Appendix: Reassessing the Public Goods Theory of Alliances

June 6, 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

I estimate three different models for the robustness check: OLS models with and without the inverse hyperbolic sine transformation of growth in spending as well as robust regression with transformed spending. Rather than interpret the coefficients, I plot the marginal effect of allied spending across the range of GDP for all three models. Figure 1 compares these results with the marginal effects plot from the robust regression in the paper. Three of the four estimation strategies produce a positive marginal effect, along with broad enough confidence intervals to suggest a conditional relationship with GDP is unlikely.[1] OLS without a transformed DV finds no positive impact of changing allied capability on growth in spending, but this result should be taken

[1] The robust regression results are so similar because even after transforming spending growth, the robust estimator heavily down-weights the most unusual observations.

with great caution due to poor model fit. Regardless, none of the results matches the expectation of Hypothesis 1— a negative effect which increases in state size.

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 2 plots the results of the binning estimator.

The binning estimator indicates that the linearity assumptions of the robust regression are acceptable. There is still no evidence state size modifies the impact of allied spending on growth in military spending, however. The marginal effect of allied spending cannot be distinguished from zero across the range of GDP.

The kernel estimator also generates little evidence of a conditional relationship. As Figure 3 demonstrates, the marginal effect of spending is null. 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 removes the positive marginal effect, it still shows little evidence of a conditional relationship, let alone the expectations of Hypothesis 1.

1.3 Robust Regression with Random Effects

In the test of Hypothesis 1, the robust regression estimator ignores clustering of observations within groups of states and years. Koller (2016) develops a robust regression estimator with random effects, which I fit as a robustness check. Table 1 summarizes the results.

Robust regression with random effects produces similar inferences about the presence of a conditional relationship between GDP and changing allied capability. The sign, magnitude and statistical significance of all the coefficients are similar to results from other estimators. Therefore,

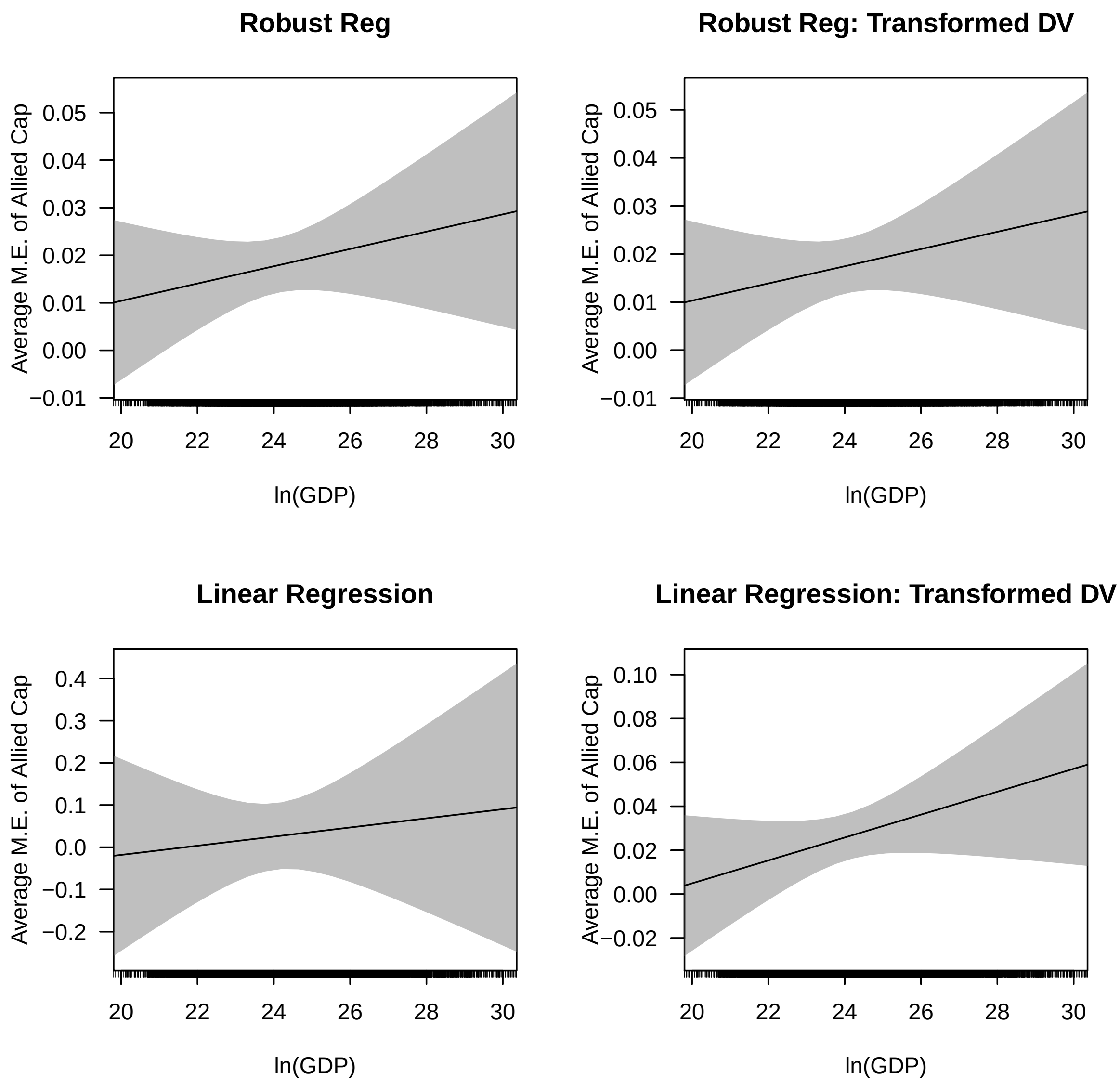

Figure 1: Comparison of marginal effect of changing allied spending on growth in military spending across the range of GDP. Each plot corresponds to an estimation strategy.

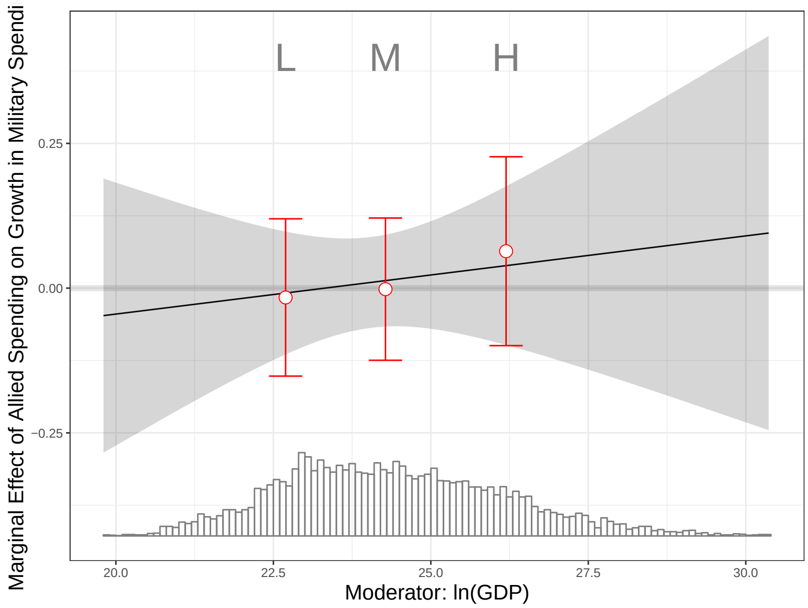

Figure 2: Binning estimates of interaction between changes in allied spending and GDP.

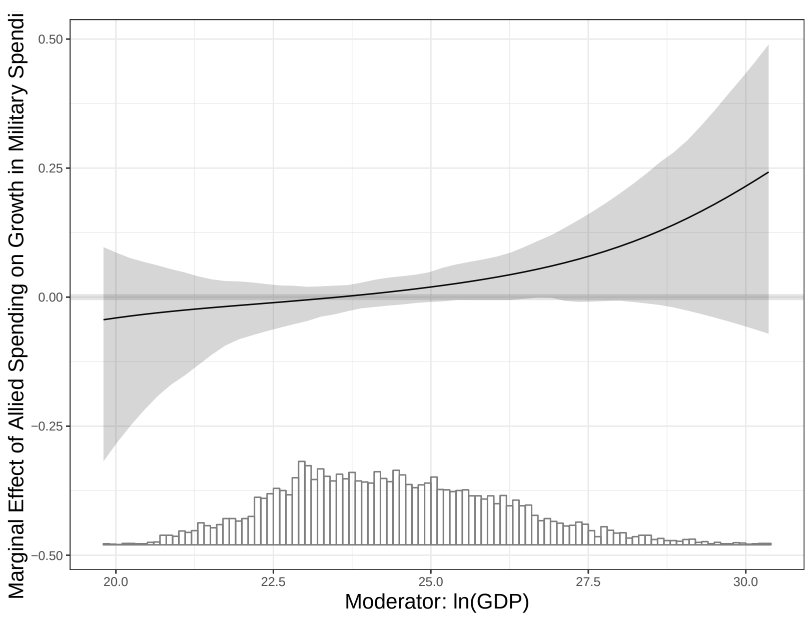

Figure 3: Kernel estimate of interaction between changes in allied spending and GDP.

	Estimate	Std. Error	t value
Change Allied Spending	-0.015	0.048	-0.310
ln(GDP)	-0.002	0.001	-1.414
Change Allied Spending $\times$ ln(GDP)	0.001	0.002	0.661
Avg. Alliance Size	-0.0001	0.0002	-0.401
Avg Alliance Democracy	-0.001	0.001	-1.354
International War	0.096	0.011	8.889
Civil War Participant	0.010	0.008	1.299
Polity	0.0002	0.0004	0.657
External Threat	0.016	0.013	1.217
Cold War	0.060	0.011	5.242
Constant	0.075	0.031	2.436

Table 1: Results from Robust Regression with State and Year random effects.

while it is difficult to use R output from Koller's robust regression estimator to plot the interaction, results from this estimator should mirror the above plots. Accounting for clustering of observations does little to change my inferences about 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 2. $p(\nu)$ is a well-behaved prior for the degrees of freedom in a t-distribution (Juárez and Steel, 2010). Given that median growth 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

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

Table 2: Summary of Priors in Multilevel Model

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 **Z** to speed up estimation.

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 military spending growth, STAN might have struggled to explore the posterior distribution.

Energy plots can diagnose this problem. Figure 4 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 4 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.

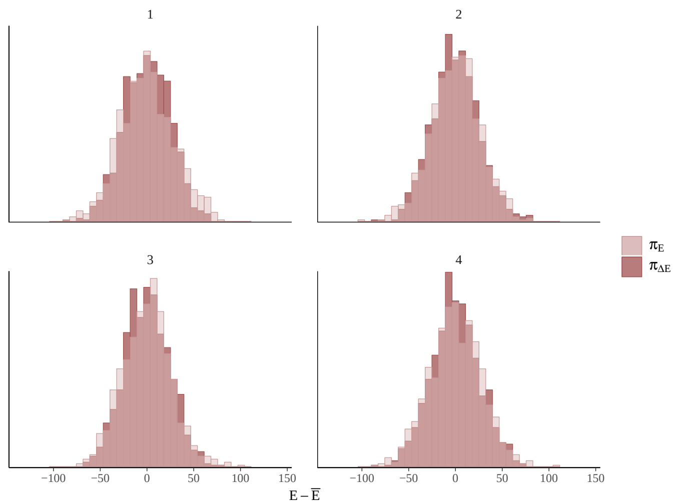

Figure 4: Energy plot of multilevel model results. Greater overlap in the two histograms indicates adequate exploration of the posterior distribution.

The standard heuristic is that an $\hat{R}$ greater than 1.1 is problematic. Figure 5 plots the $\hat{R}$ statistic for every parameter in the model. No parameters generate concern, even at a more conservative threshold of 1.05.

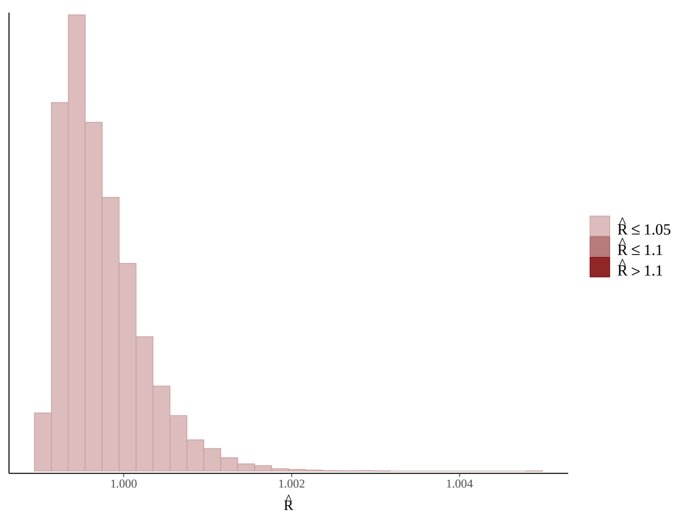

Figure 5: Histogram of split $\hat{R}$ statistic for all parameters in the multilevel model.

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 growth 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 6 and Figure 7 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.

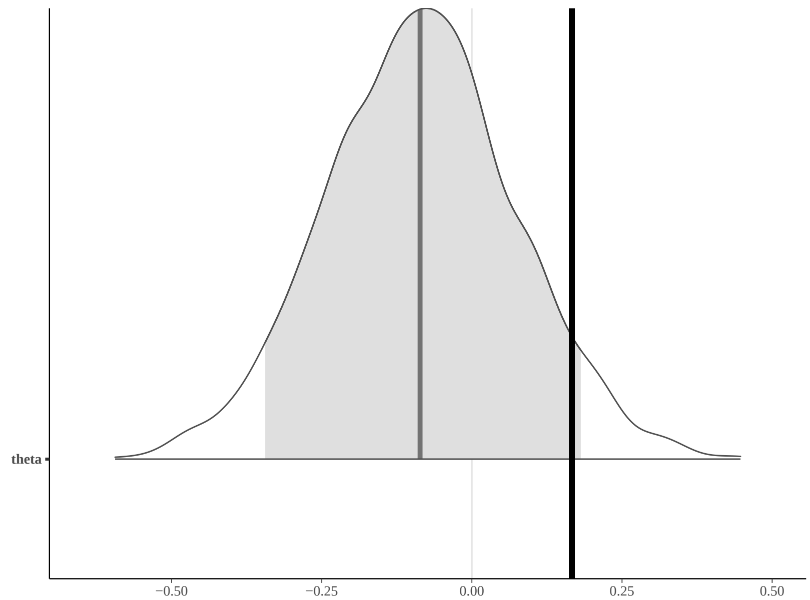

Figure 6: Posterior estimates and known parameter value for the alliance hyperparameter θ. The dark gray bar marks the posterior mean, while the shaded area captures the 90% credible interval. The black line marks the known, “true” θ value, which falls within the 90% interval.

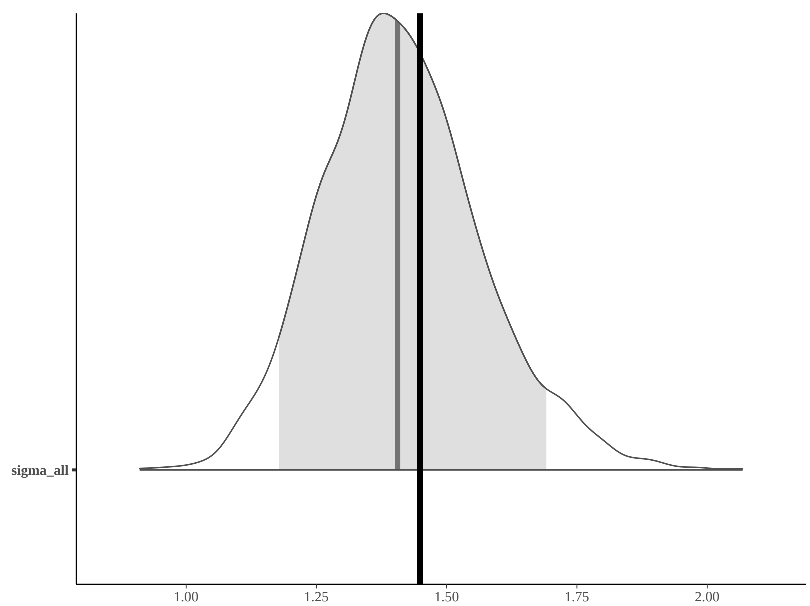

Figure 7: Posterior estimates and known parameter value for the alliance hyperparameter σ_{all}. The dark gray bar marks the posterior mean, while the shaded area captures the 90% credible interval. The black line marks the known, “true” σ_{all} value, which falls within the 90% interval.

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.

Koller, Manuel. 2016. "robustlmm: an R package for robust estimation of linear mixed-effects models." *Journal of Statistical Software* 75(6):1–24.