Happy Analysis

Angela Lee

9/4/2018

- 0.1 Introduction
- 0.2 Subset the data for using only the key features mentioned the World Happiness Report.
- 0.3 Check the new dimension for H2018 - 426 x 9
- 0.4 Checking Missing Data
- 0.5 Happiness Score Distribution
- 0.6 Extract 6 key features and put them in the new data frame
- 0.7 Let's create a multiple linear regression model using the six key features.
- 0.8 Model Intepretation:
- 0.9 Variance Inflation Factor:
- 0.10 ANOVA - MLR Hypothesis - case 1
- 0.11 ANOVA - MLR Hypothesis - case 2

0.1 Introduction

The 2018 World Happiness Report is a survey that ranks 156 countries by their happiness level. It is published by the Sustainable Development Solutions Network(SDSN). The 2018 release of Gallup World Poll covers 1,562 observations of happiness scores from 2005 to 2017. Not all the countries and territories appear in all the years. For example, Canada is not included in the 2017 Happiness Index.

0.2 Subset the data for using only the key features mentioned the World Happiness Report.

They are GDP, Social Support, Healthy Life Expectancy, Freedom, Generosity, Perceptions of Corruption for the years between 2015 and 2016, and 2017. The new data frame is called H1018.

0.3 Check the new dimension for H2018 - 426 x 9

```
dim(H2018)
```

```
## [1] 426   9
```

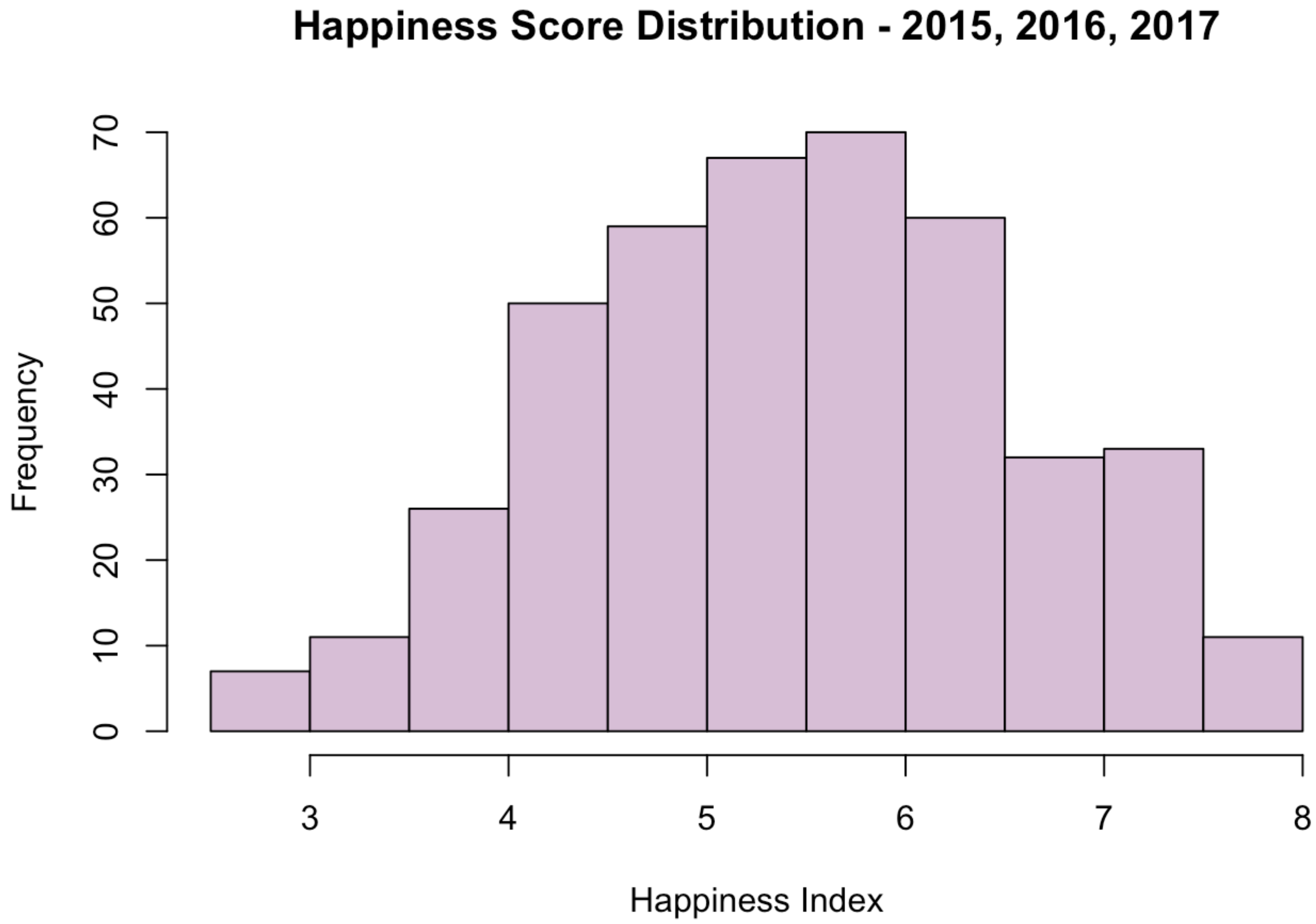
0.4 Checking Missing Data

Here are 74 missing records in the new data frame - H2018. Fill missing data with 0's.

```
## [1] 74
```

```
## [1] 0
```

0.5 Happiness Score Distribution



The standard deviation is 1.12

```
## [1] 1.123839
```

0.6 Extract 6 key features and put them in the new data frame

```
happy_df <- H2018[, 3:9]
#sum(is.na(happy_df))
```

0.7 Let’s create a multiple linear regression model using the six key features.

We look at the coefficients in the t-test. The MLR function for the six key features is the following:

Life Ladder (Happiness) = -0.54 + 0.03 GDP + 2.99 Social Support + 0.04 Healthy Life Expectancy + 1.25 Freedom - 0.32 Generosity - 0.70 Perceptions of Corruption

The coefficient of multiple determination, R squared (R^2) lies between 0 and 1 and is a descriptive measure of the utility of the regression for making predictions. Since the (R^2) always increases as we increase a variable x in the MLR model, we should look at the adjusted (R^2). The (R^2) only increases if the new varialble enhances the model beyond what would be obtained by chance and decreases when a predictor variable enhances the model less than what is expected by chance.

The adjusted for R squared is 0.6325 (R^2= 0.6325), which indicates the regression equation is SOMEWHAT useful for making predictions.

```
df <- lm(`Life Ladder` ~., data = happy_df)
summary(df)
```

```
## 
## Call:
## lm(formula = `Life Ladder` ~ ., data = happy_df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.9437 -0.4354  0.0138  0.4279  3.5141 
## 
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -0.543082   0.307539  -1.766   0.0781
## `Log GDP per capita`                0.028177   0.021439   1.314   0.1895
## `Social support`                    2.990840   0.314822   9.500  < 2e-16
## `Healthy life expectancy at birth`  0.045516   0.005238   8.690  < 2e-16
## `Freedom to make life choices`      1.247626   0.241297   5.170 3.62e-07
## Generosity                          0.320778   0.222776   1.440   0.1506
## `Perceptions of corruption`        -0.702811   0.129514  -5.427 9.74e-08
##                                       
## (Intercept)                        .  
## `Log GDP per capita`                  
## `Social support`                   ***
## `Healthy life expectancy at birth` ***
## `Freedom to make life choices`     ***
## Generosity                            
## `Perceptions of corruption`        ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6813 on 419 degrees of freedom
## Multiple R-squared:  0.6377, Adjusted R-squared:  0.6325 
## F-statistic: 122.9 on 6 and 419 DF,  p-value: < 2.2e-16
```

0.8 Model Intepretation:

Our best estimate for the common standard deviation of all life ladder (happiness index) for all countries at any particular GDP, Social Support, Healthy Life Expectancy, Freedom, Generosity, and corruption is 0.6813.

0.9 Variance Inflation Factor:

Let’s check if any of the six key variables have multicollinearity. VIF measures and indicates how much variance of an estimated regression coefficient is increased because of collinearity. If any of these variables has high VIF, then it is a highly correlated predictor variable. By looking at the variance inflation factor below, none of these key variables has VIF higher than 5, thus they are not highly correlated predictor variables.

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.1
```

```
## 
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
## 
##     recode
```

```
warning = FALSE
vif(df)
```

```
##               `Log GDP per capita`                   `Social support` 
##                           1.739878                           1.661426 
## `Healthy life expectancy at birth`     `Freedom to make life choices` 
##                           1.941618                           1.336884 
##                         Generosity        `Perceptions of corruption` 
##                           1.124961                           1.099767
```

0.10 ANOVA - MLR Hypothesis - case 1

We test if all slope parameters are 0.

The following is the hypothesis for the mutilple linear regression model for six features.

$$H_0 : \beta_1 GDP + \beta_2 SocialSupport + \beta_3 Health + \beta_4 Freedom + \beta_5 Generosity + \beta_6 PerceptionCorruption = 0$$

$$H_a : At\ Least\ One\ Of\ The\ Variables\ \beta_j \neq 0$$

###Full Model:

$$Life\ Ladder\ (Happiness\ Index) = \beta_0 + \beta_1 GDP + \beta_2 SocialSupport + \beta_3 Health + \beta_4 Freedom + \beta_5 Generosity + \beta_6 Perception\ Corruption$$

###Reduced Model:

$$Life\ Ladder\ (Happiness\ Index) = \beta_0$$

## Hypothesis Intepretation: By looking at the F statistics and the p value ($p < 2.2e - 16$), we can interpret that at the 5% significance level, the data provide sufficient evidence to conclude that at least one of the life ladder (happiness index) regression coefficients is not 0. We reject the null hypothesis. Therefore, taken together, GDP, social support, healthly life expectancy at birth, freedom, generosity, and perceptions of corruption are useful in predicting happiness for a country.

0.11 ANOVA - MLR Hypothesis - case 2

In this second scenario, based on the p value for the t test above, it looks like GDP and Generosity are not statistically significant (p > 0.05). In other words, is happiness significantly related to GDP, Soical Support, and Generosity after taking into account other factors such as Health, Freedom, and Perceptions of Corruption? Let's make another hypothesis for the second case.

$$H_0 : \beta_1 GDP = \beta_5 Generosity = 0$$

$$H_a : At\ Least\ One\ Of\ The\ Variables\ \beta_1, \beta_5 \neq 0$$

0.11.1 Full Model:

$$Life\ Ladder\ (Happiness\ Index) = \beta_0 + \beta_1 GDP + \beta_2 SocialSupport + \beta_3 Health + \beta_4 Freedom + \beta_5 Generosity + \beta_6 Perception\ Corruption$$

0.11.2 Reduced Model:

$$Life\ Ladder\ (Happiness\ Index) = \beta_0 + \beta_2 SocialSupport + \beta_3 Health + \beta_4 Freedom + \beta_6 Perception\ Corruption$$

## ANOVA for the partial model interpretation:

By looking at the p value for the f test, p value = 2.2e-16 (p < 0.05), we reject the null hypothesis. At the 5% significance level, the data provides sufficient edvidence to reject the null $beta_1 = beta_5 = 0$. Hence, in conjunction with other variables, GDP and Generosity are useful predictors of the happiness.

```
partial_H <- lm(`Life Ladder` ~ `Healthy life expectancy at birth` + `Freedom to make life choices` + `Perception
s of corruption` , data = happy_df)

anova(partial_H, df)
```

```
## Analysis of Variance Table
## 
## Model 1: `Life Ladder` ~ `Healthy life expectancy at birth` + `Freedom to make life choices` + 
##     `Perceptions of corruption`
## Model 2: `Life Ladder` ~ `Log GDP per capita` + `Social support` + `Healthy life expectancy at birth` + 
##     `Freedom to make life choices` + Generosity + `Perceptions of corruption`
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
## 1    422 240.03                                  
## 2    419 194.49  3    45.547 32.709 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```