Unconventional Confidence Bands in the Literature on the Government Spending Multiplier

Ryan H. Murphy

In recent decades, vector autoregression, especially structural vector autoregression, has been used to study the size of the government spending multiplier (Blanchard and Perotti 2002; Fatás and Mihov 2001; Mountford and Uhlig 2009). Such methods are used in a significant proportion of empirical research designed to estimate the multiplier (see Ramey 2011a). Despite being published in respected journals and cited by prominent members of the profession, much of this literature does not use the conventional standard of statistical significance that economists are accustomed to in empirical research.

Results in the literature on the fiscal multiplier are typically communicated using a graph of the estimated impulse-response functions. For instance, the effect of government spending on output may be reported by reproducing a graph of an impulse-response function of a one-unit (generally, one percentage point or one standard error) change in government spending. The graph would show the percent change in output over time following the change in government spending. To report statistical significance, authors of these studies may then draw confidence bands around the impulse response function. Ostensibly, if zero lies outside the confidence band, it is statistically distinguishable from zero. But very frequently in this literature the confidence bands correspond to only one standard error. In other words, instead of representing what corresponds to rejecting the null hypothesis at a 90% level or 95% level, the confidence bands correspond to rejecting the null hypothesis at a 68% level. By conventional standards, this confidence band is insufficient for hypothesis testing. Not every useful empirical study must achieve

1. O’Neil Center for Global Markets and Freedom, Southern Methodist University, Dallas, TX 75275.
significance at the 95% level to be considered meaningful, of course, but a pattern of studies which do not use and reach the conventional benchmark is a cause for attention and perhaps concern. Statistical significance is not the only standard by which we should judge empirical research (Ziliak and McCloskey 2008). It is, however, a useful standard, and still an important one.

Here I examine papers in the fiscal multiplier literature which apply vector autoregression methods. Sixteen of the thirty-one papers identified use narrow, one-standard-error confidence bands to the exclusion of confidence bands corresponding to the conventional standard of 90% or 95% confidence. This practice will often not be clear to the reader of a paper unless its text is read rather carefully.

**Methodological preliminaries**

It is a mistake for economists to take statistical significance too seriously. Without understanding the limits of data availability, or by too stubbornly adhering to arbitrary rules, worthwhile research may never reach the audience it deserves. The mere fact that one regression fails to achieve a certain p-value does not mean that a result is worthless, even though that notion is taught all too frequently via the frequentist interpretation of statistics.

On the other hand, statistical significance is a useful standard by which to judge empirical research. In some ways statistical significance is an arbitrary construct, but it offers one potential way for applied economists to decide which empirical results they find most credible. It is not to be ignored, or perhaps worse, obscured. And while economists should not use $p < 0.05$ as a lazy litmus test, it remains an important point of reference. The onus is on the practitioners of a subfield whose favored econometric method does not pass the conventional standard to demonstrate why they should not be suspected of simply moving the goalposts.

In certain contexts, when a lack of robustness has been made especially obvious, economists will change their minds. An example of this is the work of William Easterly in casting doubt on the then-fashionable belief that foreign aid could encourage growth in developing nations when the developing nations adopt good economic policies. Craig Burnside and David Dollar (2000) had previously developed evidence across many specifications that such an interaction term was statistically significant and in the ‘correct’ direction. Easterly (2003) showed that this result was contingent on the definitions of aid and of good policy; when these definitions were adjusted, the statistical significance of the results disappeared. Easterly’s evidence led economists to be far more skeptical of the work of Burnside and Dollar, even though in most of Easterly’s specifications the point estimate of the
aid-policy interaction term remained the same in sign and magnitude. Several of Easterly’s specifications even resulted in t-statistics for the redefined aid-policy interaction greater than one, as high as 1.41 (2003, 41-44). Yet showing the fragility of the reported statistical significance made development economists more circumspect in their support of aid.

### Unconventional standard error bands in studies of the fiscal multiplier

The standard error bands shown are only 68% bands, based on bootstrap standard errors. Although this is common practice in the government spending literature, it has no theoretical justification. … Some have appealed to Sims and Zha (1999) for using 68% bands. However, there is no formal justification for this particular choice. It should be noted that most papers in the monetary literature use 95% error bands. (Ramey 2011b, 11, 11 n.11)

The numbers in brackets are the one standard error confidence bands from the bootstrap distribution of multipliers. (Bachmann and Sims 2012, 244)

The asterisks indicate significance within the one-standard-deviation bandwidth. (Burriel et al. 2010, 265)

As Valerie Ramey (2011b, 11) notes, the decision to use confidence bands of one standard error is an ad hoc departure from professional standards. There are sometimes, to be sure, good reasons to include more bands than only the ones corresponding to two standard errors. The inclusion of bands corresponding to the 68% level, in addition to 95% bands, may communicate the “shape” of the results better than the latter alone (Sims and Zha 1999).² This paper takes no issue with such a practice of using multiple bands, but the exclusion of those bands corresponding to conventional notions of statistical significance is problematic. Authors who employ the narrower bands should provide a rationale for doing so and make their use of unconventional confidence levels obvious to the casual reader. It is not cynical to expect that most readers of journal articles skim the

---

² “We argue that the conventional pointwise bands common in the literature should be supplemented with measures of shape uncertainty… [F]or characterizing likelihood shape, bands that correspond to 50% or 68% posterior probability are often more useful than 95% or 99% bands” (Sims and Zha 1999, 1113, 1118).
articles, reading only to the point that they believe they can understand the model or results. Economists are trained to perceive confidence bands as implicit hypothesis testing, and so there is a danger that unconventional confidence bands may project a false sense of the power of the result. Figure 1 and Figure 2 are two examples of graphs that at first glance may seem to convey statistical significance as conventionally understood; however, the confidence bands employed correspond to 68%, not the usual 90% or 95%.

Regrettably, some economists have further muddled the matter by using one standard error as the threshold for deploying the term “statistically significant.” Raffaela Giordano et al. (2007) do so, despite providing both one- and two-sigma bands graphically. Giordano and collaborators provide no rationale for this use of “statistically significant” save for the remark that the same was done “in most previous studies” (2007, 716). But the argument that providing the 68% bands is desirable for certain technical reasons, such as the capacity to convey notions of shape with greater precision, does not tell us that the 68% level is also an acceptable standard for hypothesis testing. If one wishes to argue that the two-sigma confidence level is too high of a standard to apply to one’s empirical research, one should offer a compelling reason why the field of study in question can be held to a lower standard of significance or explain why economists in other fields should reevaluate their methods.

**Literature review and results**

To investigate the prevalence of these issues, I studied 31 papers which applied some variation of vector autoregression in measuring the size of the multiplier. These papers varied from unpublished manuscripts to publications in the *Quarterly Journal of Economics*. A similar exercise could be performed for the tax multiplier. I consider only the debate on the size of the government spending multiplier because the debate over that is most in vogue. Also, the most prominent line of contemporary research on the tax multiplier does not emphasize vector autoregression.

---

3. Giordano et al. (2007) cite Sims and Zha (1999) for the idea that “error bands corresponding to 0.50 or 0.68 probability…provide a more precise estimate of the true coverage probability” (Giordano et al. 2007, 716 n.8).

4. These papers were collected from work from an in-progress larger literature review which attempts to review all existing empirical literature on the size of the government spending multiplier. The thirty papers are those which employ vector autoregression as described in this paper. One additional paper was identified by a referee.

5. A recent and important paper in the literature on the tax multiplier, Romer and Romer (2010), includes a vector autoregression that uses the one-standard-error confidence bands.
A table of the 31 papers can be found in Appendix A, giving the number of papers citing them (according to Google Scholar) and whether the result was published in a top-100 journal (according to RePEc’s “simple” journal rankings). Such metadata allows me to evaluate whether, say, only the less prominent papers use the unconventional confidence bands. Full bibliographic information on the 31 papers is given in Appendix B.

Sixteen of the thirty-one papers use the unconventionally narrow confidence bands at the exclusion of other confidence bands. Papers which provide the 90% or 95% confidence bands, even once in the paper, are not included among the sixteen. Five of the six papers with at least 500 citations use the unconventionally narrow bands. As for papers with between 100 and 500 citations, there are eleven, and four of these used the unconventionally narrow bands. Looking at papers with less than 100 citations, eight of fourteen used the unconventionally narrow bands. Regarding journal placement, thirteen of the thirty-one papers were published in a top-100 journal. Of those thirteen, six use the unconventionally narrow bands and seven do not. Of the eighteen papers not published in a top-100 journal, ten use the unconventionally narrow bands and eight do not.

A narrative consistent with these facts is that the papers on the frontier of empirical macroeconomic methodology used the unconventionally narrow bands and others followed. In the middle (still respectably cited) tier of the literature, the one-standard-error practice is less pervasive. Finally, there is no strong evidence that the top journals are playing an important role in limiting the use of these narrower bands. It seems fair to say that the usage is pervasive within the literature and that those who employ the practice are following the most cited papers in the field.

**Final remarks**

Further research may bring even further into question whether this literature is held to the same standard as other areas of applied econometric analysis. As Ramey (2011a) has shown, the disagreement among macroeconomists today is whether the government spending multiplier is less than 1 or greater than 1, not whether it is greater than zero. That being the case, an important hypothesis to test is whether a multiplier point estimate is statistically different than 1, not zero, but that test is rarely performed. Where it has become common practice to lower the threshold to one standard error, it is hard to imagine point estimates are commonly

---

Another obfuscating issue is the lack of standardization in reporting the results of an impulse-response function, as quite often papers will merely report the first period (‘impact’) multiplier or the value of the impulse-response function at its highest point (‘peak’). Beyond the narrow issue of confidence bands, statistical significance generally is an issue with which the literature struggles.

Proponents of vector autoregression should be more transparent about its limitations. While in some ways it may be uniquely useful for macroeconomic topics (Sims 2010), its inability to pass conventional standards is demonstrative of its low power. When a headline point estimate is not statistically significant as conventionally defined, that fact should be made clear, not obscured. Otherwise, we could be holding this method to a lower standard than we hold competing methods. If that is the case, then other empirical methods tending to have smaller standard errors and tighter confidence intervals/bands, such as those pioneered by Valerie Ramey and Robert Barro (see, e.g., Barro and Redlick 2011), or those using clever research designs (e.g., Wilson 2012; Shoag 2013), should be held in higher relative esteem than they are presently.

Figure 1. Reproduction of “Figure 1” in Jordi Galí et al. (2007, 232)

Figure 1. The dynamic effects of a government spending shock.

Note: Estimated impulse responses to a government spending shock in the large VAR. Sample Period 1954:I-2003:IV. The horizontal axis represents quarters after the shock. Confidence intervals correspond ± 1 standard deviations of empirical distributions, based on 1,000 Monte Carlo replications. The right bottom panel plots the point estimates of both consumption (solid line) and disposable income (dashed line).
**Figure 2.** Reproduction of “Figure V” in Olivier Blanchard and Roberto Perotti (2002, 1348)

**Figure V**

Response to a Spending Shock
Appendix A.
Use of unconventionally narrow confidence bands in the literature on the fiscal multiplier

<table>
<thead>
<tr>
<th>Paper</th>
<th>Use 1 s.e. confidence bands</th>
<th>Google Scholar citations</th>
<th>Top-100 Journal (RePEc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bachmann and Sims (2012)</td>
<td>Yes</td>
<td>50</td>
<td>Yes</td>
</tr>
<tr>
<td>Beetsma and Giuliodori (2011)</td>
<td>No</td>
<td>65</td>
<td>Yes</td>
</tr>
<tr>
<td>Beetsma et al. (2008)</td>
<td>No</td>
<td>128</td>
<td>Yes</td>
</tr>
<tr>
<td>Benérix and Lane (2009)</td>
<td>Yes</td>
<td>28</td>
<td>No</td>
</tr>
<tr>
<td>Blanchard and Perotti (2002)</td>
<td>Yes</td>
<td>1999</td>
<td>Yes</td>
</tr>
<tr>
<td>Burnside et al. (2004)</td>
<td>No</td>
<td>372</td>
<td>Yes</td>
</tr>
<tr>
<td>Burriel et al. (2010)</td>
<td>Yes</td>
<td>50</td>
<td>No</td>
</tr>
<tr>
<td>Caldara and Kamps (2008)</td>
<td>Yes</td>
<td>102</td>
<td>No</td>
</tr>
<tr>
<td>Cimadomo and Bénassy-Quéré (2012)</td>
<td>No</td>
<td>3</td>
<td>No</td>
</tr>
<tr>
<td>Corsetti and Müller (2006)</td>
<td>Yes</td>
<td>183</td>
<td>Yes</td>
</tr>
<tr>
<td>Corsetti et al. (2012)</td>
<td>Yes</td>
<td>130</td>
<td>Yes</td>
</tr>
<tr>
<td>De Castro (2006)</td>
<td>Yes</td>
<td>62</td>
<td>No</td>
</tr>
<tr>
<td>Fatás and Mihov (2001)</td>
<td>Yes</td>
<td>529</td>
<td>No</td>
</tr>
<tr>
<td>Fazzari et al. (2013)</td>
<td>No</td>
<td>33</td>
<td>No</td>
</tr>
<tr>
<td>Gali et al. (2007)</td>
<td>Yes</td>
<td>1050</td>
<td>Yes</td>
</tr>
<tr>
<td>Giordano et al. (2007)</td>
<td>No</td>
<td>151</td>
<td>No</td>
</tr>
<tr>
<td>Ilzetzki et al. (2013)</td>
<td>No</td>
<td>309</td>
<td>Yes</td>
</tr>
<tr>
<td>Kirchner et al. (2010)</td>
<td>Yes</td>
<td>65</td>
<td>No</td>
</tr>
<tr>
<td>Kuttner and Posen (2002)</td>
<td>No</td>
<td>67</td>
<td>No</td>
</tr>
<tr>
<td>Monacelli et al. (2010)</td>
<td>No</td>
<td>103</td>
<td>Yes</td>
</tr>
<tr>
<td>Mountford and Uhlig (2009)</td>
<td>Yes</td>
<td>933</td>
<td>Yes</td>
</tr>
<tr>
<td>Pereira (2008)</td>
<td>Yes</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>Pereira and Lopes (2010)</td>
<td>Yes</td>
<td>22</td>
<td>No</td>
</tr>
<tr>
<td>Perotti (2005)</td>
<td>Yes</td>
<td>927</td>
<td>No</td>
</tr>
<tr>
<td>Perotti (2007)</td>
<td>Yes</td>
<td>330</td>
<td>No</td>
</tr>
<tr>
<td>Ramey (2011b)</td>
<td>No</td>
<td>733</td>
<td>Yes</td>
</tr>
<tr>
<td>Ramey (2012)</td>
<td>No</td>
<td>33</td>
<td>No</td>
</tr>
<tr>
<td>Ravnik et al. (2007)</td>
<td>No</td>
<td>132</td>
<td>No</td>
</tr>
<tr>
<td>Ravnik and Žilić (2011)</td>
<td>No</td>
<td>7</td>
<td>No</td>
</tr>
<tr>
<td>Stevans and Sessions (2010)</td>
<td>No</td>
<td>1</td>
<td>No</td>
</tr>
</tbody>
</table>
Appendix B.
Bibliographical information for the papers listed in Appendix A


References


About the Author

Ryan Murphy is a Research Associate at the O’Neil Center for Global Markets and Freedom at the Southern Methodist University Cox School of Business. He received his Ph.D. in economics from Suffolk University in 2013. His email address is rhmurphy@smu.edu.