified model also does a fairly good job of correctly identifying the nonlinear relationship present in the true regression equation. When the true model has a curvilinear quadratic term for one of the predictors, the term is correctly identified as positive and significant at least 98.7% of the time in small samples if the predictors are correlated at .6 or less. When the predictors have a correlated of .9, this rate drops down to 36.1% for small samples. In large samples, the coefficient for the correct quadratic term is identified as positive and significant in virtually every sample when the predictors are correlated at .9 or less and 85.5% of the time when predictors are correlated at .95. If the true regression equation has only an interactive term, the nonlinear effect is somewhat more difficult to detect. When predictors have a correlation of .3 or less, the correct term has a positive and significant relationship 98.5% of the time or more in both small and large samples. At .6 correlation, the small sample produces a positive and significant coefficient 85.7% of the time while the large sample does so 100% of the time. A correlation of .9 brings the rate for small samples down to 13.4%, but large samples still yield a positive, significant relationship 86.5% of the time. A correlation of .95, however, brings the rate for large samples down to 35.7%

The decision of how many nonlinear terms to include in a model requires weighing a tradeoff between the risk of type I and type II errors. Including terms that are not needed in the model will increase the risk of type II errors (null results). The risk of type II errors is particularly strong for interactive terms and when samples are small. The rate of type II errors also increases dramatically as the correlation between predictors becomes stronger. Unfortunately, the risk of type I errors (false positives) when a model is misspecified is also strongest when predictors are most strongly correlated. If quadratic or interaction terms that are part of the true regression equation are omitted, false positives on the nonlinear terms are very likely even at relatively low levels of correlation between the predictors (.3). Large samples increase the likelihood of false positives when the model is misspecified, and interactive terms appear to be slightly more vulnerable than squared terms to type I errors. In some cases, theory may justify the assumption that direct effects are strictly linear for one or both variables (or that no interaction is present), reducing the number of terms that need to be included.

Measurement Error in Predictors

A large literature has developed around the problem of measurement error and offers a variety of models to correct for measurement error (see Buonaccorsi 2010). Familiar to many political scientists may be the structural equation modeling (SEM) approach which dominates the field of psychology. Unfortunately, techniques used to correct for measurement error require information on how precisely each variables is measured (or multiple measures of a construct which allow one to derive estimates of how reliably the construct has been measured). In many applications, such information is simply not available to researchers, making measurement error models of limited use.

Estimating a regression with data that imperfectly measure one or more of the independent variables found in the true regression equation constitutes a violation of standard regression assumptions and often has severe consequences. Though some applied scholars believe that measurement error will only serve to attenuate relationships, even completely uncorrelated measurement error can actually bias coefficient estimates away from zero in a multiple regression setting (Jackman 2008; McAdams 1986).

Measurement error can be represented mathematically with the following equation:

$$\boldsymbol{w} = \boldsymbol{x} + \boldsymbol{u}$$

where $\boldsymbol{w}$ is a proxy measure for $\boldsymbol{x}$ and $\boldsymbol{u}$ is the measurement error. A particularly simple type of measurement error can be considered by imposing the assumption that the error is unrelated to the true value of the variable:

$$E(\boldsymbol{u}|\boldsymbol{x}) = \mathbf{0}$$

Suppose one has the regression equation

$$\boldsymbol{y} = \beta_1 \boldsymbol{x_1} + \boldsymbol{\beta_2} X_2 + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$ is a vector of values for the dependent variable, $\boldsymbol{x_1}$ is a vector of values of a predictor that cannot be measured, X_2 is a matrix of other predictors (that are perfectly measured), $\boldsymbol{\varepsilon}$ is a vector containing the error term, β_1 is a scalar representing a regression coefficient, and $\boldsymbol{\beta_2}$ is a vector of coefficients. This regression equation is then estimated as

$$\boldsymbol{y} = \widehat{\beta_1} \boldsymbol{w_1} + \widehat{\boldsymbol{\beta_2}} X_2 + \hat{\boldsymbol{\varepsilon}}$$

because only a proxy measure $\boldsymbol{w_1}$ is available for $\boldsymbol{x_1}$. Furthermore, it is assumed that $\boldsymbol{w_1}$ exhibits what I refer to as uncorrelated measurement error, meaning that the measurement error is independent of all predictors and other error terms:

$$\boldsymbol{w_1} = \boldsymbol{x_1} + \boldsymbol{u_1}$$

$$E(\boldsymbol{u_1}|\boldsymbol{x_1}, X_2, \boldsymbol{\varepsilon}) = \mathbf{0}$$

Since only one predictor is measured with error, the estimated coefficient for the imperfectly measured predictor ($\widehat{\beta_1}$) will be biased towards zero. Thus, the measurement error will not cause any risk of false positives in a test of statistical significance for the imperfectly measured variable ($\boldsymbol{x_1}$). However, the same cannot be said for the other (perfectly measured) independent variables. Because the model cannot fully control for the imperfectly measured variable ($\boldsymbol{x_1}$), some of the variance attributable to that variable may be misattributed to other predictors (X_2). As such, the estimates contained in $\widehat{\boldsymbol{\beta_2}}$ may be biased either towards or away from zero, depending upon signs of β_1 and $\boldsymbol{\beta_2}$ and the patterns of covariance among the predictors (Achen 1983; Buonaccorsi 2010, 112-113). Once one moves beyond the world of one imperfectly measured predictor, generalizing becomes increasingly difficult. With multiple imperfectly measured predictors, the coefficients for all predictors (including those that are measured imperfectly) can be either inflated or attenuated (Achen 1983).

In the measurement error literature, multicollinearity is occasionally referred to as one determinant of the severity of bias problems associated with measurement error. Buonaccorsi (2010, 112) notes that under assumptions similar to those used in the prior two paragraphs (except that multiple predictors can be measured imperfectly), predictors that are perfectly measured will have unbiased coefficient estimates if the perfectly measured variables are uncorrelated with the imperfectly measured variables. Operating under similar assumptions, Achen (1983, emphasis in original) briefly mentions that "high collinearity will induce relatively large asymptotic biases, *no matter how small the measurement error variance*." In the case correlated measurement errors, Achen (1985) shows that correlation between predictors with

correlated measurement errors can cause serious bias, potentially reversing the sign of the coefficient for one of the predictors.

Despite these occasional explicit references to multicollinearity in the measurement error literature, I have come across only one mention of measurement error in the multicollinearity literature. In an article on multicollinearity published in animal behavior journal, Freckleton (2011) examines the effects of measurement error by simulating data for a regression on two predictors, one of which is measured with error. The measurement error causes bias in the coefficient estimates, and he observes that: "The effects of this bias become extremely important when collinearity between the variables exists…. What is happening is that the measurement error in x_2 results in under-estimation in the effect of this variable and, as the collinearity between the predictor increases, the effect of x_2 is mis-attributed to x_1" (97).

I conduct a set of Monte Carlo simulations that build on the insights provided by the measurement error literature in order to better understand the bias caused by measurement error given various levels of multicollinearity. Specifically, I wish to determine the rate of false positives under various conditions for a predictor that is not included in the true regression equation when there are several predictors included in the model. For all conditions, the true regression equation is

$$\boldsymbol{y} = \boldsymbol{x}_2 + \boldsymbol{x}_3 - \boldsymbol{x}_4 + \boldsymbol{\varepsilon}$$

which is estimated (using OLS) with proxy measures of the predictors as

$$\boldsymbol{y} = \widehat{\beta_0} + \widehat{\beta_1}\boldsymbol{x}_1 + \widehat{\beta_2}\boldsymbol{w}_2 + \widehat{\beta_3}\boldsymbol{w}_3 + \widehat{\beta_4}\boldsymbol{w}_4 + \widehat{\beta_5}\boldsymbol{x}_5 + \hat{\boldsymbol{\varepsilon}}$$

where

$$\boldsymbol{w}_i = \boldsymbol{x}_i + \tau_{i,p}/\left(1 - \tau_{i,p}\right)\boldsymbol{u}_i \; \forall \; i \in \{2,3,4\}, p \in P$$

$\boldsymbol{\varepsilon}$ and all $\boldsymbol{u}_i$ are error terms independently drawn from a standard normal distribution, and $\tau_{i,p}$ is the simulation parameter indicating the proportion of variance in proxy measure $\boldsymbol{w}_i$ attributable to measurement error under simulation parameter set $p \in P$. 5000 samples were randomly created for each set of simulations parameters. The conditions vary in terms of sample size (50, 500) and the proportion of variance in the proxy measures of the predictors that is attributable to measurement error (0, .01, .05, .1, .25). Since $\boldsymbol{x_1}$ and $\boldsymbol{x_5}$ are unrelated to the dependent variable, adding measurement error to them should have no effect on the simulation results. After all, adding white noise to white noise only creates whites noise with wider variance.

The independent variables ($\boldsymbol{x}_i$) each have a mean of zero and a variance of one, and they are drawn with expected covariance patterns that are randomly determined separately for each sample. Specifically, five latent variables were independently drawn from a standard normal distribution. Then, twenty-five linking variables (corresponding to each combination of an independent variable and a latent variable) were created by taking the absolute value of an independent draw from the standard normal distribution. For each independent variable, five of the linking variables were summed, and then each of these five linking variables was converted to a proportion of the sum. The five proportions determined the proportion of variance in the independent variable that was attributable to each of the five latent variables. Each independent variable was then created by summing the five latent variables, with each latent variable weighted by the square root of the corresponding proportion and multiplied by a -1 if a random draw from a Bernoulli distribution (p=.5) produced a zero. This process produces bivariate correlations among the proxy predictors in the various samples that have a mean of zero and a variance of approximately .43.

Since expected patterns of covariance among the predictors were determined randomly, multicollinearity varies continuously and randomly throughout the samples generated under each set of simulation parameters. Thus, results cannot be easily summarized without estimating a model to fit the continuous variation in levels of multicollinearity. For each set of simulation parameters, I use logistic regression to predict the probability of obtaining a false positive result (significant coefficient) at the .05 alpha level (using a two-tailed t-test). I model the probability of a false positive as a function of the level of multicollinearity in the data (measured as a VIF, explained more below) with polynomial terms going up to a fifth-order polynomial in order to allow for curvilinear effects. Because of some extreme values of the VIF that led to unreasonable estimates from the logistic regression models or prevented model convergence, I only include observations (results from simulation samples) with a VIF smaller than 20 in the logistic regression models.

I measure multicollinearity by computing the VIF for the predictor of interest ($\boldsymbol{x_1}$) in the regression model that is estimated in each sample. The VIF is computed using the proxy measures of the imperfectly measured variables rather than the true values of the predictors because applied researchers will only be able to compute a VIF for the variables they have access to, not for the true values of variables for which they are using proxies in a regression. The inclusion of a second variable that is not part of the true regression equation ($\boldsymbol{x_5}$) adds further noise to the VIF measure, but this is likely to mirror many applied settings in which researchers may include extra variables which may have no real impact on the dependent variable in order to minimize the risk of omitted variable bias. In sum, it is useful to examine how much a measure that applied researchers will be able to readily compute can tell them about the likelihood of measurement error causing serious problems in their models.

Figure 1 shows the simulation results for several sets of parameters, all of which produce small samples (50 observations). The figure depicts the predicted probably of a false positive (at the .05 level) based on the results of logistic regression models. The x-axis shows the level of multicollinearity, measures as the observed VIF for $\boldsymbol{x_1}$—the main predictor of interest (which has a coefficient of zero in the true regression equation). The different lines correspond to sets of samples that have different levels of measurement error for the observed values of the three variables ($\boldsymbol{x_2}$, $\boldsymbol{x_3}$, and $\boldsymbol{x_4}$) that have a true effect on the dependent variable. It is worth noting that random measurement error and multicollinearity are generally negatively related since random measurement error always serves to attenuate bivariate relationships in expectation. For this reason, high levels of observed multicollinearity should be rare in practice when data have high levels of random measurement error.

When there is no measurement error or when measurement error makes up only 1% of the variance in the proxy measures, the rate of false positives in small samples stays more or less constant at 5%, regardless of the level of multicollinearity. When measurement error increases to 5% of the variance in the proxy measures, a slight upward trend becomes apparent in the figure. When there is no multicollinearity (VIF=1), it appears that false positives occur only 5% of the time. But samples with higher VIFs for $\boldsymbol{x_1}$ more frequently yield regression results that incorrectly find that $\boldsymbol{x_1}$ is significant. This positive relationship appears to taper off around a VIF of 6, at which point false positives appear to occur at a rate of approximately 15% and higher levels of multicollinearity do not appear to increase this probability. A similar pattern is found when measurement error is at 10% except that rates of false positives are greater, with the maximum rate probably hitting over 20%. Rates of false positives are much greater when measurement errors make up 25% of the variation in the proxy measures. Even with no

multicollinearity, false positive rates appear to be greater than 15%. The false positive rate then hits a maximum just above 40% when the VIF is at 4. Such high rates of type I errors are alarming and suggest that high levels of measurement error (25% or greater) are problematic, even in small samples and even when there is no multicollinearity (although multicollinearity certainly worsens the problem).

Figure 2 shows results for parameter settings identical to figure 1 except that the sample size is changed to 500. The most obvious difference is that rates of false positives get much larger with large samples. Unlike many other statistical problems, concerns related to measurement error often are worse in large samples. Given measurement error of 1%, type I error rates appear to increase monotonically as the VIF increases, starting at a 5% rate (VIF=1) and reaching close to 20% (VIF=10). This suggests that in large samples, even very tiny amounts of measurement error can be problematic when multicollinearity is excessive. Measurement error of 5% raises the rate of false positives to about 15% when there is no multicollinearity and 50% at a VIF of 5, after which there is not a dramatic increase as multicollinearity increases. The pattern is similar but somewhat more severe when measurement error accounts for 10% of the variance in the proxy measures. At 25% measurement error, the rate of type I errors approaches 50% even when there is no multicollinearity present. This alarming rate casts serious doubt on whether meaningful hypothesis tests can be obtained from OLS regression when measures of key variables contain such large errors (in large samples).

In figure 3, I examine whether or not containment of measurement errors to a single key variable substantially lessens the risk of false positives. The solid line reports results for samples where all three true predictors contain measurement error (which comprises 10% of each of their variances) while the dashed line shows results for samples where only one variable ($\boldsymbol{x_2}$) contains

measurement error (again with measurement error of 10%). While rates of false positives are consistently lower when only one variable contains errors, the difference between the two lines is very slight. Rates of false positives are only marginally lower when measurement error is contained to a single predictor.

Finally, in figure 4 I use an alternative measure of multicollinearity. Rather than examining the VIF for $\boldsymbol{x}_1$, I examine type I error rates across varying values of the bivariate correlation between $\boldsymbol{x}_1$ and $\boldsymbol{w}_2$ (the proxy measure for $\boldsymbol{x}_2$) when $\boldsymbol{x}_2$ is the only variable measured with error. When only one variable contains error, one might think that the bivariate correlation with the imperfectly measured variable would be a better indicator of false positives than the VIF since the VIF is affected by correlations with other variables that are perfectly measured. Comparing to the dashed line in figure 3 (which uses the VIF to predict type II errors in the same set of simulation samples), one can see that the rate of false positives is much higher when the bivariate correlation is 0 (~50%) than when the VIF is 1 (~20%). This implies that a low bivariate correlation is not sufficient to conclude that multicollinearity is not exacerbating a measurement error bias; complex covariance patterns can cause bias in the estimates for $\boldsymbol{x}_1$ even if $\boldsymbol{x}_1$ is uncorrelated at a bivariate level with the variable measured with error ($\boldsymbol{w}_2$). In contrast, a very low VIF (closer to 1 than to 2) appears to be a reliable indicator that multicollinearity is not worsening a measurement error problem. A very strong bivariate correlation (.9) with the poorly measured variable does, however, appear to be a stronger indication of a near-absolute false positive than a high VIF (10).

Guidelines for Applied Researchers

What should applied researchers do if they find high levels of multicollinearity as they are running regressions? The answer may depend on whether results are significant or not. If

significant results are found, then multicollinearity may offer reason for pause. First, if there are nonlinear terms (squared or interaction terms) of substantive interest, researchers might want to consider adding other nonlinear terms to the regression equation. Omitting a necessary nonlinear term is particularly problematic if (1) two variables are being interacted and they are reasonably correlated with one another (.3 or greater in large samples) or (2) a variables that has been squared is reasonably correlated with another variable with which it could be interacted. The safest bet for avoiding false positives is to include an interaction term for the correlated variables as well as a squared term for each of the correlated variables. However, adding unnecessary variables reduces statistical power. If adding additional nonlinear terms dramatically widens standard errors and makes it difficult to draw inferences, a judgment must be made about how important it is to include the originally omitted terms.

One can choose to assume that one or more nonlinear terms does not affect the dependent variable, perhaps guided by theoretical reasoning suggesting the absence of a curvilinear relationship or of an interaction. Imposing a statistical assumption based on a theory means that the researcher is choosing not to test that aspect of the theory. For example, a researcher who decides to omit squared terms when testing an interaction between two moderately correlated variables is choosing to impose (rather than test) the theoretical assumption that direct effects are strictly linear. The researcher can still test whether the estimate for the interaction conforms to theoretical expectations given that another aspect of the theory is assumed to be true. The results will allow the researcher to say something like "if the direct effects are truly linear as the theory assumes, these results suggest that…" This limited test can still be a valuable contribution; no empirical test is perfect or can test all aspects of a complex theory. But a researcher who has enough data to obtain precise estimates for both squared and interaction terms can test two

theoretical claims simultaneously: they can determine whether direct effects appear to be truly linear and whether the interaction conforms to expectations.

A second question for researchers to ask if they find significant results with high levels of multicollinearity is whether there might be measurement error in one or more of their independent variables. Even if it is only one important control variable that is poorly measured, this can seriously bias estimates of the variables of interest, particularly if there are reasonably high levels of multicollinearity (VIF of 4 or greater). If measurement error is severe (measurement error accounts for 25% or more of the variance in the proxy measure), results may be untrustworthy even in the absence of multicollinearity. At lower levels of measurement error, however, problems are much less likely to arise when there is very little multicollinearity.

If measurement error is a concern, one potential solution is to guess what proportion of variance is due to measurement error for any variables that likely contain error. For simple linear regressions, measurement error models can then be easily run in standard statistical packages like Stata, and users can see how the results are affected. One can try inputting various levels of measurement error in order to see how sensitive a finding is to various guesses about the degree of measurement error for a variable or set of variables.

If multicollinearity is found and results are insignificant, two general ideas can guide researchers as they interpret results and ponder whether to run different models. Both are good general practices but are especially important when dealing with data with multicollinearity. First, null (insignificant) results should not be interpreted as evidence of no effect. Since multicollinearity generally produces wide confidence intervals (large standard errors), insignificant results often do not allow one to conclude that the true value of the coefficient is close to zero. Of course, one can simply look at the confidence intervals to see whether this is the

case for a particular set of results. Researchers who have some sense of the scale of their variables should be able to make intelligent statements about whether the confidence intervals rule out the possibility of a substantively meaningful impact. Another useful tool in some situations may be to run joint significance tests. These may allow one to make claims about a set of variables have an effect in concert even if individual coefficients are not significant. For example, suppose one has measures of ideology and partisanship which are highly correlated. If large standard errors make it impossible to draw meaningful substantive results from the individual coefficients, one might still be interested in whether or not the variables are jointly significant. Joint significance would indicate that partisanship or ideology or both are significant determinants of the dependent variable, which is not a terribly specific conclusion but may still be useful for some purposes. If one is wanting to draw conclusions about whether some outcome is determined primarily by broad political ideas or by preferences for specific policies, the joint significant test results may be adequate.

Second, multiple regression (including OLS) provides estimates of the independent effects of variables. In other words, a coefficient estimate for variable X indicates the expected change in the dependent variable if X changes and all other independent variables are held constant. If high levels of multicollinearity result from variables that are so intertwined with one another that it makes little sense to think about one changing while the other(s) is (are) held constant, the individual coefficient estimates will be difficult to interpret. In some cases, multicollinearity may result because one has included multiple measures of the same concept in a regression or because one has included mediating variables in a regression. Insignificant results due to multicollinearity can sometimes serve as a reminder to make sure that one really wants to

test for the effect of some predictor independent of another one with which it is highly correlated.

Works Cited

Achen, Christopher H. 1982. *Interpreting and Using Regression*. Newbury Park, CA: SAGE.

———. 1983. "Toward Theories of Data: The State of Political Methodology." In *Political Science: The State of the Discipline*, ed. Ada W. Finifter. Washington, DC: American Political Science Association, 69-93.

———. 1985. "Proxy Variables and Incorrect Signs on Regression Coefficients." *Political Methodology* 11(3-4): 299-316.

Buonaccorsi, John P. 2010. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: Chapman & Hall/CRC.

Can, Seda, Rens van de Schoot, and Joop Hox. Forthcoming. "Collinear Latent Variables in Multilevel Confirmatory Factor Analysis: A Comparison of Maximum Likelihood and Bayesian Estimations." *Educational and Psychological Measurement* doi: 10.1177/0013164414547959.

Cortina, Jose M. 1993. "Interaction, Nonlinearity, and Multicollinearity: Implications for Multiple Regression." *Journal of Management* 19(4): 915-922.

Dormann, Carsten F., Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J. Leitão, Tamara Münkemüller, Colin McClean, Patrick E. Osborne, Björn Reineking, Boris Schröder, Andrew K. Skidmore, Damaris Zurell, and Sven Lautenbach. 2013. "Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance." *Ecography* 36(1): 27-46.

Farrar, Donald E., and Robert R. Glauber. 1967. "Multicollinearity in Regression Analysis: The Problem Revisited." *Review of Economics and Statistics* 49(1): 92-107.

Freckleton, Robert P. 2011. “Dealing with Collinearity in Behavioural and Ecological Data: Model Averaging and the Problems of Measurement Error.” *Behavioral Ecology and Sociobiology* 65(1): 91-101.

Ganzach, Yoav. 1997. “Misleading Interaction and Curvilinear Terms.” *Psychological Methods* 2(3): 235-247.

———. 1998. “Nonlinearity, Multicollinearity and the Probability of Type II Error in Detecting Interaction.” *Journal of Management* 24(5): 615-622.

Graham, Michael H. 2003. “Confronting Multicollinearity in Ecological Multiple Regression.” *Ecology* 84(11): 2809-2815.

Greene, William H. 2012. *Econometric Analysis*. 7th Edition. Upper Saddle River, NJ: Prentice Hall.

Grewal, Rajdeep, Joseph A. Cote, and Hans Baumgartner. 2004. “Multicollinearity and Measurement Error in Structural Equation Models: Implications for Theory Testing.” *Marketing Science* 23(4): 519-529.

Gujarati, Damodar N., and Dawn C. Porter. 2009. *Basic Econometrics*. 5th Edition. New York: McGraw-Hill.

Hill, R. Carter, and Lee C. Adkins. 2001. “Collinearity.” In *A Companion to Theoretical Econometrics*, ed. Badi H. Baltagi. Malden, MA: Blackwell, 256-278.

Jackman, Simon. 2008. “Measurement.” In *The Oxford Handbook of Political Methodology*, eds. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. New York: Oxford University Press, 119-151.

MacCallum, Robert C., and Corinne M. Mar. 1995. “Distinguishing Between Moderator and Quadratic Effects in Multiple Regression.” *Psychological Bulletin* 118(3): 405-421.

McAdams, John. 1986. “Alternatives for Dealing with Errors in the Variables: An Example Using Panel Data.” *American Journal of Political Science* 30(1): 256-278.

O’Brien, Robert M. 2007. “A Caution Regarding Rules of Thumb for Variance Inflation Factors.” *Quality & Quantity* 41(5): 673-690.

Shieh, Yann-Yann, and Rachel T. Fouladi. 2003. “The Effect of Multicollinearity on Multilevel Modeling Parameter Estimates and Standard Errors.” *Educational and Psychological Measurement* 63(6): 951-985.

Silvey, S. D. 1969. “Multicollinearity and Imprecise Estimation.” *Journal of the Royal Statistical Society, Series B* 31(3): 539-552.

Wold, S., A. Ruhe, H. Wold, and W. J. Dunn, III. 1984. “The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses.” *SIAM Journal on Scientific and Statistical Computing* 5(3): 735-743.

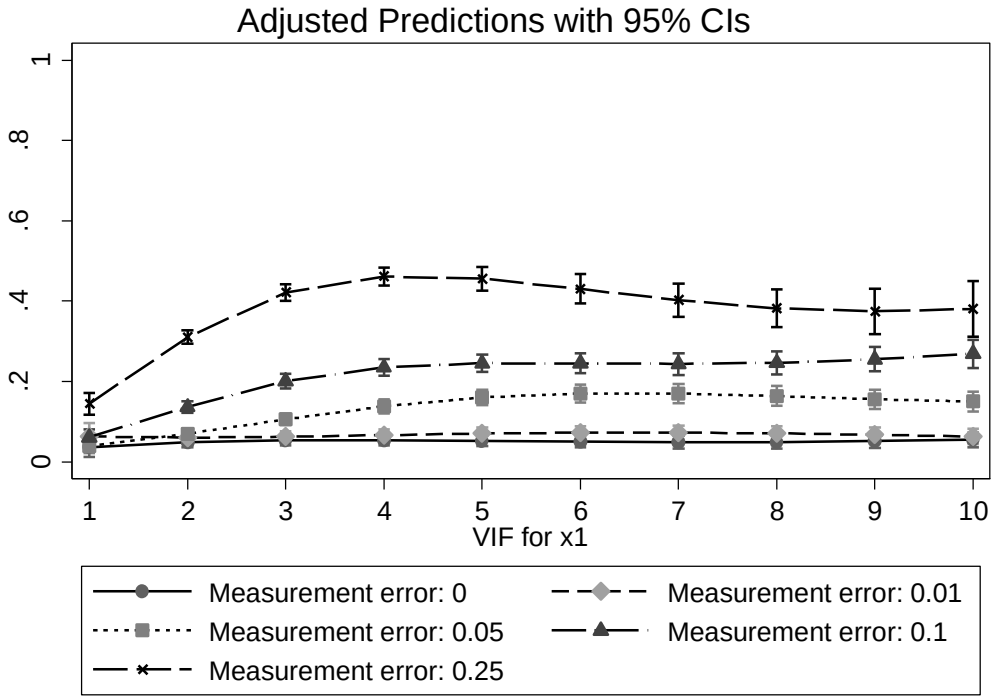
Wooldridge, Jeffrey M. 2013. *Introductory Econometrics: A Modern Approach*. 5th Edition. Mason, OH: South-Western.

Table 1. Simulations for Nonlinear Models

		True equation: quadratic curvilinear $y = x_1 + x_2 + {x_1}^2 + \varepsilon$			True equation: interactive $y = x_1 + x_2 + x_1x_2 + \varepsilon$		
		Misspecified model: $y = \beta_1x_1 + \beta_2x_2 + \beta_5x_1x_2 + \varepsilon$	Overspecified model: $y = \beta_1x_1 + \beta_2x_2 + \beta_3{x_1}^2 + \beta_4{x_2}^2 + \beta_5x_1x_2 + \varepsilon$		Misspecified model: $y = \beta_1x_1 + \beta_2x_2 + \beta_3{x_1}^2 + \varepsilon$	Overspecified model: $y = \beta_1x_1 + \beta_2x_2 + \beta_3{x_1}^2 + \beta_4{x_2}^2 + \beta_5x_1x_2 + \varepsilon$	
N	Corr(x_1, x_2)	β_5: false positives	β_5: false positives	β_3: positive & significant	β_3: false positives	β_3: false positives	β_5: positive & significant
50	0	24.8%	5.3%	99.9%	19.2%	4.9%	99.7%
50	.3	55.6%	4.7%	100.0%	49.9%	4.5%	98.5%
50	.6	94.5%	4.9%	98.7%	91.4%	5.0%	85.7%
50	.9	99.9%	4.7%	36.1%	99.8%	4.8%	13.4%
50	.95	99.9%	4.8%	13.3%	99.9%	4.9%	6.4%
500	0	29.2%	4.8%	100%	26.1%	5.0%	100%
500	.3	99.9%	5.0%	100%	99.7%	5.1%	100%
500	.6	100%	4.5%	100%	100%	5.0%	100%
500	.9	100%	5.0%	99.9%	100%	4.7%	86.5%
500	.95	100%	4.9%	85.5%	100%	5.1%	35.7%

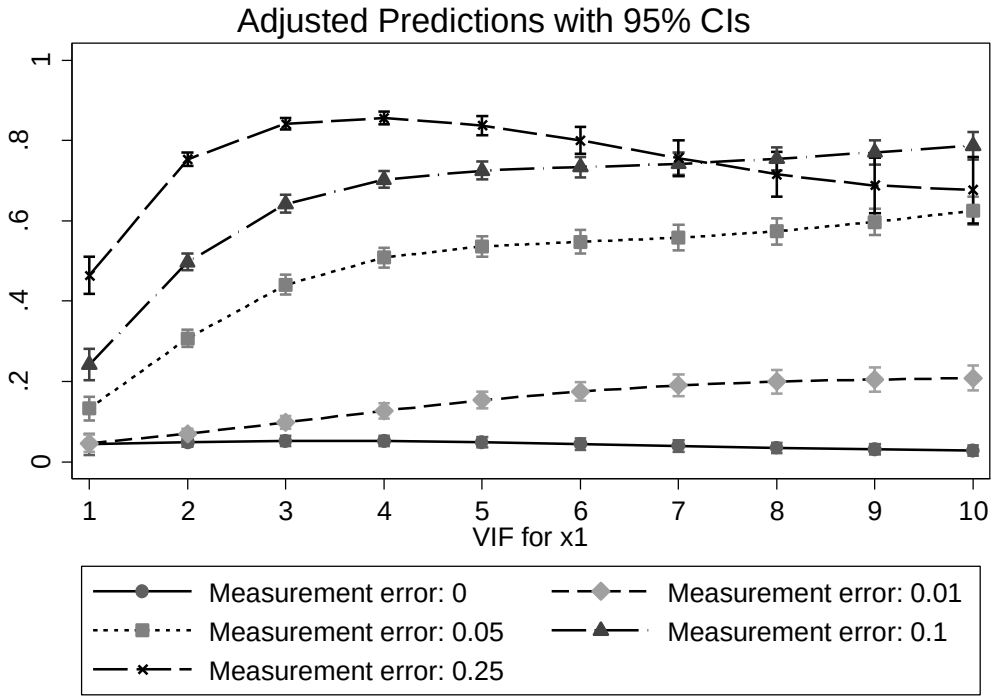
Notes:
5,000 datasets were randomly generated for each set of parameters, and all models estimated used OLS regression.
x_1, x_2, and ε were each independently drawn from a standard normal distribution (mean=0, variance=1).
Two-way tests of significance were conducted with an alpha level of .05 (null: $\beta_i = 0$).

Figure 1. Small Sample Simulations with Measurement Error in Controls



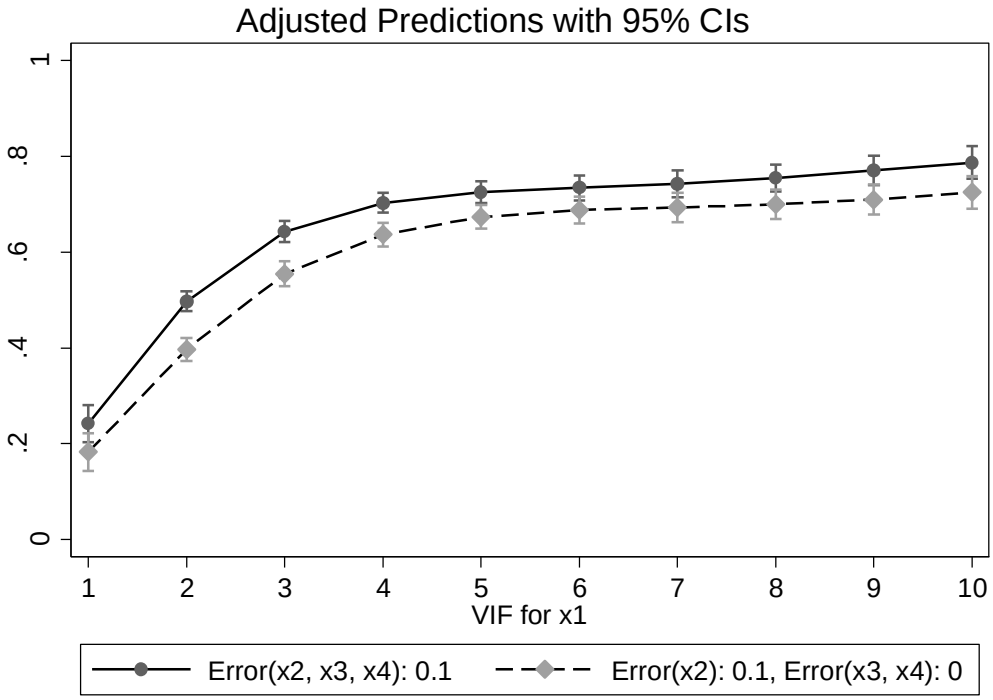
Measurement error listed is for x2, x3, and x4.
x1 and x5 have no measurement error.
N=50

Figure 2. Large Sample Simulations with Measurement Error in Controls



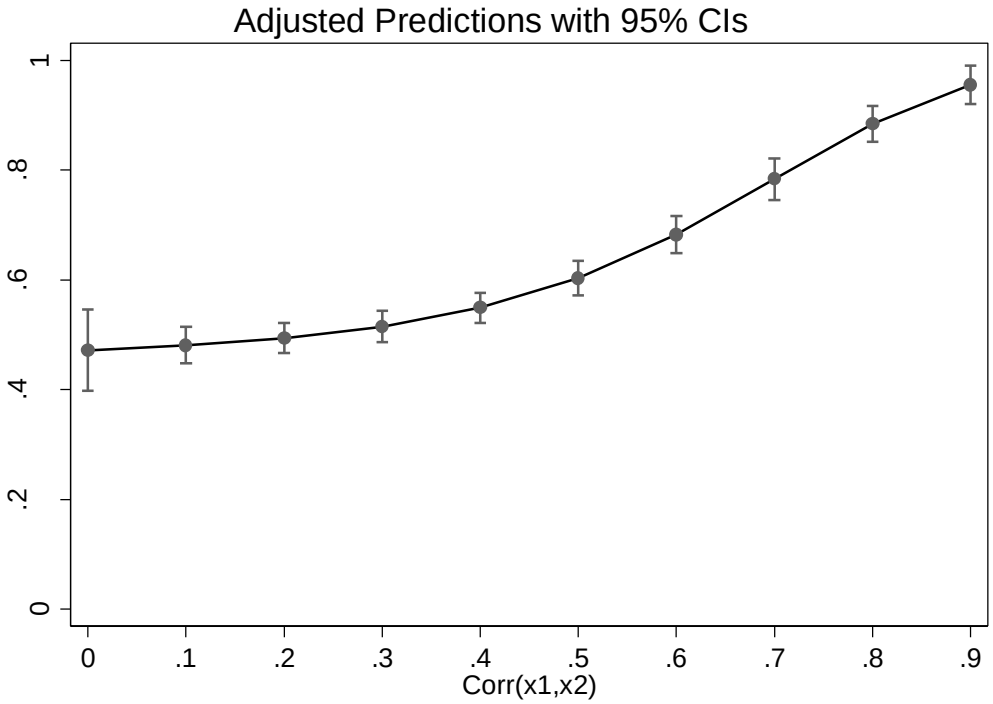
Measurement error listed is for x2, x3, and x4.
x1 and x5 have no measurement error.
N=500

Figure 3. Simulations with Measurement Error in One versus Multiple Controls



Error listed is measurement error.
x1 and x5 have no measurement error.
N=500

Figure 4. Simulations with Multicollinearity Measured Through Bivariate Correlation



Measurement error accounts for 0.1 of the variance in x2.
x1, x3, x4, and x5 have no measurement error.
N=500