Appendix: Alliance Participation, Treaty Depth, and Military Spending

This online appendix provides more detail about the multilevel model and some checks of the results.

Priors

Table 1 summarizes the prior distributions in the multilevel model. All priors are weakly informative relative to the scale of the data. ν is the degrees of freedom for the t-distribution, and the gamma prior is the recommended default prior for STAN.

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

Table 1: Summary of Priors in Multilevel Model

Hamiltonian Monte Carlo Diagnostics

There were no divergent iterations in either sample running 4 chains for 2,000 iterations with 1,000 warmup iterations. The $\hat{R}$ is less than 1.1 for all parameters in both samples. Trace plots in Figure 1 indicate good mixing of the chains for the alliance-level parameters. Taken together, all of this implies that the chains adequately explored the posterior distribution.

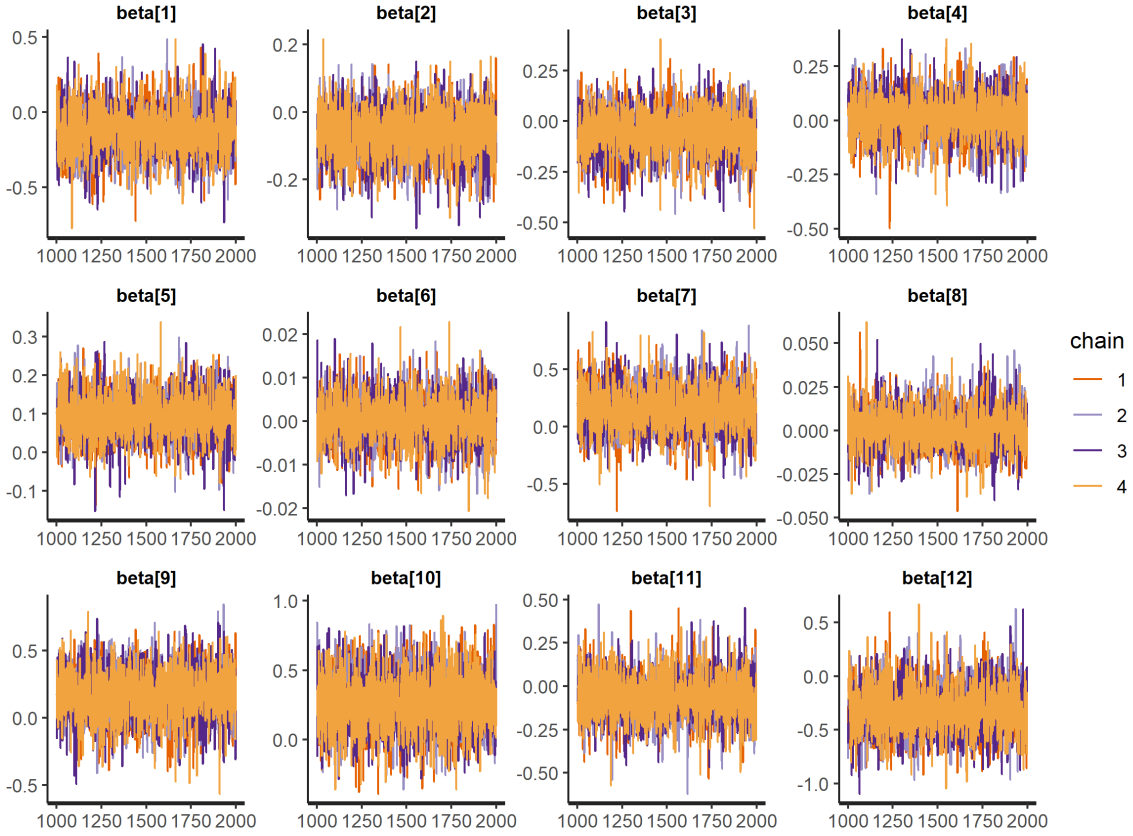

Figure 1: Traceplot of alliance level parameters in the non-major power sample.

Normalizing Allied Capability

As noted in the paper, I place allied capability in the membership matrix **Z** on the same scale as the other parameters by normalizing it by year. Within each year, I divide the total military

spending of allied states by the maximum value, so capability values within each year range from just above zero to one. This ensures that allied capability is comparable within years, and that I do not treat more modern alliances as the most capable as raw defense budget sizes increase.

The choice of this specific normalization is less theoretically informed than using capability itself. Therefore, I assessed the use of different normalizations and rescalings for allied capability by comparing model fit. I fit three models in addition to the one presented in the paper. The first rescaled allied capability by dividing each capability value by the maximum of capability without grouping alliances by year. The second rescaled alliance capability by dividing by two standard deviations, which is problematic because it introduces negative capability values. The last used total allied CINC scores instead of military spending as an indicator of allied capability, which also facilitates comparisons of allied capability within years. CINC scores measure the share of total world military capability each state has in a particular year, so it is useful for comparing allied capability within years (Singer, 1988).

After estimating these three models, I used leave-one-out (LOO) cross validation to assess model fit (Vehtari, Gelman and Gabry, 2017). LOO estimates pointwise out-of-sample prediction accuracy using the log-likelihood evaluated at the posterior simulations of the parameter values.[1] All diagnostics indicate the LOO results are not driven by unusual observations. As with other information criteria, lower values indicate better fit.

Allied Capability	elpd_diff	se_diff	elpd_loo	se_elpd_loo
Normalized by Year	0.000	0.000	-1159.513	184.714
Rescaled by Maximum	-3.165	2.643	-1162.679	184.723
Recaled by 2SD	-10.749	6.116	-1170.262	184.741
Total Allied CINC	-12.308	5.576	-1171.821	184.683

Table 2: Leave-one-out cross validation to assess model fit with different rescalings or normalizations of alliance capability.

Table 2 summarizes the assessment of each model using the expected log pointwise predictive

[1]The widely applicable information criteria (WAIC) produces similar results, but the estimates for the CINC model may be driven by an unusual observation.

density (elpd). I use the model from the paper as the comparison model: a negative elpd_diff implies the normalized model fits the data better. The difference also has some uncertainty, which is summarized by the se_diff column of Table 2. The other three models have a negative elpd_diff compared to the model with normalized capability by year. For the models with CINC and rescaling by two standard deviations the difference is large, relative to the se_diff, so there is a clear preference for the normalized model. Normalizing by year provides at best a marginal improvement over a model where capability is rescaled using the maximum. Rescaling capability by the maximum produces similar inferences about alliance characteristics, including treaty depth.

Fake Data Simulation Check

With any complicated model, simulating fake data and seeing if the model can recover known parameters is essential. Fake-data simulation helps validate results from observed data and catch potential pathologies. This section summarizes results from fitting the multilevel model to fake data.

I simulated a dataset of 2000 t-distributed observations with 50 states observed for 200 years and 100 alliances. The outcome has a different scale than the military spending outcome variable, so coefficient values here do not match reported values in the paper. I then simulated two state and alliance level variables and a sparse matrix of state membership in alliances. Last, I ran the model without evaluating the likelihood, generating a posterior prediction of the outcome based on the fake data.

To check whether the model could recover known parameters, I took the 12th draw of the posterior distribution. This draw included a simulated outcome for each observation and a set of coefficients. I then fit the multilevel model on the simulated outcome values and checked whether the credible intervals contained the corresponding parameter values. If a parameter is within the 90% credible interval, the model captures it.

The model recovers known parameters with a high degree of accuracy. As shown by Figure 2, the two credible intervals of the alliance-level regression include the known values. Credible interval coverage for the variance hyperparameters and γ parameters is also acceptable.

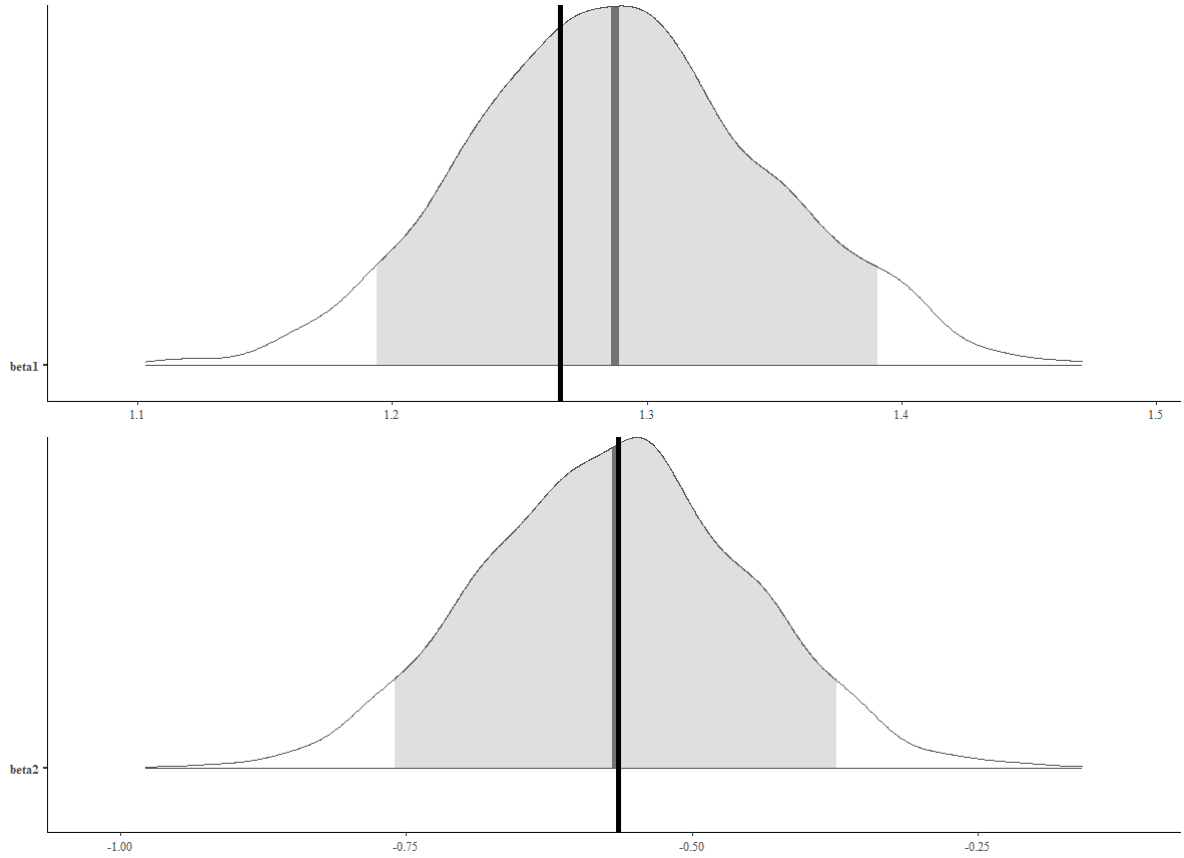

Figure 2: Posterior distributions of β parameters from fitting multilevel model to fake data. The black vertical line marks the known parameter value, and the grey area is the 90% credible interval.

Even with small multiples, the 100 λ parameters are harder to plot, so I offer a descriptive summary here. Among the λ parameters, 93 of 100 intervals contain the known λ value. Given the large number of parameters and smaller sample, this is acceptable accuracy. Even the seven inaccurate confidence intervals were quite close— all were within .015 of the known parameter.[2]

In summary, convergence diagnostics and fake data fitting both suggest that the multilevel

[2]Fine margins around these intervals implies that the exact number of accurate λ intervals is sensitive to simulation variance.

model is working well. No convergence diagnostics indicate problems exploring the posterior. Just as importantly, the model can recover known parameters from fake data. The next section provides more detail on results from the major and non-major power samples.

Robustness Check 1: Alternative Measure of Military Spending

The main findings in the manuscript rely on the Correlates of War military spending. Due to reporting issues, definition problems and measurement challenges, other measures of military spending could lead to different results. I check the robustness of my results by using Nordhaus, Oneal and Russett (2012)'s measure of military spending, which combines data from the COW project and the Stockholm International Peace Research Institute (SIPRI). DiGiuseppe and Poast (2016) use this measure of military spending in their paper.

I estimate the same multilevel model on this measure of military spending, which covers from 1949 to 2001. This model also checks whether how treaty depth modifies the impact of alliance participation on military spending changes after World War II. Because the coefficient on a lagged dependent variable in this model is close to one, implying a probable lack of stationarity in levels, I use changes in military spending as the outcome of interest.

Figure 3 summarizes the alliance-level regression parameters. As with the COW data, the credible interval for treaty depth is negative and does not overlap zero. All the parameter estimates are similar in this data, which increases my confidence that the results are not driven by the COW spending data.

Robustness Check 2: Single-Level Regression

Though the multilevel model best reflects the theory, I also fit some more standard panel data models. In what follows, I briefly present results from robust regressions of state-year percentage

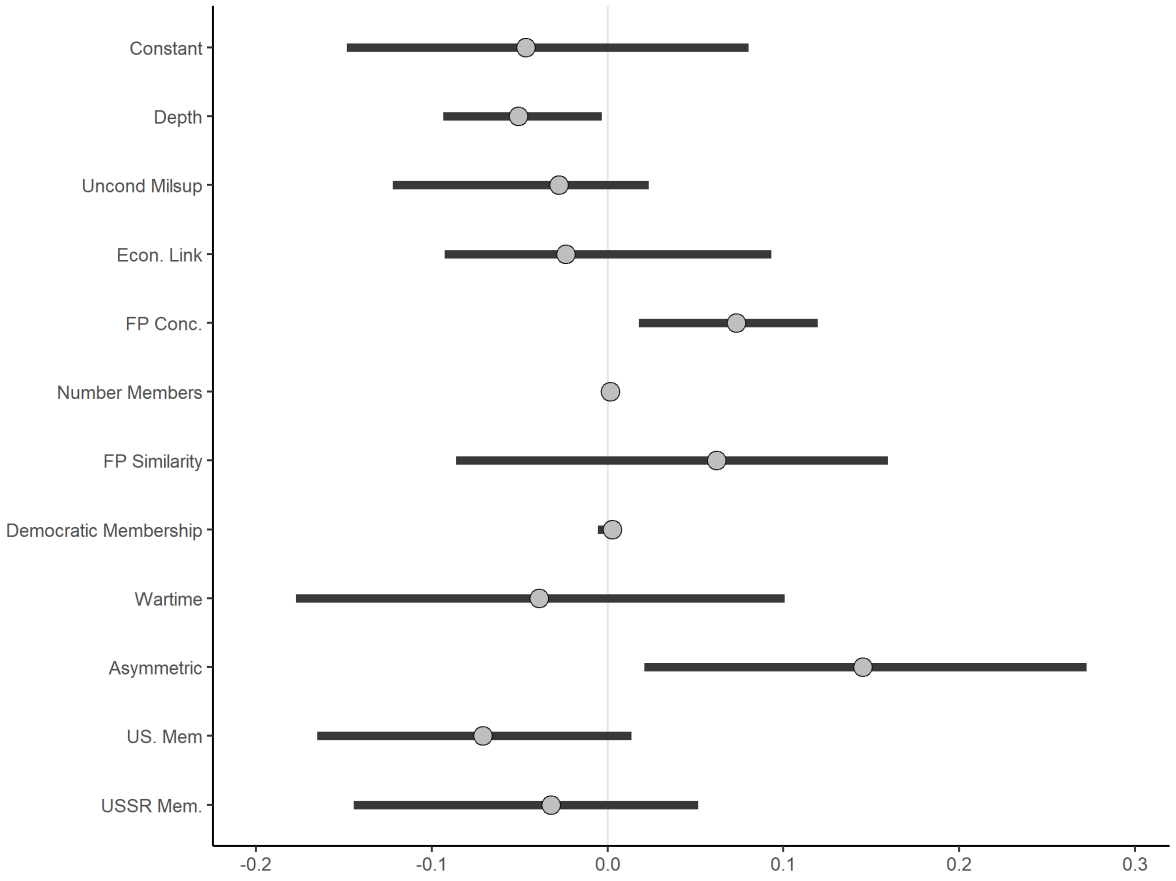

Figure 3: 90% credible intervals of the β parameters from an analysis of changes in non-major power military spending from 1949 to 2001.

changes in military spending in the same sample of non-major powers. As in the multilevel model, I applied the inverse hyperbolic sine transformation to the outcome. In these models, I employ two indicators of alliance depth. The first is the average depth of a state's alliances. The second is a dummy which equals 1 if a state has at least one alliance with greater than average depth. Both variables compare states as the depth of their alliance portfolio shifts. In addition to the state-level controls in the multilevel model, I included averages of alliance size and democracy and the log of total allied capability as controls.

I estimated several models, including robust regressions on all states, non-major powers, and non-major powers in alliances. I also applied fixed effects to an OLS model of percentage changes in defense expenditures. The estimated association between average treaty depth and military spending changes is summarized in Figure 4. Results are inconsistent- the average depth measure fails to reject the null without fixed effects. The deep alliance dummy coefficient estimate is negative and statistically significant across several samples and model specifications, however.

The analysis of non-major powers in alliances is the best approximation of the multilevel model, as it compares non-major powers with different kinds of alliances. Analyzing a sample of all non-major powers includes states with no alliances, with does not match the comparison in the multilevel model. To assess the robustness of the coefficient estimate in the sample of non-major powers with alliances, I performed Extreme Bounds analysis. Specifically, I present results from Sala-i Martin (1997)'s method of bounds analysis in Figure 5.

Figure 5 shows the distribution of the deep alliance coefficient and an indicator of whether the alliance includes economic agreements. Across many specifications, where all regression coefficients are doubtful, the CDF of the deep alliance coefficient has 99% negative mass. Even though the normality assumption is clearly violated, the histogram in Figure 5 shows little evidence having a deep alliance increases percentage changes in military spending. The bounds analysis indicates a deep alliance dummy is a robust predictor of percentage changes in military spending across over 1500 single-level model specifications.

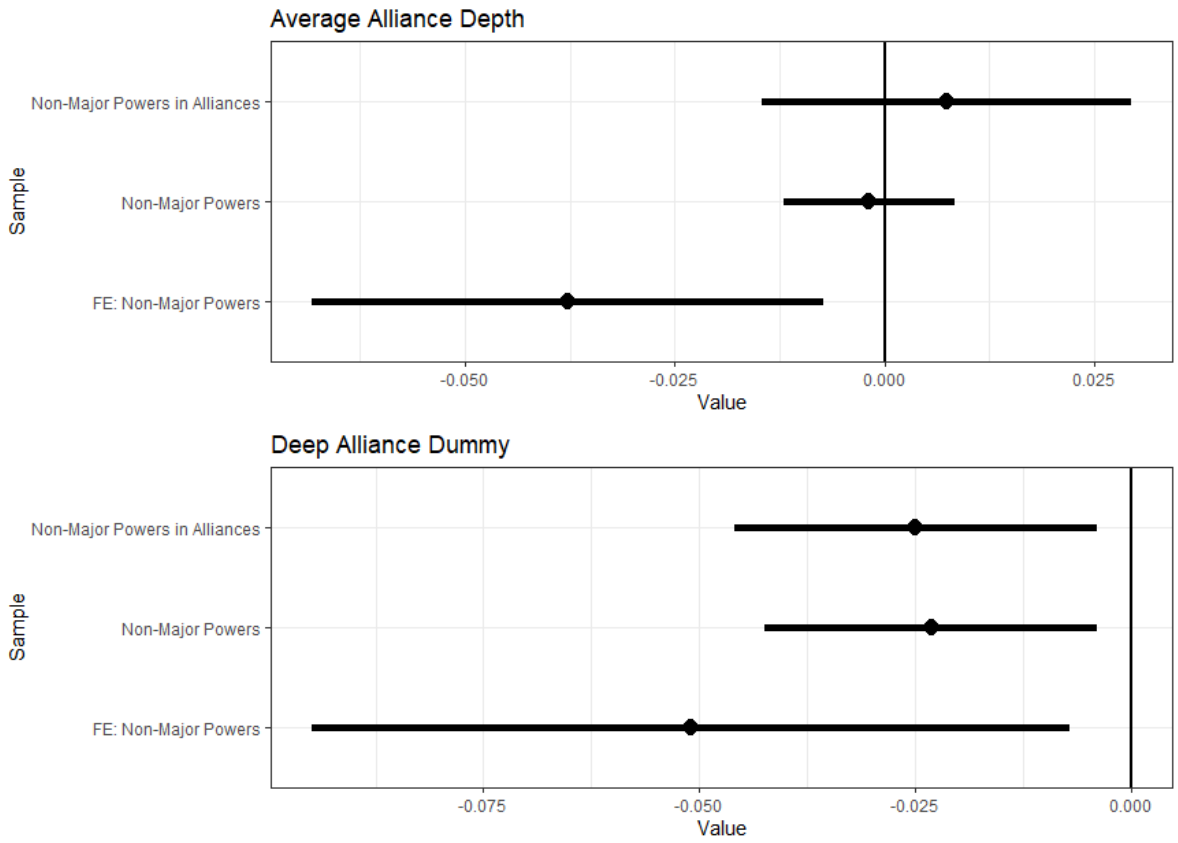

Figure 4: Estimated effect of average alliance treaty depth or a dummy indicator of participation in a deep alliance on percentage changes in non-major power military spending.

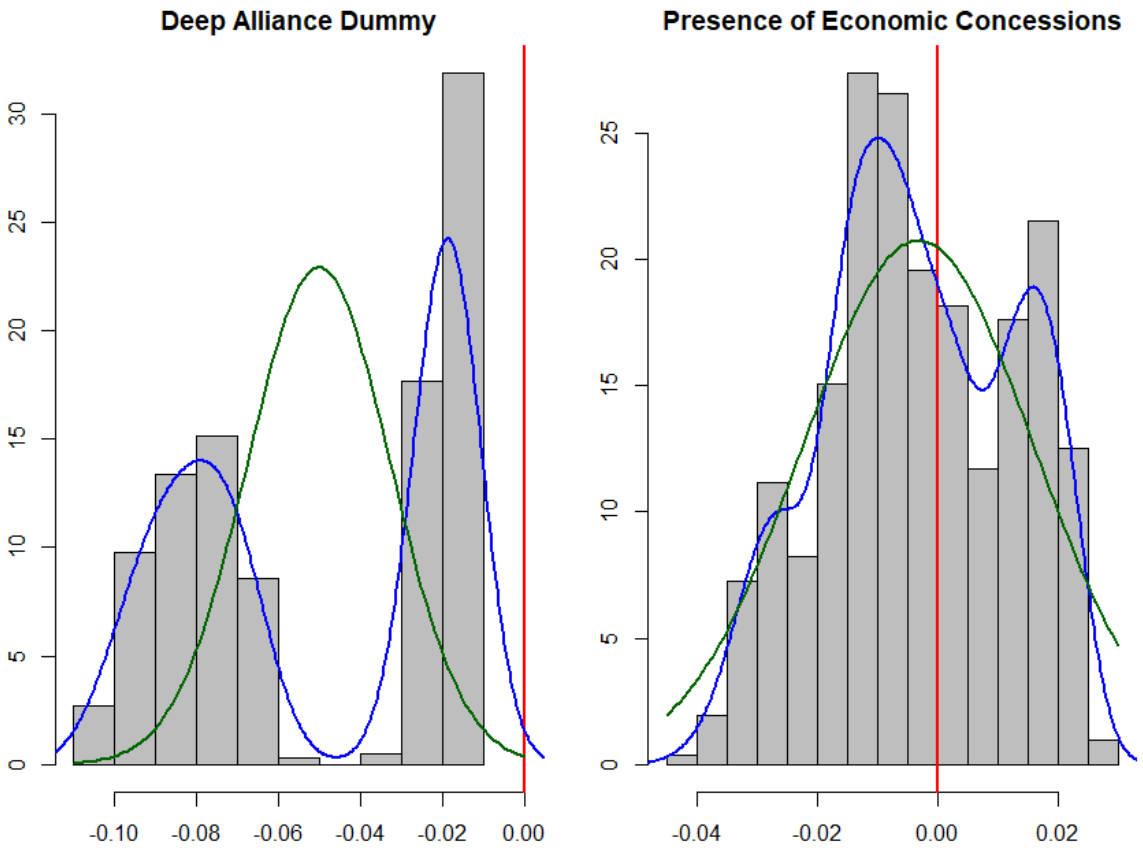

Figure 5: Histogram of coefficient values for a deep alliance dummy and economic concessions in at least one alliance in a single-level robust regression of non-major powers in alliances.

References

DiGiuseppe, Matthew and Paul Poast. 2016. "Arms versus Democratic Allies." *British Journal of Political Science* pp. 1–23.

Nordhaus, William, John R Oneal and Bruce Russett. 2012. "The Effects of the International Security Environment on National Military Expenditures: A Multicountry Study." *International Organization* 66(3):491–513.

Sala-i Martin, Xavier. 1997. "I Just Ran Two Million Regressions." *The American Economic Review* 87(2):178–83.

Singer, J David. 1988. "Reconstructing the correlates of war dataset on material capabilities of states, 1816–1985." *International Interactions* 14(2):115–132.

Vehtari, Aki, Andrew Gelman and Jonah Gabry. 2017. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing* 27(5):1413–1432.