

Freeport AP Statistics

Chapter 12 : More About Regression

12.1 Inference for Linear Regression

OBJECTIVE(S):

- Students will learn how to check conditions for performing inference about the slope β of the population (true) regression line.
- Students will learn how to interpret computer output from a least-squares regression analysis.
- Students will learn how to construct and interpret a confidence interval for the slope β of the population (true) regression line.
- Students will learn how to perform a significance test about the slope β of a population (true) regression line.

	Parameter	Statistic
Slope		
y-intercept		

How to Check the Conditions for Regression Inference

- **Linear**

- **Independent**

- **Normal**

- **Equal SD**

- **Random**

Use the acronym _____ to remember the conditions for regression inference.

1. How well does the number of beers a person drinks predict his or her blood alcohol content (BAC)? Sixteen volunteers aged 21+ with an initial BAC of 0 drank a randomly assigned number of cans of beer. Thirty minutes later, a police officer measured their BAC. Least-squares regression was performed on the data. A residual plot and a histogram of the residuals are shown on **p. 759 #3**.

- a. Check whether the conditions for performing inference about the regression model are met.

- **Linear:**

- **Independent:**

- **Normal:**

- **Equal variance:**

- **Random:**

- b. p. 760 #6 The model for regression inference has three parameters: α , β , and σ . Explain what each parameter represents in context. Then provide an estimate for each

2. In Chapter 3, we examined data on the percent of high school graduates in each state who took the SAT and the state's mean SAT Math score in a recent year. The figure given on **p. 759 #2** shows a residual plot for the least-squares regression line based on these data. Are the conditions for performing inference about the slope β of the population regression line met? Justify your answer.
3. Many people believe that students learn better if they sit closer to the front of the classroom. Does sitting closer cause higher achievement, or do better students simply choose to sit in the front? To investigate, an AP Statistics teacher randomly assigned students to seat locations in his classroom for a particular chapter and recorded the test score for each student at the end of the chapter. The explanatory variable in this experiment is which row the student was assigned (Row 1 is closest to the front and Row 7 is the farthest away). Here are the results:

Row 1: 76, 77, 94, 99

Row 2: 83, 85, 74, 79

Row 3: 90, 88, 68, 78

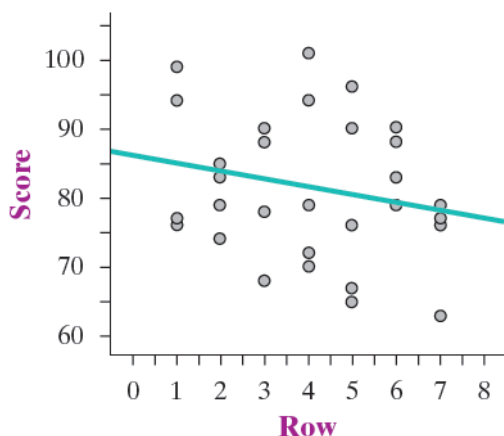
Row 4: 94, 72, 101, 70, 79

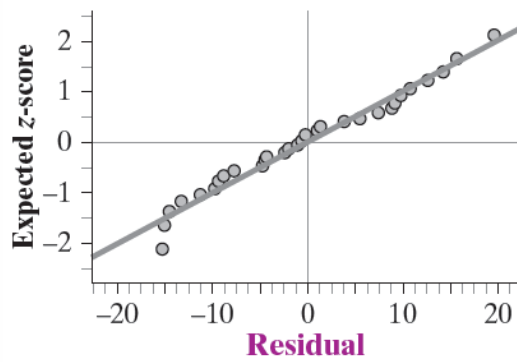
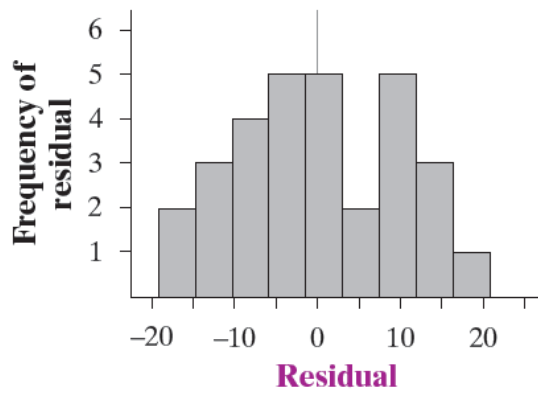
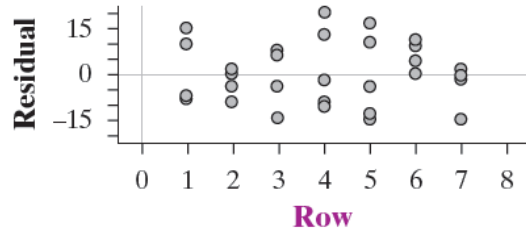
Row 5: 76, 65, 90, 67, 96

Row 6: 88, 79, 90, 83

Row 7: 79, 76, 77, 63

Check whether the conditions for performing inference about the regression model are met. A scatterplot, residual plot, histogram, and Normal probability plot of the residuals are shown.





- **Linear:**
- **Independent:**

- **Normal:**

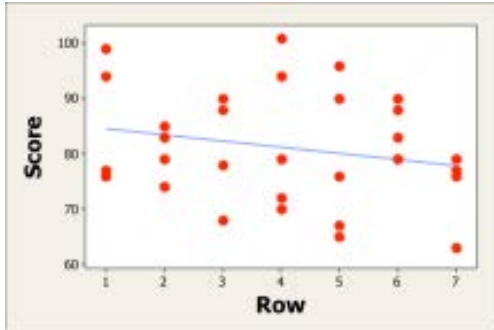
- **Equal variance:**

- **Random:**

4. What do we mean when we say our statistic, b , is unbiased.

t Interval for the Slope of a Least-Squares Regression Line

5. In a previous alternate example, we looked at the results of an experiment designed to see if sitting closer to the front of a classroom causes higher achievement. We checked the conditions for inference and there were no major violations. Here is a scatterplot of the data and some output from a regression analysis.



Regression Analysis: Score versus Row

Predictor	Coef	SE Coef	T	P
Constant	85.706	4.239	20.22	0.000
Row	-1.1171	0.9472	-1.18	0.248

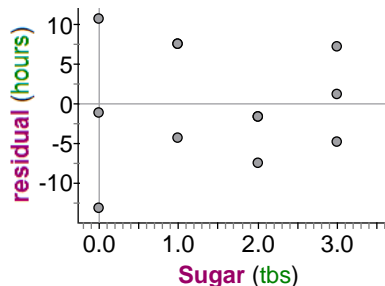
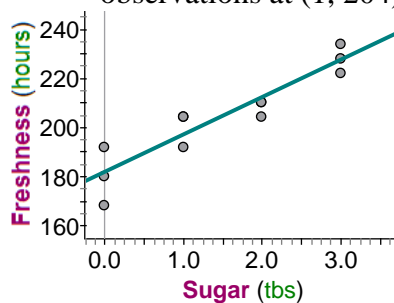
S = 10.0673 R-Sq = 4.7% R-Sq(adj) = 1.3%

- Identify the standard error of the slope SE_b from the computer output. Interpret this value in context.
- Calculate the 95% confidence interval for the true slope. Show your work.
- Interpret the interval from part b. in context.

- d. Based on your interval, is there convincing evidence that seat location affects scores?
6. For their second-semester project, two AP Statistic students decide to investigate the effect of sugar on the life of cut flowers. They went to the local grocery store and randomly selected 12 carnations. All the carnations seemed equally healthy when they were selected. When the students got home, they prepared 12 identical vases with exactly the same amount of water in each vase. They put one tablespoon of sugar in 3 vases, two tablespoons of sugar in 3 vases, and three tablespoons of sugar in 3 vases. In the remaining 3 vases, they put no sugar. After the vases were prepared and placed in the same location, the students randomly assigned one flower to each vase and observed how many hours each flower continued to look fresh. Here are the data

Sugar (tbs.)	Freshness (hours)
0	168
0	180
0	192
1	192
1	204
1	204
2	204
2	210
2	210
3	222
3	228
3	234

A scatterplot, residual plot, and computer output from the regression are shown. Only 10 points appear on the scatterplot and residual plot since there were two observations at (1, 204) and two observations at (2, 210).



CHAPTER 12

Predictor	Coef	SE Coef	T	P
Constant	181.200	3.635	49.84	0.000
Sugar	15.200	1.943	7.82	0.000

$S = 7.52596$ $R\text{-Sq} = 86.0\%$ $R\text{-Sq}(\text{adj}) = 84.5\%$

- a. Construct and interpret 99% confidence interval for the slope of the true regression line.

*P*arameter of interest

Are the conditions met?

- **Linear:**
- **Independent:**
- **Normal:**
- **Equal variance:**

- **Random:**

Name of the interval

Interval

Conclusion

- b. Would you feel confident predicting the hours of freshness if 10 tablespoons of sugar are used? Explain.

t Test for the Slope of the Population Regression Line

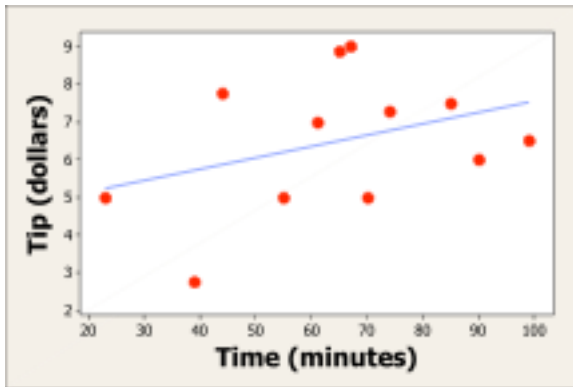
How do we write a null hypothesis that says there is no linear relationship between x and y in the population.?

7. Do customers who stay longer at buffets give larger tips? Charlotte, an AP Statistics student who worked at an Asian buffet, decided to investigate this question for her second-semester project. While she was doing her job as a hostess, she obtained a random sample of receipts, which included the length of time (in minutes) the party was in the restaurant and the amount of the tip (in dollars). Do these data provide convincing evidence that customers who stay longer give larger tips? Here are the data:

Time (minutes)	Tip (dollars)
23	5.00
39	2.75
44	7.75
55	5.00
61	7.00
65	8.88
67	9.01
70	5.00
74	7.29
85	7.50
90	6.00
99	6.50

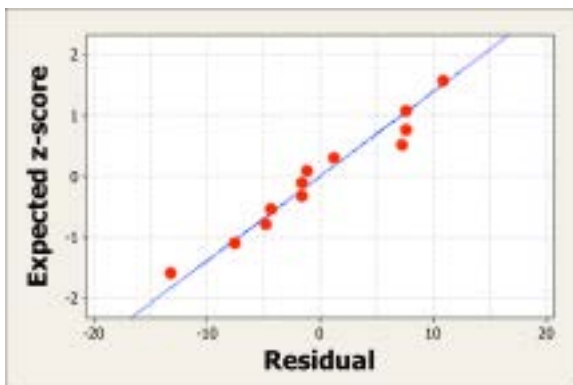
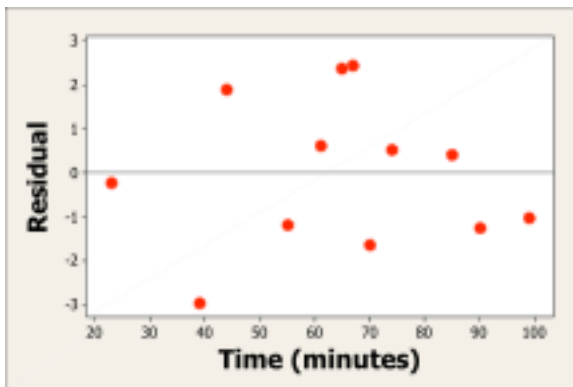
a.

Using the below scatterplots and statistical output. Describe what this graph tells you about the relationship between the two variables.



Predictor	Coef	SE Coef	T	P
Constant	4.535	1.657	2.74	0.021
Time (minutes)	0.03013	0.02448	1.23	0.247

S = 1.77931 R-Sq = 13.2% R-Sq(adj) = 4.5%



- b. What is the equation of the least-squares regression line for predicting the amount of the tip from the length of the stay? Define any variables you use.
- c. Interpret the slope and y intercept of the least-squares regression line in context.
- d. Carry out an appropriate test to answer Charlotte's question.
- e.

Parameter of interest

Hypothesis

$H_0 :$

$H_a :$

Are the conditions met?

- **Linear:**
- **Independent:**
- **Normal:**
- **Equal variance:**
- **Random:**

Name of the test

Test of statistic

Obtain a *P*-value.

P-value =

Make a decision about null

State your conclusion (H_a in context of the problem)

8. A random sample of 11 used Honda CR-Vs from the 2002-6 model years was selected from the inventory at www.carmax.com. The number of miles driven and the advertised price were recorded for each CR-V. A 95% confidence interval for the slope of the true least-squares regression line for predicting advertised price from number of miles (in thousands) driven is (-50.1, -122.3). Based on this interval, what conclusion should we draw from a test of $H_o : \beta = 0$ versus $H_a : \beta \neq 0$ at the 0.05 significance level?

9. In Chapter 3, we examined data on the body weights and backpack weights of a group of eight randomly selected ninth-grade students at the Webb Schools. See **p. 755** for the Minitab output.
 - a. Do these data provide convincing evidence of a linear relationship between pack weight and body weight in the population of ninth-grade students at the school? Carry out a test at the $\alpha = 0.05$ significance level. Assume that the conditions for regression inference are met.

Parameter of interest

Hypothesis

CHAPTER 12

$H_0 :$

$H_a :$

Are the conditions met?

Name of the test

Test of statistic

Obtain a P -value.

P -value =

Make a decision about null

State your conclusion (H_a in context of the problem)

- b. Would a 99% confidence interval for the slope β of the population regression line include 0? Justify your answer. (You don't need to do any calculations to answer this question.)

10. Here are data on the time (in minutes) Professor Moore takes to swim 2000 yards and his pulse rate (beats per minute) after swimming on a random sample of 23 days:

Time: 34.12	35.72	34.72	34.05	34.13	35.72
Pulse: 152	124	140	152	146	128
Time: 36.17	35.57	35.37	35.57	35.43	36.05
Pulse: 136	144	148	144	136	124
Time: 34.85	34.70	34.75	33.93	34.60	34.00
Pulse: 148	144	140	156	136	148
Time: 34.35	35.62	35.68	35.28	35.97	
Pulse: 148	132	124	132	139	

- a. Is there statistically significant evidence of a negative linear relationship between Professor Moore's swim time and his pulse rate in the population of days on which he swims 2000 yards? Carry out an appropriate significance test at the $\alpha = 0.05$ level.

Parameter of interest

Hypothesis

$H_0 :$

$$H_a :$$

Are the conditions met?

- **Linear:**
- **Independent:**
- **Normal:**
- **Equal variance:**
- **Random:**

Name of the test

Test of statistic

Obtain a P -value.

P -value =

Make a decision about null

State your conclusion (H_a in context of the problem)

- b. Calculate and interpret a 95% confidence interval for the slope β of the population regression line.

Parameter of interest

Are the conditions met?

Name of the interval

Interval

Conclusion

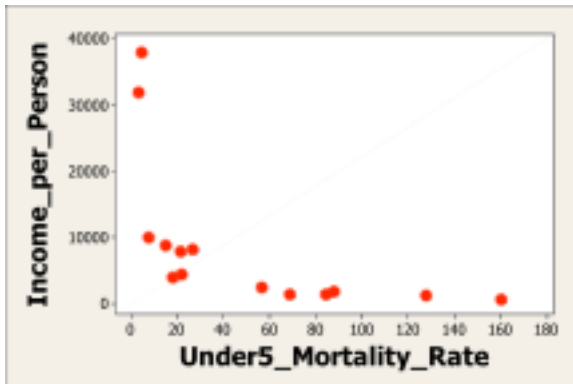
Freeport AP Statistics
 Chapter 12 : More About Regression
12.2 Transforming to Achieve Linearity

OBJECTIVE(S):

- Students will learn how to use transformations involving powers and roots to achieve linearity for a relationship between two variables.
- Students will learn how to use transformations involving logarithms to achieve linearity for a relationship between two variables.
- Students will learn how to make predictions from a least-squares regression line involving transformed data.
- Students will learn how to determine which of several transformations does a better job of producing linear relationship.

11. What does a country's mortality rate for children under five years of age (per 1000 live births) tell us about the income per person (measured in gross domestic product per person adjusted for differences in purchasing power) for residents of that country? Here are the data for a random sample of 14 countries in 2009.

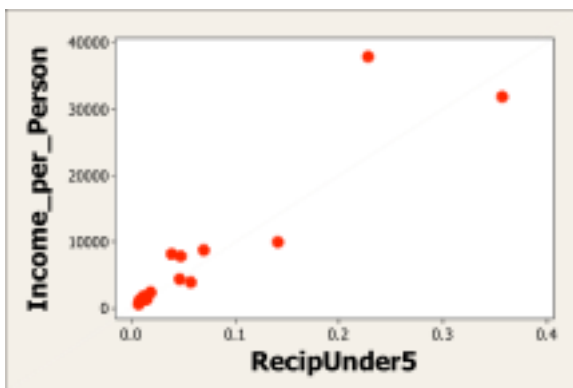
Country	Mortality Rate	Income per Person
Switzerland	4.4	38,003.9
Timor-Leste	56.4	2,475.68
Uganda	127.5	1,202.53
Ghana	68.5	1,382.95
Peru	21.3	7,858.97
Cambodia	87.5	1,830.97
Suriname	26.3	8,199.03
Armenia	21.6	4,523.44
Sweden	2.8	32,021
Niger	160.3	643.39
Serbia	7.1	10,005.2
Kenya	84	1,493.53
Fiji	17.6	4,016.2
Grenada	14.5	8,826.9



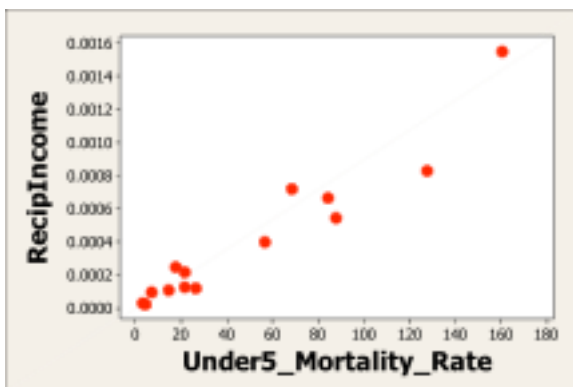
The scatterplot shows a strong negative association that is clearly nonlinear. Because the horizontal and vertical axes look like asymptotes, perhaps there is a reciprocal relationship between the variables, such as

$$\text{income per person} = \frac{a}{\text{child mortality rate}} = a(\text{child mortality rate})^{-1}.$$

If we find the reciprocal of child mortality rate, and graph income versus $1/(\text{child mortality rate})$, we find approx. a linear relationship!



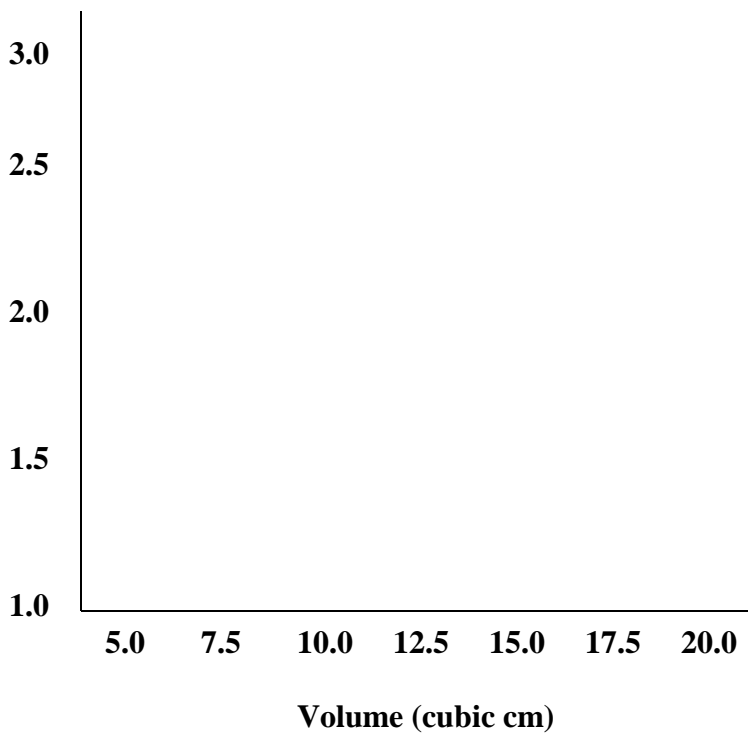
Likewise, we could calculate the reciprocal of the income per person and graph $1/(\text{income per person})$ versus child mortality rate. Again we get a linear relationship!



12. If you have taken a chemistry class, then you are probably familiar with Boyle's law: for gas in a confined space kept at a constant temperature, pressure times volume is a constant (in symbols, $PV = k$). Students collected the following data on pressure and volume using a syringe and a pressure probe.

Volume (cubic centimeters)	Pressure (atmospheres)
6	2.9589
8	2.4073
10	1.9905
12	1.7249
14	1.5288
16	1.3490
18	1.2223
20	1.1201

- a. Make a reasonably accurate scatterplot of the data by hand using volume as the explanatory variable. Describe the bi-variate distribution.

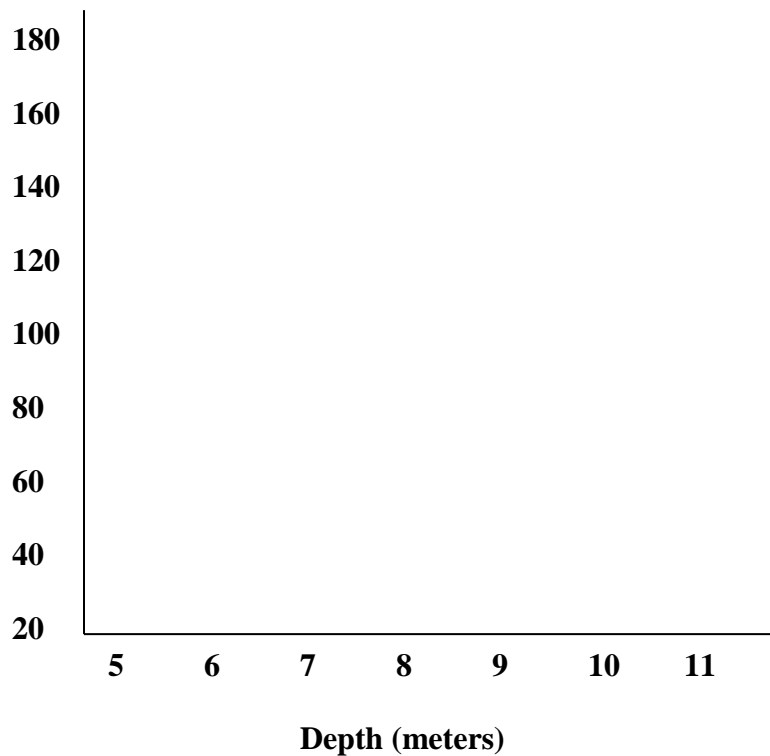


- b. If the true relationship between the pressure and volume of the gas is $PV = k$, we can divide both sides of this equation by V to obtain the theoretical model $P = \frac{k}{V}$, or $P = \frac{1}{V}k$. Use the graph on **p. 786 #32b** to identify the transformation that was used to linearize the curved pattern in part a.
 - c. Use the graph on **p. 786 #32c** to identify the transformation that was used linearize the curved pattern in part a.
13. See **p. 786 #34** for the Minitab output from separate regression analysis of the two sets of transformed pressure data from 12.
 - a. Give the equation of the least-squares regression line. Define any variables you use.
 - b. Use the model from part a. to predict the pressure in the syringe when the volume is 17 cubic centimeters. Show your work.
 - c. Interpret the value of s in context

14. Some college students collected data on the intensity of light at various depths in a lake. Here are their data:

Depth (meters)	Light intensity (lumens)
5	168.00
6	120.42
7	86.31
8	61.87
9	44.34
10	31.78
11	22.78

- a. Make a reasonably accurate scatterplot of the data by hand, using depth as the explanatory variable. Describe what you see.



- b. A scatterplot of the natural logarithm of light intensity versus depth is shown on p. 789 #42b. Based on this graph, explain why it would be reasonable to use an exponential model to describe the relationship between light intensity and depth.

- c. Minitab output from a linear regression analysis on the transformed data is shown on **p. 789 #42c**. Give the equation of the least-squares regression line. Be sure to define any variables you use.
 - d. Use your model to predict the light intensity at a depth of 12 meters. The actual light intensity reading at the depth was 16.2 lumens. Does this surprise you? Explain.
15. It is easy to measure the “diameter at breast height” of a tree. It’s hard to measure the total “aboveground biomass” of a tree, because to do this you must cut and weigh the tree. the biomass is important for studies of ecology, so ecologists commonly estimate it using a power model. Combining data on 378 trees in tropical rain forests gives this relationship between biomass y measured in kilograms and diameter x measured in centimeters:

$$\ln \hat{y} = -2.00 + 2.42 \ln x$$

Use this model to estimate the biomass of a tropical tree 30 centimeters in diameter. Show your work.

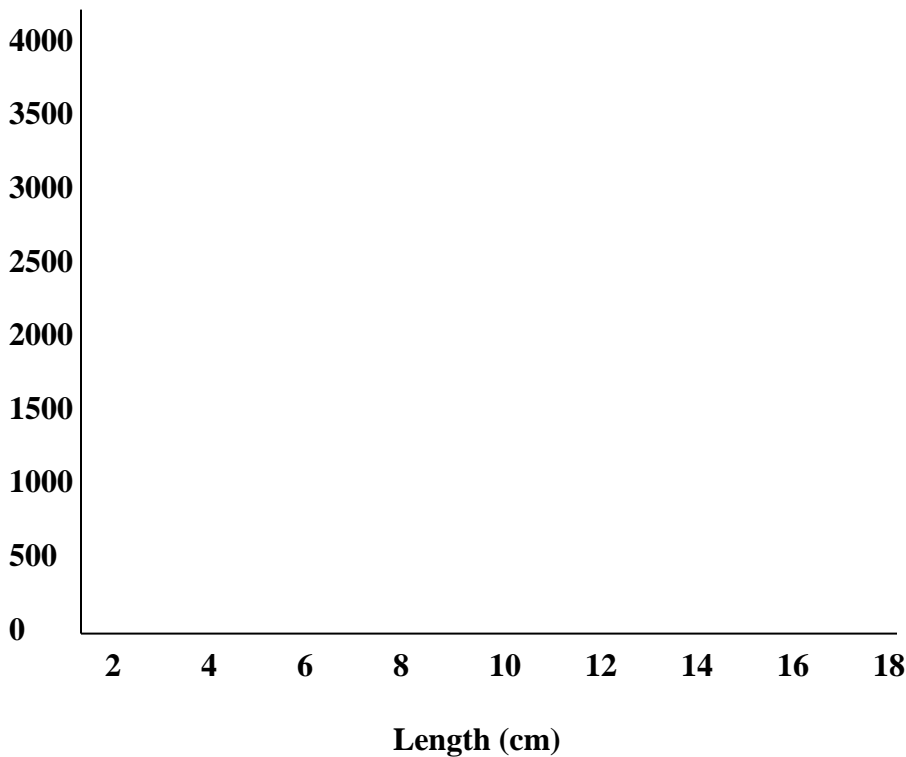
16. See p. 789 #43.

- The following graphs show the results of two different transformations of the data. Would an exponential model or a power model provide a better description of the relationship between bounce number and height? Justify your answer.
- Minitab output from a linear regression analysis on the transformed data of $\log(\text{height})$ versus bounce number is shown. Give the equation of the least-squares regression line. Be sure to define any variables you use.
- Use your model from part b. to predict the highest point the ball reaches on its seventh bounce. Show your work.
- A residual plot for the linear regression in part b. is shown below. Do you expect your prediction part c. to be too high, too low, or just right? Justify your answer.

17. Here are some data on the hearts of various mammals.

Mammal	Heart weight (grams)	Length of cavity of left ventricle (centimeters)
Mouse	0.13	0.55
Rat	0.64	1.0
Rabbit	5.8	2.2
Dog	102	4.0
Sheep	210	6.5
Ox	2030	12.0
Horse	3900	16.0

- a. Make an appropriate scatterplot for predicting heart weight from length. Describe what you see.



- b. Use transformations to linearize the relationship. Does the relationship between heart weight and length seem to follow an exponential model or a power model? Justify your answer.

