Project 2 PSTAT 126

Wilber Delgado and Mason Delan

2023-11-06

```
library("readxl")
WHR = read_excel("/Users/wilberdelgado/Downloads/WHR2023.xls")
```

To get a better set of data we are going to drop the country column, and drop any rows with missing data. Then we are going to take a random sample of 500 to get a normalized dataset.

```
# Drop the "Country name" column
WHR <- WHR[, -which(names(WHR) == "Country name")]

# Remove rows with missing data in specific columns
columns_with_missing_data <- c("year", "Life Ladder", "Log GDP per capita", "Social supp
ort",
                               "Healthy life expectancy at birth", "Freedom to make life
choices",
                               "Generosity", "Perceptions of corruption", "Positive affe
ct",
                               "Negative affect")

WHR <- WHR[complete.cases(WHR[, columns_with_missing_data]), ]
```

```
set.seed(1234)

# Take a random sample of 500 rows from the WHR dataset
sample_size <- 500
random_sample <- WHR[sample(nrow(WHR), sample_size), ]
```

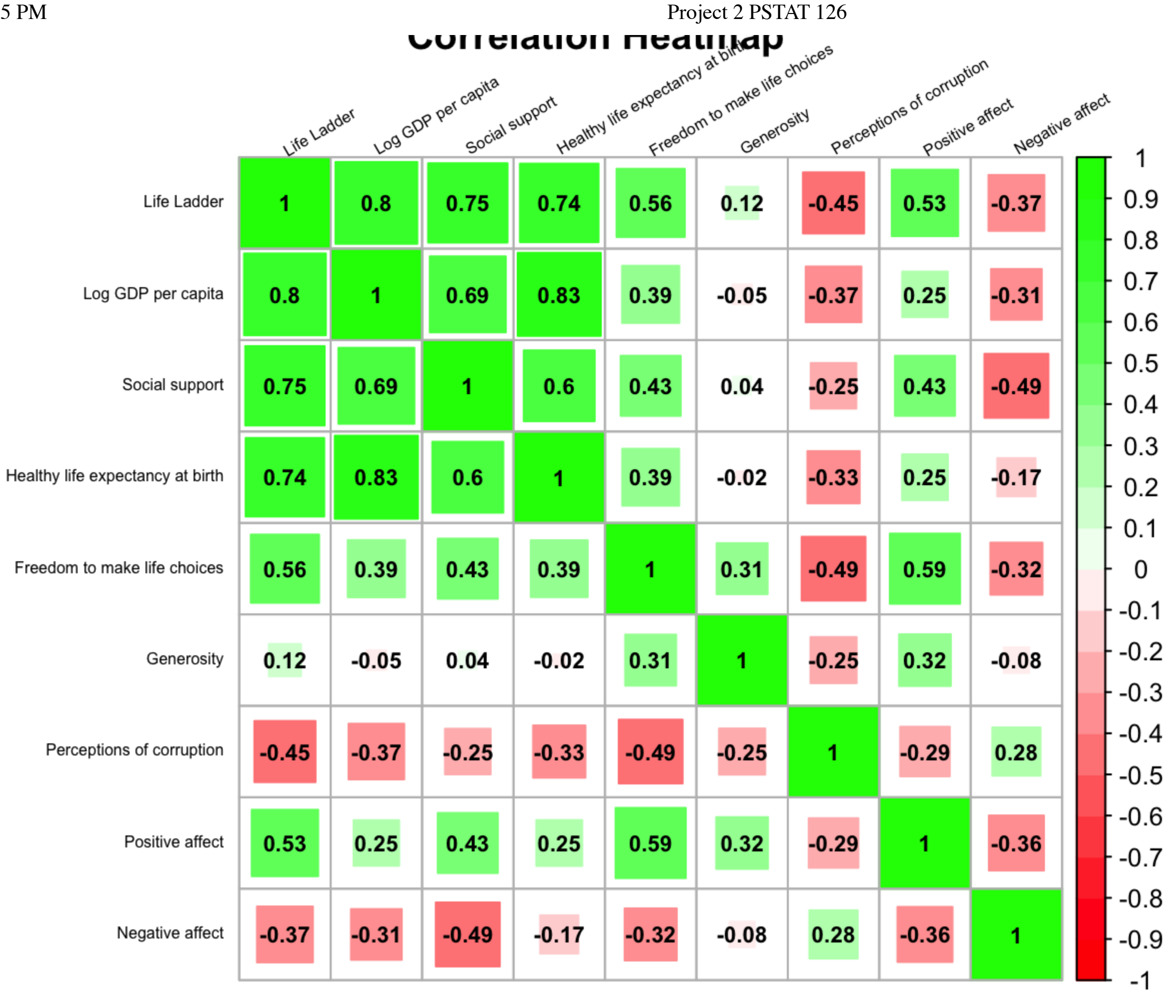
Create heatmap to see the variables' correlation

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
correlation_matrix <- cor(random_sample[, -which(names(random_sample) == "year")])

corrplot(correlation_matrix, method = "square",
         col = colorRampPalette(c("red", "white", "green"))(20),
         tl.col = "black", tl.srt = 30,
         addCoef.col = "black",
         number.cex = 0.7,
         tl.cex = 0.5,  # Reduce the text size for variable labels
         title = "Correlation Heatmap")
```



	Life Ladder	Log GDP per capita	Social support	Healthy life expectancy at birth	Freedom to make life choices	Generosity	Perceptions of corruption	Positive affect	Negative affect
Life Ladder	1	0.8	0.75	0.74	0.56	0.12	-0.45	0.53	-0.37
Log GDP per capita	0.8	1	0.69	0.83	0.39	-0.05	-0.37	0.25	-0.31
Social support	0.75	0.69	1	0.6	0.43	0.04	-0.25	0.43	-0.49
Healthy life expectancy at birth	0.74	0.83	0.6	1	0.39	-0.02	-0.33	0.25	-0.17
Freedom to make life choices	0.56	0.39	0.43	0.39	1	0.31	-0.49	0.59	-0.32
Generosity	0.12	-0.05	0.04	-0.02	0.31	1	-0.25	0.32	-0.08
Perceptions of corruption	-0.45	-0.37	-0.25	-0.33	-0.49	-0.25	1	-0.29	0.28
Positive affect	0.53	0.25	0.43	0.25	0.59	0.32	-0.29	1	-0.36
Negative affect	-0.37	-0.31	-0.49	-0.17	-0.32	-0.08	0.28	-0.36	1

From the heat map, considering that Life Ladder and Healthy life expectancy at birth have a 0.74 correlation, this is a fairly high correlation meaning that the two variable are linearly related. Therefore we are deciding to use these variables as our quantitative variables for Project 2.

Hypothesis

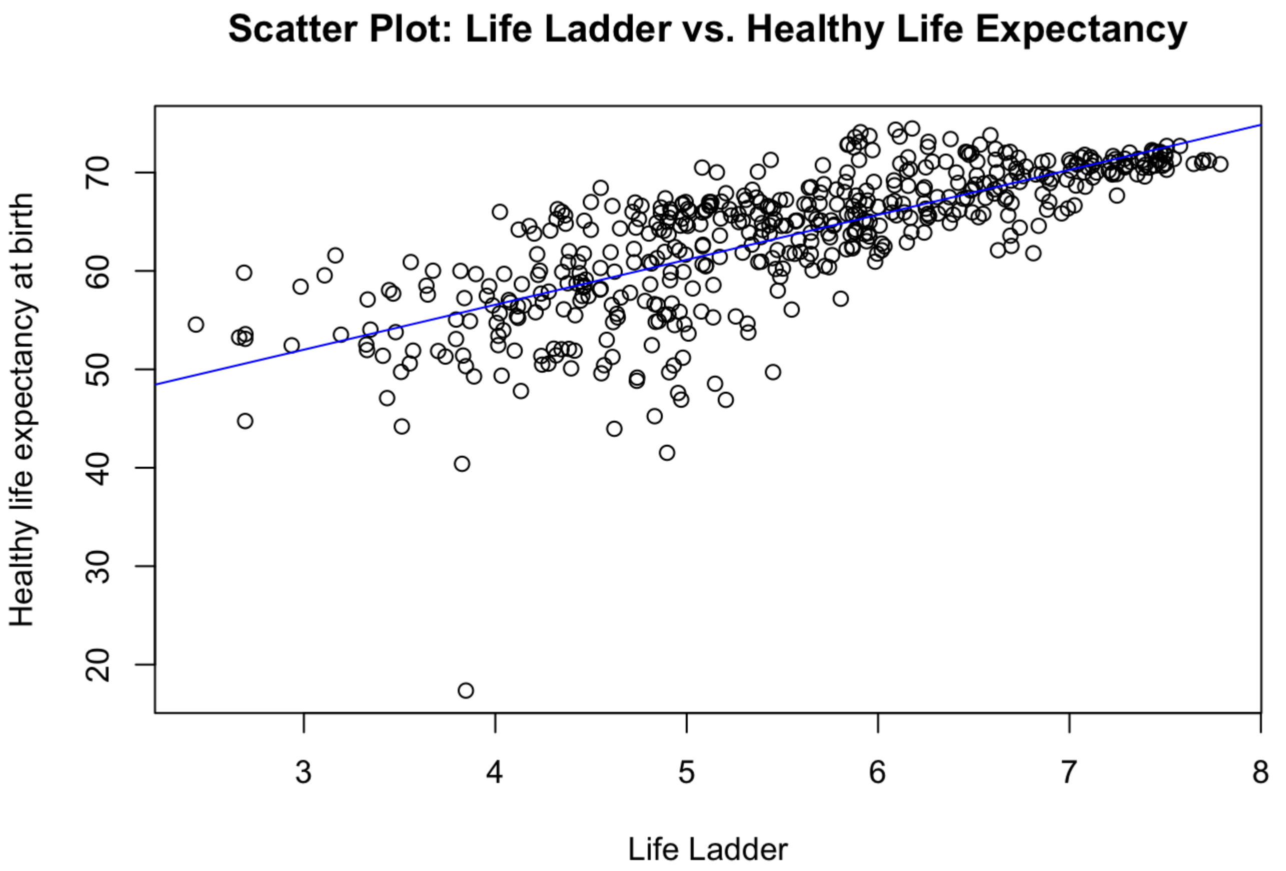
Null Hypothesis: H_0 : There is no relationship between “Life Ladder” and “Healthly life expectancy at birth”.

Alternative Hypothesis: H_1 : There is a linear relationship between “Life Ladder” and “Healthy life expectancy”.

Linearity:

```
# Perform linear regression
lm_model <- lm(`Healthy life expectancy at birth` ~ `Life Ladder`, data = random_sample)

plot(random_sample$`Life Ladder`, random_sample$`Healthy life expectancy at birth`,
     xlab = "Life Ladder", ylab = "Healthy life expectancy at birth",
     main = "Scatter Plot: Life Ladder vs. Healthy Life Expectancy")
abline(lm_model, col = "blue")
```



From the graphing of the scatter plot we are able to see that the variables “Life Ladder” and “Healthy life expectancy at birth” are positively linear and the variability seems constant due to the spread of the points.

We decided to choose “Life Ladder” as our explanatory variable and “Healthy life expectancy at birth” as the response variable. From our linear model summary we get that

$$\text{Healthy life expectancy at birth} = 38.2873 + 4.5719 * \text{Life Ladder}$$

We decided to let our “Life Ladder” value equal to 4.

```
y_hat <- 38.2873 + 4.5719*4
y_hat
```

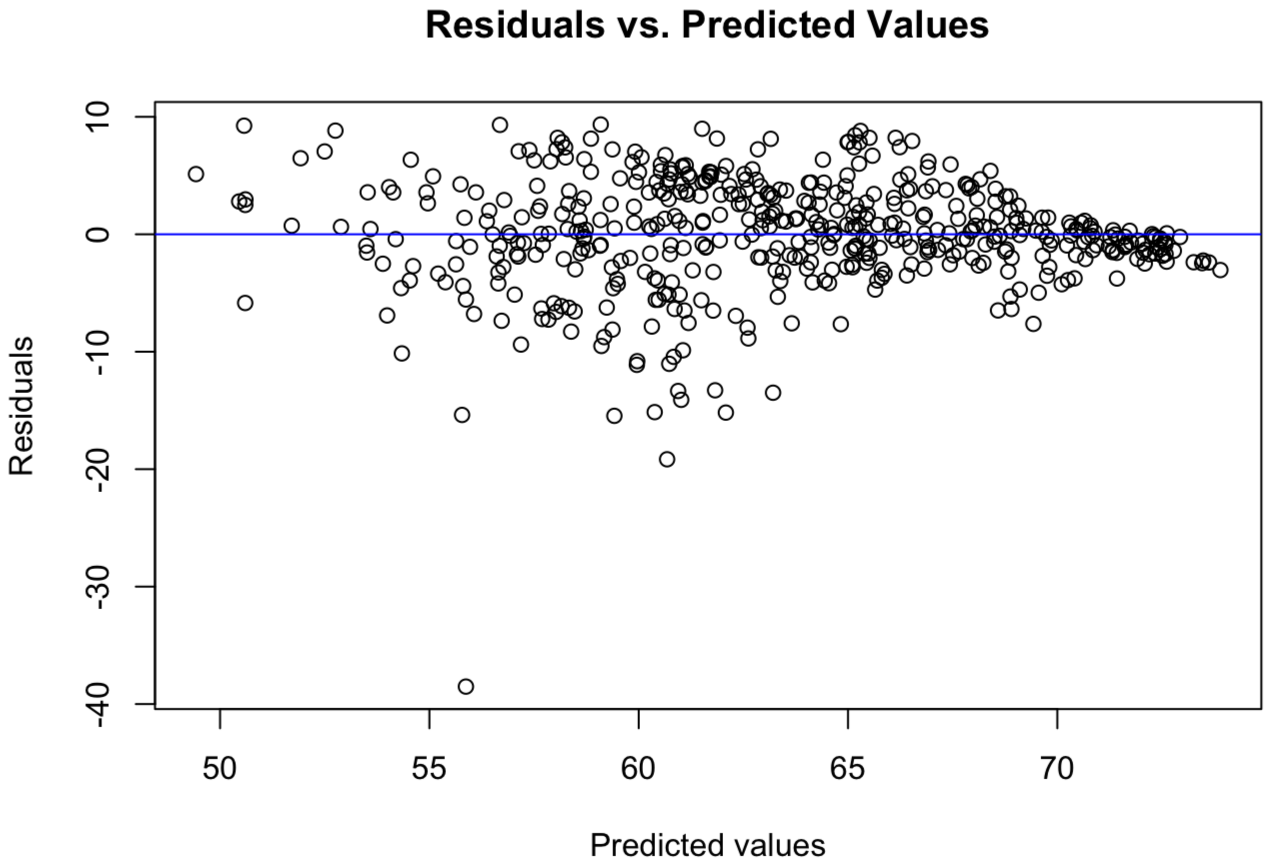
```
## [1] 56.5749
```

Based on the results we can see that if we let our explanatory variable, “Life Ladder”, equal to 4, our response variable, “Healthy Life Expectancy at birth” (or $\hat{Y}$) would equate 56.5749. This indicates that if there a low life satisfaction, then the “Healthy Life Expectancy at birth” or number of years a person could expect to live in good health would roughly be about 57 years old.

```
# Predicted values from the linear regression model
predicted_values <- predict(lm_model)

# Residuals from the linear regression model
residuals <- residuals(lm_model)

# Plot of residuals vs. predicted values
plot(predicted_values, residuals,
     xlab = "Predicted values", ylab = "Residuals",
     main = "Residuals vs. Predicted Values")
abline(h = 0, col = "blue")
```



Based on the Residuals vs Predicted values scatter plot we can see that there is a relatively consistent spread, and the residuals are centered around 0 indicating homoscedasticity.

```
summary(lm_model)
```

```
## 
## Call:
## lm(formula = `Healthy life expectancy at birth` ~ `Life Ladder`, 
##     data = random_sample)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -38.512  -2.011   0.062   3.247   9.348 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    38.2873     1.0537   36.34   <2e-16 ***
## `Life Ladder`   4.5719     0.1876   24.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.783 on 498 degrees of freedom
## Multiple R-squared:  0.544,  Adjusted R-squared:  0.5431 
## F-statistic: 594.1 on 1 and 498 DF,  p-value: < 2.2e-16
```

From the linear model we can see that the p value < 0.05, therefore we can reject the null hypothesis and conclude that there is a statistically significant linear relationship between or variables “Life Ladder” and “Healthy life expectancy at birth.”

Based on the linear model summary we can see that the coefficient estimate for “Life Ladder” is 4.5719 and its standard error is 0.1876.

Using this we can calculate the 95% confidence interval as (CI= coefficient estimate - (1.96 x SE), CI= coefficient estimate + (1.96 x SE))

Using the values from the summary we can conclude that the CI=(4.204204, 4.939596). This means that we are 95% confident that the true effect of “Life Ladder” on “Healthy life expectancy at birth” falls within this confidence interval.

We can also find the confidence interval using the confint function.

```
confint(lm_model)
```

```
##                   2.5 %   97.5 %
## (Intercept)   36.217015 40.35751
## `Life Ladder`  4.203381  4.94046
```

Using the confint function, we were able to confirm that our confidence interval is correct. From the confint function we are able to see that the confidence interval for our linear model is (4.203381, 4.94046) which is very close to the confindece interval we calculated.

The R^2 value is 0.544, meaning that approximately 54.4% of the variability in “Healthy life expectancy at birth” can be explained by the linear relationship with “Life Ladder”.

Conclusion

Overall, our chosen quantitative variables, “Life Ladder” and “Healthy life expectancy at birth”, from our random sample of the World Happiness Report, exhibit a linear relationship and meets all assumptions for regression, ultimately allowing us to reject our null hypothesis that there is no relationship between “Life Ladder” and “Healthy life expectancy at birth”.