

Global Happiness Analysis

Arvind Kandala, Ammy Lin, Lingyue Hao, Shelly Cao

Load data

```
Rows: 781
Columns: 10
$ country      <chr> "Switzerland", "Iceland", "Denmark", "Norway", "Canad-
$ region       <chr> "Western Europe", "Western Europe", "Western Europe", ~
$ score        <dbl> 7.587, 7.561, 7.527, 7.522, 7.427, 7.406, 7.378, 7.36-
$ gdp_index    <dbl> 1.39651, 1.30232, 1.32548, 1.45900, 1.32629, 1.29025, ~
$ family_index <dbl> 1.34951, 1.40223, 1.36058, 1.33095, 1.32261, 1.31826, ~
$ lifeexp_index <dbl> 0.94143, 0.94784, 0.87464, 0.88521, 0.90563, 0.88911, ~
$ freedom_index <dbl> 0.66557, 0.62877, 0.64938, 0.66973, 0.63297, 0.64169, ~
$ trust_index   <dbl> 0.41978, 0.14145, 0.48357, 0.36503, 0.32957, 0.41372, ~
$ generosity_index <dbl> 0.29678, 0.43630, 0.34139, 0.34699, 0.45811, 0.23351, ~
$ year          <int> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, ~
```

	country	region	score	gdp_index	family_index	lifeexp_index	freedom_index	trust_index	generosity_index	year
1	Switzerland	Western Europe	7.587	1.39651	1.34951	0.94143				
2	Iceland	Western Europe	7.561	1.30232	1.40223	0.94784				
3	Denmark	Western Europe	7.527	1.32548	1.36058	0.87464				
4	Norway	Western Europe	7.522	1.45900	1.33095	0.88521				
5	Canada	North America	7.427	1.32629	1.32261	0.90563				
6	Finland	Western Europe	7.406	1.29025	1.31826	0.88911				
1				0.66557	0.41978	0.29678	2015			
2				0.62877	0.14145	0.43630	2015			
3				0.64938	0.48357	0.34139	2015			
4				0.66973	0.36503	0.34699	2015			
5				0.63297	0.32957	0.45811	2015			
6				0.64169	0.41372	0.23351	2015			

Research Question 1: Life Expectancy, GDP, and Happiness

Is there a significant association between life expectancy and national happiness, after adjusting for GDP per capita?

Model Specification:

We estimate a multiple linear regression model to study how life expectancy and economic development jointly relate to national happiness. The outcome variable is the country-level happiness score reported in the World Happiness Report, which is constructed relative to a “Dystopia” baseline that represents the lowest observed levels of each well-being component. All component indices, including life expectancy, GDP per capita, family support, freedom, trust, and generosity, take values between 0 and 1 and quantify how far each country lies above this Dystopia reference point. Because the outcome is continuous and the predictors are measured on interval-like scales, a linear regression framework is appropriate. The model relies on the standard assumptions that (1) the conditional relationship between predictors and the outcome is linear, (2) residuals have constant variance across fitted values, (3) residuals are independent across observational units, and (4) residuals are approximately normally distributed.

The primary objective is to examine whether the association between life expectancy and happiness differs across levels of economic development. To address this, the model includes life expectancy and GDP per capita indices as continuous predictors along with their interaction. In this specification, the individual coefficients on life expectancy and GDP per capita cannot be interpreted as marginal effects; instead, they contribute to a combined effect that depends jointly on both indicators. The interaction term captures whether improvements in life expectancy translate into larger or smaller increases in happiness depending on a country’s economic conditions.

All remaining available well-being components—family support, freedom, trust, and generosity—are included as covariates to reduce confounding. These factors represent social, institutional, and relational environments that plausibly influence both life expectancy and subjective well-being, and adjusting for them allows the model to better isolate the association of interest. Because the dataset spans multiple time points, the model also includes the survey year as a continuous predictor, along with interactions between year and both life expectancy and GDP per capita.

Since countries contribute repeated observations to the dataset, standard errors are computed using country-level clusters—robust variance estimators. This approach addresses within-country dependence that would otherwise bias classical standard errors and ensures valid inference even in the presence of serial correlation or heteroskedasticity.

Table 1: Linear regression model for national happiness (cluster-robust SEs by country)

Variable	Estimate	Std Error	t-value	p-value	2.5% CI	97.5% CI
Intercept	138.048	85.540	1.614	0.107	-29.872	305.967
Life expectancy index	-341.906	168.642	-2.027	0.043	-672.959	-10.853
GDP per capita index	324.225	103.867	3.122	0.002	120.328	528.122
Year	-0.067	0.042	-1.583	0.114	-0.150	0.016
Family index	0.890	0.118	7.542	<0.001	0.658	1.122
Freedom index	1.480	0.274	5.394	<0.001	0.942	2.019
Trust index	0.367	0.560	0.656	0.512	-0.732	1.467
Generosity index	0.350	0.332	1.054	0.292	-0.302	1.001
Life expectancy × GDP per capita	1.195	0.486	2.461	0.014	0.242	2.148
Life expectancy × Year	0.170	0.084	2.028	0.043	0.005	0.334
GDP per capita × Year	-0.161	0.052	-3.116	0.002	-0.262	-0.059
R-squared	0.781					
Adjusted R-squared	0.778					

The main effect coefficients in Table 1, such as -341.906 for life expectancy and 324.225 for GDP per capita, appear numerically large because, in a model that includes interaction terms, each main effect represents the estimated association when all interacting variables are equal to 0. In this dataset, such zero values do not occur. The life expectancy and GDP indices are constructed relative to a Dystopia reference point and never take the value 0, and the year variable ranges from 2015 to 2019 rather than approaching 0. As a result, the main effect coefficients reflect values at hypothetical predictor combinations outside the empirical data range. They are therefore not meaningful in isolation, and interpretation must instead focus on the interaction terms that determine the estimated relationships within the actual values observed in the dataset.

The interaction terms provide the relevant information for understanding how the association between life expectancy and happiness depends on economic level and time period. The interaction between life expectancy and GDP per capita has an estimated coefficient of 1.195 with a p value of 0.014 and a confidence interval from 0.242 to 2.148 , indicating that the estimated association between life expectancy and happiness becomes stronger at higher GDP levels. The interaction between life expectancy and year shows a similar pattern. Its estimated coefficient is 0.170 with a p value of 0.043 and a confidence interval from 0.005 to 0.334 . Within the observed period from 2015 to 2019, the model therefore estimates an increasing association between life expectancy and happiness across successive years. These two interactions confirm that the effect of life expectancy cannot be summarized by a single slope and varies across both GDP per capita and year.

Additional covariates also contribute to explaining variation in happiness. The family index and freedom index have statistically significant positive coefficients of 0.890 and 1.480 respectively, indicating that they account for meaningful differences in happiness scores after

conditioning on all other predictors. Trust and generosity are not statistically significant in this specification, as their confidence intervals include 0. Overall model fit is strong, with an R squared of 0.781 and an adjusted R squared of 0.778, indicating that the model explains a substantial portion of the variation in national happiness.

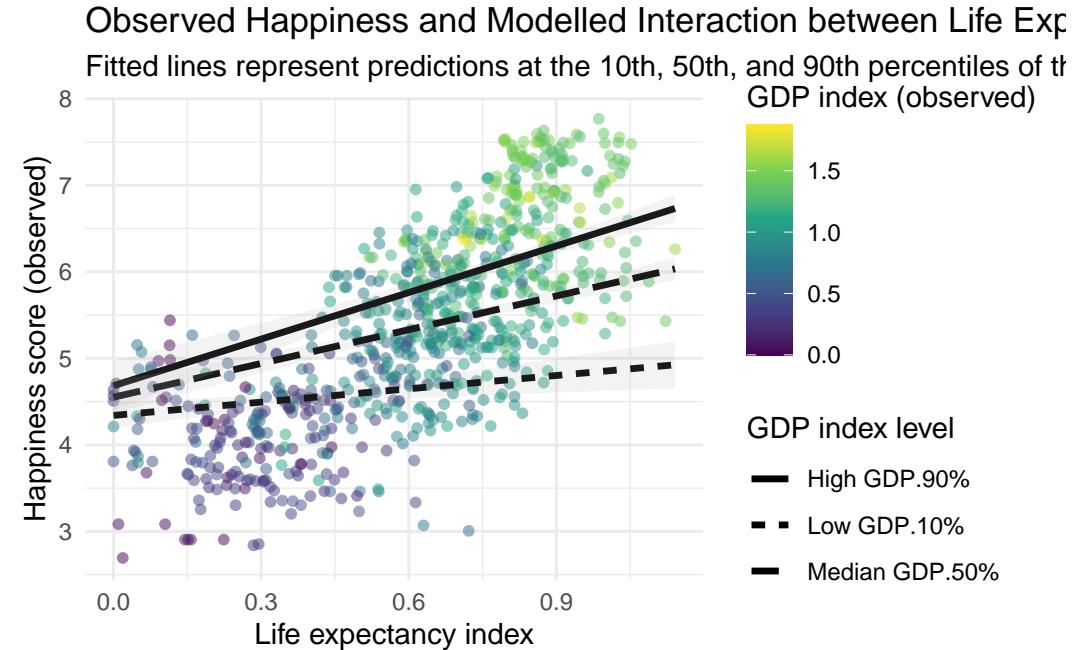


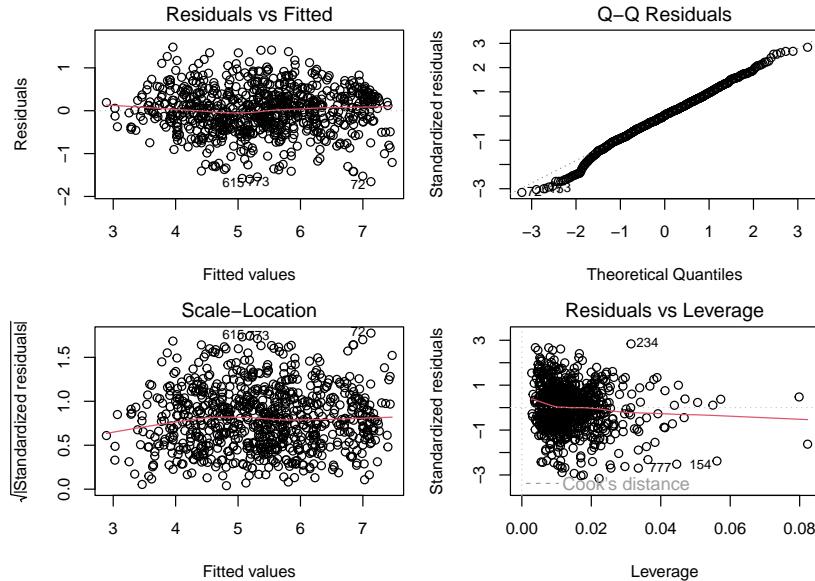
Figure above presents the observed relationship between the life expectancy index and national happiness, together with fitted regression lines that illustrate the estimated interaction between life expectancy and GDP. The horizontal axis displays the life expectancy component from the World Happiness Report. This variable is not measured in years; rather, it is a 0–approximately 1.03 scaled index representing how much a country’s life expectancy exceeds the report’s Dystopia benchmark—a hypothetical minimum reference level used across all countries.

The vertical axis reports the observed happiness score, which ranges from 0 to 10. Each point corresponds to a country–year observation, and its colour shows the observed value of the GDP component index. Similar to life expectancy, the GDP index is not GDP per capita in dollars. Instead, it is a 0–approximately 1.7 scaled measure indicating how far a country’s GDP per capita lies above the shared Dystopia baseline. The World Happiness methodology transforms GDP per capita into this index to make all component scores comparable on a common scale.

Superimposed on the scatterplot are three fitted regression lines representing predicted happiness at the 10th, 50th, and 90th percentiles of the GDP index, while all other covariates (family, freedom, trust, generosity, and year) are held at their mean levels. These lines summarize the interaction estimated in the regression model. The fitted line at the 90th percentile of the GDP index has the steepest positive slope, indicating that increases in the life expectancy

index are associated with larger increases in predicted happiness when GDP is relatively high. In contrast, the fitted line at the 10th percentile is noticeably flatter, showing that improvements in the life expectancy index correspond to smaller predicted increases in happiness when GDP levels are low.

Model Assessment:



The residuals-versus-fitted plot displays residuals scattered on both sides of zero without a strong curved trend, indicating no obvious departure from linearity in the fitted relationship. Although the spread of points varies somewhat across the range of fitted values, the plot does not show a clear funnel shape, and no single region dominates the pattern. This visual pattern suggests that the linearity assumption is reasonably supported by the observed residuals.

The Q–Q plot shows that standardized residuals follow the theoretical quantile line closely through the central portion of the distribution, while deviations are visible in both tails. The extreme points fall away from the reference line more noticeably than the bulk of the observations, indicating that the residual distribution departs from perfect normality primarily at the extremes. The majority of points remain near the line, and the departures are concentrated in the upper and lower tails as illustrated in the figure.

The scale–location plot presents standardized residuals whose average level remains relatively stable across fitted values, with a slight increase in variability toward the higher end of the fitted range. While not perfectly uniform, the pattern does not display a strong upward or downward trend, and the observed heterogeneity in spread appears modest based on the visual distribution. This suggests mild heteroskedasticity, though the figure does not show a severe or systematic change in variance.

The residuals-versus-leverage plot indicates that most observations fall within a low leverage range, forming a dense cluster near the vertical axis. A few points appear at higher leverage values, but none approach the Cook's Distance reference curves drawn in the plot. All points lie well below these curves, and no observation stands out as exerting undue influence based on its position relative to the reference boundaries shown. Thus, the figure does not provide visual evidence of highly influential observations.

We evaluated multicollinearity using Variance Inflation Factors (VIFs). In the initial model, the interaction terms produced extremely large VIF values. This occurs because an interaction constructed from two uncentered predictors is inherently highly correlated with those same predictors: the product term increases in tandem with each component, creating structural collinearity even when the underlying data themselves do not exhibit problematic relationships.

To address this, the predictors involved in the interaction were mean-centered before refitting the model. Centering removes the linear dependence between each predictor and the product term by redefining the interaction around the average levels of the predictors rather than around zero, which is not meaningful in this context. After centering, the VIFs for the two main predictors were approximately 2.68 and 2.67, and the VIF for the interaction term fell to approximately 1.15.

lifeexp_index	gdp_index	year	
6.062630e+06	6.331141e+06	8.279030e+00	
family_index	freedom_index	trust_index	
2.270764e+00	1.629919e+00	1.670560e+00	
generosity_index	lifeexp_index:gdp_index	lifeexp_index:year	
1.243699e+00	2.700847e+01	6.065838e+06	
gdp_index:year			
6.333643e+06			

Table 2: Variance Inflation Factor (VIF) for Centered Model

	Predictor	VIF
lifeexp_centered	lifeexp_centered	2.680526
gdp_centered	gdp_centered	2.672042
lifeexp_centered:gdp_centered	lifeexp_centered:gdp_centered	1.145777