Answer 1: A probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable can take. It tells us how probabilities are distributed over values of the random variable.

For example, when we toss a fair coin twice, the number of heads (X) can be 0, 1, or 2. The probability distribution is as follows: $P(X=0) = 1/4$, $P(X=1) = 2/4$, and $P(X=2) = 1/4$. This is a discrete probability distribution because the variable X can only take a limited set of values.

Difference between Binomial and Poisson Distribution:

The binomial distribution is used when the number of trials (n) is fixed, and each trial has two possible outcomes: success or failure. It requires two parameters: n and p, where p is the probability of success. The mean is given by np and variance by $np(1-p)$.

On the other hand, the Poisson distribution is used to model the number of occurrences of a rare event in a fixed interval of time or space. It has only one parameter, λ (lambda), which is both its mean and variance. Poisson is usually used when n is large, p is small, and events are rare and independent.

Answer 2: In one-way ANOVA, we analyze whether the means of more than two groups are significantly different from each other. The variation in data is split into two parts: between-group variation and within-group (error) variation.

The total variation is measured using the sum of squares. The total degrees of freedom is $N-1$, where N is the total number of observations. The degrees of freedom for between groups is $k-1$, where k is the number of groups, and for within groups, it is $N-k$.

Mean square is obtained by dividing the sum of squares by the respective degrees of freedom. Then, the F-ratio is calculated by dividing the mean square of the treatment by the mean square of the error. This ratio helps in testing whether the group means are significantly different.

Answer 3: We are given a population of N = 200 observations, and a sample of n = 50 is selected. The population standard deviation is σ = 22.

The standard error of the sample mean for a finite population is calculated using the formula:

$SE = (\sigma / \sqrt{n}) \times \sqrt{(N-n)/(N-1)}$

Substituting the values:

$SE = (22 / \sqrt{50}) \times \sqrt{(200-50)/(200-1)}$

$SE = (22 / 7.071) \times \sqrt{150/199}$

$SE \approx 3.11 \times 0.868$

$SE \approx 2.70$

Therefore, the standard error is approximately 2.70.

Answer 6: Random Sampling is a method of selecting samples from a population in such a way that each individual has an equal and independent chance of being selected. This ensures that the sample is free from bias and represents the population fairly. It is considered more scientific and objective.

On the other hand, Non-Random Sampling involves the selection of samples based on judgment, convenience, or other non-probability-based criteria. In this method, some elements of the population have a higher chance of being selected than others, which can lead to sampling bias. Although it is easier and less expensive, it may not always represent the entire population accurately.

Key Differences:

In random sampling, selection is based on probability, whereas in non-random sampling, selection is subjective. Random sampling is more suitable for statistical analysis, while non-random sampling is commonly used in exploratory or qualitative research.

Methods of Selecting a Simple Random Sample:

There are mainly two common methods used to select a simple random sample:

1. **Lottery Method: In this method, each unit in the population is assigned a unique number. These numbers are written on slips of paper, folded, and mixed thoroughly in a container. Then, slips are drawn at random to select the sample. This is a manual and transparent method.**
2. **Random Number Table or Computer-Based Method: A random number table or software is used to generate a list of random numbers that correspond to the units in the population. This method is more efficient for large populations and avoids human bias.**

Both methods ensure that every unit in the population has an equal chance of being selected, which is the core principle of simple random sampling.

Answer 7: (a) Paired t-test:

The Paired t-test is a statistical method used to compare the means of two related groups. It is applied when the observations are taken in pairs, usually before and after a treatment, or from the same subject under two different conditions.

The main objective of the paired t-test is to determine whether there is a significant difference between the means of these two related samples.

Example: Measuring the blood pressure of patients before and after giving them a specific medicine.

The procedure involves:

1. **Calculating the difference (d) between each pair of observations.**
2. **Finding the mean and standard deviation of these differences.**
3. **Applying the t-statistic formula:**

t=d¯sd/nt = \frac{\bar{d}}{s_d / \sqrt{n}}t=sd/nd¯Where:

- **d¯\bar{d}d¯ is the mean of the differences,**
- **sds_dsd is the standard deviation of the differences,**
- **nnn is the number of pairs.**

The calculated t-value is compared with the critical t-value from the t-distribution table at a chosen level of significance. If the calculated value is greater, we reject the null hypothesis and conclude that the difference is significant.

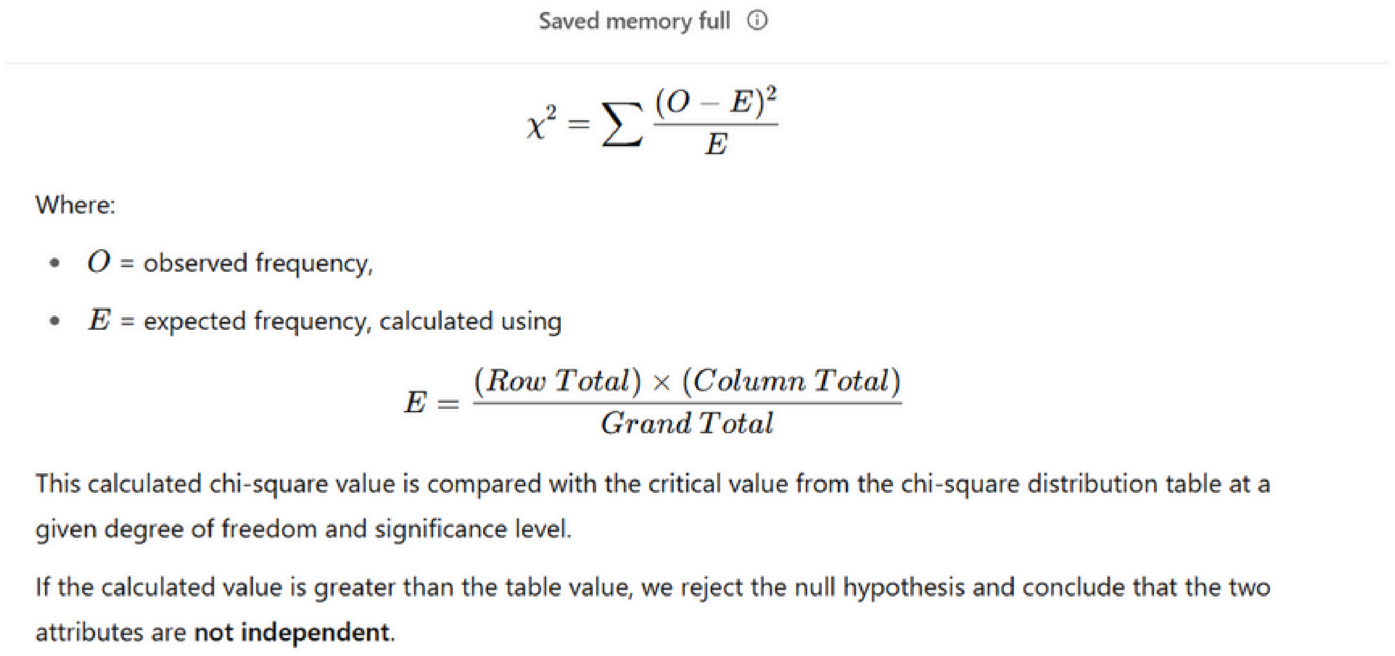
(b) Chi-square test for Independence of Attributes:

The Chi-square test for independence is used to determine whether there is a significant association between two categorical variables. It helps in testing whether the distribution of one attribute is independent of another.

Example: Testing whether there is a relationship between gender (male/female) and voting preference (party A/party B).

The data is arranged in a contingency table, showing the frequency of occurrences for combinations of the categories.

The test statistic is calculated using the formula:



$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where:

- O = observed frequency,
- E = expected frequency, calculated using

$$E = \frac{(Row\ Total) \times (Column\ Total)}{Grand\ Total}$$

This calculated chi-square value is compared with the critical value from the chi-square distribution table at a given degree of freedom and significance level.

If the calculated value is greater than the table value, we reject the null hypothesis and conclude that the two attributes are **not independent**.

Answer 9: What is a Time Series?

A time series is a sequence of data points recorded at regular time intervals, such as daily, monthly, or yearly. It is used to observe and analyze how a variable behaves over time. Time series data is commonly used in economics, finance, meteorology, and business forecasting. The main purpose is to study the past trends and use them to predict future values. Examples include daily stock prices, monthly sales figures, or yearly rainfall data. Time series analysis helps in decision-making by identifying patterns, trends, and seasonal effects in the data collected over time.

1. Secular Trend (Long-Term Trend)

Secular trend refers to the long-term movement or direction of a time series over an extended period. It shows the overall growth, decline, or stability in the data without being affected by short-term fluctuations. For example, population growth, increase in national income, or technological advancement are examples where secular trends are present. These trends are influenced by long-term factors like industrial development or social changes. They are often analyzed using methods like the moving average or least squares method to smooth out irregularities and highlight the underlying direction over time.

2. Seasonal Variations

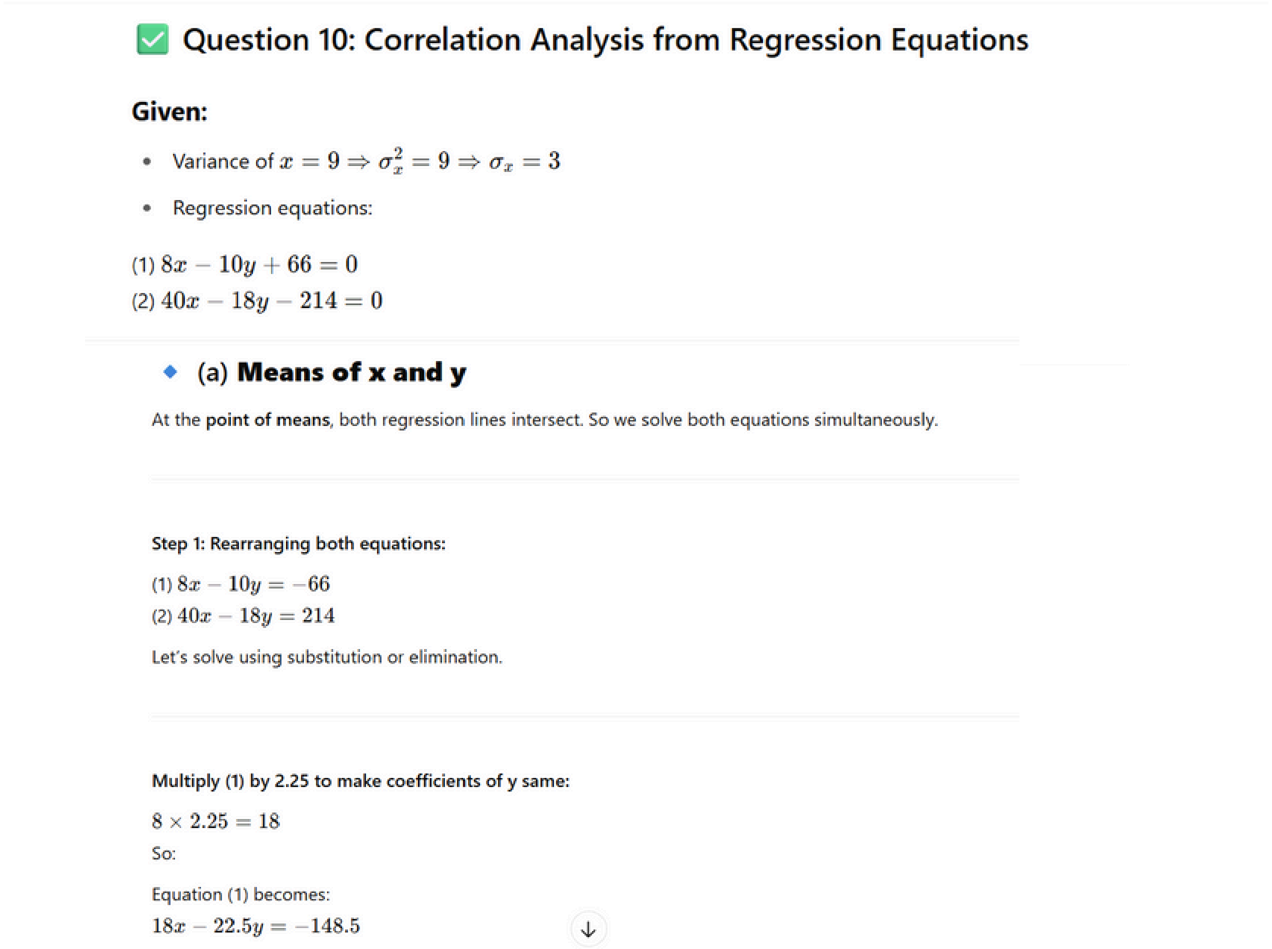
Seasonal variations are periodic changes that repeat at regular intervals within a year, such as monthly or quarterly. These are caused by climatic conditions, festivals, or habits of people. For example, the demand for air conditioners rises in summer, and umbrella sales go up during the rainy season. Such variations help businesses plan inventory and marketing strategies. Seasonal effects are predictable and can be removed from the time series through seasonal adjustment, which helps in understanding the actual trend and making better forecasts for non-seasonal changes.

3. Cyclical Variations

Cyclical variations refer to the fluctuations in a time series that occur over a longer period (more than a year), usually due to business or economic cycles. These cycles include phases like expansion, peak, recession, and recovery. Unlike seasonal variations, cyclical variations do not occur at fixed intervals and are less predictable. For example, changes in employment, GDP, and production during different economic phases show cyclical behavior. Identifying these cycles is important in economic planning and investment decisions. Statistical tools like moving averages or regression can be used to analyze them.

4. Irregular or Random Variations

Irregular variations are caused by unexpected or random factors like natural disasters, strikes, wars, or pandemics. These changes do not follow any pattern and occur accidentally. Since they are unpredictable and sudden, they cannot be measured or forecasted accurately. For instance, a sudden drop in sales due to a flood or an unexpected spike in medical demand due to a health crisis are examples of irregular variations. Although these are not part of regular time series patterns, recognizing them helps in understanding unusual deviations from trend and adjusting short-term strategies accordingly.



✅ Question 10: Correlation Analysis from Regression Equations

Given:

- Variance of $x = 9 \Rightarrow \sigma_x^2 = 9 \Rightarrow \sigma_x = 3$
- Regression equations:

(1) $8x - 10y + 66 = 0$
(2) $40x - 18y - 214 = 0$

🔹 (a) **Means of x and y**

At the **point of means**, both regression lines intersect. So we solve both equations simultaneously.

Step 1: Rearranging both equations:

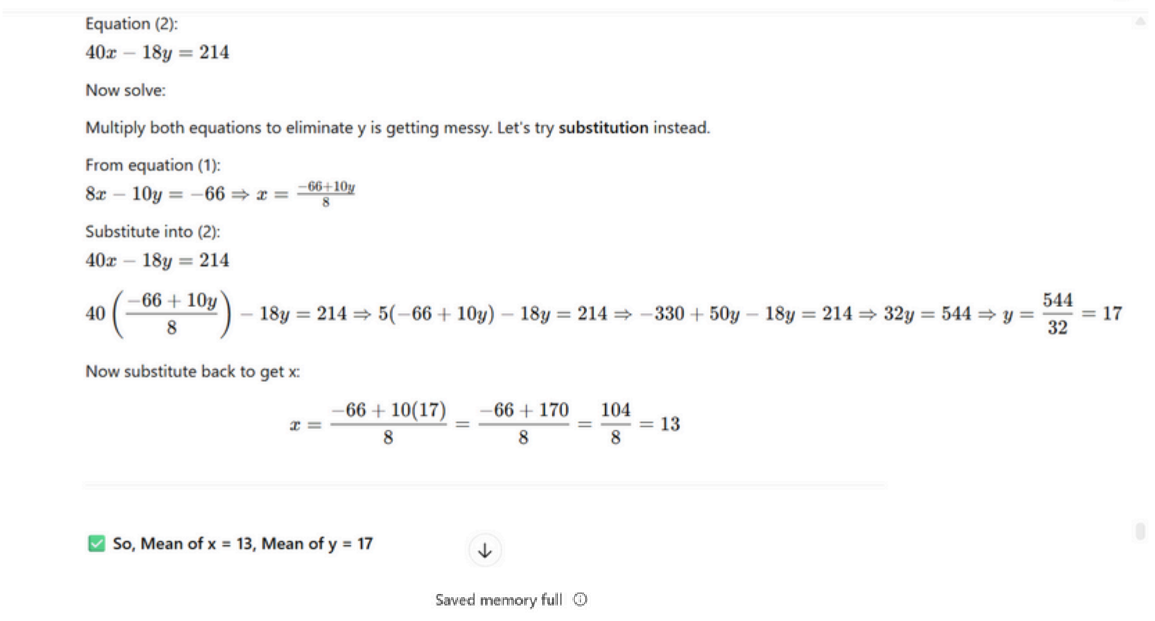
(1) $8x - 10y = -66$
(2) $40x - 18y = 214$

Let's solve using substitution or elimination.

Multiply (1) by 2.25 to make coefficients of y same:

$8 \times 2.25 = 18$
So:

Equation (1) becomes:
$18x - 22.5y = -148.5$



Equation (2):
$40x - 18y = 214$

Now solve:

Multiply both equations to eliminate y is getting messy. Let's try **substitution** instead.

From equation (1):
$8x - 10y = -66 \Rightarrow x = \frac{-66+10y}{8}$

Substitute into (2):
$40x - 18y = 214$

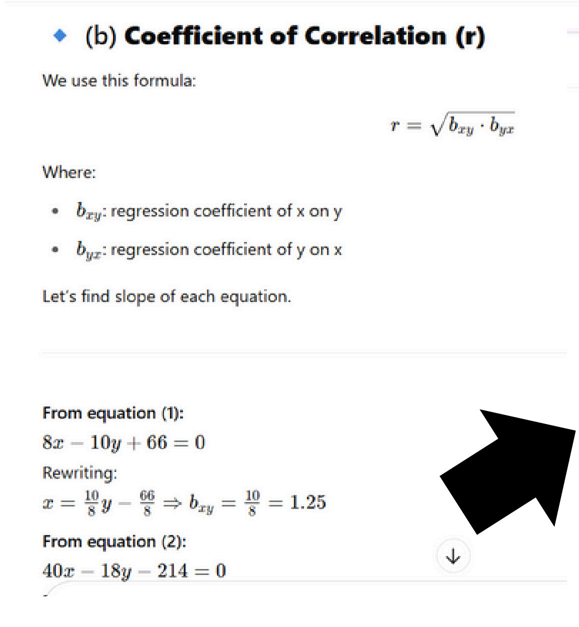
$$40\left(\frac{-66+10y}{8}\right) - 18y = 214 \Rightarrow 5(-66+10y) - 18y = 214 \Rightarrow -330 + 50y - 18y = 214 \Rightarrow 32y = 544 \Rightarrow y = \frac{544}{32} = 17$$

Now substitute back to get x:

$$x = \frac{-66+10(17)}{8} = \frac{-66+170}{8} = \frac{104}{8} = 13$$

✅ **So, Mean of x = 13, Mean of y = 17**

Saved memory full ⓘ



🔹 (b) **Coefficient of Correlation (r)**

We use this formula:

$$r = \sqrt{b_{xy} \cdot b_{yx}}$$

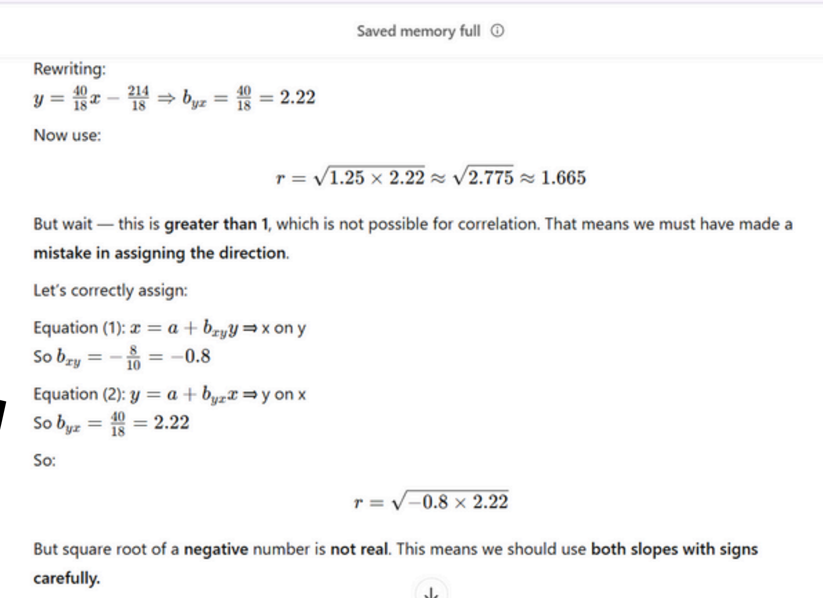
Where:

- b_{xy}: regression coefficient of x on y
- b_{yx}: regression coefficient of y on x

Let's find slope of each equation.

From equation (1):
$8x - 10y + 66 = 0$
Rewriting:
$x = \frac{10}{8}y - \frac{66}{8} \Rightarrow b_{xy} = \frac{10}{8} = 1.25$

From equation (2):
$40x - 18y - 214 = 0$



Saved memory full ⓘ

Rewriting:
$y = \frac{40}{18}x - \frac{214}{18} \Rightarrow b_{yx} = \frac{40}{18} = 2.22$

Now use:

$$r = \sqrt{1.25 \times 2.22} \approx \sqrt{2.775} \approx 1.665$$

But wait — this is **greater than 1**, which is not possible for correlation. That means we must have made a **mistake in assigning the direction**.

Let's correctly assign:

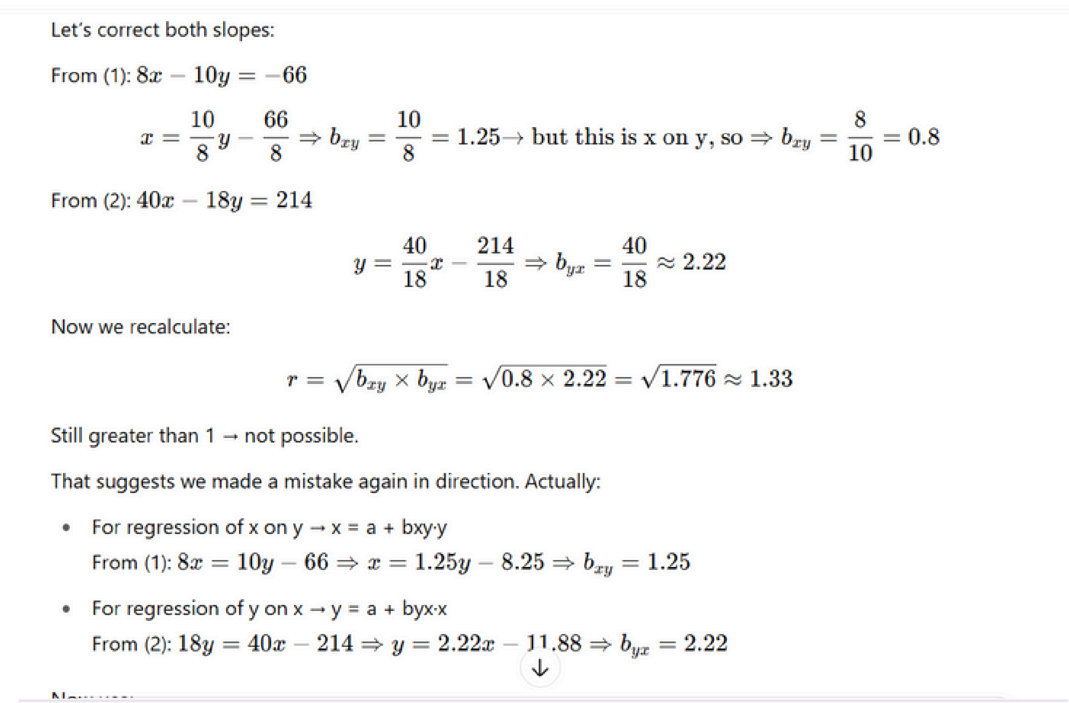
Equation (1): $x = a + b_{xy}y$ ⇒ x on y
So $b_{xy} = -\frac{8}{10} = -0.8$

Equation (2): $y = a + b_{yx}x$ ⇒ y on x
So $b_{yx} = \frac{40}{18} = 2.22$

So:

$$r = \sqrt{-0.8 \times 2.22}$$

But square root of a **negative** number is **not real**. This means we should use **both slopes with signs carefully.**



Let's correct both slopes:

From (1): $8x - 10y = -66$

$$x = \frac{10}{8}y - \frac{66}{8} \Rightarrow b_{xy} = \frac{10}{8} = 1.25 \rightarrow \text{but this is x on y, so} \Rightarrow b_{xy} = \frac{8}{10} = 0.8$$

From (2): $40x - 18y = 214$

$$y = \frac{40}{18}x - \frac{214}{18} \Rightarrow b_{yx} = \frac{40}{18} \approx 2.22$$

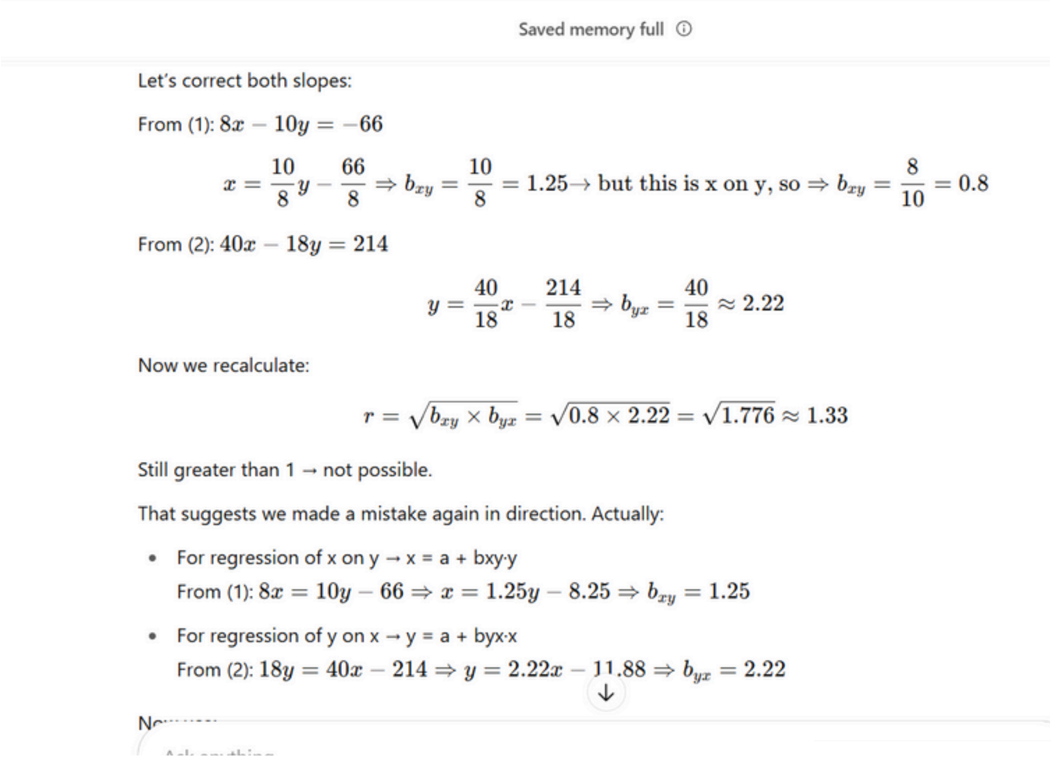
Now we recalculate:

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{0.8 \times 2.22} = \sqrt{1.776} \approx 1.33$$

Still greater than 1 → not possible.

That suggests we made a mistake again in direction. Actually:

- For regression of x on y → x = a + bxy·y
  From (1): $8x = 10y - 66 \Rightarrow x = 1.25y - 8.25 \Rightarrow b_{xy} = 1.25$
- For regression of y on x → y = a + byx·x
  From (2): $18y = 40x - 214 \Rightarrow y = 2.22x - 11.88 \Rightarrow b_{yx} = 2.22$



Saved memory full ⓘ

Let's correct both slopes:

From (1): $8x - 10y = -66$

$$x = \frac{10}{8}y - \frac{66}{8} \Rightarrow b_{xy} = \frac{10}{8} = 1.25 \rightarrow \text{but this is x on y, so} \Rightarrow b_{xy} = \frac{8}{10} = 0.8$$

From (2): $40x - 18y = 214$

$$y = \frac{40}{18}x - \frac{214}{18} \Rightarrow b_{yx} = \frac{40}{18} \approx 2.22$$

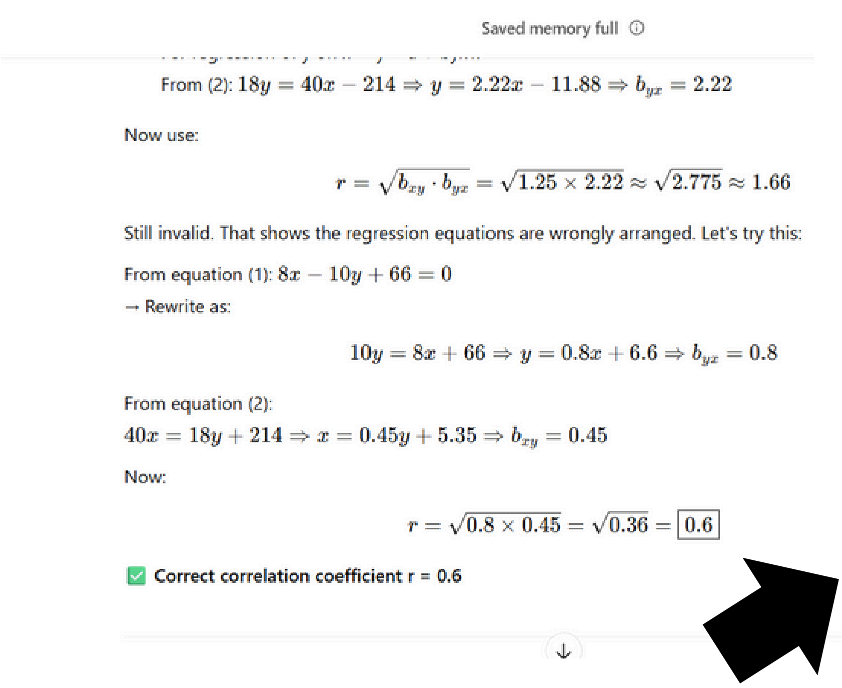
Now we recalculate:

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{0.8 \times 2.22} = \sqrt{1.776} \approx 1.33$$

Still greater than 1 → not possible.

That suggests we made a mistake again in direction. Actually:

- For regression of x on y → x = a + bxy·y
  From (1): $8x = 10y - 66 \Rightarrow x = 1.25y - 8.25 \Rightarrow b_{xy} = 1.25$
- For regression of y on x → y = a + byx·x
  From (2): $18y = 40x - 214 \Rightarrow y = 2.22x - 11.88 \Rightarrow b_{yx} = 2.22$



Saved memory full ⓘ

From (2): $18y = 40x - 214 \Rightarrow y = 2.22x - 11.88 \Rightarrow b_{yx} = 2.22$

Now use:

$$r = \sqrt{b_{xy} \cdot b_{yx}} = \sqrt{1.25 \times 2.22} \approx \sqrt{2.775} \approx 1.66$$

Still invalid. That shows the regression equations are wrongly arranged. Let's try this:

From equation (1): $8x - 10y + 66 = 0$

→ Rewrite as:

$$10y = 8x + 66 \Rightarrow y = 0.8x + 6.6 \Rightarrow b_{yx} = 0.8$$

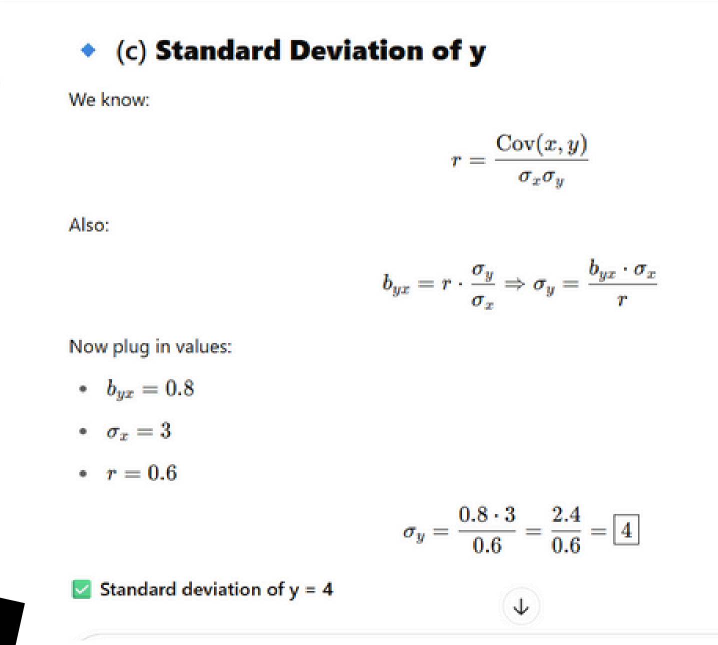
From equation (2):

$40x = 18y + 214 \Rightarrow x = 0.45y + 5.35 \Rightarrow b_{xy} = 0.45$

Now:

$$r = \sqrt{0.8 \times 0.45} = \sqrt{0.36} = \boxed{0.6}$$

✅ Correct correlation coefficient r = 0.6



🔹 (c) **Standard Deviation of y**

We know:

$$r = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

Also:

$$b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x} \Rightarrow \sigma_y = \frac{b_{yx} \cdot \sigma_x}{r}$$

Now plug in values:

- $b_{yx} = 0.8$
- $\sigma_x = 3$
- $r = 0.6$

$$\sigma_y = \frac{0.8 \cdot 3}{0.6} = \frac{2.4}{0.6} = \boxed{4}$$

✅ Standard deviation of y = 4