LIFE EXPECTANCY: CAN IT BE PREDICTED?

By John C. Erickson

Spring 2014

Statistics Project—STAT 350

Presented to Dr. Elizabeth Johnson

George Mason University

Background

- Many people & organizations are trying to help others either live longer and/or more productive lives
  - United Nations Millenium Development Goals
  - Bill & Melinda Gates Foundation
  - Gates' donates money for computers to schools
  - Clean water programs are being established in Africa and some parts of Asia
  - The Catholic church in Africa does a lot of medical work/hospitals,etc

- Many western nations have advanced medical care and a good economy
  - Yet it seems natural to think there are some relationships to be observed between life span and other variables
  - What countries have obviously low-life spans? What can they tell us?

Abstract

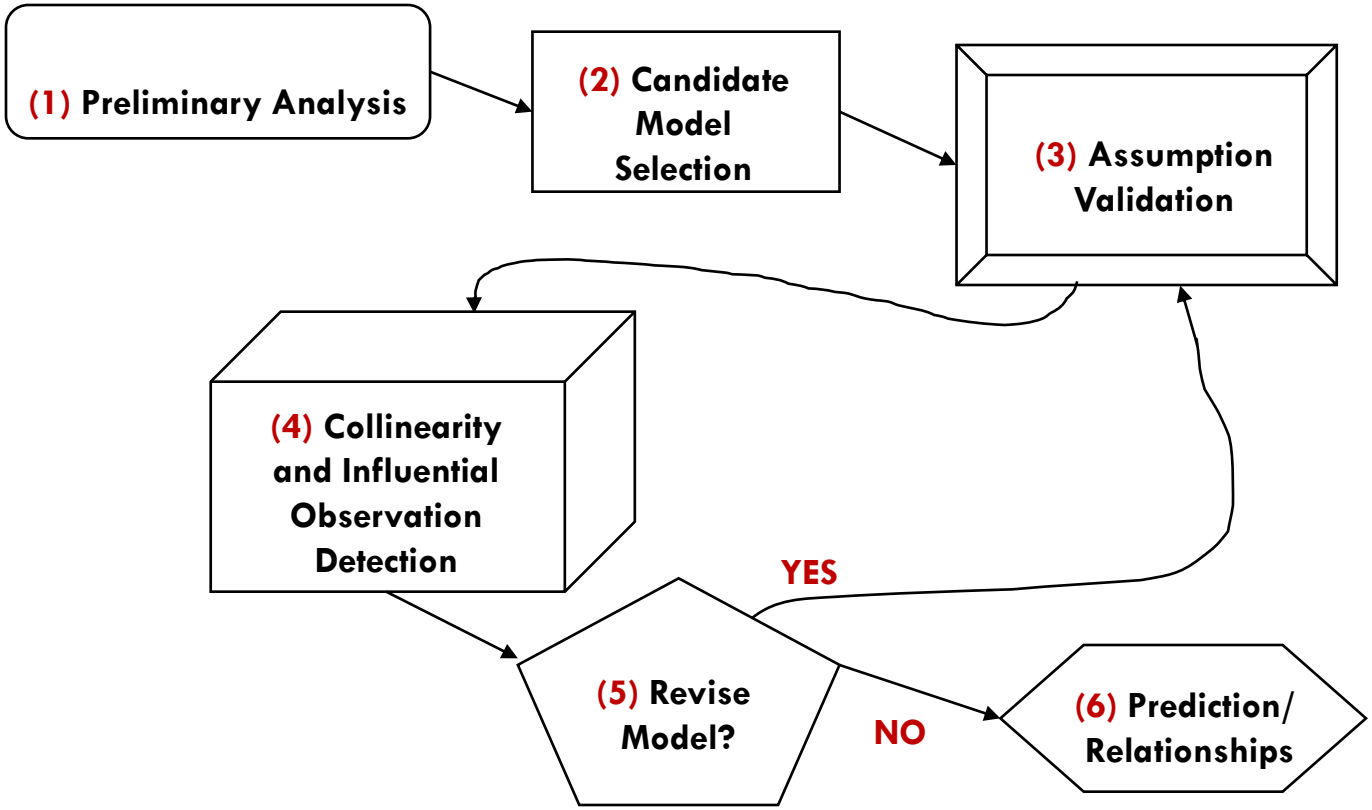
- Since most of us seem to care about the quality of life of others on this planet, it is a worthwhile goal to conduct analysis of various factors to see if lifespan can be predicted.

- **Hypothesis for this Study**: Life Span can be predicted
- **Key Questions for this Study**:
  - What factors predict life expectancy?
  - Does the country I live in, or other factors, affect my life span?
  - Does female literacy rate or GDP have anything to do with life span?

- To answer the above, we focus on three explanatory (independent) variables: 1) Birth Rate; 2) GDP; 3) Female Literacy Rate. The response (dependent) variable is life expectancy. The data collected is from over 220 countries.

Modeling Cycle*

Where do we begin?



-<u>The goal</u>: to predict life expectancy from GDP, female literacy rate, and/or birth rate. If the data doesn't let us do to not passing assumptions, produce descriptive statistics

*Flow chart taken from Dr. Larry Tang, STAT 362, Presentation on Stepwise Regression in SAS

Data Set

	Country	Birth Rate	GDP	GDP (Billions)	Life Exp.	Female Literacy
1	Afghanistan	38.84	$45,300,000,000.00	$45.30	50.49	12.6
2	Albania	12.73	$26,730,000,000.00	$26.73	77.96	95.7
3	Algeria	23.99	$284,700,000,000.00	$284.70	76.39	63.9
4	American Samoa	22.87	$575,300,000.00	$0.58	74.91	97.0
5	Andorra	8.48	$3,163,000,000.00	$3.16	82.65	100.0

217	West Bank	23.41	$8,022,000,000.00	$8.02	75.69	92.6
218	Western Sahara	30.71	$906,500,000.00	$0.91	62.27	70.0
219	Yemen	31.02	$61,630,000,000.00	$61.63	64.83	48.5
220	Zambia	42.46	$25,470,000,000.00	$25.47	51.83	51.8
221	Zimbabwe	32.47	$7,496,000,000.00	$7.50	55.68	80.1

•There are 221 countries in the total data set (see exhibit B, "Data Set" attached). The data was collected at the CIA Worldfactbook website located at https://www.cia.gov/library/publications/the-world-factbook/

• The data was collected by manipulating data in Excel and then running descriptive statistics and regression in Minitab.

Preliminary Analysis

Descriptive Statistics from Minitab

Descriptive Statistics: Birth Rate, GDP (Billions), Life Exp., Female Literacy

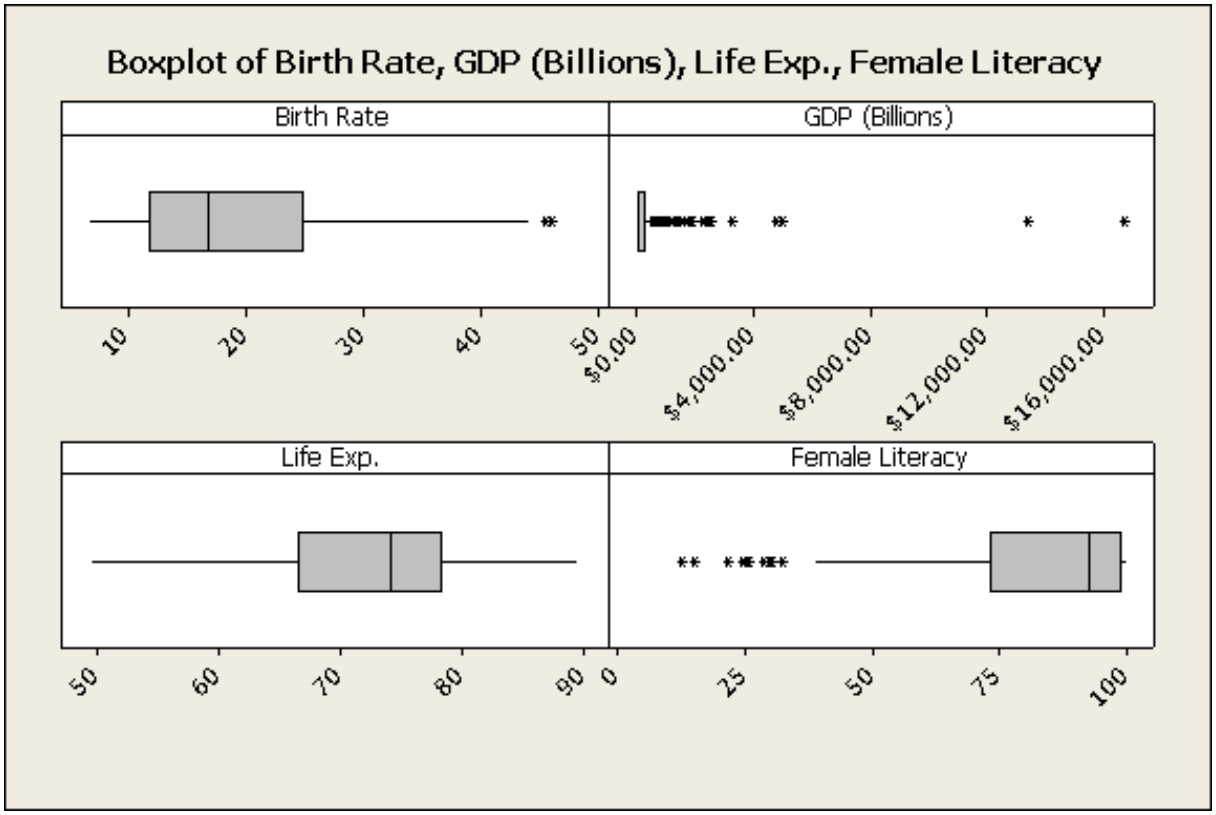
Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median
Birth Rate	221	0	19.562	0.646	9.602	6.720	11.810	16.880
GDP (Billions)	221	0	394	105	1555	0	5	32
Life Exp.	221	0	71.749	0.594	8.829	49.440	66.625	74.250
Female Literacy	221	0	83.07	1.41	20.93	12.60	73.10	92.30

Variable	Q3	Maximum	Range	IQR
Birth Rate	24.825	46.120	39.400	13.015
GDP (Billions)	198	16720	16720	192
Life Exp.	78.295	89.570	40.130	11.670
Female Literacy	98.65	100.00	87.40	25.55

-The above descriptive statistics include the IQR (in case data is not normal). See next slide for each variable's boxplot distribution and description

Preliminary Analysis

Shape, Center, Spread



Interpretations:*

•Birth Rate is right skewed; median births (per 1000 people) is 16.88 children; IQR = 13.015

•GDP(billions) is strongly right skewed ; median GDP of $32 billion; IQR = $192 billion

•Life Expectancy is slightly left skewed ; with a median age of 74.25 years; IQR = 11.67

• Female literacy rate is strongly left skewed; median percentage is 92.30; IQR = 25.55.

*Since the variables are all non-normal, the normal description (shape, center, and spread) of the data will be modified to direction of skew, median value, and the range/IQR.

Correlation

Correlations: Birth Rate, GDP (Billions), Life Exp., Female Literacy

	Birth Rate	GDP (Billions)	Life Exp.
GDP (Billions)	-0.151		
Life Exp.	-0.849	0.137	
Female Literacy	-0.789	0.101	0.701

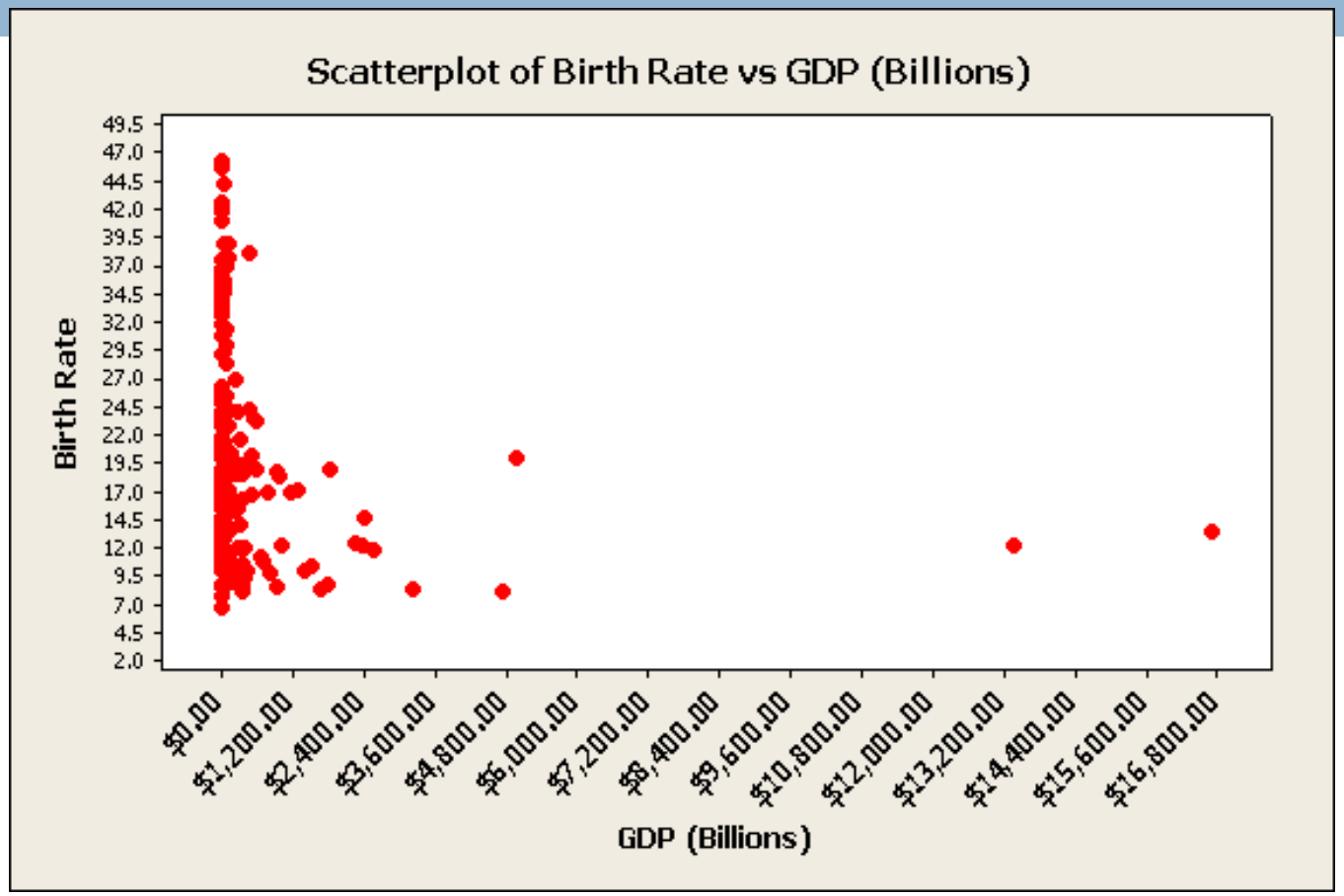
Cell Contents: Pearson correlation

Interpretation:
-There appears to be a ….
-weak, negative correlation between GDP and Birth Rate.
-strong, negative correlation between Life Expectancy and Birth Rate.
-strong, negative correlation between Female Literacy Rate and Birth Rate.
- weak, positive correlation between Life Expectancy and GDP
- weak, positive correlation between Female Literacy and GDP.

-Conclusion: There appears to be correlation between life expectancy and birth rate and between life expectancy and female literacy. (See scatterplots on the next slides).

Preliminary Analysis:

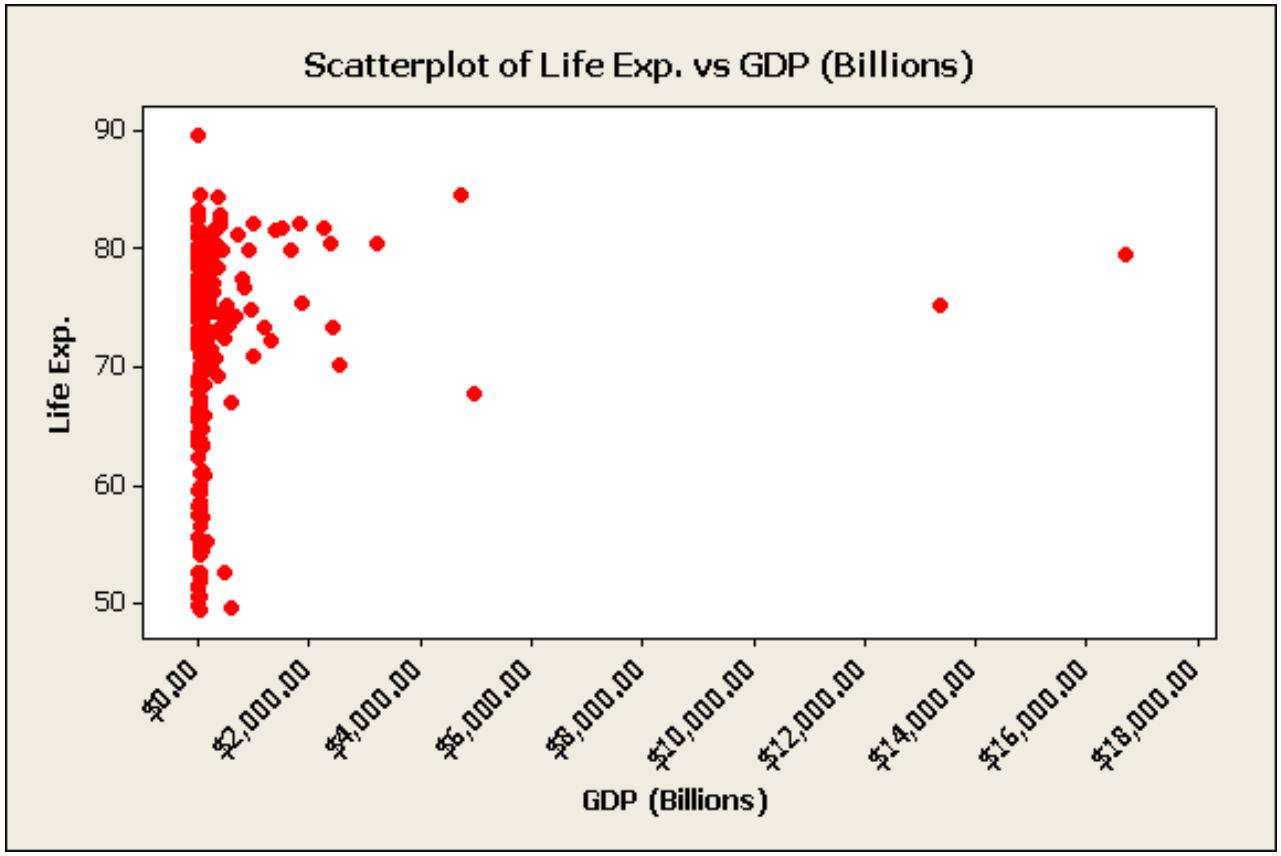
Scatterplot of Birth Rate vs GDP (Billions)



The above scatterplot is not normal. It is non-linear. This fails to meet the assumption for simple and multiple regression.

Preliminary Analysis:

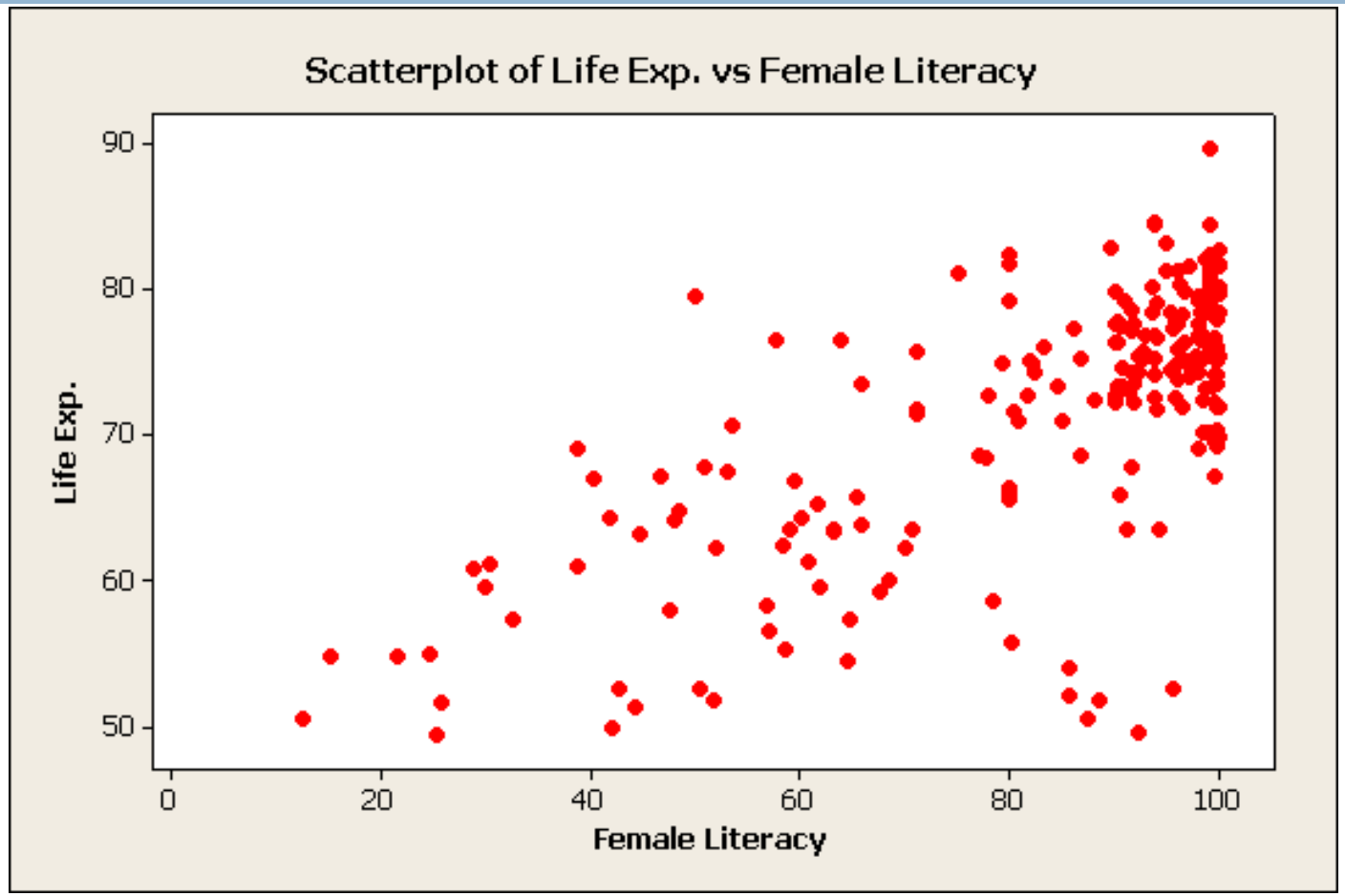
Life Expectancy vs GDP (Billions)



The above scatterplot is not normal. It is non-linear. This fails to meet the assumption for simple and multiple regression.

Preliminary Analysis:

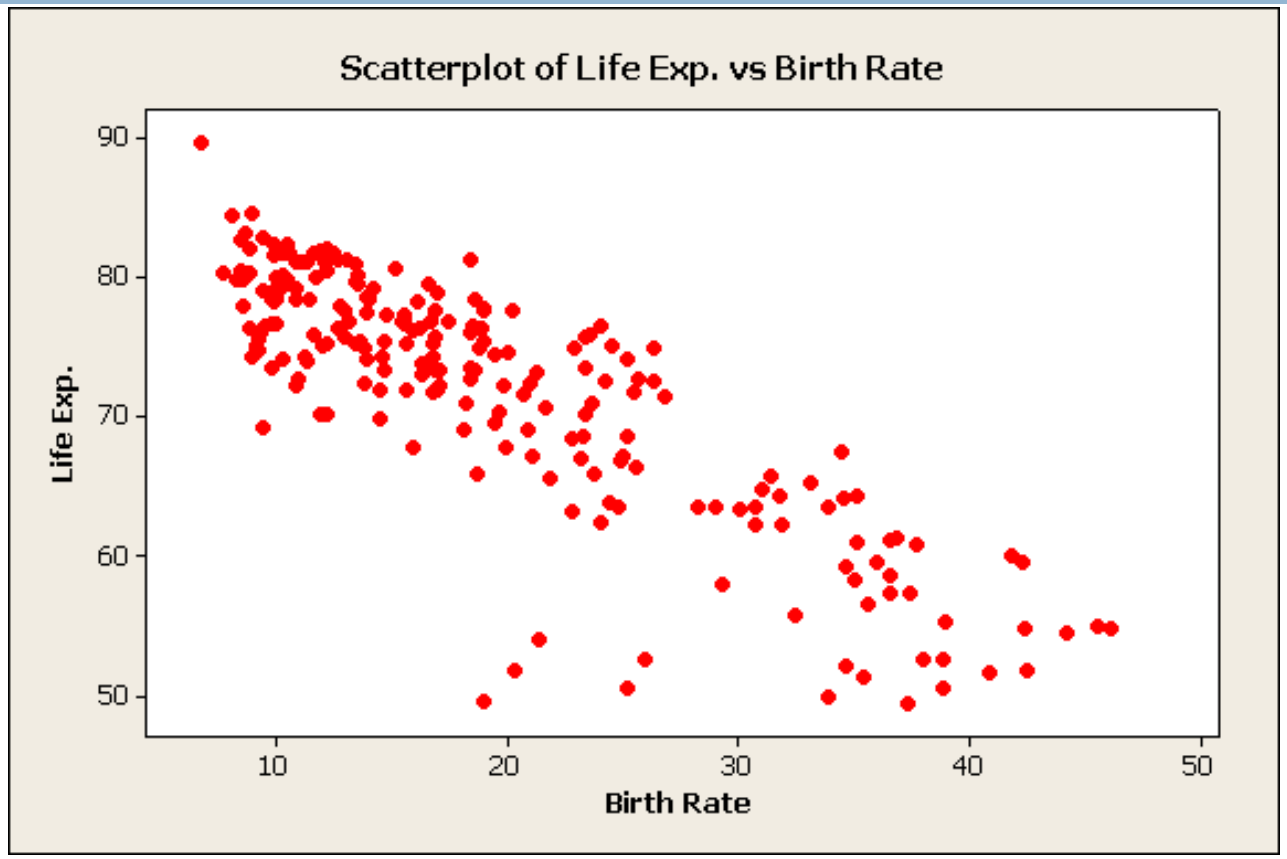
Life Expectancy vs Female Literacy Rate



The above scatterplot is not normal. It is non-linear. This fails to meet the assumption for simple and multiple regression. However, there is a noticeable cluster of high literacy rate nations with higher life expectancy nations.

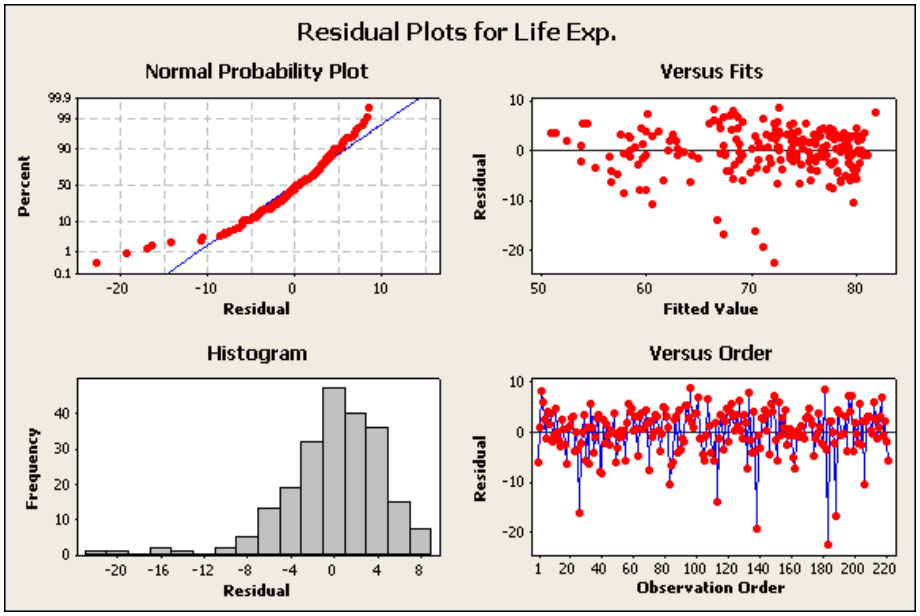
Preliminary Analysis:

Life Expectancy vs Birth Rate



The above scatterplot appears to be quasi-linear or show some promise of a relationship.

Assumptions Check



Simple Regression Assumptions

- The mean of the probability distribution of ε is 0.
- The variance of the probability distribution of ε is constant for all values of x.
- The probability distribution of ε is normal.
- The values of ε associated with any two observed values of y are independent.

The residual plots above show cause for concern. There is some minor to medium curvature in the normal probability plot and some minor appearance of a trend in the variance of the residuals (top right corner). We will run the simple regression between life expectancy and birth rate because it is the most normal of the scatterplots.

Simple Regression

(if it were normal)

Regression Analysis: Life Exp. versus Birth Rate

The regression equation is
Life Exp. = 87.0 - 0.781 Birth Rate

Predictor	Coef	SE Coef	T	P
Constant	87.0268	0.7143	121.84	0.000
Birth Rate	-0.78099	0.03279	-23.82	0.000

S = 4.67022 R-Sq = 72.1% R-Sq(adj) = 72.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	12372	12372	567.25	0.000
Residual Error	219	4777	22		
Total	220	17149			

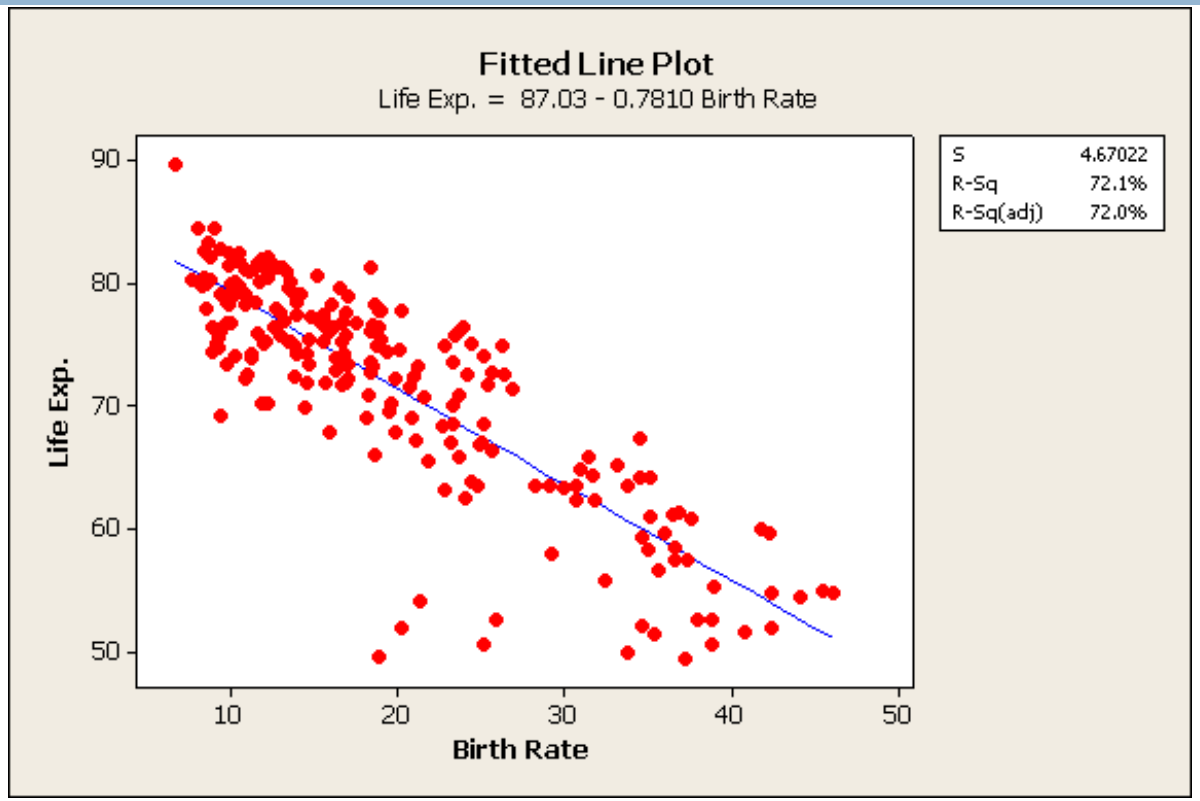
Unusual Observations

Obs	Birth Rate	Life Exp.	Fit	SE Fit	Residual	St Resid
27	21.3	54.060	70.360	0.320	-16.300	-3.50R
32	42.4	54.780	53.897	0.813	0.883	0.19 X
34	42.3	59.550	53.968	0.810	5.582	1.21 X
84	33.8	49.870	60.606	0.564	-10.736	-2.32R
113	25.9	52.650	66.784	0.377	-14.134	-3.04R
122	41.8	59.990	54.381	0.794	5.609	1.22 X
125	45.5	54.950	51.468	0.908	3.482	0.76 X
138	20.3	51.850	71.188	0.315	-19.338	-4.15R
145	46.1	54.740	51.008	0.926	3.732	0.82 X
183	18.9	49.560	72.235	0.315	-22.675	-4.87R
188	25.2	50.540	67.361	0.364	-16.821	-3.61R
205	44.2	54.460	52.531	0.866	1.929	0.42 X
206	9.4	69.140	79.678	0.458	-10.538	-2.27R
220	42.5	51.830	53.866	0.814	-2.036	-0.44 X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.

-The regression equation is Life Expectancy = 87.0 – 0.781*Birth Rate. There are a number of unusual observations, which is not surprising given the data is not normally distributed. Note the "R" and "X" values in the residuals column above.

Fitted Line Plot: Simple Regression



- There is a strong indicator that many of the countries with high birth rates in the data DO show a low life expectancy.

Possible Confounding Variables For the Simple Regression Model

- Healthcare /Hospital Care
- Women in the workforce
- War / Conflict areas
- Drinking Water (potable water by region)
- Sanitary Conditions
- HIV / Diseases
- Education level (by gender)

-This is a complex issue that most likely cannot be predicted with just birth rates. However, it is our opinion that the birth rates are indicators of something else: the level of healthcare and other factors listed above that all potentially affect lifespan.

Other potential models?

- Even if we added many more variables to avoid confounding/lurking variables, the variables will be likely non-normally distributed. This would require a non-parametric test(s).
- Non-parametric regression possibilites..
  - Regression trees and splines
  - Gaussian / Kriging
  - Penalized Least Squares
  - Kernel regression
  - Multiplicative regression

-Since the simple regression was borderline unacceptable due to some problems with passing the assumptions, we can analyze the data by way of descriptive statistics

Descriptive Statistics: High and Low Countries' Birth Rate and Life Expectancy

Highest Life Expectancy Countries

Country	Life Exp.
Monaco	89.57
Macau	84.48
Japan	84.46
Singapore	84.38
San Marino	83.18
Hong Kong	82.78
Andorra	82.65
Guernsey	82.39
Switzerland	82.39
Australia	82.07

Lowest Life Expectancy Countries

Country	Life Exp.
Chad	49.44
South Africa	49.56
Guinea-Bissau	49.87
Afghanistan	50.49
Swaziland	50.54
Central African Republic	51.35
Somalia	51.58
Zambia	51.83
Namibia	51.85
Gabon	52.06

-The lowest life expectancy countries are all in Africa or Afghanistan

Highest and Lowest Birth Rates

Highest Birth Rate Countries

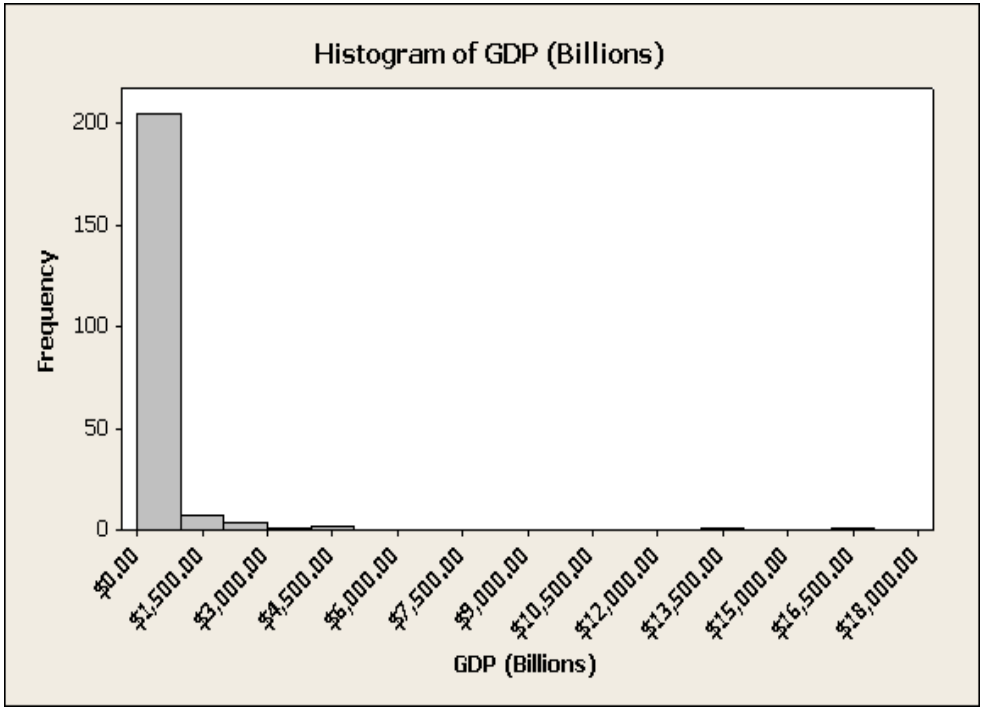
Country	Birth Rate
Niger	46.12
Mali	45.53
Uganda	44.17
Zambia	42.46
Burkina Faso	42.42
Burundi	42.33
Malawi	41.8
Somalia	40.87
Angola	38.97
Afghanistan	38.84

Lowest Birth Rate Countries

Country	Birth Rate
Monaco	6.72
Saint Pierre and Miquelon	7.7
Japan	8.07
Singapore	8.1
Korea, South	8.26
Germany	8.42
Andorra	8.48
Slovenia	8.54
Taiwan	8.55
San Marino	8.7

-The highest birth rate countries are mainly African. This is important because the previous slide had many African countries as low life expectancy.

Descriptive Statistics: GDP



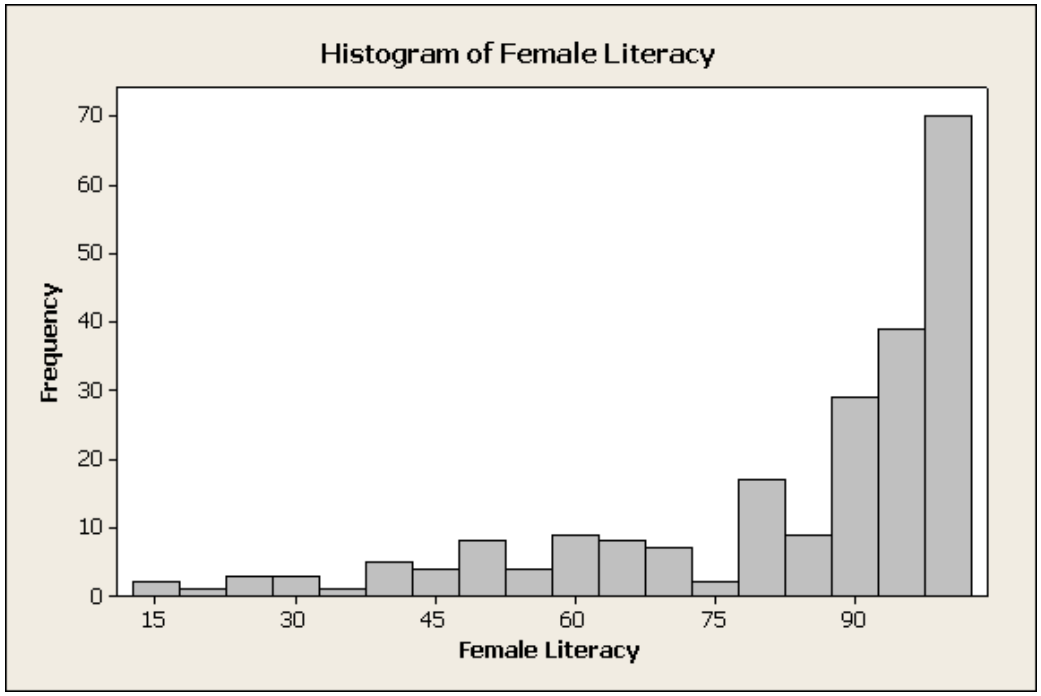
Highest GDP

Country	GDP
United States	$16,720,000,000,000.00
China	$13,370,000,000,000.00
India	$4,962,000,000,000.00
Japan	$4,729,000,000,000.00
Germany	$3,227,000,000,000.00
Russia	$2,553,000,000,000.00
Brazil	$2,422,000,000,000.00
United Kingdom	$2,378,000,000,000.00
France	$2,273,000,000,000.00

Lowest GDP

Country	GDP
Saint Helena, Ascension, and Tristan da Cunha	$31,100,000.00
Tuvalu	$40,000,000.00
Montserrat	$43,780,000.00
Wallis and Futuna	$60,000,000.00
Nauru	$60,000,000.00
Anguilla	$175,400,000.00
Cook Islands	$183,200,000.00
Saint Pierre and Miquelon	$215,300,000.00
Palau	$245,500,000.00
Sao Tome and Principe	$421,000,000.00

Descriptive Statistics: Literacy



-The Female Literacy rate in the world is left skewed.

Lowest Female Literacy Rates

Country	Female Literacy
Afghanistan	12.6
Niger	15.1
Burkina Faso	21.6
Mali	24.6
Chad	25.4
Somalia	25.8
Ethiopia	28.9
Guinea	30
Benin	30.3
Sierra Leone	32.6

Highest Female Literacy Rates

Country	Female Literacy
Andorra	100
Austria	100
British Virgin Islands	100
Cook Islands	100
Finland	100
Greenland	100
Korea, North	100
Liechtenstein	100
Luxembourg	100
Norway	100

Limitations of this Study

- Non-normality of data
- Too little variables to accurately answer the original question. This topic requires much more data and many more explantory variables.
- Skills of the student not enough for non-parametric regression

Interesting Areas for Further Research

(Life Expectancy)

- Human Development Index (HDI) developed by the UN.
- UN Millenium Development Goals
- Clean cook stoves (many women die of cooking food over dung-fires, and their children are exposed to it)
- Health care in Africa—it does seem the lowest life spans are mostly African countries and yet they have the most children

Lessons Learned

- The real world has “messy” data. This study proved to be no exception to that rule. Collecting the data was easy; the hard part is cleaning the data for statistical modeling. The most I could do in this study was descriptive statistics. If I had more time, I could break the countries down by continent, region, or by GDP groupings, and do more analysis by those groups. Once they were in those groups, I could try to do see if the data became linear in the scatterplots and run two-way ANOVA tests for means.

- Also, I stumbled upon the United Nation’s Human Development Index, which takes into account almost all the necessary variables required for a truly meaningful statistical study into life expectancy. I did not realize that my interest in life expectancy, GDP, birth rate, etc are something that the United Nations looks at very seriously each year and the HDI is a very good tool when looking at the world’s countries.

- I would like to seriously consider taking additional classes in non-parametric statistics. Since real-world data seems to be “messier” than in-class examples, this is probably a good skill to have (at least for predictions).

Conclusion

- Birth rate is a good *indicator* of life expectancy, but NOT a good *predictor*. With more variables and advanced regression techniques, a good statistical model for prediction could be found.

References

- https://www.cia.gov/library/publications/the-world-factbook/rankorder/2054rank.html (birth rate)

- https://www.cia.gov/library/publications/the-world-factbook/rankorder/2001rank.html (gdp)

- https://www.cia.gov/library/publications/the-world-factbook/rankorder/2102rank.html (life expectancy)

- https://www.cia.gov/library/publications/the-world-factbook/fields/2103.html (female literacy rate)

- http://www.cleancookstoves.org/ (clean cook stoves)

- http://hdr.undp.org/en/statistics/hdi (human development index)