University of Warwick

Applied Statistical Modelling

ST404

World Happiness Report 2017 Analysis

Authors:

Joshua Gorman

Luke Hardcastle

Nathan Quinn

Yuqui Yang

February 6, 2018

1 Introduction

Research regarding life satisfaction in various countries throughout the world has become more prominent in recent years as improving happiness and well-being has become an important goal for policy makers. This report will concisely explore the relationship between life satisfaction, as measured by the Cantril Life Ladder and various socio-economic indicators, as well as critically discussing how these indicators have varying impact in different regions of the world. We will focus on 6 main indicators; "GDP per capita" (hereafter referred to as just GDP),"Support","Healthy Life Expectancy", "Freedom", "Positive Affect", and "Negative Affect".

2 Executive Summary

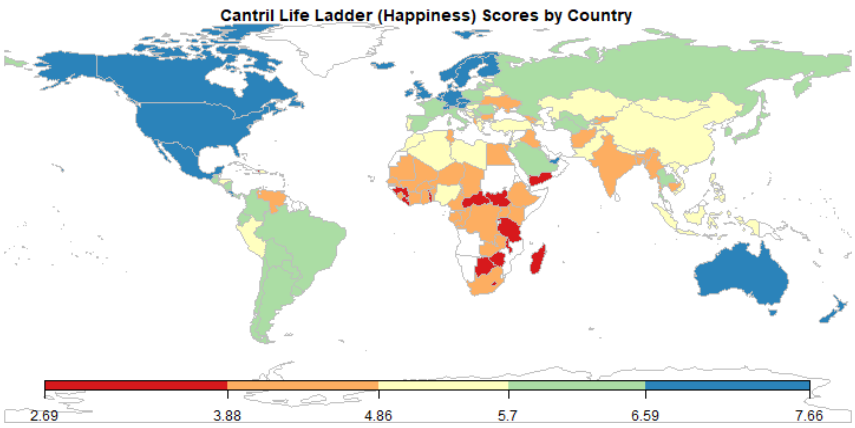

Figure 1: *Plots showing Cantril Life Ladder scores by country*

- Our findings indicate that life satisfaction is determined by four main indicators;
  1. 2016 GDP per capita, PPP (constant 2011 international $)
  2. Support - The percentage of people who felt they would have support if they were in trouble.
  3. Freedom - The percentage of people who were satisfied that they were free to choose what to do with their lives.
  4. Positive Affect - A personal account of 'happiness, laughing and enjoyment'

  In general, countries with higher values of these indicators have a higher happiness score. The happiest regions are "Western Europe" and "North America & Australia New Zealand", having scores of 6.8 and 7.1 respectively. Meanwhile those with the least life satisfaction were "Sub-Saharan Africa" and "South Asia", with scores of 4.1 and 4.8 respectively.
- The most influential indicator across *all* countries was Support, followed by Freedom. While Positive Affect was an indicator, it's counterpart Negative Affect was not - people having bad days did not effect a country's score. Similarly, anything explained by Life Expectancy was already, and more accurately, explained by GDP.
- Discrepancies exist between regions. Positive Affect has no influence in Africa, but is in fact the largest indicator in Western Europe, North American and Australia New Zealand. Furthermore GDP is the largest indicator when comparing poorer countries - mainly those in Africa. But it has relatively little effect when comparing richer countries. Every country has it's own political, cultural and societal circumstances and therefore exceptions exist compared to the general picture.

3 Findings

We received a data set collated for the "World Happiness Report 2017"[1]. For each country it contained a happiness score (known as the *Cantril Life Ladder*), and measures of: GDP per capita, Healthy Life Expectancy, Social Support, Freedom to make life choices, Positive Affect and Negative Affect. We attached to the dataset information on which continent and region it was in as laid out by the World Happiness Report. Our initial analysis revealed the relationships between these measures and the ladder score of a country to be as expected, with higher GDP, Life Expectancy, Support, Freedom and Positive Affect scores all associated with higher Ladder scores, whilst higher Negative Affect scores were associated with lower Ladder scores.

3.1 Regional Context

To study regional variation we used regions defined in the World Happiness Report. Oceania, North America and Europe were the happiest continents with mean Ladder scores of 7.24, 6.08 and 6.08 respectively, in comparison with Africa and Asia's scores of 4.24 and 5.39. In most cases the relationships between variables and Happiness do not differ from the global picture when viewed on a regional basis. There are however some regions where the size of these relationships differs. For example, in Africa Life Expectancy and Positive Affect has little effect on Happiness relative to other regions, whilst in the same region changes in GDP per capita have a much larger impact on Happiness compared to the rest of the world (Figure 2).

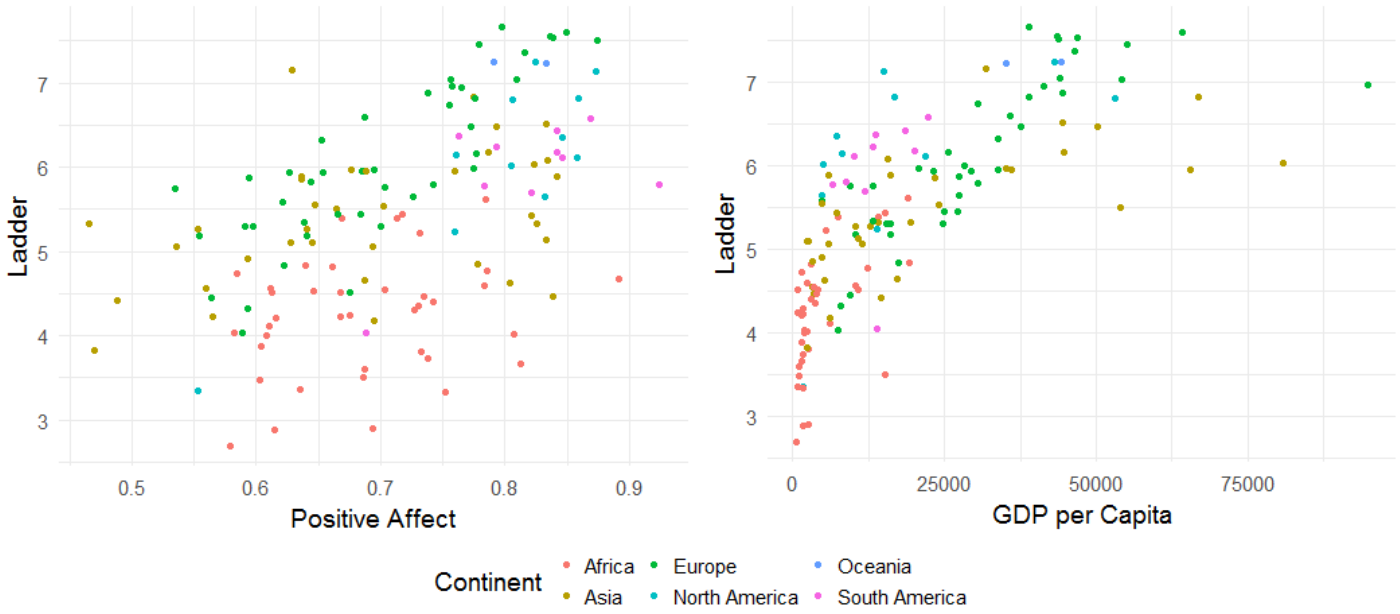

Figure 2: *Plots showing variation in relationships of Positive Affect, GDP and Happiness by continent.*

3.2 The Model

Our model recommends that to improve Happiness, richer countries should focus on improving Freedom, Support and Positive Affect, and poorer countries should focus on GDP primarily and then the other measures. Our model also shows that for African countries Positive Affect has no impact on Happiness, but for countries in Western Europe, North America and Australia & New Zealand, Positive Affect has an increased impact on Happiness.

3.3 Model Interpretation

Here we explore how to interpret the model with respect to each variable:

- Ladder: This score is based on the Cantril Life Ladder. It represents Happiness and shows how it changes when explanatory variables in our model change. It is a score from 0 to 10.
- 2016 GDP per capita, PPP (constant 2011 international $): We have taken the logarithm of GDP in our model. Here an increase in the GDP of a country corresponds to an increase in the Ladder score, but an increase in GDP for a country with high GDP will have less effect than on a country with lower GDP.
- Freedom: The Freedom score can be interpreted as a percentage. That percentage of 0.84 is then added to the Ladder score, so higher Freedom scores in countries correspond to higher Happiness scores.
- Support: The Support score can be interpreted as a percentage. That percentage of 2.26 is then added to the Ladder score, so higher Support scores in countries correspond to higher Happiness scores.
- Positive Affect: The Positive Affect score can be viewed as a percentage. That percentage of 2.20 is then added to the Ladder score, so higher Positive Affect in countries corresponds to higher Happiness scores.
- Region & Positive Affect: We have included indicator variables, they are equal to 1 if the country is in that specific region and 0 otherwise. We also include them interacting with Positive Affect. These can be interpreted as showing African country's Happiness is unaffected by Positive Affect and the Happiness in countries in Western Europe, North America or Australia New Zealand has a stronger relationship with Positive Affect than the rest of the world.

3.4 Limitations

3.4.1 Limitations of the Model

- The model does not contain any measure of Life Expectancy and so may experience problems explaining Happiness when Life Expectancy takes exceptional values.
- The model contains a negative intercept, so a hypothetical country could have a negative happiness score. However such a country would have to have such low scores in GDP, Freedom, Support and Positive Affect that we would never see such a place. Further, our model is primarily there to explain rather than predict.
- Our model implies that there is a negative relationship between happiness and positive affect in Africa. This seems counter-intuitive. However African countries in our model have little variation in their Positive Affect scores, so in practise this part of our model acts as a constant, showing Positive Affect has no effect on African countries.

3.4.2 Limitations of the Data

- The World Happiness Reports cites both Generosity and Corruption as key indicators for explaining Happiness and including these measures may have improved the explanatory power of our model. However we only had incomplete datasets for both these measures, so felt it better to exclude them.
- The question to measure freedom, asks "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?". This question is subjective and may be influenced by different cultural attitudes. This may therefore not be as useful for comparing freedom between countries.

4 Statistical Methodology

4.1 Exploratory Data Analysis

4.1.1 Missing Data

The data taken from the "World Happiness Report 2017" had a total of 12 missing data points;

- GDP - Argentina, Central African Republic, Malta, Myanmar, North Cyprus, Somalia, Taiwan and Yemen.
- Life Expectancy - North Cyprus.
- Freedom - Algeria, China and Iran.

We found alternative data sources for 8 out of 12 of the missing data points. The world bank, International Comparison Program data base[4] provided us with 5 of these as it contained 2016 GDP per capita, PPP (constant 2011 international $) values for, Argentina, Central African Republic, Malta, Myanmar and Yemen. Another 2 missing data alternatives were found using their previous years values collected by the "World Happiness Report 2017"[1]. China and Algeria had freedom scores of 0.805 in 2013 and 0.587 in 2012 respectively. The final missing data point was the GDP per capita for Taiwan which we filled using data from the 'Statistics Times'[3]. As this data point was a measurement of 2016 GDP per capita, PPP (current international $) we had to divide the value by a factor of 1.075 to take into account changes in the implied prices when switching from the PPP current prices to the PPP 2011 prices used in the data from the "World Happiness Report 2017".

There were four missing data points remaining that we couldn't find alternative data sources for. Two of these points were the missing GDP per capita and healthy life expectancy values for North Cyprus. As Turkey is the only country that recognizes this territory as an independent republic and we already had data points with no missing information for Turkey and Cyprus in our data set, we decided to remove North Cyprus from the data set. This left us with missing data points of GDP per capita for Somalia and Freedom for Iran. To avoid including inaccurate information, Iran and Somalia were removed. Our data set is large enough to ensure our conclusions are still valid.

4.1.2 Univariate Properties

We examined the density plots for each variable and the interesting cases (see figure 3) are discussed below;

- The distribution of **LifeExp** is positively skewed. There are spike points at 55 and 62.97 (the mean). We also note that 15 out of 17 data points between 50 and 55 are collected from Africa.
- **GDP** appears to have a negatively skewed Normal distribution, we observed 5 outliers above the upper boundary in the boxplot but the regions of these points varied (there was no particular pattern).
- The distribution of **Support** is a positively skewed normal distribution, with 6 outliers under the lower boundary of the boxplot. Five of them are from Africa whilst only one is from Europe.
- For **PosAffect** we observed a bimodal distribution with two spikes being at 0.66 and 0.81. This indicates our sample can be grouped according to certain characteristics. When we look at the group with PosAffect <0.7 (0.7 is the midpoint between two spikes), most samples in this group are collected from "Sub-Saharan Africa" region. For the group with PosAffect >0.7, most samples are from Western Europe region. This finding gives us the motivation to include Region or Continent as an additional variable in our model.

The graphs not shown revealed that **Ladder** is mostly Normal, but there is a small spike at 4.5, which comes

from 11 countries in Sub-Saharan Africa. **NegAffect** is lightly right skewed, **Freedom** is left skewed.

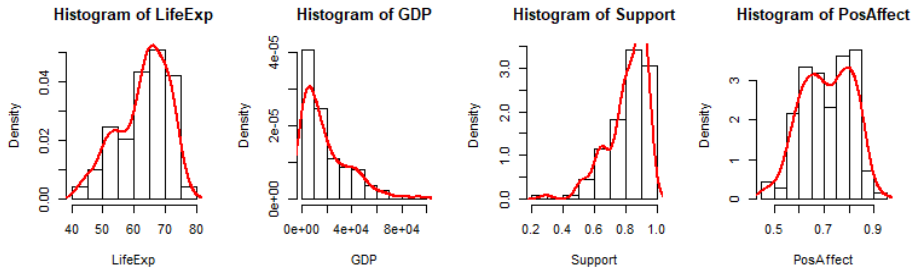

Figure 3: Density Plots of LifeExp, GDP, Support and PosAffect

4.1.3 Bivariate & Multivariate Properties

- The relationship between Ladder and every variable other than NegAffect is positive. We observed that NegAffect has a negative relationship with Ladder.
- **Predictors' Multicollinearity Analysis** - There exists several significant correlations between: GDP and LifeExp(r=0.72,p<2.2e-16), Support and LifeExp(r=0.67,p<2.2e-16), Freedom and PosAffect (r=0.59, p=2.312e-15). This gives us the motivation to remove some of them from our model, since they can be linearly predicted from the remaining variables with a substantial degree of accuracy.
- **Regional Multicollinearity** - By checking the data points in Africa, we find an insignificant correlation between Ladder and PosAffect (r=0.2401876,p=0.1464). This suggest that, Ladder varies little with PosAffect in Africa. However, in Western Europe Ladder is significantly correlated to PosAffect (r=0.805,p=1.885e-05).

4.2 Modelling Decisions

Our first model consisted of Ladder regressed on all available numerical variables. Examining a residual plot shows that the assumption of linearity is not satisfied. The 'U' shape suggests we should consider variable transformations(fig.4).

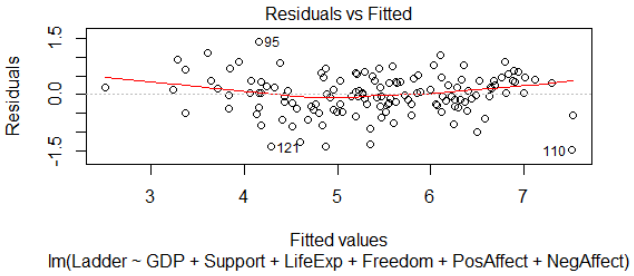

Figure 4: Residuals of inital model

4.2.1 Variable Transformations

A Box-Tidwell test informs us that a log transform of GDP may improve the linearity of the model (λ=0.215, p=0.00696). It also suggests a transform to the power 6 of PosAffect (λ=-6.47, p=0.0181) and a power transform of -3 to NegAffect (λ=-3.33, p=0.0402). However, such a high polynomial relationship between these variables is unrealistic, hence we consider the model with GDP transformed to log(GDP). There is sufficient evidence to justify a log transform of GDP. The use of the Box-Cox, Spread Level and Inverse Response plots, both before and after transforming GDP, do not provide evidence of the need to transform the response variable.

4.2.2 Variable Selection & Correlated Predictors

After transforming GDP to log(GDP), we conducted a best subset variable selection. The results show that the model without NegAffect is the best of these models, as it has the highest R^2_{adj} (0.787), the lowest Mallows CP (6.10), and the lowest BIC (254). Furthermore, NegAffect was not significant (p=0.296) which was also observed in our EDA. Looking at the VIF of our predictor variables, we find that log(GDP) and LifeExp have VIF of

4.28 and 3.77 respectively. This is larger than all other variables (Support 2.27, Freedom 1.68, PosAffect 1.72). We will therefore remove LifeExp from our model. It seems intuitive to include a GDP variable, as high GDP is likely to cause higher life expectancy, rather than the other way around. We in fact test both models with LifeExp and models with log(GDP) and we find that log(GDP) is marginally better. That being said, a model using LifeExp instead would also perform well. Therefore our current model is:

Ladder = log(GDP) + Support + Freedom + PosAffect

This model has a R^2_{adj} value of 0.787, with AIC=225 and BIC=245.

4.2.3 Polynomial & Interaction Terms

A residual plot is still "U" shaped however, and so further improvements to the model were made. The "U" shape suggests we are underestimating those countries which are low on the happiness ladder, and underestimating those at the higher end. To examine why this may be we can look at the data itself. We can see the unhappiest countries tend to be within Africa, while generally the happiest are in the regions "Western Europe" and "NA & ANZ". In particular the correlation between the continent of Africa and PosAffect is not statistically significant (p=0.146), but the correlation between the combined regions Western Europe and NA & ANZ is statistically significant (p=5.39e-07). Furthermore it is a very high correlation at 0.820. This suggests PosAffect does not impact the ladder score in Africa significantly, but for the regions "Western Europe" and "NA & ANZ" has a larger influence than other regions. We therefore introduce two interaction variables. One for the continent Africa and PosAffect, and another for the regions "Western Europe" or "NA & ANZ" again with PosAffect. This seems intuitive in that the regions Western Europe NA and ANZ share many cultural similarities.

4.2.4 The Final Model & Assumptions

Our final model is:

$$\textbf{Ladder} = -1.77529 + 0.342\log(\textbf{GDP}) + 2.257\textbf{Support} + 0.839\textbf{Freedom} + 2.201\textbf{PosAffect} + 2.027\mathbb{1}_{Africa} - 2.965\mathbb{1}_{WestEU,NA,ANZ} - 3.597\mathbb{1}_{Africa}\textbf{PosAffect} + 4.483\mathbb{1}_{WestEU,NA,ANZ}\textbf{PosAffect}$$

Every variable is significant at the 5% level, with an R^2_{adj} value of 0.829, with AIC=198 and BIC=227. Comparing this with our previous model shows this model is an improvement.

1. **Linearity & Mean Zero:** Examining the graph of residual plots against fitted values (fig. 7) shows an almost completely straight line at zero. This indicated our model is indeed linear, and further the errors have an expected value of zero.
2. **Independence:** To check for independence we look to see if the residuals are significantly associated with the regions or continents of their respective data points. A box plot of the residuals provides insight into this (figures 5 and 6). We can see that the only slight concern is the region South Asia. However, this region only has 5 data points, and in fact many of the regions have very few data points. Comparing by larger groups - as is done with continents, gives a much clearer picture, and clearly shows that residuals do not change significantly by region. We also observed residual plots by predictor variable, showing no correlation by variable.
3. **Homoscedasticity:** A plot of standardized residuals against fitted values gives insight into whether our model satisfies homoscedasticity. There appears to be some mild relationship between fitted value and size

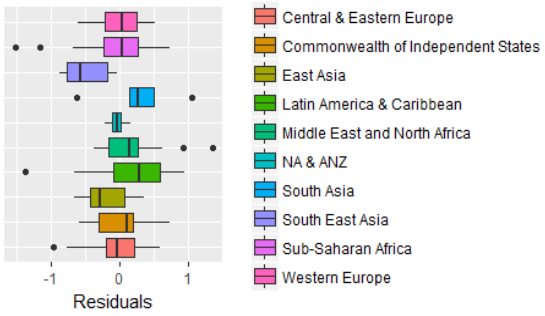

Figure 5: Residuals compared to regions

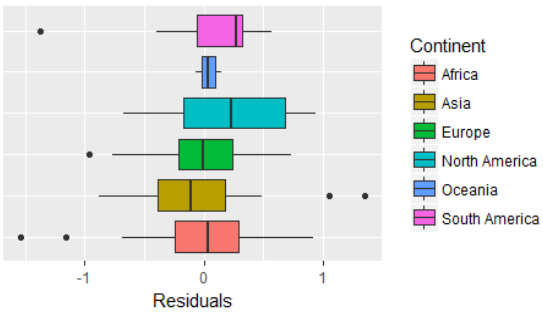

Figure 6: Residuals compared to continents

of residual – in particular the lower and highest fitted values tend to be more accurate. This has likely been caused by the introduction of interaction variables. However, performing an NCV test generates a p-value of 0.143, implying the residuals are indeed uncorrelated to fitted value. While this test is not a definitive answer, and one could argue from examining the plot that heteroscedasticity is present, we believe there is sufficient evidence to suggest the model satisfies the third assumption.

4. **Normally Distributed Residuals:** This assumption is best investigated using a Q-Q plot. The plot is not perfect, with particular problems towards the bottom tail. Further attempts to rectify this using other variables and testing other interaction terms yielded worse to no results. Nonetheless the Q-Q plot is still fairly good, with only a minority of potential outliers creating problems towards the tails. Therefore our model satisfies this assumption quite well.

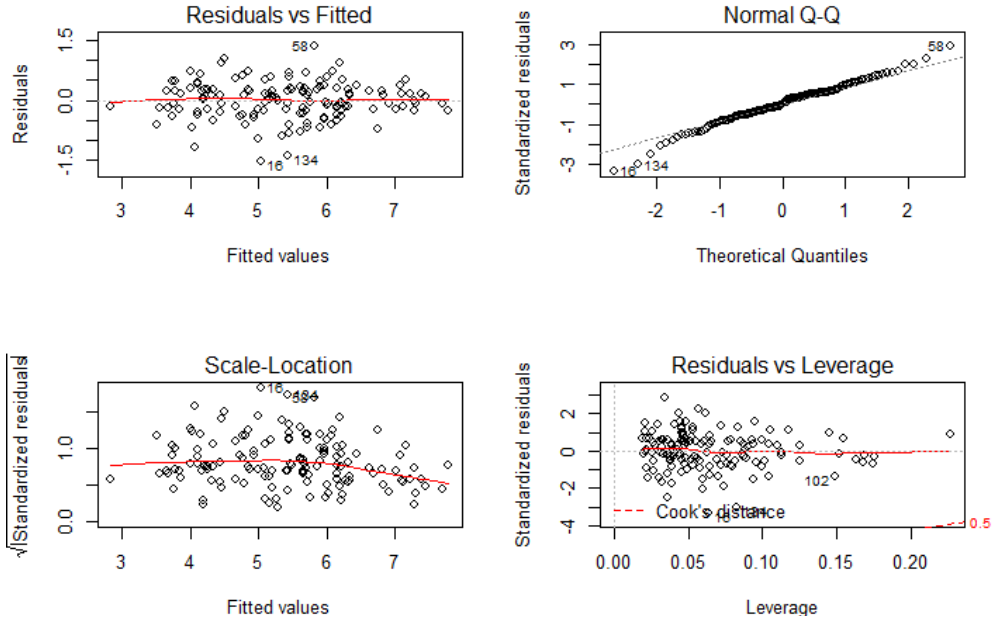

Figure 7: *Residual analysis*

4.3 Outliers and Influential Data Points

Outliers found are listed in Table 1.

Obs.	Country	Predicted Ladder	Ladder	log(GDP)	Support	Freedom	PosAffect	Continent
16	Botswana	5.033	3.499	9.625	0.768	0.852	0.686	Africa
58	Israel	5.810	7.159	10.367	0.890	0.772	0.629	Asia
134	Venezuela	5.418	4.041	9.534	0.902	0.458	0.688	S. America

Table 1: Outliers

Absolute studentized residuals much higher than 2, large cooks distances and low bonferroni p values (see figures 8 and 9) made it clear Botswana, Israel and Venezuela were outliers. Further investigation revealed:

- Botswana - Despite high freedom, average support and average log(GDP), Botswana's ladder score of happiness is low and so is over-estimated by the final model. From Figures 8 and 9 we can observe that Botswana has the highest absolute studentized residual score (greater than 3) and the the lowest Bonferroni p-value indicating that it is the most extreme observation.
- Israel - The final model under estimates the ladder score for Israel with a studentized residual score of roughly 3 which suggests this point has some leverage however it's Cook's distance score is not significant (below 0.04). Note that the Positive affect score for Israel is in the lower quartile of the data set and so is likely to play a major role in this under estimation.
- Venezuela - With a Cook's distance greater than 0.08 and the largest bubble radius (figure 9) it is clear Venezuela is a data point with high leverage. The over estimation can be explained by Venezuela's unusually high support.

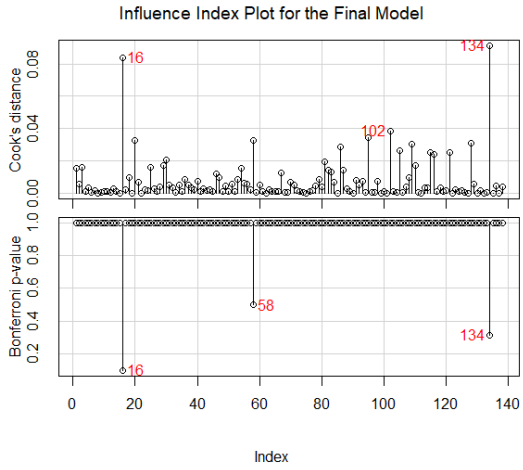

Figure 8: Residuals compared to regions

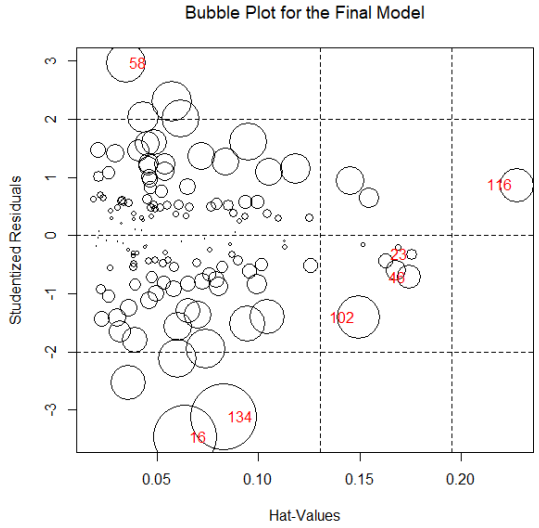

Figure 9: Residuals compared to continents

Overall we think that is important not to remove these influential points from our analysis. Although they have a high impact on the normality assumption of the model they do not influence other more essential assumptions such as linearity. In addition to this we have no reason to believe that the information collected is erroneous and so it would not be justifiable to delete the outliers.

4.4 Bibliography

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

[2]Peng RD, Exploratory Data Analysis in R[Internet],Bookdown,2016 [cited 2018 02 06]

[3] Statistics Times[internet], List of Asian countries by GDP per capita, Statistics Times 21 Feb 2017[cited 20 Jan 2018], available from: http://statisticstimes.com/economy/asian-countries-by-gdp-per-capita.php

[4] The World Bank[internet],GDP per capita PPP (constant 2011 international $), The World Bank, 2016,[cited 20 Jan 2018], Available from: https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.KD?locations=AR-CF-MT-MM-CY-YE

5 Authors' Contributions

Author	Contribution
Josh Gorman	Conducted the initial reading into Happiness, checked the data and found alternatives for missing data, performed outlier analysis. Wrote up the missing data and outlier sections. Compiled the report in LaTeX.
Luke Hardcastle	Performed the explatory data analysis for variation between regions. Performed the interactions and transformations. Checked model assumptions. Wrote up the findings section and compiled the appendix.
Nathan Quinn	Conducted the initial reading into Happiness. Performed the initial modelling, interactions and transformations. Checked model assumptions. Wrote up the modelling decisions section of the report.
Yuqiu Yang	Performed univariate and bivariate exploratory data analysis. Performed the initial modelling. Wrote up the section on univariate and bivariate exploaratory data analysis

After discussion, we agreed that we should equally distribute the final mark between the group.

Appendix

```
require(ggplot2)
require(dplyr)
require(leaps)
require(car)
require(doBy)
require(MASS)
require(gridExtra)
require(scales)
require(rio)
require(rworldmap)
require(classInt)
require(RColorBrewer)
```

```
FullData <- import("FullDataFinal.xlsx")
```

Executive summary

World Map

```
library('rworldmap')
Comp_Data <- dplyr::select(FullData, Country, Ladder) %>%
  mutate(Country=as.character(Country)) %>%
  mutate(Country=replace(Country,Country=="Palestinian Territories","Palestine")) %>%
  mutate(Country=replace(Country,Country=="Taiwan Province of China","Taiwan")) %>%
  mutate(Country=replace(Country, Country=="Hong Kong S.A.R., China","Hong Kong"))
colnames(Comp_Data)<-c("Country","Happiness Scores by Country")
sPDF <- joinCountryData2Map(Comp_Data, joinCode = "NAME",
                            nameJoinColumn = "Country", verbose="TRUE")
classInt <- classIntervals( sPDF[["Happiness Scores by Country"]]
                                      ,n=5, style = "jenks")
catMethod = classInt[["brks"]]
colourPalette <- brewer.pal(5,'Spectral')
mapDevice(device = 'png', file="worldGDP.png")
mapParams <- mapCountryData(sPDF
                            ,nameColumnToPlot="Happiness Scores by Country"
                            ,addLegend=FALSE
                            ,catMethod = catMethod
                            ,colourPalette=colourPalette )
do.call(addMapLegend
        ,c(mapParams
           ,legendLabels="all"
           ,legendWidth=0.4
           ,legendIntervals="data"
           ,legendMar = 2))
mapParams$legendText <-
  c('Lowest Scores '
    ,'Second Lowest Scores'
    ,'Medium Scores'
```

```
    ,'Second Highest Scores'
    ,'Highest Scores')
do.call( addMapLegendBoxes
         , c(mapParams
             ,x='left'
             ,bg=NA
             ,title="Happiness Score's Index"))
dev.off()
include_graphics("worldGDP.png")
```

Findings

Summary Statistics

```
tapply(FullData$Ladder, FullData$Continent, summary)
```

Function for adding a shared legend to plots

```
grid_arrange_shared_legend <- function(..., ncol = length(list(...)),
                                       nrow = 1,
                                       position = c("bottom", "right")) {
  plots <- list(...)
  position <- match.arg(position)
  g <- ggplotGrob(plots[[1]] +
                    theme(legend.position = position))$grobs
  legend <- g[[which(sapply(g, function(x) x$name) == "guide-box")]]
  lheight <- sum(legend$height)
  lwidth <- sum(legend$width)
  gl <- lapply(plots, function(x) x +
                 theme(legend.position = "none"))
  gl <- c(gl, ncol = ncol, nrow = nrow)

  combined <- switch(position,
                     "bottom" = arrangeGrob(do.call(arrangeGrob, gl),
                                            legend,ncol = 1,
                                            heights = unit.c(unit(1, "npc") - lheight,
                                                             lheight)),
                     "right" = arrangeGrob(do.call(arrangeGrob, gl),
                                           legend, ncol = 2,
                                           widths = unit.c(unit(1, "npc") - lwidth,
                                                           lwidth)))

  grid.newpage()
  grid.draw(combined)

  # return gtable invisibly
  invisible(combined)
}
```

Scatter plots by region

```
ScatterContinentPosAffect <- ggplot(FullData,
                                    aes(x=PosAffect, y=Ladder, color=Continent)) +
  geom_point() + theme_minimal(base_size = 14) + xlab("Positive Affect")
ScatterContinentGDP <- ggplot(FullData,aes(x=GDP, y=Ladder, color=Continent)) +
  geom_point() + theme_minimal(base_size = 14) +xlab("GDP per Capita")
grid_arrange_shared_legend(ScatterContinentPosAffect,ScatterContinentGDP,
                           ncol = 2,
                           nrow = ,
                           position = "bottom")
```

Statistical Methodology

Univariate Properties

```
par(mfrow=c(1,4))
hist(LifeExp,prob=T)
lines(density(LifeExp),col=2,lwd=2)
hist(GDP,prob=T)
lines(density(GDP),col=2,lwd=2)
hist(Support,prob=T)
lines(density(Support),col=2,lwd=2)
hist(PosAffect,prob=T)
lines(density(PosAffect),col=2,lwd=2)
```

Bivariate & Multivariate Properties

Matrix Scatter Plot

```
scatterplot.matrix(FullData[,c("Ladder","GDP","Support","LifeExp",
                               "Freedom","PosAffect","NegAffect")])
```

Modelling Decisions

The initial model & modelling decisions

```
attach(FullData)
model1 <- lm(Ladder ~ GDP + Support + LifeExp +
               Freedom + PosAffect + NegAffect) #The initial model
plot(model1)
```

Variable transformations

```
boxTidwell(Ladder ~ GDP+LifeExp+Support+Freedom+
             PosAffect+NegAffect)#Box-Tidwell Test

model2 <- lm(Ladder ~ log(GDP)+Support+
                          LifeExp+Freedom+
               PosAffect+NegAffect)
model3 <- lm(Ladder ~ log(GDP)+Support+LifeExp+
               Freedom+PosAffect)
boxcox(model2)
spreadLevelPlot(model2)
invResPlot(model2)
boxcox(model3)
spreadLevelPlot(model3)
invResPlot(model3)
vif(model3)
```

Variable selection & Correlated predictors

```
summary(model3)
best.subsets <- regsubsets(Ladder ~ log(GDP)+Support+
                          LifeExp+Freedom+PosAffect
                           +NegAffect, data = FullData)
summary.best.subsets <- summary(best.subsets)

AIC(model3)
BIC(model3)

model4 <- lm(Ladder ~ log(GDP)+Support+Freedom+PosAffect)
summary(model3)
summary(model4)
```

Polynomial & Interaction terms

```
W.eu.NA.ANZ <- FullData[FullData$Region == "NA & ANZ"|
                          FullData$Region == "Western Europe",]
Afric <- FullData[FullData$Continent == "Africa",]
cor.test(W.eu.NA.ANZ$Ladder,W.eu.NA.ANZ$PosAffect)
cor.test(Afric$Ladder,Afric$PosAffect)
```

The final model & assumptions

```
newmodel1<- lm(Ladder~logGDP+Freedom+Support+
                 as.factor(Continent == "Africa")*PosAffect+
                 as.factor(Region == "Western Europe"| Region == "NA & ANZ")*PosAffect)
```

```
summary(newmodel1)
AIC(newmodel1)
BIC(newmodel1)

#Linearity


#Independence

boxnew <- data.frame(na.omit(FullData), newmodel1$residuals)

#figure 4
boxregion <- ggplot(boxnew, aes(x = Region,
                                y = newmodel1.residuals,
                                fill=Region)) +
  geom_boxplot()+
  coord_flip()+
  theme(axis.title.y = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank())+
  ylab("Residuals")

#figure 5
boxcontinent <- ggplot(boxnew,
                       aes(x = Continent,
                                   y = newmodel1.residuals,
                                   fill = Continent))+
  geom_boxplot() +
  coord_flip() +
  theme(axis.title.y = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) +
  ylab("Residuals")


#Homoscedacity
ncvTest(newmodel1)

#figure 6
Assumptions <- par(mfrow=c(2,2))
plot(newmodel1)
par(Assumptions)
```

Outliers & influential data points

```
#Figures 7 & 8
influenceIndexPlot(newmodel1, id.n = 3,
                   main="Influence Index Plot for the Final Model",
                   id.col = "red",
                   vars=c("Cook", "Bonf"),cex.lab=1)
influencePlot(newmodel1, id.n = 3,
              main="Bubble Plot for the Final Model",
```

```
                id.co="red", font.main=1)
```