Jonathan Lam
Mr. Rivers
H. Alg. II p. 5
4 / 23 / 16

Food = Longevity

Modeling Project II: Function Types, Correlations, and Regressions

We Americans eat a lot. It's a fact nowadays— with our ethnic "American food" being junk food, it has become an important part of culture. Yet although junk food diminishes our health, it is counteracted by positive health benefits such as a lack of hunger and improved technology and medicine. And because of the latter, we Americans have been living longer and healthier lives, despite the increasing "junk" that we are stuffing into our bodies. I thought it would be interesting to see the relation between the average amount a country eats bersus how healthy they are, on average. I decided that, the more a country eats on average— junk food included— there would be a generally positive correlation. However, because of the increased prevalence of junk food and the "obesity epidemic," I figured that a power function would fit the data the best— in other words, when the average person eats a lot, then their health would not increase by as much, or perhaps it would actually slightly decrease.

I garnered up data charts from Wikipedia for average daily food consumption in Calories (kilocalories, or dietary calories) per country and average life expectancy— two pieces of data that I thought would accurately reflect the information I wanted to collect— and compared them for this modeling project. The data is all from 2015, and I only included the countries which existed in both data charts— therefore, the data may be slightly off because of the possible lack of some smaller, unknown countries.

[See data table, attached]

I didn't rule out any of the types of regressions except the inverse exponential function— an exponential decay— at first. An exponential decay would make no sense, because more food would not decrease life expectancy. However, a power function, should make the most sense: if the average person does not eat any food (no Calorie intake), then they would not live for any years. Therefore, it would pass through the origin of the graph, the point (0, 0). This would be highly unlikely, but true. An inverse power function would also show the kind of curve that I had predicted: it would quickly grow (as any kind of nourishment would greatly improve longevity, but food in excess would do little to increase longevity, especially with the modern diseases associated with or related to obesity. I thought it would be similar for logarithmic functions, because they can have a similar curve; however, it has a positive x-intercept, which means that eating a few Calories a day would result in a life expectancy of 0— this slightly shifts the graph so that low Calorie intakes are negative (which is impossible). This creates a slight inconsistency in the graph, but it is still very possibly a good fit for a graph because such a low Calorie intake or such a low life expectancy are very unlikely (and outside of my data range). I also considered linear functions, because of the unrealisticness of very low data points: this would give a positive y-intercept, which means that eating nothing will give a positive life expectancy, but this is out of the probable range. However, because there will be a generally

positive correlation, and because the graph is unlikely to curve too much as a result of damaging foods, a linear function is still a good candidate. An exponential growth function would be the least probable, but still possible nonetheless— if I had guessed incorrectly, and increased food caused a quickly increasing life-expectancy, then it would be the case. However, I deemed this unlikely.

[See graphs, attached]

All four graphs were very close with their r-values. The closest was the logarithmic function y=31.79ln(x)-180.1, with a correlation coefficient of 0.7032. However, it was less than 0.0078 (less than one one-hundredth) away from the power, linear, and exponential regressions: 0.7019, 0.0.7002, and 0.6954, respectively. The data wasn't very strong, hence the medium-strength correlation coefficients, but it was still arranged in an easily visible positive correlation. They are all closely arranged in an almost-linear fashion, very close at the center of the data, but slightly branching off at the beginning and end. Although it has unrealistic values near zero, I will use it for the following predictions.

To test the strength of the graph, I plugged in some x values (daily Calorie consumption) to see if the y values (life expectancy) were reasonable. For example, if a person eats approximately two thousand Calories a day (the recommended daily amount), then they should have a 62 year life expectancy.

$$y = 31.79ln(2000) - 180.1 = 31.79(7.601) - 180.1 = 241.6 - 180.1 \approx 62$$

If they eat as much as the average American at three thousand seven hundred fifty Calories a day, then they would live around 82 years old.

$$y = 31.79ln(3750) - 180.1 = 31.79(8.320) - 180.1 = 261.6 - 180.1 \approx 82$$

This means that at a life expectancy of 79 years old, the American life expectancy is actually three years under the one that the data predicts. Both of these lend us reasonable results— but if the domain is very far in either direction, the results are not so accurate. For example, with a very large Calorie intake, we would be living very long lives, according to the model: ten thousand Calories a day, for instance, would give us a 113 year life expectancy.

$$y = 31.79ln(10000) - 180.1 = 31.79(9.210) - 180.1 = 292.8 - 180.1 \approx 113$$

According to Wikipedia, there are only seventeen living people 113 years-old or older, and no amount of food could possibly guarantee such an old age. On the other hand, this model is also improbable at very Calorie intakes. Zero Calories isn't even possible, for example: there is a vertical asymptote at 0, which means that the y-intercept is effectively negative infinity. At very small Calorie intakes, there are very negative values, which are impossible: for example, at 1 Calorie a day there would be an impossible life expectancy of negative one hundred eighty.

$$y = 31.79ln(1) - 180.1 = 31.79(0) - 180.1 \approx \ -180$$

Therefore, there is a small domain in which the values are realistic, approximately one thousand five hundred to four thousand daily Calories, in which all the data lies. Therefore, it

would be reasonable to interpolate and extrapolate only slightly out of the range of the data before the predictions become unreasonable.

The data shows an obvious positive correlation, but the best type of regression is inconclusive. The correlation coefficients were all tightly grouped around 0.7, with a range of only 0.0078— this is a tiny difference that may be affected by the missing countries (the countries not in one or both of the data charts from Wikipedia). It appears that the data is linear, but the correlation coefficient for the logarithmic and inverse power functions are closer, but only by a tiny amount. Because of this ambiguity between the type of graph, and because of the medium correlation coefficient, I am not confident with the results of this graph. The data is shaped too much in a general blob that can be represented by all of these graphs, and there are numerous outliers that mess with the results. Additionally, I know that food consumption is not the only factor that affects health, nor is life expectancy the only indicator of health. Also, this data is subject to change from year to year, because it is statistical data based on billions of people. The model that I have created may be helpful to make some interpolations, but far extrapolations would probably not be too accurate.

I have never done a statistical analysis, nor have I ever used data that has involved so many people and countries. I thought it was interesting to see from statistical data that in the real world, math is not so obvious, nor is it as perfect in its relationships as in math class. It also shows the importance of taking multiple trials or data samples in order to even out the outliers— in this case, for example, I could include the data points from other years to see if more data should make my data closer to a certain type of regression, or I could just run the same regressions on different years and average the results. To further my research on the initial question, I could also work with trying different international data sets that may have been impacted by food consumption, or that may impact life expectancy, such as average height per country, GDP per capita, happiness level on average, or percentage of deaths by cardiovascular disease. If I ran all these regressions, then it could perhaps lead to a comprehensive food statistical analysis, which could inspire me to study many more different aspects. On the other hand, it might also be useful to break it down instead and compare more specific statistics, such as rates of junk food consumption, vegetable consumption, meat consumption, or consumption of sugary drinks instead of the more generic energy from food consumption.

Data

- Wikipedia: List of Countries by Daily Calorie Intake (2014)
- Wikipedia: List of Countries by Life Expectancy (WHO, pub. 2015)

Country	**Calorie Intake** (x)	**Life Expectancy** (y)
Eritrea	1590	64
Burundi	1680	56
Comoros	1840	62
Haiti	1850	63
Zambia	1880	58
Ethiopia	1950	65
Central African Republic	1960	51
Angola	1960	52
Chad	2010	52
United Republic of Tanzania	2020	63
Timor-Leste	2020	67
Kenya	2030	61
Yemen	2050	64
Mozambique	2070	54
Rwanda	2090	65
Bolivia	2100	68
Democratic People's Republic of Korea	2110	70
Sierra Leone	2120	46
Madagascar	2130	64
Togo	2150	58
Malawi	2150	60
Guatemala	2150	72
Cambodia	2180	73
Tajikistan	2190	69
Liberia	2200	62
Zimbabwe	2210	59
Uganda	2220	59
Botswana	2230	64

Cameroon	2240	57
Lao People's Democratic Republic	2240	66
Mongolia	2240	68
Armenia	2260	71
Guinea-Bissau	2270	54
Bangladesh	2270	71
Dominican Republic	2270	74
Sudan	2280	63
Senegal	2280	64
Pakistan	2280	66
Swaziland	2290	53
Djibouti	2300	62
Ecuador	2300	76
Gambia	2330	61
Antigua and Barbuda	2330	75
Nepal	2340	68
India	2360	66
Namibia	2360	68
Sri Lanka	2370	75
Niger	2390	59
Solomon Islands	2400	69
Grenada	2400	73
Seychelles	2400	74
Peru	2410	77
Nicaragua	2420	74
Panama	2450	77
Lesotho	2460	50
Saint Kitts and Nevis	2460	74
Suriname	2460	77
Côte d'Ivoire	2500	53
Benin	2510	59
Thailand	2540	75
Guinea	2550	58

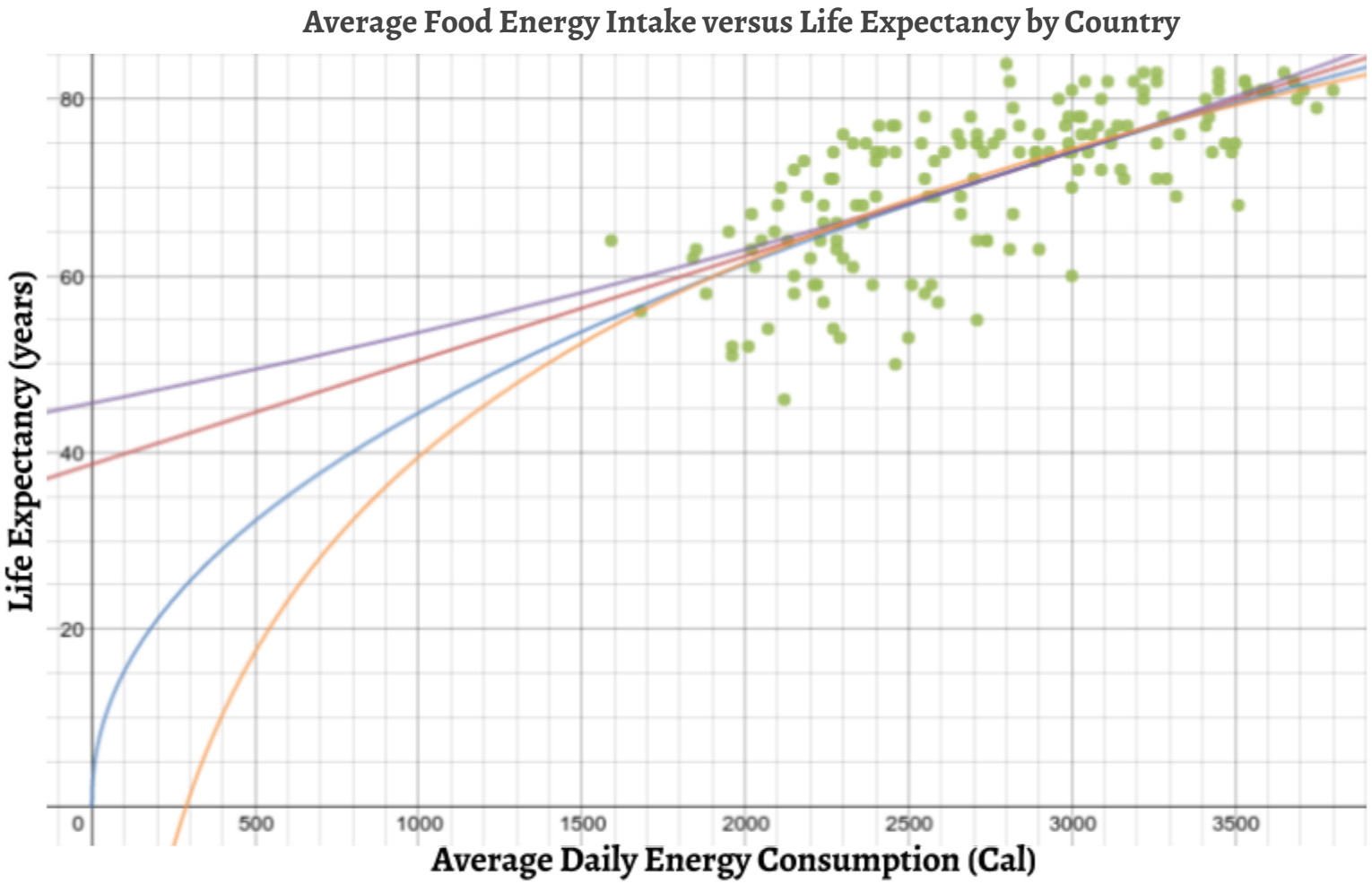
Indonesia	2550	71
Maldives	2550	78
Uzbekistan	2560	69
Congo	2570	59
Philippines	2580	69
El Salvador	2580	73
Mali	2590	57
Honduras	2610	74
Venezuela	2650	76
Sao Tome and Principe	2660	67
Kyrgyzstan	2660	69
Paraguay	2660	75
Colombia	2690	78
Trinidad and Tobago	2700	71
Nigeria	2710	55
Gabon	2710	64
Saint Lucia	2710	75
Belize	2710	75
Bahamas	2710	76
Georgia	2730	74
Turkmenistan	2740	64
Guyana	2740	64
Bulgaria	2760	75
Vietnam	2780	76
Japan	2800	84
Mauritania	2810	63
New Zealand	2810	82
Kiribati	2820	67
Costa Rica	2820	79
Jamaica	2840	74
Uruguay	2840	77
Samoa	2890	73
Albania	2890	74

Malaysia	2890	74
Ghana	2900	63
Slovakia	2900	76
Mauritius	2930	74
Chile	2960	80
Brunei Darussalam	2980	77
Latvia	2990	74
China	2990	75
Croatia	2990	78
South Africa	3000	60
Fiji	3000	70
Jordan	3000	74
Netherlands	3000	81
Azerbaijan	3020	72
Barbados	3020	78
Argentina	3030	76
Kuwait	3030	78
Republic of Korea	3040	82
Iran	3050	74
Macedonia	3060	76
Bosnia and Herzegovina	3080	77
Algeria	3090	72
Lebanon	3090	80
Sweden	3110	82
Brazil	3120	75
Saudi Arabia	3120	76
Estonia	3140	77
Belarus	3150	72
Egypt	3160	71
United Arab Emirates	3170	77
Cyprus	3190	82
Slovenia	3220	80
Finland	3220	81

Australia	3220	83
Morocco	3260	71
Mexico	3260	75
Iceland	3260	82
Spain	3260	83
Czech Republic	3280	78
Ukraine	3290	71
Russian Federation	3320	69
Tunisia	3330	76
Poland	3410	77
Denmark	3410	80
Cuba	3420	78
Lithuania	3430	74
United Kingdom	3450	81
Norway	3450	82
Switzerland	3450	83
Hungary	3470	75
Romania	3490	74
Turkey	3500	75
Kazakhstan	3510	68
France	3530	82
Israel	3530	82
Canada	3530	82
Germany	3540	81
Portugal	3580	81
Ireland	3590	81
Malta	3600	81
Italy	3650	83
Luxembourg	3680	82
Belgium	3690	80
Greece	3710	81
United States of America	3750	79
Austria	3800	81

Graphs

See graph on Desmos.



Regressions

Note: Regressions were calculated on different online calculators, not Desmos, because of its maximum data capacity. Correlation coefficients (r) were calculated from r^2 values given from those sites (using the square root).

Type	Color (in graph)	Correlation Coefficient (r)	Equation
Linear	Red	0.7002	$y = 0.01178x + 38.68$
Power	Blue	0.7019	$y = 1.812x^{0.4634}$
Exponential	Purple	0.6954	$y = 45.59e^{0.0001617x}$
Logarithmic	Orange	0.7032	$y = 31.79ln(x) - 180.1$