A CORRECTION/UPDATE TO "WHEN IS IT JUSTIFIABLE TO IGNORE EXPLANATORY VARIABLE ENDOGENEITY IN A REGRESSION MODEL?"

RICHARD A. ASHLEY AND CHRISTOPHER F. PARMETER

Abstract. This note corrects an error – pointed out in Kiviet (2016) – in the Ashley and Parmeter (2015a) derivation of the asymptotic distribution of the OLS parameter estimator in the usual k-variate multiple regression model, but where some or all of the explanatory variables are endogenous. This sampling distribution lies at the heart of the Ashley and Parmeter (2015a) sensitivity analysis of a hypothesis test rejection p-value with respect to potential endogeneity in the explanatory variables in such regression models, so this correction is of practical importance. We also discuss the settings in which Kiviet's way of displaying univariate sensitivity analysis results is an improvement (and in what settings it is not), and we provide new analytic results for our sensitivity analysis in an important special case.

JEL Classification: C2, C15.

1. Introduction

We are indebted to Professor Kiviet for pointing out (Kiviet, 2016) a defect in our recent paper (Ashley and Parmeter, 2015a) proposing a sensitivity analysis for OLS estimation/inference in the presence of unmodeled endogeneity in the explanatory variables. If this sensitivity analysis shows that the inferences one most cares about are robust with respect to reasonable amounts of endogeneity, then one can dispense with a search for valid instruments. In contrast, where the sensitivity analysis indicates that these inference results are fragile in that regard, then such a search is mandatory. And, where that search yields instrumental variables inferences which are themselves similarly fragile (Ashley and Parmeter, 2015b) with respect to a reasonable degree of endogeneity in these instruments, then one might reasonably be motivated to make the additional assumptions about this degree of endogeneity necessary for the application of Bayesian methods – as in Caner and Morrill (2013) and Kraay (2012) – for 'robustifying' the inferential machinery. Thus, our algorithm is of practical, as well as intellectual, significance. Consequently – both practically and intellectually – it is important to get the analytics right.

In retrospect, we erred in obtaining our OLS endogeneity sensitivity analysis as a special case of the GMM/IV sensitivity analysis we developed in Ashley and Parmeter (2015b), taking the possibly-endogenous explanatory variables as instruments for themselves. In this circumstance the 2SLS estimator does reduce to the OLS estimator. However, it is known (Davidson and MacKinnon,

Date: July 16, 2018.
Key words and phrases. Robustness, Exogeneity, Instruments.
Richard Ashley, Corresponding Author, Department of Economics, Virginia Polytechnic Institute and State University; e-mail: ashleyr@vt.edu. Christopher Parmeter, Department of Economics, University of Miami; e-mail: cparmeter@bus.miami.edu.

1993; Kinal, 1980) that the asymptotic variance of 2SLS estimators is distorted in this just-identified case. This is what led to the discrepancy between the GMM/IV-based sampling distribution we provided for the general OLS case and the sampling distribution Kiviet derives for the special case of the coefficient on the single endogenous explanatory variable in a Gaussian bivariate regression model.[1] We also erred in not conducting a detailed enough literature review and acknowledging the previous works of Kiviet and Niemczyk (2007, 1012) and Kiviet (2013) on the sampling distribution of the OLS estimator in a Gaussian regression with an endogenous explanatory variable, which preceded our sensitivity analysis.

In Section 2 below we derive the correct asymptotic sampling distribution of the OLS multiple regression model parameter estimator, under the assumption that the k explanatory variates are endogenous (each with a specified covariance with the model error), and not assuming Gaussianity. We then show that this sampling distribution reduces to the Kiviet (2016) result for the special case of the bivariate Gaussian regression model he considers. The importance of this more general multivariate sampling distribution result is that it allows our analysis to address the sensitivity of a hypothesis test rejection p-value result to possible simultaneous endogeneity in more than one explanatory variable. This more general sampling distribution result also allows our analysis to address multiple (joint) linear restrictions and/or to address null hypotheses involving the coefficients on multiple explanatory variables, such as equality restrictions or (in a production function context) the existence of constant returns to scale.

In Section 3 we utilize this new sampling distribution in the kind of sensitivity analysis proposed in Ashley and Parmeter (2015a,b), showing how straightforward the required calculations are in general and deriving analytic results for the special - but not uncommon - case where endogeneity is considered possible in only one explanatory variable.

Section 4 applies these new results to correct the sensitivity analysis results on several inferences in the Mankiw, Romer and Weil (1992) estimated aggregate production function example used in Ashley and Parmeter (2015a). The revised results are not markedly different, but they provide a nice illustration for the present paper.

Finally, in Section 5 we discuss in what settings Kiviet's suggested way of displaying the sensitivity analysis results is an improvement over our approach and in what settings it is not.[2]

2. The OLS Multiple Regression Estimator With Endogenous Covariates

2.1. **Sampling Distribution of the OLS Estimator.** Assume that

$$Y = X\beta + \varepsilon, \tag{1}$$

where the matrix of regressors, X, is $n \times k$ with (for simplicity) zero mean and variance-covariance Σ_{XX}. Here $\ell \leq k$ of the explanatory variables may be endogenous, being linearly related to the

[1] More broadly, GMM/IV based inference is generally suspect in the exact identification case, not just in our sensitivity analysis.

[2] A comparison of our approach to other alternatives - e.g., Caner and Morrill (2013) - will appear elsewhere.

error term, with covariance $E(\frac{1}{n}X'\varepsilon)$ given by the vector λ. Then,

$$\varepsilon = X\Sigma_{XX}^{-1}\lambda + \nu, \tag{2}$$

where ν is iid with mean zero and variance σ_ν^2, with each component assumed to be independent of each element of X.[3]

Substituting (2) into (1) results in

$$\begin{aligned} Y =& X\beta + \varepsilon \\ =& X\beta + (X\Sigma_{XX}^{-1}\lambda + \nu) \\ =& X(\beta + \Sigma_{XX}^{-1}\lambda) + \nu. \end{aligned} \tag{3}$$

This is a regression model satisfying all of the usual assumptions necessary for OLS with stochastic regressors that are independent of the error term. As such, OLS applied to the regression model (1) produces, per Johnston (1972, Section 9-2):

$$\sqrt{n}\left(\widehat{\beta}^{OLS} - (\beta + \Sigma_{XX}^{-1}\lambda)\right) \xrightarrow{D} N(0, \sigma_\nu^2\Sigma_{XX}^{-1}). \tag{4}$$

In practical applications the fitting errors for Equation (3) – the regression model actually estimated – might well show indications of heteroscedasticity and/or serial correlation. In such cases one could simply replace $\sigma_\nu^2\Sigma_{XX}^{-1}$ (the sample estimate of $\sigma_\nu^2\Sigma_{XX}^{-1}$ in Equation (4), which is consistent under the assumption that $\nu \sim iid$, by the corresponding White-Eicker or Newey-West estimate.

2.2. **Relation to the Single-Regressor Result of Kiviet.** For the bivariate regression special case explicitly considered by Kiviet (2013, 2016), the $k \times k$ variance-covariance matrix Σ_{XX} reduces to the scalar σ_x^2 and the k-vector λ reduces to the scalar

$$\lambda = Cov(x_i, \varepsilon_i) = \sigma_{x\varepsilon} = \rho_{x\varepsilon}\sqrt{\sigma_x^2\sigma_\varepsilon^2}, \tag{5}$$

for all $i \in [1, \ldots, n]$, where x_i and ε_i each denote the vector's i^{th} component, and where $\rho_{x\varepsilon}$ is the (now scalar) correlation between the single explanatory variable and the original model error term, ε.

It follows from (2) that

$$\begin{aligned} \varepsilon'\varepsilon =& \left(X\Sigma_{XX}^{-1}\lambda + \nu\right)'\left(X\Sigma_{XX}^{-1}\lambda + \nu\right) \\ =& \lambda'\Sigma_{XX}^{-1}X'X\Sigma_{XX}^{-1}\lambda + 2\lambda'\Sigma_{XX}^{-1}X'\nu + \nu'\nu. \end{aligned} \tag{6}$$

Dividing both sides of (6) by n and taking expectations yields

$$\sigma_\varepsilon^2 = \lambda'\Sigma_{XX}^{-1}\lambda + \sigma_\nu^2 = \frac{\sigma_{x\varepsilon}^2}{\sigma_x^2} + \sigma_\nu^2 = \frac{\rho_{x\varepsilon}^2\sigma_x^2\sigma_\varepsilon^2}{\sigma_x^2} + \sigma_\nu^2, \tag{7}$$

[3]Note that $E(n^{-1}X'\varepsilon) = E\left[n^{-1}X'X\Sigma_{XX}^{-1}\lambda + n^{-1}X'\nu\right] = \lambda + E(n^{-1}X'\nu) = \lambda + 0$ verifies the form of (2) and that this derivation is tacitly assuming that any dependence of ε on X is linear.

since λ is the scalar $\sigma_{x\varepsilon} = \rho_{x\varepsilon}\sqrt{\sigma_x^2\sigma_\varepsilon^2}$ and Σ_{XX} is the scalar σ_x^2 in this section. Hence,

$$\sigma_\nu^2 = \left(1 - {\rho_{x\varepsilon}}^2\right)\sigma_\varepsilon^2. \tag{8}$$

This directly implies that

$$\sqrt{n}\left(\widehat{\beta}^{OLS} - \left(\beta + \frac{\sigma_{x\varepsilon}}{\sigma_x^2}\right)\right) \xrightarrow{D} N\left(0, \left(1 - \rho_{x\varepsilon}^2\right)\frac{\sigma_\varepsilon^2}{\sigma_x^2}\right). \tag{9}$$

as in Kiviet (2016, Equation 2.7) for the special case of a bivariate regression.

3. Our Proposed Sensitivity Analysis

As Friedman (1953) has famously noted, it is both necessary and appropriate to make assumptions – notably, even assumptions which we know to be false – in any successful economic modeling effort: the usefulness of a model, he asserted, inheres in the richness/quality of its predictions rather than in the accuracy of its assumptions. Our contribution here – and in Ashley and Parmeter (2015b), which addresses similar issues in the context of GMM/IV inference using possibly-flawed instruments – is to both point out and operationalize a general proposition that is a natural corollary to Friedman's assertion: It is perfectly all right to make possibly-false (and even very-likely-false) assumptions – *if and only if one can and does show that the model results one most cares about are insensitive to the levels of violations in these assumptions it is reasonable to expect.* The present context provides an ideal setting in which to both exhibit and operationalize the quantification of the "insensitivity" alluded to in this proposition, because this setting is so very simple. This setting is also attractive in that OLS estimation of multiple regression models with explanatory variables of suspect exogeneity is very common in applied economic work.

In implementing the sensitivity analysis proposed here, we presume that the regression model equation given above as Equation (3) has been estimated using OLS, so that the sample data (realizations of Y and X) are available, and have been used to obtain a sample realization of the inconsistent OLS parameter estimator ($\widehat{\beta}_{OLS}$) – which is actually consistent for $\beta + \Sigma_{XX}^{-1}\lambda$ – and to obtain a sample realization of the usual OLS estimator error variance estimator $s^2 = \widehat{\sigma}_\nu^2$, which is in fact a consistent estimator of the Equation (3) error variance, σ_ν^2.[4] In addition, the sample length (n) is taken to be sufficiently large that both $\widehat{\beta}_{OLS}$,and $\widehat{\sigma}_\nu^2$ have essentially converged to their probability limits; thus, in particular, n is sufficiently large that $\widehat{\sigma}_\nu^2$ need not be distinguished from σ_ν^2.

Now assume – for a moment – that λ, the k-dimension vector of covariances between the columns of the X matrix and the vector of errors (ε) in Equation (1) – the model couched in terms of the true parameter vector (β) – is given; this artificial assumption will be relaxed shortly.

[4]In principle one might want to instead denote these as y and x, so as to make explicit that these are the sample realizations of the random variables Y and X, but - for simplicity - that is not done here. Professor Kiviet has pointed out that one might want to explicitly account for the fact that X is in fact a stochastic matrix in the sensitivity analysis calculations. We concede that it might conceivably be fruitful to attempt an extension along these lines, but that is not done here: the sensitivity analysis considered here - just like the empirical economic analysis it is intended to augment - is taken to be conditional on the observed X matrix.

In that case the rejection p-value for any null hypothesis specifying a linear restriction on the components of the parameter vector β can be readily obtained, using the asymptotic sampling distribution given as Equation (4) above, leading to a test statistic distributed as Student's t with n - k degrees of freedom under this null hypothesis.[5]

A consistent estimator of β – call it $\widehat{\beta}_{consistent}$ - which, from Equations (3) and (4), is clearly just $\widehat{\beta}_{OLS} - \Sigma_{XX}^{-1}\lambda$ can then be easily obtained and substituted into Equation (1) to provide a set of model residuals asymptotically equivalent to the vector ε, the original model errors. The sample variance of this implied ε vector then yields $\widehat{\sigma}_{\varepsilon}^2$ a consistent estimate of σ_{ε}^2, the variance of the original model errors.[6]

This consistent estimate of the variance of ε is then combined with the posited λ covariance vector and with the (consistently estimated) sample variances of the k explanatory variables, $(\sigma_{X1}^2, \ldots, \sigma_{Xk}^2)$ to yield $\widehat{\rho}_{X\varepsilon}$, a consistent estimate of the corresponding k-vector of correlations between these explanatory variables and the original model errors (ε) in Equation (1). As with the other sample quantities, we will assume that n is sufficiently large that the sampling errors in $\widehat{\rho}_{X\varepsilon}$ can be neglected so that, for any posited λ covariance vector, the concomitant correlation vector $\rho_{X\varepsilon}$ can be readily calculated. This vector of correlations is worth estimating because it quantifies the endogeneity posited in each of the explanatory variables in a more intuitively interpretable way than does λ, the posited vector of covariances between the explanatory variables and the original model errors. For the purpose of our sensitivity analysis we denote the Euclidean length of this implied correlation k-vector $\rho_{X\varepsilon}$ below as "$|r|$".

In summary, then, any posited λ vector of covariances between the explanatory variables and ε, the original model errors, yields an implied (asymptotically valid) rejection p-value for the null hypothesis at issue, and a consistent estimate of the k correlations between the explanatory variables and the original model errors $\rho_{X\varepsilon}$, and a value for $|r|$, its Euclidean length.

The value of the λ vector is, of course, unknown, so this calculation is repeated for a selection M of all possible values it can take on, retaining the aforementioned correlation vector ($\rho_{X\varepsilon}$), its length ($|r|$), and the concomitant null hypothesis rejection p-value, and writing these values (for each chosen λ vector) to one row of a spreadsheet file if and only if the null hypothesis is no longer rejected at some designated p-value. (Here, for clarity of exposition only, this designated p-value will be specified as 0.05.) Because the regression model need not be re-estimated for each posited λ vector, these calculations are computationally inexpensive; consequently, it is quite feasible for

[5]The rejection p-value for testing a null hypothesis instead specifying a set of $q > 1$ linear restrictions, to be jointly tested, similarly leads to a test statistic distributed $F(q, n-k)$ and leads to a sensitivity analysis which is so similar that the exposition is couched here (solely for expositional clarity) in terms of a single linear restriction. For that matter, many econometrics programs - e.g., Stata - make it very easy to test a nonlinear restriction on the components of β; and the sensitivity analysis described here further extends to the (asymptotically valid) rejection p-values from such testing in a completely straightforward way.

[6]This variance σ_{ε}^2 exceeds σ_{ν}^2, the residual variance in the model as actually estimated using OLS, because the inconsistency in the OLS parameter estimate strips out of the estimated model errors the portions of ε which are - due to the assumed endogeneity - correlated with the columns of X. Note that $\sigma_{\varepsilon}^2 > \sigma_{\nu}^2$ also follows mathematically from Equation (7), since Σ_{XX} is positive definite.

M to range up to 10^5 or even 10^6. Thus, for $\ell \leq 2$ – i.e., where the exogeneity of at most two of the k explanatory variables is taken to be suspect – it is most reasonable repeat the calculations using a straightforward ℓ-dimensional grid-search over the reasonably-possible λ vectors; for larger values of ℓ it is still feasible (and, in practice, effective for this purpose) to use a Monte-Carlo search instead, as described in Ashley and Parmeter (2015b).

This algorithm yields a spreadsheet containing $M' < M$ rows each containing an implied correlation k-vector $\rho_{X\varepsilon}$, its Euclidean length $|r|$, and its implied null hypothesis p-value – in each case, by construction, exceeding the nominal value of 0.05. For a sufficiently large value of M' this collection of $\rho_{X\varepsilon}$ vectors well approximates an ℓ-dimensional set in the vector space of correlations which we denote as the "No Longer Rejecting" or "NLR" set: these are the X-column to ε correlations (exogeneity-assumption flaws) which are sufficient to overturn the 5%-significant null hypothesis rejection observed in the original OLS regression model. Sorting this spreadsheet on the correlation-vector length $|r|$ then yields the point in the NLR which is closest to the origin – i.e., the smallest $\rho_{X\varepsilon}$ vector which represents a flaw in the exogeneity assumptions sufficient to overturn the rejection of the null hypothesis of interest at the 5% level. This vector – which we denote r_{min} – and its length, $|r|_{min}$ then quantify the sensitivity of this particular null hypothesis inference to possible endogeneity in any of these ℓ explanatory variables in the original regression model.

The computational burden of the computations described above is not large – so that `R` and `Stata` code (available from the authors) is generally quite sufficient to the task – but it is instructive to obtain r_{min} analytically for the not-uncommon special case where $\ell = 1$, and just one – the m^{th}, say – of the k explanatory variables is being taken as possibly-endogenous. Restricting attention (solely for expositional clarity) to the null hypothesis which corresponds to the single linear restriction that $\beta_j = 0$, it is easy to characterize the two values of λ_m for which the null hypothesis rejection p-value equals α, as they must each satisfy the equation

$$\left| \frac{\widehat{\beta}_j - \left(\Sigma_{XX}^{-1}(j,m)\right) \lambda_m^*}{\sigma_\nu^2 \Sigma_{XX}^{-1}(j,m)} \right| = t_{\alpha/2}^c(n-k), \tag{10}$$

so that

$$\lambda_m^* = \frac{\widehat{\beta}_j \pm \sigma_\nu^2 \Sigma_{XX}^{-1}(j,m) t_{\alpha/2}^c(n-k)}{\Sigma_{XX}^{-1}(j,m)} \tag{11}$$

where $t_{\alpha/2}^c(n-k)$ is the $\alpha/2$ critical value for a Student's t distribution with $n-k$ degrees of freedom and $\Sigma_{XX}^{-1}(j,m)$ is the $(j,m)^{\text{th}}$ element of the $(X'X)^{-1}$ matrix.[7] The two values of λ_m^* given by Equation (11) then lead to two implied values for the m^{th} component of the implied correlation vector; r_{min} is then the one of these two vectors with the smallest magnitude, which magnitude is then $|r|_{min}$. Mathematically, there are clearly two solutions to Equation (10) because of the absolute value function. Intuitively, there are two solutions because increasing λ_m increases the

[7] Equations (10) and (11) generalize in an obvious way to a null hypothesis consisting of a linear restriction on β; the generalization to the case of d linear restrictions is similar, except that the statistic now involves the usual quadratic form expression and the critical point used is then $F_\alpha^c(d, n-k)$.

bias in the j^{th} component of $\widehat{\beta}_{OLS}$ at rate $-\Sigma_{XX}^{-1}(j,m)$. Thus – supposing that this component of $\widehat{\beta}_{OLS}$ is (for example) positive – then sufficiently changing λ_m in one direction can reduce the value of $\widehat{\beta}_{consistent}$ just enough so that it remains positive and is now just barely significant at the α significance level; but changing λ_m sufficiently more in this direction will reduce the value of $\widehat{\beta}_{consistent}$ enough so that it becomes *negative* and barely significant at the α significance level.

Thus, this vector r_{min} is practical to calculate for any multiple regression model for which we suspect that one (or a number) of the explanatory variables might be endogenous to some degree, and its length ($|r|_{min}$) objectively quantifies the sensitivity of the rejection p-value for any particular null hypothesis to such possible endogeneity.

But how, precisely, is one to interpret the value of this estimated quantity? Clearly, if $|r|_{min}$ is close to zero – e.g., less than 0.20, say – then only a fairly small amount of explanatory-variable endogeneity suffices to invalidate the original OLS-model rejection of this particular null hypothesis at the 5% level. One could characterize such an inference as "fragile" with respect to possible endogeneity problems, and one might not want to place much confidence in this null hypothesis rejection unless and until one is able to find credibly-valid instruments for the explanatory variables with regard to which inference is relatively fragile.[8] In contrast, a large value of $|r|_{min}$ – e.g., greater than 0.60, say – indicates that quite a large amount of explanatory-variable endogeneity is necessary in order to invalidate the original OLS-model rejection of the null hypothesis at the 5% level. One could characterize such an inference as "robust" with respect to possible endogeneity problems, and perhaps not worry overmuch about looking for valid instruments in this case. Notably, inference with respect to one important and interesting null hypothesis might be fragile (or robust) with respect to possible endogeneity in one set of explanatory variables, whereas inference on another key inference might be differently fragile (or robust) – and with respect to a different set of explanatory variables: the sensitivity analysis results sensibly depend on the inferential question raised.

But what about an intermediate estimated value of $|r|_{min}$? Such a result is indicative of an inference for which the issue of its sensitivity to possible endogeneity issues is still sensibly in doubt. Here the analysis again suggests that one should limit the degree of confidence placed in this null hypothesis rejection, unless and until one is able to find credibly-valid instruments for the explanatory variables which the sensitivity analysis indicates are potentially problematic. In this instance the sensitivity analysis has not clearly settled the fragility versus robustness issue, but at least it provides a quantification which is communicable to others, and which is objective in the sense that any analyst will will obtain the same $|r|_{min}$ result. This situation is analogous to the ordinary hypothesis-testing predicament when a null hypothesis is rejected with a p-value of, say, 0.08: whether or not to reject the null hypothesis is not clearly resolved based on such a result, but one has at least objectively quantified the weight of the evidence against the null hypothesis.

[8]Perfectly exogenous instruments are generally unavailable also, so it is useful to note at this point that Ashley and Parmeter (2015b) provides an analogous sensitivity analysis procedure allowing one to quantify the robustness (or fragility) of IV-based inference rejection p-values to likely flaws in the instruments used.

In summary, here we have generalized the Kiviet (2016) sampling distribution result (for the Gaussian bivariate regression OLS estimator), obtaining the asymptotic distribution of the OLS parameter estimator for the general (k-variate) multiple regression model with endogenous regressors, not assuming Gaussianity. This result allows us to analyze hypothesis test rejection p-value sensitivity in the usual multiple regression model to possible endogeneity in any combination of the explanatory variables, and for any specific null hypothesis restriction (or restrictions). Notably, we have obtained essentially analytic sensitivity analysis results for the special – but not particularly uncommon – situation where only one explanatory variable at a time is considered to be possibly-endogenous, but we note that the numerical calculations required for the general case are not in fact computationally burdensome.

In Section 4 below we illustrate the value of this approach with a brief update of the Ashley and Parmeter (2015) sensitivity analysis of several key inferences from the classic Mankiw, Romer and Weil (1992) study of the impact of human capital on economic growth, based on the corrected OLS parameter estimator sampling distribution obtained here.

4. Revised Sensitivity Analysis Results for the Mankiw, Romer, and Weil (1992) Study of the Impact of Human Capital on Economic Growth

Ashley and Parmeter (2015a) provided sensitivity analysis results for two hypotheses in the classic Mankiw, Romer, and Weil (MRW, 1992) study on economic growth: first, that human capital accumulation does impact growth, and second that their main regression model coefficients sum to zero.[9] Here these results are updated using the corrected sampling obtained in Section 2 as Equation (4).

For reasons of limited space the reader is referred to Ashley and Parmeter (2015a) – or MRW (1992) – for a more complete description of the MRW model. Here we will merely note that the MRW regression model dependent variable is real per capita GDP and that the three MRW explanatory variables are the logarithm of the number of years of schooling ("ln SCHOOL", their measure of human capital accumulation), the real investment rate ("ln I/GDP"), and a catch-all variable ("ln(n+g+δ)", capturing population growth, income growth, and depreciation).

Table 1 displays our revised sensitivity analysis results for both of the key MRW hypothesis tests considered in Ashley and Parmeter (2015a, Tables 1 and 2), the main change here – in addition to using the Equation (4) sampling distribution result – is that we now include sensitivity analysis results allowing for possible exogeneity flaws in all three explanatory variables simultaneously.[10] These results indicate that MRW's inference results are robust regardless of which variable(s) are presumed correlated with the error term. Even in the case for testing $H_o : \beta_{school} + \beta_{I/GDP} + \beta_{ng\delta} = 0$ when both $\ln I/GDP$ and ln SCHOOL are assumed to have correlation with ε, our r_{min}

[9]MRW (1992, page 421) explicitly indicates that this sum is to equal zero under the null hypothesis: the indication to the contrary in Ashley and Parmeter (2015a) was only a typographical error.

[10]We note that Kiviet's procedure cannot address scenarios in which either two or all three explanatory variables are considered to be possibly endogenous, because his sampling distribution result is restricted to the special case where $\ell = 1$.

vector is determined to be $(0.228, 0.021)$. Certainly the 0.021 on its own might suggest that just a small amount of correlation between schooling and the noise in the model would undermine the inferential outcome, however, this is contingent on having a substantial amount of correlation between $\ln I/GDP$ and the model error as well. Combined this makes the outcome of the inference reasonably robust, as $|r|_{min}$ is nearly 0.23.

TABLE 1. Sensitivity Analysis Results on $H_o : \beta_{school} = 0.0$ and $H_o : \beta_{school} + \beta_{I/GDP} + \beta_{ng\delta} = 0$ from Mankiw, Romer and Weil (1992). The value of $|r|_{min}$ is reported beneath each entry in brackets.

Variable	$\ln(n+g+\delta)$	$\ln I/GDP$	ln SCHOOL	$\ln I/GDP$ & ln SCHOOL	All Three
			H_o:$\beta_{school} = 0.0$		
Implied $\rho_{X\varepsilon}$ (r_{min})	0.933	-0.571	0.444	(-0.354, 0.197)	(-0.970, 0.152, 0.160)
	[0.933]	[0.571]	[0.444]	[0.405]	[0.995]
			H_o:$\beta_{school} + \beta_{I/GDP} + \beta_{ng\delta} = 0$		
Implied $\rho_{X\varepsilon}$ (r_{min})	0.198	0.367	0.7120	(0.228, 0.021)	(0.335, -0.847, -0.242)
	[0.198]	[0.367]	[0.712]	[0.229]	[0.942]

It is useful to consider what meaning a practitioner should attach to an $|r|_{min}$ stemming from an r_{min} vector with very unequal components. Suppose, solely for clarity of exposition, that $\ell = 2$ and that the two possibly endogenous covariates are denoted x_1 and x_2, so that the area of the NLR is a 2-dimensional region, with the implied x_1-to-error correlations plotted on the vertical axis and the x_2-to-error correlations plotted on the horizontal axis. Suppose, then, that r_{min} is a 2-vector with positive components $r_{min_{x_1}}$ and $r_{min_{x_2}}$ and with $r_{min_{x_1}}$ substantially larger than $r_{min_{x_2}}$. In this case it is safe to assume that the NLR boundary crosses the vertical (x_1-to-error correlation) axis at a substantially larger value than the value at which this NLR boundary crosses the horizontal (x_2-to-error correlation) axis. This suggests that the fact that $r_{min_{x_1}}$ substantially exceeds the $r_{min_{x_2}}$ does imply that the rejection p-value for the key inference is markedly less sensitive to possible endogeneity in x_1 than it is to possible endogeneity in x_2. This kind of $\ell = 2$ result would thus alert the user to assess the two individual sensitivity analyses – one with regard to only the x_1-to-error correlation and the other with regard only to the x_2-to-error correlation – which would more directly address this relative sensitivity issue.

5. Concluding Remarks

It is interesting to note that – due to the positive definiteness of Σ_{XX} – Equation (7) in our derivation of the sampling distribution of $\widehat{\beta}^{OLS}$ makes it plain that the sampling variance of the OLS estimator is necessarily reduced by any endogeneity in the regressors. Upon reflection, this result is intuitively appealing: the correlations between the regressors and the model error term allow these endogenous regressors to 'fit' some of the sample variation in the model errors by partially proxying for them. This 'proxying' leads to an estimated model which actually fits the data better, and hence yields more precise (albeit inconsistent) parameter estimators.

The computational implementation of the Ashley and Parmeter (2015a) sensitivity analysis is essentially unaffected by the new results reported here: the main change is that the corrected sampling distribution – Equation (4) above – replaces the analogous Equation (8) in Ashley and Parmeter (2015a) in the `R`/Stata software implementations of the procedure. The software now also implements the essentially analytic computation of r_{min} in the special case of $\ell = 1$, where only one explanatory variable is being analyzed as possibly-endogenous.

Kiviet (2106) also suggests a different way of exploiting the sampling distribution of the – in his analysis, necessarily single – OLS coefficient estimator, so as to quantify the way in which unmodeled endogeneity affects OLS inference. In particular, he suggests tabulating the 95% confidence interval for this coefficient versus a set of assumed values for the correlation between this single explanatory variable and the model error. Where it is a single confidence interval which is most meaningful to the analyst, Kiviet's approach for displaying the sensitivity analysis certainly has merit. But we note that, with ℓ potentially endogenous variables, such a table becomes $(\ell + 1)$-dimensional; this is extremely unwieldy for values of ℓ greater than one. In contrast, our approach computes and displays $|r|_{min}$ – the smallest length for a correlation vector relating the model errors to a chosen set of ℓ possibly-endogenous explanatory variables which is sufficiently large as to overturn one's most crucial inferential result. This length is still convenient to compute and interpret even when ℓ substantially exceeds one. Still, we can readily envision settings where the exogeneity of just a single explanatory variable is problematic and where one's interest is in a confidence interval rather than in a hypothesis test rejection p-value; consequently, we can see either one of these sensitivity result display approaches as potentially preferable.

Finally, we would like to again thank Professor Kiviet for helping us improve this sensitivity analysis method by alerting us to the incorrect derivation in Ashley and Parmeter (2015a): it is important to get this sort of thing right. Moreover, because of his effort we were able to not only correct our distributional result, but also obtain a deeper and more intuitive analysis; the new analytic results obtained here are an unanticipated bonus.

References

[1] Ashley, R., and C. F. Parmeter 2015a. When is it Justifiable to Ignore Explanatory Variable Endogeneity in a Regression Model?", *Economics Letters*, 137(1), 70-74.

[2] Ashley, R., and C. F. Parmeter 2015b. "Sensitivity Analysis for Inference in 2SLS Estimation with Possibly-Flawed Instruments," *Empirical Economics*, 49(4), 1153-1171.

[3] Caner, M. and Morrill, M. S., 2013. "Violation of Exogeneity: A Joint Test of Structural Parameters and Correlation," Working paper.

[4] Davidson, R. and MacKinnon, J. G. 1993. *Estimation and Inference in Econometrics*, Oxford University Press, New York New York.

[5] Friedman, M. 1953. *Essays in Positive Economics*, University of Chicago Press, Chicago, IL.

[6] Johnston, J. 1992. *Econometric Methods*, McGraw-Hill, New York New York.

[7] Kinal, T. W. 1980. "The existence of moments of k-class estimators," *Econometrica*, 48, 241-249.

[8] Kiviet, J. F. 2013. "Identification and inference in a simultaneous equation under alternative information sets and sampling schemes," *The Econometrics Journal*, 16, S24-S59.

[9] Kiviet, J. F. 2016. "When is it really justifiable to ignore explanatory variable endogeneity in a regression model?," *Economics Letters*, 145, 192-195.

[10] Kraay, A. 2012. "Instrumental Variables Regression with Uncertain Exclusion Restrictions: A Bayesian Approach," *Journal of Applied Econometrics*, 27, 108-128.

[11] Mankiw, N. G., Romer, D. and D. N. Weil, 1992. "A Contribution to the Empirics of Economic Growth," *The Quarterly Journal of Economics*, 107(2) pp. 407-437.