Chapter 12: More About Regression

Section 12.2
Transforming to Achieve Linearity

The Practice of Statistics, 4th edition – For AP*

STARNES, YATES, MOORE

Section 12.2
Transforming to Achieve Linearity

Learning Objectives

After this section, you should be able to…

- ✓ USE transformations involving powers and roots to achieve linearity for a relationship between two variables
- ✓ MAKE predictions from a least-squares regression line involving transformed data
- ✓ USE transformations involving logarithms to achieve linearity for a relationship between two variables
- ✓ DETERMINE which of several transformations does a better job of producing a linear relationship

Introduction

In Chapter 3, we learned how to analyze relationships between two quantitative variables that showed a linear pattern. When two-variable data show a curved relationship, we must develop new techniques for finding an appropriate model. This section describes several simple transformations of data that can straighten a nonlinear pattern.

Once the data have been transformed to achieve linearity, we can use least-squares regression to generate a useful model for making predictions. And if the conditions for regression inference are met, we can estimate or test a claim about the slope of the population (true) regression line using the transformed data.

Applying a function such as the logarithm or square root to a quantitative variable is called **transforming** the data. We will see in this section that understanding how simple functions work helps us choose and use transformations to straighten nonlinear patterns.

Transforming with Powers and Roots

When you visit a pizza parlor, you order a pizza by its diameter—say, 10 inches, 12 inches, or 14 inches. But the amount you get to eat depends on the area of the pizza. The area of a circle is π times the square of its radius r. So the area of a round pizza with diameter x is

$$\text{area} = \pi\left(\frac{x}{2}\right)^2 = \pi\left(\frac{x^2}{4}\right) = \frac{\pi}{4}x^2$$

This is a **power model** of the form $y = ax^p$ with $a = \pi/4$ and $p = 2$.

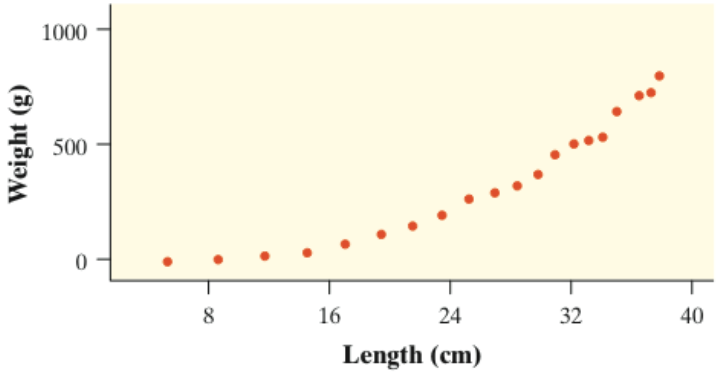
Although a power model of the form $y = ax^p$ describes the relationship between x and y in this setting, there is a *linear* relationship between x^p and y.

If we transform the values of the explanatory variable x by raising them to the p power, and graph the points (x^p, y), the scatterplot should have a linear form.

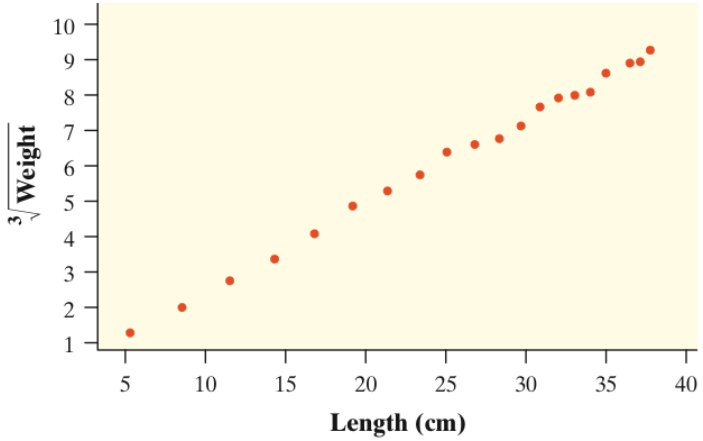
Example: Go Fish!

Imagine that you have been put in charge of organizing a fishing tournament in which prizes will be given for the heaviest Atlantic Ocean rockfish caught. You know that many of the fish caught during the tournament will be measured and released. You are also aware that using delicate scales to try to weigh a fish that is flopping around in a moving boat will probably not yield very accurate results. It would be much easier to measure the length of the fish while on the boat. What you need is a way to convert the length of the fish to its weight.

Length:	5.2	8.5	11.5	14.3	16.8	19.2	21.3	23.3	25.0	26.7
Weight:	2	8	21	38	69	117	148	190	264	293
Length:	28.2	29.6	30.8	32.0	33.0	34.0	34.9	36.4	37.1	37.7
Weight:	318	371	455	504	518	537	651	719	726	810

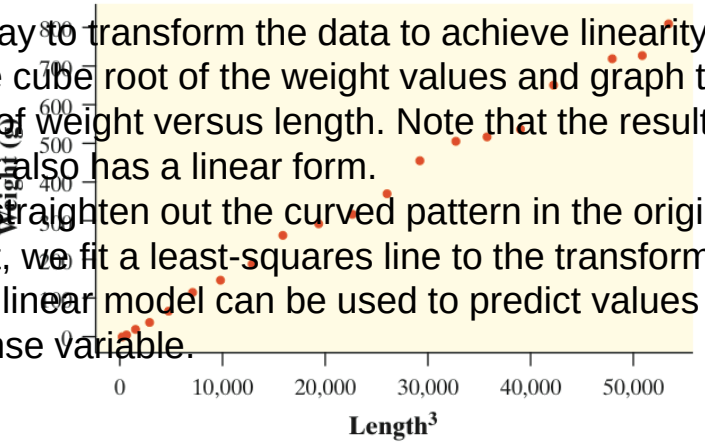


Reference data on the length (in centimeters) and weight (in grams) for Atlantic Ocean rockfish of several sizes is plotted. Note the clear curved shape.



veight (like
odel of the form
elationship
riable he
Another way to transform the data to achieve linearity is to take the cube root of the weight values and graph the cube root of weight versus length. Note that the resulting scatterplot also has a linear form.

Once we straighten out the curved pattern in the original scatterplot, we fit a least-squares line to the transformed data. This linear model can be used to predict values of the response variable.



Example: Go Fish!

Here is Minitab output from separate regression analyses of the two sets of transformed Atlantic Ocean rockfish data.

Transformation 1: (length³, weight)

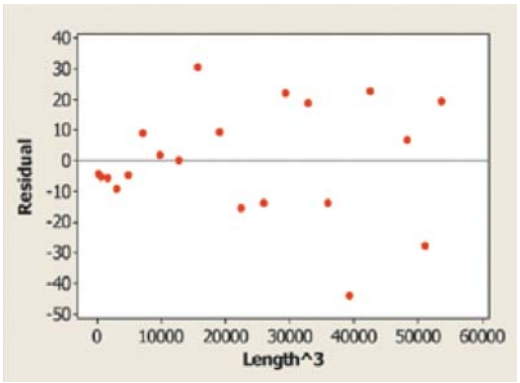
Predictor	Coef	SE Coef	T	P
Constant	4.066	6.902	0.59	0.563
Length^3	0.0146774	0.0002404	61.07	0.000

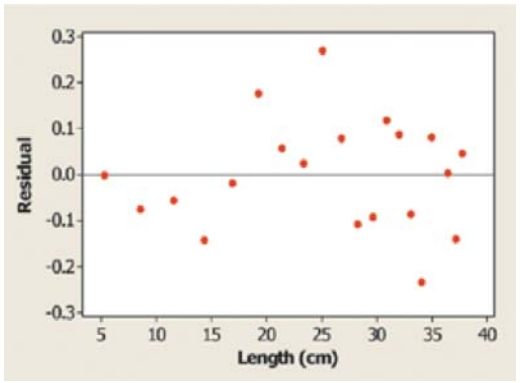
S = 18.8412 R-Sq = 99.5% R-Sq(adj) = 99.5%

Transformation 2: (length, $\sqrt[3]{\text{weight}}$)

Predictor	Coef	SE Coef	T	P
Constant	−0.02204	0.07762	−0.28	0.780
Length	0.246616	0.002868	86.00	0.000

S = 0.124161 R-Sq = 99.8% R-Sq(adj) = 99.7%





(a) Give the equation of the least-squares regression line for predicting weight from length^3 for Atlantic ocean rockfish.

(b) Interpret the value of s in the context of the problem.

(c) Suppose that a fish caught in the future is 36 centimeters long. Use the model from part (a) to predict the fish's weight. Show your work.

For transformation 1, the standard deviation of the residuals is $s = 18.841$ grams. Predictions of fish weight using this model will be off by an average of about 19 grams. For transformation 2, $s = 0.12$. that is, predictions of the cube root of fish weight using this model will be off by an average of about 0.12 grams.

Transformation 1: $\widehat{\text{weight}} = 4.066 + 0.0146774(36^3) = 399.9$ grams

Transformation 2: $\widehat{\sqrt[3]{\text{weight}}} = -0.02204 + 0.246616(36) = 8.856$

$$\widehat{\text{weight}} = 8.856^3 = 694.6 \text{ grams}$$

Transforming with Powers and Roots

When experience or theory suggests that the relationship between two variables is described by a power model of the form $y = ax^p$, you now have two strategies for transforming the data to achieve linearity.

1. Raise the values of the explanatory variable x to the p power and plot the points (x^p, y).

2. Take the p^{th} root of the values of the response variable y and plot the points $(x, \sqrt[p]{y})$.

What if you have no idea what power to choose? You could guess and test until you find a transformation that works. Some technology comes with built-in sliders that allow you to dynamically adjust the power and watch the scatterplot change shape as you do.

It turns out that there is a much more efficient method for linearizing a curved pattern in a scatterplot. Instead of transforming with powers and roots, we use **logarithms**. This more general method works when the data follow an unknown power model or any of several other common mathematical models.

Transforming with Logarithms

Not all curved relationships are described by power models. Some relationships can be described by a **logarithmic model** of the form $y = a + b \log x$.

Sometimes the relationship between y and x is based on repeated multiplication by a constant factor. That is, each time x increases by 1 unit, the value of y is multiplied by b. An **exponential model** of the form $y = ab^x$ describes such multiplicative growth.

If an exponential model of the form $y = ab^x$ describes the relationship between x and y, we can use logarithms to transform the data to produce a linear relationship.

$$y = ab^x$$

exponential model

taking the logarithm of both sides

using the property $\log(mn) = \log m + \log n$

using the property $\log m^p = p \log m$

Transforming with Logarithms

We can rearrange the final equation as **log *y* = log *a* + (log *b*)*x***. Notice that log *a* and log *b* are constants because *a* and *b* are constants.

✓ So the equation gives a linear model relating the explanatory variable *x* to the transformed variable log *y*.

Thus, if the relationship between two variables follows an exponential model, and we plot the logarithm (base 10 or base ***e***) of *y* against *x*, we should observe a ***straight-line pattern*** in the transformed data.

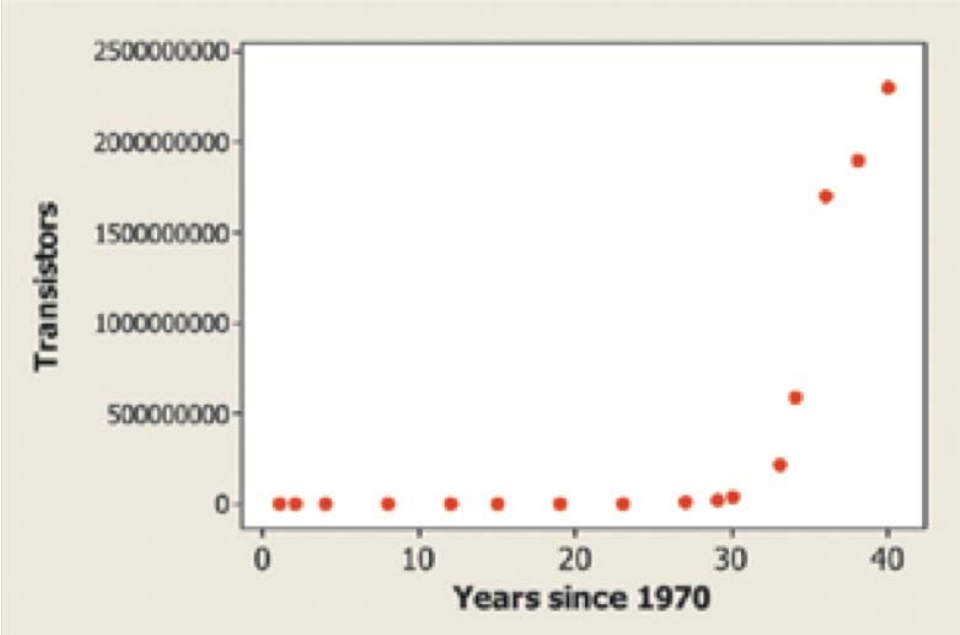
If we fit a least-squares regression line to the transformed data, we can find the predicted value of the logarithm of *y* for any value of the explanatory variable *x* by substituting our *x*-value into the equation of the line.

✓ To obtain the corresponding prediction for the response variable *y*, we have to "undo" the logarithm transformation to return to the original units of measurement. One way of doing this is to use the definition of a logarithm as an exponent:

$$\log_b a = x \Rightarrow b^x = a$$

Example: Moore's Law and Computer Chips

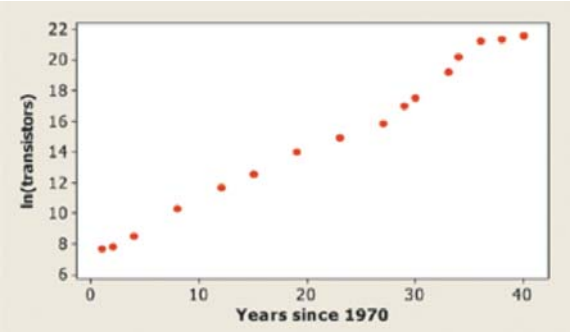
Gordon Moore, one of the founders of Intel Corporation, predicted in 1965 that the number of transistors on an integrated circuit chip would double every 18 months. This is Moore's law, one way to measure the revolution in computing. Here are data on the dates and number of transistors for Intel microprocessors:



Processor	Date	Transistors
4004	1971	2,250
8008	1972	2,500
8080	1974	5,000
8086	1978	29,000
286	1982	120,000
386	1985	275,000
486 DX	1989	1,180,000
Pentium	1993	3,100,000
Pentium II	1997	7,500,000
Pentium III	1999	24,000,000
Pentium 4	2000	42,000,000
Itanium 2	2003	220,000,000
Itanium 2 w/9MB cache	2004	592,000,000
Dual-core Itanium 2	2006	1,700,000,000
Six-core Xeon 7400	2008	1,900,000,000
8-core Xeon Nehalem-EX	2010	2,300,000,000

Example: Moore's Law and Computer Chips

(a) A scatterplot of the natural logarithm (log base *e* or ln) of the number of transistors on a computer chip versus years since 1970 is shown. Based on this graph, explain why it would be reasonable to use an exponential model to describe the relationship between number of transistors and years since 1970.



If an exponential model describes the relationship between two variables *x* and *y*, then we expect a scatterplot of (*x*, ln *y*) to be roughly linear. the scatterplot of ln(transistors) versus years since 1970 has a fairly linear pattern, especially through the year 2000. So an exponential model seems reasonable here.

(b) Minitab output from a linear regression analysis on the transformed data is shown below. Give the equation of the least-squares regression line. Be sure to define any variables you use.

Predictor	Coef	SE Coef	T	P
Constant	7.0647	0.2672	26.44	0.000
Years since 1970	0.36583	0.01048	34.91	0.000
S = 0.544467 R-Sq = 98.9% R-Sq(adj) = 98.8%				

$$\widehat{\ln(\text{transistors})} = 7.0647 + 0.36583(\text{years since 1970})$$

Example: Moore's Law and Computer Chips

(c) Use your model from part (b) to predict the number of transistors on an Intel computer chip in 2020. Show your work.

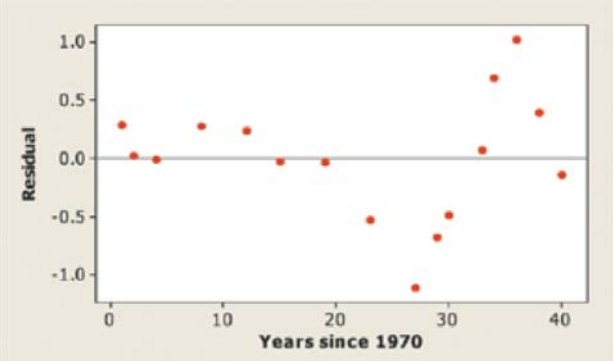
$$\widehat{\ln(\text{transistors})} = 7.0647 + 0.36583(\text{years since 1970})$$
$$= 7.0647 + 0.36583(50) = 25.3562$$

$$\log_b a = x \Rightarrow b^x = a$$

$$\widehat{\ln(\text{transistors})} = 25.3562 \Rightarrow \widehat{\log_e(\text{transistors})} = 25.362$$

$$\widehat{\text{transistors}} = e^{25.362} \approx 1.028 \cdot 10^{11}$$

(d) A residual plot for the linear regression in part (b) is shown below. Discuss what this graph tells you about the appropriateness of the model.



The residual plot shows a distinct pattern, with the residuals going from positive to negative to positive as we move from left to right. But the residuals are small in size relative to the transformed *y*-values. Also, the scatterplot of the transformed data is much more linear than the original scatterplot. We feel reasonably comfortable using this model to make predictions about the number of transistors on a computer chip.

Power Models Again

When we apply the logarithm transformation to the response variable y in an exponential model, we produce a linear relationship. To achieve linearity from a power model, we apply the logarithm transformation to both variables. Here are the details:

1.A power model has the form $y = ax^p$, where a and p are constants.

2.Take the logarithm of both sides of this equation. Using properties of logarithms,

$$\log y = \log(ax^p) = \log a + \log(x^p) = \log a + p \log x$$

The equation $\log y = \log a + p \log x$ shows that taking the logarithm of both variables results in a linear relationship between $\log x$ and $\log y$.

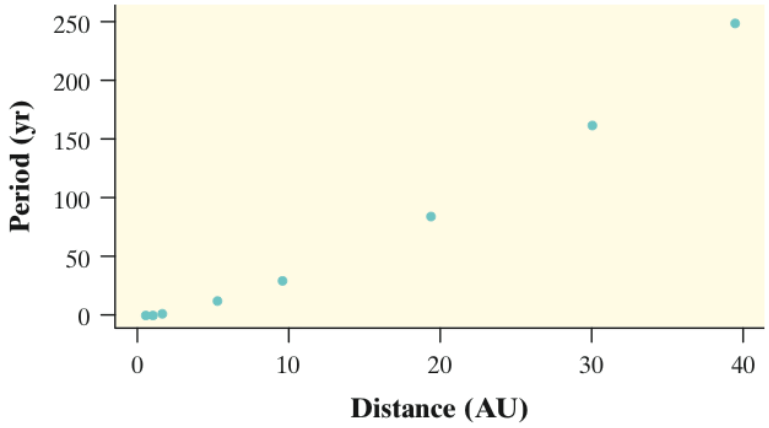
3. Look carefully: the *power* p in the power model becomes the *slope* of the straight line that links $\log y$ to $\log x$.

If a power model describes the relationship between two variables, a scatterplot of the logarithms of *both* variables should produce a linear pattern. Then we can fit a least-squares regression line to the transformed data and use the linear model to make predictions.

Example: What's a Planet, Anyway?

On July 31, 2005, a team of astronomers announced that they had discovered what appeared to be a new planet in our solar system. Originally named UB313, the potential planet is bigger than Pluto and has an average distance of about 9.5 billion miles from the sun. Could this new astronomical body, now called Eris, be a new planet? At the time of the discovery, there were nine known planets in our solar system. Here are data on the distance from the sun (in astronomical units, AU) and period of revolution of those planets.

Planet	Distance from sun (astronomical units)	Period of revolution (Earth years)
Mercury	0.387	0.241
Venus	0.723	0.615
Earth	1.000	1.000
Mars	1.524	1.881
Jupiter	5.203	11.862
Saturn	9.539	29.456
Uranus	19.191	84.070
Neptune	30.061	164.810
Pluto	39.529	248.530

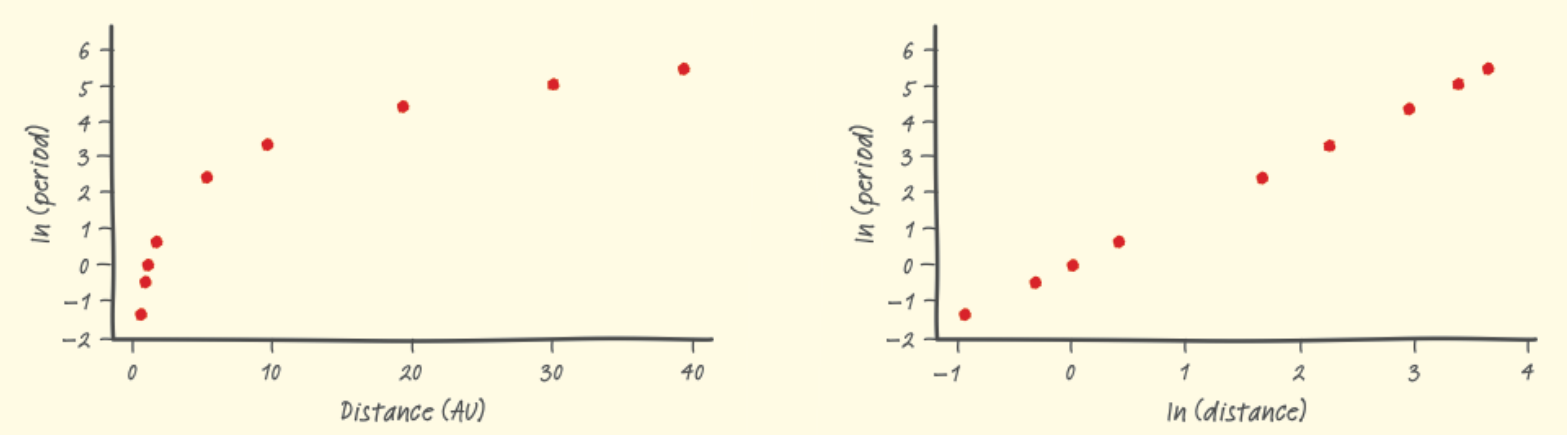


Describe the relationship between distance from the sun and period of revolution.

There appears to be a strong, positive, curved relationship between distance from the sun (AU) and period of revolution (years).

Example: What's a Planet, Anyway?

(a) Based on the scatterplots below, explain why a power model would provide a more appropriate description of the relationship between period of revolution and distance from the sun than an exponential model.



The scatterplot of ln(period) versus distance is clearly curved, so an exponential model would not be appropriate. However, the graph of ln(period) versus ln(distance) has a strong linear pattern, indicating that a power model would be more appropriate.

(b) Minitab output from a linear regression analysis on the transformed data (ln(distance), ln(period)) is shown below. Give the equation of the least-squares regression line. Be sure to define any variables you use.

Predictor	Coef	SE Coef	T	P
Constant	0.0002544	0.0001759	1.45	0.191
ln(distance)	1.49986	0.00008	18598.27	0.000

S = 0.000393364 R-Sq = 100.0% R-Sq(adj) = 100.0%

$$\widehat{\ln(\text{period})} = 0.0002544 + 1.49986(\ln(\text{distance}))$$

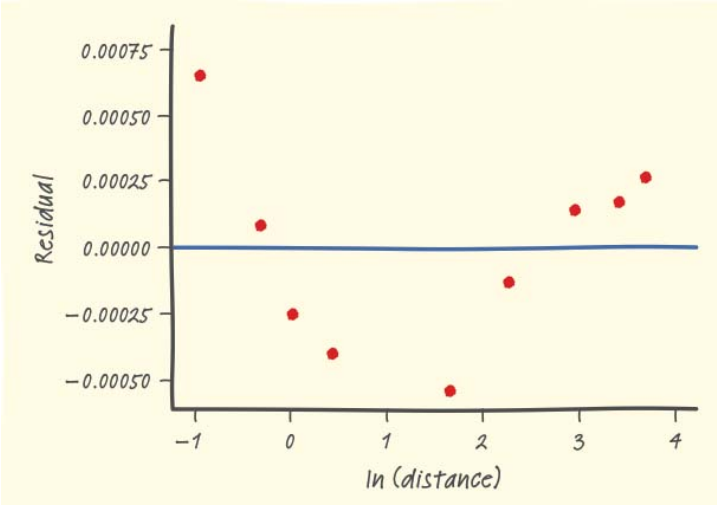
Example: What's a Planet, Anyway?

(c) Use your model from part (b) to predict the period of revolution for Eris, which is 9,500,000,000/93,000,000 = 102.15 AU from the sun. Show your work.

$$\widehat{\ln(\text{period})} = 0.0002544 + 1.49986(\text{distance})$$
$$= 0.0002544 + 1.49986(102.15)$$
$$= 6.939$$

$$\widehat{\text{period}} = e^{6.939} \approx 1032 \text{ years}$$

(d) A residual plot for the linear regression in part (b) is shown below. Do you expect your prediction in part (c) to be too high, too low, or just right? Justify your answer.



Eris's value for ln(distance) is 6.939, which would fall at the far right of the residual plot, where all the residuals are positive.

Because residual = actual *y* - predicted *y* seems likely to be positive, we would expect our prediction to be too low.

Section 12.2
Transforming to Achieve Linearity

Summary

In this section, we learned that…

- ✓ Nonlinear relationships between two quantitative variables can sometimes be changed into linear relationships by **transforming** one or both of the variables. Transformation is particularly effective when there is reason to think that the data are governed by some nonlinear mathematical model.

- ✓ When theory or experience suggests that the relationship between two variables follows a power model of the form $y = ax^p$, there are two transformations involving powers and roots that can linearize a curved pattern in a scatterplot.

  Option 1: Raise the values of the explanatory variable x to the power p, then look at a graph of (x^p, y).

  Option 2: Take the p^{th} root of the values of the response variable y, then look at a graph of (x, p^{th} root of y).

Section 12.2
Transforming to Achieve Linearity

Summary

- In a linear model of the form $y = a + bx$, the values of the response variable are predicted to increase by a constant amount b for each increase of 1 unit in the explanatory variable. For an **exponential model** of the form $y = ab^x$, the predicted values of the response variable are multiplied by an additional factor of b for each increase of one unit in the explanatory variable.

- A useful strategy for straightening a curved pattern in a scatterplot is to take the **logarithm** of one or both variables. To achieve linearity when the relationship between two variables follows an exponential model, plot the logarithm (base 10 or base e) of y against x. When a power model describes the relationship between two variables, a plot of log y (ln y) versus log x (ln x) should be linear.

- Once we transform the data to achieve linearity, we can fit a least-squares regression line to the transformed data and use this linear model to make predictions.