# STAT 510 World Happiness Report Analysis

Ted Cheng
Basanth Shankar
Rene Zamudio

## Introduction

Dataset Title: World Happiness Report (2019)

Description:
The dataset, World Happiness Report 2019, is a landmark survey of the state of global happiness and was acquired from Kaggle. Dating back to 2012, this survey has been conducted numerous times and was presented to the United Nations during a celebration of International Day of Happiness. As the popularity of the report grew, so did the number of ranked countries, which now total to 156. The happiness scores and ranking use data from the Gallup World Poll, with its goal to understand the thoughts, feelings, and behaviors of people around the world. Ranging from 0-10, the scores measure the rankings of the countries based on the similarly scored responses, with 0 being the “worst” possible life and 10 being the “best” possible Our response will be score and  predictors will be GDP, Social Support, Healthy Life Expectancy, Freedom to Make Life Choices, Generosity, and Perception of Corruption. Predictors are measured by the extent to which they contribute to the calculation of the Happiness Score. This is a more complicated calculation called the Dystopia Residual Metric and the sum of these factors make up the happiness score.

Research Goal:
What makes people happy? Of the six predictors based on the dystopia happiness score, which ones are the most significant? This information may be helpful in determining what factors country's should focus resources on to lead to a happier population.

## Questions of Interest

Question of Interest 1:
In order for us to determine which of the six predictors are the most significant, we will first have to visualize the data. Creating a scatterplot will allow us to have a first glimpse at any potential interactions amongst the predictors, and probable relationships between predictors and response. In addition, we are also able to see the exact correlation coefficients and shape of the relationships. Afterwards, we will continue testing (F-test, AIC, and BIC) to find our first potential Linear Model for further evaluation.

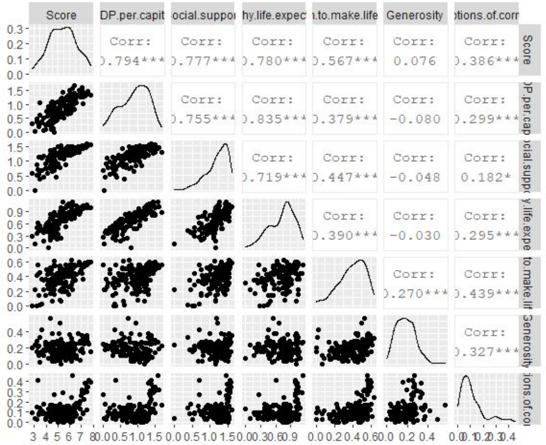
Question of Interest 2:
The final model will allow us to determine which variables are confounding. We will be able to see if there are any interaction terms that prove to be essential for the study and which non-interaction terms may be removed to make the model more accurate. Aside from the model, we can use the information to ultimately understand which predictors outweigh the rest and how/why these predictors could be different in each country or region of the world.

## Regression Analysis

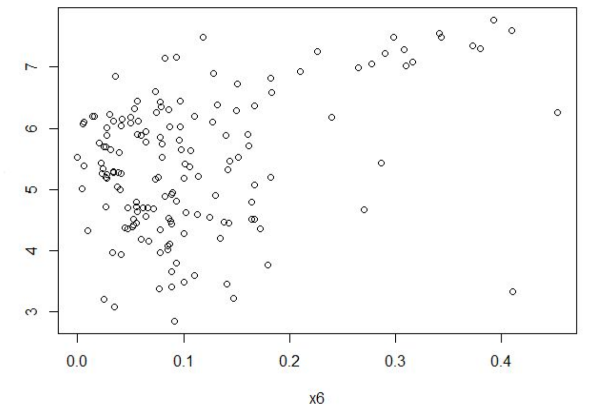
Below (on the left) is the visual for the scatterplot and starting point for finding potential relationships and interactions. We can see that certain predictors and fewer interactions have strong Correlation Coefficients. Specifically, we can see that x1, x2, x3, x4, x5 are potential appropriate predictors; we can also see that interactions x1:x2, x1:x3, and x2:x3 are potential essential interactions. At first, we thought that x6 (Perception of Corruption) had a quadratic relationship/interactions, but that was not the case.

Upon further inspection of Perceptions of Corruption, we can also see below an example of what the individual scatter plot looks like when it is zoomed in. The points appear to be more linear rather than our initial quadratic assumption.

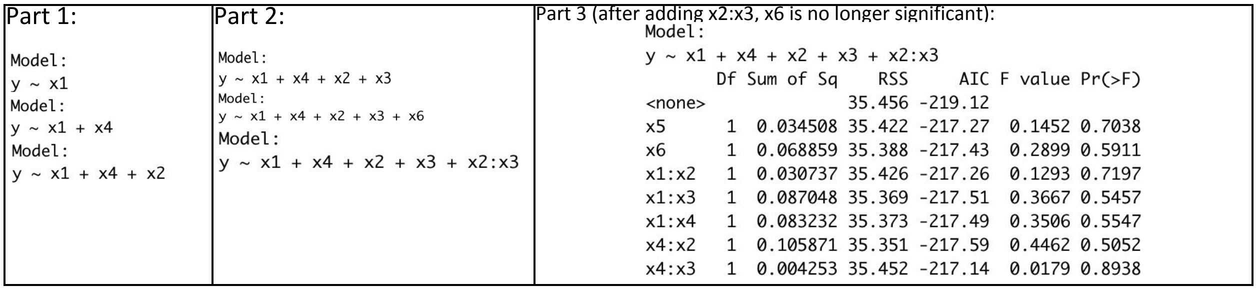
Scatter Plot



plot(x6, x4)



We now perform other assessments, starting with a Stepwise Regression for the F-test (below).



Part 1:

```
Model:
y ~ x1
Model:
y ~ x1 + x4
Model:
y ~ x1 + x4 + x2
```

Part 2:

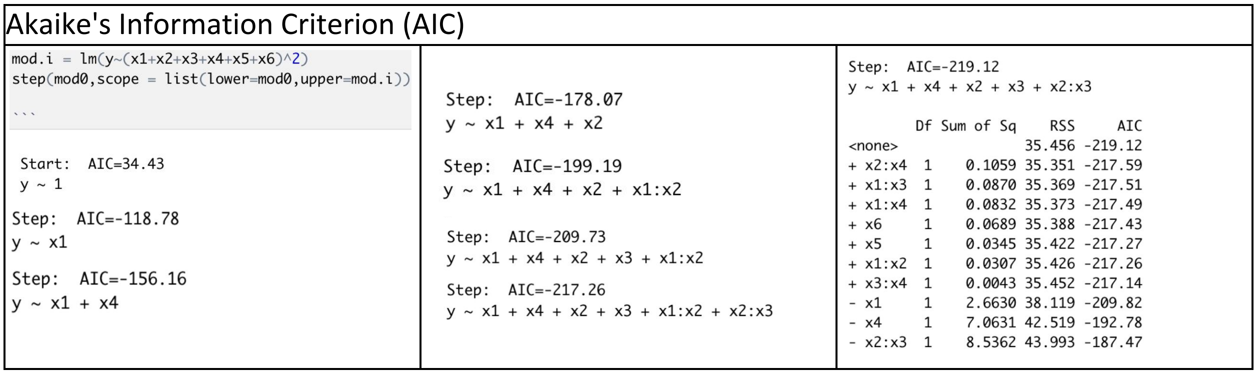
```
Model:
y ~ x1 + x4 + x2 + x3
Model:
y ~ x1 + x4 + x2 + x3 + x6
Model:
y ~ x1 + x4 + x2 + x3 + x2:x3
```

Part 3 (after adding x2:x3, x6 is no longer significant):

```
Model:
y ~ x1 + x4 + x2 + x3 + x2:x3
       Df Sum of Sq    RSS     AIC F value Pr(>F)
<none>              35.456 -219.12
x5      1  0.034508 35.422 -217.27  0.1452 0.7038
x6      1  0.068859 35.388 -217.43  0.2899 0.5911
x1:x2   1  0.030737 35.426 -217.26  0.1293 0.7197
x1:x3   1  0.087048 35.369 -217.51  0.3667 0.5457
x1:x4   1  0.083232 35.373 -217.49  0.3506 0.5547
x4:x2   1  0.105871 35.351 -217.59  0.4462 0.5052
x4:x3   1  0.004253 35.452 -217.14  0.0179 0.8938
```

The concluded “mod8” linear model is `y ~ x1 + x4 + x2 + x3 + x2 * x3`. Next is the Akaike's Information Criterion (AIC) test shown below (ordered left to right). We will run an estimation where all possible models and interactions will be compared. The test will end when the best quality model is found. Best case scenario: the conclusion of the AIC test supports the F-test.

Akaike's Information Criterion (AIC)



```
mod.i = lm(y~(x1+x2+x3+x4+x5+x6)^2)
step(mod0,scope = list(lower=mod0,upper=mod.i))

```

 Start:  AIC=34.43
 y ~ 1

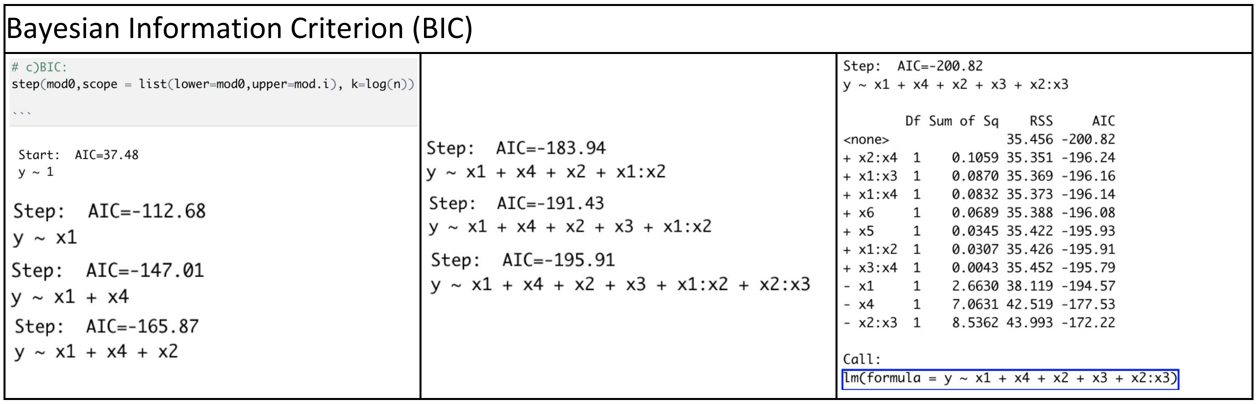
Step:  AIC=-118.78
y ~ x1

Step:  AIC=-156.16
y ~ x1 + x4
```

```
Step:  AIC=-178.07
y ~ x1 + x4 + x2

Step:  AIC=-199.19
y ~ x1 + x4 + x2 + x1:x2

Step:  AIC=-209.73
y ~ x1 + x4 + x2 + x3 + x1:x2

Step:  AIC=-217.26
y ~ x1 + x4 + x2 + x3 + x1:x2 + x2:x3
```

```
Step:  AIC=-219.12
y ~ x1 + x4 + x2 + x3 + x2:x3

        Df Sum of Sq    RSS     AIC
<none>                35.456 -219.12
+ x2:x4  1    0.1059 35.351 -217.59
+ x1:x3  1    0.0870 35.369 -217.51
+ x1:x4  1    0.0832 35.373 -217.49
+ x6     1    0.0689 35.388 -217.43
+ x5     1    0.0345 35.422 -217.27
+ x1:x2  1    0.0307 35.426 -217.26
+ x3:x4  1    0.0043 35.452 -217.14
- x1     1    2.6630 38.119 -209.82
- x4     1    7.0631 42.519 -192.78
- x2:x3  1    8.5362 43.993 -187.47
```

We have concluded the test with the model "mod8": `y ~ x1 + x4 + x2 + x3 + x2:x3`. Next is the Bayesian Information Criterion (BIC) test shown below (ordered left to right). Similar to the previous step, we now perform another comparison of the conclusion obtained from the previous two tests. Best case scenario: the conclusion of the BIC supports, both,  AIC and F-test.

Bayesian Information Criterion (BIC)



```
# c)BIC:
step(mod0,scope = list(lower=mod0,upper=mod.i), k=log(n))

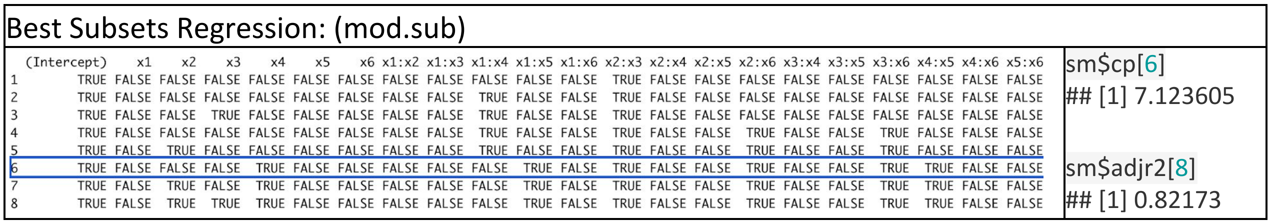
```

 Start:  AIC=37.48
 y ~ 1

Step:  AIC=-112.68
y ~ x1

Step:  AIC=-147.01
y ~ x1 + x4

Step:  AIC=-165.87
y ~ x1 + x4 + x2
```

```
Step:  AIC=-183.94
y ~ x1 + x4 + x2 + x1:x2

Step:  AIC=-191.43
y ~ x1 + x4 + x2 + x3 + x1:x2

Step:  AIC=-195.91
y ~ x1 + x4 + x2 + x3 + x1:x2 + x2:x3
```

```
Step:  AIC=-200.82
y ~ x1 + x4 + x2 + x3 + x2:x3

        Df Sum of Sq    RSS     AIC
<none>                35.456 -200.82
+ x2:x4  1    0.1059 35.351 -196.24
+ x1:x3  1    0.0870 35.369 -196.16
+ x1:x4  1    0.0832 35.373 -196.14
+ x6     1    0.0689 35.388 -196.08
+ x5     1    0.0345 35.422 -195.93
+ x1:x2  1    0.0307 35.426 -195.91
+ x3:x4  1    0.0043 35.452 -195.79
- x1     1    2.6630 38.119 -194.57
- x4     1    7.0631 42.519 -177.53
- x2:x3  1    8.5362 43.993 -172.22

Call:
lm(formula = y ~ x1 + x4 + x2 + x3 + x2:x3)
```

Once again, we have concluded with the model "mod8": `y ~ x1 + x4 + x2 + x3 + x2:x3`. Following the tests, we compare all possible models that can be created based upon an identified set of predictors with the regsubsets function below:

Best Subsets Regression: (mod.sub)



```
  (Intercept)    x1    x2    x3    x4    x5    x6 x1:x2 x1:x3 x1:x4 x1:x5 x1:x6 x2:x3 x2:x4 x2:x5 x2:x6 x3:x4 x3:x5 x3:x6 x4:x5 x4:x6 x5:x6
1        TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
2        TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
3        TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
4        TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE
5        TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE
6        TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE
7        TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE
8        TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE
```

```
sm$cp[6]
## [1] 7.123605

sm$adjr2[8]
## [1] 0.82173
```

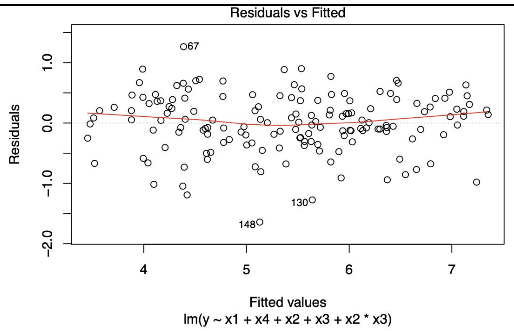
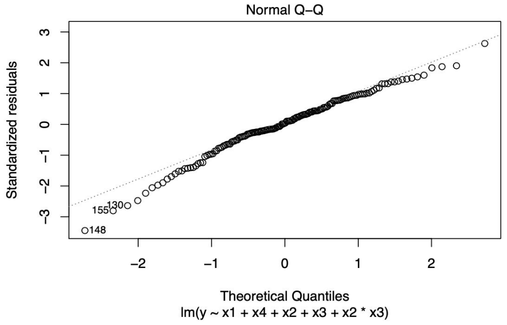
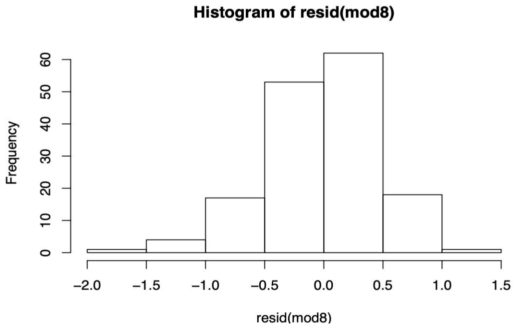
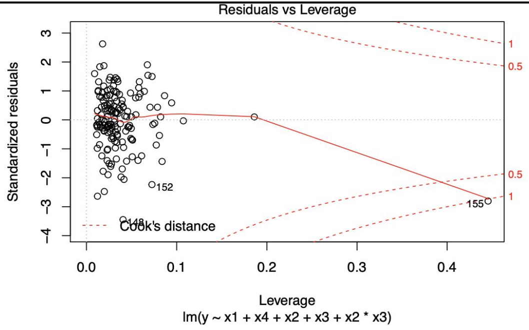
Performing the Best Subset Regression concludes with us choosing option 6. For our dataset of 6 predictors ( $p$ ), our optimal Mallow's CP is p, which is 7 (inclusive of the intercept term). We can see above that option 6 gave us a CP value closest to 7. When looking at the Adjusted $R^2$ , option 8 gives us the best value of approximately 0.822, while option 6 gives us a slightly lower value of approximately 0.816.

LINE: Our current models of interest are the following.

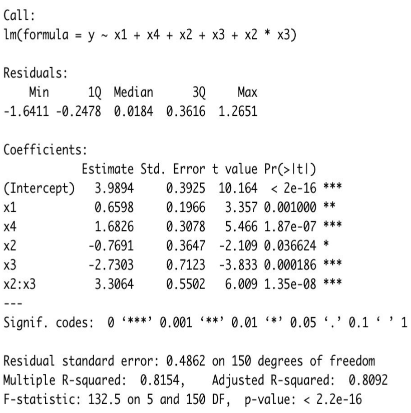
mod.sub: `y ~ x1 + x4 + x2 + x3 + x5 + x6 + x1:x5 + x2:x3 + x2:x6 + x3:x6 + x4:x5`

mod8: `y ~ x1 + x4 + x2 + x3 + x2:x3`.

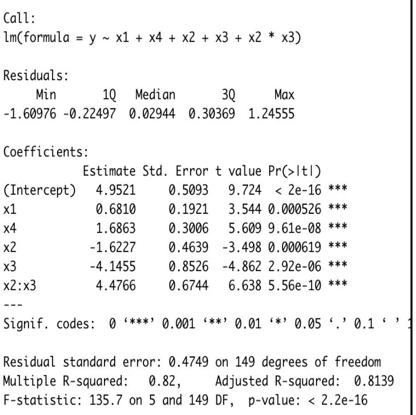
However, we will reject the subset model (mod.sub), since we want a parsimonious model and do not want it to be overspecified; it is the only model that differs from the other assessments.

| | |
|---|---|
| Notes on the Residual vs Fitted graph<br>● Well-behaved<br>● Consistent Variance<br>● Point 148 is a point of interest due to it being furthest from the red line. |  |
| Notes on the Normal Q-Q graph<br>● Fairly Linear, but slight signs of skewness is present.<br>● Skewness/non-linearity is not present enough to support the idea of outliers.<br>● Somewhat linearity supports normally distributed error terms. |  |
| Notes on the Histogram and Shapiro<br>● Noticeable left skewness.<br>● Shapiro-Wilk Normality Test: p=0.05482<br>● Shapiro test concludes with a p-value of 0.05482 > 0.05. This supports the idea of normal distribution amongst the error terms. |  |
| Notes on the Residual vs Leverage<br>● Observation 155 (Central African Republic) is the only point outside the dotted line, which means it potentially has high influence.<br>● Cook's Distance (155) = 1.0529 |  |

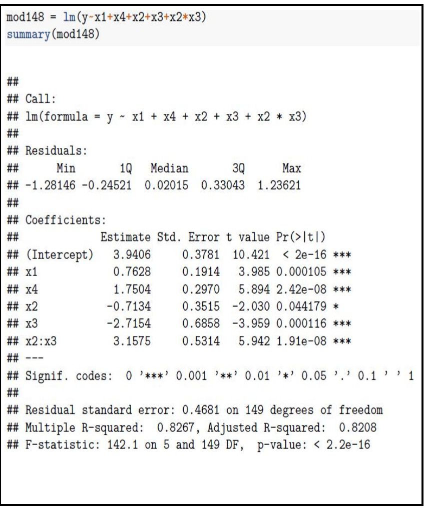
We compared the summary tables for 148 and 155:



```
Call:
lm(formula = y ~ x1 + x4 + x2 + x3 + x2 * x3)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6411 -0.2478  0.0184  0.3616  1.2651 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.9894     0.3925  10.164  < 2e-16 ***
x1            0.6598     0.1966   3.357 0.001000 ** 
x4            1.6826     0.3078   5.466 1.87e-07 ***
x2           -0.7691     0.3647  -2.109 0.036624 *  
x3           -2.7303     0.7123  -3.833 0.000186 ***
x2:x3         3.3064     0.5502   6.009 1.35e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4862 on 150 degrees of freedom
Multiple R-squared:  0.8154,	Adjusted R-squared:  0.8092 
F-statistic: 132.5 on 5 and 150 DF,  p-value: < 2.2e-16
```



```
Call:
lm(formula = y ~ x1 + x4 + x2 + x3 + x2 * x3)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.60976 -0.22497  0.02944  0.30369  1.24555 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.9521     0.5093   9.724  < 2e-16 ***
x1            0.6810     0.1921   3.544 0.000526 ***
x4            1.6863     0.3006   5.609 9.61e-08 ***
x2           -1.6227     0.4639  -3.498 0.000619 ***
x3           -4.1455     0.8526  -4.862 2.92e-06 ***
x2:x3         4.4766     0.6744   6.638 5.56e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4749 on 149 degrees of freedom
Multiple R-squared:   0.82,	Adjusted R-squared:  0.8139 
F-statistic: 135.7 on 5 and 149 DF,  p-value: < 2.2e-16
```

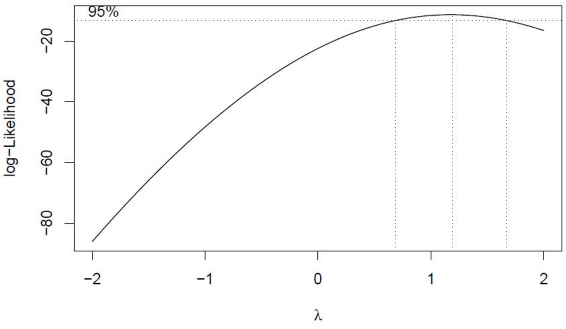


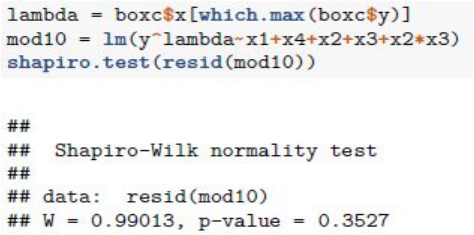
```
mod148 = lm(y~x1+x4+x2+x3+x2*x3)
summary(mod148)


## 
## Call:
## lm(formula = y ~ x1 + x4 + x2 + x3 + x2 * x3)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.28146 -0.24521  0.02015  0.33043  1.23621 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   3.9406     0.3781  10.421  < 2e-16 ***
## x1            0.7628     0.1914   3.985 0.000105 ***
## x4            1.7504     0.2970   5.894 2.42e-08 ***
## x2           -0.7134     0.3515  -2.030 0.044179 *  
## x3           -2.7154     0.6858  -3.959 0.000116 ***
## x2:x3         3.1575     0.5314   5.942 1.91e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4681 on 149 degrees of freedom
## Multiple R-squared:  0.8267, Adjusted R-squared:  0.8208 
## F-statistic: 142.1 on 5 and 149 DF,  p-value: < 2.2e-16
```

We can see that removing the influential point significantly affects x2 = Social support, x3 = Healthy life expectancy, and their interaction. Their slope and p-values are significantly affected especially. Because those factors are affected significantly, we should not remove the influential point. We also see that our shapiro was worse when we removed the influential point.

However Botswana (148) is not influential but an outlier according to studentized residual so we can remove it.

Due to error terms having potential  non-normality, we transform the model and find the optimal lambda power for y. One is contained in the interval, so it is not necessary to use Boxcox. We found our optimal lambda = 1.1919. Shapiro test concludes with a p-value of 0.3527 > 0.05. This supports the idea of normal distribution amongst the error terms. All other LINE assumptions still hold.





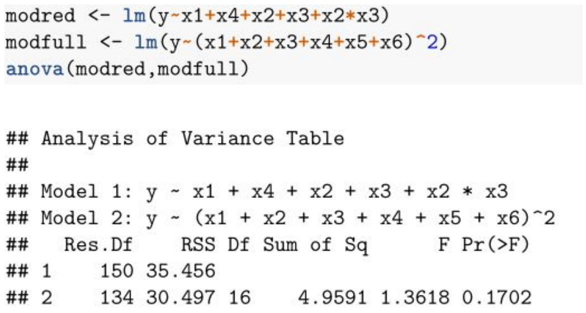
```
lambda = boxc$x[which.max(boxc$y)]
mod10 = lm(y^lambda~x1+x4+x2+x3+x2*x3)
shapiro.test(resid(mod10))
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  resid(mod10)
## W = 0.99013, p-value = 0.3527
```

Comparison ANOVA test between the reduced model containing only x1, x2, x3, x4, and x2:x3 and the full model gives us the following:

p-value = 0.1702

We fail to accept the full model over the reduced model



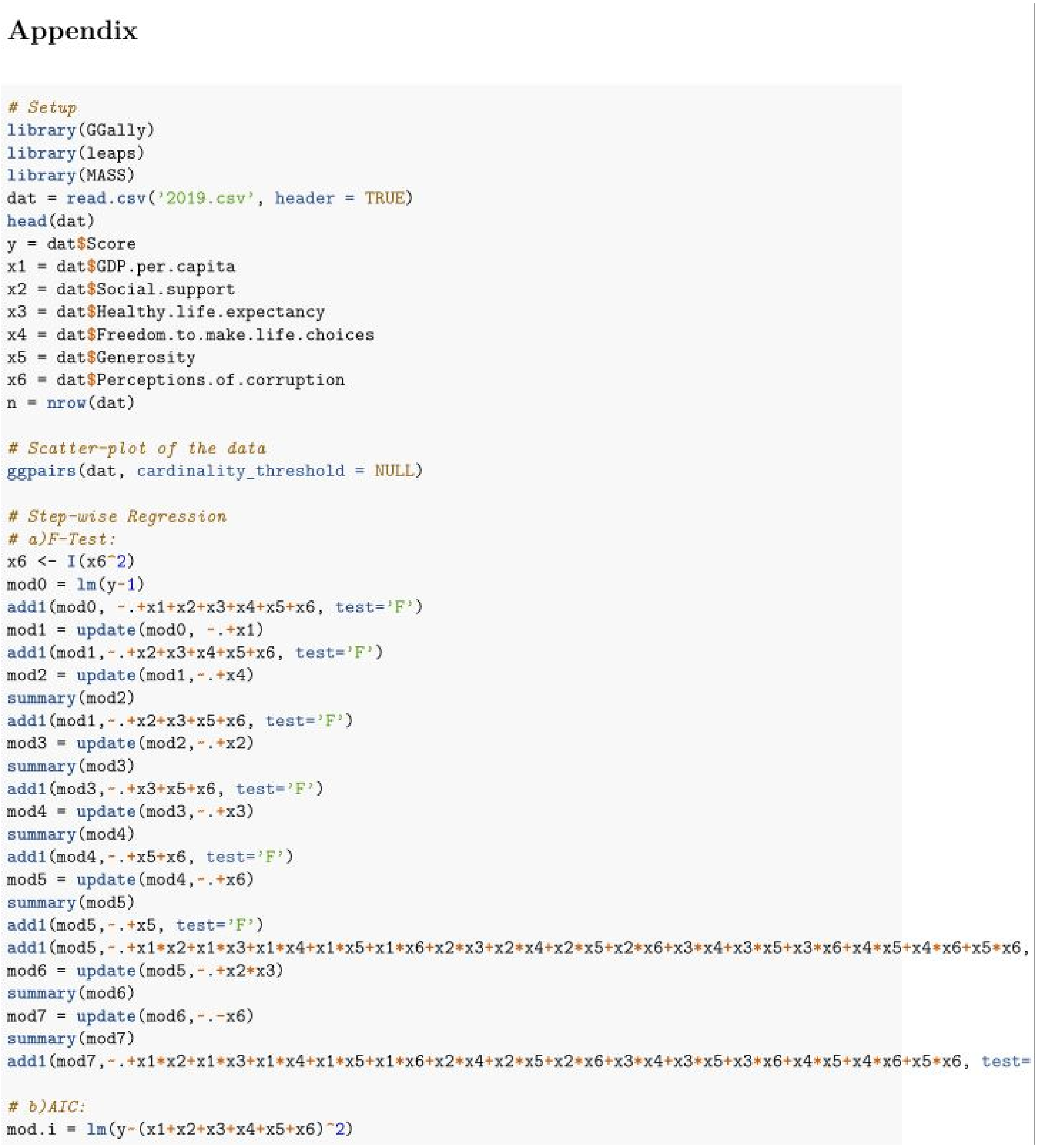
```
modred <- lm(y~x1+x4+x2+x3+x2*x3)
modfull <- lm(y~(x1+x2+x3+x4+x5+x6)^2)
anova(modred,modfull)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x4 + x2 + x3 + x2 * x3
## Model 2: y ~ (x1 + x2 + x3 + x4 + x5 + x6)^2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    150 35.456
## 2    134 30.497 16    4.9591 1.3618 0.1702
```

## **Conclusion**

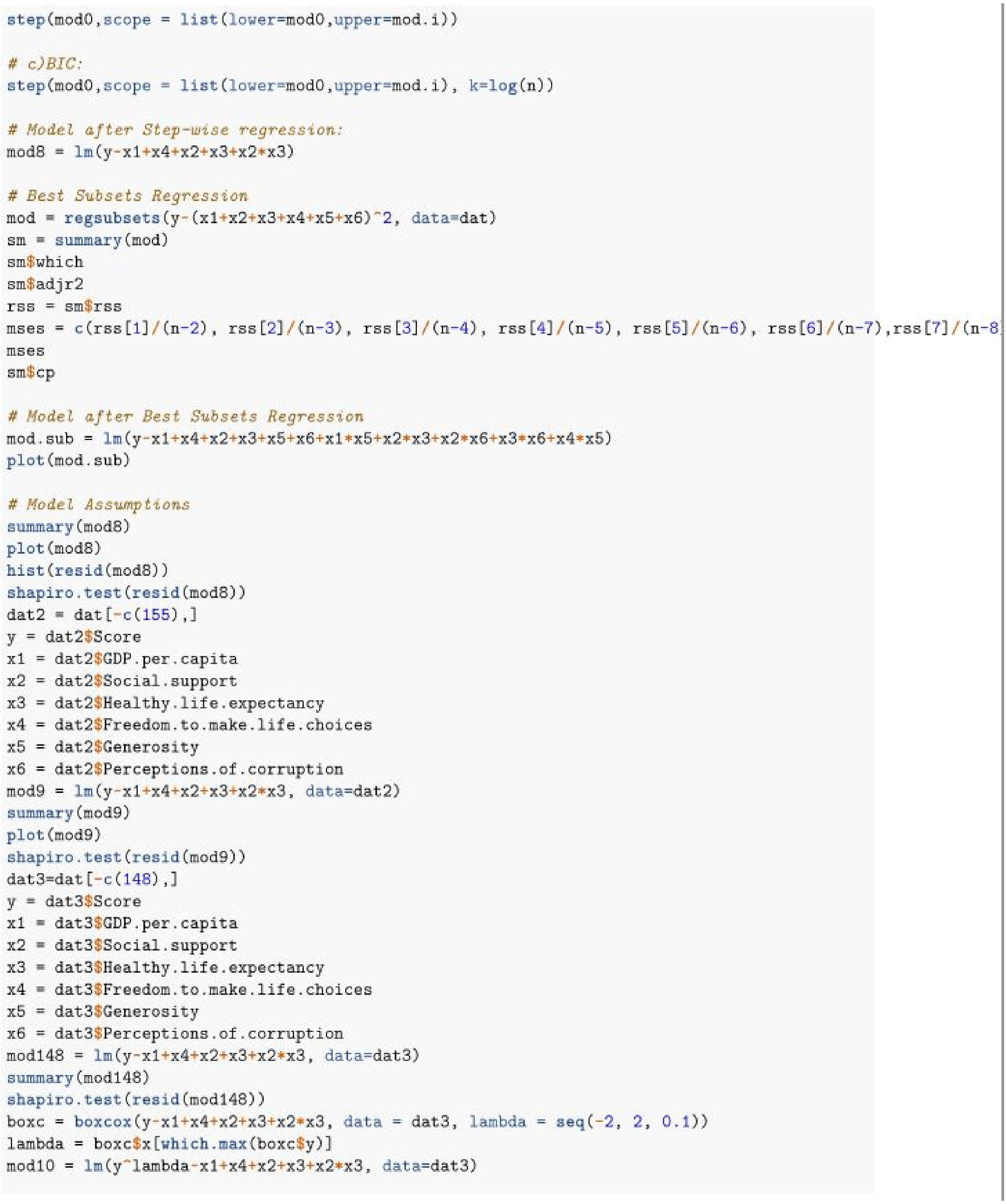
From our model assessments, we can conclude that the best model is one with x1 = GDP per capita, x2 = Social support, x3 = Healthy life expectancy, x4 = Freedom to make life choices and x2:x3 = interaction between social support and healthy life expectancy.The Central African Republic also has poor social support, GDP, and healthy life expectancy. In general, a lot of African regions show low scores. This means that those parameters are the most significant ones for our model, and give us some insight into what is important in determining a country’s happiness rank and score. We can potentially use this information to focus resources into those factors to improve a country’s standard of living.

We can try testing more models to see how standard of living factors affect other parameters like GDP. For example, we can see the Central African Republic region has an exceptionally low GDP and 0 social support and low healthy life expectancy. A new research goal can be conducted to determine how such low scores in these factors affect GDP.
The model is not extensive, and there are many potential factors that could come into play when determining happiness. For example, suicide rates in Japan could be high due to work-life balance, and that could also impact the happiness score of that region.
Each predictor could also have many other sub parameters that affect them. For example, healthy life expectancy could be affected by environment, war time, etc.

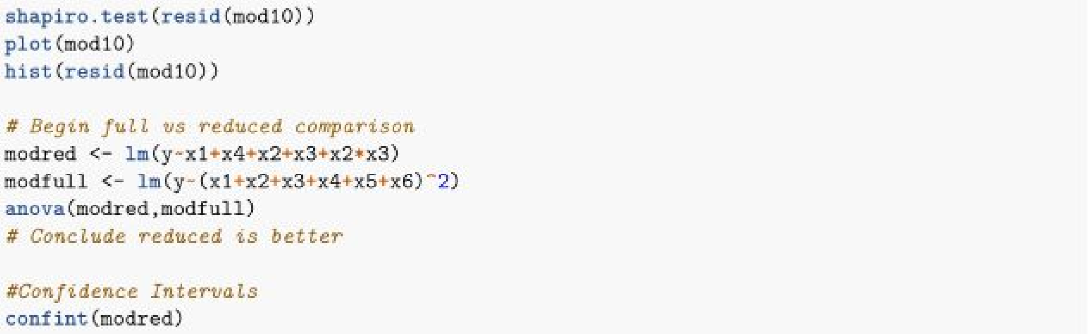


## Appendix

```
# Setup
library(GGally)
library(leaps)
library(MASS)
dat = read.csv(’2019.csv’, header = TRUE)
head(dat)
y = dat$Score
x1 = dat$GDP.per.capita
x2 = dat$Social.support
x3 = dat$Healthy.life.expectancy
x4 = dat$Freedom.to.make.life.choices
x5 = dat$Generosity
x6 = dat$Perceptions.of.corruption
n = nrow(dat)

# Scatter-plot of the data
ggpairs(dat, cardinality_threshold = NULL)

# Step-wise Regression
# a)F-Test:
x6 <- I(x6^2)
mod0 = lm(y~1)
add1(mod0, ~.+x1+x2+x3+x4+x5+x6, test=’F’)
mod1 = update(mod0, ~.+x1)
add1(mod1,~.+x2+x3+x4+x5+x6, test=’F’)
mod2 = update(mod1,~.+x4)
summary(mod2)
add1(mod1,~.+x2+x3+x5+x6, test=’F’)
mod3 = update(mod2,~.+x2)
summary(mod3)
add1(mod3,~.+x3+x5+x6, test=’F’)
mod4 = update(mod3,~.+x3)
summary(mod4)
add1(mod4,~.+x5+x6, test=’F’)
mod5 = update(mod4,~.+x6)
summary(mod5)
add1(mod5,~.+x5, test=’F’)
add1(mod5,~.+x1*x2+x1*x3+x1*x4+x1*x5+x1*x6+x2*x3+x2*x4+x2*x5+x2*x6+x3*x4+x3*x5+x3*x6+x4*x5+x4*x6+x5*x6,
mod6 = update(mod5,~.+x2*x3)
summary(mod6)
mod7 = update(mod6,~.-x6)
summary(mod7)
add1(mod7,~.+x1*x2+x1*x3+x1*x4+x1*x5+x1*x6+x2*x4+x2*x5+x2*x6+x3*x4+x3*x5+x3*x6+x4*x5+x4*x6+x5*x6, test=

# b)AIC:
mod.i = lm(y~(x1+x2+x3+x4+x5+x6)^2)
```



```
step(mod0,scope = list(lower=mod0,upper=mod.i))

# c)BIC:
step(mod0,scope = list(lower=mod0,upper=mod.i), k=log(n))

# Model after Step-wise regression:
mod8 = lm(y~x1+x4+x2+x3+x2*x3)

# Best Subsets Regression
mod = regsubsets(y~(x1+x2+x3+x4+x5+x6)^2, data=dat)
sm = summary(mod)
sm$which
sm$adjr2
rss = sm$rss
mses = c(rss[1]/(n-2), rss[2]/(n-3), rss[3]/(n-4), rss[4]/(n-5), rss[5]/(n-6), rss[6]/(n-7),rss[7]/(n-8
mses
sm$cp

# Model after Best Subsets Regression
mod.sub = lm(y~x1+x4+x2+x3+x5+x6+x1*x5+x2*x3+x2*x6+x3*x6+x4*x5)
plot(mod.sub)

# Model Assumptions
summary(mod8)
plot(mod8)
hist(resid(mod8))
shapiro.test(resid(mod8))
dat2 = dat[-c(155),]
y = dat2$Score
x1 = dat2$GDP.per.capita
x2 = dat2$Social.support
x3 = dat2$Healthy.life.expectancy
x4 = dat2$Freedom.to.make.life.choices
x5 = dat2$Generosity
x6 = dat2$Perceptions.of.corruption
mod9 = lm(y~x1+x4+x2+x3+x2*x3, data=dat2)
summary(mod9)
plot(mod9)
shapiro.test(resid(mod9))
dat3=dat[-c(148),]
y = dat3$Score
x1 = dat3$GDP.per.capita
x2 = dat3$Social.support
x3 = dat3$Healthy.life.expectancy
x4 = dat3$Freedom.to.make.life.choices
x5 = dat3$Generosity
x6 = dat3$Perceptions.of.corruption
mod148 = lm(y~x1+x4+x2+x3+x2*x3, data=dat3)
summary(mod148)
shapiro.test(resid(mod148))
boxc = boxcox(y~x1+x4+x2+x3+x2*x3, data = dat3, lambda = seq(-2, 2, 0.1))
lambda = boxc$x[which.max(boxc$y)]
mod10 = lm(y^lambda~x1+x4+x2+x3+x2*x3, data=dat3)
```



```r
shapiro.test(resid(mod10))
plot(mod10)
hist(resid(mod10))

# Begin full vs reduced comparison
modred <- lm(y~x1+x4+x2+x3+x2*x3)
modfull <- lm(y~(x1+x2+x3+x4+x5+x6)^2)
anova(modred,modfull)
# Conclude reduced is better

#Confidence Intervals
confint(modred)
```