

|  |  |
| --- | --- |
| **MODULE NAME AND CODE** | DAT7301: DATA ANALYSIS AND VISUALIZATION |
| **ASSESSMENT TITLE** | HOMEWORK 7 |
| **MARKING TUTOR** | DR EUGEN HARINDA |
| **STUDENT NUMBER** | 2423918 |
| **STUDENT NAME** | NWEKORI, FRIDAY EGEDE |
| **PROGRAMME NAME** | MSc. SOFTWARE ENGINEERING |
| **DATE OF SUBMISSION:** | 04/05/2025 |
| **WORD COUNT** | **704 WORDS** |

**Question (1).**

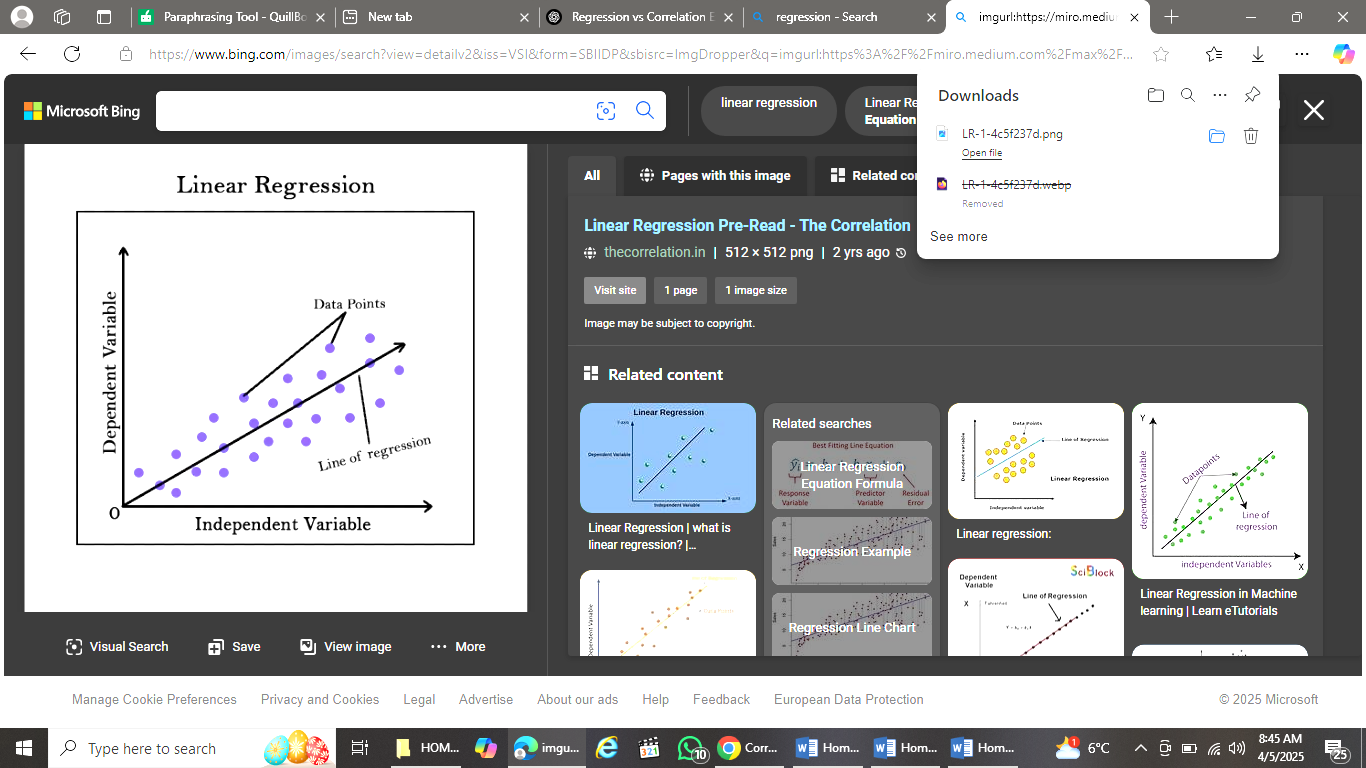
Correlation, also called correlation analysis, is a word that refers to the association or relationship between two or more quantitative variables. Gogtay and Thatte, 2017 stated that the main notion of this analysis is that the quantitative variables have a straight-line (linear) relationship. Similar to measures of association for binary variables, it assesses the "strength" or "extent" of an association between variables, as well as the direction.

Pearson's correlation coefficient, or "r," is used when both variables have a normal distribution. If not, we employ the non-parametric Spearman's correlation coefficient rho (ρ), which is more robust to outliers than the Pearson's correlation coefficient "r." Regression is used in combination with correlation analysis, which is rarely employed alone (Figer et al., 2017).

**Questions (2).**

We can describe regression as a statistical method we use when we want to predict the value of one variable based on the value of another. It is used to estimate the relationship between a dependent variable (the outcome) and one or more independent variables (the predictors) (Montgomery et al., 2021). For example, suppose a scholar wants to predict students’ examination scores (dependent variable) based on the number of hours they studied (independent variable). We could use simple linear regression to solve the problem.

Exam Score = 40 + 5 x (Hours Studied). This equation shows us that for each additional hour studied, the student’s examination score will increase by points, beginning from a base score of 40.



Graph showing regression analysis

**Question (3).**

Correlation analysis calculates the correlation coefficient and tests its significance, while regression analysis expresses the relationship as an equation and moves into prediction (Zhang et al., 2024).

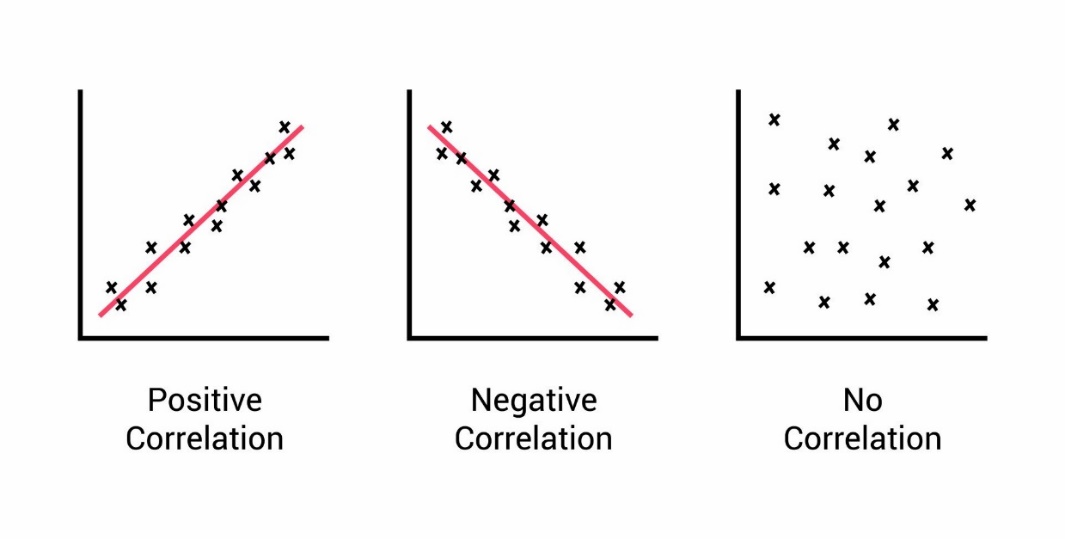
Table 1: Differences between regression and correlation

|  |  |  |
| --- | --- | --- |
| **Feature** | **Regression** | **Correlation** |
| **Purpose** | Predict outcomes | Measure strength and direction |
| **Variables** | One dependent, one/more independent | Two variables, no direction |
| **Direction** | Yes | No (symmetric) |
| **Causation** | Suggests but doesn’t confirm | Does not imply causation |

**Question (4).**

The end result of a correlation analysis is a Correlation coefficient whose values range from -1 to +1.

* +1: this means perfect positive correlation.
* - 1: perfect negative correlation
* 0: no correlation.



A scatter plot indicating the strength and direction of the correlation between the co-variables.

**Question (5).**

When performing regression, the intercept, which is denoted by βo or a, is the predicted value of the dependent variable if all independent variables are zero. Given a regression equation: Y = a + bX (where Y = what you are predicting, a = intercept, b = slope, X = independent variable). The intercept (a) is the value of Y when X = 0 (Wooldridge, 2020).

**Question (6).**

To choose between correlation analysis and regression analysis relies on your goal and the type of relationship you are trying to study. We can use correlation analysis if we want to do the following:

* To measure the strength and direction of the relationship between two variables.
* If neither variable is dependent on the other – you are treating them as equal.
* When you are performing a preliminary analysis to decide whether to proceed with regression or not.

Let us take for example, you want to know if there is a relationship that exists between sleep duration and academic performance, we can use correlation to see how strongly they are related. If, for instance, the output is r =0.73, this shows how much and in what direction two variables go together, and we know that correlation does not imply causation (Field, 2022).

**Question (7).**

A scatter plot is a graph that displays individual data points for two continuous variables. While a regression line (also called a line of best fit) is a mathematical model drawn through the scatter plot that summarizes the relationship between the variables (Montgomery, 2021).

Let us get the difference clearer using summary table

Table 2: Scatter plot vs. regression line

|  |  |  |
| --- | --- | --- |
| **Feature** | **Scatter plot** | **Regression line** |
| What is shows | Raw data points | Trend or prediction line |
| Purpose | Visualize relationship | Summarize/model relationship |
| Data used | Actual observations | Predicted values from regression |
| What it indicates | Direction form, strength, outliers | Central tendency, prediction formula |
| How it looks like | Dots scattered across a graph | A straight (or curved) line through them. |

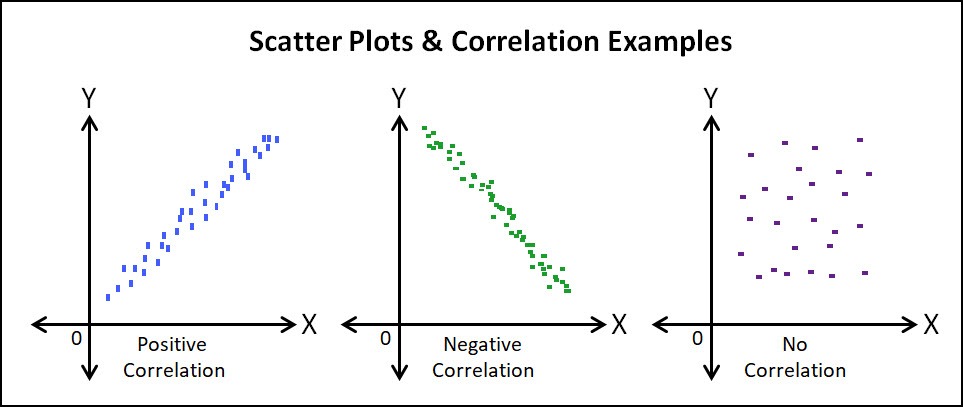
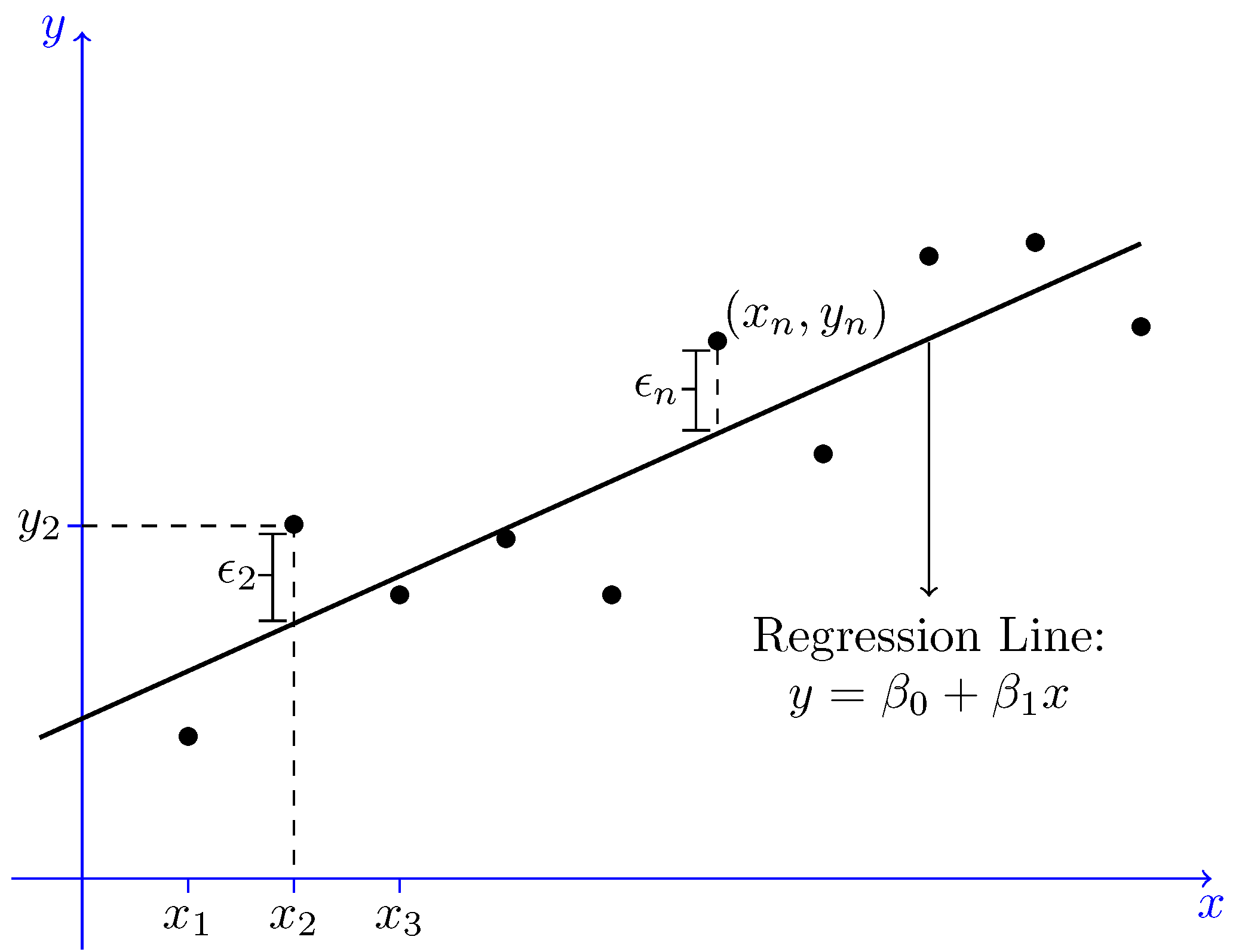


Diagram showing scatter plots



Graph showing regression line

**References**

1. Gogtay, N.J. and Thatte, U.M., 2017. Principles of correlation analysis. *Journal of the Association of Physicians of India*, *65*(3), pp.78-81. <https://www.kem.edu/wp-content/uploads/2012/06/9-Principles_of_correlation-1.pdf>.
2. Figer, B. H., Chaturvedi, M., Thaker, S. J., Gogtay, N. J. and Thatte, U. M. (2017). A comparative study of the informed consent process with or without audio-visual recording. *National Medical Journal of Industry,* 2017; in press.
3. Montgomery, D. C., Peck, E. A. and Vining, G. G. (2021). Introduction to Linear Regression Analysis. Wiley.
4. Zhang, Y., Chen, L. and Wu, Q. (2024). Predictive Analytics in Retail: A Machine Learning Approach. Journal of Artificial Intelligence Research, 68(3), 110-126.
5. Wooldridge, J. M. (2020). Introductory Econometrics: A Modern Approach (7th ed.). Cengage Learning.
6. Field, A. (2022). Discovering Statistics Using IBM SPSS Statistics (6th ed.). SAGE Publications.