IB MATHEMATICS AI HL
UNIT 4: STATISTICS

Data Collection, Validity & Non-Linear Regression

2

Instructions to Candidates

- This question booklet contains **15 questions**.
- The paper targets **AHL** syllabus components 4.12 and 4.13.
- Answer all questions, showing all step-by-step working clearly.

2

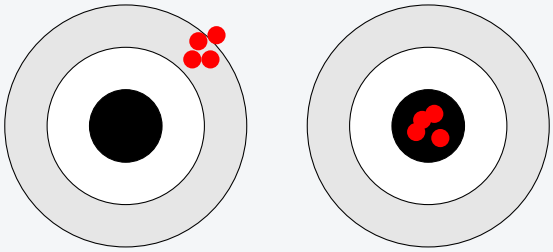
Difficulty Progression

- **Questions 1 - 5 (Easy):** Reliability vs Validity, sampling methods, R^2 values, and calculating basic residuals.
- **Questions 6 - 10 (Medium):** Exponential/Power regression via GDC, sum of squared residuals (SS_{res}), and survey bias.
- **Questions 11 - 15 (Hard):** Sine regression for seasonal data, comparing models using SS_{res}, and advanced test-retest validity analysis.

SECTION A: EASY (Fundamentals)

Question 1 (4 Marks)

he concepts of *Reliability* and *Validity* in data collection are represented by the two target boards below.



Target A **Target B**

State which target represents a data collection method that is "Reliable but NOT Valid", and explain your reasoning.

Question 2 (4 Marks)

researcher wishes to survey the students in a high school. The school has 200 freshmen, 300 sophomores, 400 juniors, and 100 seniors. The researcher selects exactly 20 freshmen, 30 sophomores, 40 juniors, and 10 seniors at random.
Name the specific sampling method used, and state one advantage of this method over simple random sampling.

Question 3 (4 Marks)

linear regression model $y = 3.2x + 1.5$ is used to model a dataset. One of the raw data points is $(4, 12)$.
Calculate the exact value of the residual for this data point.

CG50 Tip: The Coefficient of Determination (R^2)

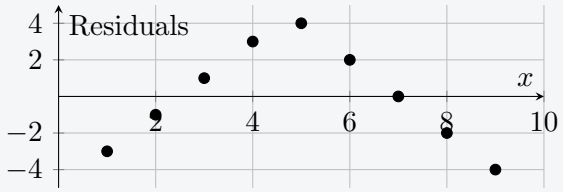
hen performing any regression (`STAT` → `CALC` → `REG`), your calculator will output an r^2 or R^2 value. This is the **Coefficient of Determination**. It tells you the percentage of the variation in the y-variable that is completely explained by your model!

Question 4 (4 Marks)

student performs a quadratic regression on a dataset and finds $R^2 = 0.942$.
State what this R^2 value means in the context of the variables x and y.

Question 5 (5 Marks)

elow is a residual plot for a linear regression model applied to a dataset.



Based on the distinct pattern in the residual plot, explain why a linear model is inappropriate for this data, and suggest a better regression model.

SECTION B: MEDIUM (Application & Modelling)

Question 6 (5 Marks)

he table below shows the population P of a bacteria colony after t hours.

Time (t hours)	1	2	3	4	5
Population (P)	15	48	130	410	1205

Using your GDC, find the exponential regression model $P(t) = a \times b^t$, giving a and b to 3 significant figures.

Question 7 (6 Marks)

sing the same bacteria data from Question 6, a student decides to try a power regression model $P(t) = a \times t^k$.
(a) Find the power regression equation using your GDC. **[2 marks]**
(b) By comparing the R^2 values of both the exponential and power models, state which model is a better fit for the data. Justify your answer. **[4 marks]**

Question 8 (6 Marks)

survey asks the question: *"Given the recent horrific increase in violent crime, do you support the mayor's new highly effective policing policy?"*
Identify two distinct types of bias present in the wording of this question, and rewrite the question to make it valid and unbiased.

CG50 Tip: Sum of Squared Residuals (SS_{res})

fter running a regression on the CG50, you can instantly find the SS_{res} without calculating residuals one by one. In `Run-Matrix`, press `VARS` → `STAT (F3)` → `Resid (F4)`. This gives you the list of residuals. To square and sum them, use `Sum List Ans²`!

Question 9 (6 Marks)

dataset has 4 points: $(1, 3), (2, 5), (3, 9), (4, 15)$.
A quadratic model $y = x^2 - x + 3$ is proposed.
Calculate the exact residual for each point, and hence calculate the Sum of Squared Residuals (SS_{res}) for this model.

Question 10 (6 Marks)

o ensure a new psychological test is reliable, a researcher gives the exact same test to the same group of students two weeks apart. They then calculate the Pearson correlation coefficient (r) between the two sets of scores and find $r = 0.88$.

State the specific name for this type of reliability testing, and conclude whether the test is reliable.

SECTION C: HARD (Synthesis & Proof)

Question 11 (7 Marks)

he average monthly temperature T (in °C) of a city for the first 6 months of the year ($m = 1$ to $m = 6$) is recorded below.

Month (m)	1	2	3	4	5	6
Temp (T)	5.1	7.2	12.0	18.5	23.1	26.8

(a) Use your GDC to find a sinusoidal regression model $T(m) = a\sin(bm + c) + d$. Give all parameters to 3 significant figures. **[4 marks]**
(b) Using your model, predict the average temperature for August ($m = 8$). **[3 marks]**

Question 12 (8 Marks)

dataset measuring the stopping distance of a car D (metres) against its speed v (km/h) is modelled by two different equations:
Model A (Quadratic): $D_A = 0.006v^2 + 0.2v$
Model B (Exponential): $D_B = 2.5 \times 1.04^v$
At $v = 100$ km/h, the actual measured stopping distance is 82 metres.
Calculate the residual at $v = 100$ for *both* models. State which model provides the most accurate prediction at this speed.

Question 13 (8 Marks)

survey is designed to measure "Math Anxiety". To check its validity, the researcher compares the survey scores of students with their actual final IB Math exam grades. The researcher finds a strong negative correlation between the survey score and the exam grade.
Explain which type of validity (Content, Criterion-related, or Parallel forms) is being demonstrated here, and justify why a negative correlation implies the survey is valid.

Question 14 (7 Marks)

cubic regression model $y = ax^3 + bx^2 + cx + d$ is fit to a dataset containing 10 points. The resulting Coefficient of Determination is exactly $R^2 = 1$.
However, the researcher notices that when predicting a value just slightly outside the data range, the model's prediction is physically impossible (e.g. a negative height).
Explain the mathematical phenomenon occurring here, referring to R^2 and the dangers of high-degree polynomial regression.

Question 15 (9 Marks)

he sales S (in thousands) of a new tech product over x months is recorded.
The data is linearized by plotting $\ln(S)$ against x.
The linear regression line of this transformed data is $\ln(S) = 0.45x + 1.2$.

(a) Show algebraically how to convert this linear equation back into an exponential model $S = A \times B^x$. Find exact values for A and B. **[4 marks]**
(b) Calculate the Sum of Squared Residuals (SS_{res}) for the original (non-linearized) data for the first two months where the raw data points are $(1, 6)$ and $(2, 9)$. **[5 marks]**