SPEC REU R Resources: Applied Introduction to T-Tests, Correlation, and OLS regression – Group Work

Alix Ziff, Gaea Morales, Zachary Johnson, Claudia Salas Gimenez and Ben Graham

January 2025

In this groupwork, you will analyze real-world data to explore the relationship between GDP per capita and the Human Development Index (HDI), addressing the question: Does greater economic wealth lead to better social and human development outcomes? You will apply concepts from previous modules, including cleaning and reshaping data, creating visualizations, and performing regression analysis by fitting a linear model to interpret potential relationships between variables. Let's get started!

Initial Setup

Begin by setting up your working directory and loading the necessary libraries, along with the `HDI_GDP_pc.csv` dataset.

To better understand the HDI variables in the `HDI_GDP_pc.csv` data, refer to the UNDP Human Development Index website, where you can explore the latest HDI dataset and its measures. Additionally, the GDP per capita data was sourced from the World Development Indicators. For more information, consult the World Development Indicators data catalog. Please note that this dataset has been modified for this groupwork, so variable names and formats may differ slightly from those in the original UNDP and World Development Indicators datasets.

```
# Set working directory
#setwd("YourFolderPath")

# Load required libraries
library(tidyverse)
library(ggplot2)
library(readr)

# Load the HDI and GDP per capita data
gdp_pc <- read_csv("HDI_GDP_pc.csv")
```

Does Economic Wealth Improve Human Development?

A key question in human development research is whether greater economic wealth leads to better social and human outcomes. Wealthier countries are generally expected to provide better education, healthcare, and overall quality of life, resulting in higher HDI scores. To analyze this relationship, we will use GDP per capita as the independent variable (a measure of economic wealth) and HDI as the dependent variable.

Exercise 1: Clean and Reshape the Data

Before analyzing the data, we need to reshape it into a country-year format and ensure that GDP per capita values are stored as numeric. When importing datasets, numeric values may sometimes be misread as text,

which can cause errors during analysis. We will use the `as.numeric()` ffunction to convert these columns before reshaping the data.

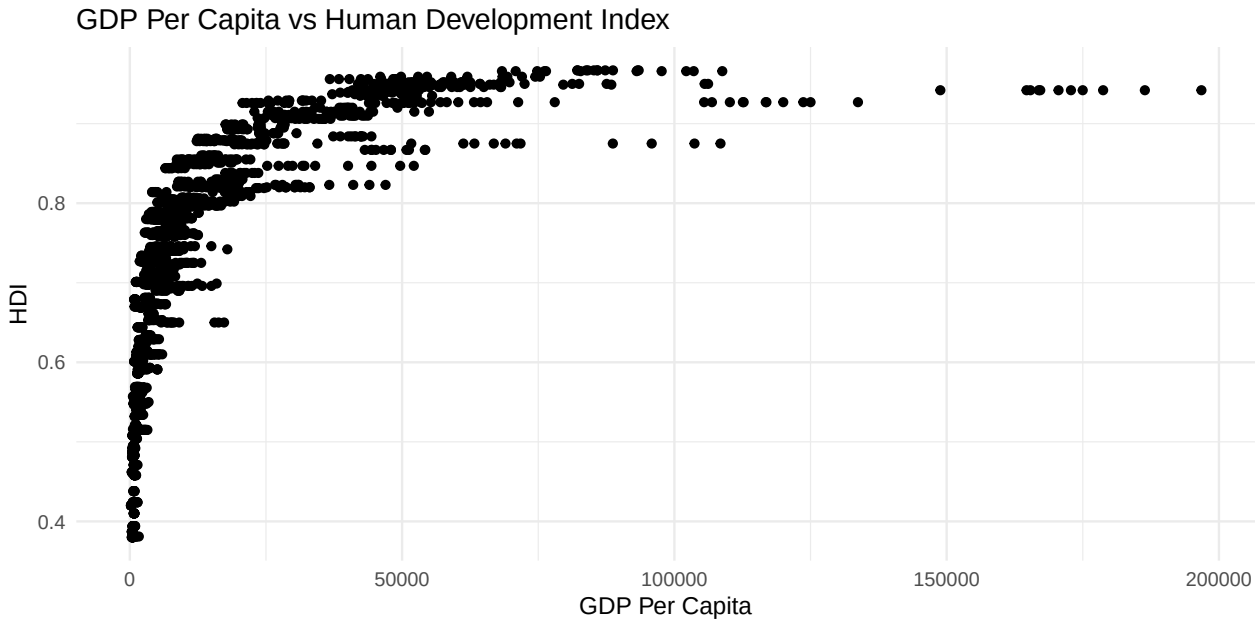
```
# Convert columns with GDP per capita (columns 5 to 15) to numeric
gdp_pc[, 5:15] <- lapply(gdp_pc[, 5:15], as.numeric)

# Reshape the data from wide to long format
gdp_pc <- pivot_longer(
  data = gdp_pc,
  names_to = "Year",
  cols = 5:15,
  values_to = "gdp_pc"
)
```

Exercise 2: Visualize Relationship Between HDI and GDP Per Capita

Now that the data is properly formatted, let's create a scatterplot to visualize the relationship between GDP per capita and HDI. Challenge yourself to customize your visualization to produce a publication-quality figure.

```
# Create the scatterplot
ggplot(data = gdp_pc, aes(x = gdp_pc, y = HDI)) +
  geom_point() +
  labs(title = "GDP Per Capita vs Human Development Index",
       x = "GDP Per Capita",
       y = "HDI") +
  theme_minimal()
```

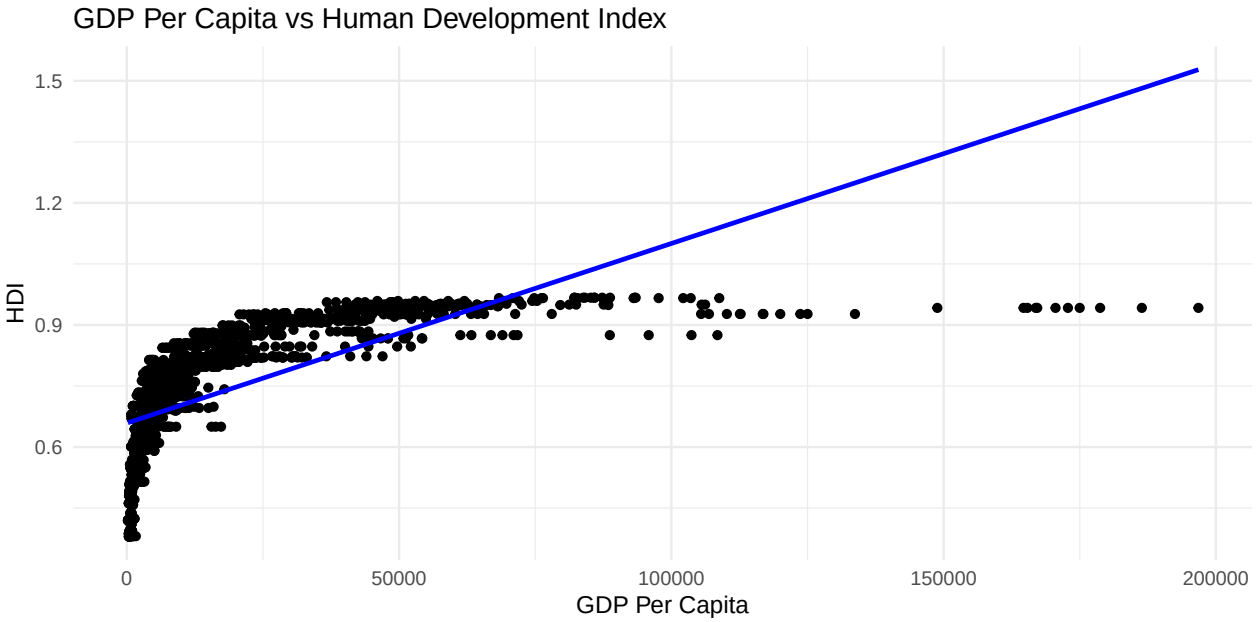



```
## The scatterplot suggests a positive relationship: countries with low GDP per capita
## exhibit lower HDI scores, whereas those with higher GDP per capita show higher HDI
## values. However, the relationship is nonlinear-after a certain threshold, further
## increases in GDP per capita result in smaller gains in HDI (diminishing returns).
```

Exercise 3: Draw an Estimated Line of Best Fit

Next, let's add a line of best fit to your scatterplot. What does this line suggest about the relationship between GDP per capita and HDI?

```
# Add a line of best fit
ggplot(data = gdp_pc, aes(x = gdp_pc, y = HDI)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "GDP Per Capita vs Human Development Index",
       x = "GDP Per Capita",
       y = "HDI") +
  theme_minimal()
```




```
## The line of best fit indicates a positive relationship between GDP per capita and HDI.
## As GDP per capita increases, HDI also tends to rise.
##
## However, given the non-linear pattern observed in the scatterplot, a simple
## linear model may not be the best fit for the data.Due to the diminishing returns
##neffect, a non-linear model (such as a logarithmic or polynomial regression) might
## better capture the true relationship between these variables.
```

Exercise 4: Interpret the Regression Results

Finally, to test this relationship, let's fit an Ordinary Least Squares (OLS) regression model. Then, interpret the summary statistics:

- Identify the slope of HDI and interpret it.
- Is the relationship between HDI rank and mean infant mortality significant? How do you know?

```
# Fit a linear model
model <- lm(HDI ~ gdp_pc, data = gdp_pc)

# Summary of the model
model_summary <- summary(model)
model_summary
```

```
##
## Call:
## lm(formula = HDI ~ gdp_pc, data = gdp_pc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58543 -0.07458  0.03795  0.08604  0.17586
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.588e-01  3.023e-03  217.89   <2e-16 ***
## gdp_pc      4.414e-06  1.094e-07   40.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1153 on 2091 degrees of freedom
##   (2846 observations deleted due to missingness)
## Multiple R-squared:  0.4377, Adjusted R-squared:  0.4374
## F-statistic:  1628 on 1 and 2091 DF,  p-value: < 2.2e-16
```

```
## The slope indicates the change in HDI for each unit increase in GDP per capita. The
## regression results suggest a slightly positive relationship between GDP per capita
## and HDI (slope = 4.414e-06). This indicates that as GDP per capita increases by one
## unit, HDI is expected to increase by approximately 0.000004414, holding other factors
## constant.
##
## However, as seen in the scatterplot, the relationship appears nonlinear, with
## diminishing returns at higher GDP levels. This suggests that beyond a certain income
## threshold, increases in GDP per capita contribute less to improvements in HDI.
```

```
## The relationship between GDP per capita and HDI is statistically significant. The
## p-value for GDP per capita is small (< 2.2e-16), indicating that we can reject the
## null hypothesis of no relationship between GDP per capita and HDI.
```