RP4

Claire Perkins

10/15/2020

```
# Load library
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse
1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## Warning: package 'readr' was built under R version 4.0.3

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(car)

## Loading required package: carData

##
## Attaching package: 'car'

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

## The following object is masked from 'package:purrr':
##
##     some

# Import dataset
dataproject <- read.csv("2015.csv")
# Select variables to include in the model
mydata <- dataproject[c("HappScore","LifeEx","Freedom","GDPperCap")]
```

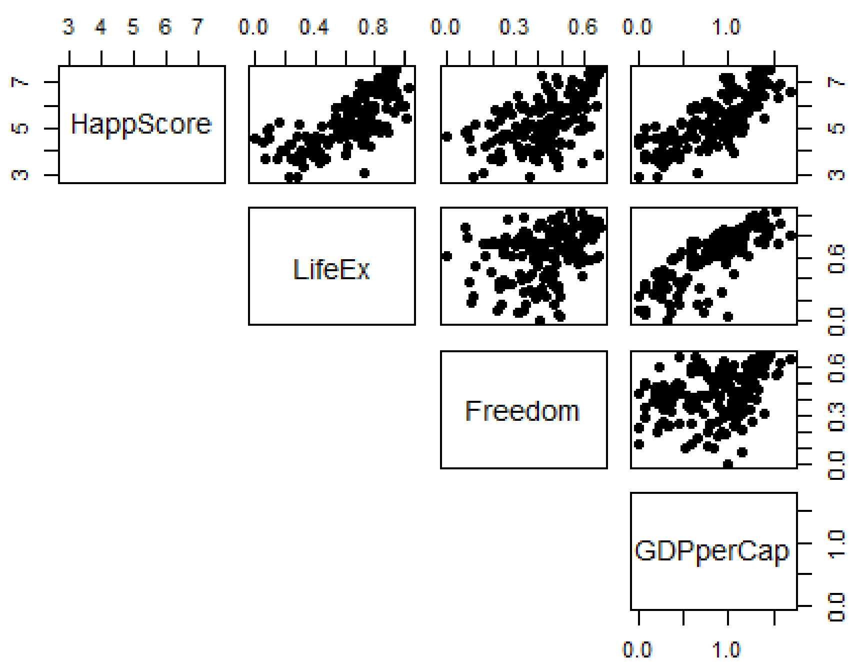
What is the topic and variables

The final topic for this study is exploring the factors that play into a country's overall happiness. Specifically, a country's life expectancy, GDP and Freedom score will all be evaluated to see what has a stronger effect on a country's overall happiness score. I accessed the data from a world happiness report from the website "Kaggle". I expect to find that life expectancy will have the greatest effect on happiness with freedom having the least effect on happiness. I will conduct a simple linear regression on each of the variables to determine which one has the strongest correlation.

```
library(knitr)
kable(head(mydata))
```

HappScore	LifeEx	Freedom	GDPperCap
7.587	0.94143	0.66557	1.39651
7.561	0.94784	0.62877	1.30232
7.527	0.87464	0.64938	1.32548
7.522	0.88521	0.66973	1.45900
7.427	0.90563	0.63297	1.32629
7.406	0.88911	0.64169	1.29025

```
# Create a scatterplot matrix
# Visualize the relationships between the response and each predictor
variable, and relationships between predictors
pairs(mydata, pch = 19, lower.panel = NULL) # pch sets the symbols to
represent the data, here 19 is a solid dot
```




```
# Create a correlation matrix
# Calculate the correlation coefficients for each relationship
cor(mydata)
```

```
##           HappScore    LifeEx   Freedom GDPperCap
## HappScore 1.0000000 0.7241996 0.5682109 0.7809655
## LifeEx    0.7241996 1.0000000 0.3604765 0.8164780
## Freedom   0.5682109 0.3604765 1.0000000 0.3702997
## GDPperCap 0.7809655 0.8164780 0.3702997 1.0000000
```

Some analysis of the data

The relationship between all three variables is positive and linear. Happiness Score and GDP had the strongest relationship with a correlation coefficient of 0.78. Happiness Score and Life Expectancy was still strong with a correlation coefficient of 0.72. Happiness Score and Freedom had a moderate relationship with a correlation coefficent of 0.568.

RP3

Create a R Markdown script and upload your pdf document including: an analysis of the relationships between predictors, and between predictors and response: investigate if there are some interactions, a nonlinear pattern, a need for transformations a model building strategy performed on a set of predictors

After running an MLR the regression equation is Y(happiness) = 2.5691 + 0.9532 X1 (Life Expectancy) + 2.3517 X2 (freedom) + 1.4157 X3 (GDP)

For every unit increase in life expectancy, the overall happiness of a country ,assuming the other variables stay constant, increases by 0.9532. For every unit increase in freedom, the overall happiness of a country ,assuming the other variables stay constant, increases by 2.3517 For every unit increase in overall GDP, the overall happiness of a country ,assuming the other variables stay constant, increases by 1.4157 GDP and freedom have a significance level of two stars and Life Expectancy has a significance level of two stars. After conducting an ANOVA, all three variables have a significance level of three starts. This suggests the possibility of some interaction effects but with all of the variables beings significant in addition to strong correlation coefficients with each predictor with happiness.

Equation from Forward and Backwards Model Y(happiness) = 2.5691 + 0.9532 X1 (Life Expectancy) + 2.3517 X2 (freedom) + 1.4157 X3 (GDP) Coefficients: (Intercept) GDPperCap Freedom LifeEx
2.5691 1.4157 2.3517 0.9532

```r
#Write the model with all 3 predictors
reg <- lm(HappScore ~ LifeEx + Freedom + GDPperCap,mydata)
# Display the summary table for the regression model
summary(reg)
```

```
##
## Call:
## lm(formula = HappScore ~ LifeEx + Freedom + GDPperCap, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.55899 -0.36672 -0.00271  0.44859  1.39676
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.5691     0.1697  15.140  < 2e-16 ***
## LifeEx        0.9532     0.3479   2.740  0.00688 **
## Freedom       2.3517     0.3546   6.633 5.27e-10 ***
## GDPperCap     1.4157     0.2141   6.612 5.88e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6182 on 154 degrees of freedom
## Multiple R-squared:  0.7141, Adjusted R-squared:  0.7085
## F-statistic: 128.2 on 3 and 154 DF,  p-value: < 2.2e-16
```

```
#Anova
anova(reg)
```

```
## Analysis of Variance Table
##
## Response: HappScore
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## LifeEx      1 107.953 107.953 282.471 < 2.2e-16 ***
## Freedom     1  22.319  22.319  58.401 2.132e-12 ***
## GDPperCap   1  16.707  16.707  43.717 5.884e-10 ***
## Residuals 154  58.855   0.382
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Fit an empty model with only the response
FitStart <- lm(HappScore ~ 1, mydata)
# Fit a full model with all predictors
FitAll <- lm(HappScore ~ LifeEx + Freedom + GDPperCap,mydata)
# Run the stepwise regression with forward selection based on the AIC
criterion
step(FitStart,direction="forward", scope = formula(FitAll))
```

```
## Start:  AIC=43.79
## HappScore ~ 1
##
##             Df Sum of Sq     RSS      AIC
## + GDPperCap  1   125.540  80.295 -102.949
## + LifeEx     1   107.953  97.882  -71.656
## + Freedom    1    66.456 139.378  -15.814
## <none>                   205.835   43.787
##
## Step:  AIC=-102.95
## HappScore ~ GDPperCap
##
##           Df Sum of Sq    RSS     AIC
## + Freedom  1   18.5711 61.723 -142.51
```

```
## + LifeEx   1    4.6261 75.668 -110.33
## <none>                80.295 -102.95
##
## Step:  AIC=-142.51
## HappScore ~ GDPperCap + Freedom
##
##           Df Sum of Sq    RSS     AIC
## + LifeEx  1    2.8687 58.855 -148.03
## <none>               61.723 -142.51
##
## Step:  AIC=-148.03
## HappScore ~ GDPperCap + Freedom + LifeEx

##
## Call:
## lm(formula = HappScore ~ GDPperCap + Freedom + LifeEx, data = mydata)
##
## Coefficients:
## (Intercept)    GDPperCap      Freedom       LifeEx
##      2.5691       1.4157       2.3517       0.9532
```

```r
# Run the stepwise regression with forward selection based on the AIC
criterion
step(FitAll,direction="backward", scope = formula(FitStart))
```

```
## Start:  AIC=-148.03
## HappScore ~ LifeEx + Freedom + GDPperCap
##
##              Df Sum of Sq    RSS     AIC
## <none>                    58.855 -148.03
## - LifeEx     1    2.8687 61.723 -142.51
## - GDPperCap  1   16.7073 75.562 -110.55
## - Freedom    1   16.8136 75.668 -110.33

##
## Call:
## lm(formula = HappScore ~ LifeEx + Freedom + GDPperCap, data = mydata)
##
## Coefficients:
## (Intercept)       LifeEx      Freedom    GDPperCap
##      2.5691       0.9532       2.3517       1.4157
```

```r
#Fitting data with a first order model
First <- lm(HappScore ~ LifeEx + Freedom + GDPperCap, mydata)
summary(First)
```

```
##
## Call:
## lm(formula = HappScore ~ LifeEx + Freedom + GDPperCap, data = mydata)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -1.55899 -0.36672 -0.00271  0.44859  1.39676 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   2.5691     0.1697  15.140  < 2e-16 ***
## LifeEx        0.9532     0.3479   2.740  0.00688 ** 
## Freedom       2.3517     0.3546   6.633 5.27e-10 ***
## GDPperCap     1.4157     0.2141   6.612 5.88e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6182 on 154 degrees of freedom
## Multiple R-squared:  0.7141, Adjusted R-squared:  0.7085 
## F-statistic: 128.2 on 3 and 154 DF,  p-value: < 2.2e-16
```

Evaluating a First order Model

The MLR regression equation for a first order model is the same as the MLR equation used from the forwards and backwards model indicating that there is not currently a model preferance and all three variables are significnt to the regression equation. While the first-order model is a basical model that uses all of the variables regaurdless of whether or not they are significant, the forwards and backwards model accounted for variables that were significant to the regression equation and provides a comprehensive justification for using Life Expectancy, freedeom, and per capita GDP to help predict a country's happiness score.

Y(happiness) = 2.5691 + 0.9532 X1 (Life Expectancy) + 2.3517 X2 (freedom) + 1.4157 X3 (GDP)

Checking for assumptions

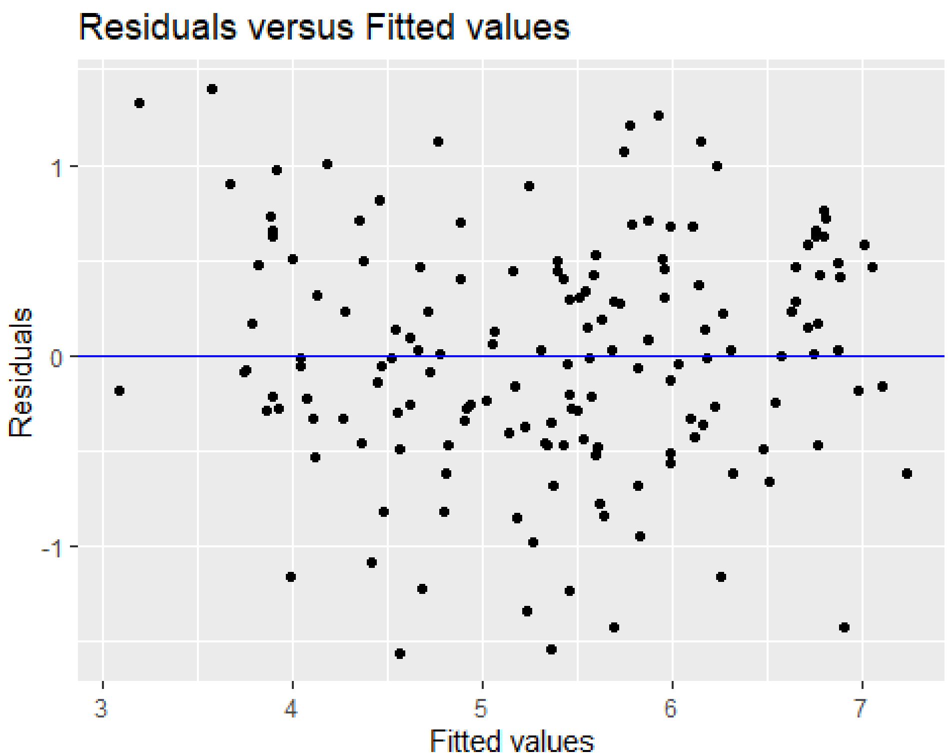
The Variance Inflatio Factor (VIF) for all three of the variables were less than 5 with lifeex at 3.035390, freedom at 1.172690 and GDP per cap at 3.060643 indicating that multicolinearity is not a problem with these variables.

The plot of the residuals versus fitted values indicates that there are a few possible outliers (at the top left and bottom right of the plot), otherwise the points seem to be scattered around randomly which indicate that the assumption of linearity or equal variance are met.

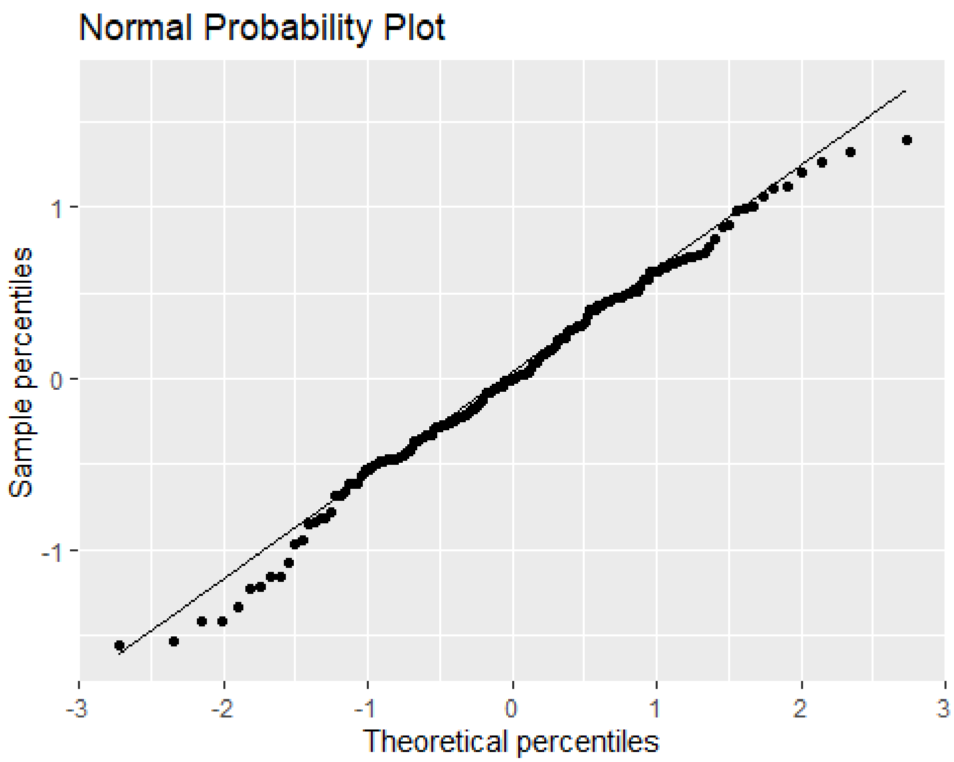
The normal probability plot of the residuals indicates that there are some heavy tails, and possibly some outliers at the top and bottom but overall the relationship between sample and theoretical percentiles seems to be approximately linear and the assumption that the residuals are normally distributed is approximately met.

```r
# Residuals versus Fitted values
mydata$resids <- residuals(reg)
mydata$predicted <- predict(reg)
ggplot(mydata, aes(x=predicted, y=resids)) + geom_point() +
geom_hline(yintercept=0, color = "blue") +
```

```
 labs(title ="Residuals versus Fitted values", x = "Fitted values", y =
"Residuals")
```




```
# Normal probability plot
ggplot(mydata, aes(sample = resids)) + stat_qq() + stat_qq_line() +
labs(title ="Normal Probability Plot", x = "Theoretical percentiles", y =
"Sample percentiles")
```




```r
# VIF Factor
Happ <- lm(HappScore ~ LifeEx + Freedom + GDPperCap, mydata)
vif(Happ)

##    LifeEx   Freedom GDPperCap
##  3.035390  1.172690  3.060643
```