NCEA Level 3 · AS 91581: Investigating Bivariate Data

Name: _____ Date: _____

This internal is all about exploring the relationship between two measurement variables using the statistical enquiry cycle. Work through the fluency questions first, then use the graded sections to practise describing, modelling, predicting and evaluating like a statistician.

Revision summary

- The PPDAC cycle: Problem (pose a relationship question), Plan, Data, Analysis (scatter plot, trend, model), Conclusion (answer in context with limitations).
- Scatter plot description covers TAFOS: Trend (linear/non-linear), Association (positive/negative), strength of correlation, Form/features, Outliers, Scatter (spread about the trend).
- Correlation coefficient r ranges from −1 to +1. Near ±1 = strong linear relationship; near 0 = weak/none. Correlation is NOT causation.
- Linear model: $y = a + bx$ (or predicted variable = intercept + slope × explanatory variable). Slope b = change in y per 1-unit increase in x; intercept a = predicted y when x = 0.
- Interpolation (predicting inside the data range) is more reliable than extrapolation (predicting outside the range), which is risky.
- Always define which variable is explanatory (x) and which is the response (y); the model predicts the response.
- Evaluate predictions by checking r, the scatter about the line, sample size, whether it is interpolation or extrapolation, and real-world context.

Key skills

1. A study of 40 Avondale cafés recorded the daily temperature (°C) and the number of iced coffees sold. The correlation coefficient is r = 0.82. Describe the strength and direction of the linear relationship.

2. A linear model for petrol used (litres) against distance driven (km) is: litres = 2.5 + 0.09 × distance. Interpret the slope in context.

3. Using the model litres = 2.5 + 0.09 × distance, predict the petrol used for a 150 km trip.

4. A scatter plot of house size (m²) versus sale price ($) shows points curving upward and getting more spread out for larger houses. Name two features you would describe from this plot.

Achieved

5. Mere collects data on the age (years) and resale value ($) of 25 used utes in West Auckland. She fits the model: value = 42000 − 3100 × age. Describe the relationship and predict the value of a 6-year-old ute.

6. A scatter plot shows the height (cm) and arm span (cm) of 30 students at a Henderson high school. The points lie close to a straight line rising from left to right, with one point far below the others. Describe the trend, direction, strength and any unusual features.

7. For a data set of weekly rainfall (mm) and weekly grass growth (cm) at a Waikato farm, r = 0.15. What does this tell you about the linear relationship?

8. A model for a car's stopping distance (m) against speed (km/h) is: distance = −4 + 0.85 × speed, fitted from data for speeds between 30 and 100 km/h. Explain why using this model to predict the stopping distance at 5 km/h could be unreliable.

Merit

9. A café owner in Avondale investigates the relationship between the number of staff on shift (2 to 8) and the average customer wait time (minutes). Her model is: wait = 14.2 − 1.6 × staff, and r = −0.88. Using the model and the correlation, discuss what the data suggests, and comment on how much you would trust a prediction of wait time for 5 staff versus for 12 staff.

10. Two variables, daily ice-cream sales and daily drowning incidents at NZ beaches, show a strong positive correlation (r = 0.9). A student concludes that buying ice-cream causes drownings. Explain why this reasoning is flawed and suggest a better explanation for the relationship.

11. A researcher compares two possible linear models for the same data on tree age (years) and trunk diameter (cm): Model A has r = 0.72; Model B (after removing one influential outlier) has r = 0.91. Explain what the change in r suggests and discuss whether removing the point is justified.

Excellence

12. A regional council models nitrogen level (mg/L) in a stream against the percentage of surrounding land used for dairy farming, based on 50 catchments. The model is nitrogen = 0.8 + 0.12 × dairy%, with r = 0.68. A colleague argues that because r is only moderate, the model is 'useless' for the council's decisions. Evaluate this claim, discussing the value AND limitations of the model for guiding policy.

13. For a linear model $y = a + bx$ fitted to data, explain algebraically and conceptually why the fitted line always passes through the point $(\bar{x}, \bar{y})$, the mean of the explanatory variable and the mean of the response. Then explain what this tells you about interpreting the intercept a.

14. Two students investigate the same relationship between weekly study hours and exam mark (%). One uses hours as the explanatory variable and gets slope 3.2; the other swaps the axes, using mark as the explanatory variable, and gets a different slope. They are confused that $1 \div 3.2 \neq$ their slope. Explain why swapping which variable is explanatory generally produces a different line that is NOT simply the algebraic inverse.