# Using Measures of Women's Peace and Security to Predict Rates of Intimate Partner Violence

STA 264: Regression Analysis

Jason D'Amico

March 11, 2022

# Contents

# 1 Executive Summary

As a student interested in the intersection of technology and the analysis of sociological phenomena (tech for social good, if you will), my primary motivation for taking Regression Analysis with Professor Hoerl was to build up my data analytics skill set such that I could use my knowledge to analyze data sets related to social issues.

When I learned about the final report project, I instantly knew that I wanted to analyze a data set relevant to a social issue. I also have a strong background in gender-based violence prevention advocacy, so I was particularly interested in analyzing a data set related to gender disparities or something of that nature. After discussing with Professor Hoerl, I reached out to a professional contact who works for the Human Rights Data Analytics Group (HRDAG), who recommended me the Georgetown WPS (Women's Peace and Security) Index data set that I ended up using for my final project. To me, this was the perfect data set for my interests, as it is a quantitative measurement of gender-based violence and related phenomena in 170 countries around the globe. Through my analysis, I ultimately hoped to gain skills of how to use data related to a social issue to quantitatively measure the impact of certain aspects of a community on gender-based violence while producing an accurate model that suggests which of the included $x$ variables in the data set have correlation with intimate partner violence.

As for my findings in the project, a logistic transformation of $y$ was determined to produce the best model fit for predicting rates of intimate partner violence based on the sample data. As found by a stepwise fit, the best model (Figure 7) contains the following five terms along with its intercept:

- Education
- Absence of legal discrimination
- Sex ratio at birth
- Absence of legal discrimination * Sex ratio at birth
- (Sex ratio at birth)$^2$

This model had the best fit of the possible $x$ variables to rates of intimate partner violence among all of the models that had low multicolinearity and did not contain any patterns in the residuals. Notable values include $R^2_{Adj} = 0.651$, a ratio of PRESS RMSE to model RMSE of 1.016 and all terms easily passing their $t$ tests with $p$ values less than or equal to 0.001.

# 2 Data source and relevant background

The data source used in this report comes from the Georgetown Institute on Women, Peace and Security and was collected for the purpose of creating a metric called the Women, Peace and Security Index (WPS Index). According to their website, the Institute was founded by Hilary Clinton in 2011 and they "engage in rigorous research, host global convenings, advance strategic partnerships, and nurture the next generation of leaders"[1]. As for the index itself, the index has been published three times, with its first publishing coming in 2017. According a document posted to summarize the 2017 inaugural WPS Index, the Institute indicated that the "primary goal of the index is to accelerate progress on both the international Women, Peace and Security agenda and the [United Nations] Sustainable Development Goals, bringing partners together around an agenda for women's inclusion, justice, and security"[2]. The WPS Index was also intended to "inspire further thought and analysis"[2] as well as encourage a deeper collection of data related to the goals of the Institute. Finally, the report identified several potential stakeholders for the Index and its associated data, including policymakers, academics and the international development community.

There were no data quantity or quality issues in this data set. Since this data set comes from an academic institution and is intended to be a relatively objective measure of women's peace and security globally, there were no missing observations nor any reason to believe that the quality of any of the observations were significantly compromised due to the reputability of each of the data sources.

It should also be noted that many of the observations should be expected to have noise or be inaccurate given the nature of the data; while Georgetown undoubtedly did their due diligence collecting this data as evident by the rigorous data pedigree provided, measures of women's peace and security are inherently difficult to measure. The 11 observed fields, referred to as "indicators", are collected from a variety of sources. Sources include the World Bank, the United Nations and the Gallup World Poll. Some of the data fields are measurable quantitative observations (such as the percentage of representation of women in a given country's parliament), while others are based on polling data (such as the perceived safety of women in a given country). The survey data in particular has potential to introduce noise in the data set.

Our $y$ variable, "Intimate partner violence", is sourced from the UN Women Global Database on Violence against Women. Unfortunately, rates of intimate partner violence is possibly the most difficult of all of the fields to accurately collect, as under-reporting of violence against women is a global problem: despite all possible measures of mitigating the effect of under-reporting, the observed values will never be truly accurate, and there is likely no way of accurately determining how far off they could be. With this in mind, a certain level of inaccuracy should be expected in the final model.

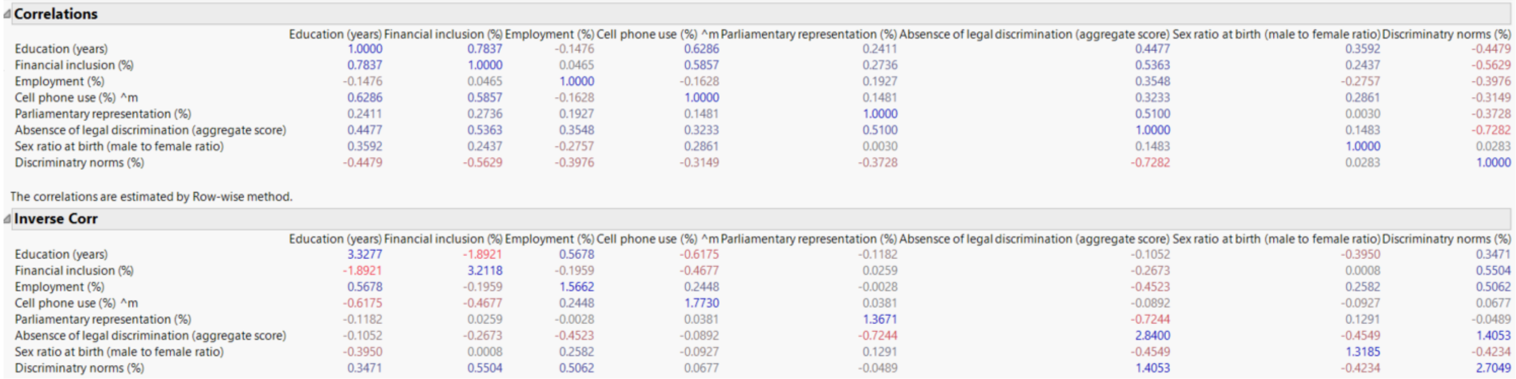
# 3 Analysis

## 3.1 Multivariate plot

To begin our analysis, we will first create multivariate plots for each of the $x$ variables in consideration for the model along with their associated correlation matrix, $X'X$, and the standardized inverse correlation matrix, $(W'W)^{-1}$.

This analysis reveals a couple of interesting relationships between the $x$ variables. First of all, we see some strong evidence for relationships between certain pairings of $x$ variables both in the multivariate plot and in the correlation matrix. The clearest relationship exists between "Education" and "Financial inclusion", both in terms of having the clearest visual relationship in their multivariate plot, the highest correlation between their $x$ variables and the highest correlation between their $\hat{\beta}_*$ values. Another notable apparent relationships between $x$ variables using the same criteria as outlined above is between "Absence of legal discrimination" and "Discriminatory norms": this is the only other relationship that has correlation values close to the above outlined relationship and also holds strong visual evidence for correlation in the multivariate plot. Several other of the plots could possibly show a non-linear correlation between their corresponding variables, but this is difficult to visually identify with confidence.

Interestingly, the VIF values produced by the $(W'W)^{-1}$ matrix are highest for the four different $x$ variables identified above as potentially having correlation with another variable, whereas all other variables appeared to have very small correlation to other variables based on their calculated VIF values. Thus, we can say that these four variables ("Absence of legal discrimination", "Discriminatory norms", "Education" and "Financial inclusion") all are significant contributors to the multicolinearity of the model as a whole and warrant particular attention in further analysis.



**Correlations**

| | Education (years) | Financial inclusion (%) | Employment (%) | Cell phone use (%) ^m | Parliamentary representation (%) | Absensce of legal discrimination (aggregate score) | Sex ratio at birth (male to female ratio) | Discriminatry norms (%) |
|---|---|---|---|---|---|---|---|---|
| Education (years) | 1.0000 | 0.7837 | -0.1476 | 0.6286 | 0.2411 | 0.4477 | 0.3592 | -0.4479 |
| Financial inclusion (%) | 0.7837 | 1.0000 | 0.0465 | 0.5857 | 0.2736 | 0.5363 | 0.2437 | -0.5629 |
| Employment (%) | -0.1476 | 0.0465 | 1.0000 | -0.1628 | 0.1927 | 0.3548 | -0.2757 | -0.3976 |
| Cell phone use (%) ^m | 0.6286 | 0.5857 | -0.1628 | 1.0000 | 0.1481 | 0.3233 | 0.2861 | -0.3149 |
| Parliamentary representation (%) | 0.2411 | 0.2736 | 0.1927 | 0.1481 | 1.0000 | 0.5100 | 0.0030 | -0.3728 |
| Absensce of legal discrimination (aggregate score) | 0.4477 | 0.5363 | 0.3548 | 0.3233 | 0.5100 | 1.0000 | 0.1483 | -0.7282 |
| Sex ratio at birth (male to female ratio) | 0.3592 | 0.2437 | -0.2757 | 0.2861 | 0.0030 | 0.1483 | 1.0000 | 0.0283 |
| Discriminatry norms (%) | -0.4479 | -0.5629 | -0.3976 | -0.3149 | -0.3728 | -0.7282 | 0.0283 | 1.0000 |

The correlations are estimated by Row-wise method.

**Inverse Corr**

| | Education (years) | Financial inclusion (%) | Employment (%) | Cell phone use (%) ^m | Parliamentary representation (%) | Absensce of legal discrimination (aggregate score) | Sex ratio at birth (male to female ratio) | Discriminatry norms (%) |
|---|---|---|---|---|---|---|---|---|
| Education (years) | 3.3277 | -1.8921 | 0.5678 | -0.6175 | -0.1182 | -0.1052 | -0.3950 | 0.3471 |
| Financial inclusion (%) | -1.8921 | 3.2118 | -0.1959 | -0.4677 | 0.0259 | -0.2673 | 0.0008 | 0.5504 |
| Employment (%) | 0.5678 | -0.1959 | 1.5662 | 0.2448 | -0.0028 | -0.4523 | 0.2582 | 0.5062 |
| Cell phone use (%) ^m | -0.6175 | -0.4677 | 0.2448 | 1.7730 | 0.0381 | -0.0892 | -0.0927 | 0.0677 |
| Parliamentary representation (%) | -0.1182 | 0.0259 | -0.0028 | 0.0381 | 1.3671 | -0.7244 | 0.1291 | -0.0489 |
| Absensce of legal discrimination (aggregate score) | -0.1052 | -0.2673 | -0.4523 | -0.0892 | -0.7244 | 2.8400 | -0.4549 | 1.4053 |
| Sex ratio at birth (male to female ratio) | -0.3950 | 0.0008 | 0.2582 | -0.0927 | 0.1291 | -0.4549 | 1.3185 | -0.4234 |
| Discriminatry norms (%) | 0.3471 | 0.5504 | 0.5062 | 0.0677 | -0.0489 | 1.4053 | -0.4234 | 2.7049 |

Figure 1: Correlation and inverse correlation matrices produced by the raw $x$ variables.

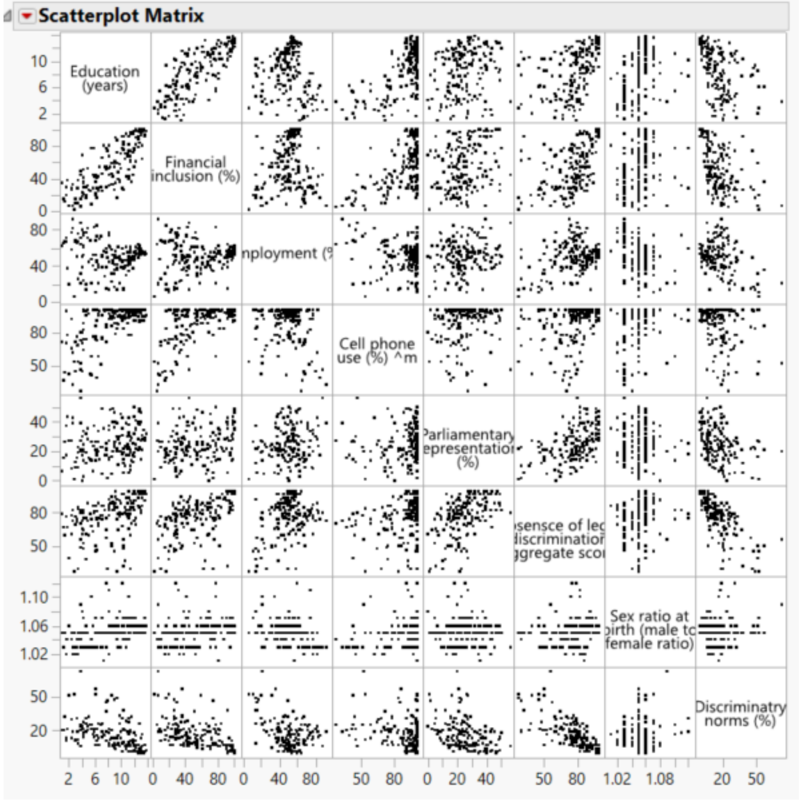
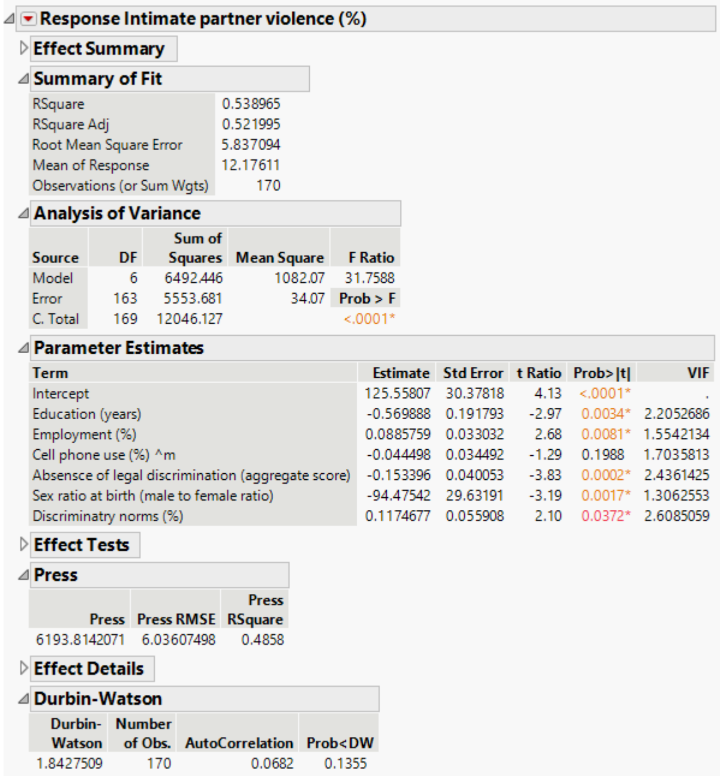


Figure 2: Scatterplot matrix for the raw $x$ variables. Note the visual correlation between "Education" and "Financial inclusion" as well as between "Absence of legal discrimination" and "Discriminatory norms".

## 3.2 Initial model build

Following the analysis of the multivariate plots, we will build an initial model by placing all eight of the $x$ variables in the data in a simple linear regression model as a marker for how well the raw data predicts intimate partner violence. A few things immediately stand out. First, the ANOVA F test is easily passed and five of the eight $x$ variables in the model pass their $t$ tests, being "Education", "Employment", "Absence of legal discrimination", "Sex ratio at birth" and "Discriminatory norms". While this does not necessarily mean that the three $x$ variables that did not pass their $t$ tests ("Financial inclusion", "Parliamentary representation" and "Cell phone use") will not be included in the final model, we note the five that did pass their $t$ tests as potential variables of interest moving forward. Further, this raw model has a relatively good fit, with a value of $R^2_{Adj} = 0.519$ and an acceptable PRESS RMSE to model RMSE ratio: while it's likely that a better model exists, this is unquestionably a good starting place given that no tweaking of the model or its parameters has been done yet.

Recalling the analysis of the correlation between the $x$ variables and the VIF values, we note that we identified that both “Financial inclusion” and “Parliamentary representation” were both flagged as potentially inflating the multicolinearity of the model due to their suspected correlation to other $x$ variables. As such, given that they are both failing their $t$ test and have high VIF values, we first fit a model with “Financial Inclusion” removed (as it had a higher p value), and upon seeing that “Parliamentary representation” still has a less than satisfactory p value in the new model ($p = 0.325$), a new model is fit with that value dropped as well. The model with two dropped variables has improved multicolinearity (as the highest VIF value is now 2.609), a slightly improved $R^2_{Adj}$ value ($R^2_{Adj} = 0.522$) and a slightly improved PRESS RMSE to RMSE ratio (improved from 1.044 to 1.034).



Response Intimate partner violence (%)

Effect Summary

Summary of Fit

| | |
|---|---|
| RSquare | 0.538965 |
| RSquare Adj | 0.521995 |
| Root Mean Square Error | 5.837094 |
| Mean of Response | 12.17611 |
| Observations (or Sum Wgts) | 170 |

Analysis of Variance

| | | Sum of | | |
|---|---|---|---|---|
| Source | DF | Squares | Mean Square | F Ratio |
| Model | 6 | 6492.446 | 1082.07 | 31.7588 |
| Error | 163 | 5553.681 | 34.07 | Prob > F |
| C. Total | 169 | 12046.127 | | <.0001* |

Parameter Estimates

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| | VIF |
|---|---|---|---|---|---|
| Intercept | 125.55807 | 30.37818 | 4.13 | <.0001* | . |
| Education (years) | -0.569888 | 0.191793 | -2.97 | 0.0034* | 2.2052686 |
| Employment (%) | 0.0885759 | 0.033032 | 2.68 | 0.0081* | 1.5542134 |
| Cell phone use (%) ^m | -0.044498 | 0.034492 | -1.29 | 0.1988 | 1.7035813 |
| Absense of legal discrimination (aggregate score) | -0.153396 | 0.040053 | -3.83 | 0.0002* | 2.4361425 |
| Sex ratio at birth (male to female ratio) | -94.47542 | 29.63191 | -3.19 | 0.0017* | 1.3062553 |
| Discriminatry norms (%) | 0.1174677 | 0.055908 | 2.10 | 0.0372* | 2.6085059 |

Effect Tests

Press

| | | Press |
|---|---|---|
| Press | Press RMSE | RSquare |
| 6193.8142071 | 6.03607498 | 0.4858 |

Effect Details

Durbin-Watson

| Durbin- | Number | | |
|---|---|---|---|
| Watson | of Obs. | AutoCorrelation | Prob<DW |
| 1.8427509 | 170 | 0.0682 | 0.1355 |

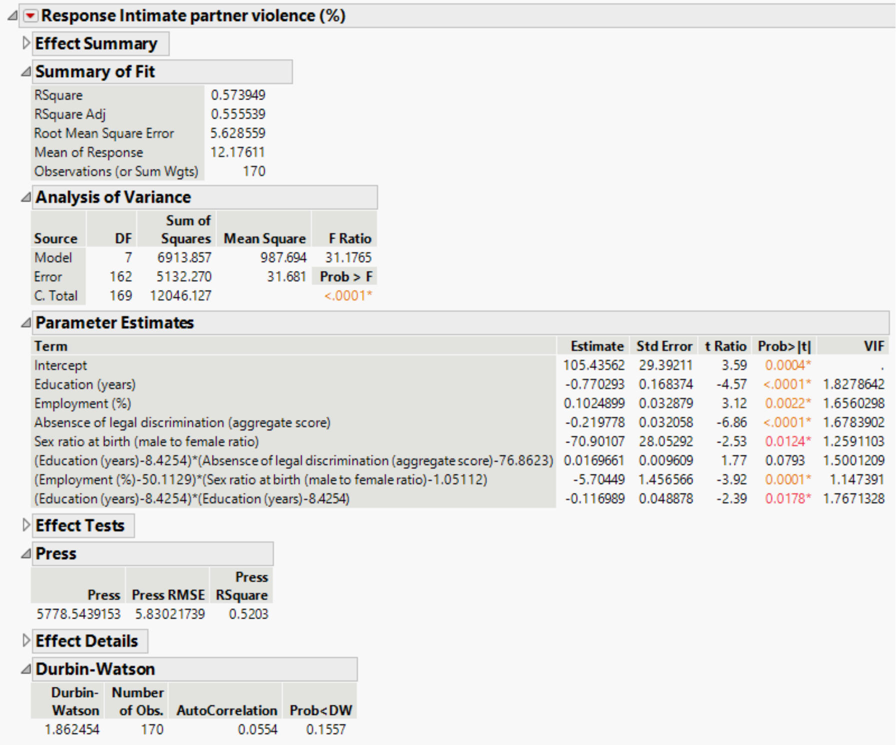
Figure 3: First attempt at creating a model, created by fitting all variables to the $y$ variable and subsequently dropping all variables that significantly failed their $t$ tests for significance.

## 3.3 Stepwise fits

### 3.3.1 Untransformed y

At this point, we will use a JMP stepwise fit to find the optimal model based on a $p$ value threshold. We will explore models with two-factor interaction variables for each combination of the $x$ variables as well as squaring each $x$ variable and finding the optimal model based on a $p$ value of 0.1. This cutoff value was decided upon as informal exploration revealed a trend of stepwise fits with a cutoff value of 0.2 or higher tending to produce models with high multicolinearity, several terms and insignificantly larger $R^2_{Adj}$ values on the "best fit" model. However, this phenomenon was not present when the cutoff value was $p = 0.1$: the VIF values tended to be smaller and there were fewer terms in the model. Only exploring two-factor models was decided upon for a similar reason: three-factor models produced high VIF values and measures of model fit that were insignificant relative to the two-factor models, suggesting that a two-factor model would be adequate.

The first two-factor model fit has a better $R^2_{Adj}$ value than the model created without the stepwise fit ($R^2_{Adj} = 0.671$), but many terms. Interestingly, several of the interaction or squared terms included use variables that are not significant themselves based on their p values and also add a significant amount of multicolinearity to the model. After removing these terms, it becomes clear that the quality of the fit is not severely compromised by nature of the changing $R^2_{Adj}$ value, yet the multicolinearity of the model is significantly improved. After removing "Cell phone use", "Financial inclusion", "Parliamentary representation", "Discriminatory Norms" and their associated interaction and exponential terms, we obtain a model with very good VIF values (with the highest being 1.667) and a worse $R^2_{Adj}$ value ($R^2_{Adj} = 0.556$). Interestingly, only removing "Cell phone use" dropped the model's $R^2_{Adj}$ value significantly, despite the fact that the value added the most multicolinearity to the model. While the $R^2_{Adj}$ value is better in the model with more terms, given the goals of this analysis, we will prefer a model where we can confidently say that the values included all pass their $t$ tests and do not add a significant amount of multicolinearity to the model.



Response Intimate partner violence (%)

Effect Summary

Summary of Fit

| | |
|---|---|
| RSquare | 0.573949 |
| RSquare Adj | 0.555539 |
| Root Mean Square Error | 5.628559 |
| Mean of Response | 12.17611 |
| Observations (or Sum Wgts) | 170 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Ratio |
|---|---|---|---|---|
| Model | 7 | 6913.857 | 987.694 | 31.1765 |
| Error | 162 | 5132.270 | 31.681 | Prob > F |
| C. Total | 169 | 12046.127 | | <.0001* |

Parameter Estimates

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| | VIF |
|---|---|---|---|---|---|
| Intercept | 105.43562 | 29.39211 | 3.59 | 0.0004* | . |
| Education (years) | -0.770293 | 0.168374 | -4.57 | <.0001* | 1.8278642 |
| Employment (%) | 0.1024899 | 0.032879 | 3.12 | 0.0022* | 1.6560298 |
| Absensce of legal discrimination (aggregate score) | -0.219778 | 0.032058 | -6.86 | <.0001* | 1.6783902 |
| Sex ratio at birth (male to female ratio) | -70.90107 | 28.05292 | -2.53 | 0.0124* | 1.2591103 |
| (Education (years)-8.4254)*(Absensce of legal discrimination (aggregate score)-76.8623) | 0.0169661 | 0.009609 | 1.77 | 0.0793 | 1.5001209 |
| (Employment (%)-50.1129)*(Sex ratio at birth (male to female ratio)-1.05112) | -5.70449 | 1.456566 | -3.92 | 0.0001* | 1.147391 |
| (Education (years)-8.4254)*(Education (years)-8.4254) | -0.116989 | 0.048878 | -2.39 | 0.0178* | 1.7671328 |

Effect Tests

Press

| Press | Press RMSE | Press RSquare |
|---|---|---|
| 5778.5439153 | 5.83021739 | 0.5203 |

Effect Details

Durbin-Watson

| Durbin-Watson | Number of Obs. | AutoCorrelation | Prob<DW |
|---|---|---|---|
| 1.862454 | 170 | 0.0554 | 0.1557 |

Figure 4: Best version of a stepwise model calculated without transforming the $y$ variable.

From there, we will analyze the residual plots for the stepwise fit model. The plot of the externally studentized residuals against the predicted values of $y$ revealed a funnel shape, and the distribution of the externally studentized residuals did not reveal a normal distribution (but rather a distribution that is more concentrated in negative externally studentized residuals). As such, it appears that the core assumptions of multiple regression models could have been compromised by this fit, leading us to look to transformations of the $y$ variable as a way of alleviating the irregularities in the residual plots.

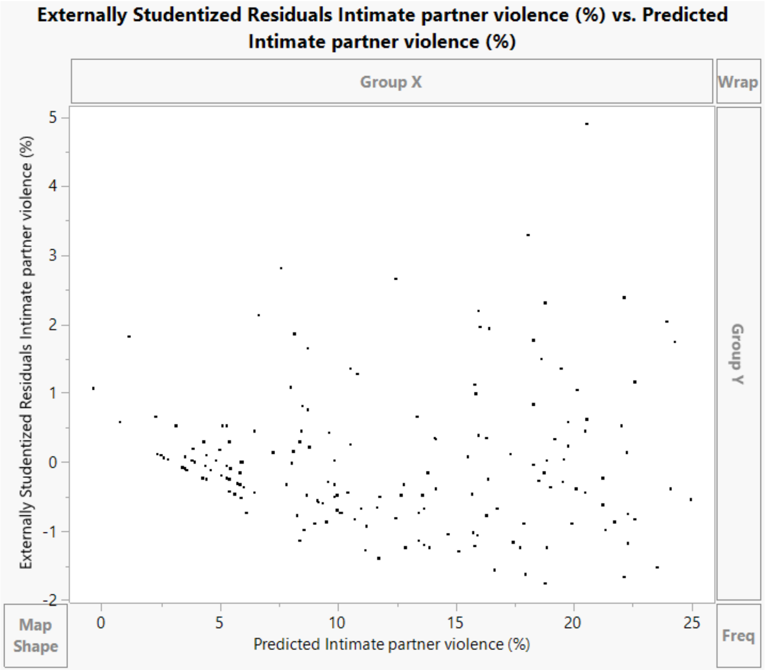
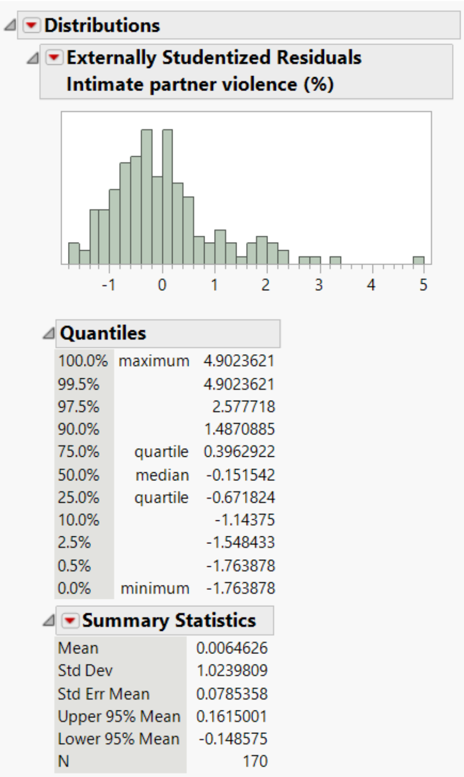


Figure 5: Externally studentized residuals plotted against $\hat{y}$ for the first untransformed model. Note that a funnel shape is visible in the plot.



Distributions

Externally Studentized Residuals
Intimate partner violence (%)

Quantiles

| | | |
|---|---|---|
| 100.0% | maximum | 4.9023621 |
| 99.5% | | 4.9023621 |
| 97.5% | | 2.577718 |
| 90.0% | | 1.4870885 |
| 75.0% | quartile | 0.3962922 |
| 50.0% | median | -0.151542 |
| 25.0% | quartile | -0.671824 |
| 10.0% | | -1.14375 |
| 2.5% | | -1.548433 |
| 0.5% | | -1.763878 |
| 0.0% | minimum | -1.763878 |

Summary Statistics

| | |
|---|---|
| Mean | 0.0064626 |
| Std Dev | 1.0239809 |
| Std Err Mean | 0.0785358 |
| Upper 95% Mean | 0.1615001 |
| Lower 95% Mean | -0.148575 |
| N | 170 |

Figure 6: Distribution of the externally studentized residuals for the non-transformed model. Note that the quantile values and the visual representation of the distribution appears to show that the externally studentized residuals are more concentrated in negative values than they would be in a normal distribution.

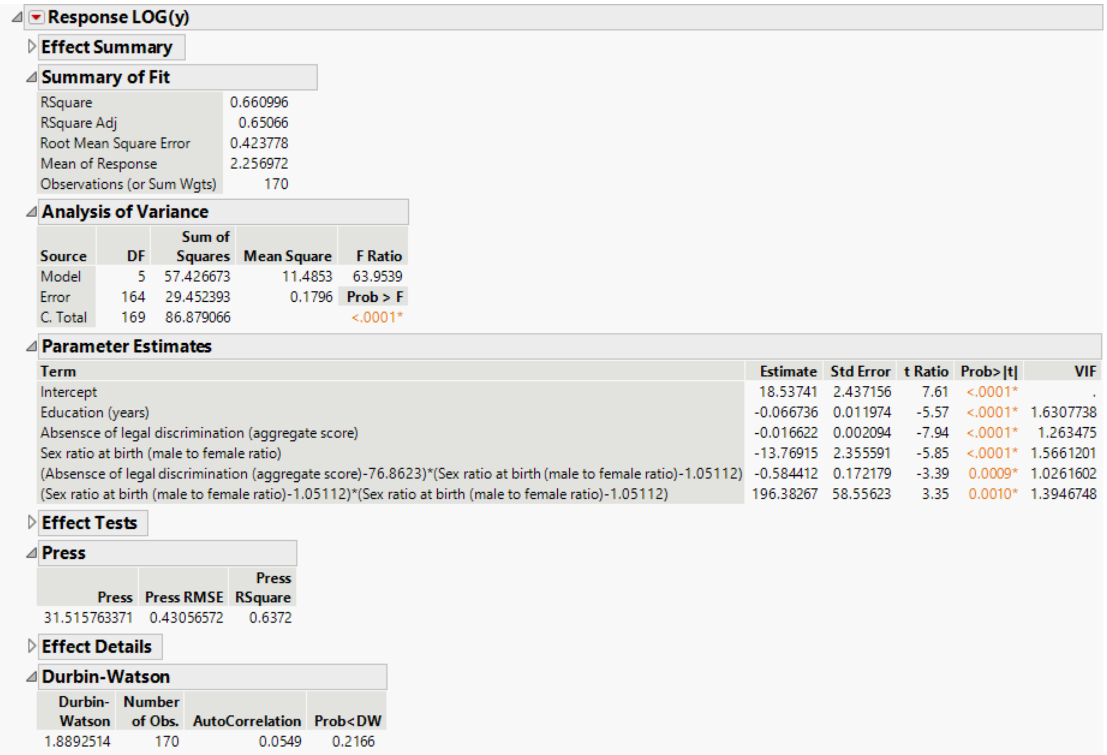
### 3.3.2 Square root transformation of y

The first transformation we will perform is taking the square root of the $y$ variable ($\sqrt{y}$) and fitting a model to it. The same logic as outlined above for choosing the optimal model using a stepwise fit for a non-transformed $y$ (i.e., a model that has the best fit as assessed by $R^2_{Adj}$, yet does not have a large amount of multicolinearity) is used in determining the best model produced by this transformed stepwise fit. Fortunately, even though the best model found using a cutoff of $p = 0.1$ in the stepwise function had a high degree of multicolinearity, the adjusted model following the dropping of terms with high VIF values only has a slightly lower $R^2_{Adj}$ value. The best model found for the $\sqrt{y}$ transformation has a value of $R^2_{Adj} = 0.635$ and its highest VIF value is 2.006. Thus, it is clear that this model is an improvement in quality of fit over the non-transformed model.

However, when looking at the residual plots for this model, a similar, although less dramatic, funnel shape appears when plotting the externally studentized residuals against the predicted values of $y$. Similarly, the distribution of the externally studentized residuals appears to be closer to a normal distribution, but still is clearly not a normal distribution (as it still appears to be concentrated more heavily in negative residuals).

### 3.3.3 Logistic transformation of y

Based on the development in the previous model's residual plots, we will explore another transformation of the $y$ variable: $log(y)$. Stepwise fit is again used to find a model that optimizes $R^2_{Adj}$ and multicolinearity. The resulting model has the best fit found so far in terms of its numeric values and also has promising residual plots. It has a value of $R^2_{Adj} = 0.651$ as a measure of its fit accuracy, and the highest VIF value present in the model is 1.631.



**Response LOG(y)**

**Effect Summary**

**Summary of Fit**

| | |
|---|---|
| RSquare | 0.660996 |
| RSquare Adj | 0.65066 |
| Root Mean Square Error | 0.423778 |
| Mean of Response | 2.256972 |
| Observations (or Sum Wgts) | 170 |

**Analysis of Variance**

| | | Sum of | | |
|---|---|---|---|---|
| Source | DF | Squares | Mean Square | F Ratio |
| Model | 5 | 57.426673 | 11.4853 | 63.9539 |
| Error | 164 | 29.452393 | 0.1796 | Prob > F |
| C. Total | 169 | 86.879066 | | <.0001* |

**Parameter Estimates**

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| | VIF |
|---|---|---|---|---|---|
| Intercept | 18.53741 | 2.437156 | 7.61 | <.0001* | . |
| Education (years) | -0.066736 | 0.011974 | -5.57 | <.0001* | 1.6307738 |
| Absensce of legal discrimination (aggregate score) | -0.016622 | 0.002094 | -7.94 | <.0001* | 1.263475 |
| Sex ratio at birth (male to female ratio) | -13.76915 | 2.355591 | -5.85 | <.0001* | 1.5661201 |
| (Absensce of legal discrimination (aggregate score)-76.8623)*(Sex ratio at birth (male to female ratio)-1.05112) | -0.584412 | 0.172179 | -3.39 | 0.0009* | 1.0261602 |
| (Sex ratio at birth (male to female ratio)-1.05112)*(Sex ratio at birth (male to female ratio)-1.05112) | 196.38267 | 58.55623 | 3.35 | 0.0010* | 1.3946748 |

**Effect Tests**

**Press**

| | | Press |
|---|---|---|
| Press | Press RMSE | RSquare |
| 31.515763371 | 0.43056572 | 0.6372 |

**Effect Details**

**Durbin-Watson**

| Durbin- | Number | | |
|---|---|---|---|
| Watson | of Obs. | AutoCorrelation | Prob<DW |
| 1.8892514 | 170 | 0.0549 | 0.2166 |

Figure 7: Best model based on a logistic transformation of $y$ found by using a stepwise fit. This model is ultimately determined to be the best fit of the data of all of the models considered in this report.

As previously noted, the residual plots for this model are the first that appear to support the core assumptions of multiple regression models. Specifically, the externally studentized residuals plotted against $\hat{y}$ values do not have a clear funnel shape, and the distribution closely resembles a normal distribution (both in its quantile values and in its visual representation of the distribution). Based on these plots, this model appears to be the best in terms of upholding the assumptions of multiple regression based on the residual plots.

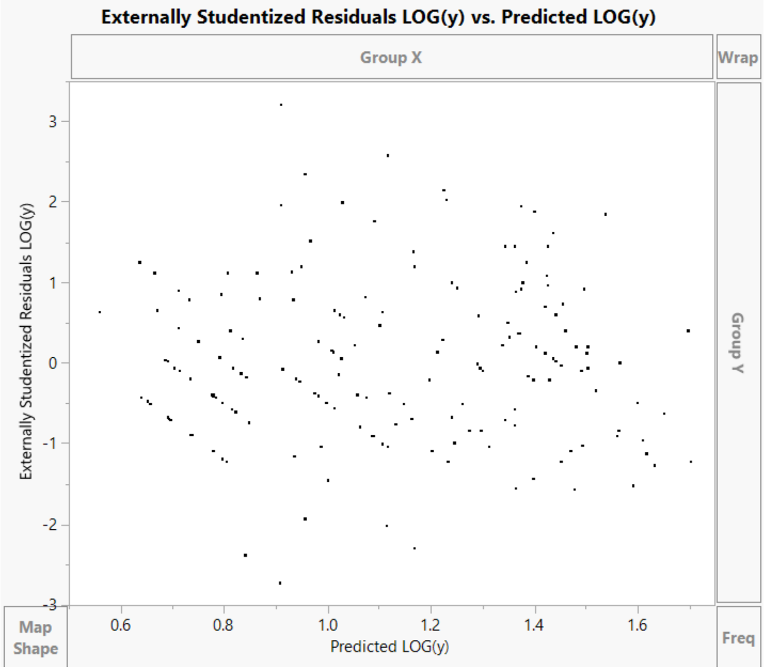
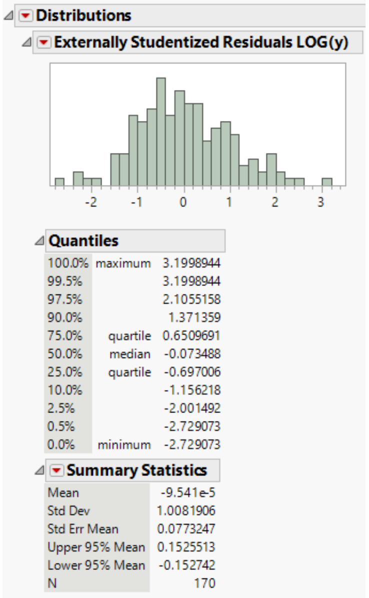


Figure 8: Externally studentized residuals plotted against $\hat{y}$ for the model fit to $log(y)$. Note that the plot relatively closely resembles a random distribution.



Quantiles

| | | |
|---|---|---|
| 100.0% | maximum | 3.1998944 |
| 99.5% | | 3.1998944 |
| 97.5% | | 2.1055158 |
| 90.0% | | 1.371359 |
| 75.0% | quartile | 0.6509691 |
| 50.0% | median | -0.073488 |
| 25.0% | quartile | -0.697006 |
| 10.0% | | -1.156218 |
| 2.5% | | -2.001492 |
| 0.5% | | -2.729073 |
| 0.0% | minimum | -2.729073 |

Summary Statistics

| | |
|---|---|
| Mean | -9.541e-5 |
| Std Dev | 1.0081906 |
| Std Err Mean | 0.0773247 |
| Upper 95% Mean | 0.1525513 |
| Lower 95% Mean | -0.152742 |
| N | 170 |

Figure 9: Distribution of the externally studentized residuals for the best model found using a logistic transformation of $y$. Note that the quantile values and the visual representation of the distribution both support that these residuals are close to a normal distribution, especially relative to the distribution in Figure 6. This plot suggests that the $log(y)$ model upholds the core assumption that the residuals follow a normal distribution and have a mean value of approximately 0 (as seen in the summary statistics).

### 3.3.4 Reciprocal transformation of y

Despite the satisfactory fit of the logistic transformation model, we will also explore the final common transformation $(1 - \frac{1}{y})$ for the purpose of checking if the resulting model has a better fit than the ones found previously. Unfortunately, not only does the $R^2_{Adj}$ value and multicolinearity of the model found by optimizing a stepwise fit fail to be an improvement over the $log(y)$ transformation ($R^2_{Adj} = 0.642$, highest VIF value of 3.22), the residual plots once again do not seem to uphold the core assumptions of multiple regression modeling (as the distribution of residuals is concentrated towards positive values and a funnel shape in the plot of residuals against $\hat{y}$ returns). Thus, we will not explore any further transformations of the data and will instead decide upon the logistic transformation model as the best model fit to the Women's Peace and Security Index data.

### 3.4 Outlier analysis

Since the "best" model is decided upon, outliers will now be considered. To determine outliers, we first consider the rule of thumb of $|r_i| > 3$ indicating an outlier. Only one observation satisfies this condition: the country of Fiji, with an externally studentized residual of 3.200. After looking at this observation, nothing appears to be particularly significant about it (i.e., none of the values in either the $x$ variables included in the calculation or the $x$ variables that aren't look irregular or extreme). Another way to consider this observation as an outlier is to look at Fiji's position in a descending rank of rates of intimate partner violence relative to its rank in the broader WPS Index score. Fiji is ranked 80th in overall WPS Index score, but 19th in descending order of rates of intimate partner violence. While this may appear to be a relatively extreme jump, there are four other observations that make similar jumps near Fiji in the ranking (Barbados, Rwanda, Tanzania and Kenya), three of which have $r$ values less than 1.1 (with the exception of Barbados, with $r_i = 2.575$). Given our understanding of the data pedigree and the rigor that went into collecting this data set, we have no reason to suspect that this outlier may be an incorrect observation. Considering the deeper analysis into this observation relative to other ones, the fact that it is only marginally over the rule of thumb value and that we have no reason in subject-matter knowledge to suspect that something peculiar is in the Fiji observation, nothing will be done about this suspected outlier.

Another measure of influence of individual observations on the model to consider is Cook's distance. Following a calculation of this measure, it becomes clear that none of the observations come close to the rule of thumb of a value of above 1 having a large influence on the model. The greatest $D_i$ value found is 0.074, belonging to Azerbaijan. The country does appear to be an interesting observation, with a relatively high average years of education and absence of legal discrimination values (which would help its WPS Index value) but one of the highest sex ratio at birth values (which would hurt its WPS Index value). Again, given the ultimately small $D_i$ value and the fact that there is not much significant reason to handle the Azerbaijan observation differently than the others, nothing is done based on the $D_i$ values.

## 4 Interpretation

From our final model fit, we can state that three of the eight possible fields proved to provide the best fit of our data to rates of intimate partner violence: "Education", "Absence of legal discrimination" and "Sex ratio at birth".

Our model also determined that the interaction between absence of legal discrimination and the sex ratio at birth collectively affects rates of intimate partner violence. This is an interesting fact given subject-matter knowledge about what creates a safe society for women: it makes logical sense that a society that favors male children that also has a biased legal system could introduce unique challenges to the wellbeing of women (particularly mothers) that would not be present if only one of the two were present in a given society.

Further, this model can predict rates of intimate partner violence in a country with the given three fields with roughly 66% accuracy. This is a fairly good quality fit given common rules of thumbs in regression analysis. When considering the potential inaccuracy of the rate of intimate partner violence measure due to how difficult it is to accuracy measure, this model accuracy is even more satisfactory.

If this project was to be continued, a way of potentially improving the model is to perform a thorough analysis of relevant sociological theories on indicators of intimate partner violence in society and adding new $x$ variables that reflect those theories. The WPS Index data is admittedly not the best to use for predicting rates of intimate partner violence; while the fact that all of the $x$ variables were collected for the purpose of measuring women's global peace and security certainly holds relevance to intimate partner violence, this data set was not collected solely for the purpose of measuring IPV, and a data set that was collected for such purpose would likely be a better source for the model. While expanding the $x$ variables is relatively attainable due to the societal interest in finding indicators for IPV, complications may arise when trying to find a measure of these indicators that match the structure of the WPS Index data (i.e., with a value for all 170 countries included in the report).

# 5 Summary

This report analyzes multiple models attempting to fit observations from the 2021 WPS Index to rates of intimate partner violence sourced from the same report. After analyzing metrics such as $R^2_{Adj}$, PRESS, residual plots and multicolinearity, a logistic transformation of the $y$ variable was determined to be the best model fit. The final model includes three of the eight $x$ variables from the data set: "Education", "Absence of legal discrimination" and "Sex ratio at birth". The squared value of "Sex ratio at birth" and the interaction between "Absence of legal discrimination" and "Sex ratio at birth" were also included in the final model. The model produced a relatively good fit, especially when considering the potential inaccuracies of measuring rates of intimate partner violence and the fact that this data set may not include all of the measures correlated with intimate partner violence (i.e., there could be other indicators of IPV that are not included in this data set).

While this analysis provided me with some takeaways about potential important indicators of intimate partner violence in a country, the majority of my takeaways from this project come from the exposure to analyzing data pertaining to a social issue. While much of the model fitting process was based on technical measures of regression model quality, it was fascinating how the nature of the data influenced how I determined which model was "better" than another (e.g., the understood inaccuracy of measuring the fields included in the data set allowed me to accept some error in the $R^2_{Adj}$ value, and how having low multicolinearity in the model allowed me to more confidently say that a given term is a significant indicator of IPV). After performing this analysis, I have greater recognition of how the desired outcomes of a project and nature of the collected data influence the most important measures of a “good” model: it is very possible that I will soon analyze a different data set relating to a social issue that requires an entirely different set of considerations when analyzing that will dictate the perception of a quality model. Finally, I found great purpose in applying my sociological subject-matter to this project, and it has inspired me to continue pursuing data analytics for social good.

## 6 References

[1] *GIWPS Home Page*. https://giwps.georgetown.edu/.

[2] *Women, Peace and Security Index 2017/18*. https://giwps.georgetown.edu/wp-content/uploads/2019/11/WPS-Index-Report-2017-18.pdf. Georgetown Institute for Women, Peace and Security, 2017.