Don't worry, be happy:
An examination of global happiness scores

Shawn Ban

December 9, 2017

1 Introduction, aim and motivation

In this project we examine world happiness scores from the Gallup World Poll. In the poll, known as the Cantril ladder, 1000 respondents from each country are asked to rate their lives on a scale of 0-10, with 10 being the best possible life imaginable and 0 being the worst possible life imaginable.

The United Nations Sustainable Development Solutions Network publishes a World Happiness Report [1] annually based on these scores. In the latest report, the authors note that three-quarters of the happiness scores can be explained by six factors: income, life expectancy, generosity, freedom, trust and social support. However, we do not have access to the raw data for these factors.

This project therefore seeks to explore global happiness scores, and replicate or improve upon the six-factor model proposed by the authors.

1.1 Analytical questions

Some questions we hope to answer are:

1. How does happiness vary by countries and by regions?
2. How has happiness changed from 2015 to 2017? Are there patterns to this change?
3. What factors explain differences in happiness among countries and by how much? Do other factors not mentioned by the report's authors such as inequality have a significant impact on happiness?
4. Are happiness scores more sensitive to absolute levels in variables or to changes in variables?

1.2 Plan of analysis

Data on global happiness scores was obtained from Kaggle [2] while data on economic variables was obtained from the World Bank [3]. For other indices, we obtain corruption ratings from Transparency International [4] and press freedom ratings from Reporters Without Borders [5].

Our plan is to first explore the happiness scores and look for patterns of interest. We then explore each independent variable in turn, looking for correlations with happiness. We also plan to check if happiness correlates to either absolute or percentage changes in the independent variables. Recent studies from the field of hedonic psychology such as Kahneman & Deaton, 2010 [6] and Dunn, et. al, 2011 [7] indicate that people adapt quickly to baseline states and that changes in states can have a larger impact on well-being than the absolute levels of states. Exploring changes in independent variables, rather than absolute levels, will help us determine if these effects show up on a macro-economic level.

2 Analytical process and design choices

In general, each variable is stored in an individual dataset, so the data wrangling involves merging multiple datasets on country codes. We first perform an in-depth analysis on happiness scores, before plotting histograms of the explanatory variables, making transformations to approximate normality as needed. We then check for correlation via a heatmap and series of scatterplots.

We then build a multivariate linear regression model using explanatory variables that were found to have high correlations, first checking for multi-collinearity. We then drop one variable at a time, running the regression analysis each time, until all remaining variables have a p-value below 0.05 significance level. Finally, we examine the residuals for normality.

We primarily perform the data wrangling in Python using the *numpy*, *pandas*, *matplotlib*, *seaborn* and *statsmodels* libraries. We perform the linear regression in both Python and R for verification. We choose to drop observations with missing values via an inner join, and this results in a loss of observations from 155 in our initial dataset to 138 for our regression model.

2.1 Exploratory data analysis

Merging multiple datasets provides us with the following variables:

- Happiness: we have three years of happiness scores from 2015 to 2017, given on a scale of 0 to 10.
- GDP: we use gross domestic product (GDP) per capita on a purchasing power parity (PPP) basis, converted to current international dollars as a proxy for levels on income. An international dollar has the same purchasing power that the US dollar has in the United States.

- Healthcare: we use life expectancy at birth and infant mortality rate per 1000 live births as proxies for healthcare.
- Inequality: we use the Gini coefficient as a proxy for inequality, with a higher score indicating a more unequal society. We note that the data quality is quite inconsistent and we take the most recent year for which we have data, making the assumption that levels of equality stay relatively consistent over time.
- Crime: we use intentional homicides per one million people as a proxy for violent crime. Again the data quality is somewhat inconsistent, and we take the most recent year for which we have data, making the assumption that violent crime rates stay relatively consistent over time.
- Press freedom: the ratings are produced by Reporters Without Borders [5] on a scale of 0 to 100, with a lower score indicating a freer press.
- Corruption: the ratings are produced by Transparency International [4] on a scale of 0 to 100, with a higher score indicating less corruption.

We also calculate the percentage change in the most recent year for GDP, life expectancy, infant mortality, press freedom and corruption.

2.1.1 Happiness scores

We find that happiness scores can be approximated by a normal distribution (Figure 1) and there are marked differences in happiness between regions (Figure 2).

Figure 1: Histogram of happiness scores

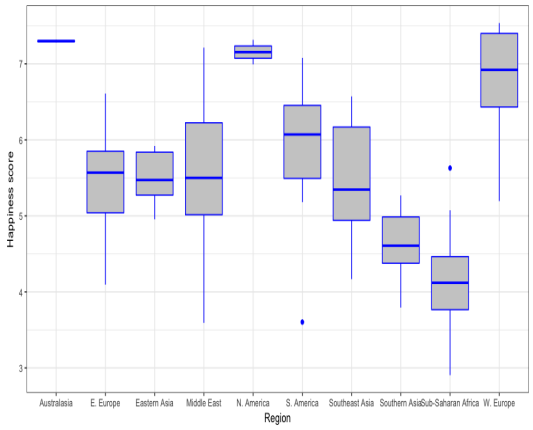

Figure 2: Happiness score 2017 by region

By examining changes over time, we see that there is a marked decline in happiness in Latin America and North America from 2015 to 2017 (Figure 3). We also find that changes in happiness are normally distributed (Figure 4), although there are a few outliers with

large declines such as Venezuela (-1.56), Liberia (-1.04) and Haiti (-0.92). A good model should aim to explain variations in happiness across regions and time.

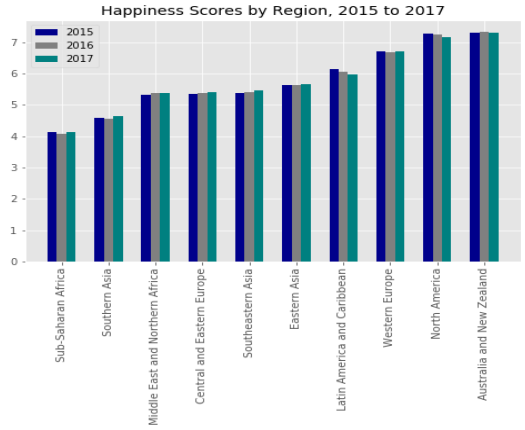

Figure 3: Change in happiness by region

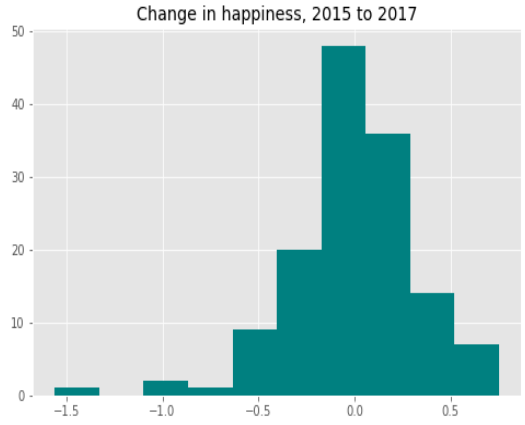

Figure 4: Change in happiness, 2015-2017

2.1.2 Histogram analysis

After merging the various datasets, we plot histograms of happiness scores in 2017 and seven explanatory variables (Figure 5).

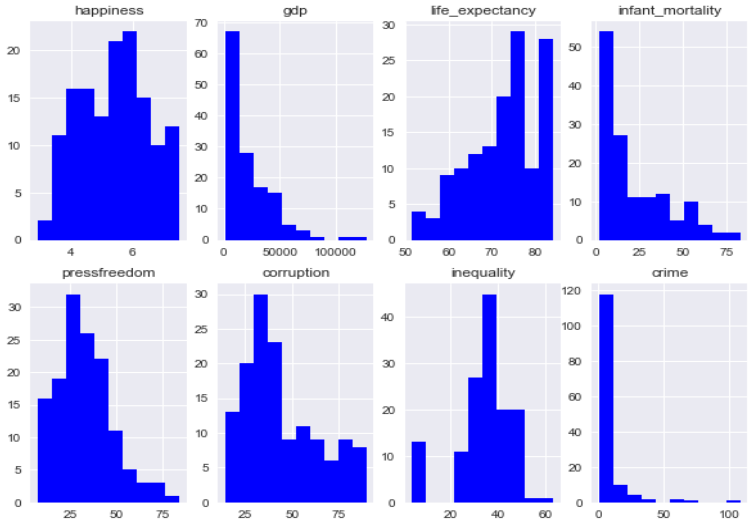

Figure 5: Histograms of happiness and explanatory variables

To reduce skew and approximate normality, we use a log transformation on GDP, crime rates, and infant mortality rates and a square root transformation on corruption and press freedom. This produces the following series of histograms (Figure 6).

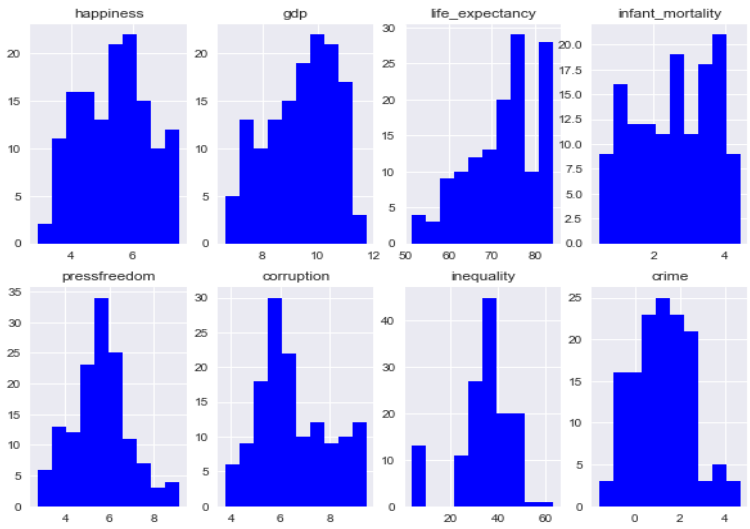

Figure 6: Histograms after transformation

2.1.3 Correlation analysis

We first plot a correlation heatmap of happiness against the explanatory variables (Figure 7). Most of the regressors seem promising, with the exception of inequality.

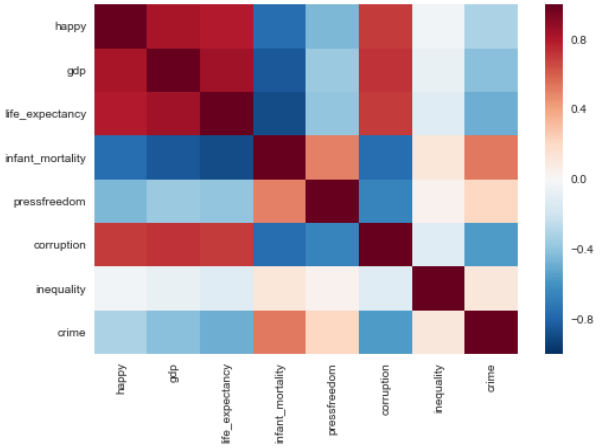

Figure 7: Correlation between happiness and regressors

The correlation heatmap of happiness against the change in variables (Figure 8) is less promising. There appears to be some correlation with changes in life expectancy and press freedom, but these are in the opposite direction as the correlation with the absolute level of the variables and runs counter to our intuition, namely that a decrease in life expectancy and press freedom improves happiness. We believe this is probably a spurious relationship that should disappear once other control variables are introduced.

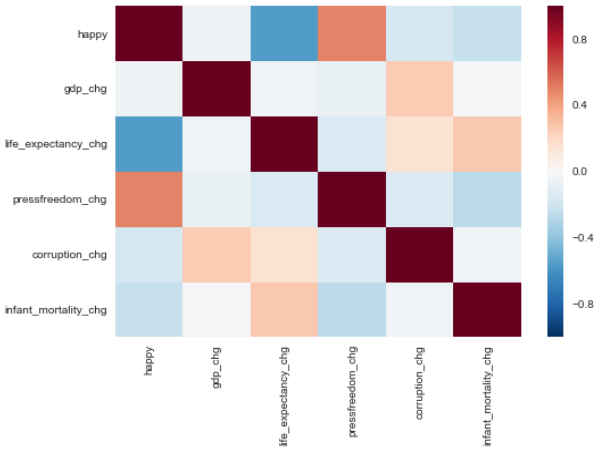

Figure 8: Correlation between happiness and % change in variables

We next check the scatterplots of happiness against the explanatory variables, reporting correlation for each pair (Figure 9).

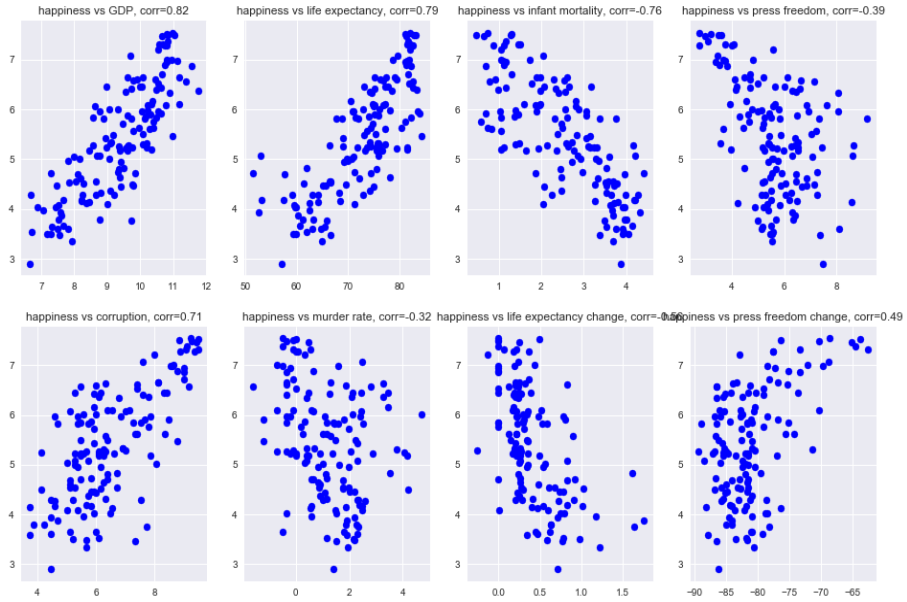

Figure 9: Scatterplots of happiness vs regressors

2.2 Regression analysis

Our aim is to build a multivariate regression model using variables that seem to have a significant effect, based on the scatterplots and correlation. Here we use a p-value of 0.05 as a cut-off for significance. For multivariate linear regression, we need the following assumptions to hold:

(i) Linear relationship between independent variable and each explanatory variable.

(ii) Explanatory variables are not multi-collinear.

(iii) Residuals are not auto-correlated.

(iv) Residuals are homoskedastic.

(v) Residuals are normally distributed.

We therefore first test for collinearity by checking the eigenvalues of the correlation matrix before running our regression analysis. We then build the regression model in a stepwise manner by dropping one explanatory variable at a time until all the remaining variables have a p-value below 0.05. Finally, we analyse the residuals to ensure the above assumptions are met.

Based on the scatterplot analysis, we choose to retain six explanatory variables for our model: GDP, life expectancy, infant mortality, crime, press freedom and corruption. Checking for multi-collinearity, we create a correlation matrix of the regressors and solve for its eigenvalues. This produces the following:

$$w = \begin{bmatrix} 4.06 \\ 0.83 \\ 0.68 \\ 0.22 \\ 0.12 \\ 0.10 \end{bmatrix}, \; C = \frac{4.06}{0.1} = 40.6$$

where w is a column vector of the eigenvalues and C is defined as the condition number, the ratio of the largest to the smallest eigenvalue. As a rule of thumb, a condition number above 100 indicates large multi-collinearity. With a condition number of 40.6, we conclude the collinearity of the explanatory variables is within acceptable limits and proceed to build the linear model.

Our first run of the linear regression yields the following output:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.15692    1.38473  -2.280  0.02424 *
gdp              0.42589    0.09103   4.679  7.1e-06 ***
life_expectancy  0.05032    0.01523   3.304  0.00123 **
infant_mort      0.04337    0.12115   0.358  0.72096
corruption       0.14942    0.07335   2.037  0.04367 *
pressfreedom    -0.05473    0.05754  -0.951  0.34333
crime           -0.13144    0.05361   2.452  0.01554 *
---
Residual standard error: 0.5871 on 131 degrees of freedom
```

```
Multiple R-squared:  0.744,Adjusted R-squared:  0.7323
F-statistic: 63.45 on 6 and 131 DF,  p-value: < 2.2e-16
```

We then proceed in a stepwise manner, dropping one explanatory variable at a time until all remaining variables have a p-value below 0.05. This yields our final four-factor model:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.23004    0.54947  -5.878 3.16e-08 ***
gdp              0.39736    0.08335   4.767 4.83e-06 ***
life_expectancy  0.04836    0.01237   3.910 0.000146 ***
corruption       0.19041    0.05541   3.436 0.000787 ***
crime           -0.14929    0.04981   2.997 0.003253 **
---
Residual standard error: 0.5847 on 133 degrees of freedom
Multiple R-squared:  0.7422,Adjusted R-squared:  0.7344
F-statistic: 95.72 on 4 and 133 DF,  p-value: < 2.2e-16
```

Finally, we examine the residuals via diagnostic plots. A quantile-quantile (Q-Q) plot (Figure 10) confirms that the residuals are fairly normally distributed overall. Since the Q-Q plot results in a fairly straight line, we conclude the residuals can be approximated by a normal distribution, and the model assumptions hold. While there does not seem to be overall bias in the residuals (Figure 11), colouring by region reveals a pink cluster of residuals representing Western Europe in the upper-right corner (Figure 12). While the model has good explanatory power overall, it seems to be under-estimating happiness scores in Western Europe. This points to further work that could be done.

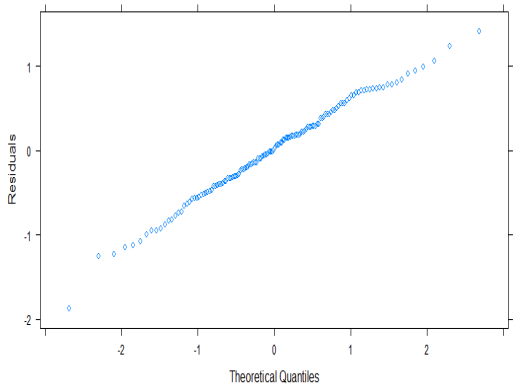

Figure 10: Q-Q plot of residuals

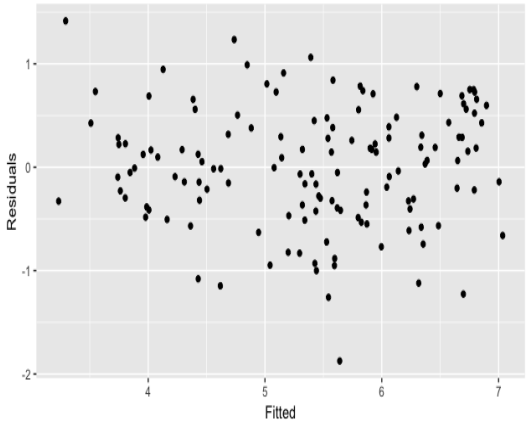

Figure 11: Residuals vs fitted values

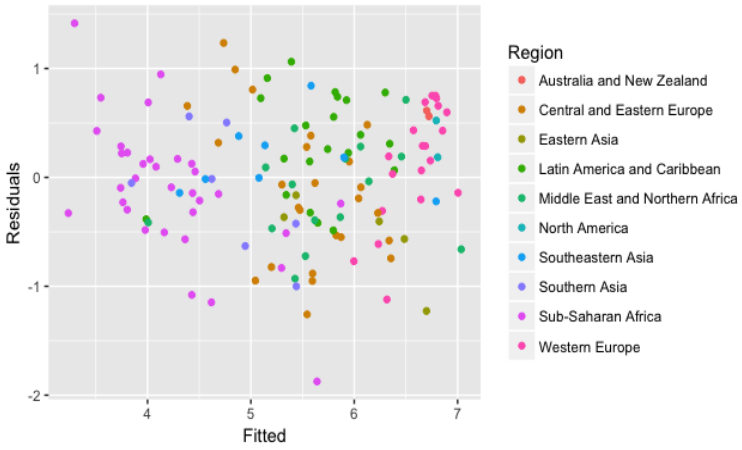

Figure 12: Residuals by region

3 Findings, reflections and future work

The regression analysis produces a significant model with an adjusted R-squared of 73.4% using four explanatory variables: GDP, life expectancy, corruption and crime. Just as importantly, we note the variables that failed to have significance in the model. In particular, happiness scores have a very low correlation with income inequality in countries. Relating back to our analytical questions, we note that happiness scores also seem to be uncorrelated with changes in macro-economic variables. For the purposes of this analysis, at least, happiness scores are more sensitive to absolute levels than to relative changes.

The findings offer some basic policy prescriptions for governments to improve the well-being of their citizenry. It suggests that governments should focus on raising levels of income, and directing expenditure towards improving healthcare and reducing crime and corruption.

For future work, we note that there is some bias in the residuals for Western Europe, so the model seems to be consistently under-estimating happiness scores there. We hypothesize that additional variables related to equality or welfare could have explanatory power for Western Europe, or perhaps the coefficients vary for this region. Further investigation could produce a more refined model for Western Europe.

There also remains the 26.6% of variation in happiness scores that is still unexplained by the model. We hypothesize that the remaining 26.6% of variation could be influenced by highly localized environmental factors, such as traffic conditions, air quality, pollution levels, levels of social care or levels of stress. Such factors are unlikely to show up in macro-economic variables, and a more granular analysis of happiness scores may be necessary to give our model more explanatory power.

We note that data in the field is relatively young, as the push for governments to measure and improve the well-being of their citizens was only initiated by a 2011 United Nations Resolution [8]. The 2017 edition of the World Happiness Report was thus only the

fifth edition, and we expect that data quality should continue improving in future editions. Producing time series of happiness scores would also allow us to examine how well-being varies over time, rather than a static analysis over geographies that we produce here.

References

[1] Helliwell, J., Layard R. & Sachs, J. (2017) *World Happiness Report 2017*, New York: Sustainable Development Solutions Network.

[2] Kaggle (2017) *World Happiness Report.*
Available at https://www.kaggle.com/unsdsn/world-happiness/data

[3] World Bank Group (2017) *World Bank Open Data.*
Available at https://data.worldbank.org

[4] Transparency International (2017) *Corruption Perception Index 2016.*
Available at https://www.transparency.org/research/cpi/overview

[5] Reporters Without Borders (2017) *2017 World Press Freedom Index.*
Available at https://rsf.org/en/ranking

[6] Kahneman D. & Deaton, A. (2010) 'High income improves evaluation of life but not emotional well-being', *Proc Natl Acad Sci USA*, vol. 107, no. 38, pp. 16489-16493.

[7] Dunn, E., Gilbert, D. & Wilson T. (2011) 'If money doesn't make you happy, then you probably aren't spending it right', *Journal of Consumer Psychology*, vol. 21, no. 2, pp. 115-125.

[8] United Nations General Assembly resolution 65/309 (19 July 2011), *'Happiness: towards a holistic approach to development'*, A/RES/65/309.