Supplemental Materials to: "Point Break: Using Machine Learning to Uncover a Critical Mass in Women's Representation"

Kendall D. Funk*
Hannah L. Paul†
Andrew Q. Philips‡

Last updated: July 19, 2021

Contents

1 Alternative Visualizations of the Relationship Between Women's Representation and Expenditures **2**
1.1 ICE Plots 2
1.2 Local Dependence and ALE Plots 2

2 Linear Regression Results **4**

3 Robustness of the Nonlinear Relationships: Alternative Modeling Strategies **8**

4 Re-Creating the Main Analysis Including Left Government **14**

5 A Re-Analysis Changing the Marginal Distribution in Regards to Quotas **17**

6 Machine Learning-Informed Parametric Testing **26**

*Assistant Professor, School of Social and Behavioral Sciences, Arizona State University, kendall.funk@asu.edu
†PhD Candidate, Department of Political Science, University of Colorado Boulder, hannah.paul@colorado.edu
‡Assistant Professor, Department of Political Science, University of Colorado Boulder, andrew.philips@colorado.edu

1 Alternative Visualizations of the Relationship Between Women's Representation and Expenditures

In the subsections below, we show several alternative depictions of our results from the main paper.

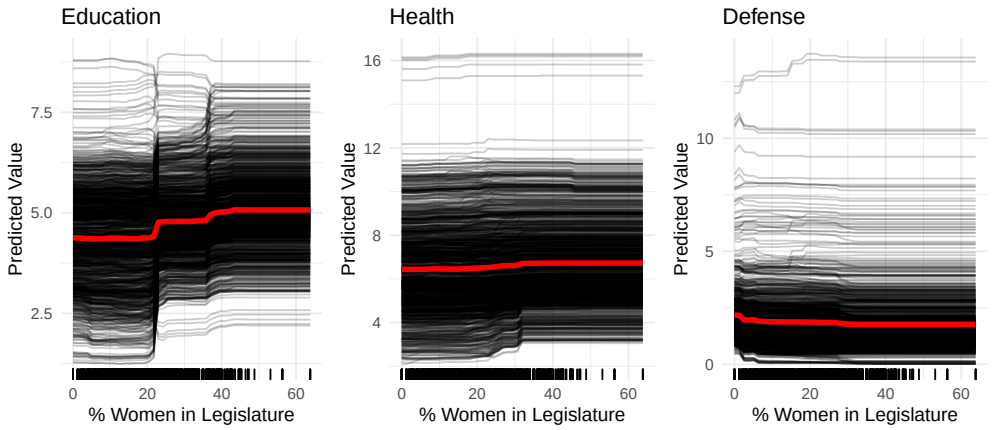
1.1 ICE Plots

In the main paper we used partial dependence plots (PDPs) to show the average expected value of expenditures, given a particular value of women's representation in the legislature, averaging over the values of all the other predictors in the model. A critique of this is that by only looking at averages, we might be overlooking underlying heterogeneity in the effects of women's representation for particular observations. A way around this is to show individual conditional expectation (ICE) plots, which show the predictions for each observation across the percentage of women in the legislature. These are shown in Figure 1. The PDP is simply the average of the ICE lines at a particular value of percent women in the legislature, and is shown by a red line in Figure 1. As is clear from the figure, although there definitely is some heterogeneity in the effects, a vast majority of cases exhibit the same relationship as that shown by the PDP (albeit with much different starting intercepts).

1.2 Local Dependence and ALE Plots

PDPs and ICE plots reveal the globalized effects of the % women in the legislature on government spending. We explore the *localized* effects of women's representation on government expenditures in two ways: by using local dependence (LD) as well as accumulated local effects (ALE) plots (Apley and Zhu 2016). One issue with PDPs and ICE plots is that they use the entire marginal distribution of all other predictors. Thus, there runs a risk of making predictions for a given value of women's representation for which there exist 'un-

Figure 1: ICE Plots for Each Expenditure



realistic' combinations of the other predictors Local dependence plots. Both LD and ALE plots "zoom in" by focusing on the conditional distribution of the predictors (other than percent women in the legislature, which is the subject of interest here). LD plots show expected values of the dependent variable over the conditional distribution. In other words, for a given value of women's representation, only values of other predictors "near" that value are used to make predictions. Because the estimates are not continuous across levels of women's representation, typically, a smoothed estimate is used (Biecek 2018).

Similar to LD, ALE plots also use the conditional distribution of other predictors, $\mathbf{x}_c$, as opposed to the entire marginal distribution, as is the case with PDPs and ICE plots. ALE plots take all of the observations in a particular k region, fits them with the value of $\mathbf{x}_c$ at the lower break point, and then refits these observations with the value of $\mathbf{x}_c$ at the upper break point. The big difference is that ALE plots "accumulate" these local effects, which avoids an issue—present in LD, PDP and ICE plots—whereby omitted variables correlated with the percent women in the legislature may bias predictions (really, we end up attributing changes due to these omitted variables as being caused by women's representation). These changes in predictions are plotted, creating an "accumulation" of local gradients in the equation to create a single plot.

In Figures 2, 3 and 4, we show the predicted values of education, health and defense,

respectively, across levels of women's representation in the legislature. Both the local dependence (dashed line) and accumulated local effects (dotted) predictions are shown, as well as the original partial dependence predictions (solid line) for reference.[1] Across all models the ALE and PDP lines show the same effect across time; in the case of defense spending in Figure 4, the predictions are nearly identical. The LD predictions appear to be slightly more different than the other two. While large jumps in predictions appear to occur at the same time—see for instance at 22 percent and about 36 percent of women legislators in Figure 2—predictions for the LD tend to have a much wider range than for the ALE or PDP predictions. One explanation for this is that the LD may be picking up the effect of omitted variables (Biecek 2018). Still, the predictions across the three approaches have the same trajectory, even if the actual predictions differ based on assumptions about local versus global dependence.

2 Linear Regression Results

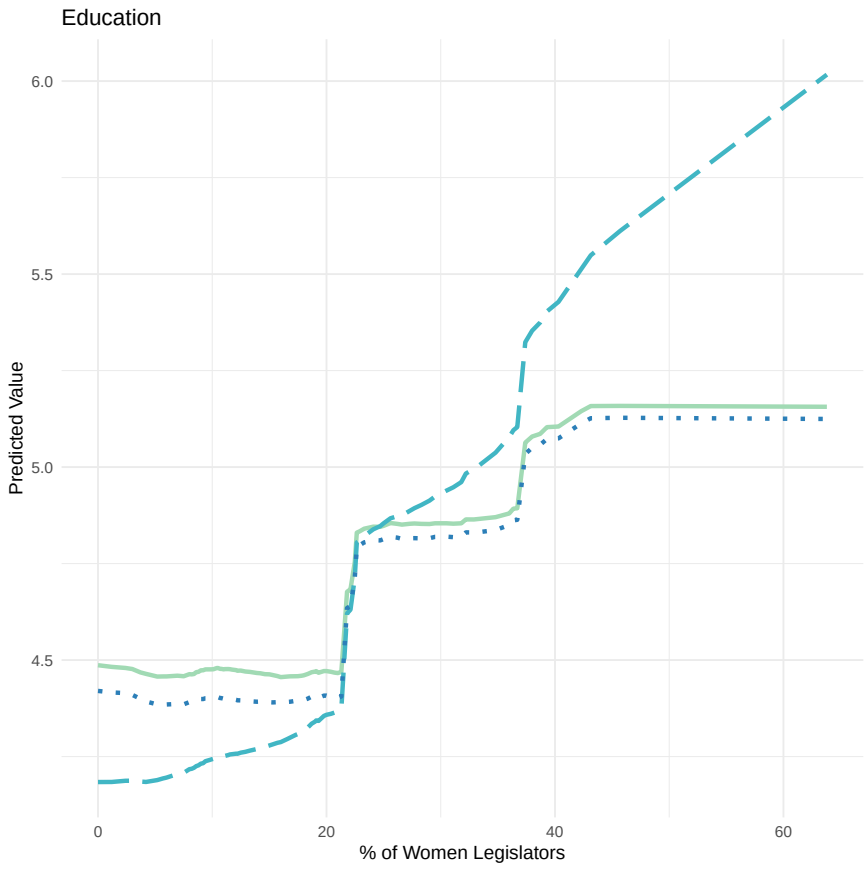
In the main paper we used an ensemble-based machine learning approach to exploring the relationship between women's representation in government and expenditure outcomes. In this section we estimate several models using OLS to test the robustness of our overall findings:

- OLS: linear regression (only regressor is % women in the legislature)
- OLS-FE: linear regression with country intercepts
- OLS-LDV: linear regression with a lagged dependent variable
- OLS-2FE: linear regression with country and year intercepts
- OLS (full): same as "OLS", but including the 34 other predictors used in the Random Forest analysis.[2]

[1] This was done using the `DALEX` package (Biecek 2018).

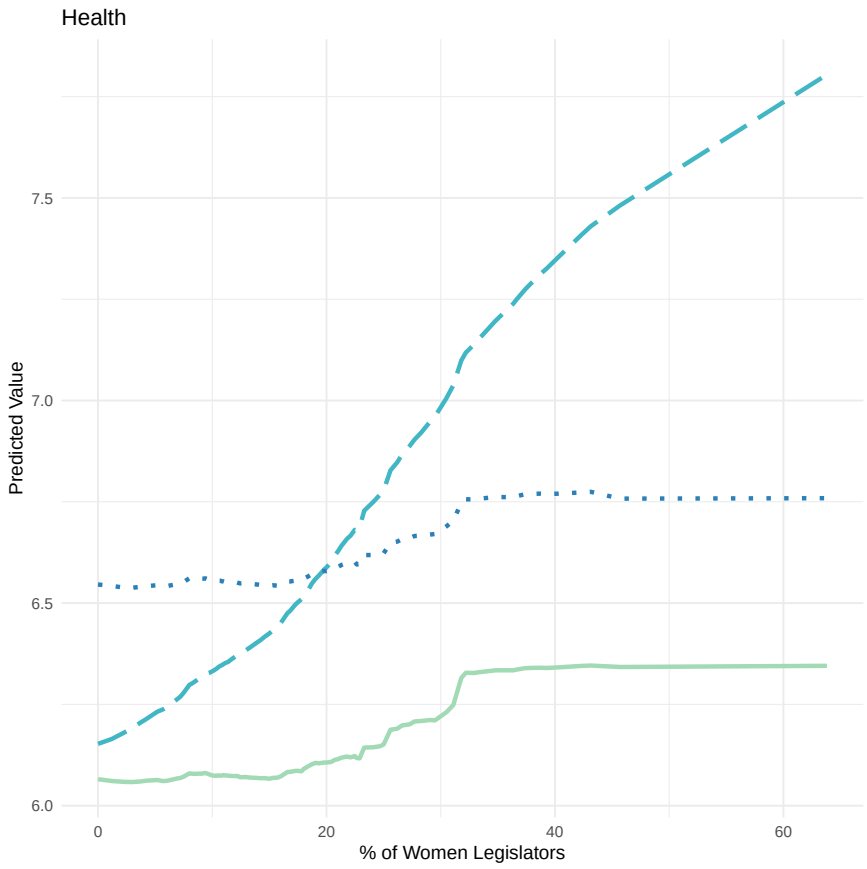
[2] Recall these were: population density, total population, % of labor force that is female, % of pop-

Figure 2: Alternative Dependence Plots for Education



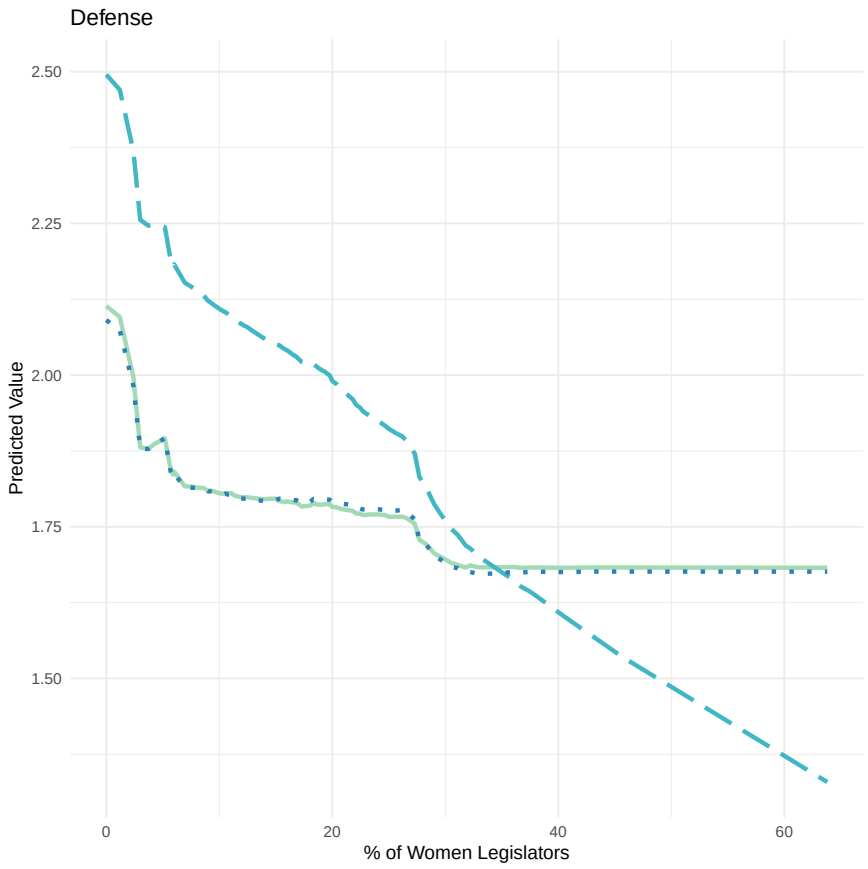
Note: partial dependence plot (solid), local dependence (dash), accumulated local effects (dotted).

Figure 3: Alternative Dependence Plots for Health



Note: partial dependence plot (solid), local dependence (dash), accumulated local effects (dotted).

Figure 4: Alternative Dependence Plots for Defense



Note: partial dependence plot (solid), local dependence (dash), accumulated local effects (dotted).

- OLS-FE (full): same as "OLS-FE", but including the 34 other predictors used in the Random Forest analysis.
- OLS-LDV (full): same as "OLS-LDV", but including the 34 other predictors used in the Random Forest analysis.
- OLS-2FE (full): same as "OLS-2FE", but including the 34 other predictors used in the Random Forest analysis.

The results for education, health, and defense spending are shown in Figures 5, 6, and 7, respectively. It is clear that, across most model specifications, the direction of the coefficients are nearly always in the same direction as what we found using the machine learning approach in the main paper. Often, the coefficients are statistically significant, even though (1). since this is linear regression, we are assuming a linear relationship between women's representation and expenditures, which was exactly the assumption we seek to relax using the Random Forest approach in the main paper, and (2). in the "full" models, there are 34 other predictors included, sharply increasing multicollinearity.[3]

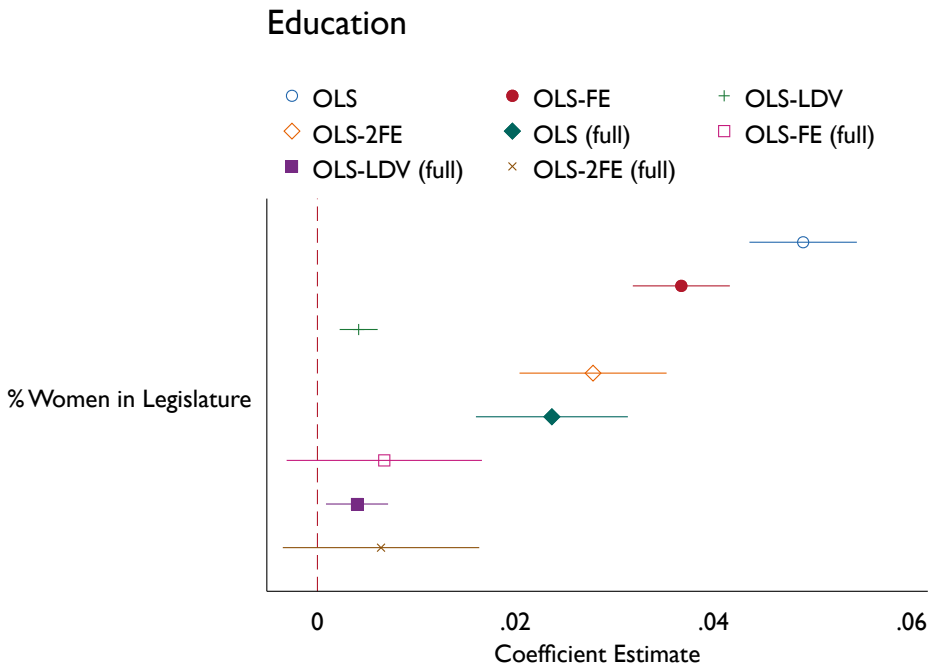
3 Robustness of the Nonlinear Relationships: Alternative Modeling Strategies

Below, we present the results from four alternative modeling strategies to Random Forests: a piecewise regression, a generalized additive model (GAM), A cubic polynomial, and a neural network. Each are described below.

ulation that is female, Polity, anemia prevalence among women, female labor force participation rate, imports, GDP per capita, male labor force participation rate, age dependency ratio, % rural population, trade, % school enrollment, lifetime maternal mortality risk, female life expectancy, employment to population ratio, year, male life expectancy, population growth, maternal mortality rate, unemployment rate, unemployment rate among males, fertility rate, % of GDP from agriculture, de facto threshold, FDI, inflation, birth rate, quota strength (1), implemented quota, GDP growth, quota strength (2), implemented quota, quota shock.

[3]As evidence of this, the average variance inflation factor for a given variable using the "OLS" models are 56 (education), 51 (health), and 51 (defense).

Figure 5: OLS Robustness Results for Education

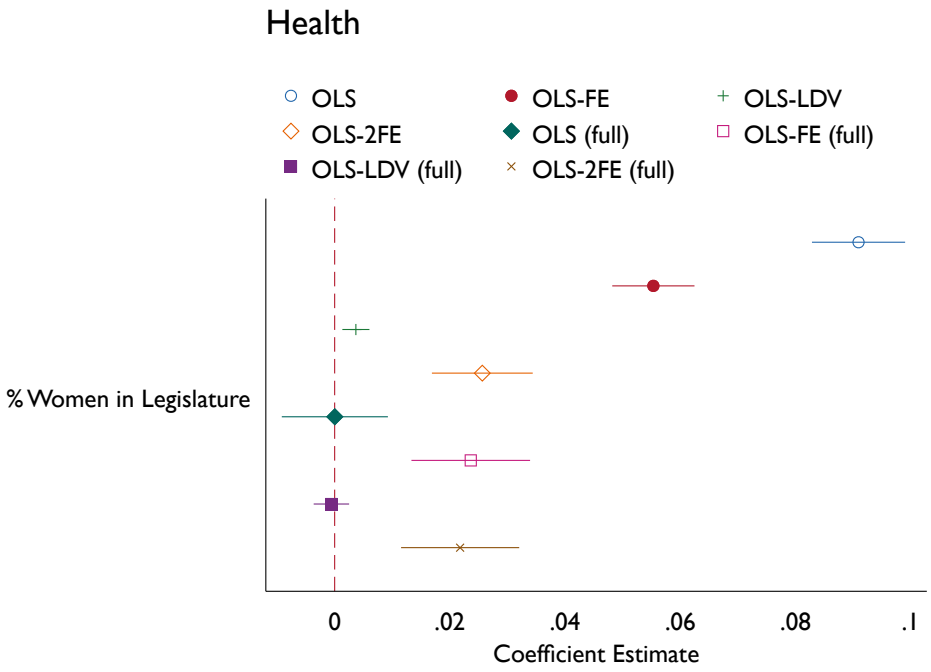


Note: Coefficient estimates with 95% confidence intervals shown. Model details described in text.

A piecewise regression is an approach to 'break' a single regression line into multiple segments. For our purposes, we can allow the relationship between women's representation and expenditures to have breakpoints; while the relationship between breakpoints is linear, looking across the entire range of women's representation allows us to model a potentially non-linear relationship. Rather than specifying *a priori* where the breakpoints lie, we used the `segmented` package in `R` to (Muggeo et al. 2008), which uses an iterative procedure to automatically determine where the breakpoints are. We specified two breakpoints; more breakpoints would better model potential non-linearities, at the risk of overfitting the data.

A GAM is a flexible approach to modeling potentially non-linear relationships by estimating a sum of smoothed functions of a variable (typically some form of spline), plus including a standard linear effect. GAMs also typically trade off complexity versus simplicity by adding a penalization term to the loss function. We used the `mgcv` package (Wood

Figure 6: OLS Robustness Results for Health



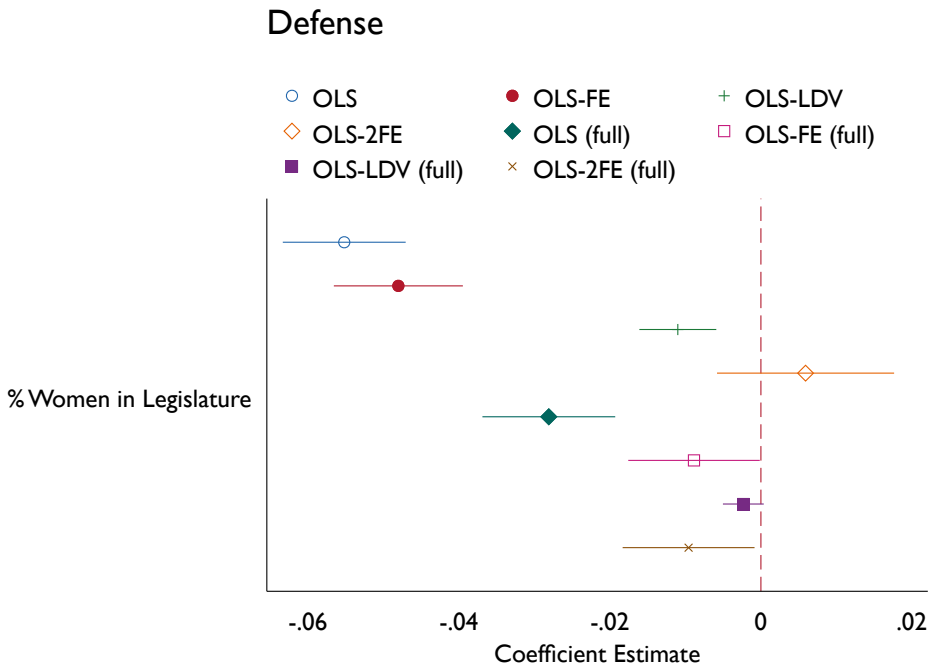
Note: Coefficient estimates with 95% confidence intervals shown. Model details described in text.

2018). Because this model—and all others discussed in this section—include all predictors (unlike Random Forests, in which it is entirely possible a predictor never appears in a given tree), we face a trade-off between assuming all relationships are linear (i.e., estimating a linear model) and assuming that all covariates have a non-linear relationship with the dependent variable, the latter of which would be heavily parameterized. As a compromise, in addition to allowing the percent of women's representation in the legislature to take on a smoothed functional form, we also include the top ten other most important predictors (as determined by the variable importance plots in the main manuscript). Thus our model is:

$$\text{Expenditure Category}_{it} = \beta_0 + f(\mathbf{X}_{it}) + \boldsymbol{\gamma}\mathbf{Z} + \varepsilon_{it} \tag{1}$$

Where $f(\mathbf{X}_{it})$ are the additive function of inputs of the top ten most important variables,

Figure 7: OLS Robustness Results for Defense



Note: Coefficient estimates with 95% confidence intervals shown. Model details described in text.

plus the women's representation variable, and $\mathbf{Z}$ are all other covariates that enter into the model linearly. For a basis function (required for estimating $f(\mathbf{X}_{it})$) we chose the default thin-plate regression spline. Smoothing parameters were chosen using generalized cross-validation. We let the model automatically choose the dimension of the basis for each variable in $f(\mathbf{X}_{it})$, the exception being the variable "Quota Strength 2", which has only six unique values, so we restricted that dimension to a maximum of three.

The cubic polynomial is perhaps the simplest approach, since it just involves adding cubed and squared terms of women's representation to the moodel:

$$\text{Expenditure Category}_{it} = \beta_0 + \beta_1 \%\text{Women in Legislature}_{it} + \beta_2 \%\text{Women in Legislature}_{it}^2 + \beta_3 \%\text{Women in Legislature}_{it}^3 + \boldsymbol{\gamma}\mathbf{Controls}_{it} + \varepsilon_{it} \tag{2}$$

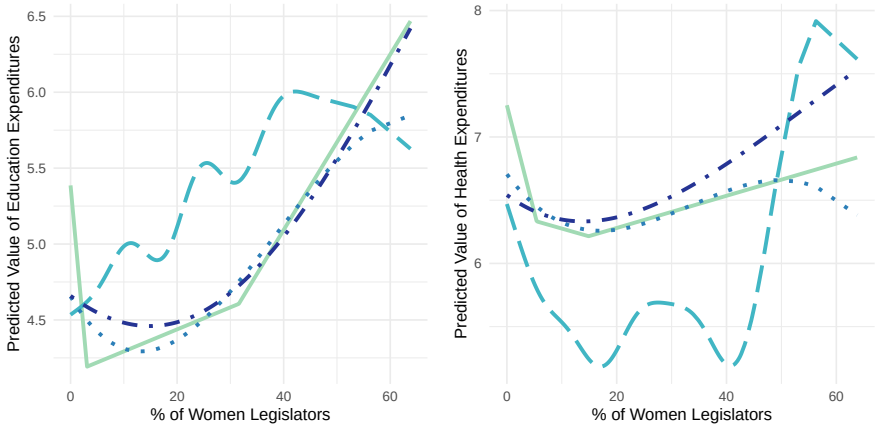
We then plotted the predicted value of each expenditure category using this model, setting all covariates to their means (or modes if dichotomous) across the plausible range of women's representation.

Neural networks start with a set of inputs (our list of covariates), pass them through one or more "hidden layers" each of which contain several neurons, which are then used to weight the inputs in order to maximize predictive ability of the output (our expenditure dependent variables). Typically—as in our model—backpropogation is used, which feeds errors back into the hidden layer in order to augment the weights to improve predictions. We estimated a relatively simple neural net using the `neuralnet` package (Günther and Fritsch 2010), which consisted of a single hidden layer with 17 neurons.[4] As is standard, we ensure all predictors are scaled before estimation (Hastie, Tibshirani and Friedman 2009). We used the default threshold of 0.01 as the stopping criteria for the partial derivatives of the error function, specified a learning rate for the backpropogated errors of 0.01, and used resilient backpropogation with weight backtracking (also the default).

Figure 8 shows the predictions for each model across levels of women's representation while holding all other covariates at their means, with the exception of the neural net, which uses partial dependence similar to the Random Forest plots described above. The results differ slightly since these are all heavily parameterized models; however, the estimates for education and health spending continue to suggest a nonlinear relationship, while the estimated effect for defense spending appears to be roughly linear. All three plots show effects similar to those produced by the Random Forest partial dependence plots, although the predictions for healthcare spending using a GAM are the most dissimilar.

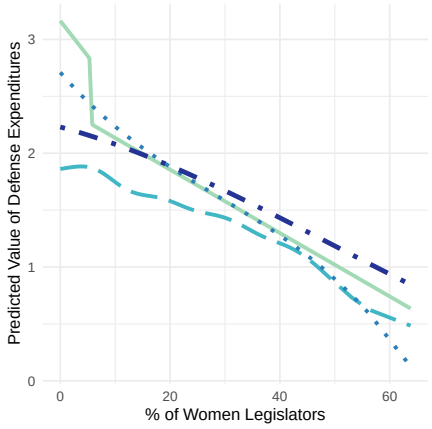
[4]How many neurons to include is often a trial and error task; we used one rule-of-thumb that the number of hidden neurons should be about two-thirds of the total number of predictors (c.f., https://stats.stackexchange.com/questions/181/how-to-choose-the-number-of-hidden-layers-and-nodes-in-a-feedforward-neural-netw. We relied on the `NeuralNetworkVisualization` package for creating the partial dependence plots (Seufert and Afanasev N.d.).

Figure 8: Predicted Values of Expenditures, Alternative Model Specifications



(a) Education

(b) Health



(c) Defense

Note: Solid = piecewise regression, dashed = GAM, dotted = cubic polynomial, dot-dash = neural net.

4 Re-Creating the Main Analysis Including Left Government

In this section, we re-run the main analysis including a dichotomous variable equal to one if the largest governing party in the legislature is left-leaning. We created this dummy variable from a categorical variable in the Database on Political Institutions that identifies the largest seat share in the legislature as left, center or right leaning. We did not include this variable in the main results due to a large loss (up to 39 percent) of data. We do not find any significant differences in results with the inclusion of this variable. The first three plots—9, 10, and 11—show the most important variables for each of the spending outcomes. Note that the left government variable is called "Largest Government Party" in the figures. The VIP plots indicate that whether or not the largest government party in the legislature is leftist or not does not appear to be of high importance for predicting government expenditures. More importantly, the position of % women in the legislator either does not change in its position of importance for the variable importance plots or changes slightly in position. For education expenditures, women's representation remains the third most important variable. In the main analysis, women's representation is the eighth most important variable in the model. When we add the largest governing party variable to the model, women's representation is the seventh most important variable. For defense spending, women's representation goes from being the tenth most important variable in the main analysis to eighth most important in the analysis that includes the largest governing party measure.

The partial dependence plots shown in Figure 12 are also generated with data including the largest government party variable. They appear to be nearly identical to those in the main paper.

As with the PDPs, we see no changes in the interaction plots when we include the largest government party variable as a predictor. These are shown in Figures 13, 14 and

Figure 9: VIP for Education Including Left Government

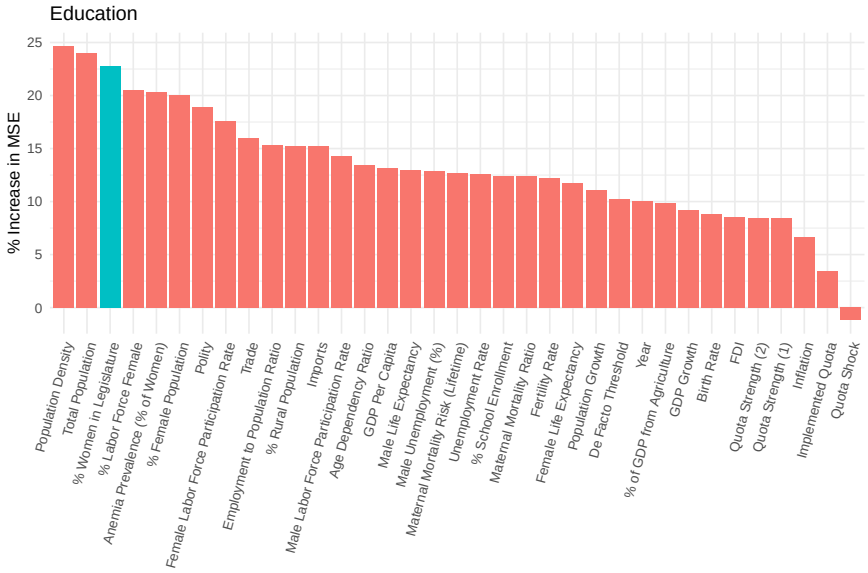


Figure 10: VIP for Health Including Left Government

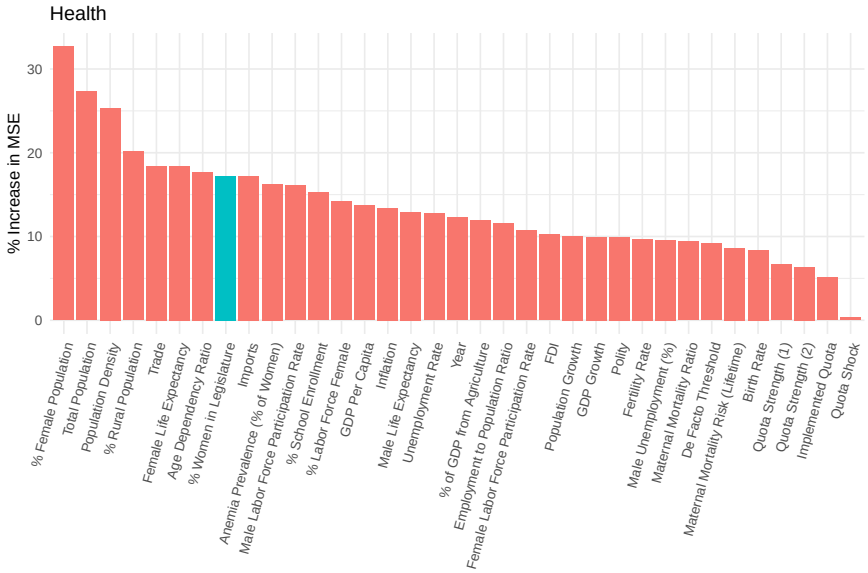


Figure 11: VIP for Defense Including Left Government

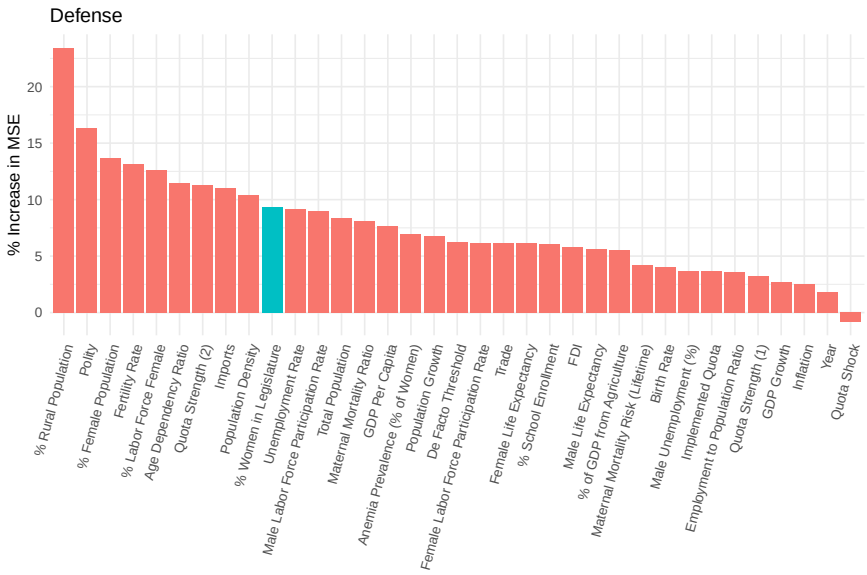
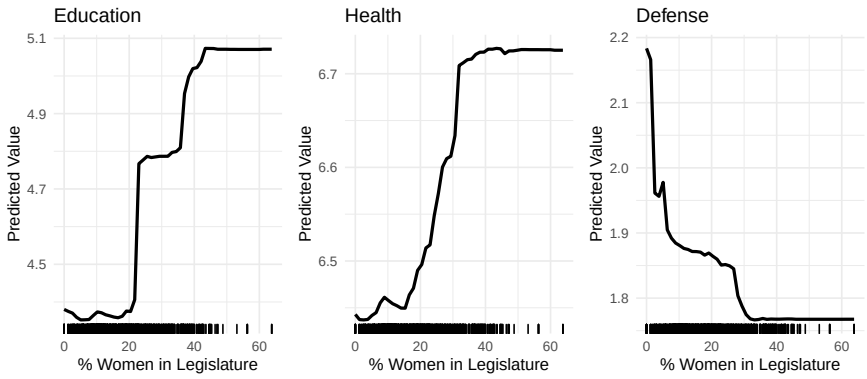


Figure 12: PDP for All Outcomes including Including Left Government



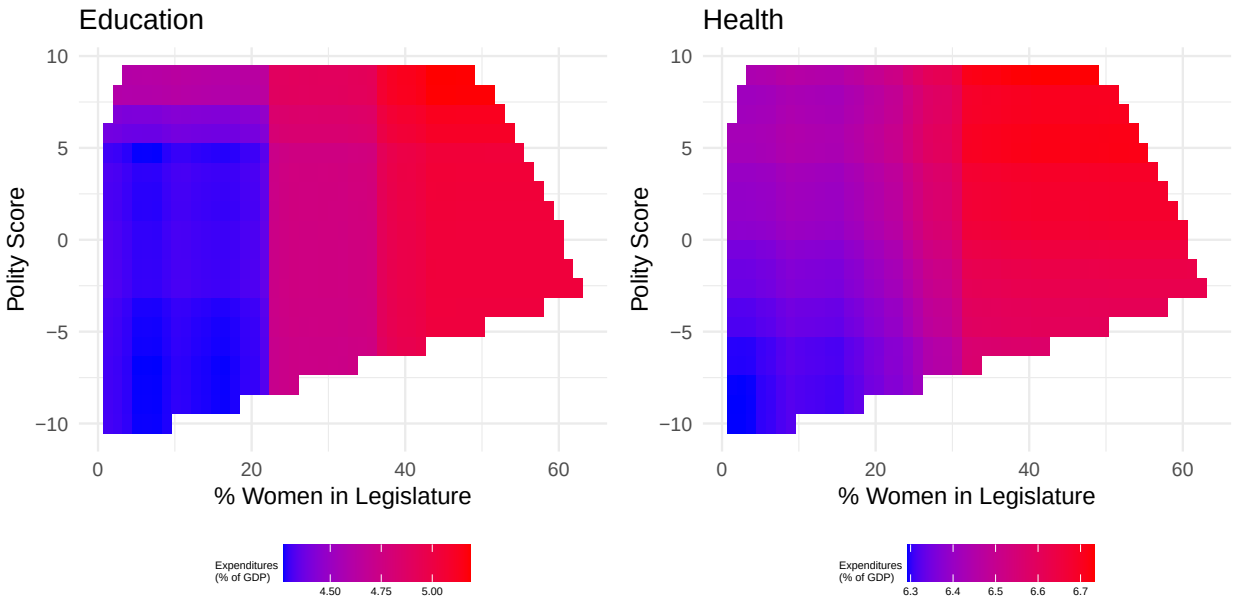
15.

5 A Re-Analysis Changing the Marginal Distribution in Regards to Quotas

As discussed in the main paper, partial dependence plots work by holding all Xc control variables at their observed values for each observation, setting the covariate of interest (percent women in the legislature, x*, in most of our applications) to a fixed value for all observations, obtaining predictions and then averaging over all predictions to obtain the partial dependence plot. While this approach—which is very similar to the observed-value approach (Hanmer and Ozan Kalkan 2013)—is much less sensitive to doing something like fixing Xc to single values (e.g., means or modes), it can lead to instances where parts of the marginal distribution of Xc for a certain value of x* are never actually observed in the data. This is more of an interpolation issue than an extrapolation one (King and Zeng 2006)—to be clear, our PDPs are not extrapolating outside of the convex hull for the variables of interest (e.g., percent women in the legislature). Still, we know that for the five quota variables we include as controls in our models, there may be implausible combinations of these marginal distributions and x*. For example, we might be averaging over countries with a 25% quota in order to calculate the predicted value given 1% women in the legislature.

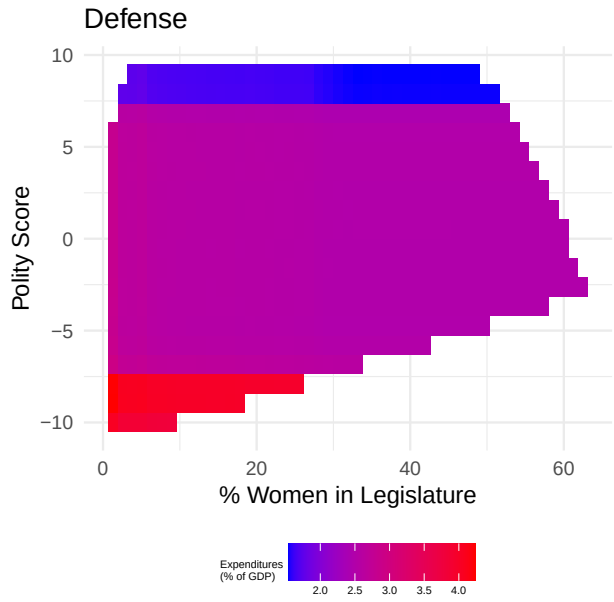
Does this affect our results? We argue that it will not for three reasons. First, even if such imbalances do occur in the data, there are many trees for which percent women will be included while these quota variables will not (or vice-versa). Second, our accumulated local effects plots above support our PDP findings, even though the 'zoom in' on more localized effects rather than the entire marginal distribution. Third, the literature on gender quotas finds that loopholes in quota legislation, lax enforcement, and incompatibility with the electoral system causes many countries to fall short of meeting the minimum threshold

Figure 13: Interactions Between Percent Women and Democracy, Including Left Government



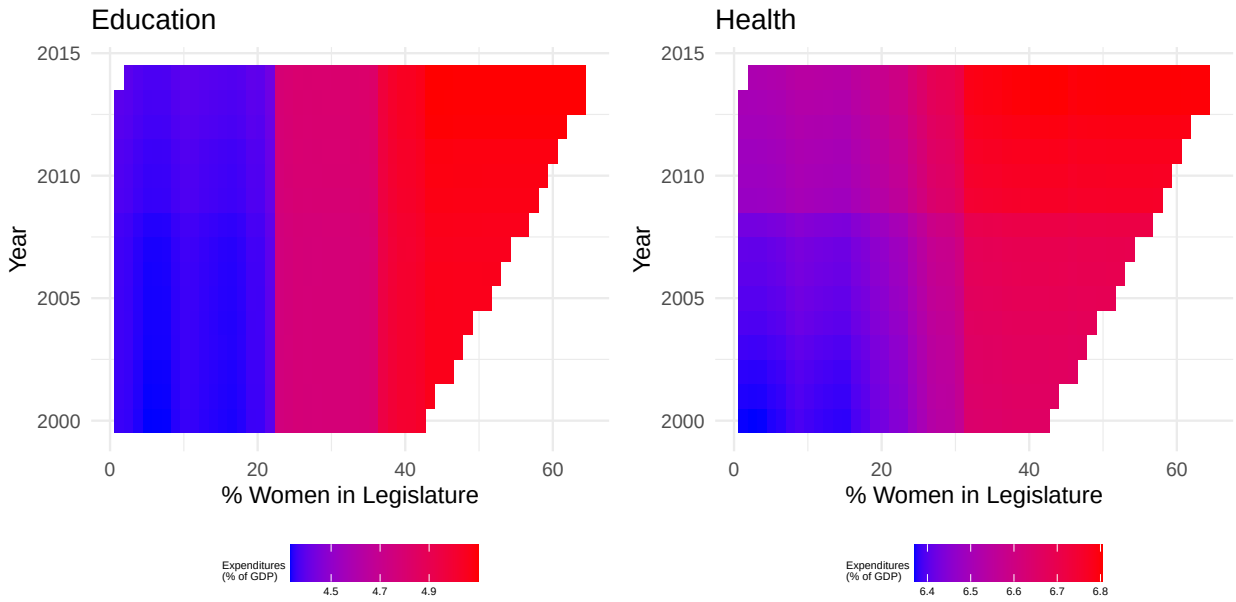
(a) Education

(b) Health



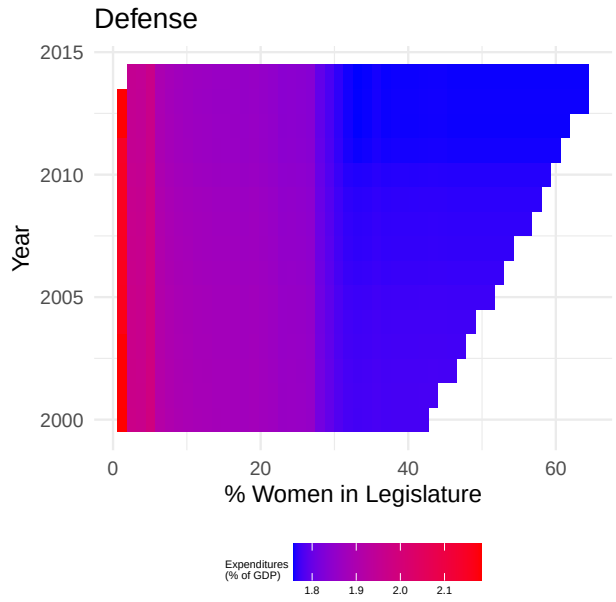
(c) Defense

Figure 14: Interactions Between Percent Women and Year, Including Left Government



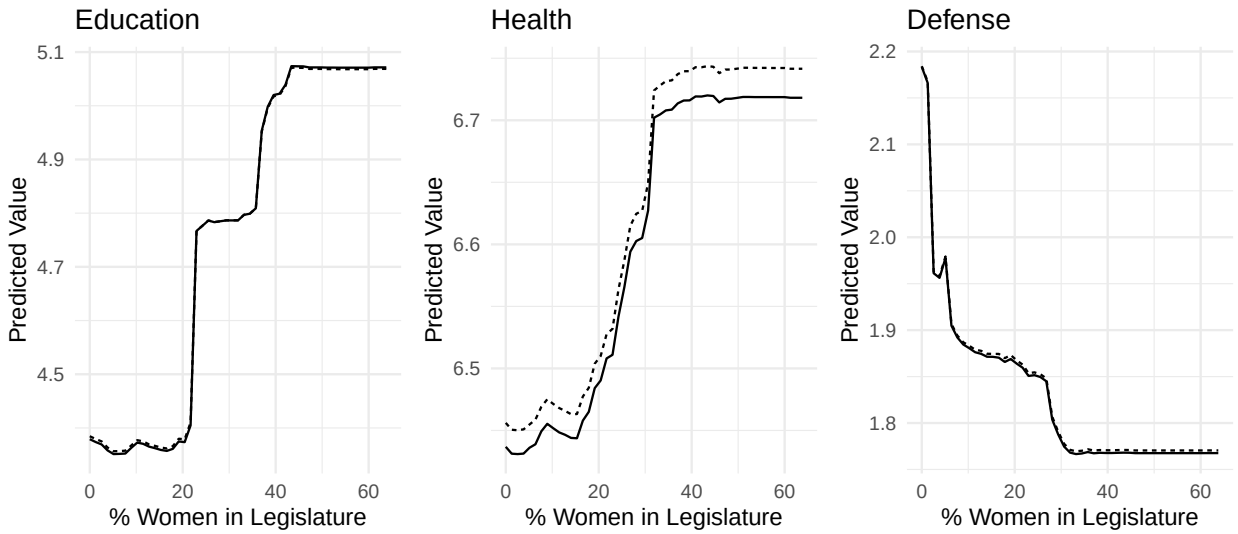
(a) Education

(b) Health



(c) Defense

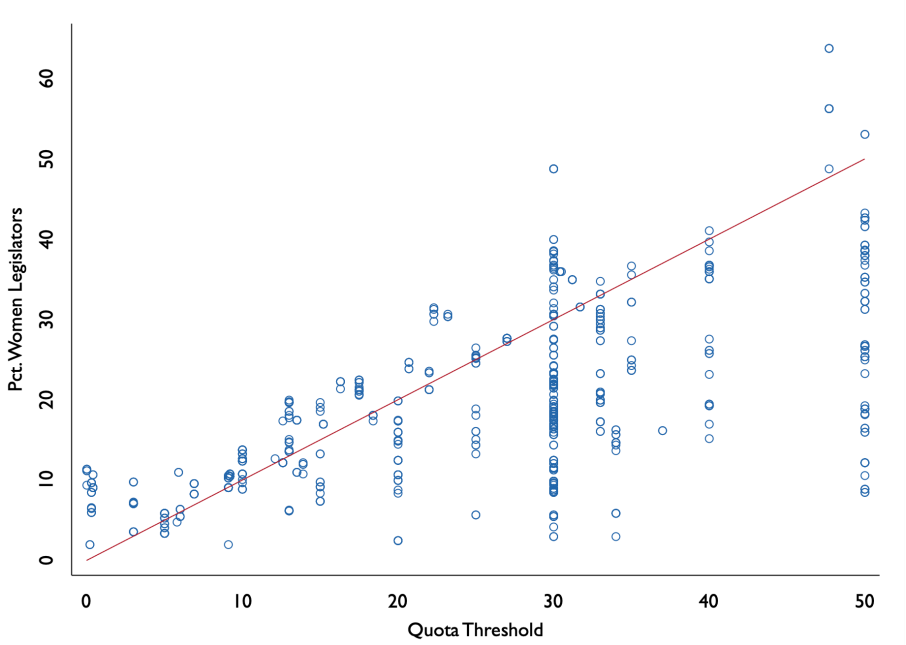
Figure 15: Interactions Between Percent Women Legislators and Quota Implementation, Including Left Government



specified by the quota law. For instance, in our data, approximately 56% of cases fall short of meeting their gender quota, 17% meet the quota (within ±1 percentage point), and 27% exceed the quota threshold. This is shown visually in Figure R1 below. The figure presents a scatterplot of the percent women legislators plotted against the quota threshold. (Country-years that lack quotas are omitted from the figure.) If the percent women legislators matches the quota threshold, the dot will fall along the diagonal line. If the percent women legislators is less than the quota threshold, the dot will fall below the diagonal line (as is most common). And if the percent women legislators is greater than the quota threshold, the dot will fall above the diagonal line. As demonstrated in the figure, the percent women legislators is frequently lower than the quota threshold. Thus, combinations such as 2.5% women legislators and a 20% quota (as occurred in Paraguay 1998-2002) or 9% women legislators and a 50% quota (DRC in 2012-2015) can exist in the real world.

We also have reasons to expect that quota laws can have effects on government spending that are independent of their effects on the percent women legislators. In other words, quota laws do not only operate through their impact on the percent women legislators (i.e., co-creating the effects), but can also have independent effects on government spend-

Figure 16: Percent women legislators plotted against the quota threshold



ing (Clayton and Zetterberg 2018). For example, gender quotas might be part of the government's larger commitment to representing women's substantive interests. Thus, we would expect quota laws to be correlated with both the percent women legislators and government spending on policy issues important to women—hence their inclusion as a control variable in order to isolate the independent effects of women legislators on government spending.

Controlling for the effects of quotas is also important, as there are questions about whether women who are elected through a quota, so-called "quota women," are able to effectively perform their legislative duties or whether they are marginalized in legislative institutions (Franceschet and Piscopo 2008; Zetterberg 2008). While quotas might increase the percent women elected, they could also inhibit women's ability to influence legislative outcomes if their fellow legislators do not view them as equals. In this case, the "origin" of women's representation (elected via a quota or not) could, in fact, affect their ability to shape government spending. Our covariates control for the presence of a quota law, as well as the strength of the quota law (e.g., presence of placement mandates, strong enforcement

mechanisms, and a high threshold) to account for the possible direct and indirect effects quotas might have on government spending.

Even though there are reasons to think that imbalance will have little effect on our substantive predictions given the discussion above, we still wanted to test the robustness of our results to this. As shown in in Figures 17 through 22, which replicate the partial dependence plots shown in Figures 3 through 6 in the main paper, the results are nearly identical when dropping these five quota variables.

Figure 17: VIP for Education (Excluding Quota Variables)

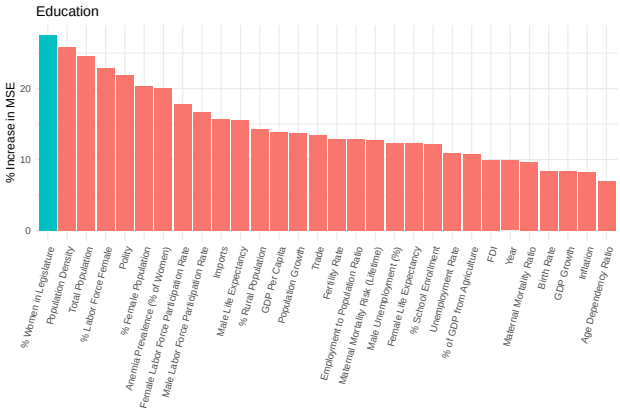


Figure 18: VIP for Health (Excluding Quota Variables)

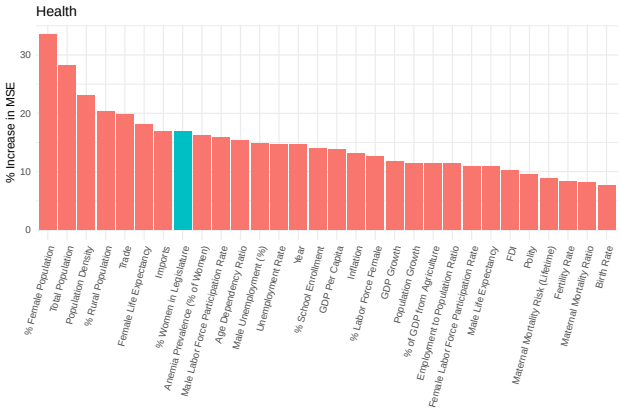


Figure 19: VIP for Defense (Excluding Quota Variables)

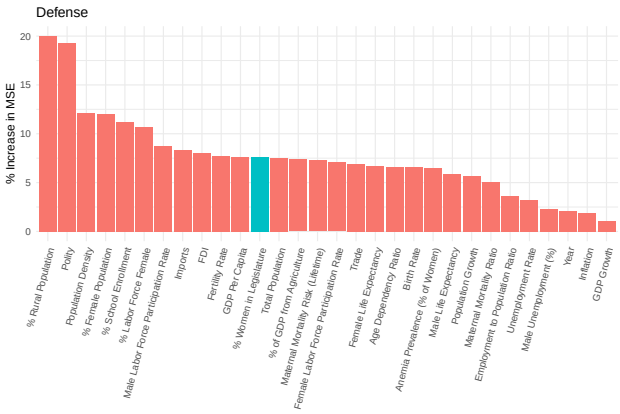


Figure 20: PDPs (Excluding Quota Variables)

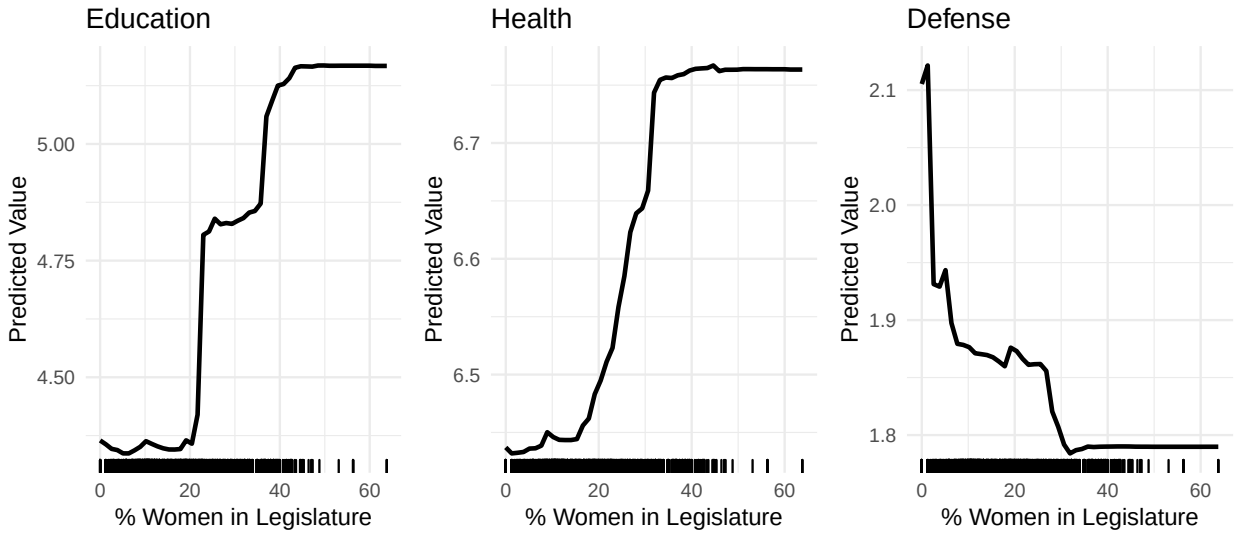
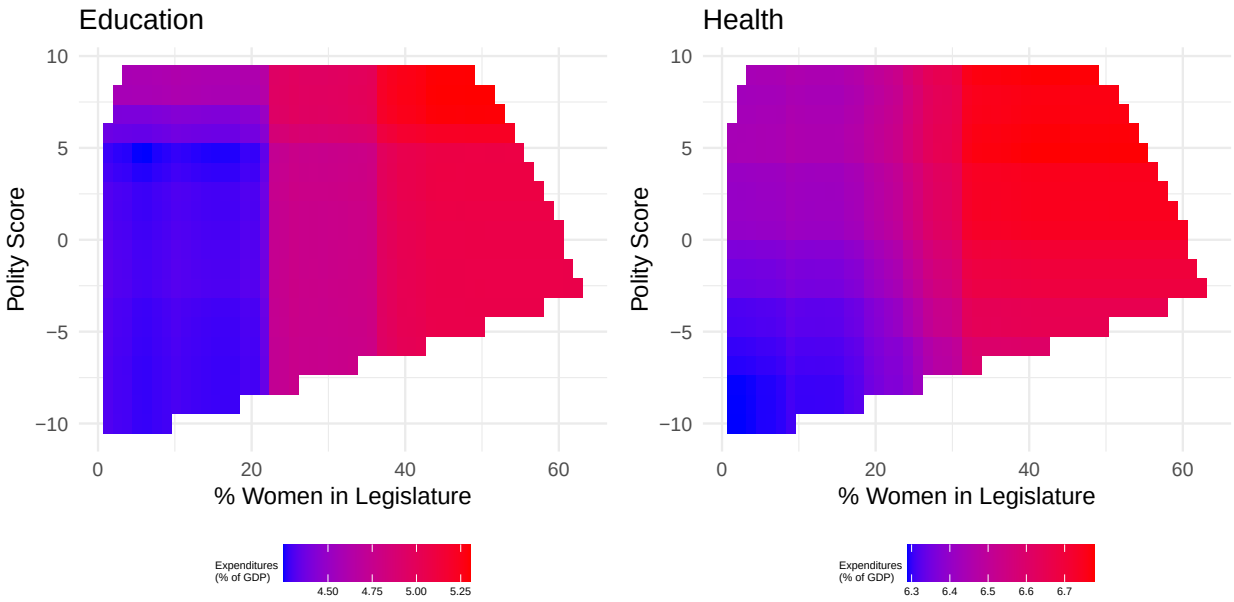
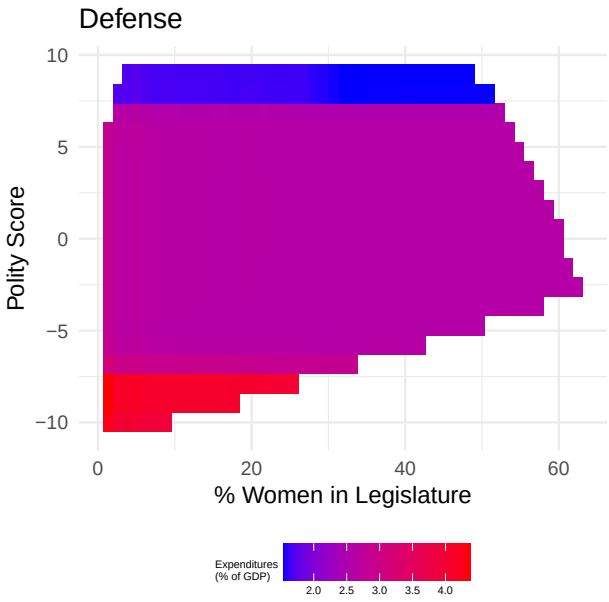


Figure 21: Interactions Between Percent Women and Democracy (Excluding Quota Variables)



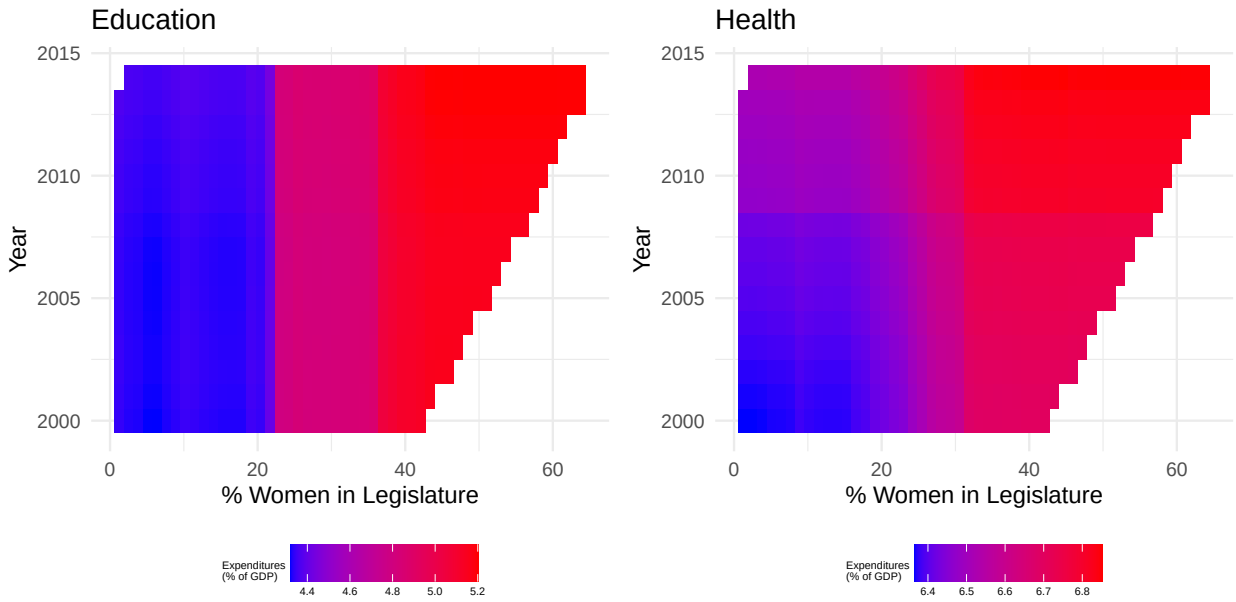
(a) Education

(b) Health



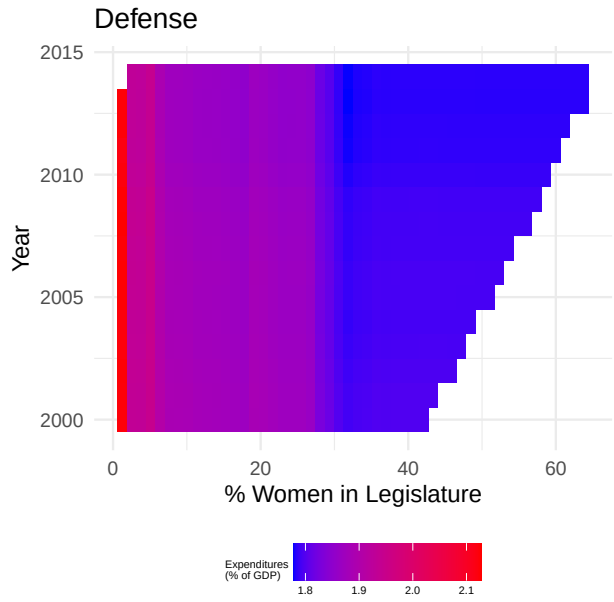
(c) Defense

Figure 22: Interactions Between Percent Women and Year (Excluding Quota Variables)



(a) Education

(b) Health



(c) Defense

6 Machine Learning-Informed Parametric Testing

Researchers who are testing hypotheses about possible non-linear relationships may wish to include measures of uncertainty, such as confidence intervals, in their analysis. Since machine learning models, like the Random Forest models used in our study, do not provide measures of uncertainty, we present an approach to specifying parametric models based on the results of non-parametric models and accompanying plots. To illustrate this approach, we use our model for education spending. We recommend referencing our companion do-file called "guide for machine learning-informed parametric testing.do" for a step-by-step guide with annotated code.

Step 1: Create knots based on partial dependence plot results using linear spline approach

If we take an exclusively parametric approach to estimating models with non-linear relationships, then we need to explicitly specify those knots or breakpoints. The machine learning-based methods discussed in the main body of this paper present a data-driven way to examine possible non-linear relationships. We recommend that users identify breakpoints or shifts in the nature of the relationship between the dependent variable and a given explanatory variable using partial dependence plots (see Figure 4 in the main text for an example).[5]

Based on PDP results for our education spending model, we find evidence of two critical mass intervals in the percentage of women: 20-21% and 38-41%. In order to avoid an overly complex parametric model specification, we create knots at the middle points for each of these intervals: 20.5% and 39.5%. We use the `mkspline` command in Stata to create two knots at 20.5% and 39.5% women in the legislature, which in turn generates three variables that will replace the original $x_{pct.w}$ in our parametric model. The first

[5]Users could also use ICE or ALE plots, but we think PDPs are the simplest to interpret for the purpose of identifying inflection points.

variable contains all values below the first breakpoint, the second includes values equal to and between both breakpoints, and the third includes all values of $x_{pct.w}$ above the second breakpoint. For further detail about this step, see the companion do-file.

Step 2: Select top control variables based on results from the variable importance plot

Unlike parametric models, non-parametric ones, such as Random Forests, do not have degrees of freedom constraints. Consequently, we were able to include control variables without regard for their number. Of course, this is not the case with parametric models, like the one we estimate below. In order to narrow down the number of covariates, we recommend selecting a subset of variables based on the variable importance plots (see Figure 3a in the main file). For this example, we select the 11 predictors (in addition to the percentage of women in the legislature) which do the best job at reducing prediction error compared to their randomly permuted counterparts.

Step 3: Estimate parametric model with explanatory variable in original form (eg. % women in parliament)

Users should estimate the parametric model that is best suited for their data. For our panel data, we started by estimating a fixed effects model with year dummies. As displayed in Table 1, the positive and statistically significant coefficient on the percentage of women in the legislature tells us that as the number of women in the legislature increases, so do the percentage that governments spend on education. Of course, these results do not tell us whether or not a non-linear relationship is at play, such as a critical mass effect. In Step 4, we show how to directly test for a non-linear relationship using parametric modeling.

Table 1: Education Spending

Model Type	**Tested Breakpoints**	**Results: β (SE)**	**Equivalency of Coefficients**	**AIC**
Two-way FE	None	$x_{pct.w}$: 0.013*** (0.004)	NA	4312.83
	20.5%, 39.5%	$x_{pct.w} < 20.5$: -0.001 (0.006) $20.5 <= x_{pct.w} <= 39.5$: 0.029*** (0.007) $x_{pct.w} > 39.5$: 0.007 (0.020)	Low vs. Middle: 9.90*** Middle vs. High: 0.89 Low vs. High: 0.15	4305.613
Linear regression	None	$x_{pct.w}$: .047(.003)***	NA	6826.65
	20.5%, 39.5%	$x_{pct.w} < 20.5$: .010(.007) $20.5 <= x_{pct.w} <= 39.5$: .089(.007)*** $x_{pct.w} > 39.5$: -.002(.022)	Low vs. Middle: 44.21*** Middle vs. High: 13.06*** Low vs. High: 0.29	6786.295
Lagged DV	None	$x_{pct.w}$: .004(.001)***	NA	2169.185
	20.5%, 39.5%	$x_{pct.w} < 20.5$: .001(.003) $20.5 <= x_{pct.w} <= 39.5$: .008(.003)*** $x_{pct.w} > 39.5$: -.008(.008)	Low vs. Middle: 2.14 Middle vs. High: 2.76* Low vs. High: 1.10	2169.797
Fixed effects	None	$x_{pct.w}$: .016(.004)***	NA	4313.003
	20.5%, 39.5%	$x_{pct.w} < 20.5$: .002(.006) $20.5 <= x_{pct.w} <= 39.5$: .032(.007)*** $x_{pct.w} > 39.5$: .011(.020)	Low vs. Middle: 10.31*** Middle vs. High: 0.88 Low vs. High: 0.17	4305.364
Random effects	None	$x_{pct.w}$: .019(.004)***	NA	4979.412
	20.5%, 39.5%	$x_{pct.w} < 20.5$: .005(.005) $20.5 <= x_{pct.w} <= 39.5$: .037(.007)*** $x_{pct.w} > 39.5$: .010(.019)	Low vs. Middle: 12.26*** Middle vs. High: 1.50 Low vs. High: 0.05	4970.939

Note: "Results" column displays $x_{pct.w}$ coefficient(s) with standard errors in parentheses. Controls included but not shown. "Equivalency of Coefficients" column display F-statistics (χ^2 for the random effects results) that the contrast of the two effects is equivalent to 0. "AIC" column displays Akaike's information criterion. p <.1 = *, p<.05 = **, p < .01 = ***.

Step 4: Estimate model with explanatory variable using linear splines (eg. % women < 20.5, 20.5 <= % women <= 39.5, % women > 39.5)

In order to assess the presence of non-linear dynamics, we test the relationship between the percentage of women in the legislature and education spending as a piecewise linear function using linear splines. Users can do this by substituting in the variables created in Step 1 for the original percentage of women variable in your model specification. Recall that based on the results we observed in the PDP for percent women and education spending, we expect that women's representation will only have a substantive effect on education spending between the range of 20.5% and 39.5% women in the legislature. If the coefficient on *wom2* is positive and statistically significant and *wom1* and *wom3* are not statistically significant, then this would suggest that there is non-linear effect—the percentage of women in the legislature only matters for how governments spend on education when the percentage of women in the legislature is between 20.5% and 39.5%. As shown in Table 1, we find evidence of a critical mass interval effect that is statistically significant, allowing for the inclusion of measures of uncertainty in our analysis of the non-linear relationship between women's representation and education spending.

Step 5: Test equivalency of breakpoint slopes to evaluate differences in breakpoints

In order to directly assess the presence of non-linear effects, we recommend that users test the equivalency of the slopes for each breakpoint region.[6] For our example, we use the `test` command in Stata, which is a post-estimation hypothesis test. With it, we perform

[6]For users interested in plotting the main effect of their model with linear splines to assess presence of non-linear effects, see (Royston 2013) and (Rios-Avila 2021).

a test that the difference between the two coefficients is equal to 0.[7] As shown in Table 1, we find that the difference in the slopes for the low breakpoint region (below 20.5% women in the legislature) and the middle breakpoint region (20.5%-39.5% women in the legislature) are different from one another, providing support for non-linear dynamics between women's representation and education spending. We find that the joint test for the difference between the middle and high (above 39.5% women in the legislature) breakpoint regions is not statistically significant, therefore, we cannot reject the null hypothesis that the difference between the slopes is equal to 0. Taken together, a conservative interpretation of these post-estimation tests is that there is strong support for a critical mass effect and mixed support for a critical mass interval effect regarding the relationship between women's representation and government spending on education.

Step 6: Use information criteria to assess whether piecewise estimator is a better fit given the data compared to model with x in its original form

Next, we suggest that users assess the relative fit of the piecewise linear function model given the data compared to the model with women's representation in its original form with criterion information statistics. For our example, we use Akaike's information criterion (AIC).[8] The smaller the AIC statistics, the better the relative fit of the model given the data. Given that we are testing the differences in relative fit based on form of an explanatory variable as apposed to a restricted versus unrestricted model, we expect to observe somewhat small differences between the test statistics across models. Based on the AIC statistics shown in Table 1 for the fixed effects model with year dummies, we find that the model with linear splines is (slightly) the more preferred model. Users might

[7] There are a number of specification options for the `test` command. Please reference '1parametric estimate.do" for the code for how we conducted the test.

[8] Stata users can use the `estat summarize` code to do this, but there are a number of ways these statistics can be calculated.

also consider how the differences in information criterion statistics vary across alternative model specifications. Across all of model specifications, do the information criterion statistics tend to be smaller for the linear piecewise function models or the ones with the original form of the explanatory variable?

Step 7: Specify models using alternative estimators

As a final robustness check, which will be familiar to most users, we recommend alternative model specifications/estimators in order to see how the results hold or change. As displayed in Table 1 we find that our fundamental finding of a critical mass interval effect of women's representation on education spending holds across a series of model specifications, increasing our confidence that a non-linear relationship does indeed exist. Last, we present results for health spending—-assuming breakpoints at 15% and 30% based on results in the main paper—in Table 2, and results for defense spending (here, assuming only a single breakpoint at 30% given our earlier results) in Table 3.

Table 2: Health Spending

Model Type	**Tested Breakpoints**	**Results: β (SE)**	**Equivalency of Coefficients**	**AIC**
Two-way FE	None	$x_{pct.w}$: .025(.005)***	NA	4784.647
	15%, 30%	$x_{pct.w} < 15$: -.005(.010)	Low vs. Middle: 8.02***	4772.779
		$15 <= x_{pct.w} <= 30$: .034(.008)***	Middle vs. High: 0.18	
		$x_{pct.w} > 30$: .039(.010)***	Low vs. High: 12.63***	
Linear regression	None	$x_{pct.w}$: .044(.004)***	NA	8183.387
	15%, 30%	$x_{pct.w} < 15$: .025(.013)*	Low vs. Middle: 2.25	8185.101
		$15 <= x_{pct.w} <= 30$: .056(.010)***	Middle vs. High: 0.71	
		$x_{pct.w} > 30$: .039(.013)***	Low vs. High: 0.64	
Lagged DV	None	$x_{pct.w}$: .004(.001)***	NA	3059.557
	15%, 30%	$x_{pct.w} < 15$: .008(.004)*	Low vs. Middle: 1.33	3061.989
		$15 <= x_{pct.w} <= 30$: .000(.003)	Middle vs. High: 1.21	
		$x_{pct.w} > 30$: .007(.004)*	Low vs. High: 0.01	
Fixed effects	None	$x_{pct.w}$: .046(.005)***	NA	4965.724
	15%, 30%	$x_{pct.w} < 15$: .024(.010)**	Low vs. Middle: 5.85**	4962.978
		$15 <= x_{pct.w} <= 30$: .059(.008)***	Middle vs. High: 0.84	
		$x_{pct.w} > 30$: .046(.010)***	Low vs. High: 2.68	
Random effects	None	$x_{pct.w}$: .045(.004)***	NA	5766.955
	15%, 30%	$x_{pct.w} < 15$: .027(.010)***	Low vs. Middle: 3.98**	5766.464
		$15 <= x_{pct.w} <= 30$: .055(.008)***	Middle vs. High: 0.35	
		$x_{pct.w} > 30$: .046(.010)***	Low vs. High: 2.35	

Note: "Results" column displays $x_{pct.w}$ coefficient(s) with standard errors in parentheses. Controls included but not shown. "Equivalency of Coefficients" column display F-statistics (χ^2 for the random effects results) that the contrast of the two effects is equivalent to 0. "AIC" column displays Akaike's information criterion. p <.1 = *, p<.05 = **, p < .01 = ***.

Table 3: Defense Spending

Model Type	**Tested Breakpoints**	**Results: β (SE)**	**Equivalency of Coefficients**	**AIC**
Two-way FE	None	$x_{pct.w}$: -.015(.008)*	NA	13806.5
	30%	$x_{pct.w} < 30$: -.014(.010) $x_{pct.w} >= 30$: -.020(.019)	Low vs. High: 0.07	13808.42
Linear regression	None	$x_{pct.w}$: -.020(.005)***	NA	15509.32
	30%	$x_{pct.w} < 30$: -.020(.007)*** $x_{pct.w} >= 30$: -.019(.017)	Low vs. High: 0.00	15511.31
Lagged DV	None	$x_{pct.w}$: .000(.003)	NA	15509.32
	30%	$x_{pct.w} < 30$: -.001(.004) $x_{pct.w} >= 30$: .006(.009)	Low vs. High: 0.44	11002.41
Fixed effects	None	$x_{pct.w}$: -.016(.008)**	NA	13811.66
	30%	$x_{pct.w} < 30$: -.017(.010)* $x_{pct.w} >= 30$: -.012(.019)	Low vs. High: 0.05	13813.61
Random effects	None	$x_{pct.w}$: -.018(.007)**	NA	14416.96
	30%	$x_{pct.w} < 30$: -.020(.009)** $x_{pct.w} >= 30$: -.013(.018)	Low vs. High: 0.08	14418.88

Note: "Results" column displays $x_{pct.w}$ coefficient(s) with standard errors in parentheses. Controls included but not shown. "Equivalency of Coefficients" column display F-statistics (χ^2 for the random effects results) that the contrast of the two effects is equivalent to 0. "AIC" column displays Akaike's information criterion. p <.1 = *, p<.05 = **, p < .01 = ***.

References

Apley, Daniel W and Jingyu Zhu. 2016. "Visualizing the effects of predictor variables in black box supervised learning models." *arXiv preprint arXiv:1612.08468* .

Biecek, Przemysław. 2018. "DALEX: explainers for complex predictive models in R." *The Journal of Machine Learning Research* 19(1):3245–3249.

Clayton, Amanda and Pär Zetterberg. 2018. "Quota shocks: Electoral gender quotas and government spending priorities worldwide." *The Journal of Politics* 80(3):916–932.

Franceschet, Susan and Jennifer M. Piscopo. 2008. "Gender Quotas and Women's Substantive Representation: Lessons from Argentina." *Politics & Gender* 4(3):393–425.

Günther, Frauke and Stefan Fritsch. 2010. "neuralnet: Training of neural networks." *The R journal* 2(1):30–38.

Hanmer, Michael J and Kerem Ozan Kalkan. 2013. "Behind the curve: Clarifying the best approach to calculating predicted probabilities and marginal effects from limited dependent variable models." *American Journal of Political Science* 57(1):263–277.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The elements of statistical learning*. 2 ed. Springer series in statistics New York.

King, Gary and Langche Zeng. 2006. "The dangers of extreme counterfactuals." *Political Analysis* 14(2):131–159.

Muggeo, Vito MR et al. 2008. "Segmented: an R package to fit regression models with broken-line relationships." *R news* 8(1):20–25.

Rios-Avila, Fernando. 2021. "Estimation of marginal effects for models with alternative variable transformations." *The Stata Journal* 21(1):81–96.

Royston, Patrick. 2013. "marginscontplot: Plotting the marginal effects of continuous predictors." *The Stata Journal* 13(3):510–527.

Seufert, Jacqueline D. and Alex Afanasev. N.d. "NeuralNetworkVisualization.". **URL:** *https://github.com/AlexAfanasev/NeuralNetworkVisualization*

Wood, SN. 2018. "Mixed GAM computation vehicle with automatic smoothness estimation. R package version 1.8–12.".

Zetterberg, Pär. 2008. "The Downside of Gender Quotas? Institutional Constraints on Women in Mexican State Legislatures." *Parliamentary Affairs* 61(3):442–460.