Introduction to Regression Group Work

Alix Ziff, Gaea Morales, Zachary Johnson, and Yuchen Gong

Fall 2023, Version 2: Sep 21

Introduction to Regression Analysis in R

In this group work, we are going to continue working with the Human Development Index data. This time we are asking: *Does a country's human development rank predict the infant mortality rate of it's population?*

The data provided has each country's most recent ranking according to the Human Development Index. The goal will be to clean the data, make a scatterplot of the relationship, draw a line of best fit, and interpret some regression coefficients.

Clean our HDI Data

Exercise 1: Reshape the data into country-year format.

```
## replace the current dataset path with a local file path
hdi_im <- read.csv("./HDI_infant_mort.csv", stringsAsFactors = TRUE)

hdi_im <- pivot_longer(data = hdi_im, names_to = "Year", cols = 3:10, values_to = "infant_mortality")

hdi_im <- hdi_im %>%
  group_by(HDI_rank) %>%
  mutate(mean_infant_mortality_year = mean(infant_mortality))
```

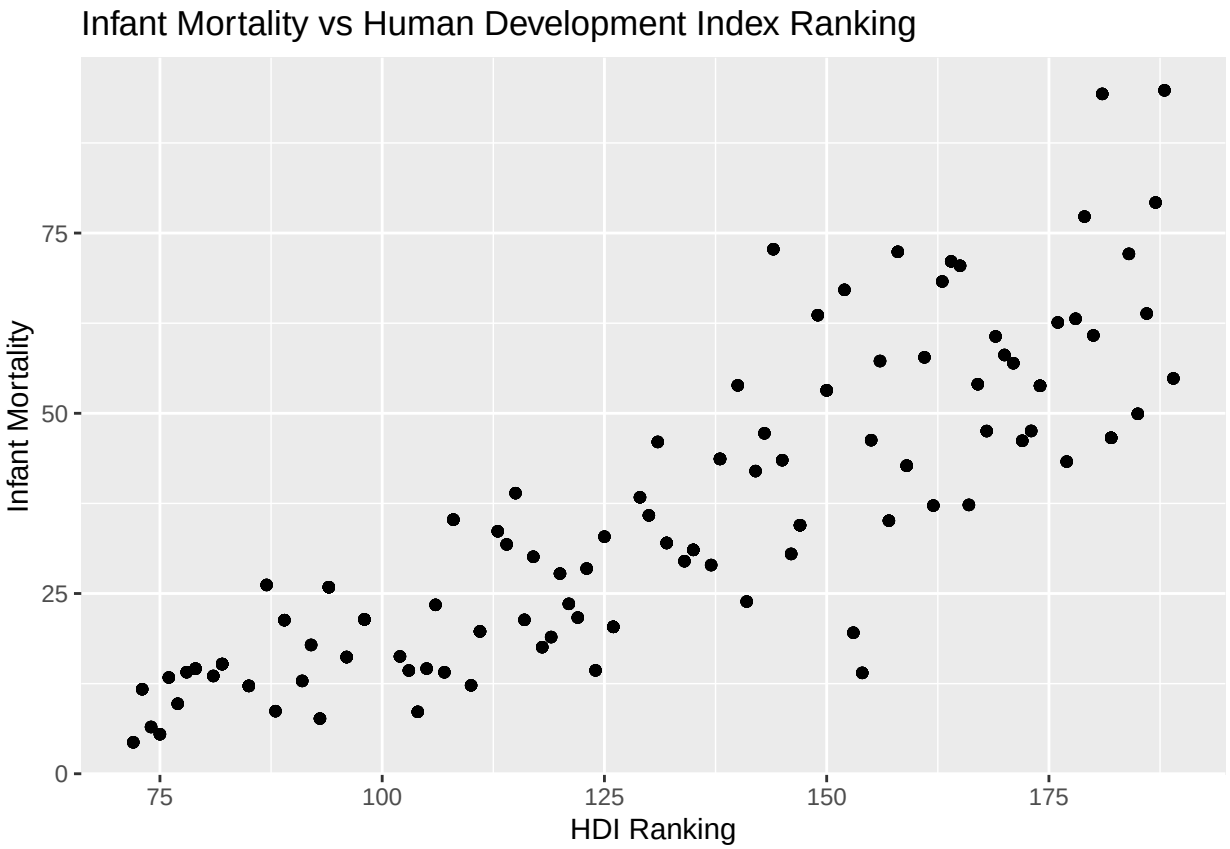
Visualize the Relationship

Exercise 2: Create a scatterplot visualizing the relationship between life expectancy and human development index ranking. Label the axes.

```
mortPlot <- ggplot(data = hdi_im, aes(y = mean_infant_mortality_year, x = HDI_rank)) +
  geom_point() +
  labs(title = "Infant Mortality vs Human Development Index Ranking") +
  ylab("Infant Mortality") +
  xlab("HDI Ranking")
```

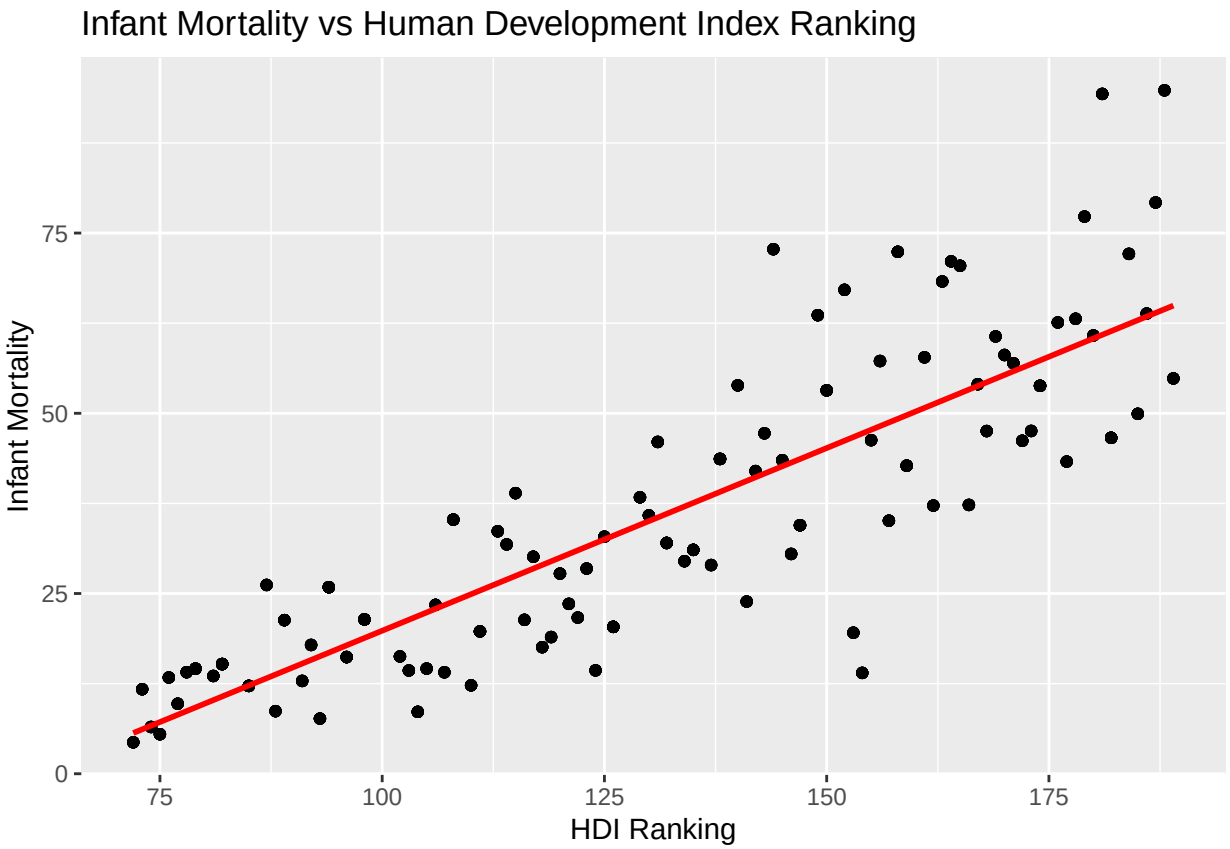
Exercise 3: If you have access to a printer, print your scatterplot. If not, copy a simplified (fewer observations) version of the scatterplot data with pen and paper.

```
mortPlot
```



Exercise 4: Draw an estimated line of best fit.

```
ggplot(data = hdi_im, aes(y = mean_infant_mortality_year, x = HDI_rank)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +  # This line adds the regression line
  labs(title = "Infant Mortality vs Human Development Index Ranking") +
  ylab("Infant Mortality") +
  xlab("HDI Ranking")
```



Interpret the Regression Results

Exercise 5: Label the predicted value, the actual value, and the error for three observations.

Bonus Exercise: Estimate the slope and the y intercept of their regression line, and explain what those mean.