STA437H1S Project Report

Yulin WANG 1003942326
Haiming Xu 1003407011

2020-04-03

I. Abstract

We analyzed the data of 141 countries from the World Happiness report of 2017. We wanted to predict the happiness score of a country using the other nine variables, including logarithm of GDP, social degree, healthy life expectancy at birth, freedom, generosity, corruption, positive effects as happiness, laugh and enjoyment, negative effects as worry, sadness and anger, and gini of household income.

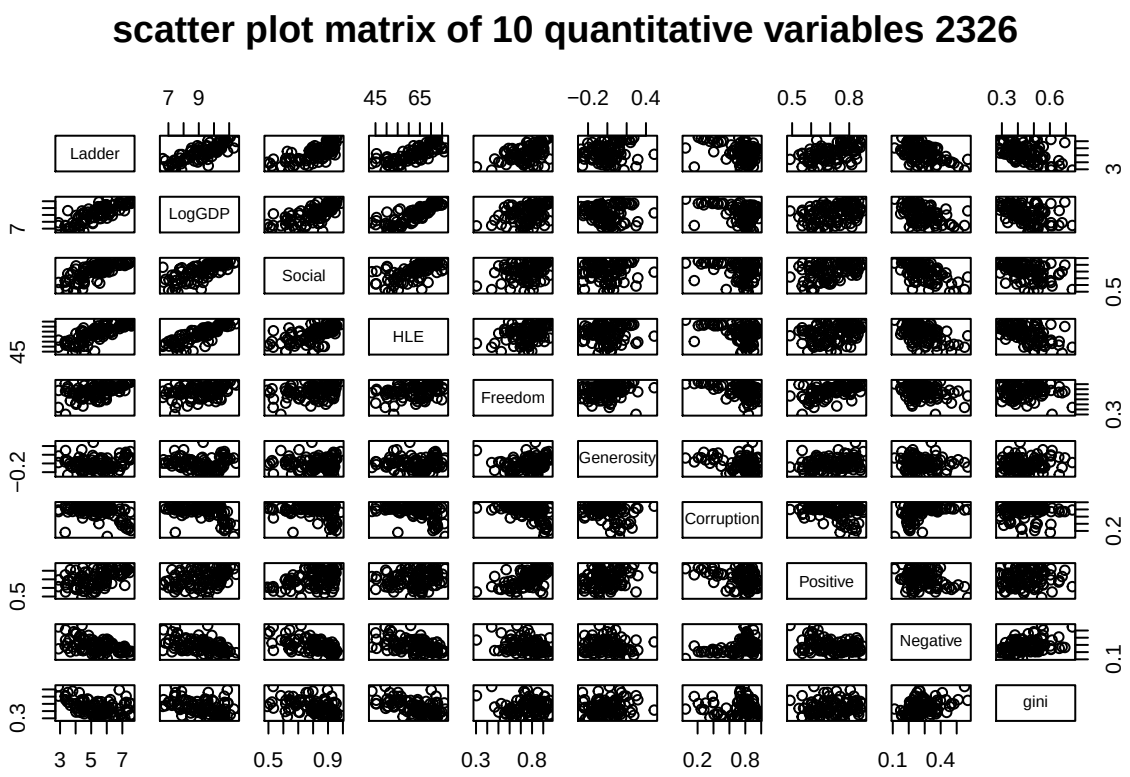
After distilling the data, we fitted a multiple linear model for the response variable, happiness score using all the nine original explanatory variables. However, this linear model violates the multicollinearity assumption. Note that principal components not only have zero correlation with each other, but also can be applied to reduce the dimension. Thus, we decided to apply Principal Components Analysis to fit another multiple linear model with the normalized principal components from the sample correlation matrix.

Moreover, we analyzed the three statistically significant PCs retained in our final model. The first PC can be viewed as a measure of the logarithm of GDP, social degree and healthy life expectancy at birth of a country. The fourth PC can be viewed as a measure of negative effects of a country. The fifth PC can be viewed as a measure of the generosity, corruption and negative effects of a country.

In conclusion, these three PCs have negative relationship with happiness score, which means the higher the PC is, the lower happiness score the country has, holding other PCs constant. More specifically, for each unit increase in the first, fourth, and fifth PC, the happiness score of the country will decrease about 0.515, 0.122 and 0.269 unit on average respectively, holding the other two PCs constant.

II. Data Manipulation and Summary

There are 141 countries in our original data set with 11 variables. Firstly, we omitted the first column(country names) since it's unnecessary for our analysis, and removed data of 22 countries with missing values. Then we took a random sample of 100 countries as our analysis data. Here is a scatterplot matrix for all pairs of 10 quantitative variables in the distilled data.



These two tables illustrate the summary statistics for the sample data.

Table 1: summary statistics for happiness data 1 2326

Ladder	LogGDP	Social	HLE	Freedom
Min. :2.888	Min. : 6.633	Min. :0.4928	Min. :43.38	Min. :0.3035
1st Qu.:4.501	1st Qu.: 8.295	1st Qu.:0.7463	1st Qu.:57.21	1st Qu.:0.6982
Median :5.439	Median : 9.362	Median :0.8361	Median :64.70	Median :0.7776
Mean :5.386	Mean : 9.229	Mean :0.8112	Mean :63.22	Mean :0.7633
3rd Qu.:6.123	3rd Qu.:10.257	3rd Qu.:0.9075	3rd Qu.:69.78	3rd Qu.:0.8544
Max. :7.596	Max. :11.459	Max. :0.9849	Max. :76.41	Max. :0.9578

Table 2: summary statistics for happiness data 2 2326

Generosity	Corruption	Positive	Negative	gini
Min. :-0.273875	Min. :0.04731	Min. :0.4887	Min. :0.1109	Min. :0.2822
1st Qu.:-0.094911	1st Qu.:0.70266	1st Qu.:0.6390	1st Qu.:0.2222	1st Qu.:0.3794
Median :-0.017674	Median :0.80966	Median :0.7083	Median :0.2671	Median :0.4318
Mean :-0.007636	Mean :0.74230	Mean :0.7119	Mean :0.2857	Mean :0.4479
3rd Qu.: 0.081096	3rd Qu.:0.86279	3rd Qu.:0.7851	3rd Qu.:0.3370	3rd Qu.:0.5204
Max. : 0.485928	Max. :0.96948	Max. :0.8739	Max. :0.5698	Max. :0.7402

From the above scatterplot, it seems to be a linear pattern between variable Ladder(happiness score) and each of the other nine variables. Moreover, there appears to be some correlation between nine different variables except Ladder. As for the summary statistics table(Table 1 and 2), some of the variables have quite different scales and the maximum values of some variables are much larger than their corresponding third quantiles, thus there may exist outliers.

After that, we handled outliers by using two methods consecutively.
Firstly, we calculated standardized values for the data and removed two outliers whose standardized values were beyond the cutoff ± 3.5. Note that standardized values $z_{ik} = \frac{x_{ik} - \bar{x}_k}{\sqrt{s_{kk}}}$ for ith row and kth column.
Next, we compared the generalized squared distances for the data with the 99% chi-square quantile($\chi^2_{10}(0.99)$) and removed three outliers with higher value than $\chi^2_{10}(0.99)$. Thus, there are 95 countries left in our sample data. Note that the generalized squared distance $d_i^2 = (x_i - \bar{x})' S^{-1} (x_i - \bar{x})$.

Furthermore, we checked multivariate normality assumption and found it was not totally satisfied for our data. However, the Shapiro-Wilk test result shows that the dependent variable of our analysis, Ladder(happiness score), satisfies the normality assumption and since Linear regression assumes that the response variable is normal, so there is no need to apply transformations.

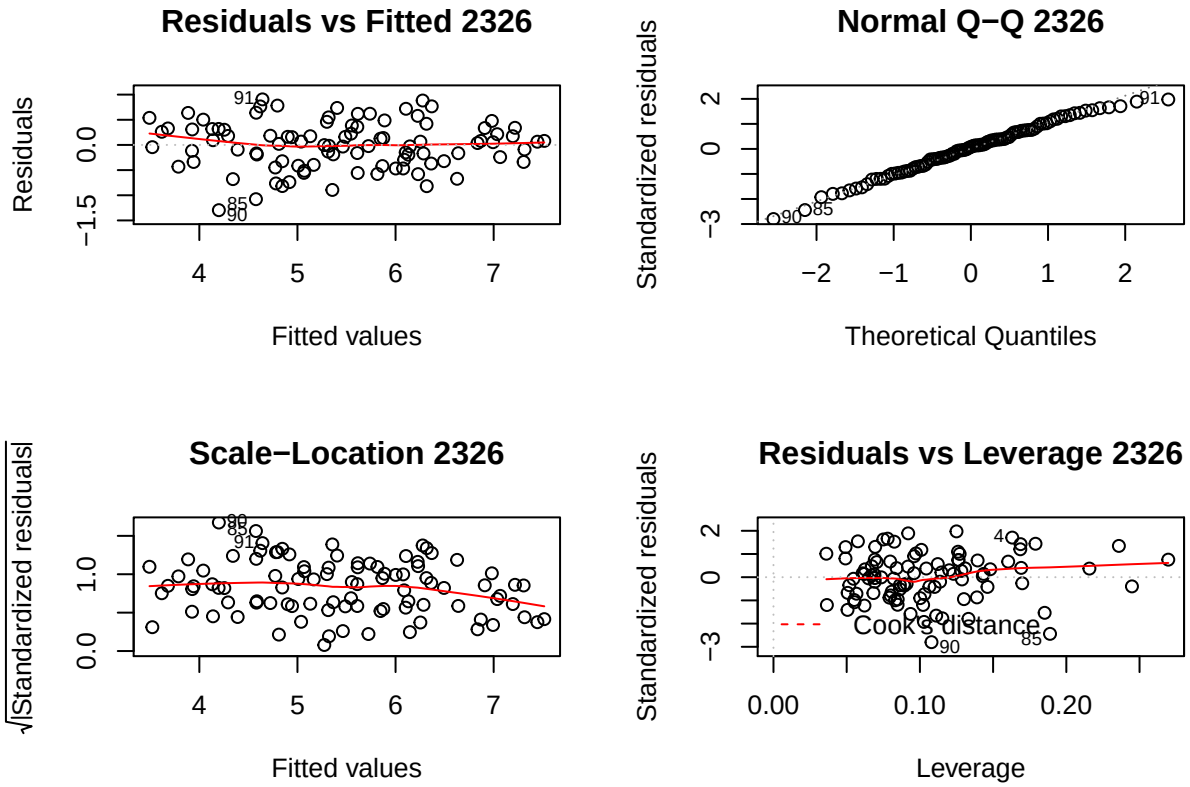
III. Multiple Linear Model using Original variables

We fiited a multiple linear model for the response variable, happiness score using all the nine original explanatory variables, with the ith country takes the form:

$$Ladder_i = \beta_0 + \beta_1 \cdot LogGDP_i + \beta_2 \cdot Social_i + \beta_3 \cdot HLE_i + \beta_4 \cdot Freedom_i + \beta_5 \cdot Generosity_i + \beta_6 \cdot Corruption_i + \beta_7 \cdot Positive_i + \beta_8 \cdot Negative_i + \beta_9 \cdot gini_i + \epsilon_i$$

where
β_0 is the intercept
$\beta_1, ..., \beta_9$ represent the 9 unknown parameters for predictors
ϵ_i is the error term

In order to check whether the assumptions of this model are satisfied, we created these four diagnostic plots.



The first plot, Residuals vs Fitted, shows that there appears equally spread residuals around a horizontal line without distinct patterns, indicating that the residuals do not have non-linear patterns.
For the second plot, Normal Q-Q, the most residuals are lined well on the straight dashed line, thus the residuals approximately follow normal distribution.
And on the third plot, Scale-Location, there exists an approximate horizontal line with most points equally randomly spread plotted. Therefore, the constant variance(homoscedasticity) assumption for residuals is satisfied.
Moreover, there seems to be no cases outside of the Cook's distance on the Residuals vs Leverage plot, so there are no influential points to the regression result.
We also applied Shapiro-Wilk test to check normality assumption for residuals. The p-value of the test result is greater than 0.05, then failed to reject the null hypothesis, thus the normality assumption is satisfied.

In addition, we checked the multicollinearity assumption for this MLR model, and here is the result of VIF(Variance Inflation Factor) test.

```
##     LogGDP     Social        HLE    Freedom Generosity Corruption
##   4.660002   2.594548   3.996853   1.906183   1.305850   1.786070
##   Positive   Negative       gini
##   1.819950   1.641615   1.753083
```

Since variance inflation factors 4.660 and 3.997 are fairly lagre, thus there exists multicollinearity, which means the explanatory variables LogGDP and HLE are highly correlated with at least one of the other predictors in the model.

As is shown above, the assumptions of this model are not fully satisfied, violating the multicollinearity assumption.

The Adjusted R-squared of this model is about 0.8136. This indicates that the model explains about 81.36% of the variability of the response data(happiness score) around its mean by taking into account how many samples we have and how many variables we used.

Plus, here is the estimated coefficient table for the MLR model with all the original explanatory variables.

Table 3: Estimated Coefficients for MLR with original predictors
2326

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.888	0.963	-2.998	0.004
LogGDP	0.295	0.092	3.203	0.002
Social	2.224	0.700	3.179	0.002
HLE	0.030	0.013	2.303	0.024
Freedom	0.897	0.610	1.471	0.145
Generosity	0.832	0.440	1.891	0.062
Corruption	-0.683	0.388	-1.757	0.082
Positive	3.065	0.740	4.144	0.000
Negative	1.346	0.772	1.743	0.085
gini	-1.838	0.708	-2.598	0.011

From the table above(Table 3), it shows that estimated coefficients for predictors LogGDP, Social, HLE, Postive and gini are statistically significant. Among these five predictors, Gini has negative relationship with happiness score while the others have positive relationship with happiness score, holding other predictors constant.

IV. Multiple Linear Model using Principal Components

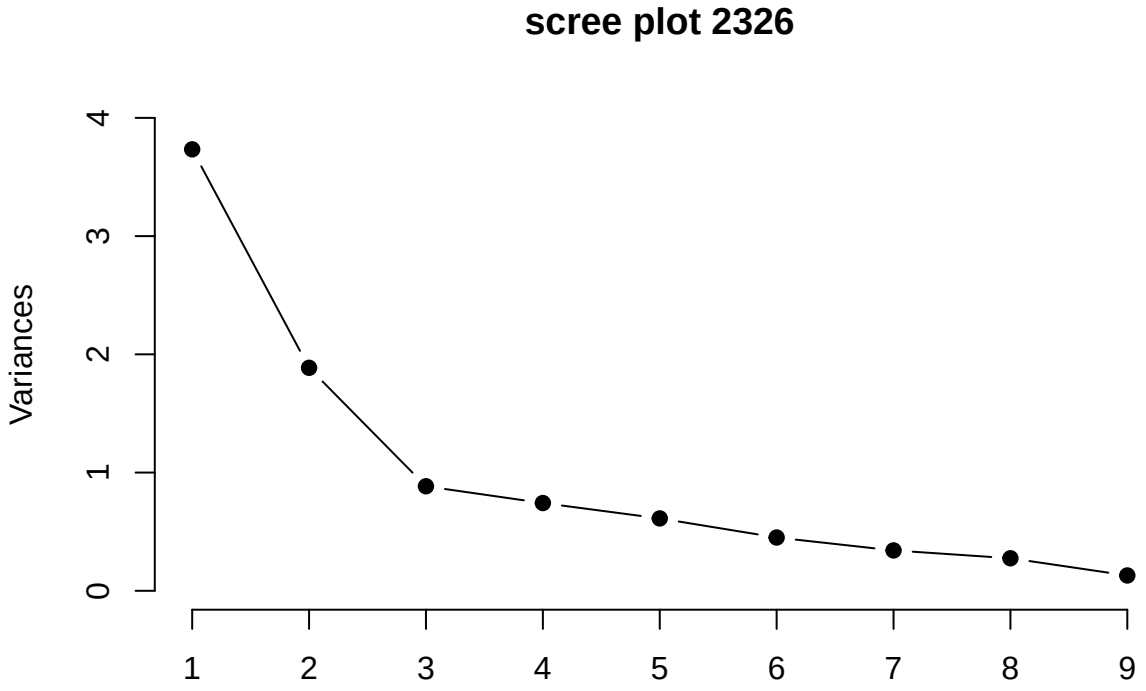
Note that a principal component is a linear combination of the original variables and principal components not only have zero correlation with each other, but also can be applied to reduce the dimension. Since the MLR model using original variables violates the multicollinearity assumption, so we decided to apply Principal Components Analysis.
After computing the sample covariance matrix of explanatory variables matrix X, we notice that these explanatory variables have quite different scales, for instance, the variance for HLE(61.982) is much larger than that of others. This suggests that we may be better off using sample correlation matrix to get the normalized principal components.

Here is the table of normalized principal components for the sampled happiness data.

Table 4: Normalized Principal Components 2326

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
LogGDP	-0.44	0.19	-0.09	-0.34	-0.06	-0.27	-0.07	-0.17	0.73
Social	-0.42	0.15	-0.21	0.25	-0.18	-0.40	-0.04	0.68	-0.21
HLE	-0.43	0.23	0.08	-0.30	-0.25	-0.06	-0.13	-0.47	-0.61
Freedom	-0.30	-0.43	-0.19	-0.25	-0.02	0.57	-0.49	0.25	0.01
Generosity	-0.11	-0.44	0.67	0.24	-0.49	-0.14	-0.09	-0.03	0.13
Corruption	0.33	0.27	-0.37	0.34	-0.54	0.05	-0.47	-0.20	0.12
Positive	-0.28	-0.39	-0.48	0.30	-0.20	0.12	0.55	-0.30	0.02
Negative	0.35	-0.08	-0.11	-0.62	-0.54	-0.05	0.32	0.27	-0.03
gini	0.20	-0.53	-0.27	-0.14	0.21	-0.63	-0.31	-0.17	-0.13

How many PCs to retain? Firstly, we applied the scree-plot method to plot eigenvalues of each component in successive order and tried to identify an elbow in the curve. The scree plot suggests keeping the first 4 or 5 PCs. Then we used the percent of variation explained method and decided to retain the first 5 PCs, which together explain about 86.78% of the total variation.



Thus, we fitted a linear model for the response variable, happiness score, using first five standardized PCs as the explanatory variables.

Table 5: Estimated Coefficients for MLR with first five PCs 2326

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.451	0.050	107.945	0.000
W[, 1]	-0.515	0.026	-19.670	0.000
W[, 2]	0.008	0.037	0.206	0.838
W[, 3]	-0.072	0.054	-1.331	0.187
W[, 4]	-0.122	0.059	-2.072	0.041
W[, 5]	-0.269	0.065	-4.159	0.000

From the summary table above(Table 5), the second and third components seem not to be statistically significant, so we fitted another nested linear model for the response variable, happiness score, using the other three statistically significant standardized PCs as the explanatory variables.

Table 6: Estimated Coefficients for MLR with three significant PCs 2326

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.451	0.050	108.057	0.000
W[, 1]	-0.515	0.026	-19.683	0.000
W[, 4]	-0.122	0.059	-2.065	0.042
W[, 5]	-0.269	0.065	-4.163	0.000

Then we checked the multicollinearity assumption for the nested model, and here is the result of the VIF(Variance Inflation Factor) test.

```
##   W[, 1]   W[, 4]   W[, 5]
## 1.000026 1.000026 1.000020
```

Since all these three variance inflation factors are quite small, so the multicollinearity assumption is satisfied. Furthermore, we did a likelihood ratio test between above two models.

```
## Likelihood ratio test
##
## Model 1: happiness_sample$Ladder ~ W[, 1] + W[, 2] + W[, 3] + W[, 4] +
##     W[, 5]
## Model 2: happiness_sample$Ladder ~ W[, 1] + W[, 4] + W[, 5]
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -64.350
## 2   5 -65.307 -2 1.9142      0.384
```

Since the p-value for the likelihood ratio test is $0.384 > 0.05$, so there is no statistically significant difference between these two models, thus the nested model is preferred to be our final model.
The Adjusted R-squared of our final model is about 0.8123. This indicates that our final model explains about 81.23% of the variability of the response data(happiness score) around its mean by taking into account how many samples we have and how many variables we used.

Here are the analysis of these three retained principal components in our final model. Note that here we determine that a correlation above 0.4 is deemed important.(Table 4)

The 1st PC is mostly about the difference between LogGDP, HLE and Social.
The first principal component increases with decreasing LogGDP, HLE and Social. This suggests that these three criteria vary together. If one decreases, then the remaining ones tend to decrease as well. Thus, this component can be viewed as a measure of the logarithm of a country's GDP, social degree and healthy life expectancy of a country.

The 4th PC is mostly about Negative.
The fourth principal component increases with only one of the values, decreasing Negative. Thus, this component can be viewed as a measure of negative effects-worry, sadness and anger of a country.

The 5th PC is mostly about the difference between Generosity, Corruption and Negative.
The fifth principal component increases with decreasing Generosity, Corruption and Negative.This suggests that these three criteria vary together. If one decreases, then the remaining ones tend to decrease as well. Thus, this component can be viewed as a measure of the generosity, corruption and negative effects-worry, sadness and anger of a country.

The estimated coefficients from Table 6 show that these three principal components have negative relationship with response variable happiness score, indicating the higher the principal component is, the lower happiness score the country has, holding the other pincipal components constant.
Here are the details.
For each unit increase in the first principal component, the happiness score of this country will decrease about 0.515 unit on average, holding the other two principal components constant.
For each unit increase in the fourth principal component, the happiness score of this country will decrease about 0.122 unit on average, holding the other two principal components constant.
For each unit increase in the fifth principal component, the happiness score of this country will decrease about 0.269 unit on average, holding the other two principal components constant.

Appendix

II. Data Manipulation and Summary

```
# set working directory
setwd("~/Desktop/2020 Winter/STA437/project")

# load the data and delete variable 'country' (first column),
# 141 observations(countries) in total.
happiness_data <- read.csv('happiness2017.csv')[,-1]
# remove countries with missing values
happiness_dt <- na.omit(happiness_data) # 119 countries left

# set the seed of randomization
set.seed(2326)
# take a random sample of 100 countries for our data
library(dplyr)
happiness_sample <- sample_n(happiness_dt, 100)

# display multivariate data graphically
plot(happiness_sample, main = "scatter plot matrix of 10 quantitative variables 2326")
# there appears a linear pattern between variable Ladder and other variables.
# And there appears to be some correlation between nine different variables except Ladder

# obtain summary statistics, create two tables
knitr::kable(summary(happiness_sample)[,1:5], digits = 3,
             caption = "summary statistics for happiness data 1 2326")
knitr::kable(summary(happiness_sample)[,6:10], digits = 3,
             caption = "summary statistics for happiness data 2 2326")
```

```
# handle outliers

# 1.calculate the standardized values
standardized_values <- scale(happiness_sample)
# use +-3.5 as cutoff to claim outliers
m <- nrow(standardized_values)
outliers <- NULL # record the index of row that is a outlier
for (i in 1:m){ # m rows
  for (j in 1:10){ # 10 columns
    if (standardized_values[i,j] > 3.5 | standardized_values[i, j] < -3.5){
      outliers <- c(outliers, i)
    }
  }
}
# reomove the outliers
happiness_sample <- happiness_sample[-outliers,]
# 98 countries left, removed 2 outliers (65th, 83th country)

# 2.calculate the generalized squared distance d^2
generalized_squared_distance <- mahalanobis(
  happiness_sample, colMeans(happiness_sample), cov(happiness_sample))
# 99% chi-square quantile cutoff for the outlier
cutoff <- qchisq(0.99, 10) # degree of freedom p=10
# reomove the outliers
```

```r
happiness_sample <- happiness_sample[-which(generalized_squared_distance > cutoff),]
# 95 countries left, removed 3 outliers (14th, 92th, 94th country)
m <- nrow(happiness_sample) # 95 countries left

# check multivariate normality assumption for our sample data
library(MVN)
mvn(happiness_sample)
# The multivariate normality assumption for our data is not satisfied.
# However, the Shapiro-Wilk test result showed that the dependent variable of our analysis,
# Ladder, satisfies the normality assumption and Linear regression assumes that
# the response is normal, so there is no need to apply transformations.
# Note that: a Box Cox transformation is a way to transform
# non-normal dependent variables into a normal shape.
```

III. Multiple Linear Model using Original variables

```r
# Fit a linear model for the response variable- happiness score
# using all the original explanatory variables.
yulin.model.1 <- lm(Ladder ~ LogGDP + Social + HLE + Freedom + Generosity +
                Corruption + Positive + Negative + gini, data = happiness_sample)
summary(yulin.model.1)
knitr::kable(summary(yulin.model.1)$coef, digits=3,
             caption = "Estimated Coefficients for MLR with original preictors 2326")

# create plots to check model assumptions:
# Residuals are normally distributed with constant variance.
# (and are independent of one another)
# Using a 2-by-2 layout, show the 4 diagnostic plots.
par(mfrow = c(2,2))
plot(yulin.model.1, 1, title(main = "Residuals vs Fitted 2326"))
# residuals don't have non-linear patterns
plot(yulin.model.1, 2, title(main = "Normal Q-Q 2326"))
# residuals are approximately normally distributed
plot(yulin.model.1, 3, title(main ="Scale-Location 2326"))
# randomly spread -> equal(constant) variance
plot(yulin.model.1, 5, title(main ="Residuals vs Leverage 2326"))
# no cases are outside of the Cook's distance
# -> no influential points to the regression result

# Check normality assumption for residuals by Shapiro-Wilk test
shapiro.test(yulin.model.1$residuals)
# p-value > 0.05 -> fail to reject null hypothesis -> normal

# Check Multicollinearity assumption by VIF test
library(car)
vif(yulin.model.1)
# violates multicollinearity
# some explanatory variables are highly correlated with each other
```

IV. Multiple Linear Model using Principal Components

```
# exclude the response variable to get the explanatory variables
X <- as.matrix(happiness_sample[2:10])

# Obtain the sample mean vector and sample covariance matrix of X:
x.bar <- apply(X,2,mean)
S <- round(cov(X), 3) # round to 3 decimal
# these explanatory variables have quite different scales
# (the variance for HLE is much larger than others),
# suggesting that we may be better off using sample correlation matrix R to get the PCs

# obtain the standardized variables:
Z <- X
for(i in 1:9){
  Z[,i] <- (X[,i]-x.bar[i])/sqrt(diag(S)[i])
  }

# obtain correlation matrix
R <- cor(X)
```

```
# obtain eigenvalues and eigenvectors of R
eigen_values <- round(eigen(R)$values, 2) # round to 2 decimal

eigen_vectors <- eigen(R)$vectors
rownames(eigen_vectors) <- colnames(X)
colnames(eigen_vectors) <- c("PC1", "PC2", "PC3", "PC4", "PC5", "PC6", "PC7", "PC8", "PC9")
eigen_vectors <- round(eigen_vectors, 2) # round to 2 decimal

# generate a table of principal components
knitr::kable(eigen_vectors, digits = 3, mdToTex = TRUE,
               guessGroup = TRUE, caption = "Normalized Principal Compenents 2326")
# Here we determine that a correlation above 0.4 is deemed important.
# The 1st PC is mostly about the difference between LogGDP, HLE and Social.
# The 2nd PC is mostly about the difference between gini, Generosity and Freedom.
# The 3rd PC is mostly about the difference between Positive and Generosity.
# The 4th PC is mostly about Negative.
# The 5th PC is mostly about the difference between Generosity, Corruption and Negative.
```

```
# Obtain sample PC values:
# Note that a principal component is a linear combination of the original variables.

W <- X  # just to create a data matrix of the same size of X
colnames(W) <- c("PC1", "PC2", "PC3", "PC4", "PC5", "PC6", "PC7", "PC8", "PC9")

# now fill in the entries by calculating sample PCs
for(i in 1:9){ # 9 PC's
  for(j in 1:95){ # 95 rows
    W[j,i] <- eigen_vectors[,i] %*% Z[j,]   # no need to center when using normalized PCCs
    }
}

# How many principal components should be retained?

# ceate a scree plot for standardized PCs
```

```r
screeplot(prcomp(Z), npcs = 9, type = "lines", pch = 19,
          main = "scree plot 2326", ylim = c(0,4))
# the scree plot suggests keeping the first 4 or 5 PCs.

# Proportion of variation explained by each PC
# by using built-in functions in R for a summary:
summary(prcomp(Z))
```

```r
# Regression with first 5 standardized PCs as the explanatory variables
yulin.PC.model.1 = lm(happiness_sample$Ladder ~  W[,1] + W[,2] + W[,3] + W[,4] + W[,5])
knitr::kable(summary(yulin.PC.model.1)$coef, digits=3,
             caption = "Estimated Coefficients for MLR with fist five PCs 2326")

# W2 and W3 seem not to be significant
# Let's remove W2 and W3
yulin.PC.model.2 = lm(happiness_sample$Ladder ~  W[,1] + W[,4] + W[,5])
knitr::kable(summary(yulin.PC.model.2)$coef, digits=3,
             caption = "Estimated Coefficients for MLR with three significant PCs 2326")
summary(yulin.PC.model.2) #Adjusted R-squared:  0.8123

# Note that removing PCs from the model nearly does not change the coefficients!
# Check Multicollinearity assumption
vif(yulin.PC.model.2)

# likelihood ratio test
library(lmtest)
lrtest(yulin.PC.model.1, yulin.PC.model.2)
# no significant difference btw two models -> prefer the nested model PC.model.2
```