Women and The Environment
A Bayesian Logistic Regression Analysis

Kaylee B. Hodgson

December 10, 2018

1 Introduction

Subordination of women in the home seems to be indicative of a violent culture that does not treat its valuable assets well, including the environment. This analysis tests the rising theory that countries that are more concerned with gender equality are also more concerned with the wellbeing of their country in general, and therefore are more likely to exert higher levels of concern for environmental wellbeing. I predict that countries that more systematically subordinate women in the household also experience lower air quality, lower levels of environmental protection, and higher risks from environmental factors.

The purpose of the analysis is to identify whether higher levels of gender inequality in the household are significantly tied to poor environmental wellbeing outcomes. I employ a Bayesian method for logistic regression models to test this theory.

2 The Data

2.1 Introduction of the Variables

All variables included in this analysis are measured on the national level. The main independent variable of interest is the "Syndrome" from The WomanStats Project's database [9], which indicates a country's level of systematic subordination of women in the household, mapped in Figure 1. The variable is scaled from 0 to 16 (where 16 indicates the worst levels of subordination), using an algorithm that combines 11 indicators of the subordination of women in the home: 1) the prevalence of patrilocal marriage, 2) the prevalence of brideprice or dowry, 3) the prevalence and legality of polygyny, 4) the presence of counsin marriage, 5) the age of marriage for girls, 6) the laws and practices surrounding women's property rights, 7) the presence of son preferences or sex ratio alteration, 8) the presence of inequity in family law or custom that favors males, 9) the overall level of violence against women in society, 10) the presence of societal sanction for femicide, and 11) whether there is legal exoneration for rapists who offer to marry their victims [9].

There are, of course, other factors that affect countries' performance on environmental indicators. I include seven control variables in the regression analyses, which may also explain the variation in the environment variables. The control variables are: 1) the percentage of the population that lives in urban areas [10], 2) the aggregated civilization identification based on Samuel Huntington's civilizational identity [6], 3) colonial heritage (dichotomous, indicates whether a country was colonized) [5], 4) the percentage of land that is arable [10], 5) the number of unique land neighbors [8], 6) the level of ethnic fractionalization [1], and 7) the level of religious fractionalization [1].

Finally, I include six indicators of environmental wellbeing as the dependent variables. These six indicators are a result of an exploratory factor analysis performed in Hudson et al. (2018), where they combine multiple variables into a single variable, and keep some separately [5]. Multiple indicators are from the Social Progress Index: Foundations and Wellbeing, Outdoor Air Pollution Attributable Deaths, Household Indoor Air Pollution Attributable Deaths, Wastewater Treatment, Biodiversity and Habitat, and Greenhouse Gasses [7]. I also include some environmental indicators from the Environmental Performance Index: Environmental Performance Index (overall scale), Air Quality, and Water and Sanitation [2]. Finally, I include the Global Climate Risk Index from German Watch's database [4]. These are commonly used variables for

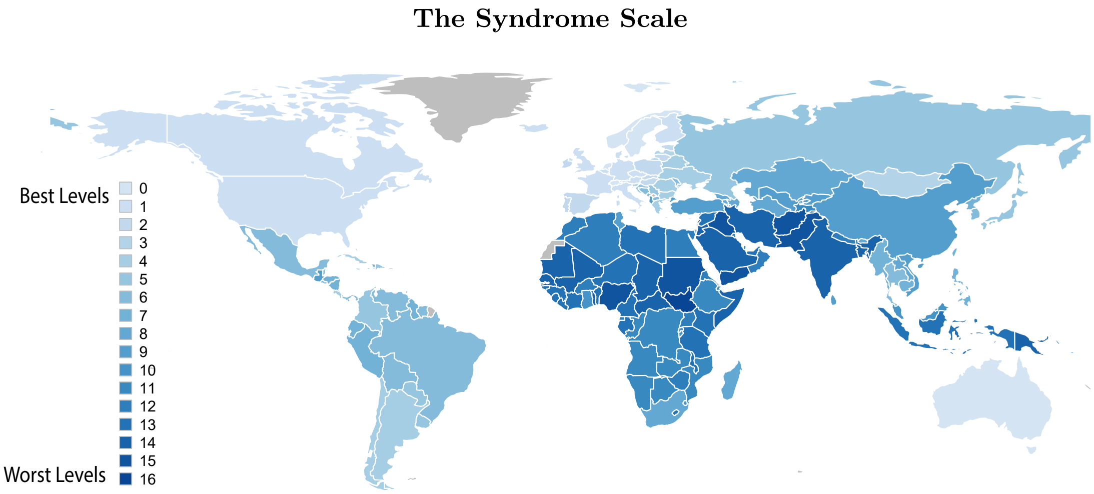

Figure 1: Map of countries' levels of systematic subordination of women at the household level

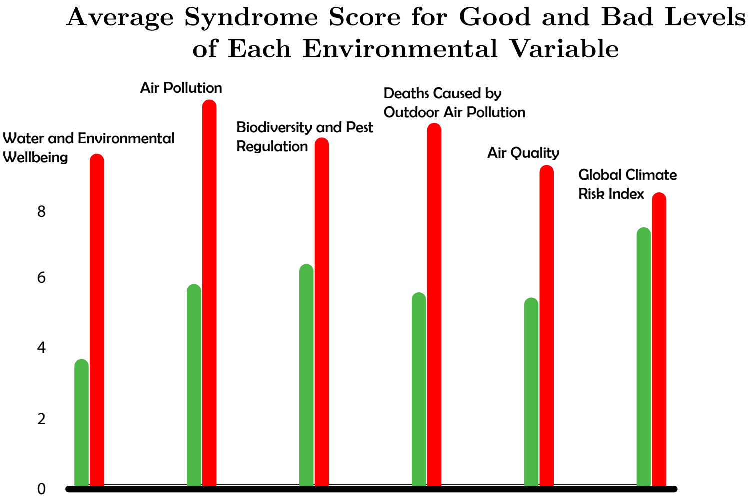

Figure 2: Red bars represent worse levels of each environmental indicator, green for better levels.

indications of environmental wellbeing, and comprehensively cover the areas of environmental health that I predict are affected by countries' levels of systematic subordination. These original eleven variables are condensed into: 1) Water and Environmental Wellbeing, 2) Air Pollution, 3) Biodiversity and Pest Regulation, 4) Deaths caused by outdoor air pollution, 5) Air Quality, and 6) Global Climate Risk Index.

2.2 Standardization of the Variables

Because I am interested in exploring how levels of subordination affect whether a country performs poorly on environmental indicators, I first dichotomize all of the environmental variables, where 0 indicates "good" levels of the environmental indicator and 1 indicates "bad" levels. The variables are generally split at the mean, or at some "natural" split in distributions if the distribution of the original variable is bimodal. The average values of the Syndrome are plotted for each level of the six environmental variables in Figure 2, which clearly shows a pattern where countries with worse environmental outcomes are, on average, those with higher levels of subordination of women.

This dichotomization of each of the dependent variables and their consistent directional meaning allows me to set common prior distributions of each independent variables' coefficients in each model. Therefore, the coefficient corresponding to each independent variable will have the same prior distribution for each of the six models.

In addition to standardizing the environmental indicators, I also standardize the independent variables following the methods put forward in Gelman et al. (2014) [3]. I first scale all of the continuous independent variables to have a mean of 0 and a standard deviation of 0.5. Then I take the colonization status variable, which is dichotomous, and shift it so that the variable also has a mean of 0 and a standard deviation of approximately 0.5. This is done by finding the proportion of 0's (0.13) and proportion of 1's (0.87), then redefining the 0's as 0.87 and the 1's as -0.13. In order standardize the civilization variable, which is a categorical variable with 4 levels, I created 3 dichotomous variables which indicated 3 of the levels, then standardized each of those in the same way as colonial status.

3 Methods

In the preliminary analyses, I built a hierarchical logistic regression model, with hyperparameters. However, that model had issues that prevented proper estimation in many different MCMC proposal schemes. One of the main issues is that the covariance matrix for the posterior draws was not positive definite. There are multiple explanations for this, including not having enough data for the number of parameters and the number of levels that those are parameters are being estimated. I would recommend that as more data becomes available, that this analysis be redone using a hierarchical model. However, for this analysis, I remove the hierarchy and get much better results.

3.1 The Model

I implement Bayesian methods to analyze the data in six different logistic regression models. The same independent variables are used in each model (the Syndrome and the seven control variables), but I analyze a different dichotomous environmental indicator as the control variable in each one.

The model is structured as follows:

$$\begin{aligned} &y_{im}|p_{im} \sim \text{ Bernoulli}(p_{im}),\ \text{logit}(p_{im}) = \beta_{m0} + \sum_{j=1}^{7} x_{imj}\beta_{mj} + \sum_{k=1}^{3} x_{imk}\alpha mk + \epsilon_{im} \\ &\beta_0 \sim \text{ Cauchy}(0,\sigma),\ \beta_j \sim \text{ Cauchy}(0,\theta),\ \alpha_k \sim \text{ Cauchy}(\mu,\phi) \\ &\sigma = 5, \theta = 2.5, \mu = \phi = 2 \end{aligned} \tag{1}$$

where $i = 1, ..., n$ corresponds with the country (or observation), $m = 1, ..., 6$ corresponds with the regression model (one model for each environmental variable), $j = 1, ..., 7$ corresponds with the independent variables I have less prior knowledge about, and $k = 1, ..., 3$ corresponds with the independent variables I have more prior knowledge about. The Cauchy distribution is chosen over the normal or the t distributions because the Cauchy is more flexible with extreme values since there is more room in the tails. Additionally, the values for the β coefficients are non-informative, and allow for the coefficients to be either negative or positive with an equal probability. The α coefficients correspond with the variables the I have more prior information for: Terrain, Urbanization, and Syndrome. The prior values for these variables indicate an expectation that as each of these increase (there is more arable land, higher proportion of the population living in urban areas, and worse subordination of women), the environmental performance of the county gets worse, on average. The scale parameter is chosen for the β priors based on suggestions from Gelman et al. (2014) [3].

The posterior distribution is given, up to proportationality:

$$\begin{aligned} &p(.|y) \propto p(\beta_0, \beta_1, ..., \beta_6, \alpha_1, ..., \alpha_4, \sigma, \theta, \mu, y) \propto \left[\prod_{i=1}^{n} \left(\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right)^{y_i} \left(\frac{1}{1+e^{X_i\beta}}\right)^{1-y_i}\right] \\ &\left[\frac{1}{\pi\sigma\left[1+(\beta_0/\sigma)^2\right]}\right] \left[\prod_{j=1}^{6} \frac{1}{\pi\theta\left[1+(\beta_i/\theta)^2\right]}\right] \left[\prod_{k=1}^{4} \frac{1}{\pi\phi\left[1+((\alpha_i-\mu)/\phi)^2\right]}\right] \end{aligned} \tag{2}$$

Posterior draws are pulled from this posterior distribution in the computation for this analysis.

3.2 Computational Methods

I use Markov chain Monte Carlo (MCMC) methods to pull posterior draws from the posterior distribution. I implement a multivariate update scheme, with a proposal density. The algorithm proceeds as follows:

- Set initial values for the parameters.
- Use normal approximation to obtain the hessian estimate for the covariance matrix ($\mathbf{S}$) and the optimized parameter estimates for the mean vector ($\mathbf{m}$).
- For each iteration:
  - Set the proposal: $\mathbf{P} = \mathbf{m} + (\text{chol}(\mathbf{S}))'\text{rnorm}(11)$.
  - Use a Metropolis-Hastings update to determine whether to keep the last samples or update with the proposal, using the proposal density: $f(\mathbf{P}) = -0.5(\mathbf{P} - \mathbf{m})'\mathbf{S}^{-1}(\mathbf{P} - \mathbf{m})$.

I run three chains of 60,000 each, with three different sets of initial values: one where they all start at 0, one where all of the parameters start at the mean estimates from their prior distributions, and one where they all start at their frequentist estimates. The first 20,000 draws for each chain are removed as burn-in, then I remove every fifth draw to thin the posterior draws. The chains are then combined to estimate the parameters.

This process is repeated for each of the six logistic regression models. The computation time for each model and for each chain is under 5 seconds, indicating that the MCMC function is very efficient.

4 Model Diagnostics

I verify the model by running both visual and statistical diagnostics on the posterior draws. I display the results for the Syndrome coefficient specifically since that is the main independent variable of interest in this analysis. Figure 3 plots the posterior draws from the MCMC algorithm for the Syndrome coefficient for each environmental variable's logistic regression model. The plot includes the three chains run from the three different sets of starting values. There is some cause for concern specifically for the Water and Environmental Wellbeing model, where the chains appear to have all individually converged, but not to the same spot. The model for Air Quality is also concerning for the same reason, although the differences are not as dramatic. These different places of conversion indicate that the posterior draws for these models are too sensitive to the starting values chosen. The other coefficients follow a similar pattern, where many of the plots look good, but some are also sensitive to the initial values.

I also calculate the acceptance rates, effective samples sizes, and $\hat{R}$ values for each of the coefficients in each model. In Table 1 I report these diagnostics for the Syndrome. The acceptance rate for the first two models looks decent, but is low for the other four. Note that the acceptance rates are the same for every coefficient within each model because my MCMC updates are multivariate. I turn to effective sample size and find that the Air Pollution model performs best in that regard. Low effective sample sizes indicate that there is too much autocorrelation between the draws. The very low sample size for the Water and Environmental Wellbeing model is especially concerning. Finally, I look at the $\hat{R}$ values, which are calculated with each of the three chains split in half. $\hat{R}$ values should be around 1 and indicate whether the chains (and within the chains) have converged to the same place as each other. The most concerning value, the $\hat{R}$ for the Water and Environmental Wellbeing model is unsurprising given the plot of the posterior draws in Figure 3, which shows that each of the chains converged to different places.

5 Results

I estimate a total of 66 parameters in this analysis, 11 parameters for each of the 6 regression models. While I estimate all of these in the analysis, I mainly report the results in terms of the Syndrome coefficient estimate,

MCMC Posterior Draws for Syndrome coefficient in each Environmental Model

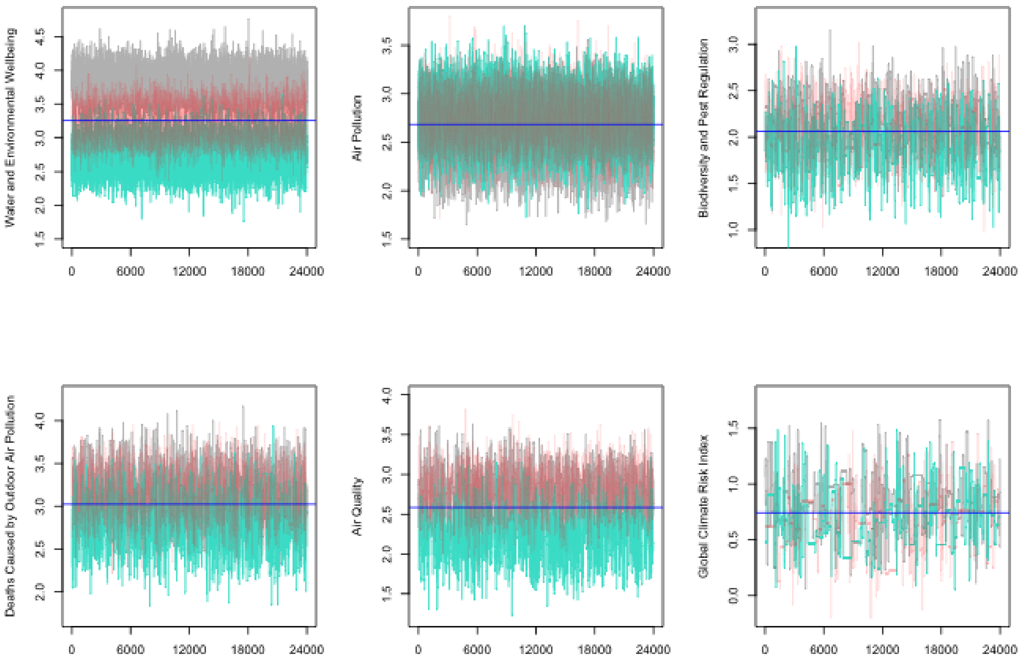

Figure 3: Draws are plotted for each of the 3 chains, after removal of burn-in and thinning.

since the main question in the analysis is how the subordination of women affects environmental outcomes. The first plot in Figure 4 shows the Syndrome coefficient for each of the six models, as well as the 95% and 99% credible intervals. As indicated by the estimates, Syndrome appears to have a significant effect on all of the environmental variables at both the 95% and the 99% level.

Since all of the estimates are positive, I can conclude that countries with higher levels of subordination of women in the home are significantly more likely to experience poor environmental outcomes.

These results are fairly consistent with the logistic regression results using a frequentist analysis. The Syndrome coefficient estimates and their 95% and 99% confidence intervals are plotted next to the Bayesian analysis results in Figure 4. From the plot, we can see that there are several advantages in the conclusions of the Bayesian analysis. First, while the estimates are similar in both analyses, the credible intervals are much tighter than the confidence intervals. This results in different significance conclusions for the Global Climate Risk Index model. In the Bayesian analysis, we conclude that the Syndrome significantly affects these index scores, but we cannot conclude that from the frequentist model. Additionally, the Bayesian approach has the obvious advantage in that we can make probability statements regarding the coefficients since we are

Table 1: Diagnostics for Syndrome coefficient

	Syndrome Coefficient Diagnostics		
Model	Acceptance Rate	Effective Sample Size	$\hat{R}$
Water and Environmental Wellbeing	0.34	53.58	2.77
Air Pollution	0.34	2615.55	1.10
Biodiversity and Pest Regulation	0.04	666.72	1.12
Deaths Caused by Outdoor Air Pollution	0.10	667.66	1.31
Air Quality	0.09	357.73	1.67
Global Climate Risk Index	0.01	325.19	1.07

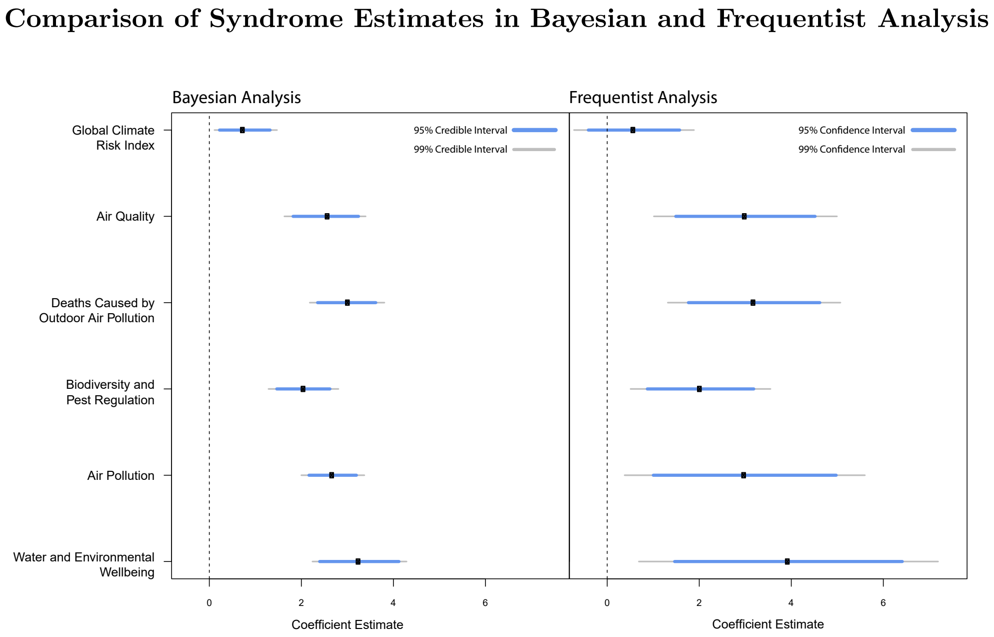

Figure 4: Coefficients have tighter intervals in the Bayesian analysis, and more are significant.

treating them as random variables.

6 Model Sensitivity

I verify the sensitivity of the model by specifying both different prior values and prior distributions. I choose both a less informative prior and a more informative prior and compare the results and diagnostics with the model outcomes in the previous section. The priors chosen for these two models are:

Less Informative Model

$$\beta_0 \sim \text{ Cauchy}(0,\sigma),\ \beta_j \sim \text{ Cauchy}(0,\sigma),\ \alpha_k \sim \text{ Cauchy}(0,\sigma) \quad (3)$$
$$\sigma = 5$$

More Informative Model

$$\beta_0 \sim \text{ Normal}(0,\sigma),\ \beta_j \sim \text{ Normal}(0,\theta),\ \alpha_k \sim \text{ Normal}(\mu,\phi) \quad (4)$$
$$\sigma = 5, \theta = 2.5, \mu = \phi = 2$$

The less informative priors have the same distributions as the original model, but with different prior values. All of the means are set to 0, but the scale parameter is 5 for all of them. This makes the model less informative in two ways: 1) the mean values of 0 imply that we do not have prior knowledge regarding the direction that the variables affect the environmental indicators, and 2) the large scale parameter allows for much more room for variation. In the more informative model, a normal distribution is used instead of the Cauchy, because the normal distribution has less room in the tails, so the prior information constrains the model more.

Again in this analysis, the results for the Syndrome parameter are given to compare the models. The Syndrome coefficients are plotted in Figure 5. In these other two models in this sensitivity analysis, the Syndrome is only significant in predicting environmental indicators in five out of the six models at the 99% level. This is because the credible intervals are a little wider for the less informative model, and the estimates

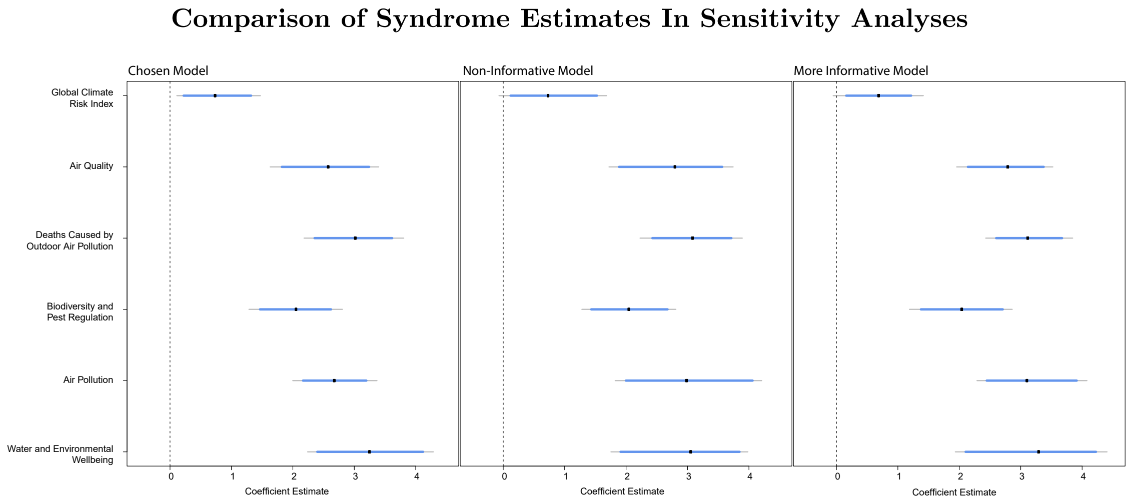

Figure 5: The results for each model are very similar, although for the Global Climate Risk Index, the Syndrome is only significant at the 0.05 level in the two other models.

are more modest in the more informative model. However, there is not too much difference in the results between the analyses, indicating that the model is not terribly sensitive to mispecification in the priors.

The statistical diagnostics are computed for each model and compared in Table 2. The diagnostic comparisons are interesting because they indicate that, in some cases, the model specified for this analysis fits best, but in others, the other model specifications perform better. Specifically, the more informative model appears to fit better for Deaths from Outdoor Air Pollution, Air Quality, and the Global Climate Risk Index.

Table 2: Diagnostic Comparisons for Syndrome coefficient in Sensitivity Analyses:
M1=Chosen Model, M2=Less Informative Model, M3=More Informative Model

	Acceptance Rate			**Effective Sample Size**			**R-hat**		
Model	M1	M2	M3	M1	M2	M3	M1	M2	M3
Water and Environmental Wellbeing	0.34	0.35	0.34	**53.58**	22.33	22.52	3.36	**1.00**	2.13
Air Pollution	0.34	0.35	0.29	**2615.55**	22.73	123.23	**1.32**	5.00	3.73
Biodiversity and Pest Regulation	0.04	0.04	0.04	**666.72**	474.21	426.77	**1.15**	1.39	1.45
Deaths from Outdoor Air Pollution	0.10	0.11	0.11	667.66	429.87	**2120.84**	**1.03**	1.37	1.05
Air Quality	0.09	0.10	0.09	357.73	112.37	**577.40**	**1.01**	1.55	1.05
Global Climate Risk Index	0.01	0.02	0.01	325.19	199.05	**407.38**	1.42	2.24	**1.26**

7 Discussion and Conclusion

The results of this analysis confirm my hypothesis for the study. The one model that should be interpreted more cautiously is the Global Climate Risk Index, because the results for the intervals around the coefficient estimate differed the most across analyses.

Future researchers should work to fill in the missing values in the data used in this analysis, and to collect observations for more countries, so that a hierarchical approach can be used to analyze this data. More work should also be done in improving the convergence and model fit diagnostics.

Overall, the results of this analysis indicate that the subordination of women in the household is a significant determinant of a country's environmental performance. While the diagnostic results are not ideal, because the results regarding the Syndrome's effect on environmental performance remain fairly consistent across different approaches (frequentist vs. Bayesian) and different prior specifications, I conclude that there is sufficient evidence to conclude that the treatment of women is significantly related to a country's environmental performance.

References

[1] Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg. “Fractionalization.” Journal of Economic growth 8, no. 2 (2003): 155-194.

[2] Environmental Performance Index (2016). Yale Center for Environmental Law & Policy. https://epi.envirocenter.yale.edu.

[3] Gelman, Andrew, Hal S. Stern, John B. Carlin, David B. Dunson, Aki Vehtari, and Donald B. Rubin. Bayesian data analysis. Chapman and Hall/CRC, 2013.

[4] Global Climate Risk Index (2016). German Watch. https://www.germanwatch.org/en/cri.

[5] Hudson, Valerie M., Donna Lee Bowen, and Perpetua Lynne Nielsen. The First Political Order: Sex, Governance, and National Security. In Preparation.

[6] Huntington, Samuel. “The clash of civilizations and the making of a new world order.” New York (1996).

[7] Social Progress Index (2016). The Social Progress Imperative. https://www.socialprogress.org

[8] Wikipedia (2018). “List of countries and territories by land borders.” https://en.wikipedia.org/wiki/List_of_countries_and_territories_by_land_borders.

[9] WomanStats Project Database (2018). The WomanStats Project. http://www.womanstats.org.

[10] The World Bank. 2014-2015. World Development Indicators. Washington, D.C.: The World Bank (producer and distributor). http://data.worldbank.org/data-catalog/world-development-indicators