Appendix: Reassessing the Public Goods Theory of Alliances

October 9, 2019

This appendix contains supporting materials for the tests of Hypothesis 1 and 2. The first section includes material for the test of Hypothesis 1. The second section describes priors, convergence diagnostics and results from simulated data for the multilevel model I used to test Hypothesis 2.

1 Other Estimates of Panel Data Interaction

1.1 Alternative Estimators

Robust regression is appropriate for residuals from the percentage changes in spending variable. Several observations of states during war see gigantic increases in spending— the largest value is 140, relative to a median of .063. This generates extremely heavy-tailed residuals, so OLS is inefficient. I do not transform the outcome for either test in the paper, but do so here as part of the robustness checks.

Even after applying the inverse hyperbolic sine (IHS) transformation, the residuals deviate strongly from normality, as Figure 1 shows. I use the IHS transformation because it accommodates positive, negative and zero values. Figure 1 shows the residuals from an OLS model with a transformation outcome, which is part of the robustness checks.

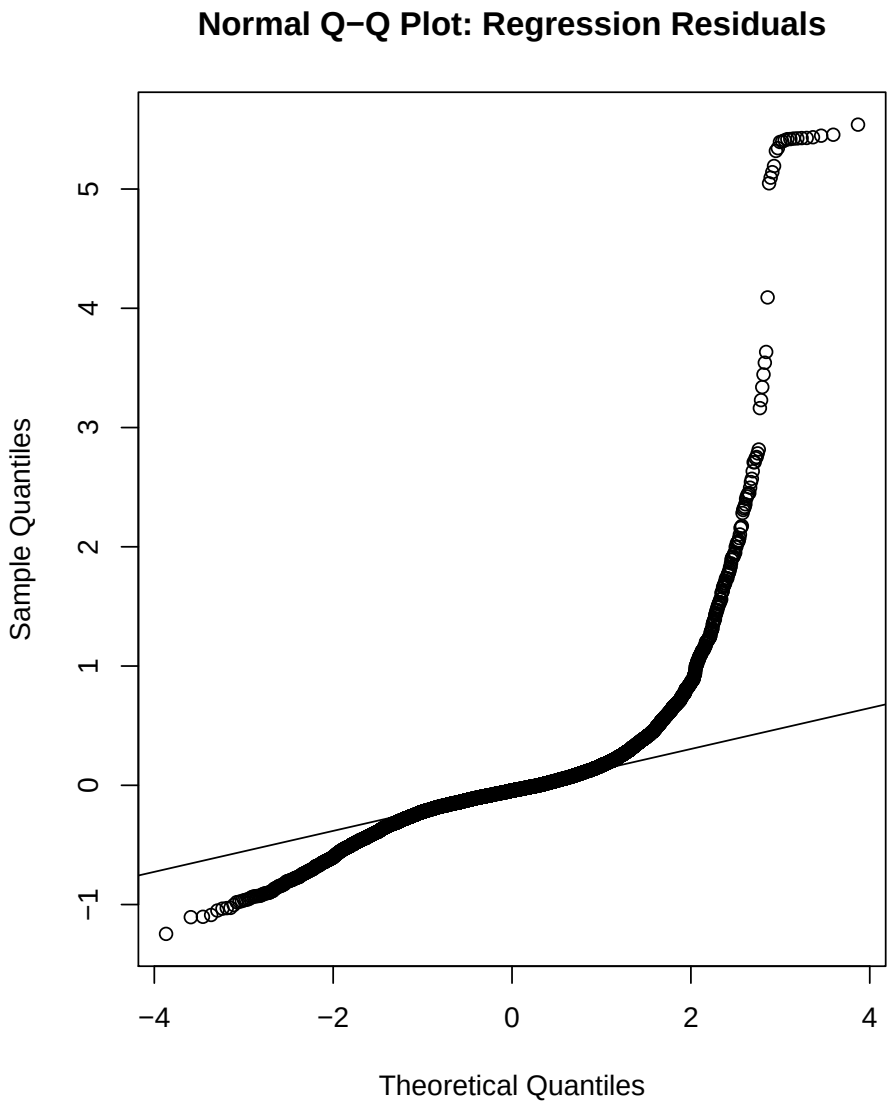

Figure 1: Plot of residuals against normal quantiles. Deviations from the straight line are deviations from the normal distribution.

I estimate three different models for the robustness check: OLS models with and without the inverse hyperbolic sine transformation, followed by robust regression with transformed spending. Rather than tabulate the coefficients, I plot the marginal effect of allied spending across the range of GDP for all three models. Figure 2 compares these results with the marginal effects plot from the robust regression in the paper. Transforming percentage changes in military spending has little effect on the robust regression results because even after transforming the dependent variable, the robust estimator heavily down-weights the most unusual observations.

The OLS estimates reverse the patterns in the robust regression, and the scale of the estimates is sensitive to whether the variable is transformed. The marginal effect for smaller states is statistically significant in the OLS estimates, but it is decreasing in GDP. For small states, expanding allied capability *increases* percentage changes in military spending, and the effect of allied capability is lower in small states. This is the opposite of Hypothesis 1, but it should be treated with extreme caution. Without the inverse hyperbolic sine transformation, the predicted marginal effect of changes in allied spending is implausibly large relative to the scale of the outcome. OLS may be too sensitive to the unusual observations in military spending data. In any case, the robust regression estimates in the manuscript are the best case for Hypothesis 1.

1.2 Continuous Modifying Variable

Hainmueller, Mummolo and Xu (2019) show linearity assumptions and a lack of support in the range of the modifying variable can generate misleading inferences in interactive models. They suggest estimating interactions with binning and kernel estimators to check for non-linearity and adequate support. Figure 3 plots the results of the binning estimator. Like the OLS margins plots above, these results may be sensitive to unusual observations.

The binning estimator indicates potential deviations from linearity. This pattern matches the OLS estimates, which are contrary to Hypothesis 1. Given the possible non-linearity, the kernel regression is a useful check.

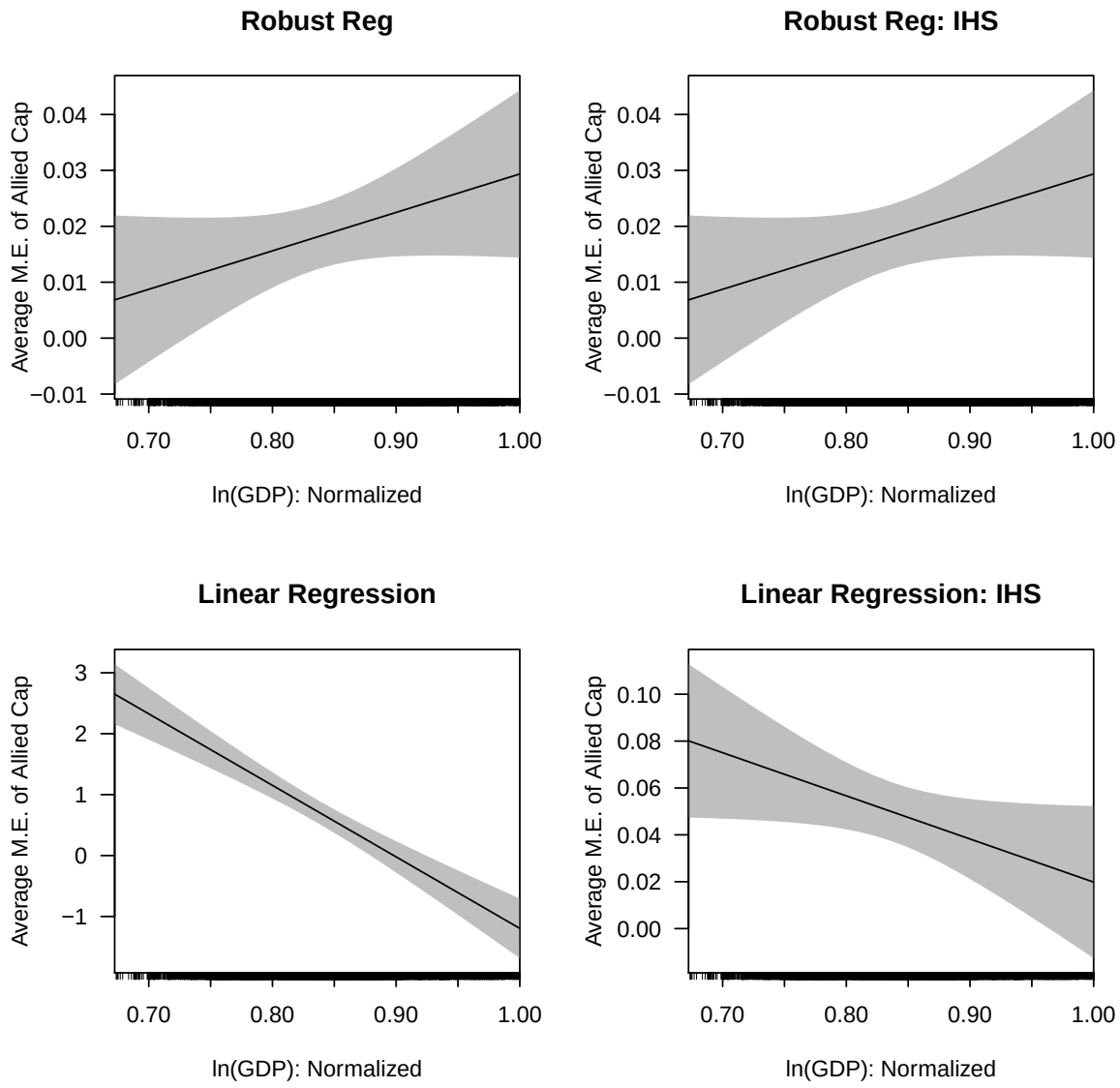

Figure 2: Comparison of marginal effect of changing allied spending on percentage changes in military spending across the range of GDP. Each plot corresponds to an estimation strategy. The top left plot is the marginal effects plot from the manuscript.

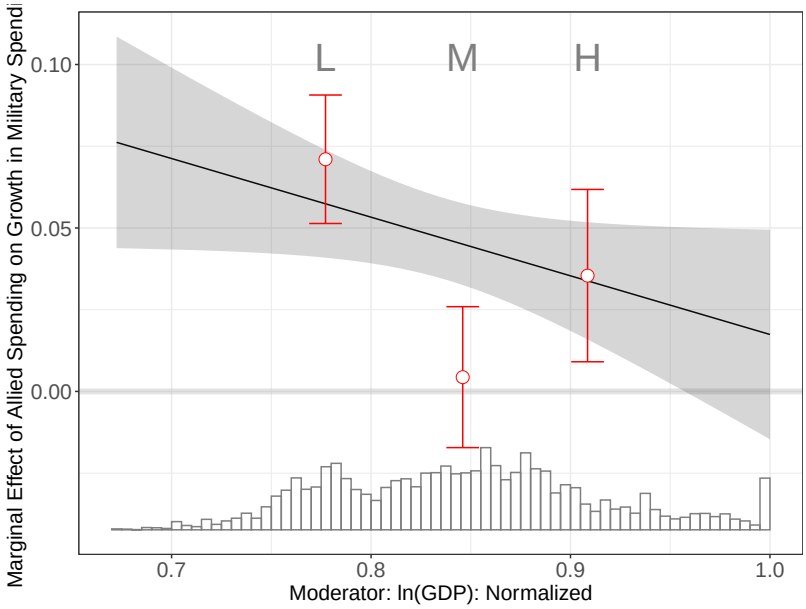

Figure 3: Binning estimates of interaction between changes in allied spending and GDP.

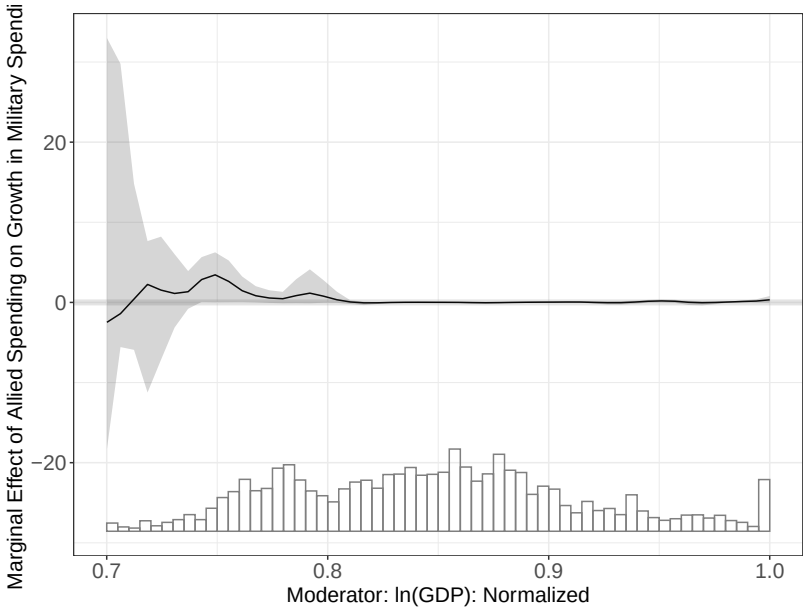

Figure 4: Kernel estimate of interaction between changes in allied spending and GDP.

The kernel estimator generates less evidence of a conditional relationship. As Figure 4 demonstrates, the marginal effect of spending is highly uncertain for small states, and close to zero otherwise. Though it is difficult to see from the plot, both positive and negative point estimates of the marginal effect are indistinguishable from zero. Allowing non-linear changes in the marginal effect of allied capability still shows little evidence of a conditional relationship, let alone the expectations of Hypothesis 1.

2 Multilevel Model

This section describes the priors on the multilevel model, convergence diagnostics for the Hamiltonian Monte Carlo, and results from running the same model on a sample of only states with at least one alliance.

2.1 Priors

All priors are specified to be weakly informative relative to the scale of the data (Gelman, Simpson and Betancourt, 2017). I summarize the prior distributions for each set of parameters in Table 1. $p(\nu)$ is a well-behaved prior for the degrees of freedom in a t-distribution (Juárez and Steel, 2010). Given that the median percentage change in military expenditures is 0.06, the priors are quite diffuse.

To facilitate estimation, I use a non-centered parameterization for the state and year varying intercepts, as well as the γ parameters (Betancourt and Girolami, 2015). A non-centered parameterization decouples the mean and variance to express an equivalent prior, which makes sampling easier. I also employ a sparse matrix representation of the alliance membership matrix $\mathbf{Z}$ to speed up estimation.

$$
\begin{gathered}
p(\alpha) \sim N(0,1) \\
p(\sigma) \sim \text{half-}N(0,1) \\
p(\alpha^{yr}) \sim N(0,\sigma^{yr}) \\
p(\sigma^{yr}) \sim N(0,1) \\
p(\alpha^{st}) \sim N(0,\sigma^{st}) \\
p(\sigma^{st}) \sim \text{half-}N(0,1) \\
p(\gamma) \sim N(\theta,\sigma^{all}) \\
p(\theta) \sim N(0,.5) \\
p(\sigma^{all}) \sim \text{half-}N(0,1) \\
p(\beta) \sim N(0,1) \\
p(\nu) \sim gamma(2,0.1)
\end{gathered}
$$

Table 1: Summary of Priors in Multilevel Model

2.2 Convergence

There were no divergent iterations in sampling. However, there are other threats to inference from the posterior samples. Given heavy tails in percentage changes of military spending, STAN might have struggled to explore the posterior distribution.

Energy plots can diagnose this problem. Figure 5 plots the marginal energy distribution and the first differenced distribution. If the two histograms do not overlap, sampling was impeded by heavy tails. The substantial overlap in the distributions for all four chains in Figure 5 indicates this was not a problem.

The split $\hat{R}$ statistic is another way to assess convergence. $\hat{R}$ compares the behavior of each chain by measuring the ratio of the average variance of draws within each chain to the variance of the pooled draws across chains. When $\hat{R}$ is close to 1, all the chains have similar variance, and are therefore in equilibrium.

The standard heuristic is that an $\hat{R}$ greater than 1.1 is problematic. Figure 6 plots the $\hat{R}$ statistic for every parameter in the model. No parameters generate concern, even at a more conservative threshold of 1.05.

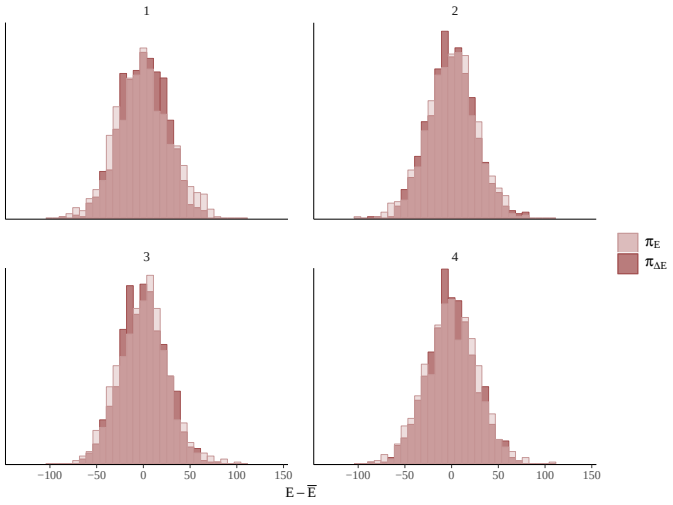

Figure 5: Energy plot of multilevel model results. Greater overlap in the two histograms indicates adequate exploration of the posterior distribution.

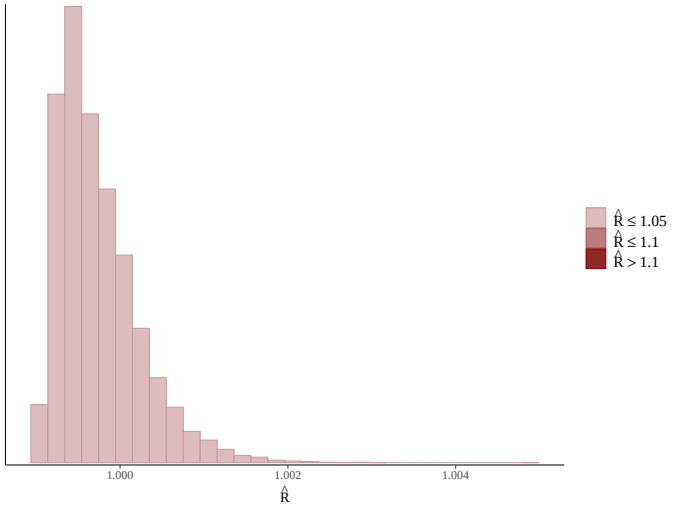

Figure 6: Histogram of split $\hat{R}$ statistic for all parameters in the multilevel model.

2.2.1 Inferences from Simulated Data

To assess if the model gives reasonable answers, I simulated data and associated parameters, then re-estimated the model on the simulated data. The model is a good fit if the credible intervals contain the known parameter values for the simulated data. This process checks whether the model can recover parameters from a known data-generating process that matches the model.

I simulate a hypothetical dataset with 5000 observations of 50 states observed over 200 years. There are 200 alliances in this data, and 2 state-level control variables. The hypothetical outcome is drawn from a Cauchy distribution with mean 0 and a scale of .25, which is more heavy-tailed than even my observed data.

I then simulate 2,000 draws of the outcome using the generated quantities block in STAN. The next step is selecting one of those draws of the outcome— which includes the value of the outcome for each observation and the associated parameter values. I select the 12th draw from the posterior and check whether after estimating the model on these data, the credible intervals include zero.

I focus on inferences about the γ, θ and σ_{all} parameters, because these are essential to testing the public goods argument. As Figure 7 and Figure 8 show, the posteriors accurately capture the known values of the hyper-parameters θ and σ_{all}. In these figures, the true parameter value is marked with a thick black line, while the light gray shaded area shows the 90% credible interval.

Because graphical presentation of the 200 γ parameters is more difficult, I calculated whether the credible interval contained the known parameter. 184 of the 200 intervals include the "true" γ value, which is a 92% success rate. Given the number of parameters and potential simulation variance, such accuracy is tolerable. Simulating data and recovering known parameters shows that the model estimates are reasonable approximations of the data-generating process.

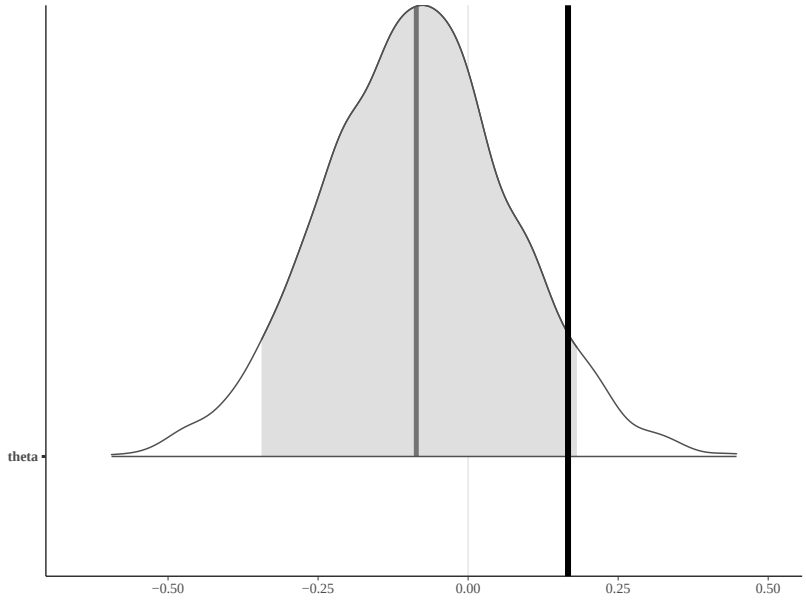

Figure 7: Posterior estimates and known parameter value for the alliance hyperparameter θ. The dark gray bar marks the posterior mean, while the shaded area captures the 90% credible interval. The black line marks the known, "true" θ value, which falls within the 90% interval.

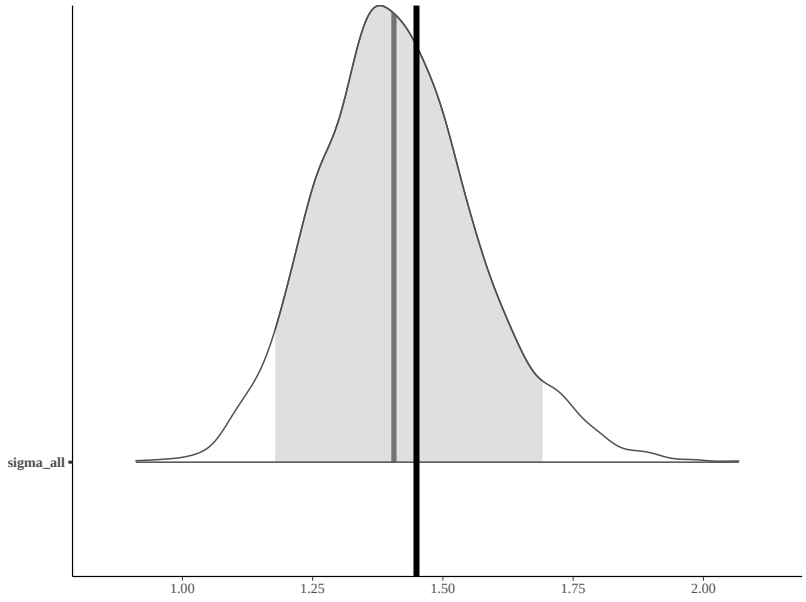

Figure 8: Posterior estimates and known parameter value for the alliance hyperparameter σ_{all}. The dark gray bar marks the posterior mean, while the shaded area captures the 90% credible interval. The black line marks the known, "true" σ_{all} value, which falls within the 90% interval.

2.3 Alternative Sample

It is possible that estimating a model on the full sample of states makes misleading comparisons by including states with no alliances. This adds many zeros to the membership matrix **Z**. To check whether inferences are sensitive to including states with no alliance participation, I re-fit the multilevel model on a sample of only states with at least one alliance. This reduced the sample size from 9,961 observations to 5,222, but the results are relatively unchanged.

All 285 treaties have a negative mean γ estimate, and none have a 90% credible interval that excludes zero. The overall mean θ is more negative in this sample and the γ estimates remain tightly clustered around that mean. Thus as Figure 9 shows, the distribution of the association between treaty contribution and spending across alliances is more negative in the alliance members sample. This figure overlays the distribution of the mean γ parameters in each sample.

Therefore, inferences about the impact of treaty contribution on percentage changes in military spending are unchanged if the sample is restricted only to alliance members. Increasing a state's share of total allied GDP leads to a more negative effect on treaty spending in expectation. I am still unable to identify any reliably positive γ parameters, which contradicts the free-riding hypothesis.

References

Betancourt, Michael and Mark Girolami. 2015. Hamiltonian Monte Carlo for hierarchical models. In *Current Trends in Bayesian Methodology with Applications*, ed. Satyanshu K. Upadhyay, Umesh Singh, Dipak K. Dey and Appaia Loganathan. Chapman and Hall/CRC Press pp. 79–102.

Gelman, Andrew, Daniel Simpson and Michael Betancourt. 2017. "The prior can generally only be understood in the context of the likelihood." *arXiv preprint arXiv:1708.07487* .

Hainmueller, Jens, Jonathan Mummolo and Yiqing Xu. 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models?: Simple Tools to Improve Empirical Practice." *Political Analysis* .

Juárez, Miguel A and Mark FJ Steel. 2010. "Model-Based Clustering of Non-Gaussian Panel Data Based on Skew-t Distributions." *Journal of Business & Economic Statistics* 28(1):52–66.

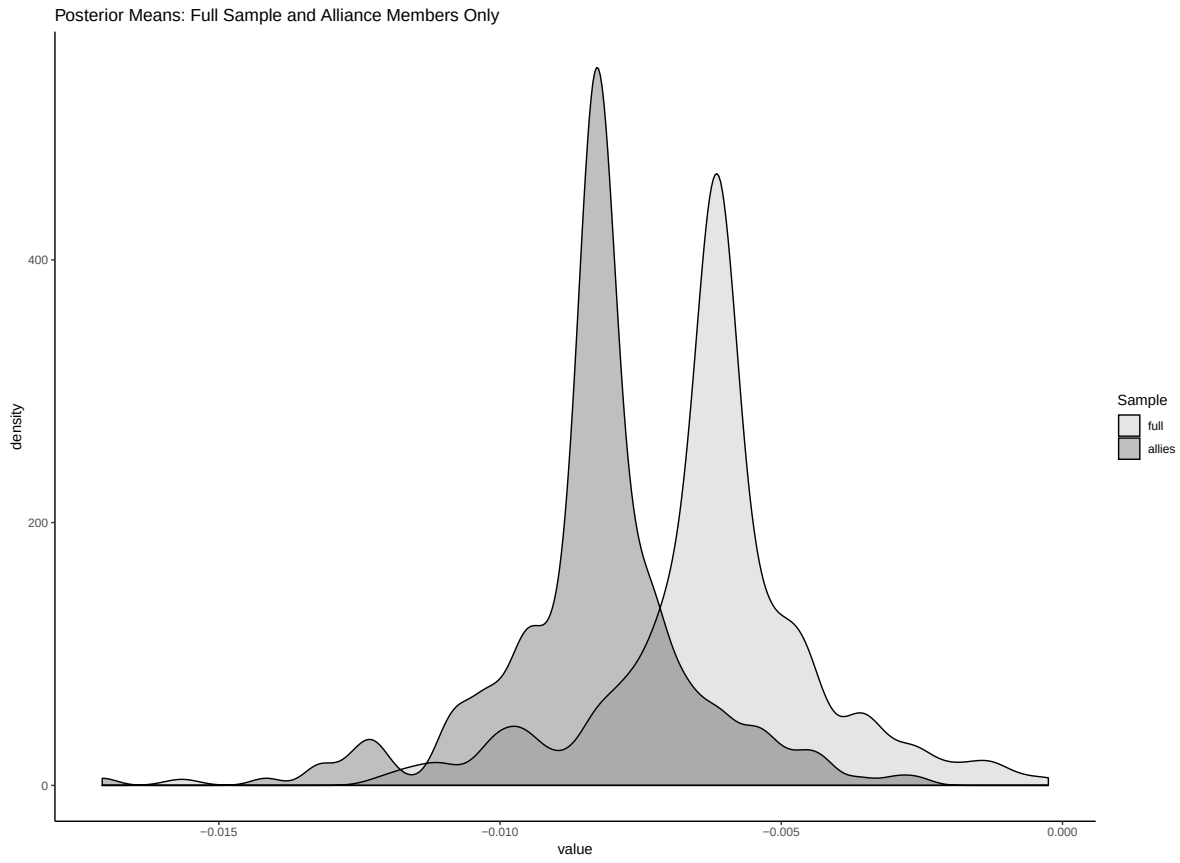

Figure 9: Comparison of the distribution of posterior mean γ parameters in full sample and a sample of only alliance participants. The darker gray distribution is the mean of the γ parameters in the sample of alliance members. The impact of treaty contribution on percentage changes in military spending is more negative in the alliance-members only sample.