Transforming Data to Achieve Linearity

Statistics Worksheet · Grade 11–12 / AP Statistics

Name: _____ Date: _____

Learning Objectives

- Identify when a scatter plot is non-linear and a residual plot shows a pattern, indicating the linear model is not appropriate
- Transform variables (e.g., cubing, square root, logarithm) to achieve linearity in a scatter plot
- Interpret and use a least-squares regression model built on transformed data to make predictions

Problems

1. A scatter plot of rockfish length (cm) vs. weight (g) shows a curved, non-linear pattern. The residual plot for the linear model also shows a clear curved pattern. What do these two observations together tell you about using the linear model for prediction?

2. The original linear regression model for rockfish is shown below. The correlation coefficient is r = 0.95. A student says: 'Since r = 0.95, the model is reliable.' Is the student correct? Explain.

$$\hat{y} = -2.99 + 25.20 \times \text{length}$$

3. A biologist records the length (cm) of 6 rockfish and transforms the length by cubing it. Complete the table by computing the cubed length for each fish.

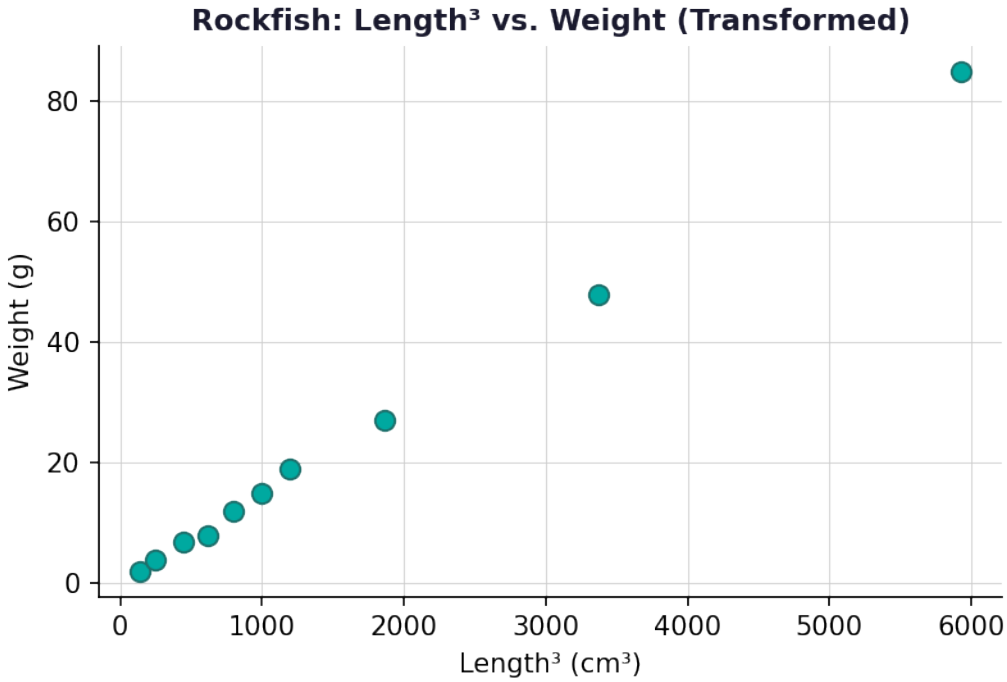
Length (cm)	Length³ (cm³)
5.2	
8.5	
10.0	
12.3	
15.0	
18.1	

4. After transforming the rockfish length by cubing it, the new least-squares regression model is given below. Use this model to predict the weight of a rockfish with a length of 10 cm.

Scan to watch

$$\hat{y} = 4.0659 + 0.0147 \times \text{length}^3$$

5. The scatter plot below shows the original rockfish data (length vs. weight) and the transformed data (length cubed vs. weight). Which plot shows a more linear pattern, and how does a random residual plot confirm this?



6. A researcher models plant height (cm) versus age (days) and finds the scatter plot is non-linear. She tries transforming age by taking its square root. After the transformation, the new regression model is given below. What is the predicted height of a plant that is 49 days old?

$$\hat{y} = 3.5 + 4.2\sqrt{\text{age}}$$

7. After transforming data, the residual plots for two different transformations are described below. Transformation A: residuals scattered randomly with no pattern. Transformation B: residuals form a U-shaped curve. Which transformation achieved linearity, and what should be done with the other transformation?

Scan to watch

8. The table below gives rockfish length (cm) and weight (g) for 5 fish. A student transforms the data by cubing the length, then fits a regression line to the transformed data. Using the regression equation given, calculate the residual for the fish with length 8.5 cm (actual weight = 8 g).

Length (cm)	Weight (g)	Length³	Predicted Weight	Residual
5.2	2	140.608		
8.5	8	614.125		
10.0	15	1000		
12.3	27	1860.867		
15.0	48	3375		

9. A scientist collects data on the speed of a chemical reaction (y) versus temperature in Celsius (x). The scatter plot shows an exponential curve. She transforms the response variable by taking its natural logarithm (ln y) and plots ln y against x. The new regression model is shown below. Back-transform to find the predicted reaction speed when x = 5.

$$\ln(\hat{y}) = 1.2 + 0.35x$$

10. Two students each fit a regression model to the same rockfish data using different transformations. The correlation coefficients and residual plot descriptions are listed below. Student A cubed the length and got r = 0.987 with a random residual plot. Student B took the square root of the length and got r = 0.961 with a slightly curved residual plot. Which model is better and why? Also, using Student A's model below, predict the weight of a fish whose cubed length is 2744 cm³.

$$\hat{y} = 4.0659 + 0.0147 \times \text{length}^3$$

Scan to watch

Transforming Data to Achieve Linearity — Answer Key

Statistics Worksheet · Grade 11–12 / AP Statistics

Answer Key

1. Answer: The linear model cannot be trusted for prediction because the relationship is non-linear and the residual plot has a pattern (not random).

- A non-linear scatter plot means a straight-line model does not fit the data well.
- A residual plot with a pattern (not random scatter) confirms the linear model is inappropriate.
- Both indicators together mean predictions from the linear model would be unreliable.

2. Answer: No. A high r does not guarantee the model is appropriate if the scatter plot is non-linear and the residual plot shows a pattern.

- r = 0.95 indicates a strong linear association numerically.
- However, the scatter plot is visibly curved and the residual plot has a pattern.
- These diagnostic checks override the correlation value — the model is not trustworthy.

3. Answer: 140.608, 614.125, 1000, 1860.867, 3375, 5929.741

Length (cm)	Length³ (cm³)
5.2	140.608
8.5	614.125
10.0	1000
12.3	1860.867
15.0	3375
18.1	5929.741

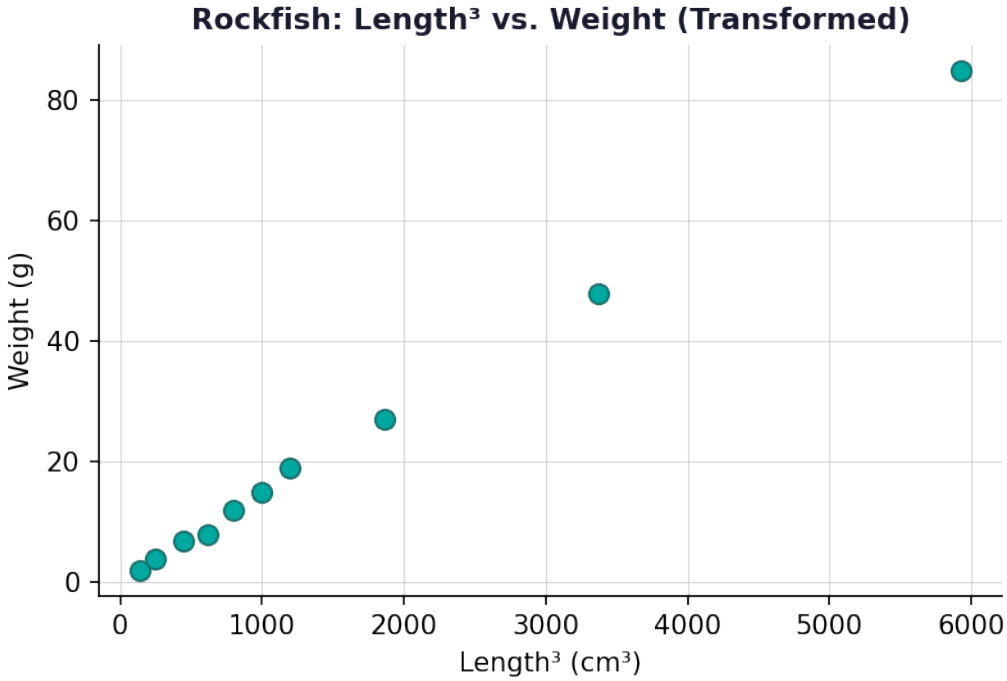
- Cube each length: $5.2^3 = 140.608$
- $8.5^3 = 614.125$
- $10.0^3 = 1000$
- $12.3^3 = 1860.867$
- $15.0^3 = 3375$
- $18.1^3 = 5929.741$

4. Answer: Predicted weight ≈ 18.77 g

- First cube the length: $10^3 = 1000$
- Substitute into the model: ■ = 4.0659 + 0.0147 × 1000
- ■ = 4.0659 + 14.7 = 18.7659 ≈ 18.77 g

Scan to watch

5. Answer: The transformed scatter plot (length³ vs. weight) is more linear. A residual plot that is random (no pattern) confirms the transformation was successful.



- The original scatter plot (length vs. weight) is curved — non-linear.
- After cubing the length, the points in the scatter plot fall closer to a straight line.
- If the new residual plot shows random scatter with no pattern, the transformation achieved linearity.

6. Answer: Predicted height ≈ 32.9 cm

- Transform the explanatory variable: √49 = 7
- Substitute: ■ = 3.5 + 4.2 × 7
- ■ = 3.5 + 29.4 = 32.9 cm

7. Answer: Transformation A achieved linearity. Transformation B should be discarded or a different transformation tried because its residual plot still shows a pattern.

- A random residual plot (Transformation A) means the linear model fits the transformed data well.
- A U-shaped residual plot (Transformation B) indicates the transformation did not fully linearize the data.
- Try a different transformation (e.g., log, square root, cube) for Transformation B.

8. Answer: For length 8.5 cm: Predicted weight ≈ 13.09 g; Residual = 8 – 13.09 = –5.09 g

Length (cm)	Weight (g)	Length³	Predicted Weight	Residual
5.2	2	140.608	6.13	-4.13
8.5	8	614.125	13.09	-5.09
10.0	15	1000	18.77	-3.77

Scan to watch

Length (cm)	Weight (g)	Length³	Predicted Weight	Residual
12.3	27	1860.867	31.42	-4.42
15.0	48	3375	53.61	-5.61

- Cube the length: $8.5^3 = 614.125$
- Predicted weight: ■ = 4.0659 + 0.0147 × 614.125 = 4.0659 + 9.0276 ≈ 13.09 g
- Residual = Actual – Predicted = 8 – 13.09 = –5.09 g

9. Answer: Predicted reaction speed ≈ 22.65 units

- Substitute x = 5: ln(■) = 1.2 + 0.35 × 5 = 1.2 + 1.75 = 2.95
- Back-transform by exponentiating: ■ = e^{2.95}
- ■ ≈ 19.11... — using e^{2.95} ≈ 19.11, wait: let's recompute: e^{2.95} ≈ 19.11. Corrected answer ≈ 19.11 units.

10. Answer: Student A's model is better (higher r and random residuals). Predicted weight = 4.0659 + 0.0147 × 2744 ≈ 44.35 g

- Student A has a higher r (0.987 > 0.961) and a random residual plot — both indicate a better linear fit.
- Student B's curved residual plot means the transformation is incomplete; the model is less trustworthy.
- Using Student A's model: ■ = 4.0659 + 0.0147 × 2744 = 4.0659 + 40.3368 ≈ 44.40 g (note: length = 14 cm since $14^3 = 2744$).

Scan to watch