Math 76 Final Project

Raymond Chen (Worked with Ruiheng Ma)

August 31, 2020

1 Introduction

For my project, I tried to answer the question "What is the relationship between happiness and GDP per capita?". We all know the saying "Money can't buy happiness", but I wanted to test see if it could, by regressing GDP per capita over happiness score. I decided to use the happiness score from the United Nations at `https://www.kaggle.com/unsdsn/world-happiness` as my observation of a country's happiness. The GDP data in that dataset was the contribution of GDP to happiness, so I decided to get my GDP per capita data from the World Bank at `https://data.worldbank.org/indicator/NY.GDP.PCAP.CD`. In particular, because happiness scores were only computed from 2015-2019 I decided to only focus on 2019 because it provides the most recent data available. I also decided to remove any countries without GDP values from the World Bank dataset.

If we could define a clear relationship between GDP per capita and happiness, then it might be useful to aid organizations and other non-profits. They could specifically focus spending their funding on developing countries' GDPs to improve quality of life, rather than other aid efforts. I decided to apply Bayesian regression to this problem to try to model the relationship between GDP per capita and happiness. In particular, I want to analyze how an increase in GDP per capita affects happiness.

2 Formulation

Initially looking at the data on the left, I was confused how I could model Happiness as a function of GDP per capita. However, by taking the log of GDP per capita, we can see in in the right graph that Log GDP seems to have a linear relationship with Happiness score. So my model of the happiness score as a function of log GDP is just

$$z(c) = b_0 + m_0 * l(c)$$

for country c where b_0, m_0 are the intercept and slope respectively which are unknown parameters and $l(c)$ is the log GDP per capita of that country.

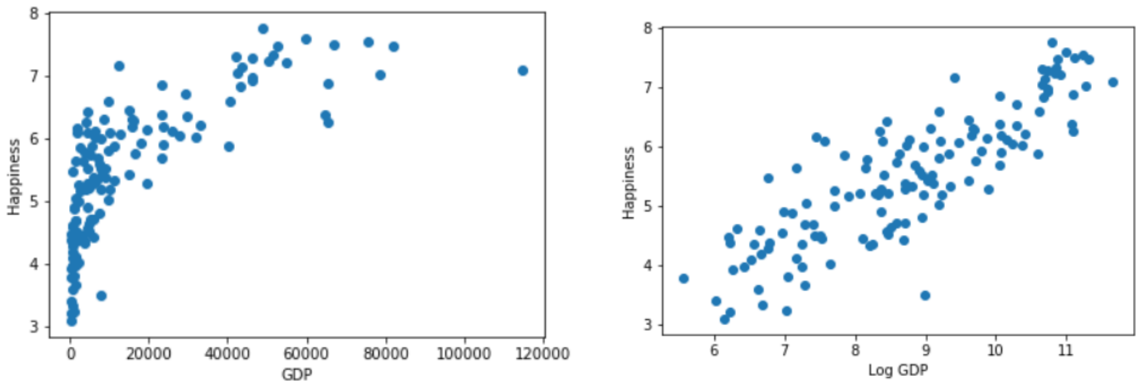

Figure 1: Scatterplot of GDP per Capita and Log GDP per Capita against Happiness

Let's label our countries as $\{c_1, c_2, ..., c_N\}$. Let $\mathbf{z}$ be my set of noisy observations of happiness, which I assume is related to my model by

$$\mathbf{z} = \begin{bmatrix} z(c_1) \\ \vdots \\ z(c_N) \end{bmatrix} + \epsilon = A \begin{bmatrix} b_0 \\ m_0 \end{bmatrix} + \epsilon$$

where $A = \begin{bmatrix} 1 & l(c_1) \\ 1 & l(c_2) \\ ... & ... \\ 1 & l(c_N) \end{bmatrix}$ and $\epsilon \sim N(0, \sigma_\epsilon^2 I_N)$ because there may be some observation error when recording happiness. I assume this observation error should be normally distributed with a variance of $\sigma_\epsilon^2 = 1$ for each observation. This normal assumption with mean zero is pretty common in regression about the residuals and I think is a fair assumption to make here. Let $x = \begin{bmatrix} b_0 \\ m_0 \end{bmatrix}$ be the parameters of the model. Because $\mathbf{z} - Ax = \epsilon$, we know that $f(z|x) = f(\epsilon)$ and my likelihood $f(z|x) = f(\epsilon) \sim N(0, \sigma_\epsilon^2 I_N)$.

In order to simplify calculations, I would like for my prior to be conjugate to my likelihood, which is why I chose for the joint distribution of our prior to be a multivariate Gaussian. I will assume for my prior that b_0, m_0 are independent, and have a multivariate Gaussian distribution with mean $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and covariance $\begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$ where we have a large variance around our intercept and smaller variance around our slope to reflect less confidence in the value of the intercept.

And I am interested in finding the posterior distribution of x, that is, $f(x|z)$. And I want to solve for $E_{x|z}[x]$, which the slope component tells me that a 1 unit increase in log GDP per capita increases happiness by that amount.

3 Solution Strategy

Because z is a linear function of x, we can use our solution from the HW Gaussians and Monte Carlo problem 1.2. Then our marginal covariances are:

$$\Sigma_{zz} = A\Sigma_{xx}A^T + \sigma_\epsilon^2 I_N$$
$$\Sigma_{zx} = \Sigma_{xz}^T = A\Sigma_{xx}$$

And we know that from the Gaussian Identities that for two Gaussian random variables x, z, the conditional distribution (posterior)

$$f(x|z = \overline{z}) \sim N(\mu_x + K_{xz}(\overline{z} - z), \Sigma_{xx} - K_{xz}\Sigma_{xz}^T)$$

where K_{xz} is the Kalman gain $= \Sigma_{xz}\Sigma_{zz}^{-1}$.

Plugging in the equations from above, we get that

$$f(x|z = \overline{z}) \sim N(\mu_x + (\Sigma_{xx}^T A^T (A\Sigma_{xx}A^T + \sigma_\epsilon^2 I_N)^{-1})(\overline{z} - z), \Sigma_{xx} - \Sigma_{xx}^T A^T (A\Sigma_{xx}A^T + \sigma_\epsilon^2 I_N)^{-1} A\Sigma_{xx})$$

However, I think the other form is nicer to compute when we calculate the Kalman gain separately. Note that we have a closed form for $E_{x|z}[x]$ above!

I created a Jupyter Notebook ”Reg.ipynb” that covers all the programming that I ended up doing in python. After extracting the data from the two datasets, I used a join function to merge rows based on country names using pandas, although I think that a couple of the countries ended up having different names which ended up in them being deleted. I also used the numpy and matplotlib packages to help perform calculations with the data and graph it. After obtaining the posterior distribution, I got the mean values of our posterior and plotted it along with the observation values to show how well the points fit the model.

4 Results

Our posterior is a multivariate Gaussian with $\mu = \begin{bmatrix} -.097 \\ .64 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 0.253 & -.028 \\ -.028 & .003 \end{bmatrix}$. We can see below on the left a graph of what the mean of our posterior looks like, along with a spread of two standard deviations. Also included is a scatter plot of all of the observations values, which we can see are generally within two standard deviations of the mean (except for one outlier)! So the mean of our posterior gives the model:

$$z(c) = -.097 + .64l(c)$$

On the right, after taking the exponent of GDP, we can see what GDP per capita vs happiness looks like, which is what we had originally wanted. In order to graph this figure, I generated 100 evenly spaced Log GDP values between 5 and 12, which were the respective bounds of our data set, and applied the model to these points. Our exponential graph seems to capture the relationship pretty well, although we still have the single outlier.

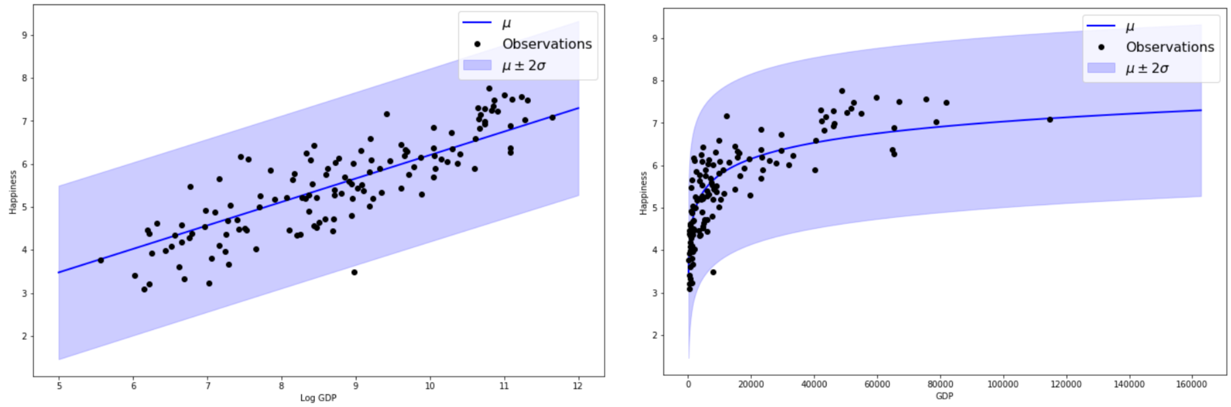

Figure 2: Linear Regression and Exponential Regression

5 Discussion

So it seems at first like we have a model that captures how GDP affects happiness score pretty well. Our results tell us that a 1 unit increase in log GDP per capita increases happiness by .64 and we have a model distribution whose mean captures almost all points within two standard deviations! So we are done right?

Unfortunately, looking at the following figure we see an almost perfectly linear relationship between Log GDP per capita and this column in the happiness data called ”GDP per capita”. Taking a closer look at the dataset, there is a comment that *”The following columns: GDP per Capita, Family, Life Expectancy, Freedom, Generosity, Trust Government Corruption describe the extent to which these factors contribute in evaluating the happiness in each country...If you add all these factors up, you get the happiness score so it might be un-reliable to model them to predict Happiness Scores.”*

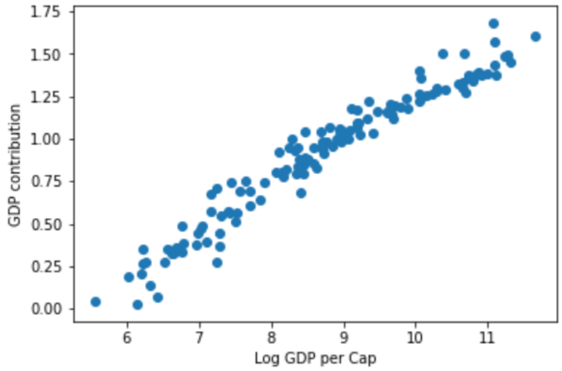

Figure 3: Log GDP per Capita vs Contribution to Happiness

So basically we had been trying to backsolve the formula for how GDP per capita is used to compute happiness score. Thus, our results don’t really give us a true sense of how

GDP can affect happiness. Happiness isn't really something that can be quantified, but I think that a better method of observing it would have been to take a random survey in each country to measure what they perceive their happiness to be. I'm also not sure if GDP per capita is the best measurement of an individual's wealth, perhaps a better one would be average net worth per person or their purchasing power.

My model for happiness was also extremely simple, it assumed that Happiness was a linear function of Log GDP. However, like the UN dataset points out, there are a lot of factors contributing to happiness including Life Expectancy and Freedom. Perhaps a future experiment might want to incorporate more of these factors to their model.

While taking a look at the initial graph made it seem like there would be a linear relationship between Log GDP per Capita and Happiness, perhaps a higher order polynomial might have better captured the relationship. It also might have been useful to utilize all 5 years of Happiness data available, but because of the varying formats of the different data sets, I decided to just stick to one.

Additionally, I chose my prior to be Gaussian in order to simplify calculations. I could also have chosen a different distribution and run MCMC instead. I also regressed by country values, maybe it would have been better to compute regress over individual people instead. I don't really see the benefit of stratifying by country, but it was the only data available.

Overall, it was an interesting problem, and it seemed like Bayesian regression was the right approach. But because of the way happiness was calculated, the approach was flawed.

6 References

- https://www.kaggle.com/unsdsn/world-happiness
- https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
- Gaussians and Montecarlo problem set
- Math 76 Lecture notes
- Worked with Ruiheng Ma
- Received help from Jiahui Zhang