MA123 – Project 3: Statistical Modeling (Report)

Name: _____ **Score:** ____ pts.

This project relates to statistical modeling with focus on using regression techniques to model and predict health expenditure in Africa.

Resources (Make sure to review these resources before starting the project):

- Calculating mean, median, mode, standard deviation, minimum, and maximum (in Excel).
  - https://www.youtube.com/watch?v=7A8dAYeulL4
  - https://www.youtube.com/watch?v=rk09wUnW61Q
- Calculating the coefficient of variation (in Excel):
  - https://www.youtube.com/watch?v=OI1wzkuKS6Y
  - https://www.youtube.com/watch?v=kRLTxYPI3_4
- Scatter Diagram and Trendlines (in Excel)
  - https://www.youtube.com/watch?v=kLROcLFzH8o
  - https://www.youtube.com/watch?v=xp-2kvsHJ4U
  - https://www.youtube.com/watch?v=rgs57VedJLU
  - https://www.youtube.com/watch?v=aw-GluLZlWA

Background and Purpose

There are many elements that influence both population health and health expenditures, such as income level, pollution related to the level of industrialization, environmental quality, etc. Offering quality health care services should be one of the most essential objectives of governments because they can lead to improved life expectancy, social and economic welfare, etc. The deterioration of environmental quality globally has a substantial effect on what we call "healthy living". Therefore, this project is purposed on the development of regression models to predict Health Expenditure taking into consideration the effect of environmental degradation in Africa using 2019 World Bank data (www.data.worldbank.org). This project assumes current health expenditure per capita (CHEPC) measured in international dollars at purchasing power parity as a proxy for the health variable (i.e., response variable) and carbon dioxide emissions (CO2) measured in metric tons per capita as a proxy for the environmental variable (i.e., predictor variable).

Your Task (Use Microsoft Excel throughout):

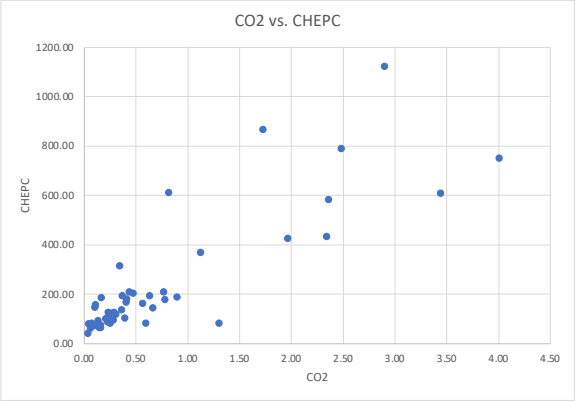
1. (15 points) Compute the mean, median, standard deviation, coefficient of variation, minimum, and maximum values of CHEPC and CO2 (Round all values to two decimal places).

Table 1: Mean, Median, Standard Deviation, Coefficient of Variation, Minimum, and Maximum of CHEPC and CO2

Measure	CHEPC (in International $)	CO2 (in Metric Tons per Capita)
Mean	238.58	.76
Median	145.27	.37
Standard Deviation	244.28	.95
Coefficient of Variation	1.02	1.25
Minimum	40.61	.04
Maximum	1122.14	4.01

2. (15 points) Draw a scatter diagram showing the relationship between CHEPC (response/dependent variable) and CO2 (explanatory/independent variable). From the scatter diagram, briefly describe the relationship between CHEPC and CO2.

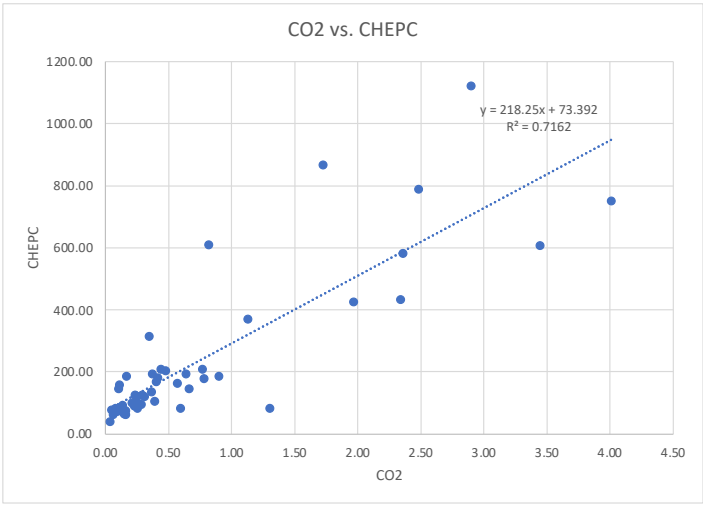
As CO2 increases, CHEPC increases as well.



3. (5 points) Compute the linear correlation coefficient (round the coefficient to 3 decimal places) and use it to describe the strength and direction of the relationship between CHEPC and CO2.

Linear correlation coefficient is 0.85. This is an indication that there is a very strong and positive linear relationship between CHEPC and CO2.

4. (15 points) Construct a scatter diagram showing the fitted linear regression line (line of best fit), fitted (estimated) model, and coefficient of determination (r_1^2).



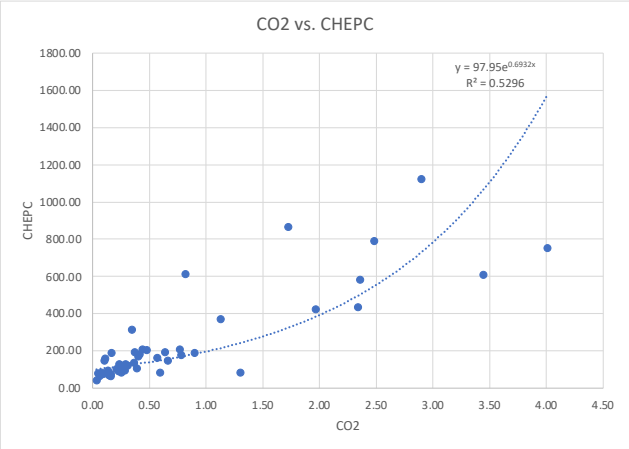
5. (5 points) From (4), write down the equation of the line of best fit and the r_1^2. Interpret the r_1^2.

$$y = 218.25x + 73.39$$

$$r_1^2 = 0.7162$$

This can be interpreted to mean the 71.62% of the variability in CHEPC can be explained by the fitted regression line.

6. (10 points) Construct a scatter diagram showing the fitted exponential regression curve (curve of best fit), fitted (estimated) model, and coefficient of determination (r_2^2).



7. (5 points) From (6), write down the equation of the curve of best fit and the r_2^2. Interpret the r_2^2.

$$y = 97.95 \times e^{0.6932x}$$

$$r_2^2 = 0.5296$$

This can be interpreted to mean the 52.96% of the variability in CHEPC can be explained by the fitted exponential regression curve.

8. (10 points) Between the fitted linear and exponential regression models, which one provides a better fit to the data? Use the coefficient of determinations (r_1^2 and r_2^2) to justify your choice.

The linear regression model provides a better fit to the data sincer_1^2 is greater than r_2^2.

9. (10 points) Use the fitted regression model that provides a better fit to the data from (8) to determine the expected CHEPC (round the value to two decimal places) when CO2 is 3.850 (i.e., predict the value of CHEPC when CO2 is 3.850). Interpret the results.

The expected value of CHEPC when CO2 is 3.850 is 913.65. When C02 increases to 3.850 metric tons per capita, the CHEPC increases to $913.65.

10. (10 points) Assume the observed CHEPC for a given CO2 of 4.01 is 750.45. Calculate the residual of this observation. Briefly comment on the residual value. **Hint:** A residual is the difference between an observed value and a predicted value in regression analysis.

When you put 4.01 into the linear equation, the CHEPC is 948.57. The residual between the two values is -$198.12.

$$Observed\ value - predicted\ value = 750.45 - 948.57 = -198.12$$