Multiple Linear Regressions on World Happiness Score

Yingyi Lin (Sunny)

California State University, East Bay

Abstract

The purpose of this project was to find out the determinant factors for happiness score in the World Happiness Report 2015. After obtaining raw data from the official website (World Happiness Report 2017, 2017), we observed linear relation between Happiness Score and almost all predictors, so we decided to apply multiple linear regression model. The first model from Akaike information criterion (AIC) test only had a few coefficients significant. In order to get a significant model with more significant coefficients, we developed our own regression model based on correlation. After dropping all the insignificant terms, we obtained a final model with six predictors. The final model passed all the assumptions. We used the final model to predict happiness score in 2016, and the model had good prediction performance. We concluded that emotional changes and feeling supportive affect the most on happiness.

Keywords: linear regression, happiness score, linear prediction

Introduction

Happiness is increasingly considered the proper measure of social progress and the goal of public policy (Helliwell, Layard and Sachs, 2017). The first World Happiness Report was published in 2012 (Helliwell et al, 2017). The Report published once a year containing many factors other than happiness score such as GDP, freedom, and age. In a recent speech, the head of the UN Development Program (UNDP) spoke against what she called the "tyranny of GDP", arguing that what matters is the quality of growth. "Paying more attention to happiness should be part of our efforts to achieve both human and sustainable development", she said (Helliwell et al, 2017). Since happiness is important to social progress, we decided to use the World Happiness Report as our study data. In this project, we will try to see if we can construct a regression model among the different reported factors and find out what affects happiness the most.

Materials and Methods

First, we obtained the raw data from Kaggle.com (World Happiness Report 2016, 2017). However, Kaggle.com doesn't provide any description or reference about the raw data. With a basic summary (See Appendix Figure 1), we found that some of the data ranges were strange. Life expectancy, which was supposed to be the age, had a range of 0 to 1.0252. Family, which we were not sure if it was the number of family member, had a range of 0 to 1.4022. Because the data was vague, we assumed that it has been modified and decided not to use it.

After researches, we obtained the original raw data and their corresponding description from the official website for the World Happiness Report (World Happiness Report 2017, 2017). The World Happiness Report 2015 ranked 143 countries by happiness score following with 8 factors. Happiness score was measured by the national average response to the question of life evaluations "Imagine a 10-step ladder. Top of ladder represents the best possible life for you, and the bottom of the ladder represents the worst. Which step of the ladder you personally feel you stand at this time?" (Helliwell, Huang, and Wang, 2017) Happiness score was also referenced as life ladder in a range of 0 to 10. GDP per capita was measured in log scale. Health (Life Expectancy) was the age. Social support, Freedom, and Government Trust were the national average response to different questions. Social support asked about "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?" Freedom was about "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?" Government Trust was the question "Is corruption widespread throughout the government or not?" Generosity was the residual of regressing national average of response to "Have you donated money to a charity in the past month?" Each of Positive affect and Negative affect were defined as the average of three negative affect measures. "Did you experience the following feeling during a lot of the day? Happiness, laugh, and enjoyment were for positive affect. Sadness, worries, and anger were for the negative affect (Helliwell et al., 2017). From the summary (See Appendix Figure 2), all factors were in a reasonable range with a few not available values (NA).

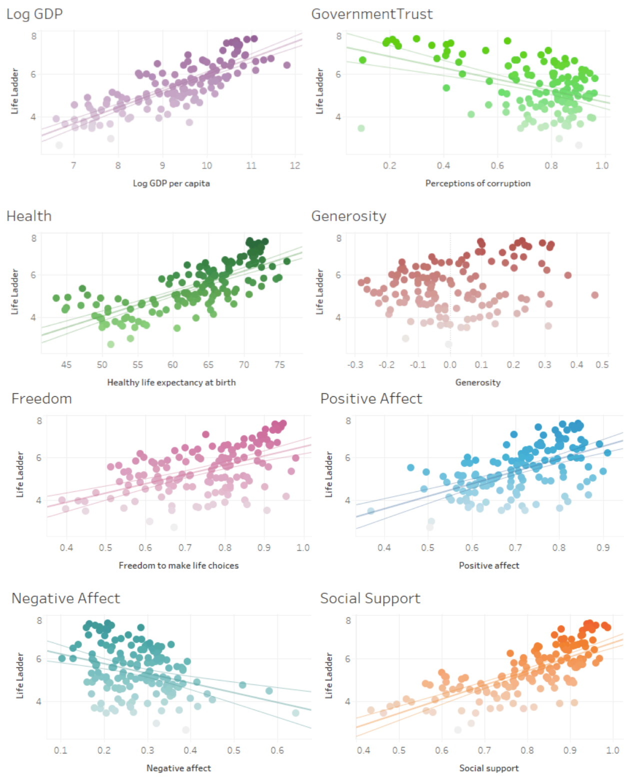


Figure A Scatter plot for Happiness Score vs. Predictors

After removing all the NA's, 124 out of 143 data were left. Since Happiness Score was our response variable, we decided to use the 8 factors as our predictor variables. First, we did histograms and boxplots for all predictors (See Appendix Figure 3). Government Trust had a few obvious outliers. Generosity and Negative Affect has one outlier. Second, we plot Happiness Score versus each predictor (See Figure A). The result showed that, Freedom, Positive Affect, Social Support, LogGDP, and Health showed a positive linear correlation. Negative Affect and Government Trust had a negative linear correlation. Generosity didn't show any obvious linear correlation. Based on the linear correlation, we decided to do a linear regression model.

First, we did the AIC test to obtain a model with the lowest AIC score. From the summary (See Appendix III AIC test), the AIC model was significant but with only three out of nine significant coefficients. Because we wanted a model with more significant coefficients, we decided to do another regression model based on correlation. The heat map (See Figure B)

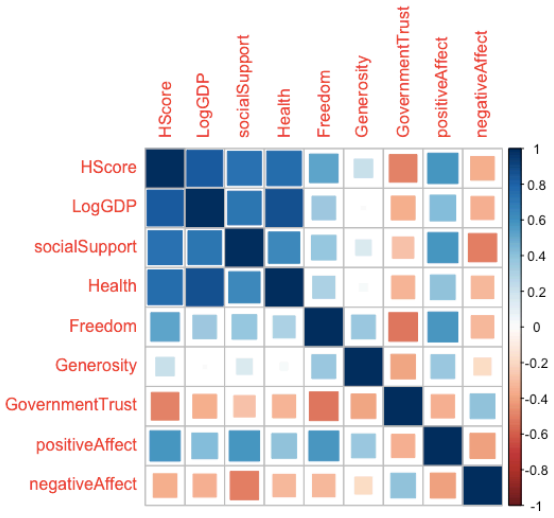


Figure B Correlation Heat Map between all factors

showed that correlations were not only between Happiness Score and predictors but also among predictor interactions, so we started from the additive linear model with all interaction terms. However, the model had similar performance as the AIC model. We started to drop not significant interaction terms and do ANOVA test to see if the dropping was valid. Eventually, we obtained a significant model with all coefficients significant (See Appendix IV Final Model). The final model was Happiness Score = LogGDP + Social Support + Health + Government Trust + Positive Affect + Negative Affect.

Results

We did six tests to diagnose our final model (See Appendix V Model Diagnosis). From the residual plot (See Appendix Figure 4), residuals were independent and normally distributed with equal variance. Shapiro test proved the residual normality. Both Breusch-Pagan test and Non-constant Variance Score Test concluded that residual had equal variance. Residual against predictor plots showed that all predictors fitted the null plot except Health (See Appendix Figure 5). Plots of data against model indicated model fitted the data very well (See Appendix Figure 6). Outlier test showed that our data had no significant outliers. From the qqPlot and influenceIndexPlot (See Appendix Figure 7 and 8), 14th and 96th data had the highest residuals and leverages, so they should have the largest influence among the data. The influencePlot proved our thought (See Appendix Figure 9). The 14th data was from Botswana, and the 96th data was from Rwanda. Overall, our final model passed all the assumptions and had very good fit to the data, so it was a good model.

In order to test model performance, we used the final model and 2015 data to predict the Happiness Score in 2016. Before prediction, we compared the Happiness Score between 2015 and 2016. In both world maps and box plots (See Figure C), there was no significance difference between two years, so we can use happiness score from 2015 to predict happiness score in 2016.

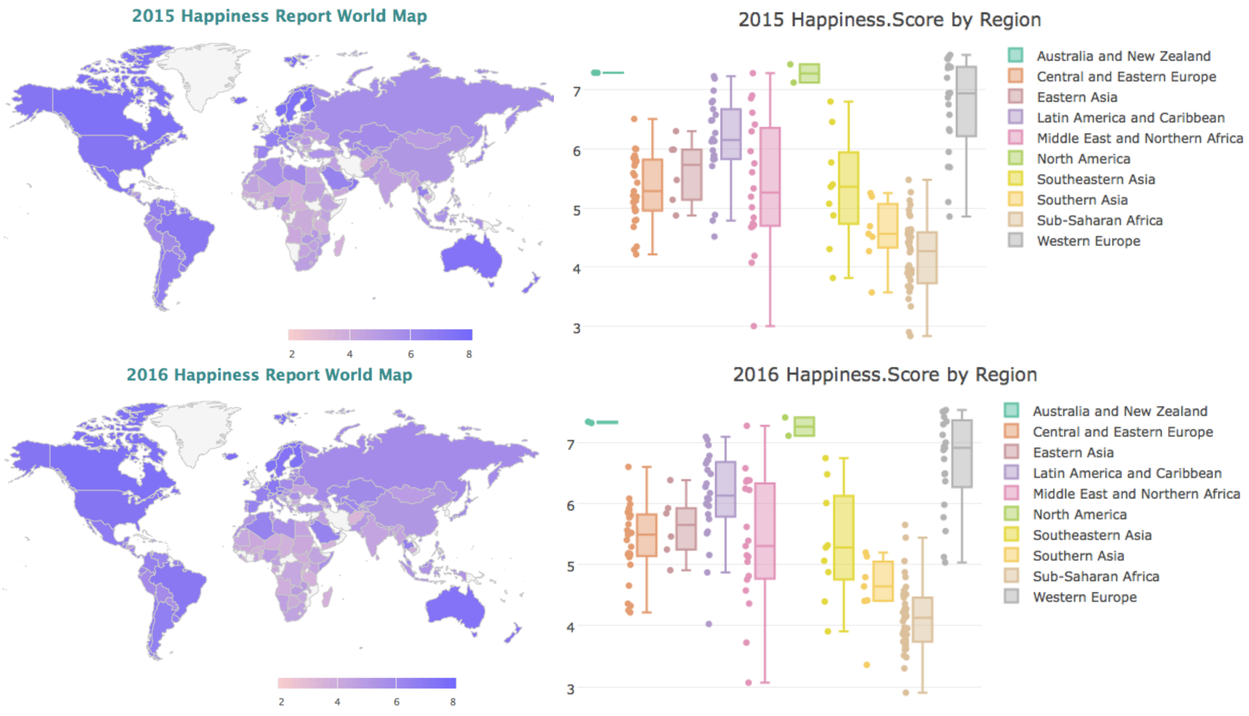
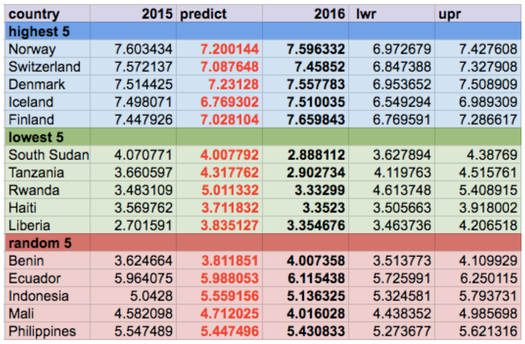


Figure C Happiness Score world map and box plot based on region

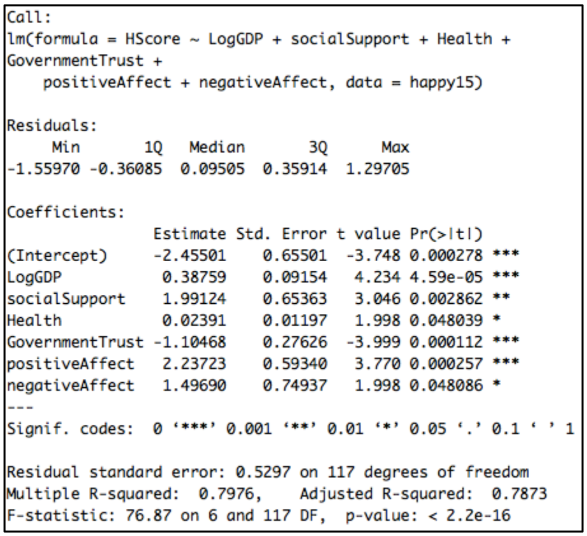
After prediction, we did summaries for the actual and predicted 2016 happiness score (See Appendix VI Linear Prediction). The actual minimum was 2.888, where the predicted minimum was 3.555. The actual maximum was 7.660, where the predicted maximum was 7.551. The prediction had a smaller range than the actual value. If we predict the lower ranking scores, we expect some higher prediction. If we predict the higher ranking scores, we expect some lower prediction. We observed fifteen predictions from



country	2015	predict	2016	lwr	upr
highest 5					
Norway	7.603434	7.200144	**7.596332**	6.972679	7.427608
Switzerland	7.572137	7.087648	**7.45852**	6.847388	7.327908
Denmark	7.514425	7.23128	**7.557783**	6.953652	7.508909
Iceland	7.498071	6.769302	**7.510035**	6.549294	6.989309
Finland	7.447926	7.028104	**7.659843**	6.769591	7.286617
lowest 5					
South Sudan	4.070771	4.007792	**2.888112**	3.627894	4.38769
Tanzania	3.660597	4.317762	**2.902734**	4.119763	4.515761
Rwanda	3.483109	5.011332	**3.33299**	4.613748	5.408915
Haiti	3.569762	3.711832	**3.3523**	3.505663	3.918002
Liberia	2.701591	3.835127	**3.354676**	3.463736	4.206518
random 5					
Benin	3.624664	3.811851	**4.007358**	3.513773	4.109929
Ecuador	5.964075	5.988053	**6.115438**	5.725991	6.250115
Indonesia	5.0428	5.559156	**5.136325**	5.324581	5.793731
Mali	4.582098	4.712025	**4.016028**	4.438352	4.985698
Philippines	5.547489	5.447496	**5.430833**	5.273677	5.621316

Figure D Actual and Predicted value for 15 countries

different rankings: the highest five, the lowest five, and the random five from the middle. Result matches our assumption (See Figure D). The scatter plot showed that predicted value and actual value are highly positive correlated (See Appendix Figure 11). Residuals of the predicted value had a small range and symmetric variance (See Appendix Figure 12). Mean absolute error (MAE), which takes the mean of the absolute value of the errors, was used to measure the model prediction performance. The MAE of our prediction was 0.4403506, which was about 10% of the prediction value range. Therefore, we consider the prediction had fairly good performance.



```
Call:
lm(formula = HScore ~ LogGDP + socialSupport + Health +
GovernmentTrust +
    positiveAffect + negativeAffect, data = happy15)

Residuals:
     Min       1Q   Median       3Q      Max
-1.55970 -0.36085  0.09505  0.35914  1.29705

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.45501    0.65501  -3.748 0.000278 ***
LogGDP           0.38759    0.09154   4.234 4.59e-05 ***
socialSupport    1.99124    0.65363   3.046 0.002862 **
Health           0.02391    0.01197   1.998 0.048039 *
GovernmentTrust -1.10468    0.27626  -3.999 0.000112 ***
positiveAffect   2.23723    0.59340   3.770 0.000257 ***
negativeAffect   1.49690    0.74937   1.998 0.048086 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5297 on 117 degrees of freedom
Multiple R-squared:  0.7976,    Adjusted R-squared:  0.7873
F-statistic: 76.87 on 6 and 117 DF,  p-value: < 2.2e-16
```

Figure E Summary of Final Regression Model

Recall the model summary, Positive Affect had the highest coefficient 2.23723, followed by Social Support, 1.99124, and Negative Affect 1.49690. It indicated that feeling supportive and emotional changes play an important role in happiness. Before starting the project, we guessed GPD would be the determinant factor on Happiness Score. Surprisingly, logGDP had a very small coefficient, 0.38759, which means economy is not so important for happiness. Age, referenced as Health, had the lowest coefficient, 0.02391, which is reasonable because logically happiness is not much affected by age. Government Trust had the only negative coefficient, -1.10468. If people think corruption is widespread throughout the government, certainly they will be less happy. To sum up, our final model fitted the data well and had good prediction performance. We conclude that Positive Affect, Negative Affect, and Social Support, predictors that involved emotion, are the determinant factor for Happiness Score.

Discussions

One interesting thing in our final model is that the Negative Affect had a negative coefficient, which means if you feel experience worry, sadness, and anger a lot of the day, you will feel more happy, which is illogical. It is possible that some of the dropped interaction terms are actually important for our model. However, we did ANOVA tests for the full and reduced model before we eliminated any terms. All the ANOVA tests indicated the dropping was valid. In our opinions, instead of analyzing predictors one by one, we should analyze the model as a whole and consider all predictors while interpreting the model.

Recall the scatter plots between Happiness Score and all predictors (See Figure A), Negative Affect had a negative linear correlation with Happiness Score. However, the linear correlation was mainly caused by the three data on the right bottom part. It is possible that Negative Affect have no linear relation with Happiness Score, like Generosity. One improvement will be doing transformation, such as quadratic or logarithmic, on Negative Affect and Generosity before adding them to the linear model.

Reference

Helliwell, J. F., Huang, H., & Wang, S. (2017, March 21). Statistical Appendix for "The social foundations of world happiness", Chapter 2, World Happiness Report 2017. Retrieved June 8, 2017, from http://worldhappiness.report/wp-content/uploads/sites/2/2017/03/StatisticalAppendixWHR2017.pdf

Helliwell,, J. F., Layard, R., & Sachs, J. D. (2017, March 21). WORLD HAPPINESS REPORT 2017, Chapter 1, Overview. Retrieved June 8, 2017, from http://worldhappiness.report/wp-content/uploads/sites/2/2017/03/HR17.pdf

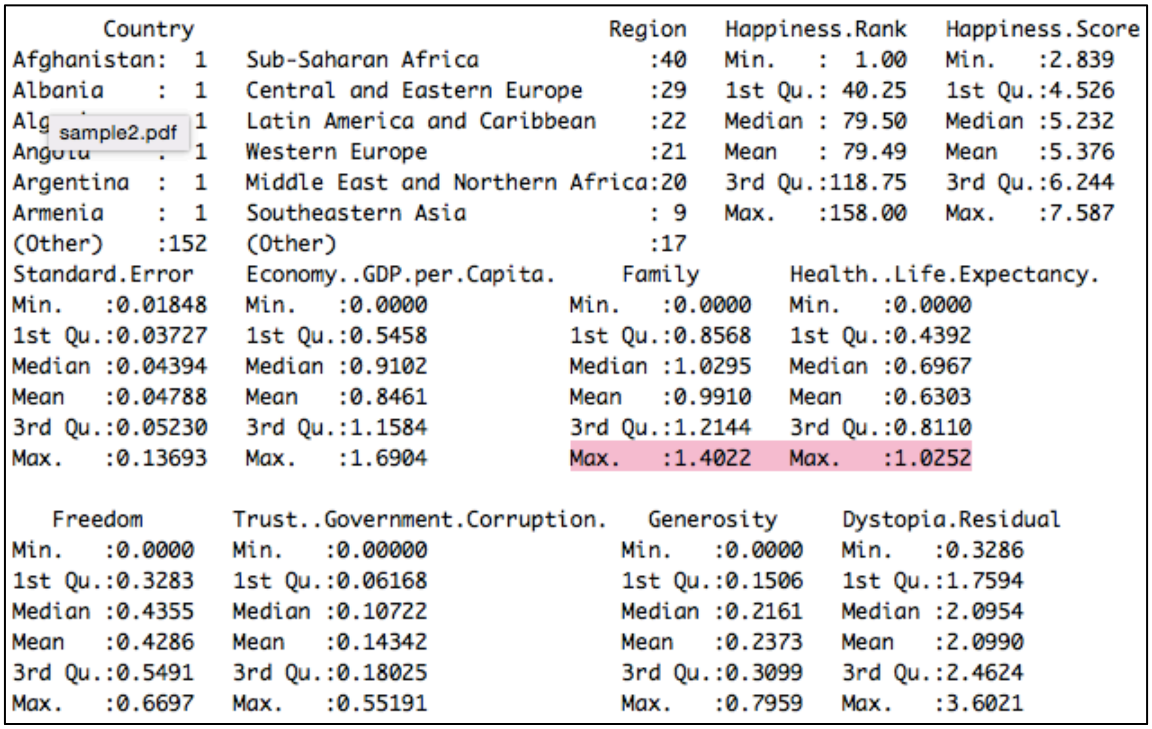
World Happiness Report 2016. (2017, March). Retrieved June 08, 2017, from https://www.kaggle.com/unsdsn/world-happiness

World Happiness Report 2017. (2017, March 21). Retrieved June 08, 2017, from http://worldhappiness.report/ed/2017/

Appendix

I. Data Collection and Preparation

```
raw <- read.csv("2015.csv")
summary(raw)
```



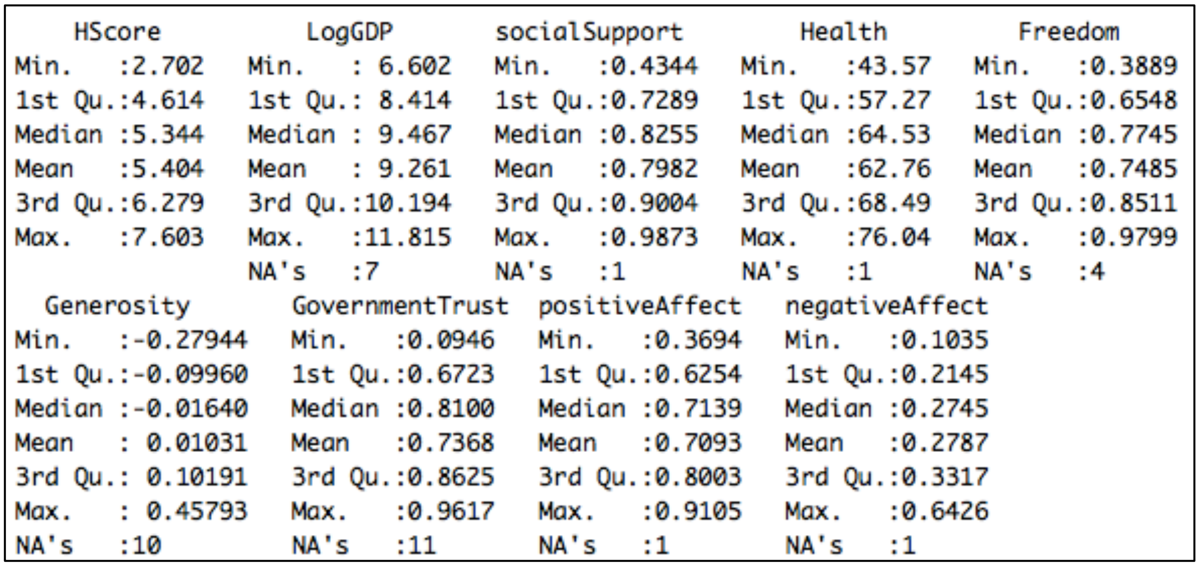
```
        Country                              Region      Happiness.Rank    Happiness.Score
 Afghanistan:  1   Sub-Saharan Africa             :40   Min.   :  1.00   Min.   :2.839
 Albania    :  1   Central and Eastern Europe     :29   1st Qu.: 40.25   1st Qu.:4.526
 Alg sample2.pdf 1   Latin America and Caribbean    :22   Median : 79.50   Median :5.232
 Angola     :  1   Western Europe                 :21   Mean   : 79.49   Mean   :5.376
 Argentina  :  1   Middle East and Northern Africa:20   3rd Qu.:118.75   3rd Qu.:6.244
 Armenia    :  1   Southeastern Asia              : 9   Max.   :158.00   Max.   :7.587
 (Other)    :152   (Other)                        :17
 Standard.Error    Economy..GDP.per.Capita.     Family       Health..Life.Expectancy.
 Min.   :0.01848   Min.   :0.0000          Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.03727   1st Qu.:0.5458          1st Qu.:0.8568   1st Qu.:0.4392
 Median :0.04394   Median :0.9102          Median :1.0295   Median :0.6967
 Mean   :0.04788   Mean   :0.8461          Mean   :0.9910   Mean   :0.6303
 3rd Qu.:0.05230   3rd Qu.:1.1584          3rd Qu.:1.2144   3rd Qu.:0.8110
 Max.   :0.13693   Max.   :1.6904          Max.   :1.4022   Max.   :1.0252

    Freedom        Trust..Government.Corruption.   Generosity     Dystopia.Residual
 Min.   :0.0000   Min.   :0.00000                Min.   :0.0000   Min.   :0.3286
 1st Qu.:0.3283   1st Qu.:0.06168                1st Qu.:0.1506   1st Qu.:1.7594
 Median :0.4355   Median :0.10722                Median :0.2161   Median :2.0954
 Mean   :0.4286   Mean   :0.14342                Mean   :0.2373   Mean   :2.0990
 3rd Qu.:0.5491   3rd Qu.:0.18025                3rd Qu.:0.3099   3rd Qu.:2.4624
 Max.   :0.6697   Max.   :0.55191                Max.   :0.7959   Max.   :3.6021
```

Figure 1 Summary of Kaggle.com raw data

```
library(readxl)
happy15_original = read_excel("2015.xlsx")
#str(happy15_original)

happy15_ml = happy15_original[,c(1,5:13)]
colnames(happy15_ml) <- c("Region","HScore", "LogGDP", "socialSupport", "Heal
th", "Freedom", "Generosity","GovernmentTrust","positiveAffect","negativeAffe
ct")
happy15_ml = na.exclude(happy15_ml)
happy15 = happy15_ml[,-c(1)]
summary(happy15)
```



```
     HScore          LogGDP       socialSupport        Health         Freedom     
 Min.   :2.702   Min.   : 6.602   Min.   :0.4344   Min.   :43.57   Min.   :0.3889  
 1st Qu.:4.614   1st Qu.: 8.414   1st Qu.:0.7289   1st Qu.:57.27   1st Qu.:0.6548  
 Median :5.344   Median : 9.467   Median :0.8255   Median :64.53   Median :0.7745  
 Mean   :5.404   Mean   : 9.261   Mean   :0.7982   Mean   :62.76   Mean   :0.7485  
 3rd Qu.:6.279   3rd Qu.:10.194   3rd Qu.:0.9004   3rd Qu.:68.49   3rd Qu.:0.8511  
 Max.   :7.603   Max.   :11.815   Max.   :0.9873   Max.   :76.04   Max.   :0.9799  
                 NA's   :7        NA's   :1        NA's   :1       NA's   :4       
   Generosity        GovernmentTrust  positiveAffect   negativeAffect  
 Min.   :-0.27944   Min.   :0.0946   Min.   :0.3694   Min.   :0.1035  
 1st Qu.:-0.09960   1st Qu.:0.6723   1st Qu.:0.6254   1st Qu.:0.2145  
 Median :-0.01640   Median :0.8100   Median :0.7139   Median :0.2745  
 Mean   : 0.01031   Mean   :0.7368   Mean   :0.7093   Mean   :0.2787  
 3rd Qu.: 0.10191   3rd Qu.:0.8625   3rd Qu.:0.8003   3rd Qu.:0.3317  
 Max.   : 0.45793   Max.   :0.9617   Max.   :0.9105   Max.   :0.6426  
 NA's   :10         NA's   :11       NA's   :1        NA's   :1       
```

Figure 2 Summary of official raw data

II. Investigate for predictor variables

```
par(mfrow=c(2,2))
boxplot(happy15$LogGDP, xlab = "LogGDP")
hist(happy15$LogGDP, xlab = "LogGDP",main ="")
boxplot(happy15$socialSupport, xlab = "socialSupport")
hist(happy15$socialSupport, xlab = "socialSupport",main ="")

boxplot(happy15$Health, xlab = "Health")
hist(happy15$Health, xlab = "Health",main ="")
boxplot(happy15$Freedom, xlab = "Freedom")
hist(happy15$Freedom, xlab = "Freedom",main ="")

boxplot(happy15$GovernmentTrust, xlab = "GovernmentTrust")
hist(happy15$GovernmentTrust, xlab = "GovernmentTrust",main ="")
boxplot(happy15$Generosity, xlab = "Generosity")
hist(happy15$Generosity, xlab = "Generosity",main ="")

boxplot(happy15$positiveAffect, xlab = "positiveAffect")
hist(happy15$positiveAffect, xlab = "positiveAffect",main ="")
boxplot(happy15$negativeAffect,xlab="negativeAffect")
hist(happy15$negativeAffect,xlab = "negativeAffect",main ="")
```

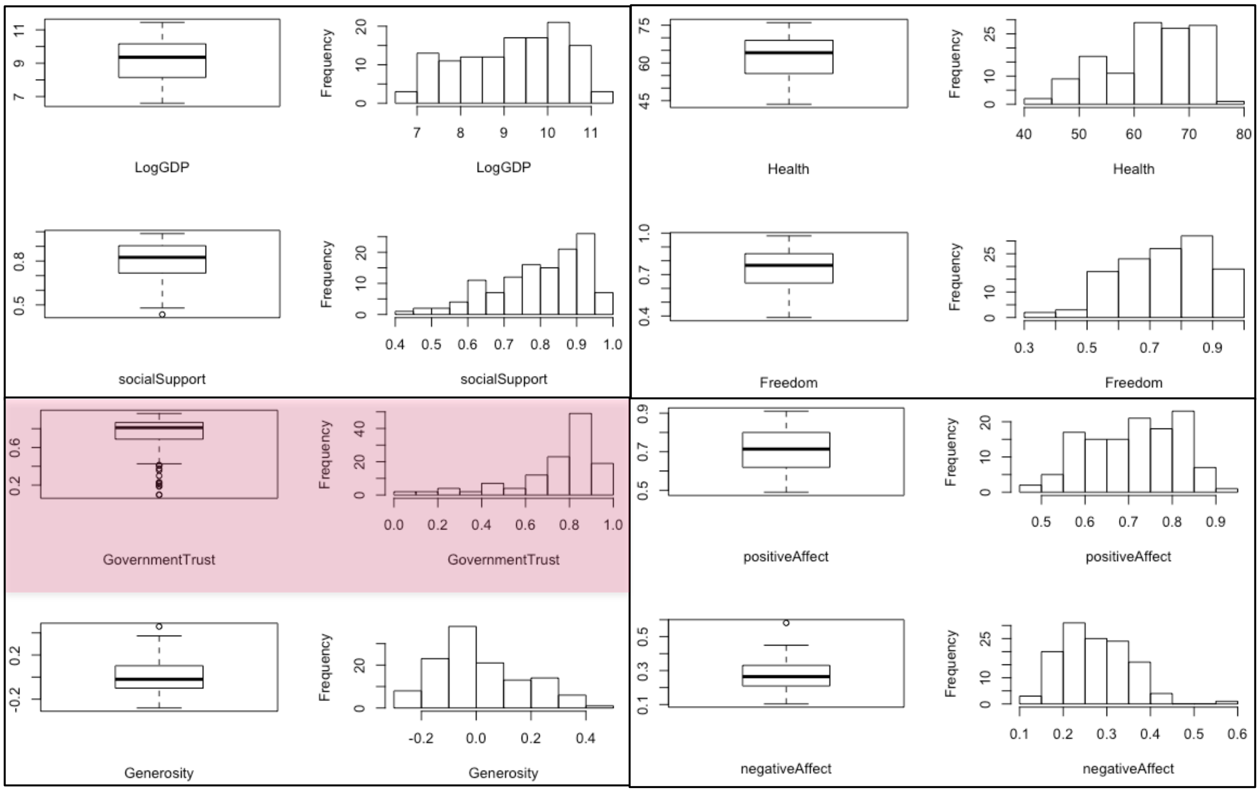


Figure 3 Histograms and boxplots between Happiness Score and all predictors

III. AIC test

```
# from 1 to all interactions
null <- lm(happy15$HScore~1,data = happy15)
full <- lm(happy15$HScore~.^2,data = happy15)
step(null,scope = list(lower = null,upper =full),direction = 'forward')
```

```
## Step:  AIC=-161.75
## happy15$HScore ~ LogGDP + positiveAffect + socialSupport + Freedom +
##     Health + LogGDP:positiveAffect + positiveAffect:Health +
##     socialSupport:Freedom
##
##                                Df Sum of Sq    RSS     AIC
## <none>                                      29.098 -161.75
## + socialSupport:Health          1  0.170539 28.927 -160.48
## + GovernmentTrust               1  0.151178 28.947 -160.40
## + negativeAffect                1  0.136559 28.961 -160.34
## + Health:Freedom                1  0.096428 29.001 -160.16
## + Generosity                    1  0.067653 29.030 -160.04
## + LogGDP:Freedom                1  0.029401 29.068 -159.88
## + LogGDP:Health                 1  0.029214 29.068 -159.88
## + Freedom:positiveAffect        1  0.024193 29.073 -159.86
## + socialSupport:positiveAffect  1  0.021775 29.076 -159.85
## + LogGDP:socialSupport          1  0.011135 29.087 -159.80
## Call:
## lm(formula = happy15$HScore ~ LogGDP + positiveAffect + socialSupport +
##     Freedom + Health + LogGDP:positiveAffect + positiveAffect:Health +
##     socialSupport:Freedom, data = happy15)
##
## Coefficients:
##           (Intercept)                 LogGDP         positiveAffect
##               10.4284                 0.6727               -14.6097
##         socialSupport                Freedom                 Health
##               -1.3937                -2.4232                -0.1935
## LogGDP:positiveAffect  positiveAffect:Health  socialSupport:Freedom
##               -0.4283                 0.3198                 4.4371
```

```
# lowest AIC model
happy15_lmAIC = lm(HScore ~ LogGDP + positiveAffect + socialSupport + Freedom
 + Health + LogGDP*positiveAffect + positiveAffect*Health + socialSupport*Fre
edom,data = happy15)
summary(happy15_lmAIC)
```

```
##
## Call:
## lm(formula = HScore ~ LogGDP + positiveAffect + socialSupport +
##     Freedom + Health + LogGDP * positiveAffect + positiveAffect *
##     Health + socialSupport * Freedom, data = happy15)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.53348 -0.27262  0.06489  0.32905  1.35479
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.42839    3.12462   3.337  0.00114 **
## LogGDP                  0.67274    0.62061   1.084  0.28064
## positiveAffect        -14.60968    4.86061  -3.006  0.00325 **
## socialSupport          -1.39373    2.10762  -0.661  0.50975
## Freedom                -2.42319    2.29986  -1.054  0.29426
## Health                 -0.19350    0.09901  -1.954  0.05308 .
## LogGDP:positiveAffect  -0.42835    0.88762  -0.483  0.63031
## positiveAffect:Health   0.31979    0.14257   2.243  0.02681 *
## socialSupport:Freedom   4.43709    2.87369   1.544  0.12533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.503 on 115 degrees of freedom
## Multiple R-squared:  0.8207, Adjusted R-squared:  0.8082
## F-statistic: 65.78 on 8 and 115 DF,  p-value: < 2.2e-16
```

IV. Final Model

```
happy15_lm5 = lm(HScore~LogGDP+socialSupport+Health+GovernmentTrust+positiveA
ffect+negativeAffect,data = happy15)
summary(happy15_lm5)
```

```
## 
## Call:
## lm(formula = HScore ~ LogGDP + socialSupport + Health + GovernmentTrust + 
##     positiveAffect + negativeAffect, data = happy15)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.55970 -0.36085  0.09505  0.35914  1.29705 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     -2.45501    0.65501  -3.748 0.000278 ***
## LogGDP           0.38759    0.09154   4.234 4.59e-05 ***
## socialSupport    1.99124    0.65363   3.046 0.002862 ** 
## Health           0.02391    0.01197   1.998 0.048039 *  
## GovernmentTrust -1.10468    0.27626  -3.999 0.000112 ***
## positiveAffect   2.23723    0.59340   3.770 0.000257 ***
## negativeAffect   1.49690    0.74937   1.998 0.048086 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5297 on 117 degrees of freedom
## Multiple R-squared:  0.7976, Adjusted R-squared:  0.7873 
## F-statistic: 76.87 on 6 and 117 DF,  p-value: < 2.2e-16
```

V. Model Diagnosis

```
# plot residuals
plot(happy15_lm)
```

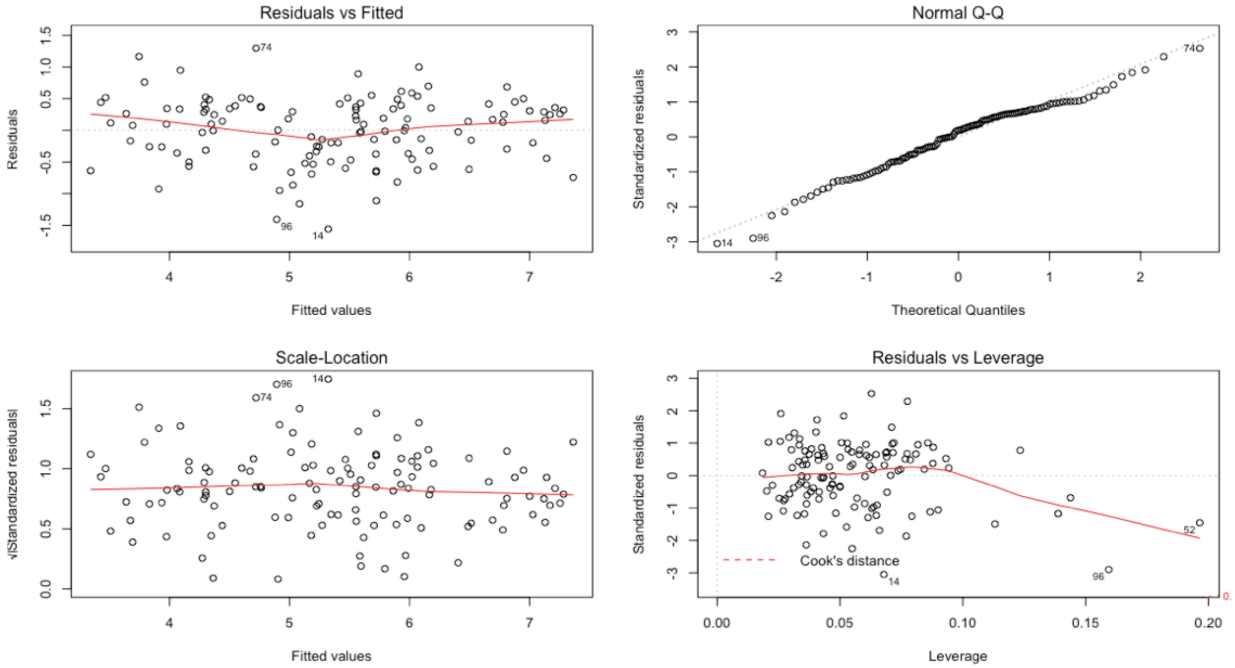


Figure 4 Residual plot from plot(lm)

```
# normality test for residual => REJECT normality of error variance
shapiro.test(happy15_lm$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  happy15_lm$residuals
## W = 0.98205, p-value = 0.09871

# Breush-Pagan test & non Consistent variance => REJECT means equal variance
library(lmtest)

bptest(happy15_lm)

##
##  studentized Breusch-Pagan test
##
## data:  happy15_lm
## BP = 3.7111, df = 6, p-value = 0.7157

ncvTest(happy15_lm)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.420122    Df = 1     p = 0.2333834
```

```
# residual plot
library(car)

residualPlots(happy15_lm)
```

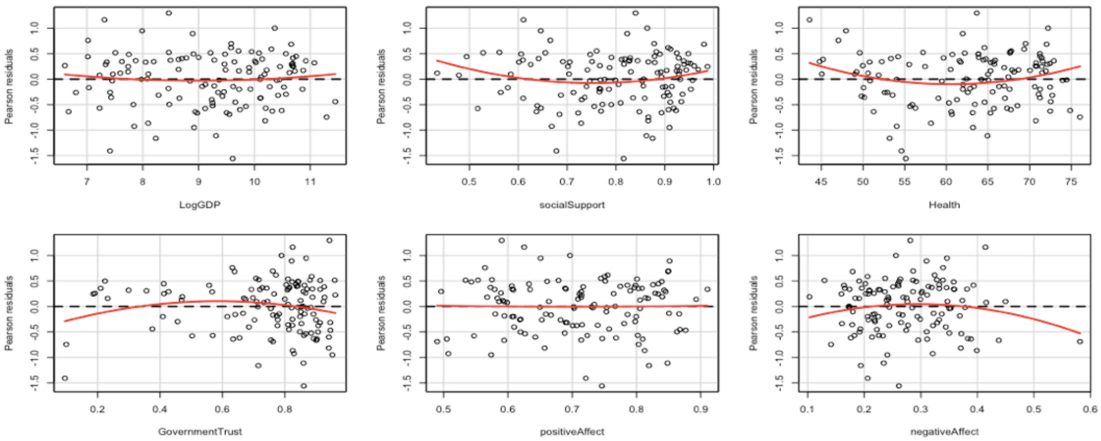


Figure 5 Residual vs. Predictor plots from residualPlots(lm)

```
##                  Test stat Pr(>|t|)
## LogGDP               0.743    0.459
## socialSupport        1.783    0.077
## Health               2.210    0.029
## GovernmentTrust     -1.616    0.109
## positiveAffect       0.098    0.922
## negativeAffect      -1.470    0.144
## Tukey test           2.618    0.009
```

```
# marginal model plots
marginalModelPlots(happy15_lm)
```

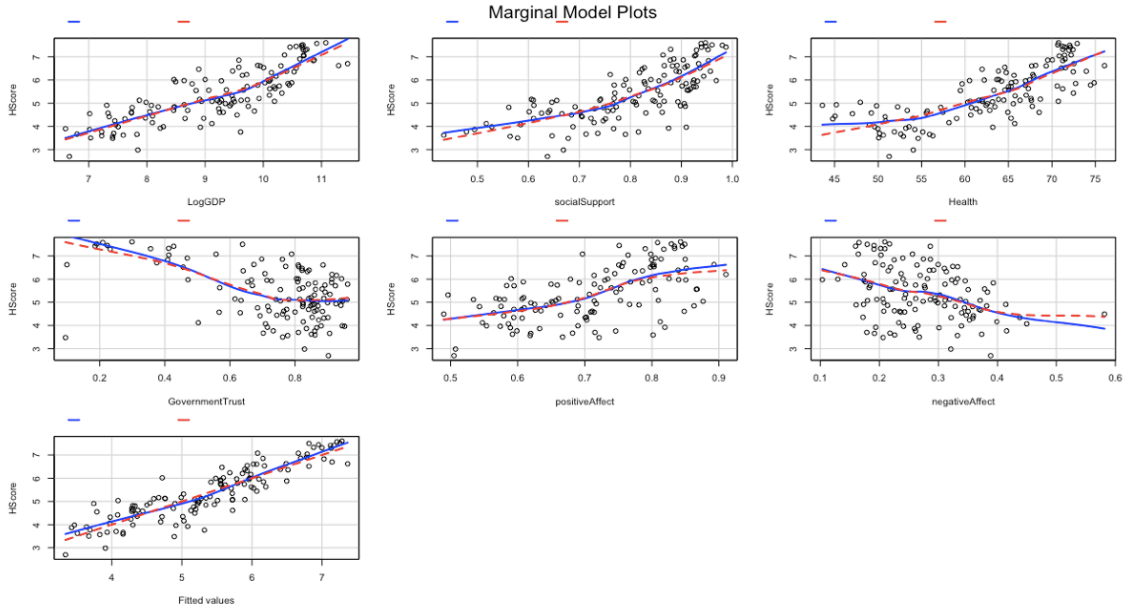


Figure 6 Data vs. Model Plots from marginalModelPlots(lm)

```
# qqPlot
qqPlot(happy15_lm, id.n = 2)
```

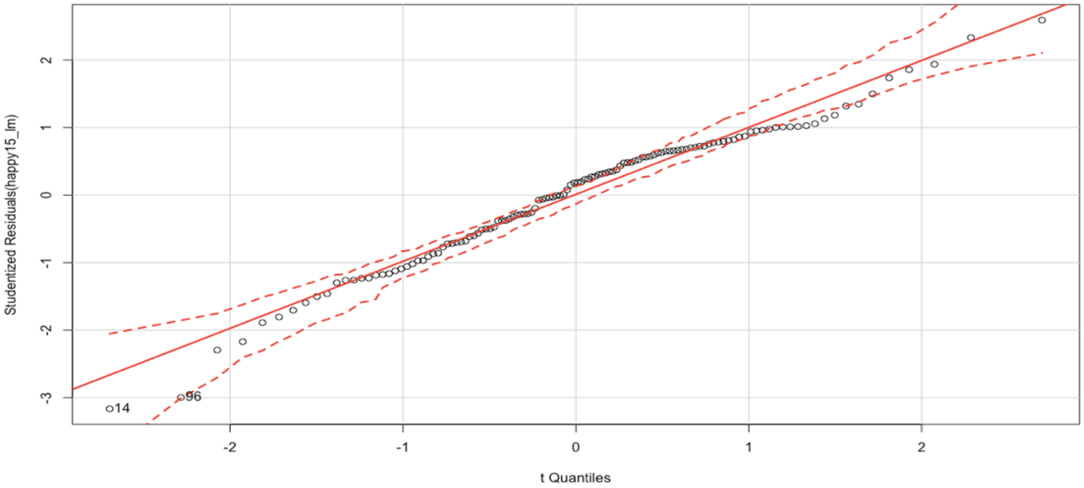


Figure 7 Highest Residual Plot from qqPlot(lm)

```
## 14 96
##  1  2

# outlierTest
outlierTest(happy15_lm)
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 14 -3.164906          0.0019816      0.24572

# infludenceIndexPlot
influenceIndexPlot(happy15_lm, id.n=2)
```

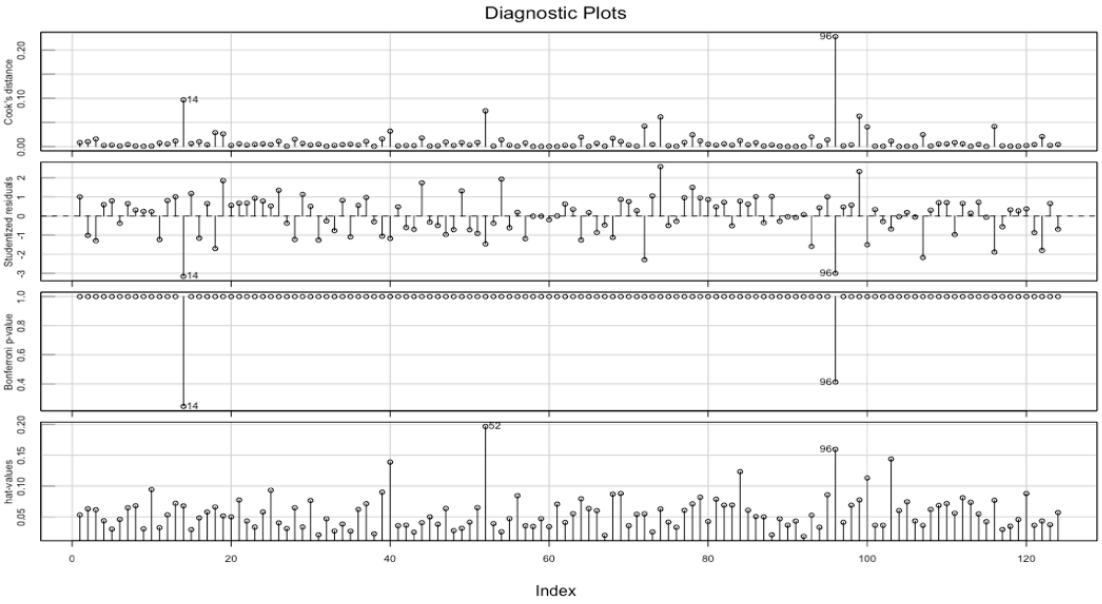


Figure 8 Highest Leverage Plots from influenceIndexPlot(lm)

```
# influencePlot
influencePlot(happy15_lm)
```

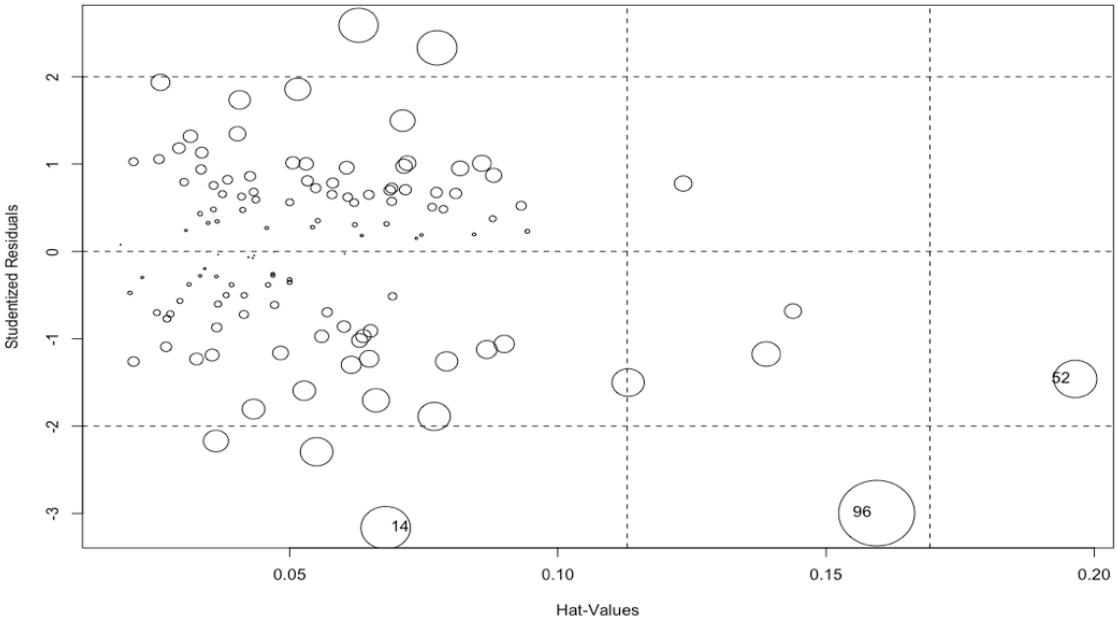


Figure 9 influencePlot(lm)

```
##       StudRes        Hat      CookD
## 14 -3.164906 0.06788099 0.09675164
## 52 -1.460571 0.19646950 0.07379954
## 96 -2.997443 0.15944620 0.22791958
```

VI. Linear Prediction

```
happy_pred_linear <- predict(happy15_lm,happy16_ml)
happy_pred_residual = happy_pred_linear-happy16_ml$HScore
summary(happy_pred_linear)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.555   4.726   5.439   5.434   6.059   7.551

summary(happy16_ml$HScore)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.888   4.472   5.386   5.385   6.118   7.660

summary(happy_pred_residual)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.18200 -0.32200  0.02940  0.04935  0.42280  1.73200

plot(happy_pred_linear, happy16_ml$HScore)
abline(0,1)

plot(happy_pred_residual)
abline(0,0)
```

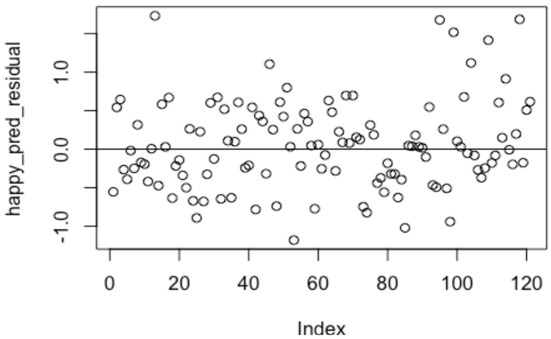


Figure 10 Prediction 2016 vs. Actual 2016

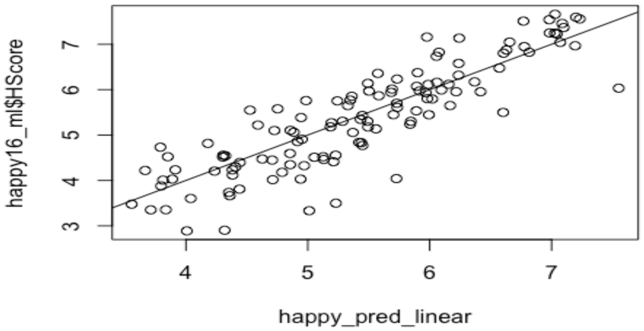


Figure 11 Residuals Plot from Prediction 2016

```
# Mean Absolute Error
MAE <- function(actual,predicted) {mean(abs(actual-predicted))}
MAE(happy16_ml$HScore, happy_pred_linear)

## [1] 0.4403506
```