Predictors of Happiness Across Developed & Developing Countries

Applied Regression Analysis
2/28/2017

Beth Dawoodi | Ana Sofia Guzman | Kristie Karnadi | Katherine Li | Lucia Mijares

Agenda

- **Project Overview**
- Finding the Best Model
- Key Takeaways & Recommendations

Project Overview

Part I

- Understand the relationship between the Happiness Index rating for a country and a series of independent variables

Part II

- Understand the relationship between Happiness Index rating for a country and independent variables when we **control for level of development** (developed vs. developing) based on GDP

Setting up our analysis

Objective

- Determine the factors that most significantly affect individual happiness and understand how they are different between developing and developed countries

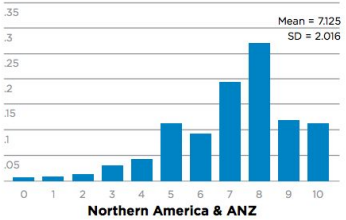
Process

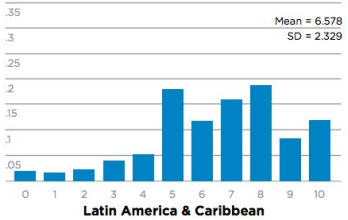
1. Correlation Analysis
2. Determined a best fit regression for all countries
3. Separated Developing vs Developed countries based on GDP Distribution
4. Ran regression models to determine significance

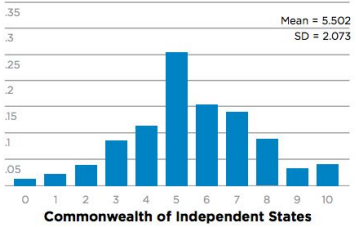
Dependent Variable: The Happiness Index

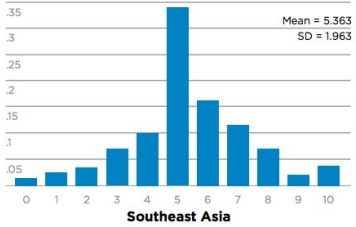
- The World Happiness Report (www.worldhappiness.report)
- Methodology: Survey of roughly 3,000 random respondents in each of more than 150 countries from 2012-2015 asking them to evaluate their current lives on a ladder where 0 represents the worst possible life and 10, the best possible.
- The overall global happiness distribution is very normally distributed about the median answer of 5, with the population-weighted mean being 5.4.

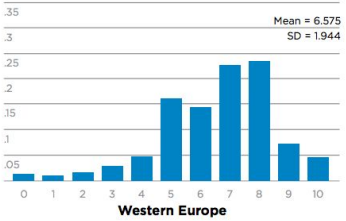
Population-Weighted Distributions of Happiness Based on Region

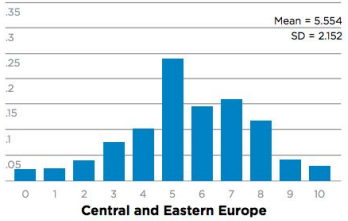


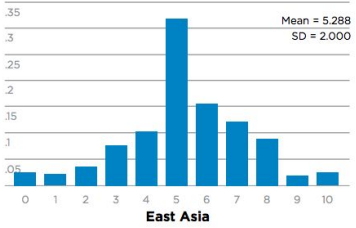


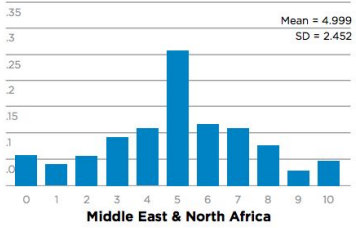


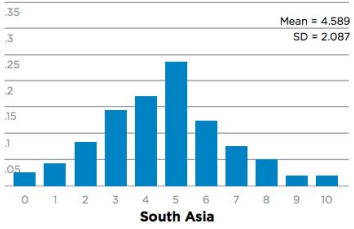


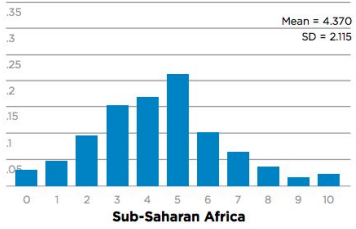












Looking at the Independent Variables

Indices	Macro Data	Internet Trends	Intl. Food Chains
• Cost of Living • Groceries • Restaurant Price • Property Price • Healthcare • Traffic • Pollution	• GDP per Capita • Population • Literacy Rate • Murder Rate • Obesity Rate	• Int. Penetration • Facebook Users	• Total Starbucks locations • Total McDonald’s locations

- Project Overview
- **Finding the Best Model**
- Key Takeaways & Recommendations

Descriptive Statistics

	Happiness Index	GDP	Cost of Living Index	Groceries Index	Restaurant Price Index	Property Price Index	HealthCare Index	Pollution Index	Traffic Index
Mean	6.08	27,055	55.87	49.41	49.81	14.62	66.17	50.69	143.32
Median	6.04	18,415	51.33	44.06	44.12	10.60	68.48	51.81	140.48
St. Dev	0.95	23,922	20.83	20.45	25.12	**24.76**	10.23	21.36	41.85

	# of McDonalds	Literacy Rates	Population (M)	Internet Penetration Rate	FB Users (M)	Murder Rate per million	# of Starbucks	Obesity
Mean	566	96%	88	72%	22	46	383	22.50
Median	142	98%	21	75%	8	17	25	22.00
St. Dev	**1,850**	8%	**234**	18%	**36**	**84**	**1,760**	9.95

Step 1: Initial Correlation Table

- Compiled a series of variables against which we wanted to regress the Happiness Index.
- Found many issues of multi-colinearity and used this guidance to get rid of a couple of redundant variables.

	Happiness Index	GDP	Cost of L	Groceries	Restaurant	Property	Health C	Pollution	Traffic	McD's	Literacy	Population	Internet	FB Users	Murder	SBUX	Obesity
Happiness Index	1																
GDP	0.736787481	1															
Cost of Living Index	0.713406875	0.876614	1														
Groceries Index	0.650533947	0.807253	0.956363	1													
Restaurant Price Index	0.725352898	0.882385	0.927554	0.814609	1												
Property Price Index (Price to Income F	-0.05892914	-0.151428	-0.165995	-0.099644	-0.148171	1											
HealthCare Index	0.459123788	0.414428	0.491112	0.512982	0.395811	-0.394161	1										
Pollution Index	-0.59892938	-0.628276	-0.606294	-0.515972	-0.62886	0.20451	-0.422546	1									
Traffic Index	-0.30376716	-0.393006	-0.378119	-0.27408	-0.424614	0.231891	-0.155166	0.687868	1								
# of McDonalds	0.177058379	0.172403	0.181359	0.246625	0.120427	-0.066659	0.100525	-0.130628	0.07805	1							
Literacy Rates	0.211014917	0.192434	0.277142	0.2004	0.263814	0.019421	0.141255	-0.317422	-0.366552	0.028089	1						
Population	-0.24646238	-0.223154	-0.221202	-0.124542	-0.255012	0.016685	-0.031964	0.342454	0.330112	0.242063	-0.308447	1					
Internet Penetration Rate	0.729946617	0.748862	0.748297	0.639743	0.747529	-0.168252	0.407952	-0.684472	-0.515543	0.151503	0.433819	-0.378929	1				
FB Users	-0.01075763	-0.13147	-0.140472	-0.049292	-0.186903	-0.030081	0.015388	0.191348	0.420341	0.658982	-0.271039	0.475594	-0.211358	1			
Murder Rate per million	-0.08668479	-0.306068	-0.333805	-0.290226	-0.273735	0.624678	-0.336935	0.27974	0.425652	-0.028796	-0.069508	0.013848	-0.301353	0.17542	1		
# of Starbucks	0.143330124	0.13123	0.124469	0.190129	0.083171	-0.059894	0.05022	-0.081932	0.087792	0.981091	0.006549	0.266454	0.102652	0.638361	-0.023986	1	
Obesity	0.157110868	0.152856	0.024932	-0.107177	0.200489	-0.035386	-0.208722	0.065017	0.13716	0.087456	0.058165	-0.31262	0.247289	-0.067154	0.17027	0.112575	1

Regression 1: Developed + Developing Countries

- Best regression that fit the data included:
  - GDP
  - Healthcare Index
  - Internet Penetration Rate
  - Murder Rate
- Our predictive power for this regression is 59%

Regression Analysis: Happiness In versus GDP, HealthCare I, Internet Pen, Murder Rate

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	36.951	9.2378	29.79	0.000
GDP	1	4.187	4.1871	13.50	0.001
HealthCare Index	1	1.622	1.6217	5.23	0.026
Internet Penetration Rate	1	3.712	3.7119	11.97	0.001
Murder Rate per million	1	2.464	2.4644	7.95	0.007
Error	57	17.676	0.3101		
Total	61	54.627			

Model Summary

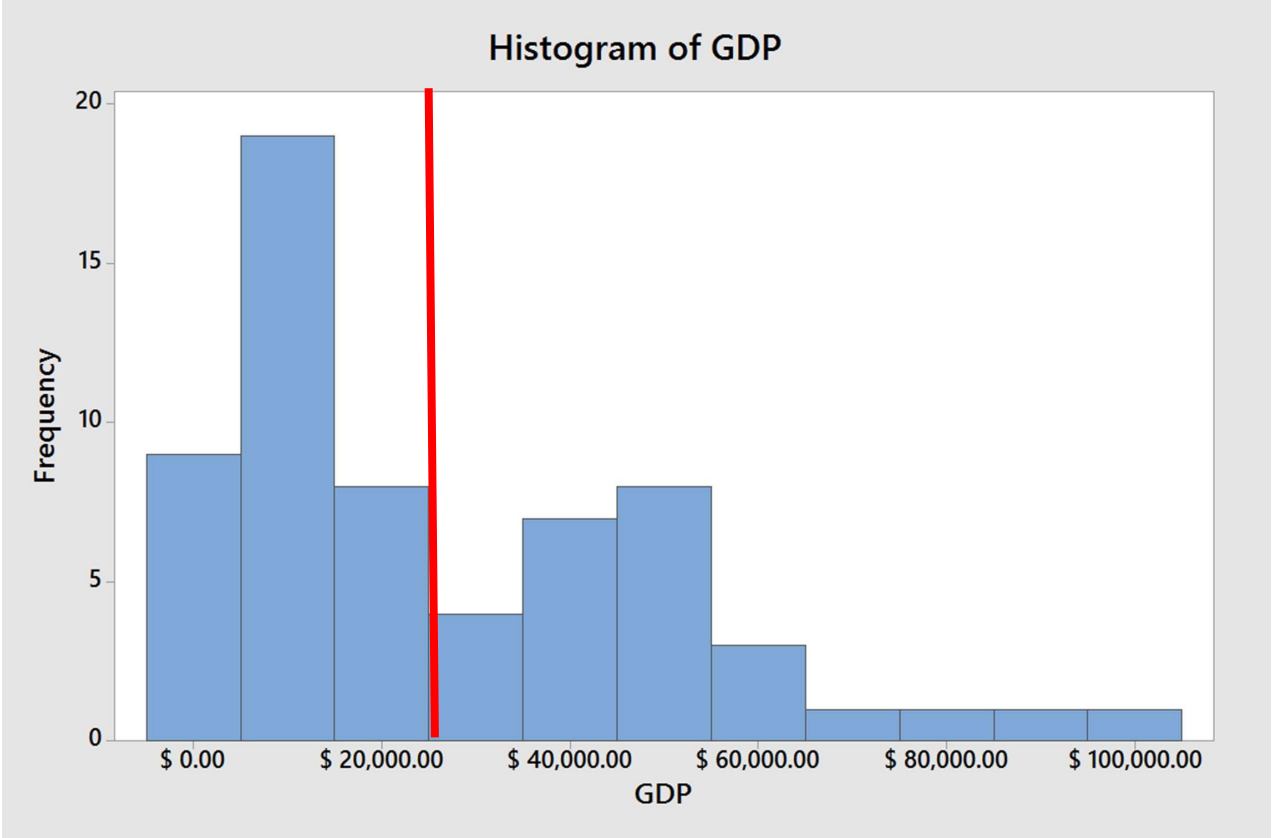
S	R-sq	R-sq(adj)	R-sq(pred)
0.556864	67.64%	65.37%	59.35%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.809	0.585	4.80	0.000	
GDP	0.000017	0.000005	3.67	0.001	2.37
HealthCare Index	0.01823	0.00797	2.29	0.026	1.31
Internet Penetration Rate	2.057	0.595	3.46	0.001	2.35
Murder Rate per million	0.002597	0.000921	2.82	0.007	1.18

However...

- We realized countries in our data set were widely different in terms of GDP and overall economic power.
- Divided our data set into “Low GDP” and “High GDP” countries



Regression 2: Developed High GDP Countries

- Happiness in High GDP countries is best determined by:
  - GDP
  - HealthCare Index
  - Pollution
  - Obesity
- Predictive power of our regression dropped to 32%
- *Interesting: Obesity as a predictor of happiness? = Correlation does not equal causation*

Regression Analysis: Happiness Index versus GDP, HealthCare Index, Pollution Index, Obesity

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	9.764	2.4410	11.02	0.000
GDP	1	2.108	2.1080	9.52	0.005
HealthCare Index	1	1.411	1.4110	6.37	0.019
Pollution Index	1	1.777	1.7766	8.02	0.009
Obesity	1	1.268	1.2676	5.72	0.025
Error	24	5.317	0.2215		
Total	28	15.081			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.470662	64.75%	58.87%	32.73%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.66	1.06	3.46	0.002	
GDP	0.000015	0.000005	3.08	0.005	1.25
HealthCare Index	0.0325	0.0129	2.52	0.019	1.43
Pollution Index	-0.01752	0.00619	-2.83	0.009	1.35
Obesity	0.0280	0.0117	2.39	0.025	1.61

Regression Equation

Happiness Index = 3.66 + 0.000015 GDP + 0.0325 HealthCare Index - 0.01752 Pollution Index
+ 0.0280 Obesity

Regression 3: Developing Low GDP Countries

- Happiness in Low GDP countries is best determined by:
  - Internet Penetration Rate
  - Murder Rate
- Predictive Power of our regression is 24.66%

Regression Analysis: Happiness Index versus Internet Penetration Rat, Murder Rate per million

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	7.987	3.9936	10.32	0.000
Internet Penetration Rate	1	6.640	6.6404	17.15	0.000
Murder Rate per million	1	1.902	1.9023	4.91	0.034
Error	30	11.615	0.3872		
Total	32	19.602			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.622219	40.75%	36.80%	24.66%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.670	0.438	8.37	0.000	
Internet Penetration Rate	2.787	0.673	4.14	0.000	1.01
Murder Rate per million	0.00229	0.00103	2.22	0.034	1.01

Regression Equation

Happiness Index = 3.670 + 2.787 Internet Penetration Rate + 0.00229 Murder Rate per million

Analyzing Regression Results

- There are fewer independent variables than expected in best fit models. Possible reasons for this:
  - Countries that are developed are more homogenous and therefore factors that contribute to happiness are similar across countries.
  - Less developed economies have many different variables that contribute or detract from happiness so it is difficult to regress on these.
- There were other subset variables that were more significant, but their predictive values were 0.

- Project Overview
- Finding the Best Model
- **Key Takeaways & Recommendations**

Key Takeaways

- Independent variables used were not highly predictive of Happiness:
  - For both developed and developing countries, best model accounts for 59.4% of variability
  - Separating developed and developing, best model accounts for 32.7% and 24.7%, respectively
- Many independent variables in the database are multicollinear, which undermined model predictive value
- Attempt to control for economic development still did not capture variability in happiness within the country groups
- Happiness in developed countries likely more homogenous to be captured in one model than happiness in developing countries

Suggestions for Future Analysis

- A more complete data set that includes as many countries as possible
- A more descriptive way to divide countries (i.e. not just by GDP)
- Find data on levels of depression, time spent outdoors, religious observance, incidence of meditation, etc.
- Run residual plots to visualize errors in data

Questions?

Appendix

Sources of Data

Indices	Macro Data	Internet Trends	Intl. Food Chains
Cost of Living	GDP per Capita	Internet & FB Users	McDonald's - US/Canada
Groceries Index	Population		McDonald's - Latam
Restaurant Index	Literacy Rate		McDonald's - Europe
Property Price Index	Murder Rate		McDonald's - Asia
Healthcare Index	Obesity Rate		Startbucks Locations
Pollution Index			Starbucks by Country
Traffic Index			
World Happiness Report			

Description of Independent Variables (I)

Indices

- **Cost of Living Index:** A relative indicator of consumer goods price, including groceries, restaurants, transportation and utilities. Cost of Living Index doesn't include accommodation expenses such as rent or mortgage (Higher = Higher Cost of Living)
- **Groceries Index:** is an estimation of grocery prices in the city compared to New York City. (Higher = More Expensive)
- **Restaurant Price Index:** Comparison of price of meals at restaurants compared to NYC (Higher = More Expensive)
- **Property Price Index:** Price to Income Ratio; compares median apartment prices to median family dispoable income (Lower = More Affordable)
- **Healthcare Index:** is an estimation of the overall quality of the health care system, health care professionals, equipment, staff, doctors, cost, etc. (Higher = Higher Quality)
- **Pollution Index:** is an estimation of the overall pollution in the city. The biggest weight is given to air pollution, than to water pollution/accessibility, two main pollution factors. Small weight is given to other pollution types. (Higher = Higher Pollution)
- **Traffic Index:** is a composite index of time consumed in traffic due to job commute, estimation of time consumption dissatisfaction, CO_2 consumption estimation in traffic and overall inefficiencies in the (Higher = More Traffic)

Source: Numbeo

Description of Independent Variables (II)

Macro Data

- **GDP per capita:** a measure of the total output of a country that takes gross domestic product (**GDP**) and divides it by the number of people in the country
- **Population:** Number of people living in a country
- **Literacy Rate:** Percentage of total population that can read and write
- **Murder Rate:** Murders per one million inhabitants
- **Obesity Rate:** WTO female obesity data by country for 2014

Internet Trends

- **Internet Penetration Rate:** Percentage of the population with access to the Internet
- **Total Facebook Users:** Number of Facebook accounts registered in a country

Description of Independent Variables (III)

International Food Chains

- **Number of McDonald's:** Total number of McDonald's locations by country
- **Number of Starbucks:** Total number of Starbucks locations per coutnry