Analysis of National Happiness with Linear Regression

Noah B Johnson, Data Science, Luther College

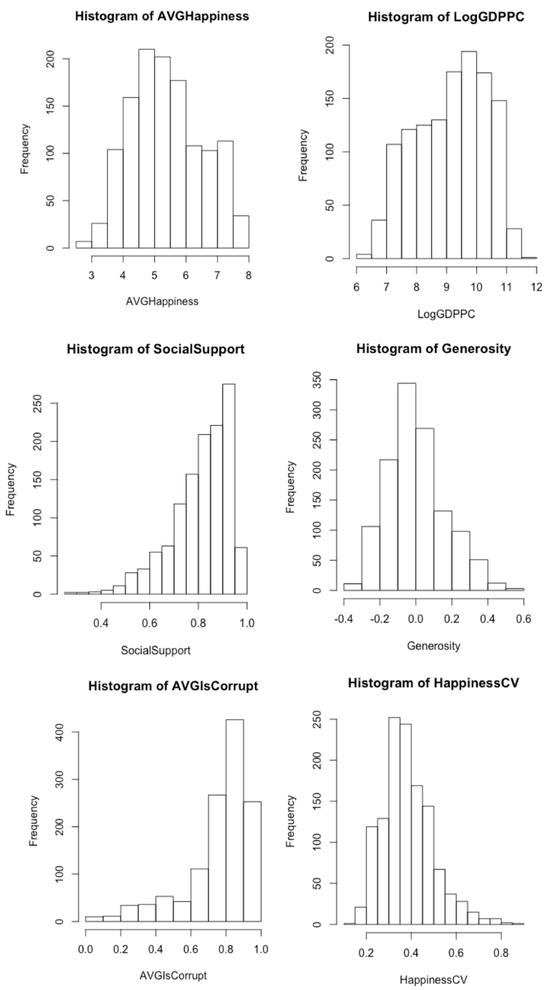
ABSTRACT

This report examines possible contributing factors to the Happiness of a nation's populace ("AVGHappiness") rated on a scale of 1-10. Most variables are national averages, mostly of survey responses, and the rest are national indicators. The primary goal of this report is determine the most predictive measure of average national Happiness. The data for this analysis comes from the Gallup World Poll and the World Bank.

DATA CHARACTERISTICS

The following data were used in the analysis presented here:

Variable Name	Description
AVGHappiness	1-10, averaged nationally
LogGDPPC	The log of the GDP per capita
SocialSupport	How much social support citizens feel they receive
Generosity	Average of 1=gave to charity this year and 0=did not
AVGIsCorrupt	The average of how corrupt citizens view their government to be
HappinessCV	The coefficient of variation of the response variable (essentially a measure of happiness inequality)



METHODOLOGY

Linear regression is a type of statistical modeling where the response variable is continuous. In multiple linear regression, more than one predictor is used to develop a model. The basic model is expressed by the following equation.

$$\Upsilon = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + \varepsilon$$

INITIAL MODEL FIT

A first-order model of AVGHappiness vs LogGDPPC, SocialSupport, Generosity, AVGIsCorrupt, and HappinessCV results in the following.

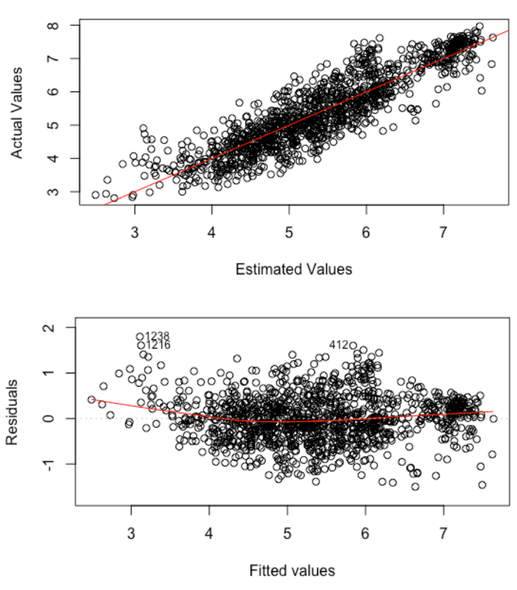
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.89921    0.21251   8.937  < 2e-16 ***
LogGDPPC      0.42761    0.01809  23.634  < 2e-16 ***
SocialSupport 1.57855    0.17433   9.055  < 2e-16 ***
Generosity    0.83932    0.09951   8.435  < 2e-16 ***
AVGIsCorrupt -0.39328    0.09090  -4.327 1.64e-05 ***
HappinessCV  -3.63235    0.17546 -20.702  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From this model, we can see that all of the variables appear to be statistically significant. Happiness appears to increase with LogGDPPC, SocialSupport, and Generosity. A nation's happiness appears to decrease with a higher AVGIsCorrupt and HappinessCV.

The adjusted R-squared for this model is .7898, which means that the model can explain almost 80% of the error in the data.

The p-value for a F (lack of fit) test is less than 2.2e-16, which indicates that this model is not a poor fit for the data.

INITIAL MODEL FIT:



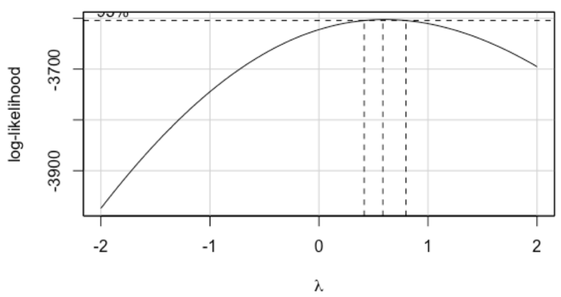
MODEL SELECTION

The initial model assumed there are no interactions between the predictor variables. After assessing whether or not any interactions should be included in the final model, there was little evidence that the model was a better fit with interactions included. In fact, the f-test for lack of fit indicated that including interactions may make the model a worse fit. The final model consists of:

X_1 = LogGDPPC
X_2 = SocialSupport
X_3 = Generosity
X_4 = AVGIsCorrupt
X_5 = HappinessCV

Box-Cox Transformation

When plotting the residuals vs the fitted values like we did in the initial model fit, there is some potential curvature in the plot. This indicates that a transformation of the response variable may increase the accuracy of the model. In order to determine the type of transformation needed, a box-cox analysis was performed.



The box-cox plot indicates that a transformation of around .6 would be appropriate. After application of this transformation, the r-squared value doesn't change much, but the standard error is much smaller. This means that the model is now a better fit.

Final model and parameter estimates:

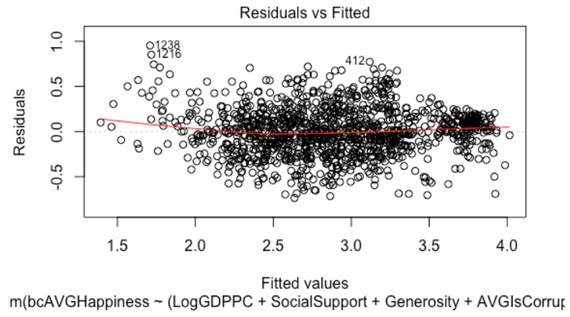
$$\Upsilon^{.6} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \varepsilon$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.045051   0.108564   9.626  < 2e-16 ***
LogGDPPC      0.219342   0.009243  23.730  < 2e-16 ***
SocialSupport 0.827084   0.089059   9.287  < 2e-16 ***
Generosity    0.393765   0.050837   7.746 1.97e-14 ***
AVGIsCorrupt -0.138227   0.046438  -2.977  0.00297 **
HappinessCV  -1.888129   0.089637 -21.064  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2652 on 1237 degrees of freedom
Multiple R-squared:  0.7899,    Adjusted R-squared:  0.789
F-statistic: 930.1 on 5 and 1237 DF,  p-value: < 2.2e-16
```

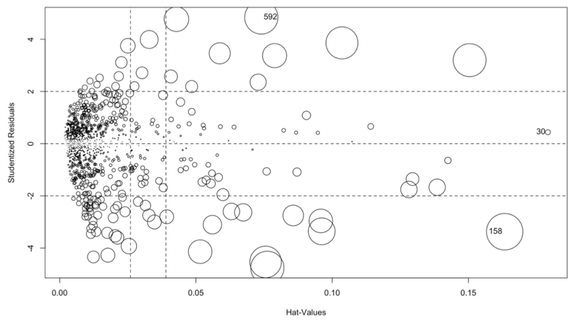
MODEL DIAGNOSTICS

A plot of residuals vs. fitted values shows marked improvement over the initial model as the curvature is much less apparent.



The p-value for the lack of fit test was less than .000000001, indicating no significant lack of fit for the model.

The influence analysis chart below has the residuals on the y-axis and the influence values on the x-axis. The larger the circle, the more influence an observation has on the model. In this case, there are some points that look like outliers, but they are not obvious enough to warrant removal.



CONCLUSIONS

This analysis shows that the average happiness of a nation was affected by that nation's GDP, Social Support, Generosity, Corruption, and Inequality.

- **More economic prosperity (GDP) usually means happier nations. (.219 LogGDPPC / AVGHappiness) Increased generosity had a similar effect.**
- **Nations in which citizens report higher levels of social support from their peers tend to be happier**
- **Governments perceived as corrupt have slightly unhappier citizens.**
- **More inequality among the happiness of a nation's populace goes hand-in-hand with a lower average happiness level.**