Stat 139 Final Project:
Factors Behind Disparities in Gender Ratio at Birth Around the World
By: Michelle Hu, Alex Pai, Varun Krishnan, Catherine Myong

Introduction:

Currently in the world, there is a gender disparity between the countries of the world, relying on a whole spectrum of confounding factors. One of the most disturbing and striking gauges is the ratio of childbirth between males and females. Quite shockingly, the general ratio between male and female births is 1.07:1, and this varies a lot from country to country - in some countries, it reaches as high as 1.26:1; using the UN convention in our paper, which compares female to male births out of 100, these ratios are 93.5 and 79.4.[1] What factors can explain this massive inequality? Are they health-related factors, demographic-related factors, or policy-related factors?

We study 7 social, political, and biological factors tied to gender imbalance, and determine each of their significances in turn. We also extrapolate for non-study confounding factors, and potential underlying correlations between each of our grouping variables. We hope that this will help us partway in reaching a conclusion regarding the gender normative imbalance in countries, measured by proportion of births of the male and female population measured in various countries.

Research Question:

Our question is the following: which national characteristics are associated with gender birth disparity in countries, measured as the ratio of female/male births in the given year 2005?

Hypothesis:

We predict that high abortion rates and lower status of women are associated unequal gender ratios at birth. Gender at birth has also been linked to biological factors - for example, younger mothers are more likely to have sons, and greater availability of food is associated with more male births. We predict that high female adult obesity rate and GDP are associated with more male births.

Methods:

For our response variable, we plan on using the female/male birth ratio by country, as acquired from UN Data and most recently updated in 2005. We plan on using data acquired from the United Nations database for each of our predictors:

1. abortion rate
2. gender inequality index

[1] https://en.wikipedia.org/wiki/List_of_countries_by_sex_ratio

3. female enrollment in primary education
4. female share of government positions
5. GDP per capita
6. female adult obesity
7. prevalence of teenage births

Our response variable was female/male birth ratio, parsed by each country. We will fit a multiple regression model using these 7 predictors.

We confronted the difficulty of finding data without too many missing values, because predictors such as abortion rate and the gender inequality index had been recorded inconsistently throughout the years. For example, there were fewer than 10 countries with recorded abortion rates in 2005 that we could find in the UN database. Instead of leaving out the abortion predictor, we decided to find a way to fill in the missing data and use the predictor, because selective abortions are a direct reason why there would be a major effect observed in female/male gender imbalance. We had two methods of filling in the missing data: (1) using data from a wider range of years, from as far back as 1985, and assuming the variable stayed constant over time, or (2) using data from neighboring countries to estimate the missing data. We decided to use method (2), since we decided that a given country's characteristic at a given time, such as abortion rate, is more likely to resemble that of the country's neighbors at the time rather than that of the country 20 years ago. We investigated this assumption by looking at the simple regression model between gender birth ratio and abortion rate using Method (1).

Method (1) consisted of not doing imputations, but instead trying to aggregate all the crude data from across the years 1996-2005. It will likely lead to very inaccurate estimates the missing data, but it serves as an interesting comparison point for factors where not much data is available so imputations are also highly suspect.

Here is a scatter plot of gender imbalance vs abortions (1996 - 2005) using Method (1):

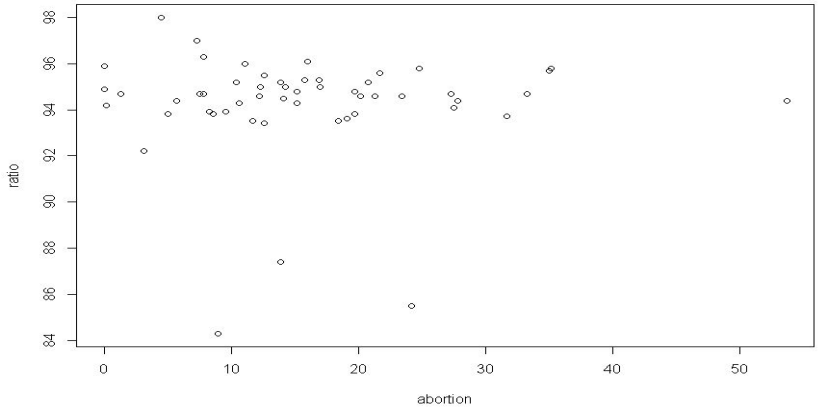


Figure 1

Indeed, there is almost no correlation whatsoever.
The following R output validates this:

```
Call:
lm(formula = ratio ~ abortion, data = Abortion.vs.ratio)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9865 -0.3300  0.4075  0.9265  3.7077

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 94.298131   0.587232 160.581   <2e-16 ***
abortion    -0.001291   0.031284  -0.041    0.967
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.341 on 52 degrees of freedom
Multiple R-squared:  3.273e-05,         Adjusted R-squared:  -0.0192
F-statistic: 0.001702 on 1 and 52 DF,  p-value: 0.9672
```

A p-value of .96 indicates that there's almost no impact of abortions on male/female ratio in this case whatsoever. Thus, we needed to do imputations to get more accurate data and hopefully get a statistically significant result.

Method (2) in detail consisted of sampling 20 estimates for a missing predictor value, from a truncated Normal distribution with mean and standard deviation of the predictor values of countries in the same region. The estimated beta coefficient for the predictor was the average of the 20 different beta coefficients given by the 20 different estimated values, and its adjusted variance was the sum of the variance of the 20 estimated coefficients and the average of the 20 estimated variance of the beta coefficients.

$$SD_{\text{Adjusted}} = \sqrt{(\bar{\sigma}_{\beta 1})^2 + (\mu_{\sigma_\beta})^2}$$

Model:

I. Visualizing the Data:

The first step is to visualize the data. We found, on the internet, a list of developing countries as compiled by the 26th Assembly of the International Union of Geodessy and Geophysics[2]

In order to examine whether Male/Female Birth Rates are related to the development of a country, we made side by side boxplots comparing Male/Female Birth rates to the indicator of whether or not a country is developed.

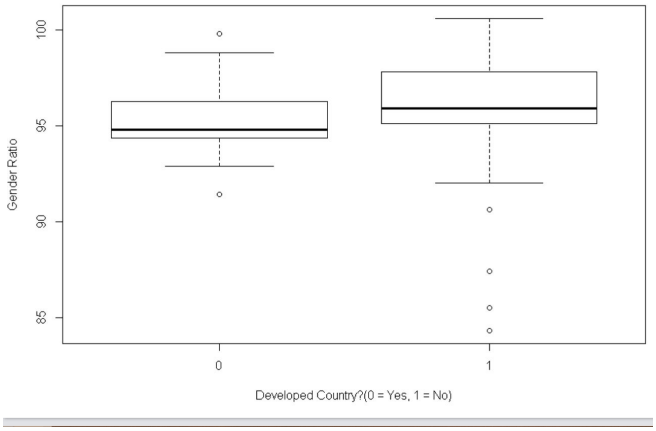


Figure 2

There are a few interesting observations that can be made from this plot. First, we can infer that on average, developing countries tend to have a higher median ratio of female to male births than developed countries. There are many reasons why this could be the case, which we will go into using more specific predictors in this report. Also note that the tail is much larger in the boxplot for developing countries. This is mainly due to the fact that certain developing countries such as Azerbaijan, Armenia, and India have overwhelmingly negative attitudes towards females, and the very bottom outlier among Developing Countries is China, due to the One-Child Policy.

II. Checking Assumptions:

<u>Checking Imputed Model Assumptions:</u>

In order to validate our statistical conclusions, we will check

1. **Normality of Residuals:**
2. **Linearity of Yi|Xi:**
3. **Homoscedasticity (Constant Variance)**

[2] E-WORKS.CZ. "List of Developing Countries." *List of Developing Countries | 26th IUGG GENERAL ASSEMBLY 2015*. N.p., n.d. Web. 07 Dec. 2016.

4. **Independence:**

1. **Normality of the Residuals:** The normality assumption is violated. We can fix this by a right-skewing transformation on Y. Let us see that the dependent variable Y, ratio of female to male births, is left-skewed (Figures 1.1-1.2). We can see this from the following histogram, as well as the qq-plot of Y quantiles to theoretical quantiles. This is because the tendency of the ratio is closer to a normalized level of 100 (percent) than to zero (percent), which is an impossibility (means zero females in a population of males). We can test transformations (Figures 1.3-1.6) to see which transforms to normal.

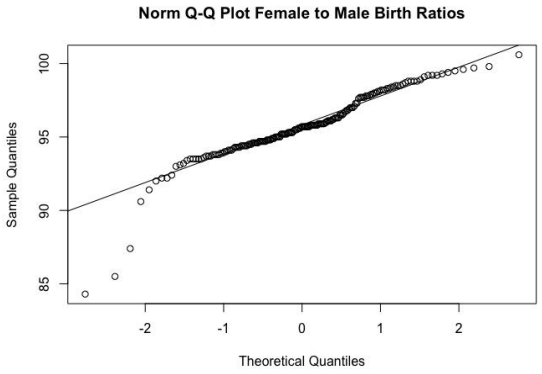


Figure 3.1

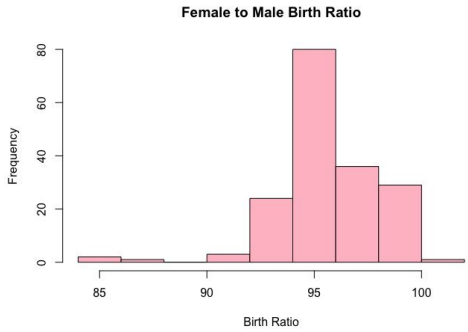


Figure 3.2

The following transformations are to correct skew. We choose the square transformation, because it fits the histogram of the Y's most closely to the standard normal distribution.

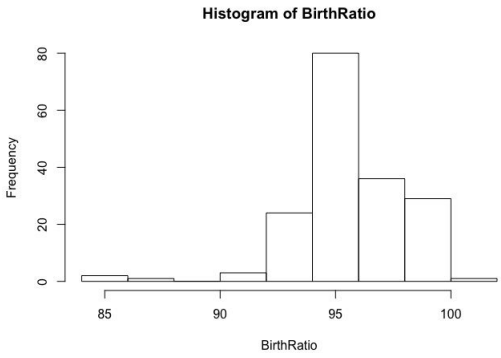


Figure 3.3

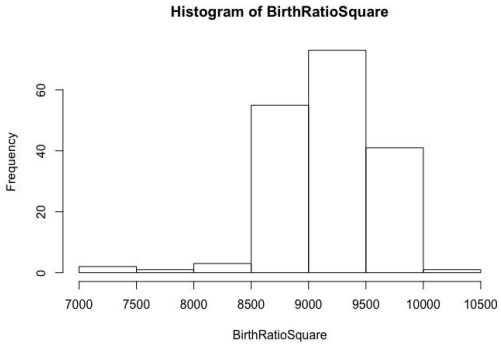
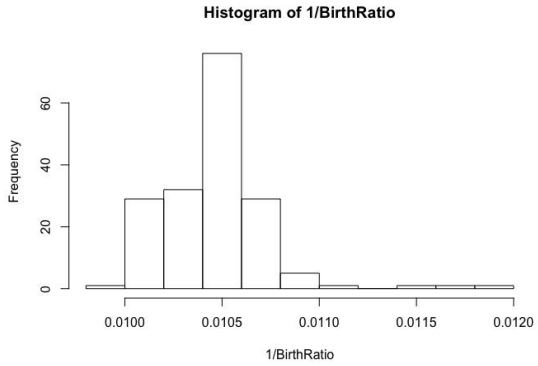


Figure 3.4



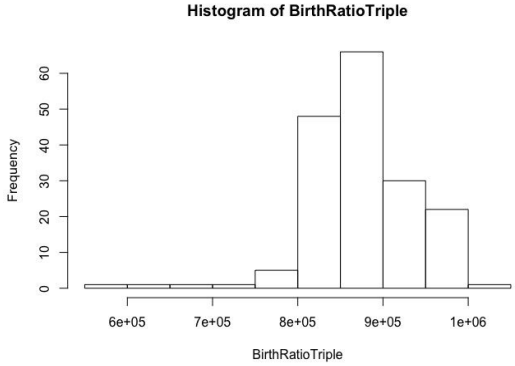


Figure 3.5 **Figure 3.6**

After we transform the data, we get three outlier points, according to Residual Normal Q-Q plot.

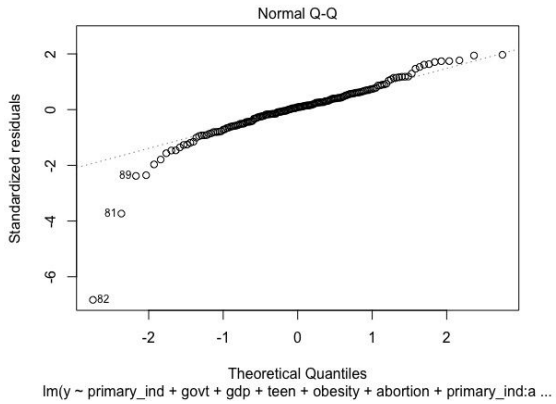


Figure 4

These outliers are Armenia, China, and Azerbijan-- which are known to have extreme gender policies regarding the birth of children. In China, there was the one-child policy, which was believed to contribute heavily to the single-sex selection of fetus. In societies with high boy preference, like Armenia and Azerbijan, with hereditary regimes, the decrease of desired number of children may cause sex-selection, as parents would try to have their preferred gender of children within the small family size.[3]

These outliers will have significant results in abortion rate, gender inequality index, and female share of government positions, as we will later describe. We will leave these points in for this reason.

2. Linearity of Regression (Yi|Xi)

Let us check the linearity of Y|X plots, using each of the predictor variables:

[3] Hayruni, Mariam. "Thesis Paper: Sex-selective Abortions in Armenia." *Thesis Paper: Sex-selective Abortions in Armenia*. American University of Armenia, May 2013. Web. 7 Dec. 2016.

Gender Inequality Index

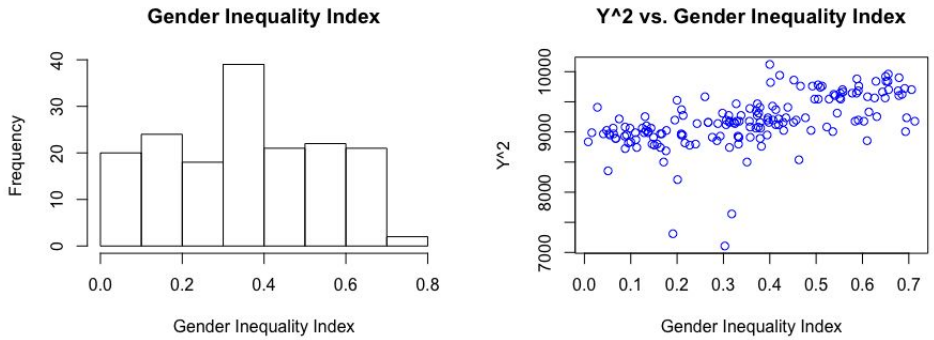


Figure 5

The Gender Inequality Index is an index measuring gender disparity developed by the UN. It incorporates measures of reproductive health, access to education and government leadership, and labor participation. The distribution seems symmetric, and its relationship with gender ratio at birth (Y) looks somewhat linear with a few outliers.

Primary School Enrollment Gender Ratio: Indicator Variable

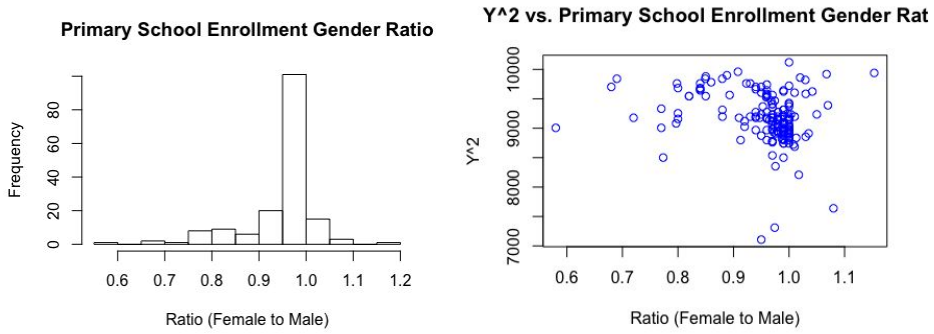


Figure 6.1

The primary school enrollment gender ratio is stacked near 1.0, so no transformation makes a difference worth the loss of interpretability. So we created an indicator variable - if primary school enrollment gender ratio > 0.95, then `primary_ind` = 1, else `primary_ind` = 0. A boxplot of the indicator variable against the Y shows a difference in the means of the two groups.

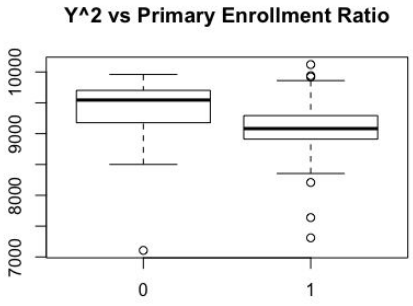


Figure 6.2

Female Seats in Government: Square root transformation

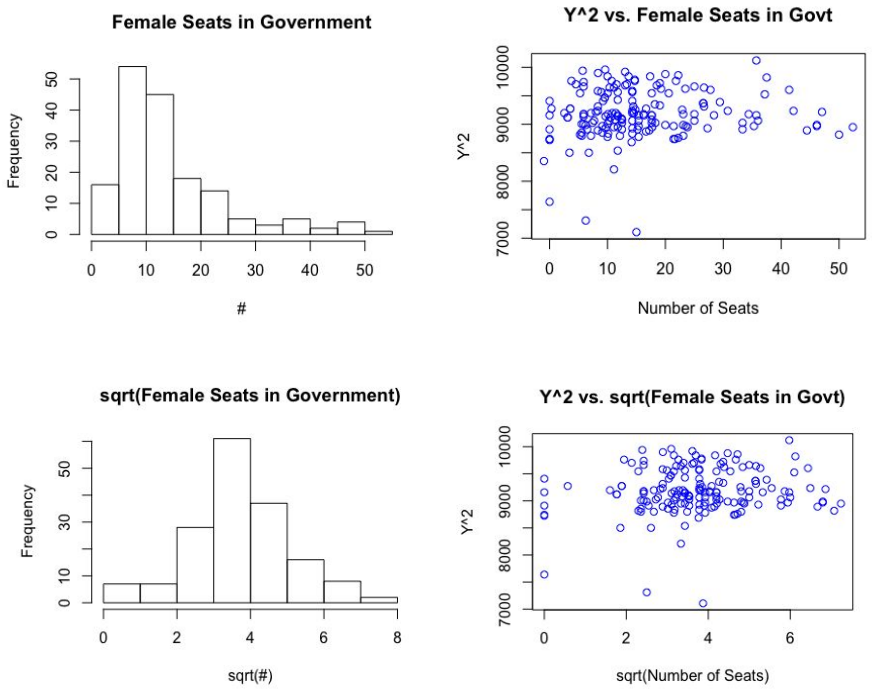


Figure 7

The female seats in government variable is somewhat right-skewed, so it benefits from a square root transformation, which improves the linearity of its relationship with Y.

GDP: Log transformation

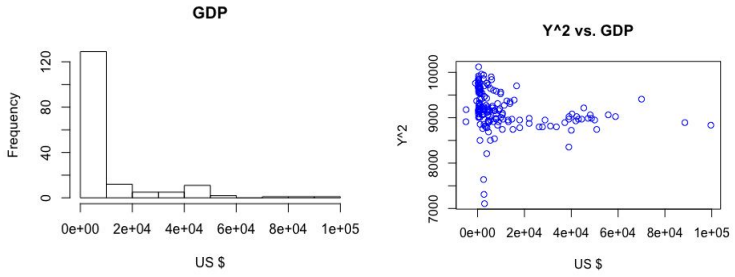


Figure 8.1

GDP is extremely right-skewed, so the log transformation improves its linearity with Y^2.

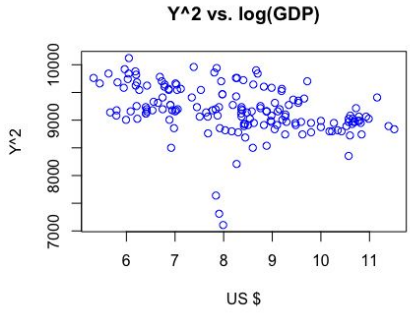


Figure 8.3

Teen Birth Rate: Log Transformation

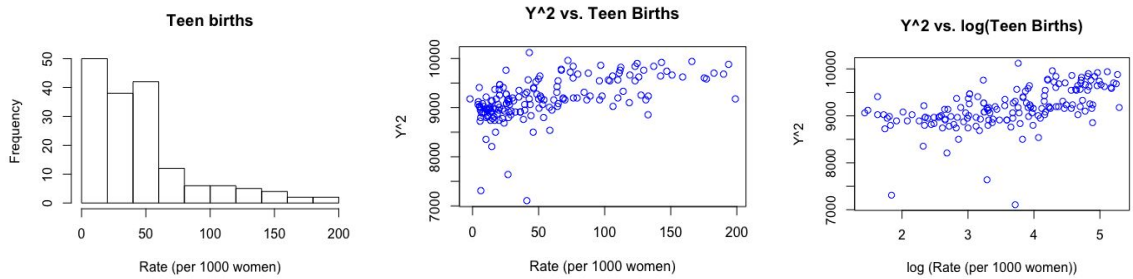


Figure 9

The teen birth rate is stacked close to 0, but the log transformation seems to reduce the concentration of the variable near 0 and improve linearity with Y.

Female Adult Obesity: Square root transformation

Figure 10

The square root transformation adjusts for the few large outliers in female adult obesity.

Abortion

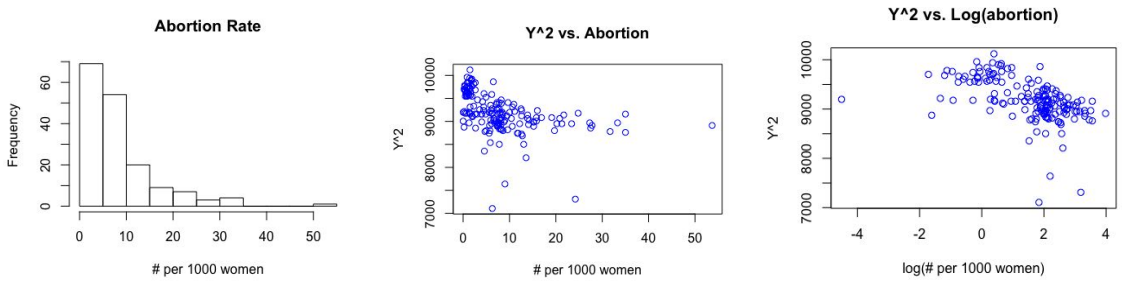


Figure 11

Abortion rate is the variable that was affected the most by using the truncated Normal distribution instead of a regular one, so it benefits from a log transformation that adjusts for the right skew.

3. Homoscedasticity: We can check this assumption by plotting the residuals against our fitted values from the step selection plot we select (titled "Model fit with data1")

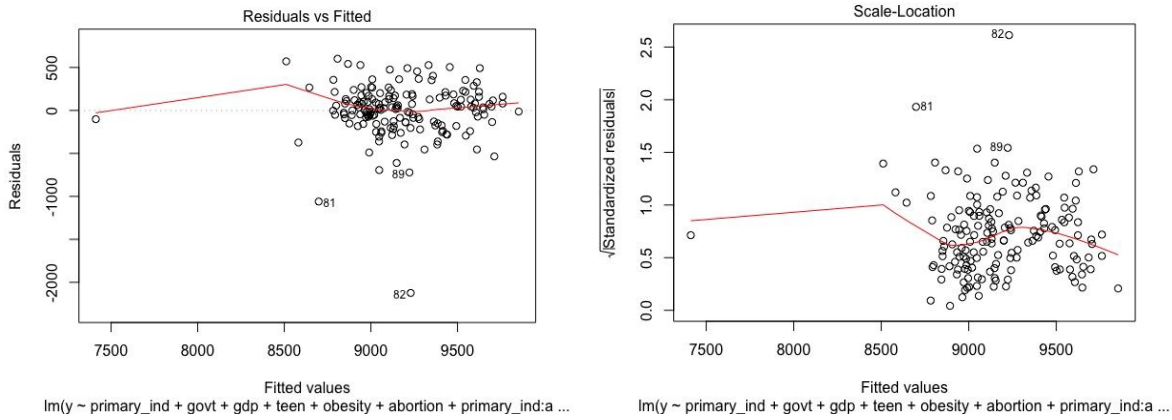


Figure 12

From the above plots, we can see our residuals are randomly assorted around the zero line, with the exception of the outlier on the far left. This outlier is Azerbijan, which has the lowest ratio of female to male births, and the lowest response value of Y for each value of X. Thus, it skews the normal distribution of Y data at each value of X, and the resulting residual distribution. We will leave this point in, because it is an important data point for trends in our set.

4. Independence of Data Points: The statistical independence of the errors is likely violated; this can be checked by study design, or model dependencies. Independence of errors is synonymous with independence of Y, because var(Yi|Xi) = var(Ei) = σ^2. It appears Y is not

independent, because there is extremely dense interaction of units, despite accounting for some extent of interactions, for ex; female enrollment rate in primary education may affect share of females in government, GDP per capita may affect prevalence of teenage births or female obesity rate. There is also spatial and temporal proximity, for ex; people in similar countries of the world living similar conditions, being proximal to one another's political beliefs, holding a high school educational degree, etc., and clustering, where communities exhibit similar demographic behavior and income level. However, much of this is accounted for in our imputed model, which accounts for similarities by region in extrapolating abortion rates; however, imputation is not used for other predictor variables, which could experience the same types of correlation. In addition, in our final step-wise model, we account for interactions between primary_ind:abortion + gdp:abortion + govt:obesity + govt:gdp + govt:teen, which eliminates some of this dependency.

III. Multiple Regression Model

To select the best model, we use one of the 20 datasets (datasets with 20 different estimates per region for missing values) for stepwise selection at a time.

Main Effect

Model:----------

lm(formula = y ~ gender_index + primary_ind + govt + gdp + teen + obesity + abortion)

Residuals:

Min	1Q	Median	3Q	Max
-2094.03	-148.80	34.71	219.32	666.95

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8986.491	321.534	27.949	< 2e-16	***
gender_index	-8.511	187.181	-0.045	0.963792	
primary_ind1	34.238	66.813	0.512	0.609056	
govt	52.125	21.731	2.399	0.017615	*
gdp	-35.390	31.093	-1.138	0.256744	
teen	152.522	41.691	3.658	0.000345	***
obesity	-31.253	26.126	-1.196	0.233391	
abortion	-14.654	3.788	-3.868	0.000160	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
Residual standard error: 354.6 on 159 degrees of freedom
Multiple R-squared:  0.3613,  Adjusted R-squared:  0.3331
F-statistic: 12.85 on 7 and 159 DF,  p-value: 4.955e-13
```

Stepwise Selection

Model:----------

Step: AIC=1939.45

y ~ primary_ind + govt + gdp + teen + obesity + abortion + primary_ind:abortion + gdp:abortion + govt:obesity + govt:gdp + govt:teen

Model fit with data1

lm(formula = y ~ primary_ind + govt + gdp + teen + obesity + abortion + primary_ind:abortion + gdp:abortion + govt:obesity + govt:gdp + govt:teen)

Residuals:

Min	1Q	Median	3Q	Max
-2033.91	-150.88	18.12	167.23	646.53

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8401.239	902.982	9.304	< 2e-16	***
primary_ind1	-184.271	81.059	-2.273	0.024382	*
govt	453.231	264.263	1.715	0.088330	.
gdp	81.032	78.432	1.033	0.303145	
teen	288.147	118.611	2.429	0.016269	*
obesity	-187.875	64.656	-2.906	0.004200	**
abortion	-102.951	27.165	-3.790	0.000215	***
primary_ind1:abortion	42.215	9.842	4.289	3.14e-05	***
gdp:abortion	6.128	3.367	1.820	0.070684	.
govt:obesity	45.513	17.878	2.546	0.011881	*
govt:gdp	-49.203	21.731	-2.264	0.024947	*
govt:teen	-56.025	35.387	-1.583	0.115409	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
Residual standard error: 321.2 on 155 degrees of freedom
Multiple R-squared:  0.4889,  Adjusted R-squared:  0.4526
F-statistic: 13.48 on 11 and 155 DF,  p-value: < 2.2e-16
```

20 Iterations: Why Did We Do It? How Does It Work?

In order to impute our values as the average of that region's, we need to average our regression model coefficients across n=20 datasets bootstrapped from our regional data.

1) Mean estimate of coefficients

```
> c(intercept,b1,b2,b3,b4,b5,b6,b7,b8,b9,b10,b11)
 [1] 8516.640969 -210.641876  433.429964   72.299418  267.868490
 [6] -175.447223 -104.960825   43.924211    6.156183   42.456614
[11]  -46.994640  -51.907868
```

2) SD of 20 estimates of each coefficient Beta_i for i = [1,12]

```
> sdlist
 [1] 132.4627118  15.9776191  42.6553786  10.3269147  18.0762893
 [6]   8.4603172   3.2340162   2.1141206   0.3728635   2.3239136
[11]   3.3749961   5.4934418
```

3)Average SD of 20 coefficient estimates: average of the 20 standard deviations of coefficients

```
> sdlist
 [1] 895.647059  81.765830 262.237715  77.899083 117.344903  63.878358
 [7]  27.023765   9.828923   3.347836  17.680048  21.584405  35.067017
```

$$SD_{\text{Adjusted}} = \sqrt{(\bar{\sigma}_{\beta 1})^2 + (\mu_{\sigma_\beta})^2}$$

4) Adjusted p-values

```
                mean_est adj_sd_coefs       tval      pvalues
(intercept)             8516.640969   905.389432  9.4066052 0.000000e+00 ***
```

```
primary_ind1           -210.641876    83.312275 -2.5283414 1.246077e-02 *
govt                    433.429964   265.684212  1.6313727 1.048419e-01
gdp                      72.299418    78.580611  0.9200669 3.589675e-01
teen                    267.868490   118.729013  2.2561334 2.546096e-02 *
obesity                 -175.447223    64.436182 -2.7228060 7.215196e-03 **
abortion                -104.960825    27.216589 -3.8565018 1.682346e-04 ***
primary_ind1:abortion    43.924211    10.053717  4.3689523 2.275709e-05 ***
gdp:abortion              6.156183     3.368535  1.8275547 6.953950e-02
govt:obesity             42.456614    17.832124  2.3809061 1.848451e-02 *
govt:gdp                 -46.994640    21.846673 -2.1511119 3.301678e-02 *
govt:teen                -51.907868    35.494698 -1.4624119 1.456530e-01
```

According to the adjusted p-values, the indicator variable for primary school enrollment gender ratio, the teen birth rate, the obesity rate, abortion rate, the interaction term between primary school indicator and abortion, the interaction between government share and obesity, and the interaction between government and GDP are significantly associated with the response variable, gender ratio at birth.

Analysis:

We report that an increased prevalence of teenage pregnancies is positively correlated with sex ratio at birth when controlled for all other predictors, which means the ratio is closer to 100 (50% male births, 50% female births), indicating more female births compared to the average because the ratio on average is skewed to a greater proportion of male births. Some studies suggest that proportion of male births increases with maternal age,[4,5] while other studies conclude there is no relationship.[6] However, incidences where the sex ratio is greater than 100 (more female live births than males) *are* associated with mothers under the age of 15.[7] Biologically speaking, this makes sense. There is evidence that female births are overrepresented in "stressful" pregnancies,[8] and teenage mothers are less likely to seek and receive prenatal care.[9]

[4] Takahashi E. The effects of the age of the mother on the sex ratio at birth in Japan. Ann N Y Acad Sci 1954;57:531-550.
[5] Hytten FE, Leitch I. The Physiology of Human Pregnancy. 2nd edn. Oxford, UK: Blackwell Scientific Publications Ltd; 1971.
[6] Rueness J, Vatten L, Eskild A (2012) The human sex ratio: effects of maternal age. Hum Reprod 27(1):283–287
[7] Mathews TJ, Hamilton BE. Trend analysis of the sex ratio at birth in the United States. Natl Vital Stat Rep 2005;53:1-17.
[8] Catalano R, Bruckner T, Anderson E, Gould JB. Fetal death sex ratios: a test of the economic stress hypothesis. Int J Epidemiol 2005;34:944-948.
[9] Makinson C. The health consequences of teenage fertility. Fam Plann Perspect. 1985 May-Jun;17(3):132–139.

We found a negative correlation between obesity rate and sex ratio at birth when controlled for all other predictors, meaning proportionally more males are born in countries with a greater obesity rate. When we chose this predictor, we intended it to serve as a proxy for measuring the average diet and abundance of food in a given country. Our findings are consistent with studies that show well-fed mothers giving birth to more males than females compared to less well-fed mothers.[10]

Abortion rate was also negatively correlated with sex ratio at birth when controlled for all other predictors. This is consistent with the observation that in countries like China and Azerbaijan, sex-selective abortions favor male births due to cultural preferences.[11]

The indicator variable for gender parity index of female enrollment in primary schools is negatively correlated to sex ratio at birth, when controlled for all other predictors. A possible explanation for this result is the evidence that higher education is correlated with greater maternal age and fewer teenage births,[12] and as provided above, the proportion of male births increases with maternal age.

An interaction term of significance worthwhile noting is `primary_ind:abortion`. The coefficient is 43.92, meaning that in countries where the primary school female to male ratio is greater than 0.95, the slope between abortion rate and gender birth ratio is less negative. We can draw the conclusion that in countries where women have more access to education, the prevalence of abortion is less associated with gender ratio at birth. A possible explanation for this might be that more access to education for women is an indicator for more equal societal status for women, therefore sex-selective abortion is less likely to take place regardless of abortion prevalence.

The gender disparity index predictor is not included in the final stepwise selection model, which is reasonable considering that the index is a holistic measure of female reproductive health, access to education and government leadership, and labor participation, which are already taken into account by the `primary_ind1`, `govt`, and `abortion` predictor variables.

[10] Mathews F, Johnson PJ, Neil A. You are what your mother eats: evidence for maternal preconception diet influencing foetal sex in humans. Proc Biol Sci. 2008;275:1661–68.
[11] Esfandiari G. Sex-Selective Abortions Take A Toll In Azerbaijan. *RadioFreeEurope*. Oct 15, 2016.
[12] Livingston G. For most highly educated women, motherhood doesn't start until the 30s. *Pew Research Center*. Jan 15, 2015.

Conclusion:

Note that our results do not in any way imply causation: in fact, it's very likely that they were caused by other proxy factors, such as the development level in a country. Instead, they're useful in implying association, which is still very interesting to understand from a statistical point of view.

Our study has many interesting implications. First, understanding predictors which cause Male/Female Birth ratio to change is very interesting in the context of early childhood development. For example, we found a statistically negative correlation between obesity and female/male birth ratio. This indicates that when doing a study about infant mortality deaths for an obesity-related reason by country, the results will be a bit skewed unless you control for gender levels.

Second, the significant negative association of abortion rates with gender birth ratio, even after factors such as GDP and indicators of general societal status of women are held constant, is interesting and worth further investigation, since a commonplace assumption would be that only in strongly patriarchal cultures would sex-selective abortion be prevalent.

Finally, examining predictors of the Female/Male Birth ratio is potentially useful many years in the future, if we have a situation where there's a major male/female birth imbalance. In this case, our research would need to be taken further, examining specific causes by doing more in depth studies about the predictors which we isolated. Thus, our research could help save the future of humanity!

References

Catalano R, Bruckner T, Anderson E, Gould JB. Fetal death sex ratios: a test of the economic stress hypothesis. Int J Epidemiol 2005;34:944-948.

Esfandiari G. Sex-Selective Abortions Take A Toll In Azerbaijan. *RadioFreeEurope*. Oct 15, 2016.

Hayruni, Mariam. "Thesis Paper: Sex-selective Abortions in Armenia." *Thesis Paper: Sex-selective Abortions in Armenia*. American University of Armenia, May 2013. Web. 7 Dec. 2016.

Hytten FE, Leitch I. The Physiology of Human Pregnancy. 2nd edn. Oxford, UK: Blackwell Scientific Publications Ltd; 1971.

List of Developing Countries. http://www.iugg2015prague.com/list-of-developing-countries.htm

Livingston G. For most highly educated women, motherhood doesn't start until the 30s. *Pew Research Center*. Jan 15, 2015.

Makinson C. The health consequences of teenage fertility. Fam Plann Perspect. 1985 May-Jun;17(3):132–139.

Mathews F, Johnson PJ, Neil A. You are what your mother eats: evidence for maternal preconception diet influencing foetal sex in humans. Proc Biol Sci. 2008;275:1661–68.

Mathews TJ, Hamilton BE. Trend analysis of the sex ratio at birth in the United States. Natl Vital Stat Rep 2005;53:1-17.

Rueness J, Vatten L, Eskild A (2012) The human sex ratio: effects of maternal age. Hum Reprod 27(1):283–287

Takahashi E. The effects of the age of the mother on the sex ratio at birth in Japan. Ann N Y Acad Sci 1954;57:531-550.

Wikipedia. List of Countries by Sex Ratio. https://en.wikipedia.org/wiki/List_of_countries_by_sex_ratio.