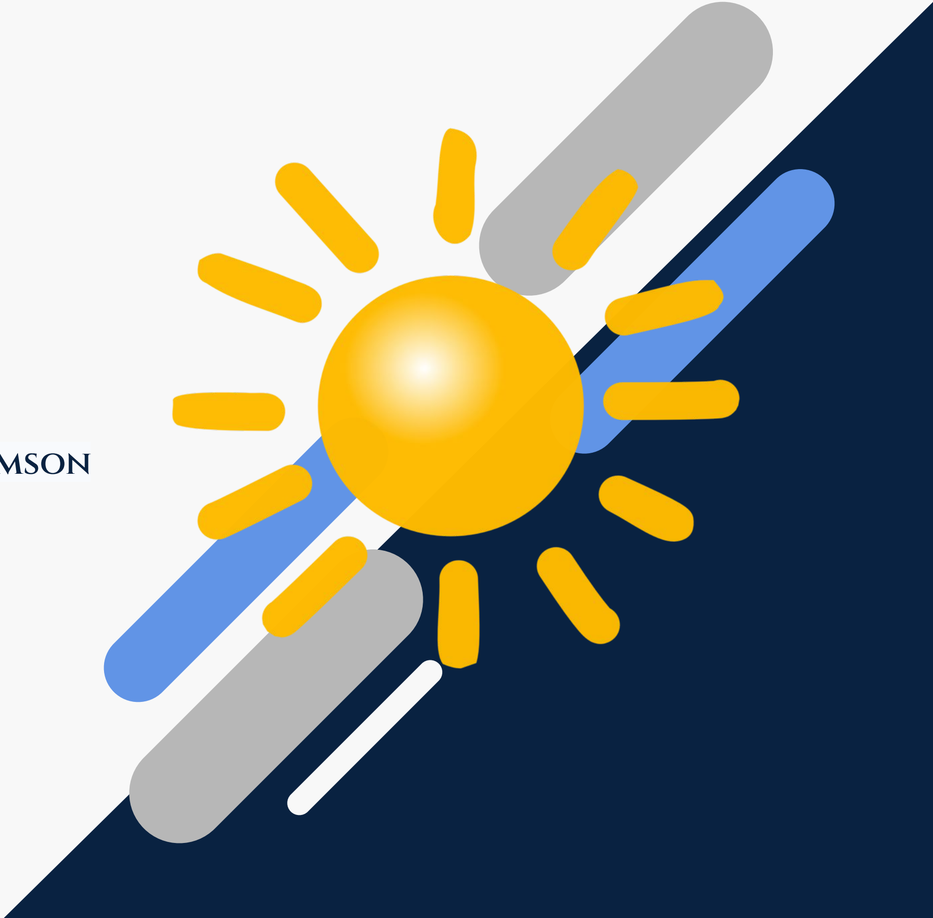
Predicting Life Expectancy

Stats 101A – Group 25

By: Lucy Borkowski, Ngoc Nguyen, Anika Chowdhury, Sriya Donepudi, Divya Kumar, Samson Huynh



Introduction

- **Motivation**:
  - how can worldwide policymakers improve public health?
  - investigate well-being using data from different countries*
  - factors known to be qualitatively linked to health: happiness, economic well-being, perceptions of corruption
- **Investigation**:
  - How do **happiness** score, **economic** well-being, and **perceptions of corruption** quantitatively affect life expectancy?
- **Data***:
  - Kaggle: World Happiness Report
  - 143 observations
  - international data from Gallup World Poll

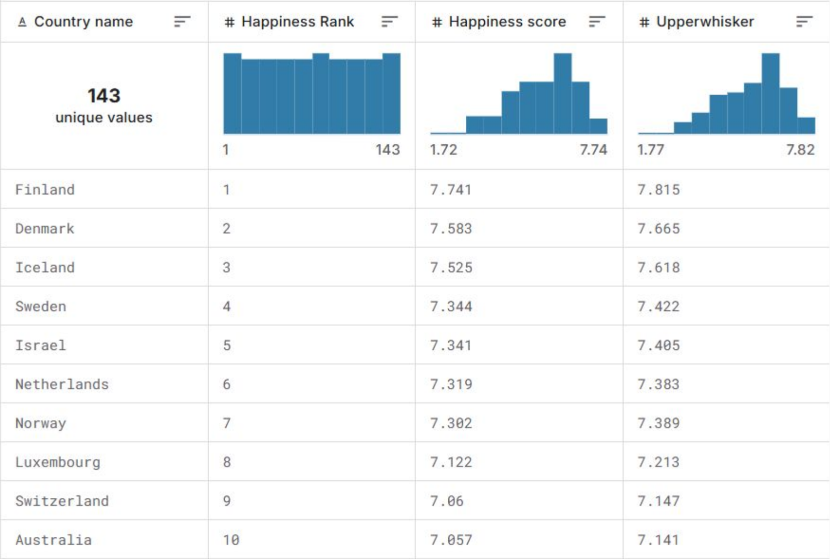
Set Up

Variables

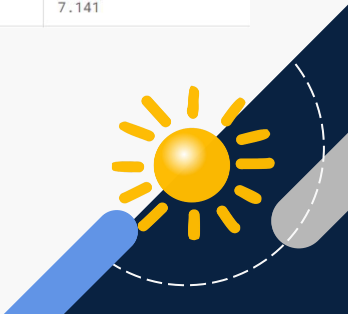
- **Predictors**:
  - happiness score (1-10 scale)
  - economy (GDP per capita)
  - perceptions of corruption (public trust in government)
- **Response**:
  - healthy life expectancy

Model

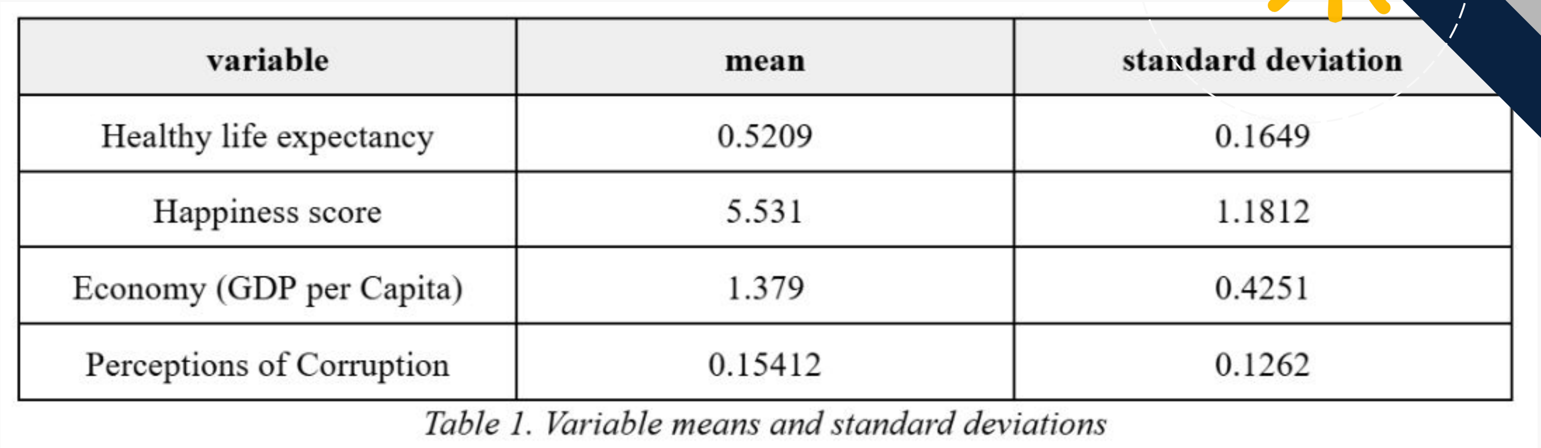
- Multiple linear regression



Country name	Happiness Rank	Happiness score	Upperwhisker
Finland	1	7.741	7.815
Denmark	2	7.583	7.665
Iceland	3	7.525	7.618
Sweden	4	7.344	7.422
Israel	5	7.341	7.405
Netherlands	6	7.319	7.383
Norway	7	7.302	7.389
Luxembourg	8	7.122	7.213
Switzerland	9	7.06	7.147
Australia	10	7.057	7.141



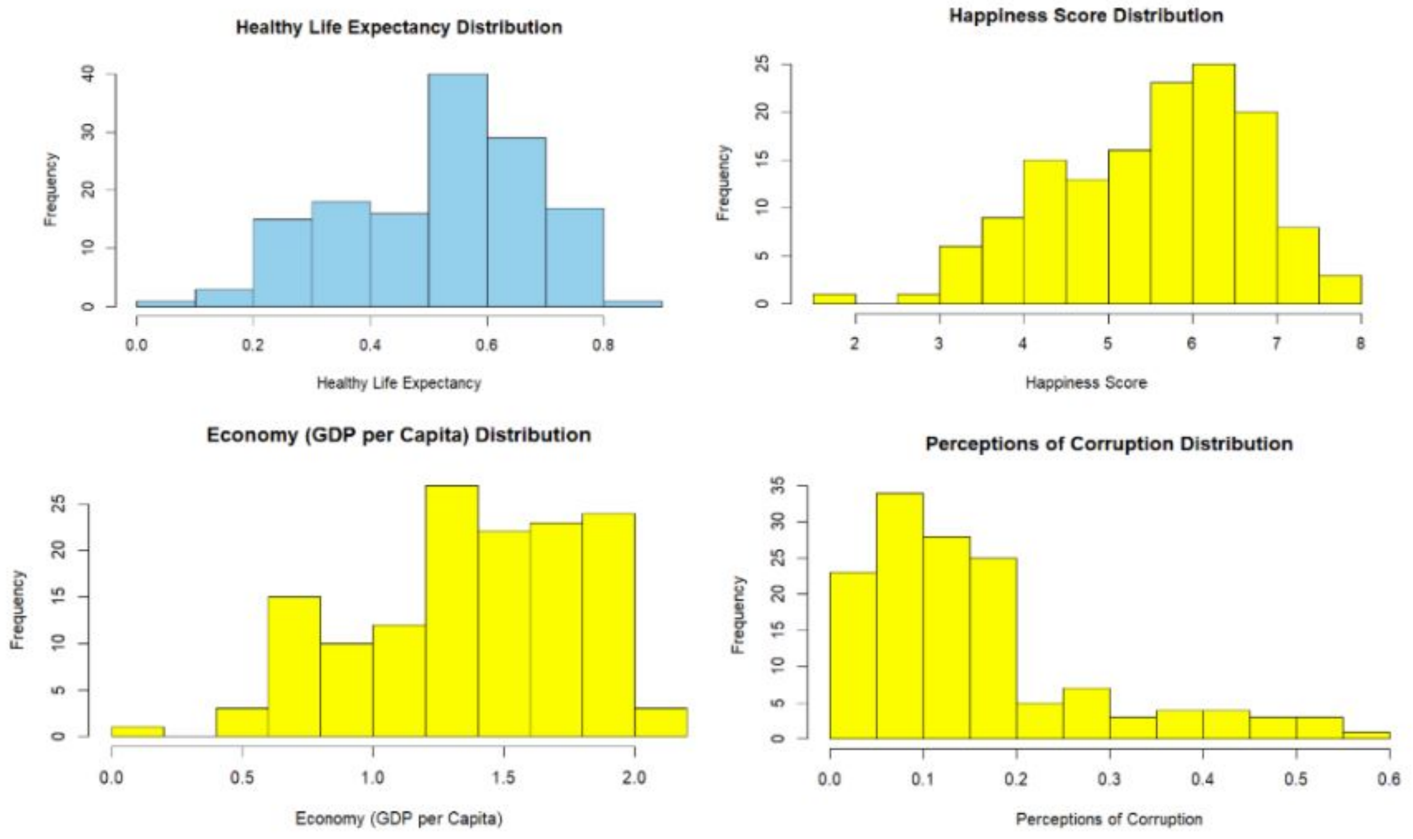
Data Analysis



variable	mean	standard deviation
Healthy life expectancy	0.5209	0.1649
Happiness score	5.531	1.1812
Economy (GDP per Capita)	1.379	0.4251
Perceptions of Corruption	0.15412	0.1262

Table 1. Variable means and standard deviations

Data Analysis



Healthy Life Expectancy Distribution

Happiness Score Distribution

Economy (GDP per Capita) Distribution

Perceptions of Corruption Distribution

Figure 1. Distribution of variables

- left skewed: life expectancy, Economy, Happiness
- right skewed: Perceptions of corruption

Data Analysis

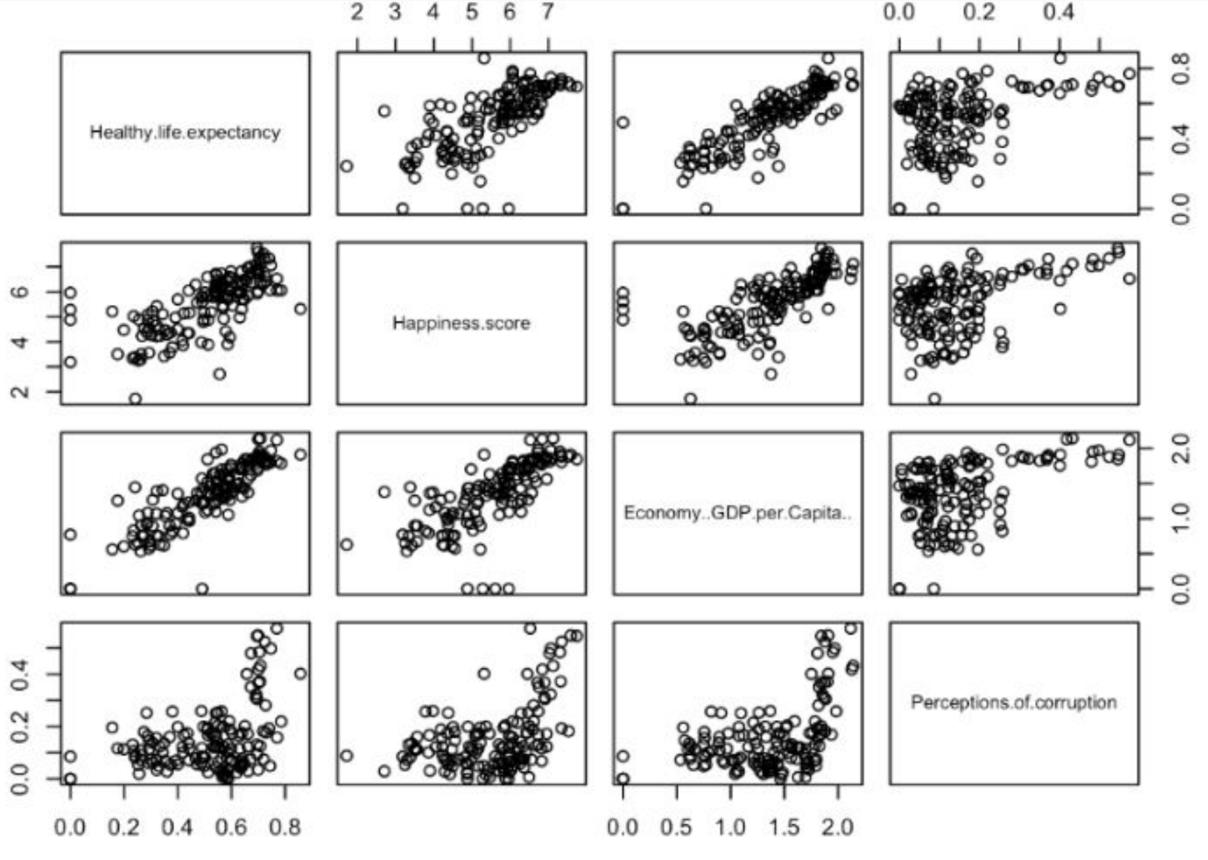


Figure 2. Scatterplot matrix of variables

Observations

Positive correlation:

- life expectancy & happiness score
- life expectancy & GDP per capita

No strong relationship:

- life expectancy & perceptions of corruption

No major collinearity

Methods

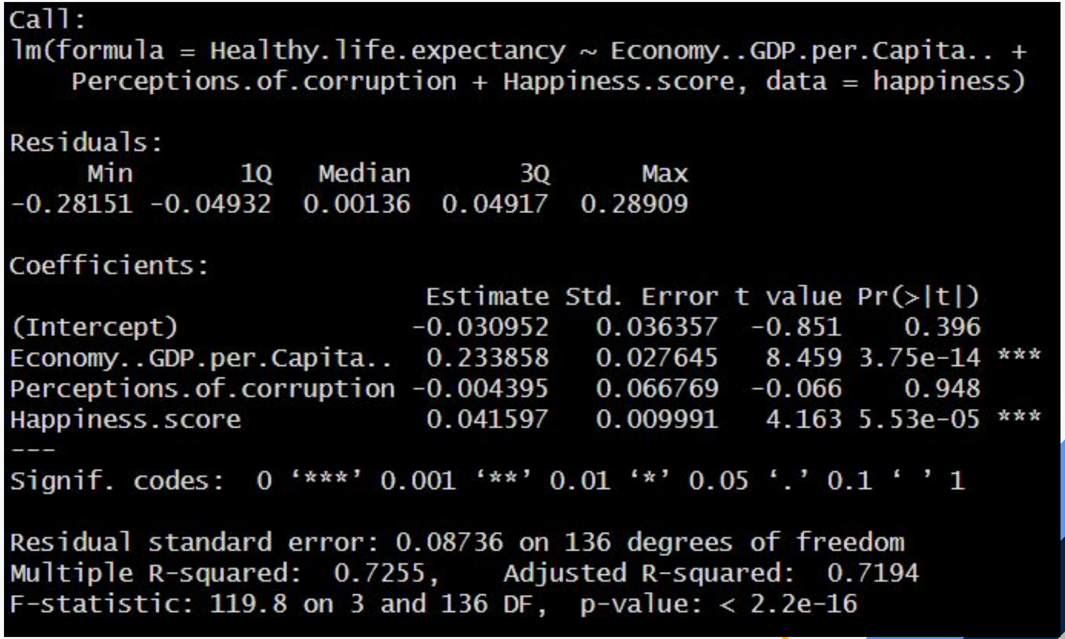
Model:

Healthy.Life.expectancy =

-0.0310 + 0.234 * Economy -0.00439 * Corruption + 0.0416 * Happiness.score

Output:

- $R^2 = 0.7255$ = › 72.55% of variance explained
- ANOVA p-value ‹ 0.05
    - model stat. significant
- p-values of predictors
    - significant: economy & happiness
    - not significant: perception of corruption



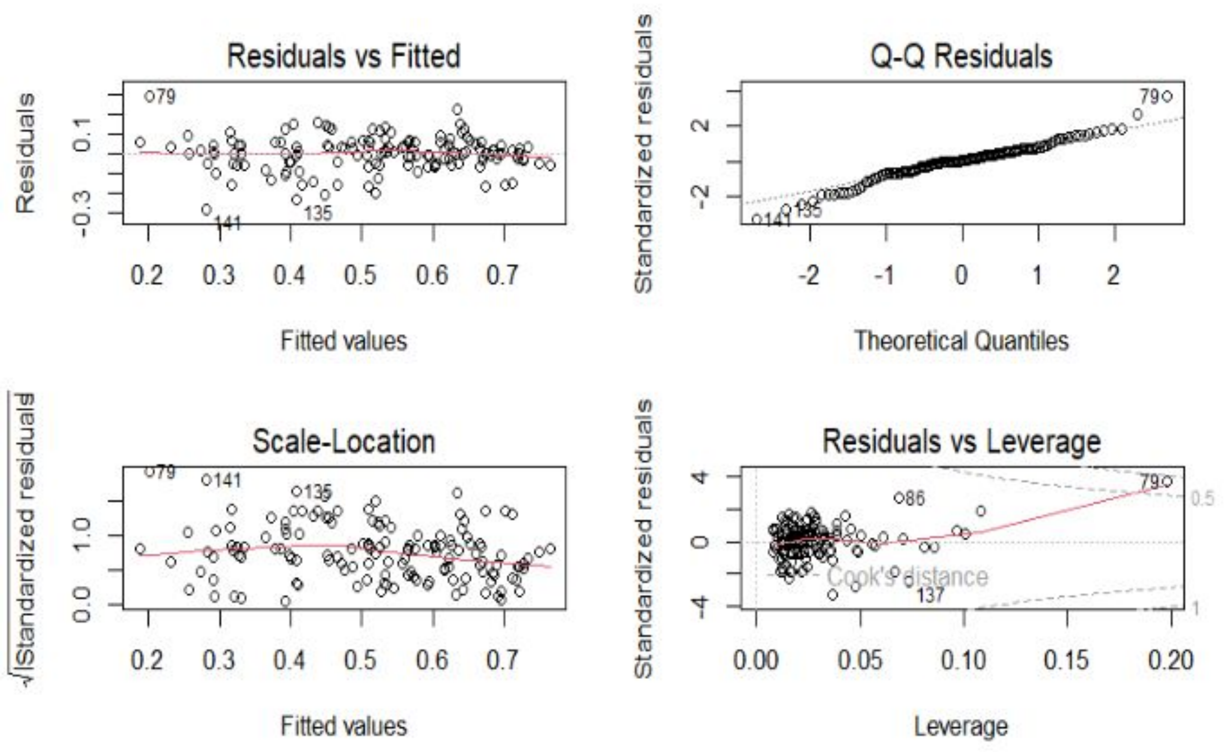
```
Call:
lm(formula = Healthy.life.expectancy ~ Economy..GDP.per.Capita.. +
    Perceptions.of.corruption + Happiness.score, data = happiness)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28151 -0.04932  0.00136  0.04917  0.28909

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -0.030952   0.036357  -0.851    0.396
Economy..GDP.per.Capita..  0.233858   0.027645   8.459 3.75e-14 ***
Perceptions.of.corruption -0.004395   0.066769  -0.066    0.948
Happiness.score            0.041597   0.009991   4.163 5.53e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08736 on 136 degrees of freedom
Multiple R-squared:  0.7255,    Adjusted R-squared:  0.7194
F-statistic: 119.8 on 3 and 136 DF,  p-value: < 2.2e-16
```

METHODS

Diagnostic Plots:

- assumptions of multiple linear regression (linearity and homoscedasticity) largely met in residual vs. fitted and scale-location plot but…
- <u>some issues</u>:
  - slight right skew & heavy tails in Q-Q (resids not normal)
  - a few outliers (obs 79, 86, 137) in residual vs leverage

Try: Improve violations with a transformation!



Methods

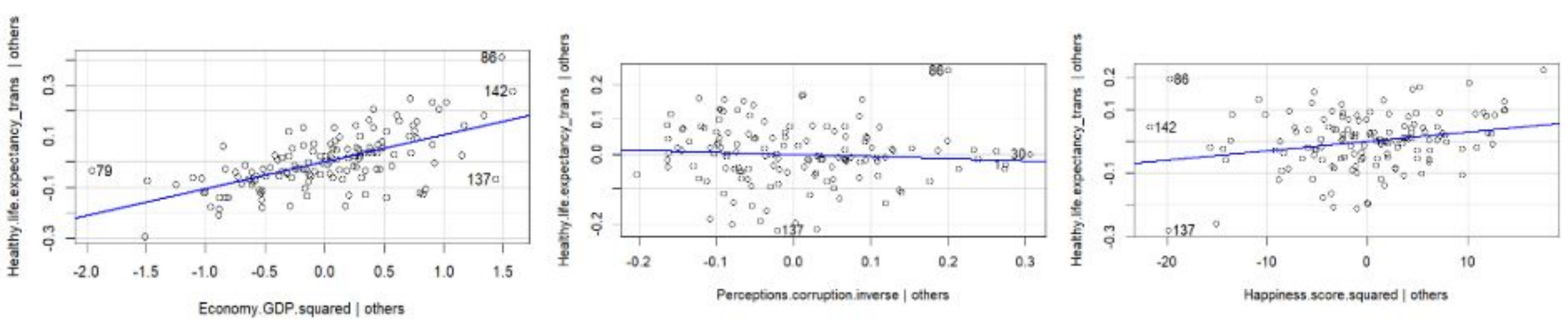
- **Used Box-Cox**
    - helps determine optimal power transformation
    - improve normality and make variance more constant
    - p-value < 0.05 which suggests log transformation *could* be used
    - Box-Cox: systematically test different transformations
        - shows: individual variable transformations = more effective
        - log transformation **not** best choice

- **Optimal transformations:**
    - life expectancy => $\lambda = 1.43$
    - economy => $\lambda = 2$
    - happiness score => $\lambda = 2$

- **Variable selection*:**
    - perceptions of corruption => not significant, removed

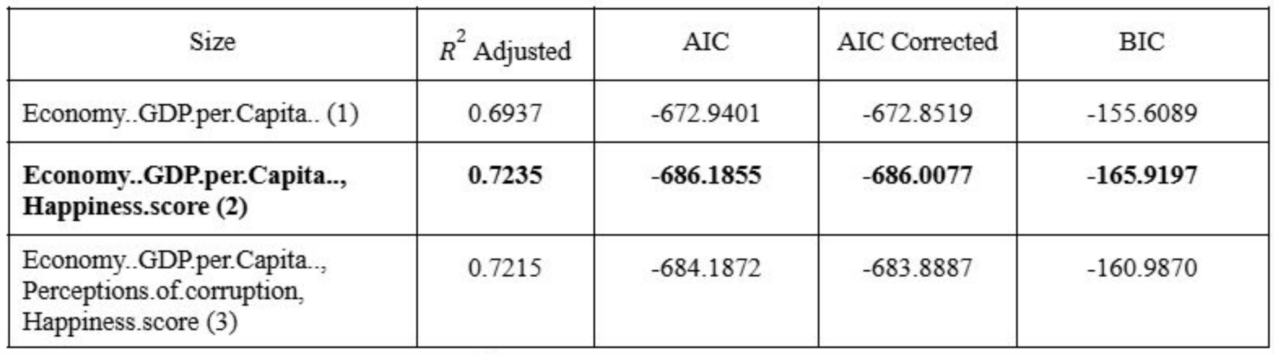
METHODS

- **Variable selection***: used the added variable plots and the Adjusted R^2, AIC, AICc, and BIC tests



Why remove perception of corruption?

- p-value high
- added variable plot: weak correlation
- AIC/BIC shows preference for this simpler model



Size	R^2 Adjusted	AIC	AIC Corrected	BIC
Economy..GDP.per.Capita.. (1)	0.6937	-672.9401	-672.8519	-155.6089
Economy..GDP.per.Capita.., Happiness.score (2)	**0.7235**	**-686.1855**	**-686.0077**	**-165.9197**
Economy..GDP.per.Capita.., Perceptions.of.corruption, Happiness.score (3)	0.7215	-684.1872	-683.8887	-160.9870

Final Model

After the entire process, **final model** we arrive at:

$(\text{Healthy.Life.expectancy})^{1.43}$ = **-0.0087** + **0.235** * $(\text{Economy})^2$ + **0.038** * $(\text{Happiness.score})^2$

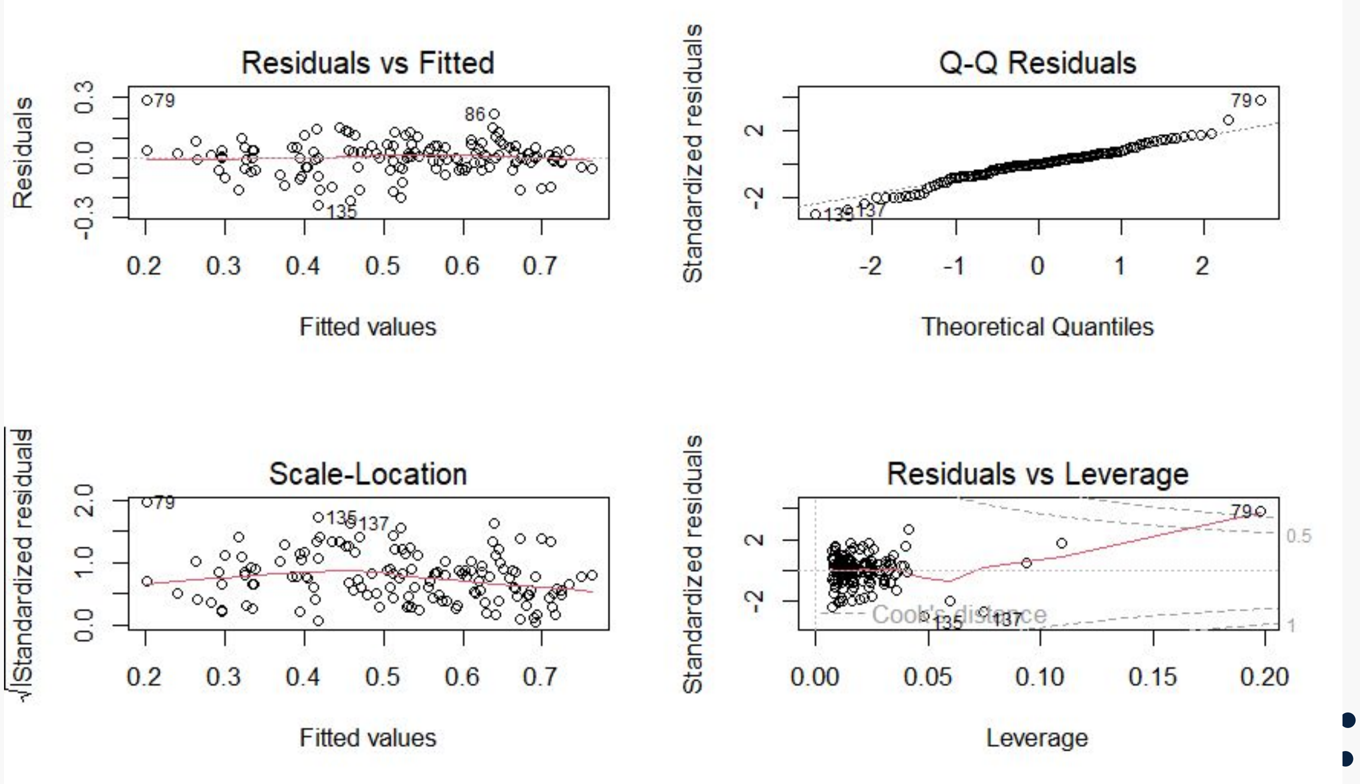
Coefficients & Interpretation:

- $\square_0$= -0.0087
  - Life expectancy is ~ 0 when Economy and Happiness score are both 0
- $\square_1$ = 0.235
  - holding all else constant, for one-unit increase in Economy there is a 0.235 increase in Healthy.Life.expectancy.
- $\square_2$ = 0.038
  - holding all else constant, for a one-unit increase in Happiness.score there is a 0.038 increase in Healthy.Life.expectancy.

Output:

- $R^2 = 0.7235$ => 72.35% of variance explained
- ANOVA p-value < 0.05 => model is statistically significant
- all predictors now significant

Final Model


Residuals vs Fitted
Residuals
-0.3
0.0
0.3
79
86
135
0.2
0.3
0.4
0.5
0.6
0.7
Fitted values
Q-Q Residuals
Standardized residuals
-2
2
79
135
137
-2
-1
0
1
2
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
0.0
1.0
2.0
79
135
137
0.2
0.3
0.4
0.5
0.6
0.7
Fitted values
Residuals vs Leverage
Standardized residuals
-2
2
79
0.5
1
Cook's distance
135
137
0.00
0.05
0.10
0.15
0.20
Leverage

Conclusion

- **Findings:**
  - Higher GDP & Happiness Score = > Increased healthy life expectancy (policy implications)
  - Perception of Corruption was not a significant predictor of life expectancy

- **Limitations & Challenges**:
  - bias from omitted variables
  - concerns about generalizability (emphasis on upper-middle class countries)
  - outliers (observations 79, 86, and 137) affected assumptions
  - our initial predictor of perception of corruption had weak predictive power

- **Future improvements to consider:**
  - testing models on future years' datasets
  - incorporating more variables
  - including a wider variety of countries' data