Past-paper walk-through

Least Squares Regression ▪ 2.8

eXercise

Questions in this set

- 2022 Q1
- 2018 Q1
- 2017 Q1
- 2015 Q5

Reading the mark scheme

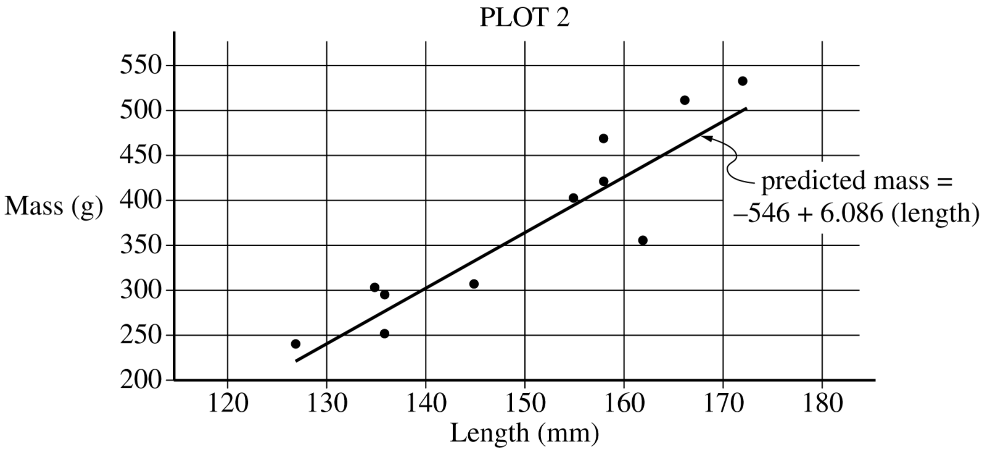
- **B1** a mark that stands on its own – the point is either made or not.
- **M1** a *method* mark: the right approach, even if the number is wrong.
- **A1** an *answer* mark, and it usually needs its M mark first.
- **C1** a working mark on the way to the answer.
- // separates answers the examiner accepts equally.
- **ecf** = error carried forward: a later part is marked on your own earlier answer.
- **owtte** = "or words to that effect" – the wording need not match.

Question 1

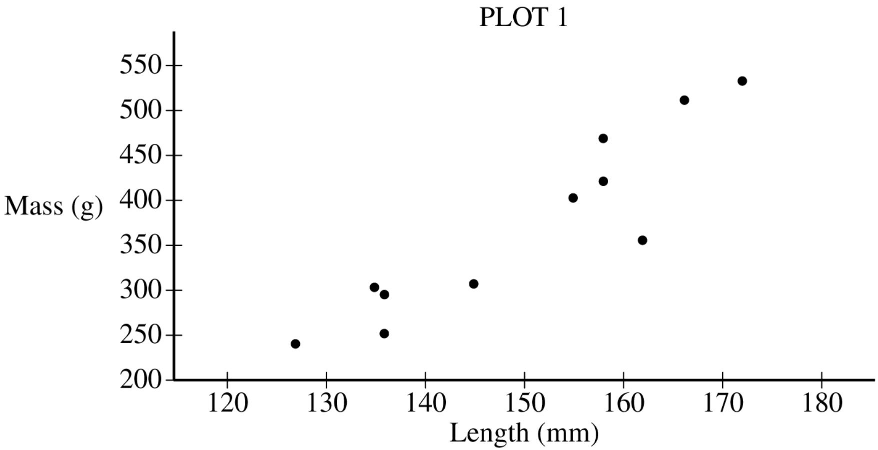
2022 ▪ Q1

A biologist gathered data on the length, in millimeters (mm), and the mass, in grams (g), for 11 bullfrogs. The data are shown in Plot 1.
[Plot 1: Scatterplot titled “PLOT 1”; x-axis “Length (mm)” from 120 to 180; y-axis “Mass (g)” from 200 to 550. Approximate points (length, mass): (127, 240), (135, 295), (136, 300), (136, 255), (145, 310), (155, 400), (157, 470), (158, 420), (162, 355), (166, 510), (172, 530).]
(a) Based on the scatterplot, describe the relationship between mass and length, in context.

Question 1


PLOT 2
Mass (g)
550
500
450
400
350
300
250
200
120
130
140
150
160
170
180
Length (mm)
predicted mass =
–546 + 6.086 (length)

Question 1


PLOT 1
Mass (g)
550
500
450
400
350
300
250
200
120
130
140
150
160
170
180
Length (mm)

Question 1

2022 ▪ Q1

A biologist gathered data on the length, in millimeters (mm), and the mass, in grams (g), for 11 bullfrogs. The data are shown in Plot 1.

[Plot 1: Scatterplot titled "PLOT 1"; x-axis "Length (mm)" from 120 to 180; y-axis "Mass (g)" from 200 to 550. Approximate points (length, mass): (127, 240), (135, 295), (136, 300), (136, 255), (145, 310), (155, 400), (157, 470), (158, 420), (162, 355), (166, 510), (172, 530).]

(b) From the data, the biologist calculated the least-squares regression line for predicting mass from length. The least-squares regression line is shown in Plot 2.

[Plot 2: Same scatterplot as Plot 1 with a least-squares regression line drawn through the points, labeled "predicted mass = -546 + 6.086 (length)"; x-axis "Length (mm)" from 120 to 180, y-axis "Mass (g)" from 200 to 550.]

Identify and interpret the slope of the least-squares regression line in context.

Question 1

2022 ▪ Q1

A biologist gathered data on the length, in millimeters (mm), and the mass, in grams (g), for 11 bullfrogs. The data are shown in Plot 1.

[Plot 1: Scatterplot titled "PLOT 1"; x-axis "Length (mm)" from 120 to 180; y-axis "Mass (g)" from 200 to 550. Approximate points (length, mass): (127, 240), (135, 295), (136, 300), (136, 255), (145, 310), (155, 400), (157, 470), (158, 420), (162, 355), (166, 510), (172, 530).]

(c) Interpret the coefficient of determination of the least-squares regression line, $r^2 \approx 0.819$, in context.

Question 1

2022 ▪ Q1

A biologist gathered data on the length, in millimeters (mm), and the mass, in grams (g), for 11 bullfrogs. The data are shown in Plot 1.

[Plot 1: Scatterplot titled “PLOT 1”; x-axis “Length (mm)” from 120 to 180; y-axis “Mass (g)” from 200 to 550. Approximate points (length, mass): (127, 240), (135, 295), (136, 300), (136, 255), (145, 310), (155, 400), (157, 470), (158, 420), (162, 355), (166, 510), (172, 530).]

(d)(i) From Plot 2, consider the residuals of the 11 bullfrogs.

Based on the plot, approximately what is the length and mass of the bullfrog with the largest absolute value residual?

(d)(ii) Does the least-squares regression line overestimate or underestimate the mass of the bullfrog identified in part (d-i)? Explain your answer.

Solution 1

(a) Mark scheme

- The scatterplot reveals a strong, positive, roughly linear association between the mass and length of bullfrogs. There are no points that seriously deviate from the straight-line pattern of the points in the plot. Full credit needs a description including at least three of: (1) direction of association (positive/increasing), (2) strength of association (strong), (3) form (linear or approximately linear), (4) no unusual features (no points with large discrepancies from the pattern), plus (5) context (association between length and mass of bullfrogs, or between length and mass in general).

Solution 1

(b) Mark scheme

- The slope of the least-squares regression line is 6.086 (may be rounded to 6.09 or 6.1, but not to 6). This indicates the predicted mass of a bullfrog increases by 6.086 grams for each additional millimeter of length. Full credit needs all three: (1) identifies the slope value as 6.086, (2) an interpretation referencing an increase of a number of grams of mass for each one-millimeter increase in length (correct units for both mass and length), (3) uses non-deterministic language such as 'predicted', 'estimated', 'expected', or 'average' rather than stating the change as certain.

Solution 1

(c) Mark scheme

- The coefficient of determination is $r^2 \approx 0.819$. This value indicates that 81.9% of the variation in bullfrog mass can be explained by variation in bullfrog length as described by the least-squares regression line. The interpretation must explicitly relate to the dependent (response) variable, mass, being explained by the independent (explanatory) variable, length – a generic definition with no context, an interpretation of r (not r^2) instead, or an interpretation that references the predicted y-values or the data (rather than mass explained by length) earns only partial credit.

Solution 1

(d)(i)

Mark scheme

- From Plot 2, the largest residual in absolute value belongs to the bullfrog with length approximately 162 millimeters and mass approximately 356 grams (any length between 160 and 165 mm and mass between 350 and 375 g, read from the plot, is accepted as reasonable). This identification is the first of two components that together earn the single point shared with part (d-ii): full credit for that shared point additionally requires correctly stating in (d-ii) whether the regression line overestimates or underestimates this bullfrog's mass.

Solution 1

(d)(ii) Mark scheme

- The least-squares regression line overestimates the mass of the bullfrog identified in part (d-i): Plot 2 shows the point for the bullfrog with length 162 mm lies below the least-squares regression line (the predicted mass at length 162, using $-546 + (6.086)(162) = 439.9$ grams, is larger than the bullfrog's actual mass of about 356 grams). This single point is shared between (d-i) and (d-ii) and requires both: (1) correctly identifying the bullfrog in (d-i) (length between 160-165 mm, mass between 350-375 g), and (2) explicitly stating in (d-ii) whether the line overestimates or underestimates the mass, with a justification comparing the identified point to the regression line (a numerical predicted-mass value is not required, but if given must be reasonable, e.g. between 425 and 450 g).

Question 1

2018 ▪ Q1

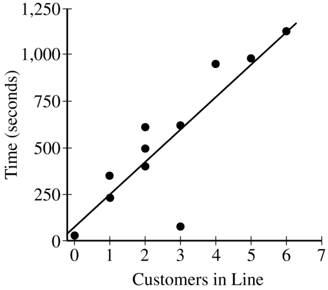
The manager of a grocery store selected a random sample of 11 customers to investigate the relationship between the number of customers in a checkout line and the time to finish checkout. As soon as the selected customer entered the end of a checkout line, data were collected on the number of customers in line who were in front of the selected customer and the time, in seconds, until the selected customer was finished with the checkout. The data are shown in the following scatterplot along with the corresponding least-squares regression line and computer output. [Scatterplot titled "Time (seconds)" vs. "Customers in Line"; y-axis (Time, seconds) from 0 to 1,250 in increments of 250; x-axis (Customers in Line) from 0 to 7. A least-squares regression line is drawn from about (0, 25) up through the point cloud to about (6.2, 1,150). Approximate data points (Customers in Line, Time): (0, 25), (1, 230), (1, 350), (2, 410), (2, 500), (2, 610), (3, 75), (3, 610), (4, 950), (5, 985), (6,

Question 1

1,150). The point at approximately (3, 75) lies well below the regression line and the rest of the data cloud.]
| Predictor | Coef | SE Coef | T | P | |—|—|—|—|—| | Constant | 72.95 | 110.36 | 0.66 | 0.525 | | Customers in line | 174.40 | 35.06 | 4.97 | 0.001 |
$S = 200.01$ $\qquad$ $R\text{-}Sq = 73.33\%$ $\qquad$ $R\text{-}Sq\ (adj) = 70.37\%$
(a) Identify and interpret in context the estimate of the intercept for the least-squares regression line.
(b) Identify and interpret in context the coefficient of determination, r^2.
(c) One of the data points was determined to be an outlier. Circle the point on the scatterplot and explain why the point is considered an outlier.

Question 1

checkout. The data are shown in the following scatterplot along with the corresponding least-squ
line and computer output.



Predictor	Coef	SE Coef	T	P
Constant	72.95	110.36	0.66	0.525
Customers in line	174.40	35.06	4.97	0.001

S = 200.01 R-Sq = 73.33% R-Sq (adj) = 70.37%

Solution 1

(a) Mark scheme

- The estimate of the y-intercept is $\hat{\beta}_0 = 72.95$. In context: it is the predicted (estimated/average) time to finish checkout when there are zero other customers in line ahead of the selected customer. Full credit requires three things: (1) correctly identifying 72.95 as the intercept, (2) communicating the y-intercept concept in context that includes both time AND zero customers, and (3) indicating the value is a prediction/estimate/average (not a guaranteed value). A bare regression equation such as $\hat{y} = 72.95 + 174.40x$ does not by itself satisfy (1) unless the intercept is explicitly labeled, and cannot satisfy (3) at all. A response that calls 72.95 the slope earns no credit for (1) or (2).

Solution 1

(b) Mark scheme

- The coefficient of determination is $r^2 = 73.33\%$. Interpretation in context: about 73.33% of the variability in the time to finish checkout (including time waiting in line) is explained by (accounted for by) the linear relationship with the number of customers in line ahead of the selected customer. Full credit requires: (1) correctly identifying 73.33%, (2) a correct (possibly generic) interpretation of r^2 as the percent/proportion of variability in y explained by the linear relationship with x, and (3) that the interpretation includes context (time and number of customers) without reversing the roles of x and y. Common wrong interpretations that earn no credit for (2): variability in the *predicted* y explained, variability in 'the data' explained, or variability explained 'on average'.

Solution 1

(c) Mark scheme

- The outlier is the point with $x = 3$ and y close to 0 (a y-value strictly between 0 and 250). It is an outlier because the combination of its x and y values does not fit the pattern of the rest of the data: specifically its y-value (checkout time) is much lower than would be expected/predicted for x=3 customers in line, given the remaining data (equivalently, it has an unusually large residual relative to the other residuals). Full credit requires (1) correctly identifying the outlier (circled on the graph, or valid coordinates given: x=3 and 0<y<250) and (2) describing an unusual feature relative to the REST of the data sufficient to justify calling it an outlier. A statement with no comparison to the remaining points (e.g. just 'it has a large residual' or 'it's nowhere near the line') does not satisfy (2), nor does an

Solution 1

appeal to the outlier's influence on the regression coefficients or correlation without explicit numerical calculation.

Question 1

2017 ▪ Q1

Researchers studying a pack of gray wolves in North America collected data on the length x, in meters, from nose to tip of tail, and the weight y, in kilograms, of the wolves. A scatterplot of weight versus length revealed a relationship between the two variables described as positive, linear, and strong.

(a)(i) For the situation described above, explain what is meant by each of the following words.

Positive:

(a)(ii) Linear:

(a)(iii) Strong:

(b) The data collected from the wolves were used to create the least-squares equation $\hat{y} = -16.46 + 35.02x$.

Interpret the meaning of the slope of the least-squares regression line in context.

Question 1

(c) One wolf in the pack with a length of 1.4 meters had a residual of -9.67 kilograms. What was the weight of the wolf?

Solution 1

(a)(i) Mark scheme

- 'Positive' means that wolves with higher values of length also tend to have higher values of weight – low length goes with low weight, and high length goes with high weight (as length increases, weight tends to increase). Acceptable phrasing: "As length increases, so does weight"; "Longer wolves weigh more"; "The points on the graph go up as you move left to right." NOT acceptable: "As length goes up, weight changes"; "the correlation is greater than 0" (must describe the low-low/high-high pattern, not just restate the term). Full part (a) credit needs a definition like this for positive, one for linear, one for strong, AND at least one of the three stated in context (using length/weight or meters/kilograms).

Solution 1

(a)(ii) Mark scheme

- 'Linear' means that as length increases by one meter, weight tends to change by a constant amount, on average – the data points follow the pattern of a straight line (equivalently, the rate of change of weight with respect to length is constant). Acceptable: "The points generally follow a straight line"; "the relationship between x and y is straight"; "length and weight have a constant slope." NOT acceptable: "the points all line up"; "you can draw a straight line through the points"; "there is a positive correlation"; merely restating that every increase in x yields a 35.02 increase in y. See the context/four-component requirement noted under 1ai for part (a) as a whole.

Solution 1

(a)(iii) Mark scheme

- 'Strong' means the data points fall close to the line (or curve) – observed weights are close to predicted weights, i.e. residuals (deviations) from the least-squares line are small, or equivalently the correlation coefficient is close to 1 in magnitude. Acceptable: "Observed values are close to predicted values"; "Deviations from the least-squares regression line are small"; "The correlation coefficient is close to 1." NOT acceptable: "the points are close together"; "the scatterplots are clustered together"; "there is a high positive correlation" (these describe density of points, not closeness to a line). See the context/four-component requirement noted under 1ai for part (a) as a whole.

Solution 1

(b) Mark scheme

- The slope of 35.02 means: two wolves whose lengths differ by one meter are predicted to differ in weight by 35.02 kilograms, with the longer wolf having the greater weight. Full credit requires three components: (1) correctly cites the slope value 35.02; (2) interprets it as a specified increase in weight for each unit increase in length; (3) indicates the relationship is not exact/deterministic, using language such as "predicted" or "on average" (e.g. "...predicted to differ by 35.02 kg" or "...weight increases by 35.02 kg on average"). A response that only says "weight increases by 35.02 kg per meter" with no averaging/predicted language misses component 3. Note: if the response instead identifies -16.46 (the

Solution 1

intercept) as the slope, component 2 is satisfied only if it states weight DECREASES by 16.46 kg on average for each meter increase.

Solution 1

(c) Mark scheme

- In general, residual = actual weight - predicted weight, so actual weight = predicted weight + residual. For the wolf with length 1.4 m: predicted weight $= -16.46 + 35.02(1.4) = 32.568$ kilograms. Actual weight $= 32.568 + (-9.67) = 22.898$ kilograms (≈ 22.9 kg). Full credit needs both: (1) the correct predicted-weight computation (32.568 kg), and (2) the correct actual-weight computation obtained by ADDING the (negative) residual to the predicted value, giving 22.9 kg. A common error is subtracting instead of adding the residual (giving 42.24 kg), which does not earn full credit. Minor arithmetic slips that don't lead to an unreasonable value (e.g. a negative weight) can still be treated as showing the correct method.

Question 5

2015 ▪ Q5

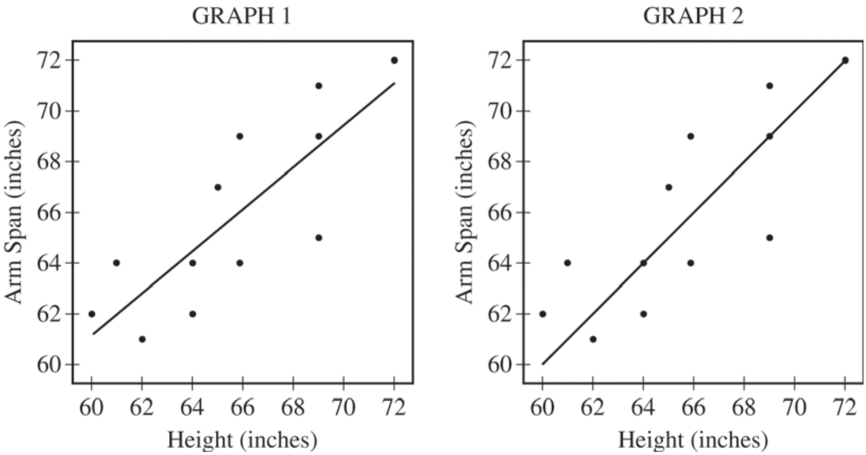
A student measured the heights and the arm spans, rounded to the nearest inch, of each person in a random sample of 12 seniors at a high school. A scatterplot of arm span versus height for the 12 seniors is shown.
[Scatterplot of Arm Span (inches) on the y-axis, 60 to 72, versus Height (inches) on the x-axis, 60 to 72. Approximate points: (61,64), (62,61), (64,64), (64,62), (65,67), (65,64), (66,69), (69,65), (69,71), (69,69), (72,72), (60,62).]
Let x represent height, in inches, and let y represent arm span, in inches. Two scatterplots of the same data are shown below. Graph 1 shows the data with the least squares regression line $\hat{y} = 11.74 + 0.8247x$, and graph 2 shows the data with the line $y = x$.
[Two side-by-side scatterplots, both with the same 12 points as above, axes Arm Span (inches) 60 to 72 vs. Height (inches) 60 to 72. GRAPH 1: shows the least-squares

Question 5

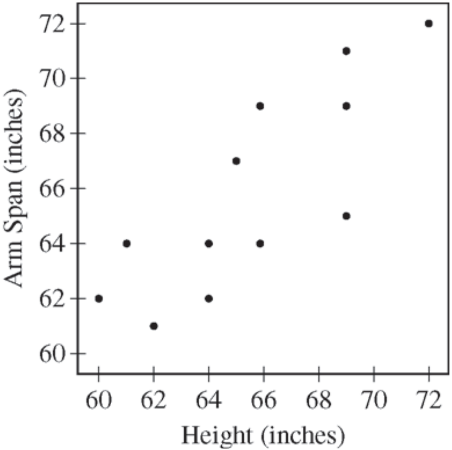
regression line $\hat{y} = 11.74 + 0.8247x$ passing through the data, from about (60,61) to (72,71). GRAPH 2: shows the line $y = x$ passing through the data, from (60,60) to (72,72).]

(a) Based on the scatterplot, describe the relationship between arm span and height for the sample of 12 seniors.

Question 5


GRAPH 1
Arm Span (inches)
Height (inches)
60
62
64
66
68
70
72
GRAPH 2
Arm Span (inches)
Height (inches)

Question 5



Arm Span (inches)
72
70
68
66
64
62
60
60 62 64 66 68 70 72
Height (inches)

Question 5

Classification	Square	Tall Rectangle	Short Rectangle
Frequency			