Chapter 12: More about Regression

Objectives: Students will:

Identify the conditions necessary to do *inference for regression*.
Explain what is meant by the *standard error about the least-squares line*.
Given a set of data, check that the *conditions for doing inference for regression* are present.
Compute a *confidence interval for the slope of the regression line*.
Conduct a *test of the hypothesis that the slope of the regression line is 0* (or that the correlation is 0) in the population.

AP Outline Fit:

IV. Statistical Inference: Estimating population parameters and testing hypotheses (30%–40%)

A. Estimation (point estimators and confidence intervals)

8. Confidence interval for the slope of a least-squares regression line

B. Tests of significance

7. Test for the slope of a least-squares regression line

Note: This chapter builds on the material developed in section 3.2.

What you will learn:

A. Preliminaries
   1. Make a scatterplot to show the relationship between an explanatory and a response variable.
   2. Use a calculator or software to find the correlation and the least-squares regression line.

B. Recognition
   1. Recognize the regression setting: a straight-line relationship between an explanatory variable *x* and a response variable *y*.
   2. Recognize which type of inference you need in a particular regression setting.
   3. Inspect the data to recognize situations in which inference isn't safe: a nonlinear relationship, influential observations, strongly skewed residuals in a small sample, or nonconstant variation of the data points about the regression line.

C. Doing Inference Using Software and Calculator Output
   1. Explain in any specific regression setting the meaning of the slope β of the true regression line.
   2. Understand computer output for regression. From the output, find the slope and intercept of the least-squares line, their standard errors, and the standard error about the line.
   3. Use the output to carry out tests and calculate confidence intervals for β.

Chapter 12: More about Regression

Chapter 12-1: Inference for Regression (slope)

Knowledge Objectives: Students will:

Check the conditions for performing inference about the slope β_1 of the population (true) regression line
Interpret the values of b_0, b_1, s and SE_{b1} in context, and determine these values from computer output
Construct and interpret a confidence interval for the slope β_1 of the population (true) regression line
Perform a significance test about the slope β_1 of the population (true) regression line

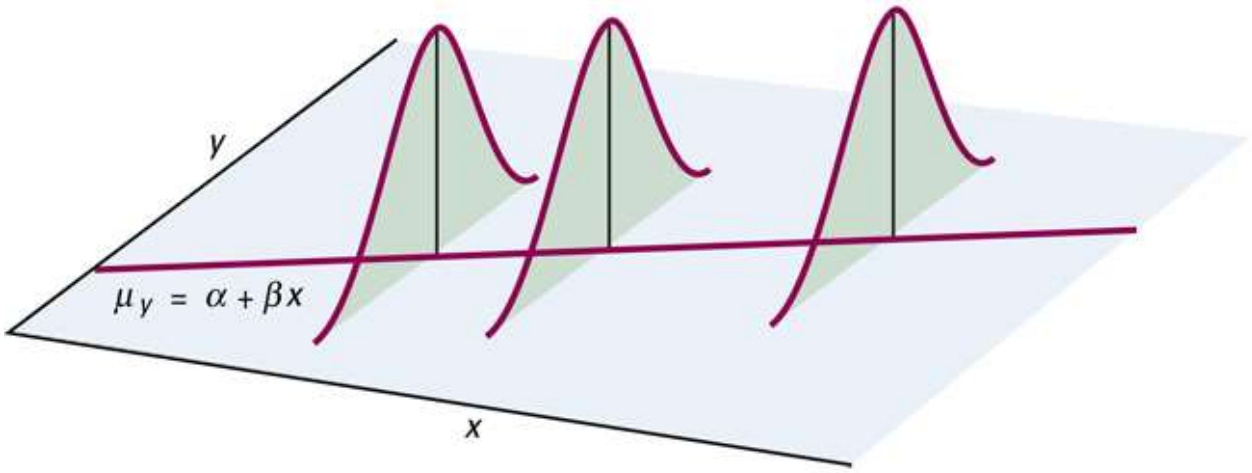
Vocabulary:

Population regression line – cc
Sample regression line – cc

Key Concepts:

Conditions for doing inference on regression coefficient for the slope

- Repeated responses y are independent of each other
- The mean response, μ_y, has a straight-line relationship with x: $\mu_y = \alpha + \beta x$
  where the slope β and intercept α are unknown parameters
- The standard deviation of y (call it σ) is the same for all values of x. The value of σ is unknown.
- For any fixed value of x, the response variable y varies according to a Normal distribution



- L: The true relationship is linear
  - Scatter plot the data to check this
  - Remember the transformations to make non-linear data linear
- I: Independent Observations
  - No repeated observations on the same individual
- N: Response varies Normally about the true regression line
  - To check this, we look at the residuals (since they must be Normally distributed as well) either with a box plot or normality plot (a post-hoc – after the test procedure)
  - These procedures are robust, so slight departures from Normality will not affect the inference
- E: Equal response standard deviation everywhere
  - Check the scatter plot to see if this is violated
- R: Random sample is needed for inference testing.
  - Often not the case in regression settings, where researchers often fix in advance the values of x being tested

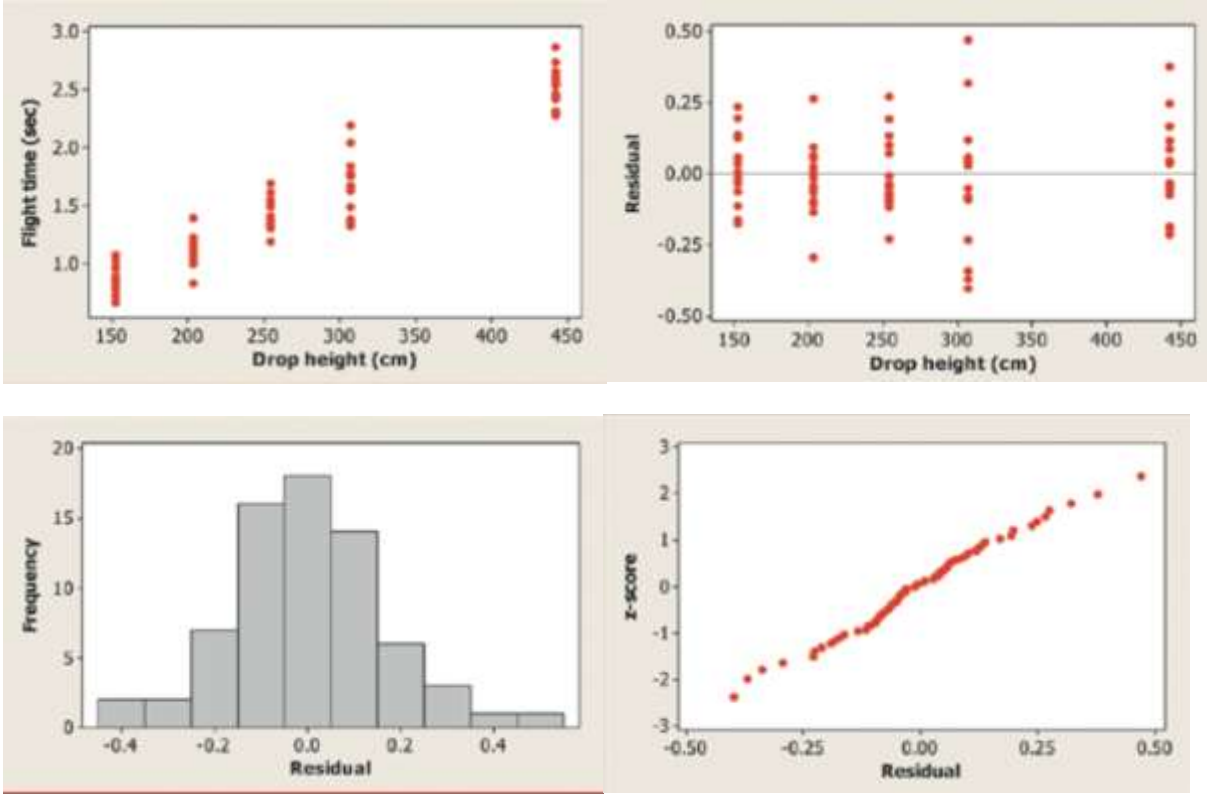
Inference Testing

- Since the null hypothesis cannot be proved, our hypotheses for tests on the regression slope will be:
  H_0: $\beta = 0$ (no correlation between x and y)
  H_a: $\beta \neq 0$ (some linear correlation)

Chapter 12: More about Regression

Example: Mrs. Barrett's class did a variation of the helicopter experiment on page 738. Students randomly assigned 14 helicopters to each of five drop heights: 152 centimeters (cm), 203 cm, 254 cm, 307 cm, and 442 cm. Teams of students released the 70 helicopters in a predetermined random order and measured the flight times in seconds.

The class used Minitab to carry out a least-squares regression analysis for these data. A scatterplot, residual plot, histogram, and Normal probability plot of the residuals are shown below. Construct and interpret a 95% confidence interval for the slope of the population regression line.



Conditions:

Regression Analysis: Flight time (sec) versus Drop height (cm)

Predictor	Coef	SE Coef	T	P
Constant	−0.03761	0.05838	−0.64	0.522
Drop height (cm)	0.0057244	0.0002018	28.37	0.000

S = 0.168181 R-Sq = 92.2% R-Sq(adj) = 92.1%

Interpretation:

Chapter 12: More about Regression

Example: Infants who cry easily may be more easily stimulated than others. This may be a sign of higher IQ. Child development researchers explored the relationship between the crying of infants 4 to 10 days old and their later IQ test scores. A snap of a rubber band on the sole of the foot caused the infants to cry. The researchers recorded the crying and measured its intensity by the number of peaks in the most active 20 seconds. They later measured the children's IQ at age three years using the Stanford-Binet IQ test. A scatterplot and Minitab output for the data from a random sample of 38 infants is below.

Do these data provide convincing evidence that there is a positive linear relationship between crying counts and IQ in the population of infants? We want to perform a test of

$H_0 : \beta = 0$

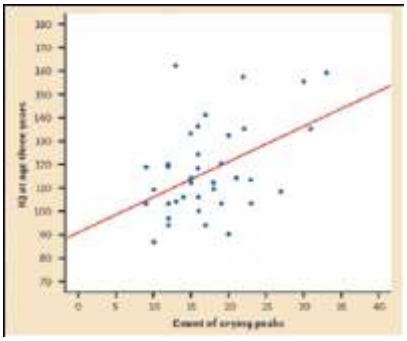
$H_a : \beta > 0$

where β is the true slope of the population regression line relating crying count to IQ score. No significance level was given, so we'll use $\alpha = 0.05$.

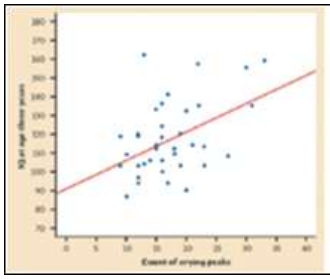
Regression Analysis: IQ versus Crycount

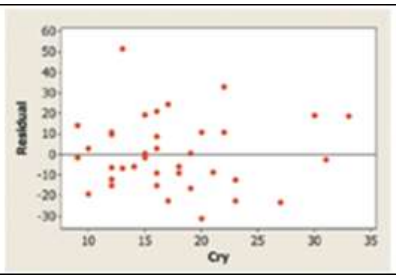
Predictor	Coef	SE Coef	T	P
Constant	91.268	8.934	10.22	0.000
Crycount	1.4929	0.4870	3.07	0.004

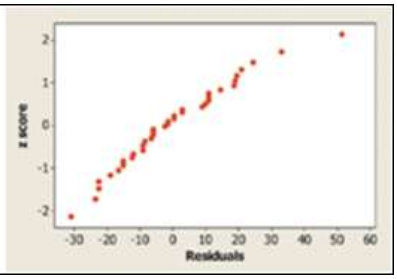
S = 17.50 R-Sq = 20.7% R-Sq(adj) = 18.5%



Conditions:





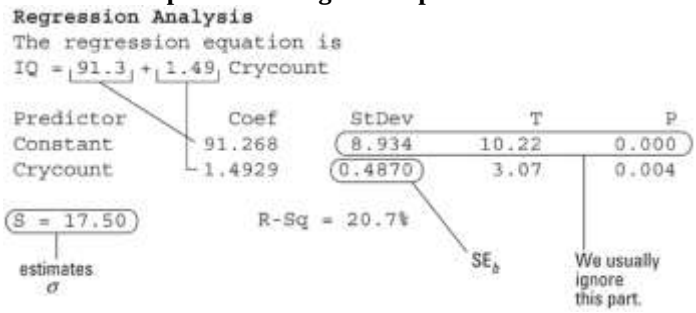


Calculations:

Interpretation:

Chapter 12: More about Regression

MINITAB output from a regression problem:

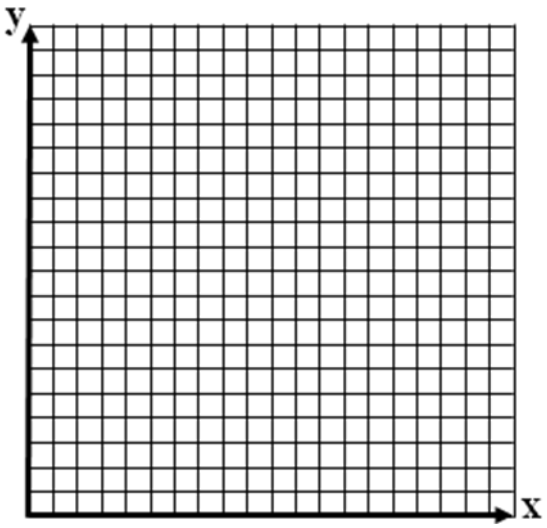


Example: 16 student volunteers at Ohio State drank a randomly assigned number of cans of beer. Thirty minutes later, a police officer measured their BAC. Here are the data:

Student	1	2	3	4	5	6	7	8
Beers	5	2	9	8	3	7	3	5
BAC	0.10	0.03	0.19	0.12	0.04	0.095	0.07	0.06
Student	9	10	11	12	13	14	15	16
Beers	3	5	4	6	5	7	1	4
BAC	0.02	0.05	0.07	0.10	0.085	0.09	0.01	0.05

Enter the data into your calculator.

a) Draw a scatter plot of the data and the regression line
LinReg L1, L2, Y1



b) Conduct an inference test on the effect of beers on BAC using your calculator

t =
p =
df =
a =
b =
s =
r^2 =
r =

Minitab Output:

```
The regression equation is
BAC = - 0.0127 + 0.0180 Beers


Predictor       Coef      StDev          T          P
Constant   -0.01270    0.01264      -1.00      0.332
Beers       0.017964   0.002402      7.48      0.000

S = 0.02044     R-Sq = 80.0%
```

Chapter 12: More about Regression

Sample Computer Output:
Problem 15-10:

Microsoft Excel

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.6021					
5	R Square	0.4652					
6	Adjusted R Square	0.4270					
7	Standard Error	0.1886					
8	Observations	16.0000					
9							
10		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
11	Intercept	0.1205	0.0927	1.2999	0.2146	-0.0783	0.3193
12	Perch	0.0086	0.0025	3.4899	0.0036	0.0033	0.0138
13							

Sheet5 Sheet1 Sheet2 Sheet2

Problem 15-20:

CrunchIt!

Simple Linear Regression

Simple linear regression results:
Dependent Variable: veloc
Independent Variable: thick
veloc = 70.436874 + 274.7821 thick
Sample size: 12
R (correlation coefficient) = 0.7019
R-sq = 0.49266976
Estimate of error standard deviation: 56.364124

Parameter estimates:

Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value
Intercept	70.436874	52.898945			
Slope	274.7821	88.17712			

Predicted values:

X value	Pred. Y	s.e.(Pred.y)	95% C.I.	95% P.I.
0.5	207.82793	17.428537	(168.99474, 246.66113)	(76.374, 339.28186)

Summary:

- Inference Conditions Needed (LINER):
  1) Linear Model appropriate
  2) Independent Observations
  3) Normally distributed response values about a given x
  4) Equal standard deviation (or spread); σ is constant
  5) Randomly selected values
- Confidence Intervals on β can be done
- Computer output needs to be understood
- Inference testing on β use the t statistic = b/SE_b

Homework: Problems 1, 5, 7, 11, 15

Chapter 12-2: Transforming to Achieve Linearity

Objectives: Students will be able to:

- Use transformations involving powers and roots to find a power model that describes the relationship between two quantitative variables, and use the model to make predictions
- Use transformations involving logarithms to find a power model that describes the relationship between two quantitative variables, and use the model to make predictions
- Use transformations involving logarithms to find an exponential model that describes the relationship between two quantitative variables, and use the model to make predictions
- Determine which of several transformations does a better job of producing a linear relationship

Vocabulary: None new

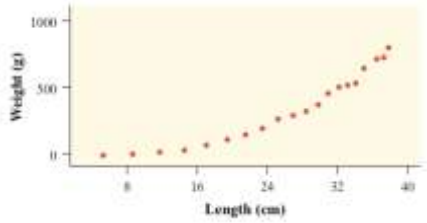
Key Concept:
Although a **power model** of the form $y = ax^p$ describes the relationship between x and y in this setting, there is a *linear* relationship between x^p and y.

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.

Example 1: 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



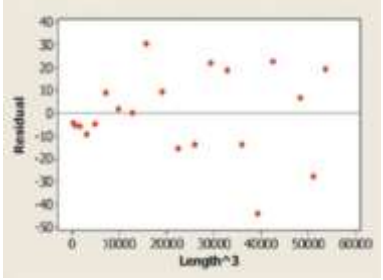
Does the data look linear?

Output from Minitab

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

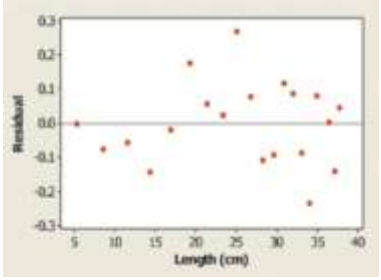
S = 18.8412 R-Sq = 99.5% R-Sq(adj) = 99.5%



Transformation 2: (length, ∛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%



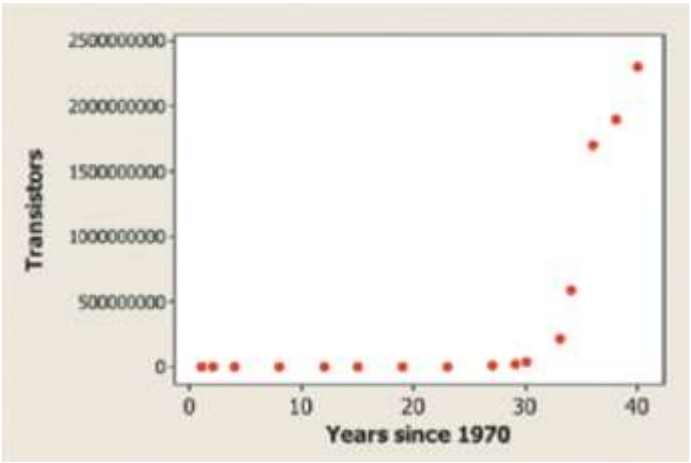
(a) Give the equation of the least-squares regression line. Define any variables you use.

(b) Suppose a contestant in the fishing tournament catches an Atlantic ocean rockfish that's 36 centimeters long. Use the model from part (a) to predict the fish's weight. Show your work.

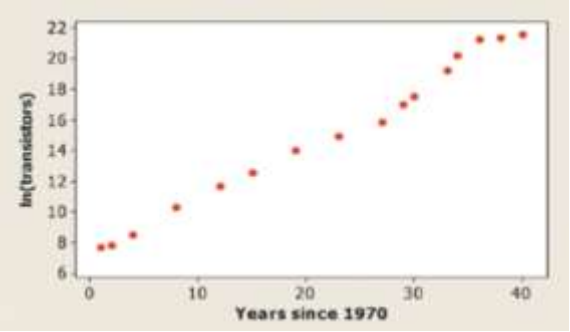
(c) Interpret the value of *s* in context.

Example 2: 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



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.



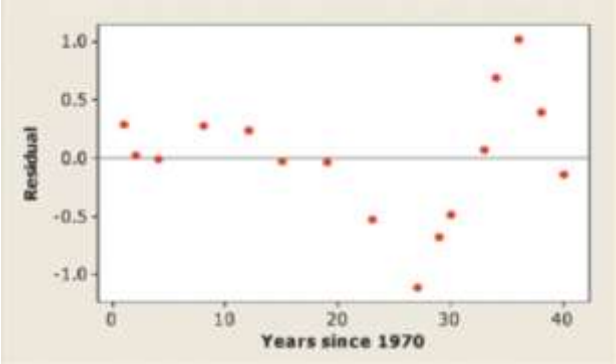
(b) Minitab output from a linear regression analysis on the transformed data is shown below. Give the equation of the least-squares regression line. Be sure to define any variables you use.

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%

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

(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.



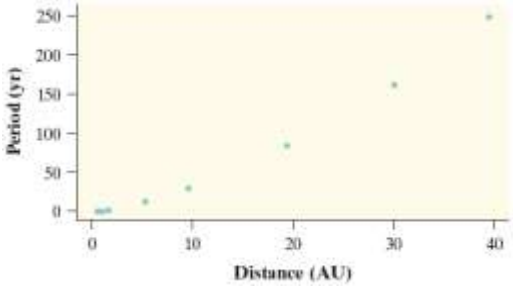
Example 3: Eris

On July 31, 2005, a team of astronomers announced that they had discovered what appeared to be a new planet in our solar system. They had first observed this object almost two years earlier using a telescope at Caltech's Palomar Observatory in California.

Originally named UB313, the potential planet is bigger than Pluto and has an average distance of about 6.3 billion miles from the sun. (For reference, Earth is about 93 million miles from the sun.)

Could this new astronomical body, now called Eris, be a new planet?

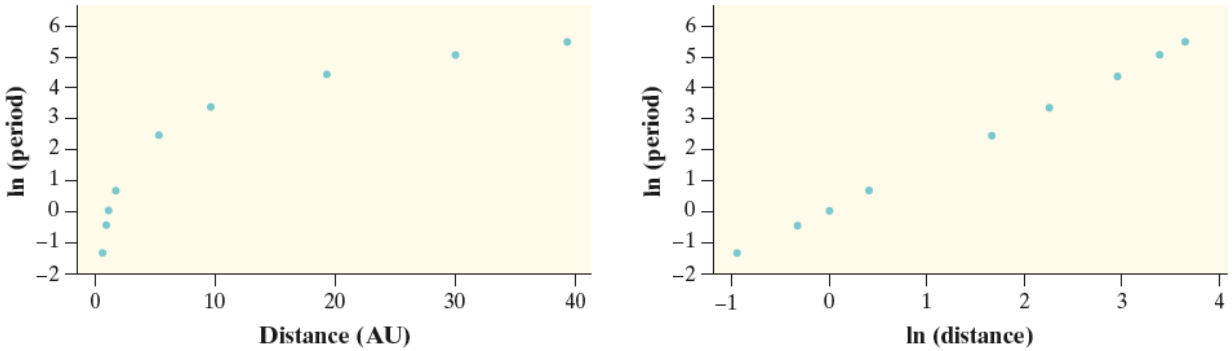
At the time of the discovery, there were nine known planets in our solar system. Here is a scatterplot on the distance from the sun and period of revolution of those planets. Note that distance is measured in astronomical units (AU), the number of Earth distances the object is from the sun. There appears to be a strong curved relationship between distance from the sun and period of revolution.



Chapter 12: More about Regression

The following graphs show the results of two different transformations of the data.

- The graph on the left plots the natural logarithm of period against distance from the sun for all nine planets.
- The graph on the right plots the natural logarithm of period against the natural logarithm of distance from the sun for the nine planets.



(a) Based on the scatterplots, would an exponential model or a power model provide a better description of the relationship between distance and period? Justify your answer.

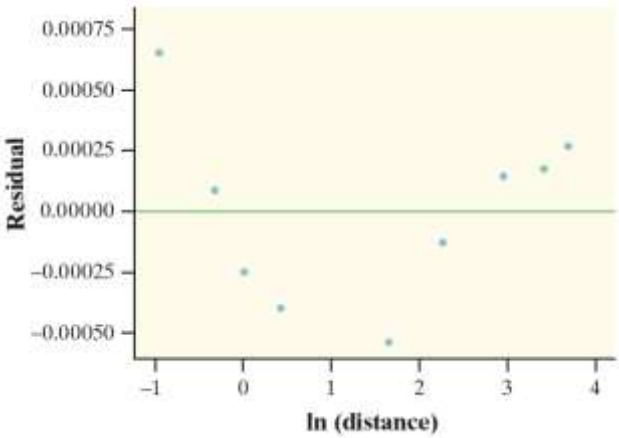
(b) Here is computer output from a linear regression analysis of the transformed data in the graph above. Give the equation of the least-squares regression line. Be sure to define any variables you use.

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%

(c) Use your model from part (b) to predict the period of revolution for Eris, which is about 68.05 AU from the sun

(d) Here is a residual plot for the linear regression in part (b). Do you expect your prediction in part (c) to be too large, too small, or about right? Justify your answer.



(e) Based on the residual graph to the right. Is the power model a good model?

Summary:

- 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: look at a graph of (x^p, y).
  - Option 2: Take the pth root of the values of the response variable y: look at a graph of (x, pth root of y).
- Power: if we use Ln(y) = a + bLn(x) to linearize the data, then a power function model is appropriate
  - $Y = (e^a)x^b$
- Exponential: if we use just Ln(y) = a + bx to linearize the data, then an exponential function model is appropriate
  - $Y = ab^x$

Homework: Problems 33, 35, 37, 43, 47

Chapter 12: Review

Objectives: Students will be able to:

Summarize the chapter
Define the vocabulary used
Know and be able to discuss all sectional knowledge objectives
Complete all sectional construction objectives
Successfully answer any of the review exercises

Identify the conditions necessary to do *inference for regression*.
Explain what is meant by the *standard error about the least-squares line*.
Given a set of data, check that the *conditions for doing inference for regression* are present.
Compute a *confidence interval for the slope of the regression line*.
Conduct a *test of the hypothesis that the slope of the regression line is 0* (or that the correlation is 0) in the population.

Vocabulary: None new