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.

Answer key (for parents and tutors)

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.
   r = 0.82 is close to +1, so there is a fairly strong, positive linear relationship: as daily temperature increases, the number of iced coffees sold tends to increase.
   Achieved: interprets a correlation coefficient for strength and direction.

2. A linear model for petrol used (litres) against distance driven (km) is: litres = 2.5 + 0.09 × distance. Interpret the slope in context.
   The slope is 0.09. For every extra 1 km driven, the model predicts the petrol used increases by 0.09 litres (about 9 litres per 100 km).
   Achieved: interprets slope as a rate of change in context.

3. Using the model litres = 2.5 + 0.09 × distance, predict the petrol used for a 150 km trip.
   litres = 2.5 + 0.09 × 150 = 2.5 + 13.5 = 16 litres.
   Key skill: substitutes into a linear model to predict.

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.
   Two features: (1) the trend is non-linear (curved/increasing), and (2) the scatter increases for larger houses (non-constant spread / fanning out). Either of these plus mentioning a positive association would be valid.
   Key skill: identifies features of a scatter plot (form and scatter).

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.
   The negative slope shows a negative relationship: as age increases, resale value decreases. Prediction: value = 42000 − 3100 × 6 = 42000 − 18600 = $23 400.
   Achieved: applies a linear model to describe a relationship and make a prediction.

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.
   Trend: linear. Direction: positive (taller students tend to have larger arm span). Strength: strong, because the points lie close to a straight line. Unusual feature: one outlier lies well below the pattern, possibly a measurement error or an unusual individual.
   Achieved: describes trend, association, strength and outliers from a scatter plot.

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?
   r = 0.15 is close to 0, so there is a very weak (essentially no) linear relationship between weekly rainfall and grass growth in this data. Knowing the rainfall would not help predict grass growth well using a straight line.
   Achieved: interprets a weak correlation coefficient.

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.
   5 km/h is well outside the data range of 30–100 km/h, so this is extrapolation. The model was only fitted for that range; the relationship may not hold at very low speeds. In fact it predicts distance = −4 + 0.85 × 5 = 0.25 m, which is a very small (and the intercept gives a negative value at 0 km/h), showing the model breaks down outside the data range.
   Achieved: recognises extrapolation and explains reduced reliability.

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.

The slope −1.6 means each extra staff member is associated with about a 1.6-minute decrease in average wait time. r = −0.88 shows a strong negative linear relationship, so the trend is reliable within the data. Prediction for 5 staff: wait = 14.2 − 1.6 × 5 = 6.2 minutes, this is interpolation (within 2–8), and with strong correlation it is fairly trustworthy. Prediction for 12 staff: wait = 14.2 − 1.6 × 12 = −5 minutes, this is extrapolation beyond the data and gives an impossible negative wait time, so it should NOT be trusted; the linear relationship clearly breaks down outside the observed range.

Merit: links slope, correlation and interpolation/extrapolation to judge prediction reliability.

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.
Correlation does not imply causation. A strong r only shows the two variables move together, not that one causes the other. A more sensible explanation is a lurking (confounding) variable: hot summer weather. On hot days more people buy ice-cream AND more people swim, so more drownings occur. The temperature drives both variables, creating the correlation without any direct causal link.
Merit: connects correlation to the causation issue and identifies a confounding variable.

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.
The rise from r = 0.72 to 0.91 shows the single removed point was an influential outlier weakening the apparent linear relationship; without it the remaining points follow a straight line much more closely. Removing it strengthens the fit. However, removal is only justified if there is a good reason, e.g. a measurement or recording error. If the point is a genuine observation, it should be kept and investigated, because deleting valid data just to improve r is misleading and hides real variation in the population.
Merit: links the effect of an outlier on r to a reasoned judgement about data handling.

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.
The claim is too strong. r = 0.68 indicates a moderate positive linear relationship, dairy land percentage explains a meaningful part of the variation in nitrogen ($r^2 \approx 0.46$, so about 46% of the variability), which is genuinely useful for identifying at-risk catchments and informing policy direction. However, limitations must be acknowledged: (1) 54% of variation is unexplained, so other factors (rainfall, soil type, fertiliser practice, riparian planting) matter and the model alone cannot predict any single catchment precisely; (2) correlation is not proof that dairy farming causes the nitrogen, though there is a plausible causal mechanism here; (3) predictions should stay within the observed range of dairy%. Conclusion: the model is a useful screening tool for policy but should guide, not solely determine, decisions, it must be combined with other evidence and treated as showing a tendency rather than an exact rule.
Excellence: justifies a nuanced evaluation using r^2, explained variation, causation and context.

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.
The least-squares line satisfies $\bar{y} = a + b\cdot\bar{x}$, because the line is fitted to minimise the total squared vertical distance to the points, and the balancing point of that fit is the centroid $(\bar{x}, \bar{y})$, the mean of x and mean of y. Rearranging gives $a = \bar{y} - b\cdot\bar{x}$, showing the intercept is fully determined once the slope and the two means are known. Conceptually, the line 'pivots' about the mean point. For the intercept: a is the predicted y when x = 0. This is only meaningful if x = 0 is within or near the data range; often x = 0 is far outside the data (e.g. a house of 0 m^2), so a is a mathematical anchor for the line rather than a realistic prediction. This is exactly why extrapolation to x = 0 can be misleading.
Excellence: generalises and justifies a structural property of the least-squares line and links it to interpretation.

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.

When fitting a linear model, the line is chosen to minimise the squared distances measured VERTICALLY, in the direction of the response variable. So the model of mark-on-hours minimises errors in mark, while the model of hours-on-mark minimises errors in hours; these are two different optimisation problems using different residuals, so they give two different lines. The slope of one is only the reciprocal of the other in the perfect case $r = \pm 1$ (all points exactly on a line). In general the product of the two regression slopes equals r^2, which is less than 1, so $1 \div 3.2$ will not equal the reversed slope. Conclusion: the choice of explanatory variable matters and must match the investigative question, you fit the model to predict the variable you care about, and you cannot simply invert one model to get the other.

Excellence: abstract reasoning about asymmetry of regression, residual direction and the r^2 relationship.