Gender Inequality and Women's Soccer Success: Utilizing Principal Component Analysis to Isolate Inequality

Matthew Yeaton*

December 14, 2012

Abstract

The effect of gender inequality on success in international women's soccer is not well known, and earlier studies have suffered from multicollinearity. This study introduces a novel model and utilizes principal component analysis and principal component regression to estimate the model in order to surmount the multicollinearity problem. The study finds that gender inequality is both important and significant in determining success in women's international soccer, and we are able to make convincing and robust arguments about the validity of the results, while at the same time fitting a better model as compared to earlier studies.

I. Introduction

The focus of this paper is to understand the explanatory impact of gender inequality in understanding the differential success of various countries' women's soccer programs over the past decade. Gender inequality is an important area of focus in the study of modern economies, and while the harms of gender inequality are several, they can be summarized by two main points: first, gender inequality is a societal harm that unfairly limits the opportunities of fifty-one percent of any country's population in a way that most would consider arbitrary and unjust, and second, gender inequality places constraints on GDP and GDP growth due to labor inefficiency such that we experiences losses in the real economy due to the inequality. More specifically, Dollar and Gatti (1999) assert that if one interprets

*Matthew Yeaton, Columbia University. Email:mry2105@columbia.edu.

gender inequality as evidence of either prejudice or market failure (or both), the gap between males and females is effectively a distortionary tax that has a negative impact on GDP and economic growth. This problem is well studied, and many others, such as Boserup (1970) or Duflo (2010) have gone on to characterize and quantify the actual impact empirically. Hence, we have both a clear social justice argument and a clear economic growth argument. However, the impacts of gender inequality are not always completely obvious, and in certain aspects of life and culture, such as sport, estimating the negative impact of gender inequality can become particularly challenging.

But why might we be interested in the relationship between gender inequality and sport? What can this tell us that the GDP-focused arguments cannot? In general, why study sport? Why study women's sports? To begin with, sport is largely a manifestation of culture, so studying the relationship between gender inequality and sport can help us to better understand underlying social attitudes. If we better understand the impacts and the factors, we can address the issue more directly. By showing that gender inequality has an impact in determining success in soccer, or any other sport, we are showing that gender inequality is real, and that it has an impact in the culture as a whole, which is itself interesting. The universality of the rules of sports like soccer makes our conclusions more dramatic. The rules are the same everywhere in the world, and the rules are the same in men's and women's soccer, so one is unable to make relativistic or hedging arguments downplaying the impacts of gender inequality if the conditions are otherwise homogeneous. Second, if we can establish that gender inequality has an impact on the success of women's soccer programs, we have established that gender inequality has an impact within the sport as a whole. Considering how lucrative the men's soccer industry is, evaluating the impact of gender inequality on other aspects of women's soccer may help us to understand why there has never been a profitable women's professional soccer league, and certainly none the magnitude of the Premier League or La Liga. Hence, this is not purely interesting as a cultural argument, but also from a monetization perspective. Furthermore, one may be interested purely for the model and methodology used in coming to the conclusion that gender inequality matters. Hence, wider applicability of the method in measuring the effect of gender inequality or other hard-to-measure ideas in totally unrelated realms of study may be interesting even if the reader has no interest in either gender inequality or sport.

In order to understand gender inequality in sport, we must of course study women's sports. However, this alone provides a set of difficulties. In particular, much of the existing research on sport and economics has focused on men's sports, both due to the relative popularity of men's sports and sports leagues compared to their female counterparts (which may itself be a symptom of gender inequality), and due to the relative dearth of available data with which to study women's sports. As such, women's sports are not particularly well studied. Furthermore, it is entirely possible the determinants of success in men's sports may be different than the determinants of success in women's sports. For example, we would not expect gender inequality to have a significant impact in analyses of men's sports, but it is entirely possible that it would affect women's sports.

In order to gain international breadth, as we would need to do a rigorous analysis of the subject on a cross-country scale, we need an international sport or sporting event. Hence, there are really only two options that provide the necessary international presence: soccer and the Olympics. Even so, the Olympics suffer from certain regional proclivities and interests that soccer does not (i.e. with the inclusion of baseball or curling). Soccer is the world's most truly international sport[1], and as such is interesting to us because it is the only sport that we can truly compare nation-to-nation.

[1]Obviously, there are certain counties who care about soccer more or less, and devote varying amounts of the national attention to it, but even so, nearly every nation in the world plays soccer, and the low capital requirements required to begin the sport (essentially just a ball-shaped object) mean that even developing countries and poor regions are still able to play the sport, which no doubt contributes to its popularity. However, this purpose of this paper is not to conjecture about the relative popularity of soccer in different regions of the world, and we will attempt to control for differences in popularity in our analysis.

A. Women and Soccer

Hence, we are led to focus on women's soccer. In many parts of the world even today, soccer is a sport dominated by men. In fact, FIFA held the first women's World Cup in 1991, almost sixty years after the first men's World Cup. However, this is not to say that the sport is not popular among women around the world. Murray (1996) and Williams (2002) note that there has been female participation essentially since the beginning of the sport, and that as early as the 1930s women's soccer leagues were formed in countries like Italy and Germany. In fact, Italy was the first country to form a national women's soccer team in 1950. Over the rest of the century, the rest of northern and western Europe, and eventually the rest of the world, would follow suit, and today we see numerous amateur domestic leagues and all but a few countries fielding women's national teams.

While women's soccer still lags in popularity behind men's soccer, the sport still draws high viewership for its largest events. For example, the 1999 Women's World Cup in the United States averaged about 38,000 per game in attendance over its seventeen games, for a total of 658,000 fans. This was larger than the average attendance of English Premier League (the most popular men's professional league) during the same year. Moreover, the final game of the World Cup saw attendance of 92,000, which to this day is the largest crowd to ever witness a women's sporting event. The television audience, which is perhaps the most quickly growing and lucrative component of sports-related monetization, was over 40 million in the United States alone[2] While we should note that the United States was in, and ended up winning, this game, the significance of the magnitude of these numbers still stands. We need only glance at the men's soccer industry to understand the magnitude of how lucrative women's soccer *could* be, but to date, the world has not seen a successful women's professional league, despite the prevalence of semi-professional and amateur women's soccer. Hence, or perhaps because, according to FIFA (2012) "soccer for young girls in many parts

[2] This is larger than an audience for a National Hockey League Finals game, and comparable to Major League Baseball World Series game or a National Basketball Association Finals game. I.e, these are *big* viewership numbers.

of the world is often considered a solely recreational activity due to cultural barriers, social mores and the lack of any financial hope for a future in the game." This is undesirable for the same reasons that gender inequality in general is undesirable, and is certainly anecdotal evidence that there may be some component of gender inequality that contributes to the differing outcomes of countries' women's national teams.

B. Determinants of Success in Sport

However, it is unlikely that gender inequality is the *sole* determinant of differences in women's national team quality. This leads us to ask what the determinants of success in sport actually are. What determines the relative strengths of various countries' athletic programs? Several studies, such as Johnson and Ali (2000), Hoffman, et al (2004), and Bernard and Busse (2004) have examined the factors that impact a country's success in the Olympics. Others, such as Hoffman, et al (2002), and Leeds (2009) examine factors that determine success in men's international soccer. Both of these types of studies will give us a good starting point for understanding determinants of national sporting success in general (although there is a clear focus on men's sports in nearly all of these). A few studies, such as Torgler (2008), Hoffman, et al. (2006), and Matheson and Congdon-Hohman (2011) specifically address the factors that help explain differential success in women's soccer. Torgler (2008) establishes that the determinants of women's soccer success may be different from men's, and Hoffman, et al (2006) and Matheson and Congdon-Hohman (2011) each attempt to elaborate on this work, including attempting to estimate the impact of gender inequality. However, both of these studies appear to suffer from multicollinearity in their regression estimations. It is well known that development and gender inequality are highly correlated, for example as documented in Cuberes (2011). However, in Matheson and Congdon-Hohman (2011) this problem was not significantly addressed, and led to insignificance of the coefficients of the development variables for both the men's and the women's rankings. Hoffman, et al (2006) dealt with the problem by excluding variables, but even so, the gender inequality variables are only significant at the 10% level. Furthermore, both of these studies suffer from relatively poor fit of their models. In fact, the R-squared value on Matheson and Congdon-Hohman (2011) primary estimation was below 50%.

Hence, neither study is able to make robust arguments about the importance of gender inequality in determining women's soccer success. This study utilizes a novel approach in order to address the multicollinearity problem in the regression estimation. Through this approach, we are able to make convincing and robust arguments about the significance and importance of gender inequality in determining success in women's international soccer, while at the same time fitting a better model as compared to the earlier studies. In Section II., I introduce the data used in the study, in Section III. I introduce a novel, but slight, modification to the traditional linear model in order to avoid the problems in estimation inherent in the earlier studies. In Section IV., I describe principal component analysis and principal component regression, the empirical methods that will allow me to estimate this model. Section V. presents the results of the implementation of principal component analysis and principal component regression, and finally in Section VI. I conclude, discuss the implications of the findings, and discuss avenues of further research.

II. Data

This study uses data from a few different sources, primarily: FIFA's men's and women's historical rankings data, macroeconomic country-level development, gender-inequality, and environmental indicators compiled by the United Nations Development Programme, and country-level average temperature data from Mitchell, et al (2003).

A. FIFA Rankings Data

The FIFA women's national team ranking data is compiled and published monthly beginning in 1993 for the men's teams, and several times per year for the women's teams beginning in 2003 (approximately quarterly, although it varies year-to-year). As of year-end 2011, the last full year for which we have data, the men's rankings included 206 national teams that have been active enough recently to merit being assigned a ranking. The women's rankings include 100 national teams who have been active for the entirety of the panel (included in the original 2003 women's FIFA rankings), and 72 who have not been active throughout the entire time period, for a total of 172 total teams. There are no countries for which we have women's data but not men's. This is an advantage over several earlier studies, like Hoffman, et al (2006) who was limited by the time period to a sample of only 88. Furthermore, because of the systematic and non-random[3] inlay that studies like these were forced to include or exclude certain countries, they were subject to a certain amount of sample selection bias that this study is able to avoid. Regardless, since the women's sample is incomplete over the period, even this study is susceptible to a certain level of unavoidable sample selection bias.

The ranking methodology for both men's and women's teams is public and available on FIFA's website (`www.fifa.com`). For men's rankings, the points for a given national team can be determined by adding the number of points front gained from matches in the current twelve-month period to a discounted value of the points gained from matches for the past thirty-six months outside of the current twelve-month period (hence, the past four years of performance are considered). For a single match, the ranking considers the importance of the match (i.e. friendly, qualifier, etc.), the strength of the opposing team based on their current FIFA ranking, and the strength of the team's FIFA confederation. The full explanation of the men's ranking system can be found in Appendix A. Summary statistics for the men's national team average points can be found in Table IV.

For women's national team's the calculation is slightly different, and the full extent of the methodology is not publicly available. However, the spirit of the calculations is very similar, and essentially follows the same pattern as the men's data. For a single match, the ranking again considers the importance of the match (i.e. friendly, qualifier, etc.), the strength of

[3]Socio-economic factors were related to whether a team had played enough games to be included in the sample, and whether the country had a women's team at all

the opposing team based on their current FIFA ranking (and uses the ranking to predict win probabilities for each team), and whether the teams are playing home or away, or both on neutral ground. The full explanation of the women's ranking system can be found in Appendix B.

It is important to note that these criteria are not identical, and hence are not directly comparable. However, this is largely due to the conference "fixed effects" that we tend to see in the men's data and which does not exist in the same way in the men's data. All this means in practice, though, is that the two ranking systems are not on the same scale (women's scale gets as high as the low 2000s while men's is essentially capped at around 1500), but as long as we consider each in the context of its own scale we will be fine. This can be seen in Table IV. Note that they are both on a different scale and also have very different relative variance, indicating that it is not just that the women's is different from the men's by a scalar multiple, but rather that they likely have different underlying distributions. Hence, we should include points rather than ranking (where $d(team_i, team_{i+1}) = 1 \ \ \forall i \geq 1$) in our model to capture the relevant underlying differences in variation.

For both men's and women's points data, the autocorrelations are extremely high over the 2003-2011 period. For the women's points, the autocorrelations are typically larger than 99.5%. For the men's points, the values of the autocorrelations are slightly lower, but still larger than 95%. This is not particularly surprising, as most earlier studies mentioned above suggest that the differential success of a given country's soccer program, for men or women, is largely determined by macroeconomic factors, and other components that do not change quickly over time. Because our window is somewhat small from a perspective of macro change and economic growth, we should not be surprised to see such astronomical autocorrelations. Hence, to help keep the sample of countries included in the study large (since some countries are not included in the rankings until later years in the sample period), we have averaged the points of both the men's and women's teams over the period. Because the autocorrelations are so high, we are able to do this with minimal loss of rigor. Furthermore, as was already mentioned, including the larger sample of countries will help us to at least partially avoid sample selection bias in the study. Table IV shows summary statistics for both the men's and women's teams' points averaged over the period 2003-2011.

Figure 1 shows the relative average successes of countries' women's national soccer teams over the period from 2003 to 2011. Here we see that North American countries like the United States and Canada, the Scandinavian countries, countries in Western Europe like France and Germany, and China and Brazil have dominated over the past decade. In general, these countries are some combination of rich and large (and in some cases both). Note that there are not many successful teams in South America, and almost no successful teams in Africa, central Europe, and the Middle East. These regions are typically not as wealthy as the North America and Western Europe, but also exhibit typically higher levels or gender inequality indices. This is an empirical regularity that will be expanded upon later in the paper.

Figure 1 shows the relative average successes of countries' men's national soccer teams over the period for the same period (2003 to 2011). Here we see that some of the same regions are successful, particularly Western Europe and Brazil, but that others are either less or more prominent. For example, South America, Africa, and the Middle East are much more successful in the men's game than the women's, while the United States and Canada, and the Scandinavian countries are much more successful in the women's game than the men's.

FIgure 2 shows the relationship between men's and women's national team's average ranks over the 2003 to 2011 period. We see a fairly strong positive relationship between the two, although there is certainly some spread. This is about what we would expect looking at the mapped data. There are some similarities between the two, but there is much variance left unexplained.

There are a few things we can begin to understand from this brief glance at the data alone. First, there are some regularities between the men's and women's rankings. We can observe this from both the mapped data and the scatter plot of the average rankings. Hence, we can say that it is likely that some of the same variables that explain the men's rankings explain the women's rankings, but there are other variables needed to understand the differential performance in the women's rankings (otherwise, we would expect a nearly perfect relationship between the two). These data are also in line with the findings of Matheson and Congdon (2011) who found that indicators for Muslim countries had explanatory significance in the women's rankings. This paper will argue, though, that this is significant because of the pervasive gender inequality in many of these countries.

B. Macroeconomic and Other Country Level Data

The United Nations Development Programme (UNDP) macroeconomic country-level development, gender-inequality, and environmental indicators are compiled and published annually, and are publicly available on the UNDP's website (`www.undp.org`). We match these by country name with the FIFA rankings data, and hand match for any remaining, which results in 129 countries who have both rankings data and all indicators data. Several of the variables are rather limiting, however, and if we only require that a country have the more common of these, our sample expands to 153 countries. The coverage of these variables, which can be seen in Table IV, is rather broad, and covers several aspects of all of these general categories.

The categories of these variables, with the exception of the gender inequality indicators, were chosen in line with the previous studies discussed above. Having such a variety of variables to cover various aspects of each of these broad categories provides a significant advantage over these earlier studies, and, for example, in the case of development, including HDI or its constituents variables will allow us to better understand the the overall effect of development than would including GDP per capita alone.

The Human Development Index (HDI) is a composite measure of overall development compiled by the UNDP on an annual basis. It is composed of indicators for life expectancy, education, and income. Specifically, the HDI is drawn from average life expectancy, adult literacy rate (ages 15 and older), gross enrollment ratio (ages 6-22), and GDP per capita. The full formulation of the HDI can be found in Appendix C. The HDI formula was recently revised, but all of the data for this paper is taken from 2011 and earlier, so the measures are comparable from year to year.

The population data contained several outliers, so we have first obtained the natural logarithm, and then subsequently Winsorized the data at the 99% level. The CO_2 per capita, ecological footprint, and maternal mortality rate data also contained outliers, so they too were Winsorized at the 99% level. This level is sufficiently high in order to ensure the rigorousness of any results we have with this data. This has stabilized the data as can be seen in Table IV. It is notable that we were able to solve this problem with a Winsorization threshold as high as 99%, further confirming the data points were in fact outliers.

We have also compiled country-level average temperature data from Mitchell, et al (2003). Following earlier studies, such as Hoffman (2006) and Matheson and Congdon (2011), who found that the deviation from the "ideal" temperature of 14 degrees Celsius to be significant, we have included the squared difference between the country's average temperature and this ideal temperature.

Again, as in the FIFA data, the autocorrelations are extremely high over the 2003-2011 period. For example, HDI, which is representative of the data as a whole as a composite index, typically has autocorrelations over 99%. Macroeconomic indicators change over the span over fifty years or a century, rather than year to year, so this is far from surprising. Because the autocorrelations are so high, we are able to do this with minimal loss of rigor.

The rest of the macro-level variables are self-explanatory, and summary statistics for each can be found in Table IV. Only a few of these are particularly noteworthy. Maternal Mortality Rate has a very high variance compared to its actual values, and the mean is significantly higher than the median. However, it does not appear to have outliers, per se, but rather has a somewhat sizable population of countries (that we might consider a "cluster") that are significantly skewing the mean to the right. GDP per capita and CO_2 per capita also exhibits a similar phenomena of clustering, with the mean significantly higher than the median due to clusters of very rich and very poor countries (for both).

III. Models

We begin with a basic estimation before moving on to our primary models. The basic estimation equation is as follows:

$$p_i = \beta_0 + \beta X_i + \epsilon_i \tag{1}$$

where p_i is country's women's program's points, X_i is a vector of explanatory variables for each country, i, and ϵ_i is the country-specific error term. We begin by defining X_i in line with intuition and the previous literature, and so we include GDP per capita, the logarithm of the population of the country, the squared deviation from the ideal temperature of 14 degrees Celsius, points of the men's program, and an indicator for gender inequality.

In this simple model, GDP per capita is a proxy for the level of economic development in a country, which is related to the leisure time available to developing athletes to hone their craft, in addition to infrastructure like youth leagues and stadiums in which to play. Furthermore, because the citizens of more wealthy, developed countries tend to be wealthier and have more leisure time themselves, there may be larger potential financial remuneration in developed countries[4]. Population is included because more populous countries have a larger pool of potential athletes to draw from. Deviation from the ideal temperature is included as soccer is largely played outdoors, so we may assume that countries with more moderate temperatures will have more time during the year to train as compared to very hot

[4]Even, and perhaps especially, for women where there is not an efficient international market for soccer labor. As such, these women must depend on national federations (like the United States Soccer Federation in the United States or the Football Association in England) or universities to support them while they train.

or very cold countries. The points of the men's team are included to try to capture soccer tradition or overall interest in the country. However, it is likely that these earlier variables are at least somewhat related to men's rankings (as indicated in the literature), and as such we will likely have some multicollinearity in estimation of the model. The last is a gender inequality variable. A few have been used in the literature, including the gender inequality index, ratio of women's to men's earnings, and ratio of women's to men's enrollment rates in secondary schools. Each of these appears to be valid candidates for capturing gender inequality in a country for the simple model.

As discussed earlier, though, there is a significant problem in using a linear model to estimate the impact of gender inequality on women's soccer success. All of the earlier studies who included a gender inequality indicator exhibited what seems to be symptoms of multicollinearity in their regression results when using OLS, even after we ignore the relatively smaller multicollinearity problem resulting from including the men's teams' points. We know that more developed countries tend to have lower levels of gender inequality, and vice versa. This is well documented as discussed above, and is borne out again in our data. For example, the pairwise correlation between the human development index and the gender inequality index, another composite index constructed by the UNDP, is larger in magnitude than 85%. Other measures of gender inequality, like secondary education ratio between women and men and adolescent fertility rate (among others) have slightly lower pairwise correlations with HDI, but still average around 75% in magnitude. If we use GDP per capita, a less complete measure of development as suggested in the models above, the pairwise correlations are slightly lower, but still extremely high, and hover around 50-60% in magnitude. This is unacceptable in any model that hopes to be estimated by OLS, and we predict that our estimates of this model will also suffer from multicollinearity.

Multicollinearity results from having high correlations between two or more explanatory variables in a multiple regression model like the one above. It can manifest itself in a few ways, all of them important in economic interpretation of data. Most notably, it tends to obscure the results of β for individual predictors since two or more of the explanatory factors comove, and it also causes the standard errors of the βs to be unnaturally high, and hence in many cases, to appear insignificant, even though joint tests of significance of the affected variables (such as the F-test) will successfully reject the null hypothesis that all of their coefficients are zero. It also causes the model to be extremely sensitive to both variable selection and sample selection, and both coefficients and standard errors can fluctuate quite dramatically even through random sampling of segments of the full data, or swapping two variables that should be capturing the same effect. This causes any discussion of the impact of a particular explanatory variable to be pointless, as we are unable to assess the validity of either its coefficient or its significance. As noted above, the earlier studies that tried to incorporate gender inequality as an explanatory variable seemed to suffer from this problem, and we predict that our simple linear model will also suffer from these shortfalls.

Let us now conceptualize a slightly different model, where we both try to encapsulate the various explanatory factors in a more broad sense while eliminating the estimation shortcomings inherent in the previous model. The model is formulated as follows:

$$p_i = \beta_0 + \beta Y_i + \epsilon_i \tag{2}$$

where p_i is country's women's program's points, Y_i is a vector of independent variables for each country, i, and ϵ_i is the country-specific error term. While similar to (1), model (2) ensures the independence of the explanatory variables. Let $Y = (y_1, y_2, \ldots, y_n)$ be a matrix such that the vectors y_k are linearly independent. Each of the vectors y_k will encapsulate a particular aspect of women's sporting success we are interested in. For example, y_1 may represent development, while y_2 may represent gender inequality, etc. The overall explanatory story behind understanding differential success in women's soccer has not changed from model (1), and as such we would like to incorporate development level of a country, size of the country's potential talent pool, level of gender inequality in the country, soccer tradition and interest in the sport, and to what extent the country has a climate that accommodates the playing of soccer throughout much of the year. The justification of these categories is as above.

This model assumes that these categories have at least some elements that are not dependent upon each other. Let d_i denote actual development level for a given country, and g_i denote actual gender inequality level in a given country. These values may or may not be measurable. However, we do assume that there exists some measurable value that may conflate the values of the two. Let us call that value c_i, and let us assume that there exits some function f such that $f(d_i, g_i) = c_i$ that denotes the relationship between $d_i, g_i \;\; \forall i$. Let us assume there exists some other function, γ, such that:

$$\gamma\Big(f(d_i, g_i) = c_i\Big) = d_i + g_i + \xi(d_i, g_i) \quad \forall i \tag{3}$$

Hence, all we assume is that there is *some* component of d_i that is independent of g_i, and vice versa. As equation (3) indicates, there may be some term, $\xi(d_i, g_i)$, that embodies the portion of each that is truly codependent upon the other (in the way that introducing more women into the workforce may increase economic growth, or more developed nations may have well-funded courts to protect the legal rights of women once such laws are passed; even these are not perfect examples). However, our goal is to be able to estimate the various effects on their own, and as such to create $Y = (y_1, y_2, \ldots, y_n) = (d, g, \xi(d, g), \ldots)$ with the components linearly independent, so that when we estimate (2) we will be able to accurately measure the effect of g on women's soccer success. We can also relax these assumptions such that both f and γ are approximations and contain some modest error term, and this will still be an advantage over the previous model, which did not allow us to accurately estimate the impact of g at all.

IV. Methodologies

In order to effectively use this model, we will utilize principal component regression. Instead of directly regressing the dependent variable against the explanatory variables, we will regress the dependent variable against the principal components of the explanatory variables. Under the assumptions of OLS, the use of principal component analysis will help us to deal with the collinearity of the data discussed above. Furthermore, the use of principal component analysis may help us to better understand the underlying relationships and comovements within the data, which may itself be a goal.

A. Principal Component Analysis

As per Jolliffe (2002), the principal component is given by:

$$ {}_nY_p^T = {}_nX_p^T W_p \tag{4} $$

where ${}_nY_p^T = (y_1, y_2, \ldots, y_p)$ is the principal component, ${}_nX_p^T$ is the mean-centered transpose of the data matrix, and ${}_pW_p$ is the matrix resulted from singular value decomposition of the data matrix, ${}_pX_n$, as shown in equation (5):

$$ {}_pX_n = {}_pW_p \Sigma_n V_n^T \tag{5} $$

where ${}_pW_p$ is the matrix of eigenvectors of the covariance matrix ${}_pX_n X_p^T$, ${}_p\Sigma_n$ is a rectangular diagonal matrix with the diagonal composed of elements of $\mathbb{R}_+$, and ${}_nV_n$ is the matrix of of eigenvectors of ${}_nX_p^T X_n$.

Essentially, principal component analysis uses a particular orthogonal transformation to turn our potentially correlated explanatory variables into a set of linearly independent vectors called “principal components.” We can think of this as finding a new orthogonal basis for the space of our explanatory variable that preserves the “structure” of the data in the sense of maintaining variance. Furthermore, this new basis is constructed to help us identify the most important gradients in the data. Hence, the first principal component is in the direction of maximum variance in the data, the second principal component is in the direction of maximum variance such that it is orthogonal to the first, and as such will be in the direction of second most variance, etc. We are in effect "rotating" our data to identify the directions of maximum variance. Note that, as mentioned above, the principal components are simply a linear combination of the original variables, albeit chosen such that the first component contains more information than the second, the second more than the third, etc. for all p components.

Because the principal components are sensitive to the relative scaling of the vectors in the data matrix, and can give misleading results if not mean-centered, we have standardized the explanatory variables before implementing the method.

There are several advantages to principal component analysis that we will be able to take advantage of in the context of this study. First, and most importantly, is the orthogonality of the principal components. Hence, assuming normality of the data, we have independence. This eliminates the multicollinearity problem discussed above. Furthermore, because principal component analysis by design creates the new basis' coordinates in order of importance, we can project our original data onto this new basis, and can visualize what was originally high dimensional data in the space of $\mathbb{R}^2$ or $\mathbb{R}^3$, which is a significant advantage in helping us to interpret the meaning principal components.

V. Results

We begin by presenting the results of the estimation of the simple linear model, and observe multicollinearity, as expected. We proceed to analyze the results of principal component analysis on the data, followed by some robustness checks. Finally, we present the results of the estimation of the second model using principal component regression.

A. Simple Model

Table V presents the results of the estimation the model given by equation (1). We estimate the model using five different proxies for gender inequality: gender inequality index in regression (1), adolescent fertility rate in regression (2), labor force participation ratio in (3), maternal mortality rate in (4), and secondary school education ratio in (5). Notice that the coefficient on GDP per capita appears to be highly dependent on our choice of gender inequality proxy. In fact, the maximum value the coefficient takes is *seven times* the minimum value. Furthermore, the significance of the coefficient on GDP per capita varies quite dramatically as well, with p-values as low as 0.8% and as high as 62.9%. In fact, when the significance level of the GDP per capita coefficient is high, the gender inequality proxy tends to be low, as can be seen in regression (3). When we combine these estimation results with the extremely high pairwise correlations discussed above, both the variation in coefficient and instability of significance level of GDP per capita and the gender inequality proxy point to multicollinearity in the data. Hence, it is appropriate to apply the model given in equation (2) by implementing principal component analysis, and subsequently principal component regression.

B. Principal Component Analysis

As mentioned above, principal component analysis is sensitive to the relative scaling and centering of the data, so we standardize the data so that each of the original variables now has mean of zero and a variance of one. Principal component analysis is also sensitive to the inclusion of outliers in the data, but our earlier treatment of outliers using Winsorization should have rectified this issue. The box and whisker plot presented in Figure 3 confirms this. The table shows the variance-standardized data, and the whiskers here are one-and-a-half of the interquartile range. Hence, we see some points that fall outside of this range, but none so egregious as to distort the general shape of the variance of the data. The only one that might come close is CO_2 per capita, but even here we see that there are many points at the right tail, and these are spread relatively far apart, indicating more that they simply have significant spread at the right tail rather than outliers. Overall, the data is generally balanced, and we observe that each variable is on a similar scale, as desired.

In order to simplify our interpretation of the data, we have constructed two different models for purposes of principal component analysis. The first uses HDI as a proxy for development, as this will give more full coverage of human development level than simply GDP per capita, and will be easier to interpret than including a number of development indicators, but with minimal loss of information. As above, we will include the logarithm of population, squared deviation from ideal temperature, Men's FIFA points, and instead of choosing just one gender inequality proxy, we will include four: labor force participation ratio, secondary education ratio, adolescent fertility rate, and maternal mortality rate. We will then construct a model that includes more explanatory variables as a robustness check. The second PCA model includes additional development (public health spending as a percentage of GDP, life expectancy, GDP per capita, expected years of schooling, educational enrollment rate, and adult literacy rate) and environmental (ecological footprint, and CO_2 per capita) variables. We suspect all of these additional development and environmental variables will comove with each other, and can generally be summarized using HDI alone.

B.1. Primary PCA Model

One of the benefits of principal component analysis is that we can project the original variables onto the space of the first few components in order to make meaningful interpretations. We can do this using the eigenvectors of the principal components, which can be seen in Table VI. When the eigenvectors of the principal components are discussed in the context of the original explanatory variables, we call them scoring coefficients, and can use these coefficients to interpret projections of the original data into reduced dimension spaces of our new orthogonal basis. We can also interpret these projections visually, and it is the combination of seeing the projection of the original vectors onto these simplified spaces, and looking at the specific scoring coefficients that we can make the best interpretations.

Figure 4 shows the our original explanatory variables and original data points projected onto the space of the first two principal components. The data seems to form a "cone" in the space of these components, and We note that there seem to be many countries with high values of both the first and second components, with relatively fewer that have high values of just one or the other, and even fewer that have low values of both. Once we look at the projection of our original explanatory variable vectors, we see that the development variables, and the some portion of the gender inequality variables related to development have high magnitudes in the first component. This is immediately in line with the assumptions of our model detailed in equation (3). Recall that:

$$\gamma\Big(f(d_i, g_i) = c_i\Big) = d_i + g_i + \xi(d_i, g_i) \quad \forall i$$

where a variable like maternal mortality rate (c_i), which naturally embodies both elements of development (d_i) and gender inequality (g_i) in a country is broken into into its constituent pieces by some process γ, in this case principal component analysis. Hence, we can think of component one as the **development component**, which generally embodies the various aspects of human development from our set of variables. It is in effect *extracting* development from each of our variables, so that later components can extract other elements. Recall, too, that principal component analysis constructs each component to be orthogonal to all others. Any later components will therefore be independent of this one under normality. Hence, it makes sense that in the development component we see very large positive values of HDI, and very large negative values of adolescent fertility and maternal mortality, two very development-focused measures of gender inequality, and moderate positive values of secondary education ratio (recalling the situation described above, where having a more educated population in general seems as if it would increase productivity, and hence economic growth and development). We also notice a moderate positive value of men's FIFA points in the development component. This also falls in line with our earlier assumption that the men's FIFA points were related to, and likely somewhat determined by, the level of development in a country. Hence, by extracting the component of men's points determined by development, we will be able to better understand and un-obscure what portion of men's points is related to soccer tradition and national interest in the sport.

The second component is dominated by population and men's FIFA points. We might think of this as the **tradition and size component**. As we acknowledged above in equation (3), it was entirely possible that there were some aspects of these conflated variables that were simply inherently conflated, i.e. $\xi(d_i, g_i)$. What we might understand from this is that the interaction of population and tradition together is more important than either of these separately, and in fact is enormously important overall, as the second principal component. This makes sense in a logical sense, too: being large, like India, is not particularly useful without at least some soccer tradition, while having a cherished tradition and national interest in the sport, like in Portugal, is not particularly helpful, as least in this time frame, as the population is small.

Now, interpreting Figure 4, which shows projection of data and original explanatory variables projected onto these components, we see that there are many countries who are both high development levels and high population and tradition. We predict that both of these variables will have a positive impact on women's FIFA points, and a quick glance back at Figure 1 seems to suggest that this is likely to be the case.

The third component, which can be seen in Figures 5 and 6, is actually what we are most interested in for the purposes of this paper. We see that this component is absolutely dominated by labor force participation ratio, with moderate magnitudes of secondary education rate and men's FIFA points. Hence, we can think of this as the **gender inequality** component[5]. It is intuitive that we would see high levels of both labor force participation

[5]Or rather gender *equality* component since we have positive values of labor force participation ratio and secondary education ratio. However, the rotation of the components is rather trivial, so we will stay with the original name and focus on the relative directions of its constituent pieces.

ratio and secondary education ratio, but perhaps less intuitive that we would see moderate values for men's FIFA points. It may be that this component does not necessarily measure gender inequality, per se, but rather measures a country's willingness to allocate resources to non-traditional sources, i.e. sport or the promotion of gender equality. This is a broader way of thinking about gender inequality, and in a sense considers it "discretionary spending." Economics is at its heart the study of the allocation of scarce resources, and this component embodies a country's willingness to allocate both time and money to promote essentially cultural goals. This is not a comment on the value of these goals, or the extent to which they may actually be good investments (as gender equality has been shown to be), but rather reflects a general cultural sentiment within a country that it is acceptable to allocate resources to promote cultural goals. However, while the component is perhaps slightly broader than gender inequality alone, it is still dominated by it, and we will continue to refer to it as such.

Figure 5 shows the projection onto the first and third principal components, and again we see a bit of a cone in the data focused in the first quadrant. This is in line with the earlier discussion of the strong relationship between development and gender equality. Figure 6 shows the projection onto the second and third principal components. We see no distinctive data shape here, but it is interesting to note that the development variables are nearly invisible due to having almost no representation in these components, which is further evidence that we are successfully extracting the various elements of interest. Furthermore, note that while men's FIFA points did have some positive value in the gender inequality component, it is nearly twice as large in the tradition and population component. Hence, we can remain confident in our interpretations of the second and third components discussed above.

Figure 7 shows the projection onto the space of the first three components. We can see that the variables generally cluster in the direction of one of the components discussed above, which seems to support our understanding of the interpretation of the components and justification for the use of the method as a solution for multicollinearity in general.

B.2. Robustness PCA Model

The results from the primary model above seem promising, but we should perform some simple robustness checks to help ensure the validity of our results. Hence, Figures 8, 9, and 10 are analogous to the figures presented above. In Figure 8 we see the projection of the robustness model onto the space of the first two components. Again, we see that the development and environmental variables comove quite significantly along the first component, as we predicted, with maternal mortality rate and adolescent fertility rate still highly negative. Hence, it appears that we are robust with respect to the development component. When we consider the second component, however, we notice that it seems to be the gender inequality component, and if we check Figures 9 and 10 we see that the second and third components have swapped. We can infer that they are similar to each other in importance, and as long as one embodies tradition and size while the other embodies gender equality their specific order is immaterial. Closer inspection of these figures confirms the similarity of their compositions to the gender inequality, and tradition and size components above. The robustness PCA model seems to confirm the results of the primary PCA, although it is more difficult to interpret directly due to the high number of original variables. Hence, we will focus on the primary model for purposes of principal component regression.

C. Principal Component Regression

Note that while the principal components are ordered in descending importance, it is with respect to the explanatory data, *not* the dependent variable, so while we can make hypotheses about the impact of these components on the data, we must actually regress the components against the data in order to draw conclusions about their significance. Typically, one does not use the full set of principal components when conducting a principal component regression, but rather uses a subset of the most important principal components. Hence, we must choose which principal components on which to perform the regression.

Roughly speaking, the eigenvalues denote the stretch of the transformation in the direction of the corresponding eigenvector, and in this case, we can use the relative magnitude of the eigenvalues as a way to understand the variance explained by the principal component. By design, the first principal component contains more information than the first, the first more than the second, and so forth, however, we must select a point at which to stop including components. Figure 11 shows a screeplot of the eigenvalues of the principal components. We see that the first component is roughly twice as important as the second, the second is slightly more important than the third, etc. Another way of thinking about this is by seeing the actual variance explained by each component. Figure 12 shows the actual variance explained by each principal component, as well as the cumulative variance explained by the principal components. Hence, we can see that the first three principal components (the development component, the tradition and size component, and the gender inequality component) explain over 80% of the variance in the data, and that because the distance between the third and the fourth is relatively small, and the eigenvalue of the fourth component is well below one, we will choose to include these first three principal components for use in the principal component regression.

Table VII shows the results of estimation of the principal component regression using OLS with heteroskedasticity-robust standard errors using the first three principal components as independent variables. Recall that the first component is generally the development component, the second is the tradition and size component, and the third is the gender inequality component. We note first that all three coefficients are significant at the 1% level, and that their corresponding p-values are incredibly small. This is already a large advantage compared to the simple linear model described above.

The coefficients of our principal components are as we predicted: the development component is positive, the tradition and country size component is positive, and the gender inequality component is positive. Recall that while we are calling this component the gender inequality component, gender *equality* actually increases as we increase in this component, so this is as expected.

Furthermore, recall that the first three principal components explained about 80% of the variance in the original explanatory variables. Thus, it is notable that the R-squared value of the regression is nearly 70%. Hence, nearly all of the variance that we were able to capture in the original data with these three components is useful, and our regression results end up with an excellent fit. Compare this to the R-squared values of our simple linear model, where it was only in one instance higher (regression (1)), and in that case the p-value of the coefficient on GDP per capita was 63%, and hence entirely unusable for purposes of interpretation. Thus we have improved both the interpretability and validity of the model without sacrificing fit.

VI. Conclusion

We have a model with significant, interpretable coefficients, and a fit approximately as good as, and often better than, the simple linear model. We can say with some confidence that this model and implementation is better than the simple linear models used in the previous literature. Furthermore, we see that although the development component embodies about twice the variance in the data as the gender inequality component, their coefficients are very close in magnitude. This allows us to draw the general conclusion that gender equality is very important in determining success of women's national soccer teams, and perhaps approximately as important as development, population size, and soccer tradition and affinity. This is something that no previous study has been able to say with a high level of certainty, and this, too, provides an advantage over previous work on the subject. Our robustness checks have confirmed our results, and the assumptions of our primary model appear to have been met, all of which is reassuring in confirming the validity of our study.

But what is the *importance* of these results? We have shown that gender inequality is important in women's soccer, perhaps even very important. As we discussed earlier, sport is largely a manifestation of culture, so studying the relationship between gender inequality and sport can help us to better understand underlying social attitudes. We have established that gender inequality is a fundamental component of success in the sport, so now what? Now we can address it. We must acknowledge that part of creating a profitable league in a women's team sport will likely involve confrontation of this problem. There may also be implications in terms of our understanding of gender inequality issues in the labor market as whole. Finally, the use of this model and technique likely has use outside of this very narrow purview, and even those with no interest in the topic of this study may find the methodologies enlightening and interesting within another field.

References

[1] Dollar, D. and Gatti, R. (1999), *Gender inequality, income and growth: Are good times good for women?* Mimeograph, World Bank, Washington, DC.

[2] Boserup, E. (1970), *Womans role in economic development*, New York: St. Martins Press.

[3] Duflo, E. (2010), *Gender inequality and development?* ABCDE Conference, Stockholm, May.

[4] Murray, B. (1996). *The Words Game: A History of Soccer*, Urbana, University of Illinois Press.

[5] Williams, J. (2002). *Women and Soccer, Factsheet 5*, Sir Normal Chester Centre for Soccer Research, University of Leicester.

[6] Federation Internationale de Soccer Association (2012). Website, `http://www.fifa.com`.

[7] Johnson, D.K.N. and A. Ali (2000). *Coming to Play or Coming to Win: Participation and Success at the Olympic Games*. Working Paper.

[8] Hoffmann, R, Lee, C. G. and Ramasamy, B (2004). *Olympics Success and ASEAN Countries: Analysis and Policy Implications*. Journal of Sports Economics, Vol. 5:3, 262-276.

[9] Bernard, A.B., and M.R. Busse (2004). *Who Wins the Olympic Games: Economic Development and Medal Totals*. Review of Economics and Statistics, 86:1, 413-417.

[10] Hoffmann, R., Lee, C.G. and B. Ramasamy (2002). *The Socio-Economic Determinants of International Football Performance*. Journal of Applied Economics, Vol. 5:2, 253-272.

[11] Leeds, M. and E. Marikova Leeds (2009). *International Soccer Success and National Institutions*. Journal of Sports Economics, Vol. 10:4, 369-390.

[12] Torgler, B. (2008). *The determinants of womens international soccer performances.* International Journal of Sport Management and Marketing, Vol. 3:4, 305-318.

[13] Hoffmann, R., Lee, C. G., Matheson, V. and Ramasamy, B (2006). *International Womens Football and Gender Inequality.* Applied Economics Letters, Vol. 13:15, 999-1001.

[14] Matheson, V. and Congdon-Hohman, J. (2011). *International Womens Soccer and Gender Inequality: Revisited*. Working Paper.

[15] Cuberes, D. (2011). *Gender Inequality and Economic Growth*. World Development Report 2012.

[16] Mitchell, T.D., T.R. Carter, P.D. Jones, M. Hulme, M. New (2003). *A comprehensive set of high-resolution grids of monthly climate for Europe and the globe: the observed record (1901-2000) and 16 scenarios (2001-2100)*. Journal of Climate.

[17] Jolliffe I.T. (2002) *Principal Component Analysis*, Series: Springer Series in Statistics, 2nd ed., Springer, NY, XXIX

Appendices

A FIFA Ranking Methodology for Men's Teams

The formula for points for a given match are given as follows:

$$P = M * I * T * C$$

where M denotes the gross points for a match in the standard way (3 points for a win, 1 point for a draw, and 0 for a loss. If the game results in a penalty shootout, the winning team receives 2 points and the losing team receives 1 point), I denotes the importance of the match (as shown in table I), T denotes the strength of the opposing team based on the formula:

$$T = \begin{cases} 200 - Rnk & \text{if } Rnk < 150 \\ 50 & \text{if } Rnk \geq 150 \end{cases}$$

and C denotes the strength of the Confederation (as shown in table II. The strength of the confederation is based on the success of that confederation in the previous three World Cups). These points per game are summed over the current twelve-month period, and are discounted on a year-by-year basis for the previous three years before the current one. This information is all publicly available on FIFA's website (`www.fifa.com`).

Table I. Determination of I (Importance of Match)

Match Importance	Weight
Friendly Match	1.0
World Cup Qualifier or Confederation-Level Qualifier	2.5
Confederation-Level Final Competition or Confederations Cup	3.0
World Cup Final Competition	4.0

Table II. Determination of C (Strength of Confederation)

Confederation	Weight
UEFA/CONMEBOL	1.00
CONCACAF	0.88
AFC/CAF	0.86
OFC	0.85

B FIFA Ranking Methodology for Women's Teams

The formula for the women's world rankings is as follows:

$$WWR_{new} = WWR_{old} + (A - P)$$

where WWR_{new} denotes the country's updated women's world ranking, WWR_{old} denotes the country's women's world ranking prior to the match, A denotes the actual result of the match, and P denotes the predicted result of the match.

The "actual" result of the match is primarily determined by whether the team won or lost, but also considers offensive factors such as goals scored and goal differential. The "predicted" result of a match is largely determined by the difference in current rating converted into a percentage of likelihood of victory. In order to account for the inherent advantage given to a team playing at a home field, the home team is given an additional 100 points in the predicted likelihood calculation (which FIFA asserts is approximately equivalent to 64%, and close to the 66% winning percentage we see for home teams empirically). This within game point value is then scaled by the same importance factor, I, as in the men's rankings, and can be seen in table I. This information is all publicly available on FIFA's website (`www.fifa.com`).

C Human Development Index Methodology

The HDI is the combination of three separate indices, the life expectancy index (LEI), the education index (EI), and the income index (II). The first of these follow the general formula:

$$v = \frac{x - \min(X)}{\max(X) - \min(X)}$$

where v is the country-specific value of the index, $v \in [0, 1]$ by design, x is the country-specific raw value of the statistic in question (i.e. literacy rate or life expectancy, etc.), $x \in X$, where X is the set of all country values for a given statistic. See table III for domains on the various sets X attributed by the UNDP. The second of these, the education index, is a weighted sum of the adult literacy index (ALI), and the gross enrollment index (GEI), both of which follow the basic index construction method. The education index is constructed as follows:

$$v_{EI} = \frac{2}{3} v_{ALI} + \frac{1}{3} v_{GEI}$$

The final index, the income index, is also constructed to still be bounded between zero and one and follows a similar form:

$$v_{II} = \frac{\log(GDPpc) - \log(100)}{\log(40,000) - \log(100)}$$

The final value of the HDI is a simple averaging of these three indices:

$$v_{HDI} = \frac{v_{LEI} + v_{EI} + vII}{3}$$

Hence, the final HDI value, $v_{HDI} \in [0, 1]$, can be used directly to compare across countries. This information is all publicly available on the UNDP's website (`www.undp.com`).

Table III. Domains on X used in HDI Index Calculations

X	Domain
$X_{\text{life expectancy}}$	$[25, 85]$
$X_{\text{adult literacy rate}}$	$[0, 100]$
$X_{\text{gross enrollment ratio}}$	$[0, 100]$
X_{GDPpc}	$[100, 40, 000]$

VARIABLES	Mean	Sd	Median	P10	P90
HDI	0.64	0.18	0.67	0.38	0.87
Adolesc. Fertil. Rate*	55.97	45.19	42.30	9.73	124.43
Adult Literacy Rate	87.14	9.54	90.16	73.91	98.00
CO_2 per cap.	11.79	16.98	6.01	0.31	25.11
Ecolog. Footprint	2.87	1.78	2.50	1.07	5.15
Education. Enroll.	75.39	14.33	77.46	55.22	91.11
Exp. Years School.	12.18	2.95	12.55	8.04	15.84
GDP per cap.	11984.65	13724.11	6938.90	982.10	32368.00
Labor Force Part. Ratio	0.69	0.18	0.73	0.41	0.88
Life Expec.	68.45	10.04	72.24	51.68	79.64
Mater. Mort Rate**	210.02	289.31	75.25	7.50	605.00
ln(Population)	8.99	1.77	9.03	6.68	11.18
Public Health Spend.***	3.53	1.90	3.15	1.55	6.59
Second. Educ. Ratio	0.84	0.23	0.93	0.47	1.03
Temperature	88.03	64.03	76.56	9.01	171.61
Women's FIFA Points	1367.78	336.7	1321.75	946.83	1872
Men's FIFA Points	448.61	266.16	414.65	122.82	797.19

Table IV. **Summary Statistics for Explanatory Variables:** This table shows summary statistics for the explanatory variables used in the various models. It includes the Human Development Index (HDI), the development variables, the gender inequality variables, the environmental variables, the log of population, squared deviation from the ideal temperature of 14 degrees Celsius, and Men's average FIFA points over the period.
*Births per 1,000 women aged 15-19
**Deaths of women per 100,000 live births
***As a percentage of GDP

VARIABLES	(1) pts	(2) pts	(3) pts	(4) pts	(5) pts
GDP per cap.	0.00110	0.00360	0.00728***	0.00536*	0.00662**
	(0.629)	(0.220)	(0.00820)	(0.0582)	(0.0200)
ln(Population)	58.68***	49.42***	51.15***	54.11***	57.76***
	(9.06e-07)	(0.000116)	(0.000205)	(3.73e-05)	(2.17e-05)
Temperature	0.916***	0.820***	0.311	0.580**	0.345
	(0.000448)	(0.00400)	(0.283)	(0.0305)	(0.233)
Men's FIFA Points	0.475***	0.652***	0.648***	0.603***	0.623***
	(1.66e-08)	(0)	(0)	(6.63e-11)	(0)
GII	-917.2***				
	(2.51e-07)				
Adolesc. Fertil. Rate		-2.295***			
		(2.78e-05)			
Labor Force Part. Ratio			229.3*		
			(0.0532)		
Mater. Mort Rate				-0.285***	
				(8.63e-05)	
Second. Educ. Ratio					216.8**
					(0.0231)
Constant	864.2***	597.1***	296.0**	503.4***	226.0
	(2.52e-10)	(3.85e-07)	(0.0415)	(6.55e-06)	(0.127)
Observations	128	129	129	129	129
R-squared	0.715	0.673	0.633	0.656	0.633

Robust p-values in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table V. **Regression Results for Simple Model:** This table shows regression results for our simple first model with several different options for the gender inequality proxy. As we predicted, the model appears to be sensitive to variable selection, even though all of these variables should capture similar things. Hence, the coefficient and significance of GDP per capita in particular fluctuate enormously, both of which are indicative of multicollinearity.

VARIABLES	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
Adolesc. Fertil. Rate	-0.47	0.05	0.07	0.57	-0.18	-0.61	0.18
HDI	0.54	-0.01	0.00	-0.02	-0.17	-0.14	0.81
Labor Force Part. Ratio	-0.06	-0.24	0.88	-0.34	0.03	-0.20	0.00
Mater. Mort Rate	-0.52	0.01	0.17	0.13	-0.08	0.69	0.45
ln(Population)	-0.04	0.74	0.14	-0.05	0.63	-0.09	0.15
Second. Educ. Ratio	0.37	-0.31	0.24	0.68	0.44	0.20	-0.10
Men's FIFA Points	0.29	0.54	0.33	0.25	-0.58	0.19	-0.27

Table VI. **Eigenvectors/Scoring Coefficients of the Principal Components:** This table presents the eigenvectors of the principal components. When discussed in the context of the original explanatory variables, the eigenvectors are called the scoring coefficients of the data, and can be used to interpret the economic significance of the principal components. Here, we can generally consider the first principal component to be the development component, the second to be the tradition and interest/country size component, and the third to be the gender equality component.

VARIABLES	(1) pts
PC_1	118.5***
	(0)
PC_2	127.5***
	(0)
PC_3	113.7***
	(3.29e-09)
Constant	1,327***
	(0)
Observations	131
R-squared	0.689

Robust p-values in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table VII. **Principal Component Regression Results:** This table shows the result of OLS estimation of the women's FIFA points against the first three principal components. We can generally consider the first principal component to be the development component, the second to be the tradition and interest/country size component, and the third to be the gender equality component. Hence, we see that all of these have a significant, sizable impact in predicting the success of women's soccer programs, and in particular we can say with some confidence that *both* development and gender equality have a significant impact in determining the success of women's soccer programs.

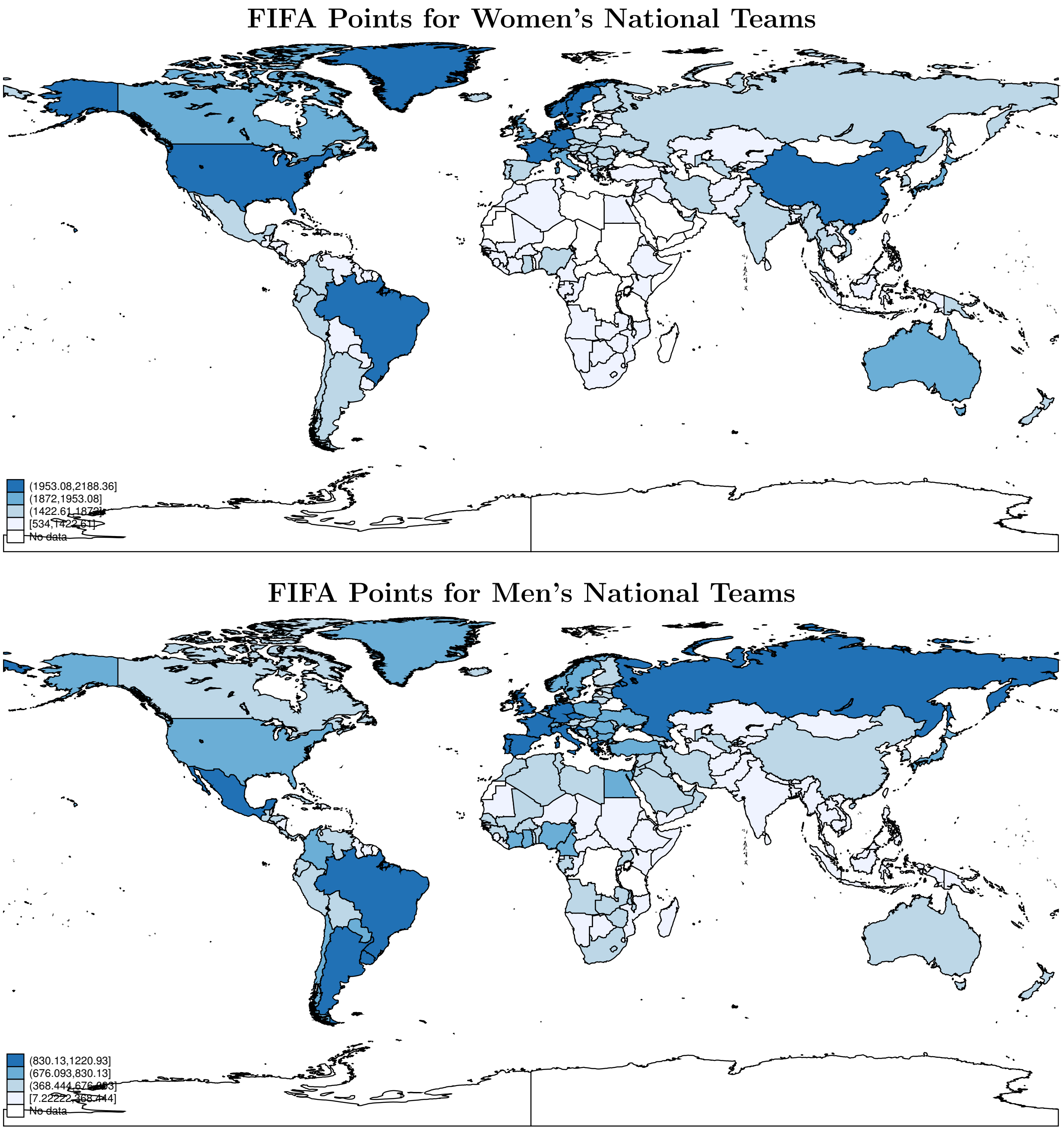

Figure 1. **Countries by Average FIFA points for National Soccer Teams:** This figure shows the average FIFA points earned by the women's national team of each country for the period from 2003-2011. Darker countries have performed better on average. We can see that for women's programs the U.S and Canada, Western and Northern Europe, Australia, China, and Brazil have performed the best over the period, and that for men's programs Europe and South America have performed the best on average.

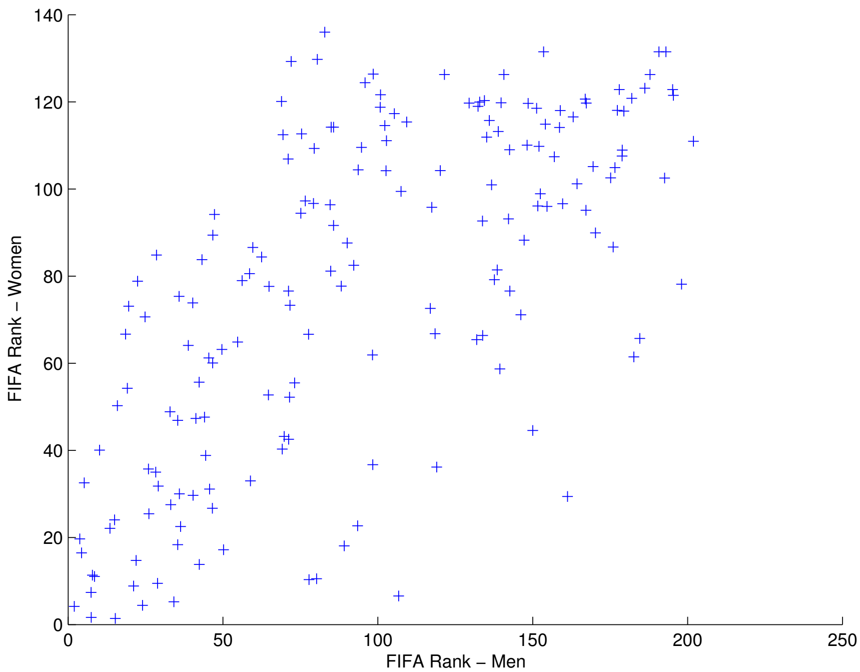

Figure 2. **Relationship between Men's and Women's Average National Team Ranks:** This figure shows the relationship between the average FIFA Rank of the Men's national team of each country for the period from 2003-2011 and the average FIFA Rank of the Women's national team during the same period. We observe a distinct positive correlation between the two, indicating that similar factors may be relevant in predicting both.

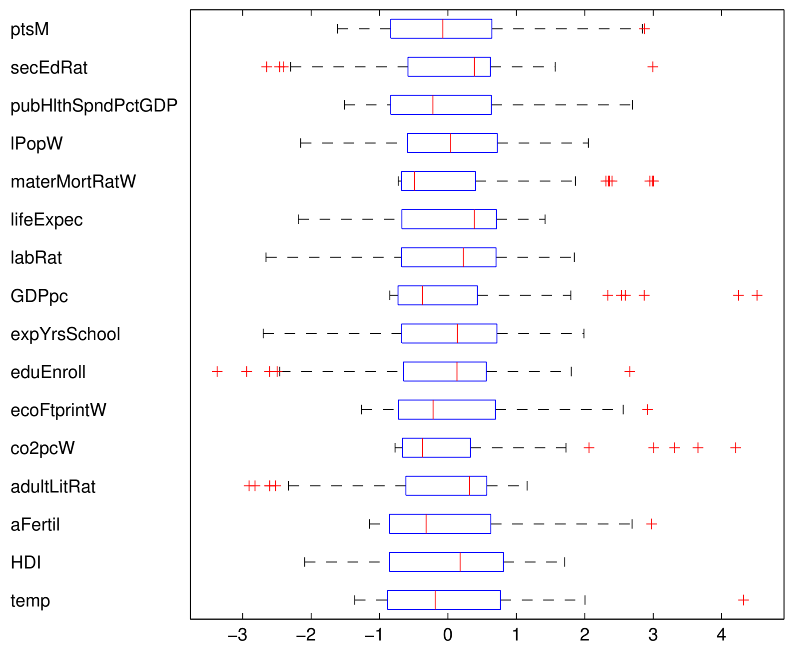

Figure 3. **Box and Whisker Plot for Mean-and-Variance-Standardized Explanatory Variables:** This figure shows the distributions of each of the variance-standardized explanatory variables on a box and whisker plot. The boxes are the interquartile ranges (IQR), the line is the median, and the whiskers are $(1.5 * IQR)$. We can see that the explanatory variables are on similar scales, and there are no instances of serious outliers (except for the marginal case of CO_2 per capita, which is close, but still within acceptable limits).

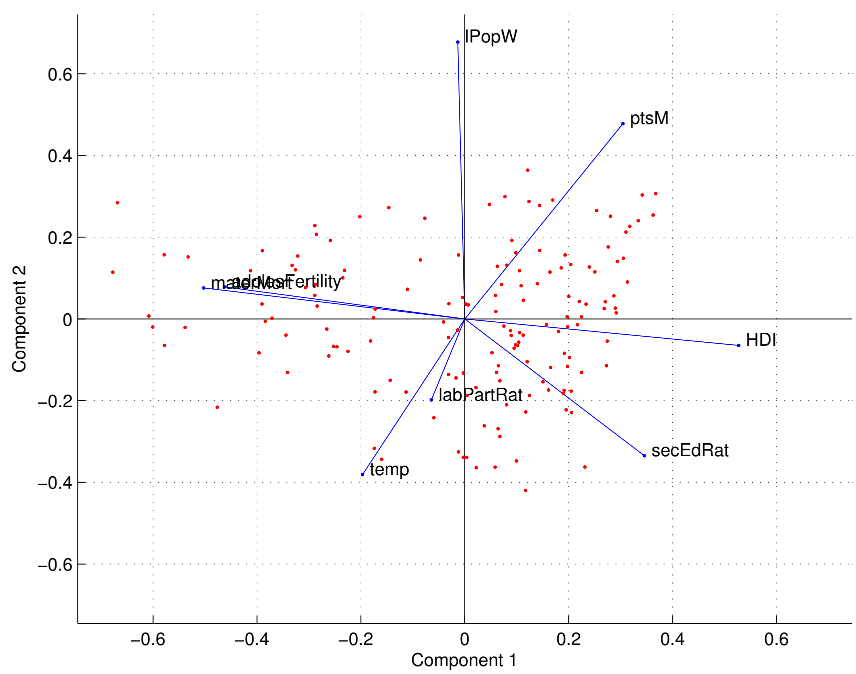

Figure 4. **First vs. Second Principal Components for Primary Model:** This figure shows the projection of the original data and explanatory variables onto the space of the first two principal components. We notice that the first component primarily corresponds with development, while the second component primarily corresponds with soccer tradition and national interest, and population size. We observe a cone of countries such that many are in the first quadrant (high development, tradition, and size), while relatively fewer are in the remaining quadrants.

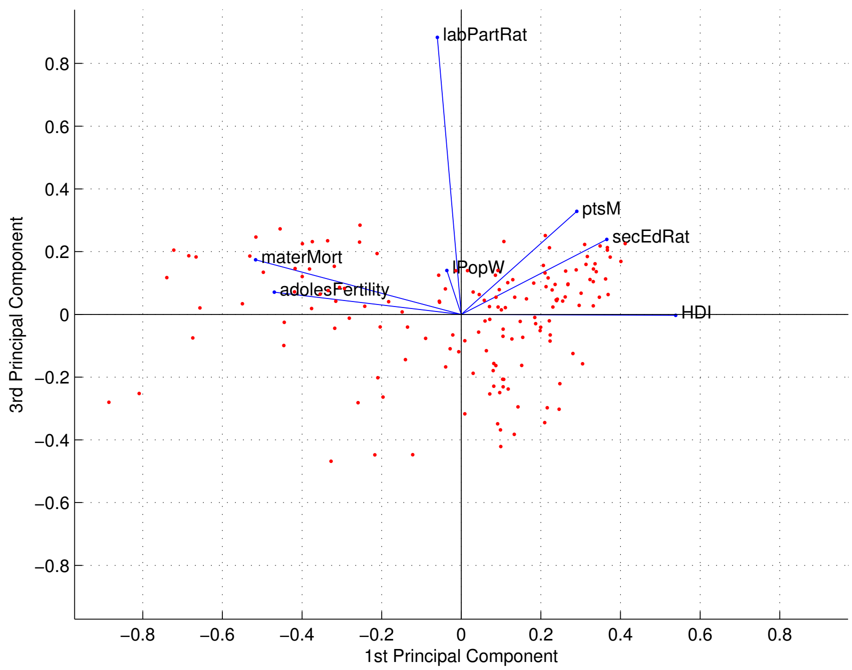

Figure 5. **First vs. Third Principal Components for Primary Model:** This figure shows the projection of the original data and explanatory variables onto the space of the first and third principal components. We notice that the first component primarily corresponds with development, while the second component primarily corresponds with gender inequality. We observe a cone of countries such that many are in the first quadrant (high development and gender equality) which corresponds to the stylized fact that more developed countries tend to have lower levels of gender inequality as noted above, while relatively fewer are in the remaining quadrants.

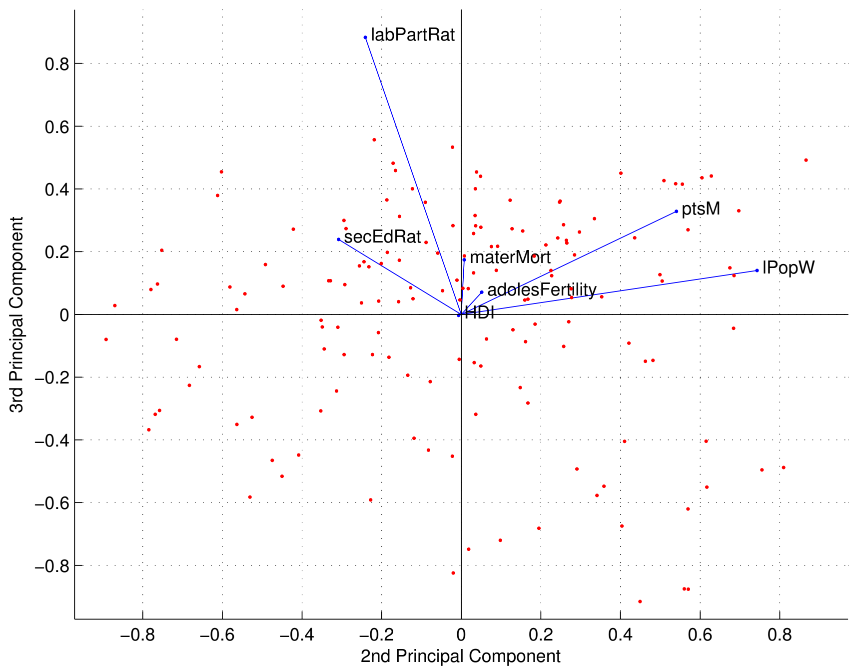

Figure 6. Second vs. Third Principal Components for Primary Model: This figure shows the projection of the original data and explanatory variables onto the space of the second and third principal components. We notice that the second component primarily corresponds with soccer tradition and national interest, and population size, while third component primarily corresponds with gender inequality. We do not observe a distinctive data shape.

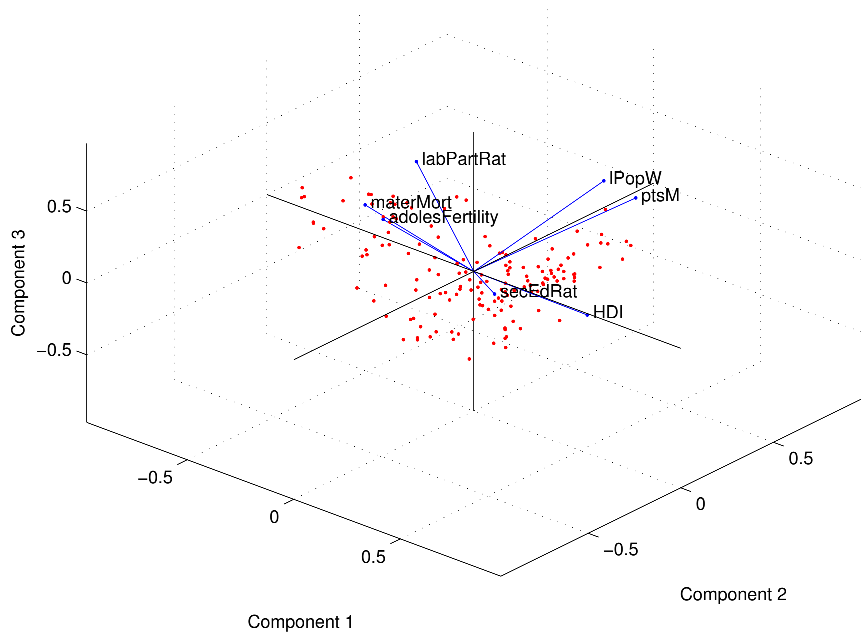

Figure 7. **First Three Principal Components for Primary Model:** This figure shows the projection of the original data and explanatory variables onto the space of the first three principal components. We notice that the first component primarily corresponds with development, the second component primarily corresponds with soccer tradition and national interest, and population size, while third component primarily corresponds with gender inequality. We observe distinct clusterings of the original explanatory variables, which is in line with the assumptions of the primary model given in (3).

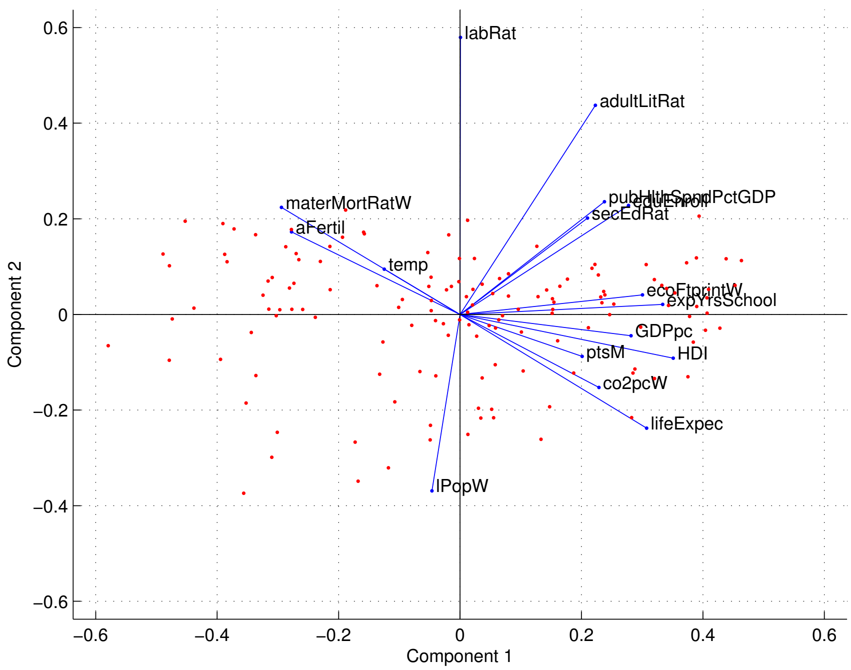

Figure 8. **First vs. Second Principal Components for Robustness Model:** This figure shows the projection of the original data and explanatory variables onto the space of the first and second principal components. We notice that the second and third components have swapped, so that this figure is analogous to Figure 5, verifying the robustness check.

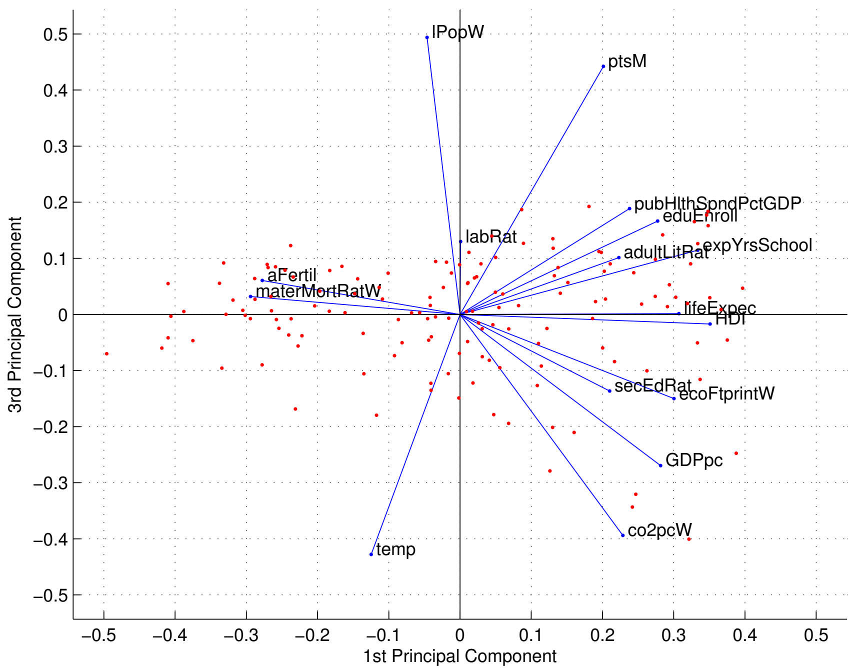

Figure 9. **First vs. Third Principal Components for Robustness Model:** This figure shows the projection of the original data and explanatory variables onto the space of the first and third principal components. We notice that the second and third components have swapped, so that this figure is analogous to Figure 4, verifying the robustness check.

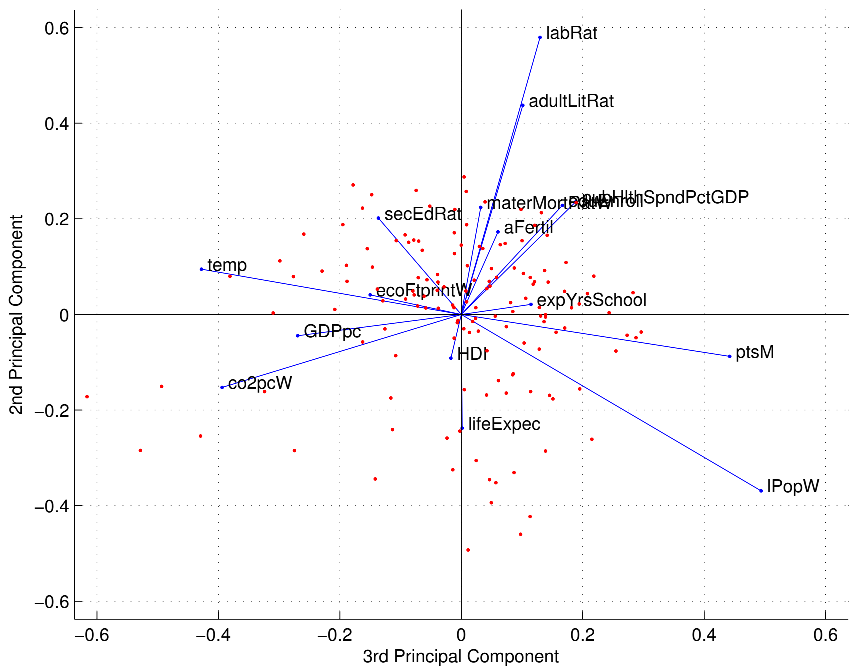

Figure 10. Second vs. Third Principal Components for Robustness Model: This figure shows the projection of the original data and explanatory variables onto the space of the second and third principal components. We notice that the second and third components have swapped, so we have swapped the axes for the viewer's convenience. This figure is analogous to Figure 6, verifying the robustness check.

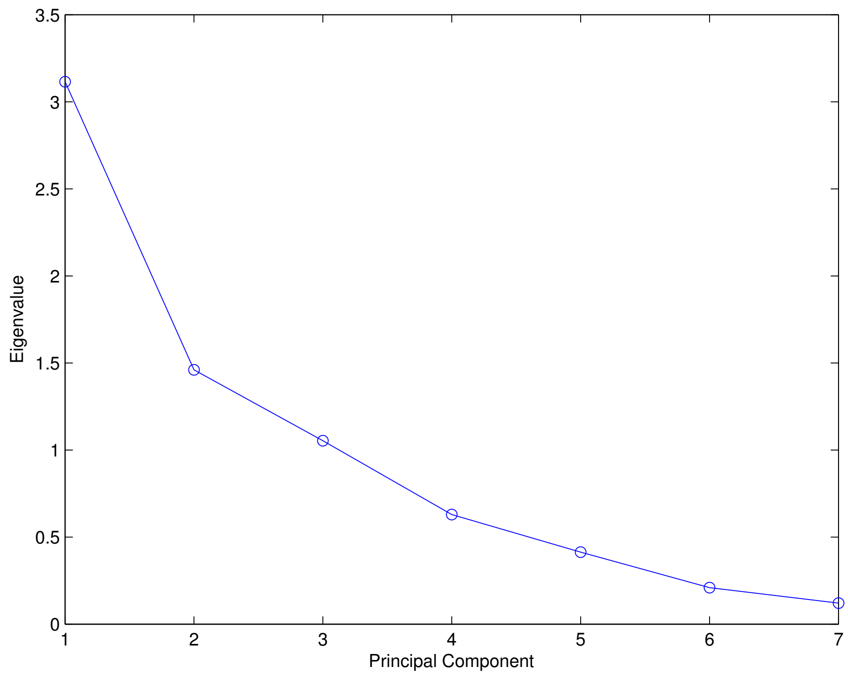

Figure 11. **Screeplot of Eigenvalues of the Principal Components of the Primary Model:** This figure shows a screeplot of the eigenvalues of all seven principal components of the primary model. The eigenvalues are an indication of the amount that the transformation stretches the data, so because we are performing principal component analysis, the eigenvalues tell us the relative importance of the principal components in explaining the variance within the data.

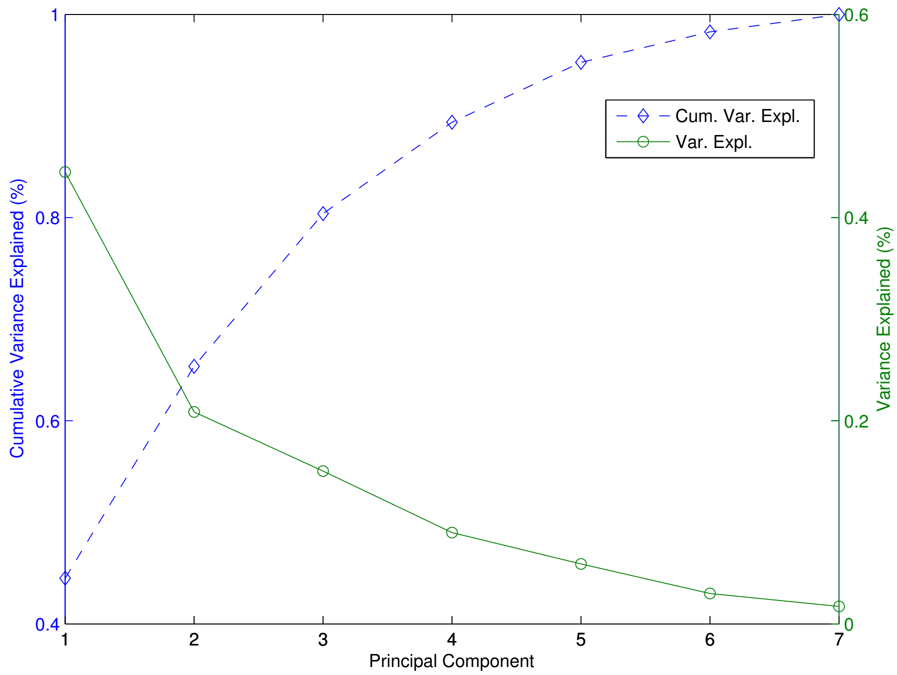

Figure 12. **Percent Variance Explained by Principal Components of the Primary Model:** This plot shows the variance explained by the individual principal components (right axis) and the cumulative variance of the principal components thus far (right axis). We see that the development component, tradition and country size component, and gender inequality component explain over 80% of the variance within the data.