## Human Development Index and Regression

Human Development Index is a way of quantifying the development of a country by taking into account the development of the people in the country and not only economic activity and growth. The dataset below has data on the HDI rankings of countries in 2023, as well as the variables listed below:

* Life expectancy at birth in years
* Mean years of schooling
* Gross national income (GNI) per capita (2021 PPP $)
* HDI rank in 2022

**[Link to Excel file with the data](https://github.com/laurennelsen/MiscellaneousFiles/raw/refs/heads/main/HumanDevelopmentIndexData.xlsx)**

**A qr code with a white background

AI-generated content may be incorrect.**

**(Note:** The original data file can be found at [this link](https://hdr.undp.org/data-center/human-development-index#/indicies/HDI).)

## We want to create a model to predict the life expectancy at birth using the other variables we have.

## Exercise 1

## **Let be life expectancy at birth (in years) and let be mean years of schooling (in years).**

1. In the scatter plot below, does it look like there is a relationship between these two variables? (Click on the image below to view an interactive form of this plot.)

[](https://public.tableau.com/views/HDIRegression/Scatterplot?:language=en-US&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link)

1. Use Excel to find the equation of the least squares line.
2. What is the correlation coefficient?
3. Find the value and write a sentence explaining what it means in the context of this problem.

## Exercise 2

## **Let be life expectancy at birth (in years) and let be GNI per capita.**

1. In the scatter plot below, does it look like there is a relationship between these two variables? (Click on the image below to view an interactive form of this plot.)

[](https://public.tableau.com/views/HDIRegression/Scatterplot2?:language=en-US&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link)

1. Use Excel to find the equation of the least squares line.
2. What is the correlation coefficient?
3. Find the value and write a sentence explaining what it means in the context of this problem.

## Exercise 3

Now we're going to use more of our variables to try to create a better model than the least squares line we found above. Let

* life expectancy at birth,
* mean years of schooling, and
* gross national income per capita.

1. Use the Data Analysis tool in Excel to create the multiple regression model. What is the equation of the model, with each coefficient rounded to four decimal places?

1. Use the model created above to predict the life expectancy at birth for a country with mean years of schooling of 15 years and GNI per capita of .
2. Consider a country with a GNI per capita of . Using this model, what mean years of schooling would predict a life expectancy of 90 years?
3. Find the coefficient of determination () for this model and explain what that means in the context of this problem.
4. Test the overall significance of the model at a significance level of :

1. Identify the p-value for this hypothesis test.
2. What do you conclude and why?

1. In part (a) you found the regression model in the form .

Remember that is a point estimate for . Find a confidence interval for .

1. A politician claims that “years of schooling do not impact life expectancy”. Use your regression analysis to evaluate this claim.

## Exercise 4

Now we're going to use more of our variables to try to create a better model than the least squares line we found above. Let

* life expectancy at birth,
* mean years of schooling,
* gross national income per capita, and
* HDI rank from 2022.

There is one country in the dataset that does not have an HDI ranking from 2022. Remove that country from the following analysis.

1. Use the Data Analysis tool in Excel to create the multiple regression model. What is the equation of the model, with each coefficient rounded to four decimal places?
2. Find the coefficient of determination () for this model and explain what that means in the context of this problem.
3. Test the overall significance of the model at a significance level of :

1. Identify the p-value for this hypothesis test.

1. What do you conclude and why?

1. In part (a) you found the regression model in the form . Remember that is a point estimate for . Use your output in Excel to perform the following hypothesis test:

1. Identify the p-value for this hypothesis test.

1. What do you conclude and why?

## Exercise 4 (continued)

1. Remove GNI per capita from your model, and create a new regression model with the following variables:

* life expectancy at birth,
* mean years of schooling, and
* HDI rank from 2022.

1. Find the coefficient of determination () for this model and explain what that means in the context of this problem.
2. Test the overall significance of the model at a significance level of :

1. Identify the p-value for this hypothesis test.

1. What do you conclude and why?

## Exercise 5

Look back at all the models you created in the previous exercises. Which one do you think is the best model for predicting life expectancy at birth? Justify your answer using your regression analyses.

## Exercise 6

What other variable(s) do you think would be helpful for predicting life expectancy at birth? (What data would you want to have for each of these countries?