Using Linear Regression to Predict Country Happiness Scores

Andrew Hong TR 6-650
Luis Aragon TR 5-550
PSTAT 126 Xiyue Liao

Introduction

The World Happiness Report is a survey that ranks 155 countries based on their happiness. The score is calculated by asking people the question of how happy they are on a scale of 1-10. This data set contains information from the 2015 survey including factors such as each country's average happiness score, GDP per capita, average life expectancy, freedom index, etc. Our first question involves using the survey factors to quantitatively understand the effect they have on the response (happiness score) by building and choosing the best model and also seeing if the GDP per capita has interaction with any other variable. Our second question involves using the previous model to predict the happiness scores for hypothetical countries with minimal, average, and maximum variable values.

Questions of Interest

1. Which is the best set of predictors for performing linear regression and predicting happiness score? Does the effect of GDP per capita on Happiness Score depend on any other predictors (Are there interaction terms involving Economy/GDP)?
2. What happiness score would a country with minimum, average, or maximum values on all predictors have?

Regression Method

1. For this question we will first isolate the variables we want to use as predictors and response. Our predictors are Economy/GDP per capita (economy), Family Index (family), Health and Life Expectancy (health), Generosity Index (generosity), Trust in Government (trust), and Freedom Index (freedom). Our response is Happiness Score (score). We will use three methods to decide the 'best' model: stepwise, best subsets, and Mallow's CP. For stepwise we will choose the model with the lowest AIC. For best subsets we will choose the subset with the highest adjusted R^2 value. For Mallow's CP, we will select the model whose Mallow's CP value is less than or equal to the number of betas. After employing all three methods, we will judge to see which model we can deem 'best'. After selecting a model, we will analyze the residuals to ensure that a linear model is usable by creating a residuals vs fit, creating a QQ-plot, and doing a Shapiro Wilks test for normality. If the errors are normally distributed and have equal variance throughout, then we can use a linear model.

For the second part of this question, we want to see if econ has an interaction term with any other variable in this model. We can check for this by using an anova-table that includes both our best model and the model that includes everything in the best model plus all the potential econ interaction terms.

2. For this question, we will use the variables in our 'best' model found in the previous question to create 95% prediction intervals for a country with minimal values, average values, and maximum values. We use the variables from our 'best' model because it gives us insight as to which variables are most useful for predicting response and also so that we won't use too few or too many variables. We will set the values to minimum, average, or maximum then create prediction intervals with the predict() function.

Regression Analysis, Results, and Interpretation

Question 1

***Important Details of Analysis*:**

For the first question, our goal was to find a reasonable linear regression model using Happiness Score as a response. After obtaining the data frame we needed with the predictors we wanted to test, we performed a stepwise regression and a subsets regression. The stepwise regression gave us a model that includes economy, family, freedom, health, and trust (in government). The stepwise regression used the criteria of lowest AIC for the next best model (Due to the length of stepwise regression, the full output and each step is in the Appendix). The equation we got for estimated happiness score is:

$$\hat{Y} = 1.90 + 0.81x_{econ} + 1.4x_{family} + 1.4x_{freedom} + 1.0x_{health} + 0.8x_{trust}$$

For subsets regression, we used the regsubsets() function and compared the adjusted R^2 of the best possible models from 1, 2, 3, 4, 5, and 6 predictors. The model with 5 predictors had the highest adjusted R^2 value at 0.7684. These 5 predictors were the same as the predictors found using the stepwise method.

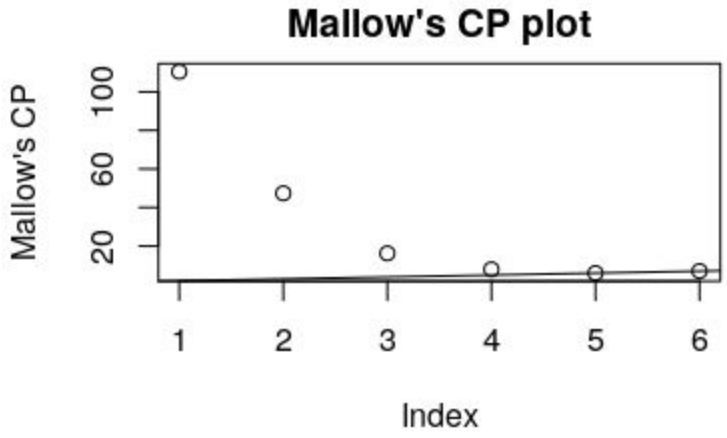
Additionally, to answer our follow up question if the effect of Economy on Happiness Score depends on other predictors, we tested our new model with 5 predictors (reduced) against one with interaction terms (econ multiplied against the other 4 predictors):

$$Reduced : x_{econ} + x_{family} + x_{freedom} + x_{health} + x_{trust}$$
$$Larger : x_{econ} + x_{family} + x_{freedom} + x_{health} + x_{trust} + x_{econ*family} + x_{econ*freedom} + x_{econ*health} + x_{econ*trust}$$

Our F Test came out with a p-value of 0.053 and F statistic of 2.40. This suggests (assuming a 95% confidence level) that we cannot reject the reduced model with no interaction terms for the larger model with interaction terms.

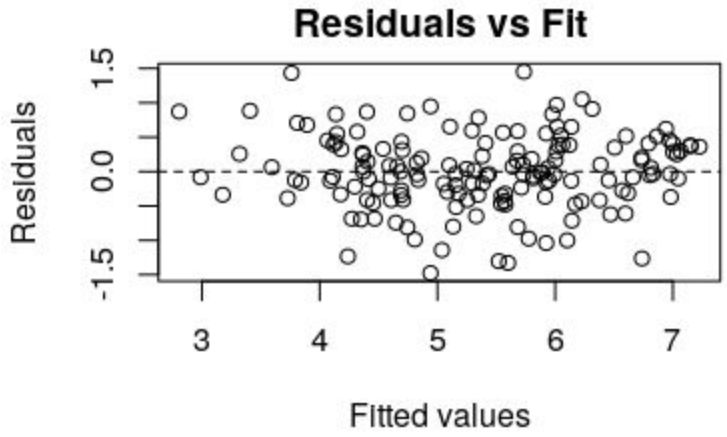
***Diagnostic Checks*:**

This model with 5 predictors is plausible. After getting the same model with these 5 predictors from stepwise and subsets regression, we double checked subsets regression using Mallows C_p Statistic as criteria instead of adjusted R^2 . This confirmed that the model with 5 predictors is optimal for linear regression.

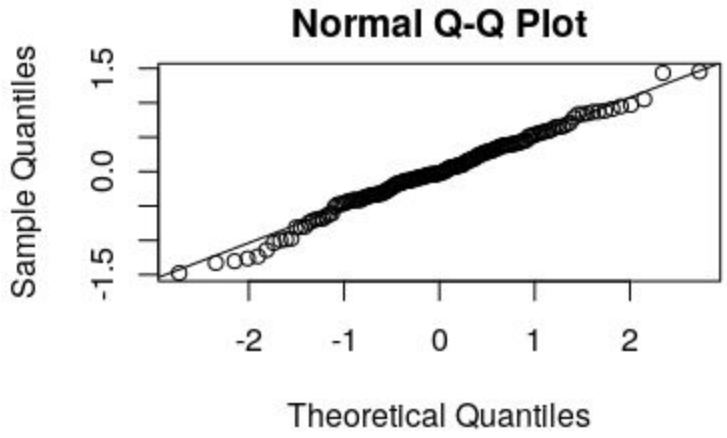


Furthermore, we did an F Test comparing a model with no predictors with a model with all 5 predictors. The p-value we got was less that $2 * 10^{-16}$ and F statistic of 105.18, so we confirmed that we could reject the reduced model (with no predictors) for the one with 5 predictors.

Lastly, after finding the model with 5 predictors, we checked to see if it met the 4 assumptions for linear regression. The Residual vs. Fit plot confirmed that the mean of the responses created a linear function and that the residuals had equal variances.



The Normal QQ-Plot showed that the residuals are normally distributed. This is confirmed by the Shapiro Wilk Test with a p-value of 0.41. This means we cannot reject the null hypothesis that the residuals are normally distributed. Lastly, we know the errors are independent because the values of the predictors are independent.



***Interpretation*:**
According to stepwise regression and subsets regression, the model with 5 predictors is the best one for linear regression. This is a multiple linear regression model which does not overfit or underfit the data. The model which best predict happiness score is

$$\hat{Y} = 1.90 + 0.81x_{econ} + 1.4x_{family} + 1.4x_{freedom} + 1.0x_{health} + 0.8x_{trust}$$

This means, that given values for the five predictors, we can estimate a country's happiness score or predict a country's happiness score. We can also estimate the change of happiness score if one of the values of the predictors were to change. For example, if health in a country rose by one unit, we could estimate that the expected countries happiness score would rise by 1 unit.

Additionally, the effect of Economy/GDP on Happiness Score does not depend on other variables. This means that after controlling for the other for predictors, with every 1 unit increase in Economy, there is an expected 0.81 increase in Happiness score.

Question 2

***Important Details of Analysis*:**

For this research question, we are trying to predict the happiness score of an individual country if they had minimum (average or maximum) values for there variables. This is different than predicting the expected or average happiness score of countries with minimum values on all variables. To do this, we first created a data frame with minimum values from each predictor. We used the min() function on the data of each predictor. We also used the mean() and max() functions. Then we used the predict() function and a 95% confidence level to create a prediction interval for that country. Below are the prediction intervals we got for an individual country.

Values of Predictors Set To	Prediction Interval of Happiness Score
Minimum	[0.749, 3.048]
Average	[4.283, 6.468]
Maximum	[6.607, 8.879]

An important thing to note is that we used the 5 predictor model we found from stepwise and subset regression from the first research question.

***Diagnostic Checks*:**

Our assumptions were plausible for an individual country because we used a larger interval than if we were to use a confidence interval. This prediction interval includes an extra MSE value. The model we used from the first research question was found using various methods and criteria.

***Interpretation*:**

The importance of these prediction intervals is that we are able to predict the happiness score of an individual country given the values of family, freedom, economy, trust in government, and health. In other words, the predictors serve as good indicators on how happy a country may be. They can predict the score of a new country, or estimate the score of an existing country in the data set. Furthermore, if one predictor were to increase k-fold, after controlling for the other predictors, we could find the new confidence interval of the happiness score. We would multiply

the bounds of the confidence interval by k. This can show how changes in certain predictors of individual countries can affect the response of their happiness scores. This does not imply causation, but is a reflection of the model we created from a set of data.

Conclusion

Based on our findings we conclude that Economy, Family, Health and Life Expectancy, Trust in the Government, and Freedom are the best factors from the dataset that we can use to predict or estimate a country's happiness index. Although the correlations don't necessarily suggest causation, these five variables give us insights into not only an entire country's general happiness, but also the impact factors beyond one's personal life can have on an average person. A few predictors that could have also given us insights are ones such as crime index, diversity index, or a pollution index. However, overall we were satisfied with what the dataset gave us and had very little trouble learning from it.

Appendix

```r
library(leaps)
# Access Data
data_dir <- "/home/lma/Documents/STATS/PSTAT126/Project/data"
happy_file <- file.path(data_dir, "2015.csv")
data <- read.csv(happy_file, header = TRUE)

#Get rid of irrelevant columns
data$Country = NULL
data$Happiness.Rank = NULL
data$Standard.Error = NULL
data$Dystopia.Residual = NULL
data$Region = NULL

#Set Variables
attach(data)
econ = Economy..GDP.per.Capita.
fam = Family
health = Health..Life.Expectancy.
gen = Generosity
score = Happiness.Score
trust = Trust..Government.Corruption.
freedom = Freedom


#Question 1: Find the best model and check line conditions. Check for econ interaction.
#########################################################################################################################
#Stepwise method
mod.lower = lm(score ~ 1, data = data)
mod.upper = lm(score ~ econ + fam + health + gen + trust + freedom, data = data)
step(mod.lower, scope = list(lower = mod.lower, upper = mod.upper))
```

```
## Start:  AIC=43.79
## score ~ 1
##
##           Df Sum of Sq     RSS      AIC
## + econ     1   125.540  80.295 -102.949
## + fam      1   112.899  92.935  -79.850
## + health   1   107.953  97.882  -71.656
## + freedom  1    66.456 139.378  -15.814
## + trust    1    32.148 173.687   18.956
## + gen      1     6.693 199.142   40.565
## <none>                205.835   43.787
##
## Step:  AIC=-102.95
## score ~ econ
##
##           Df Sum of Sq     RSS      AIC
## + fam      1    19.752  60.542 -145.562
## + freedom  1    18.571  61.723 -142.509
## + gen      1     7.314  72.981 -116.039
```

```
## + trust    1     5.445  74.849 -112.045
## + health   1     4.626  75.668 -110.325
## <none>                  80.295 -102.949
## - econ     1   125.540 205.835   43.787
##
## Step:  AIC=-145.56
## score ~ econ + fam
##
##           Df Sum of Sq    RSS     AIC
## + freedom  1    10.067 50.476 -172.29
## + trust    1     5.250 55.292 -157.89
## + gen      1     4.720 55.822 -156.39
## + health   1     4.445 56.097 -155.61
## <none>                 60.542 -145.56
## - fam      1    19.752 80.295 -102.95
## - econ     1    32.393 92.935  -79.85
##
## Step:  AIC=-172.29
## score ~ econ + fam + freedom
##
##           Df Sum of Sq    RSS     AIC
## + health   1    3.1304 47.345 -180.41
## + gen      1    1.0278 49.448 -173.54
## + trust    1    0.9629 49.513 -173.34
## <none>                 50.476 -172.29
## - freedom  1   10.0665 60.542 -145.56
## - fam      1   11.2478 61.723 -142.51
## - econ     1   27.5824 78.058 -105.41
##
## Step:  AIC=-180.41
## score ~ econ + fam + freedom + health
##
##           Df Sum of Sq    RSS     AIC
## + trust    1    1.1928 46.152 -182.44
## <none>                 47.345 -180.41
## + gen      1    0.5125 46.833 -180.13
## - health   1    3.1304 50.476 -172.29
## - econ     1    5.3849 52.730 -165.39
## - freedom  1    8.7520 56.097 -155.61
## - fam      1   11.5095 58.855 -148.03
##
## Step:  AIC=-182.44
## score ~ econ + fam + freedom + health + trust
##
##           Df Sum of Sq    RSS     AIC
## <none>                 46.152 -182.44
## + gen      1    0.3004 45.852 -181.47
## - trust    1    1.1928 47.345 -180.41
## - health   1    3.3603 49.513 -173.34
## - econ     1    4.3337 50.486 -170.26
## - freedom  1    4.6399 50.792 -169.31
## - fam      1   12.3001 58.453 -147.11
##
```

```
## Call:
## lm(formula = score ~ econ + fam + freedom + health + trust, data = data)
##
## Coefficients:
## (Intercept)         econ          fam      freedom       health
##      1.8982       0.8053       1.4164       1.4426       1.0338
##       trust
##      0.8540
```

```
#Best Subsets method
mod = regsubsets(cbind(econ, fam, health, gen, trust, freedom), score)
summary(mod)$which
```

```
##   (Intercept) econ   fam health   gen trust freedom
## 1        TRUE TRUE FALSE  FALSE FALSE FALSE   FALSE
## 2        TRUE TRUE  TRUE  FALSE FALSE FALSE   FALSE
## 3        TRUE TRUE  TRUE  FALSE FALSE FALSE    TRUE
## 4        TRUE TRUE  TRUE   TRUE FALSE FALSE    TRUE
## 5        TRUE TRUE  TRUE   TRUE FALSE  TRUE    TRUE
## 6        TRUE TRUE  TRUE   TRUE  TRUE  TRUE    TRUE
```

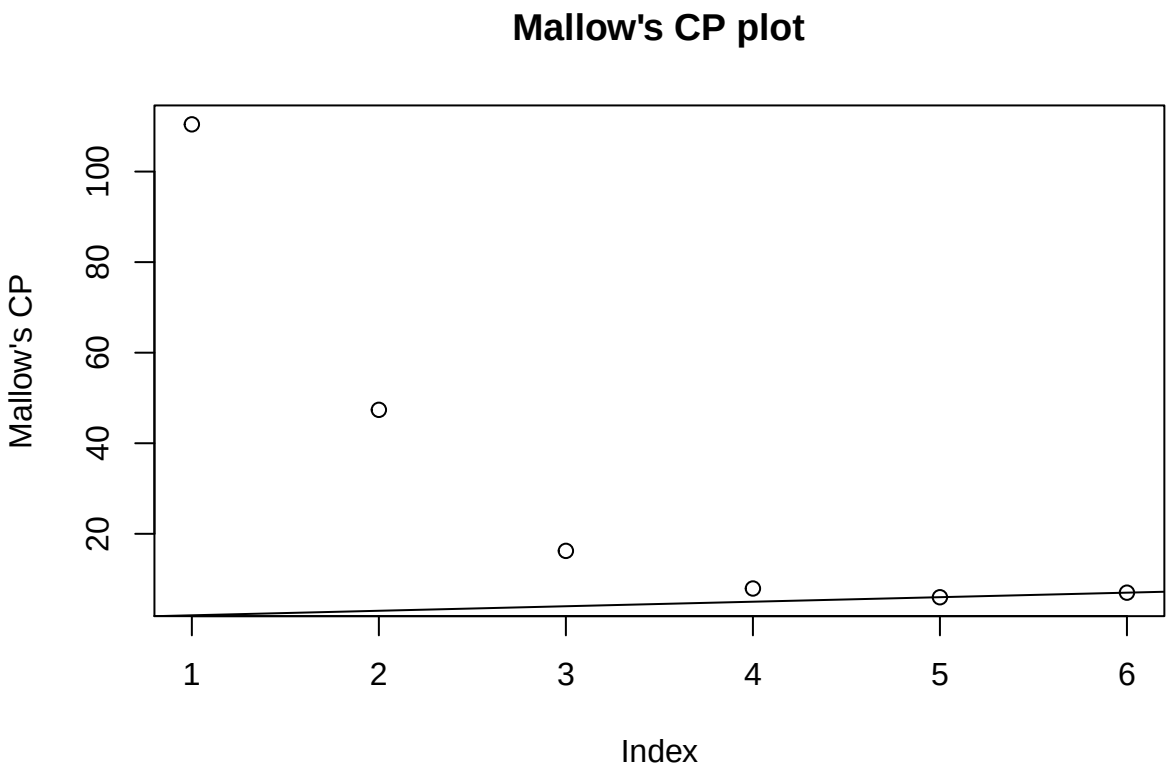
```
summary(mod)$adj
```

```
## [1] 0.6074066 0.7020743 0.7499983 0.7639704 0.7684031 0.7683870
```

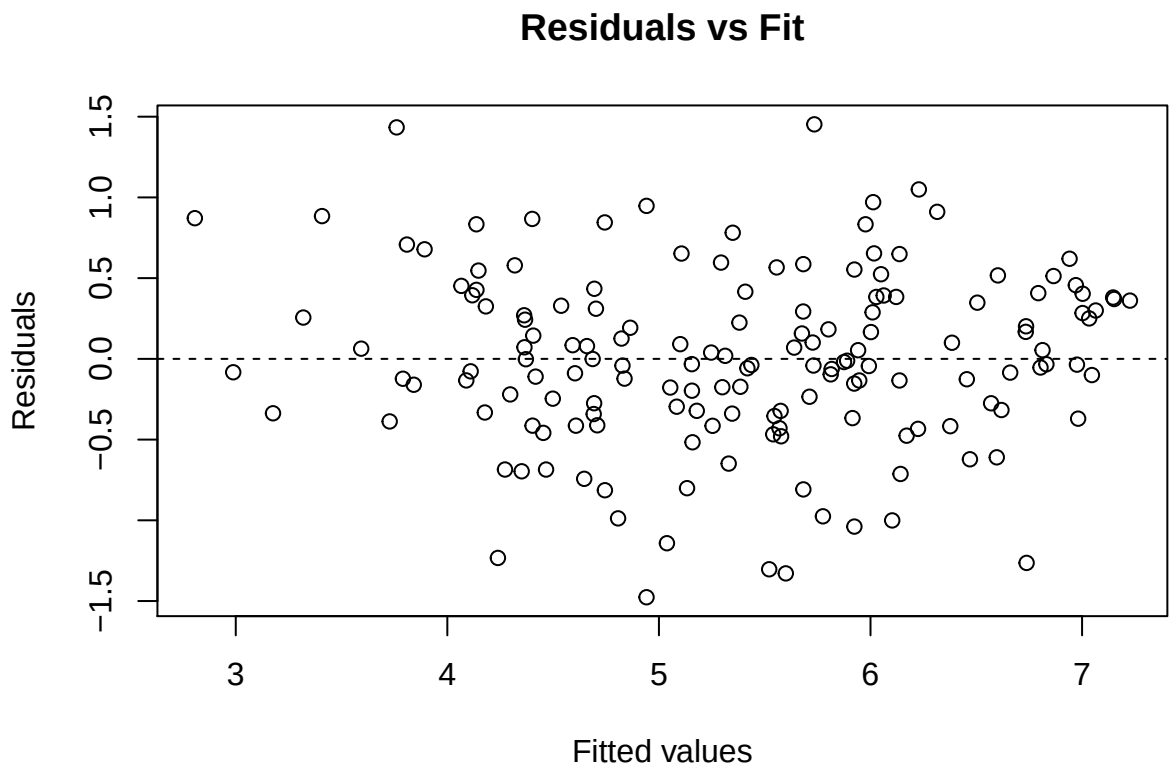
```
#Mallows CP Method and plot
summary(mod)$cp
```

```
## [1] 110.426379  47.377792  16.226709   7.917546   5.989440   7.000000
```

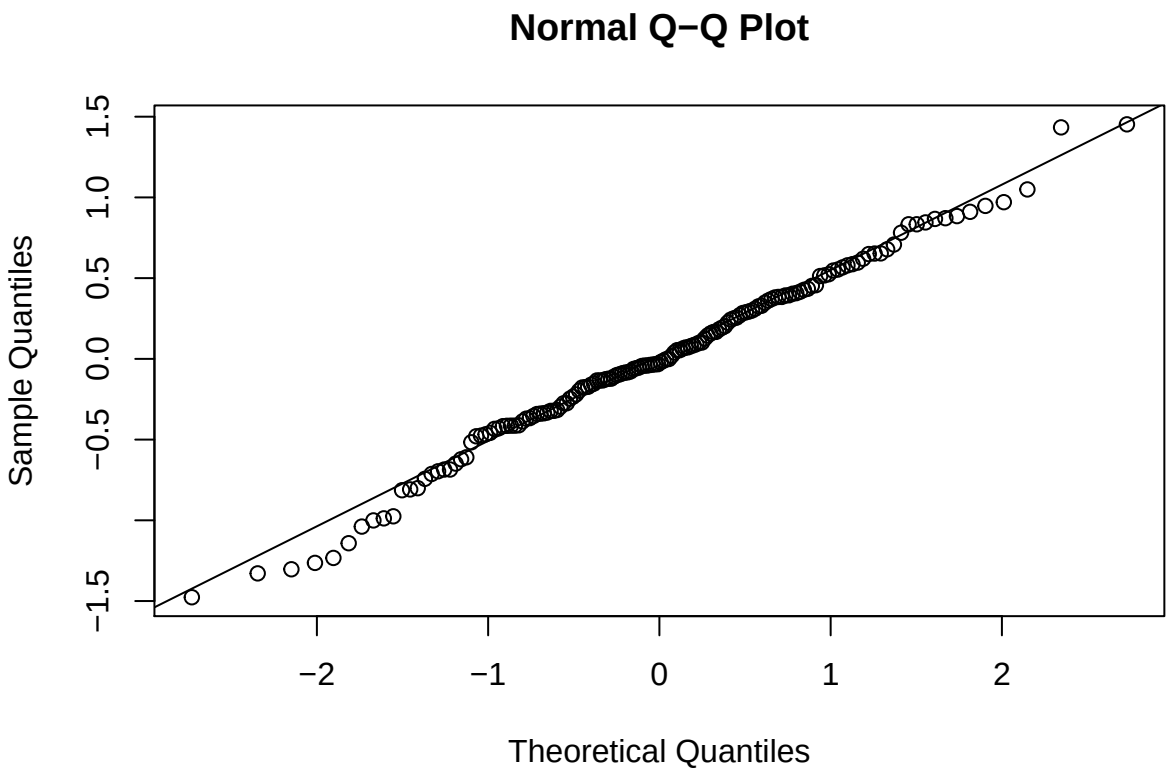
```
plot(summary(mod)$cp, ylab = "Mallow's CP", main = "Mallow's CP plot")
abline(1,1)
```




```r
#Check LINE conditions for best model (Residuals vs fit, QQ-plot)
fit = lm(score ~ econ + fam + health + trust + freedom)
yhat = fitted(fit)
e = score - yhat
plot(yhat, e, xlab = "Fitted values", ylab = "Residuals", main = "Residuals vs Fit")
abline(h = 0, lty = 2)
```




```
qqnorm(e)
qqline(e)
```




```
shapiro.test(e)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  e
## W = 0.99092, p-value = 0.4107
```

```
#F-test for model and for econ interaction
anova(lm(score ~ 1),fit)
```

```
## Analysis of Variance Table
##
## Model 1: score ~ 1
## Model 2: score ~ econ + fam + health + trust + freedom
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    157 205.835
## 2    152  46.152  5    159.68 105.18 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit, lm(score ~ econ + fam + health + trust + freedom + econ*fam + econ* health + econ*trust + ec
```

```
## Analysis of Variance Table
##
## Model 1: score ~ econ + fam + health + trust + freedom
## Model 2: score ~ econ + fam + health + trust + freedom + econ * fam +
##     econ * health + econ * trust + econ * freedom
```

```
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1    152 46.152
## 2    148 43.340  4    2.8124 2.401 0.05254 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#########################################################################################################################


#Question 2: Find prediction interval for a country with lower, average, and max values
#########################################################################################################################
#Prediction for min stats
minnew = data.frame(econ = min(econ), fam = min(fam), health = min(health), trust = min(trust), freedom
minint = predict(fit, minnew, se.fit = TRUE, interval = 'prediction', level = .95)
minint
```

```
## $fit
##        fit       lwr      upr
## 1 1.898207 0.7488108 3.047604
##
## $se.fit
## [1] 0.1866011
##
## $df
## [1] 152
##
## $residual.scale
## [1] 0.5510305
```

```
#Prediction for mean stats
meannew = data.frame(econ = mean(econ), fam = mean(fam), health = mean(health), trust = mean(trust), fr
meanint = predict(fit, meannew, se.fit = TRUE, interval = 'prediction', level = .95)
meanint
```

```
## $fit
##        fit      lwr      upr
## 1 5.375734 4.283627 6.467842
##
## $se.fit
## [1] 0.04383764
##
## $df
## [1] 152
##
## $residual.scale
## [1] 0.5510305
```

```
#Prediction for max stats
maxnew = data.frame(econ = max(econ), fam = max(fam), health = max(health), trust = max(trust), freedom
maxint = predict(fit, maxnew, se.fit = TRUE, interval = 'prediction', level = .95)
maxint
```

```
## $fit
##        fit      lwr      upr
## 1 7.743114 6.606753 8.879475
##
## $se.fit
```

```
## [1] 0.1648827
##
## $df
## [1] 152
##
## $residual.scale
## [1] 0.5510305
```

```
#########################################################################################################################
detach(data)
```