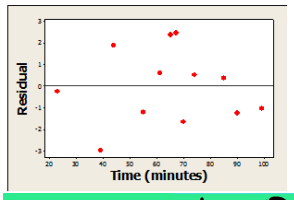
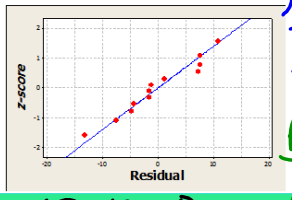
Bellwork

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? Assume all conditions are met.

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%





State $H_0: \beta = 0$ $\alpha = .05$
$H_a: \beta > 0$
β the true slope of the reg. line relating time & tip $.

PLAN - problem assumes conditions met. we will use t test for β.

DO: $t = \frac{b - \beta_0}{SE_b} = \frac{.03013 - 0}{.02448}$

$t = 1.23$

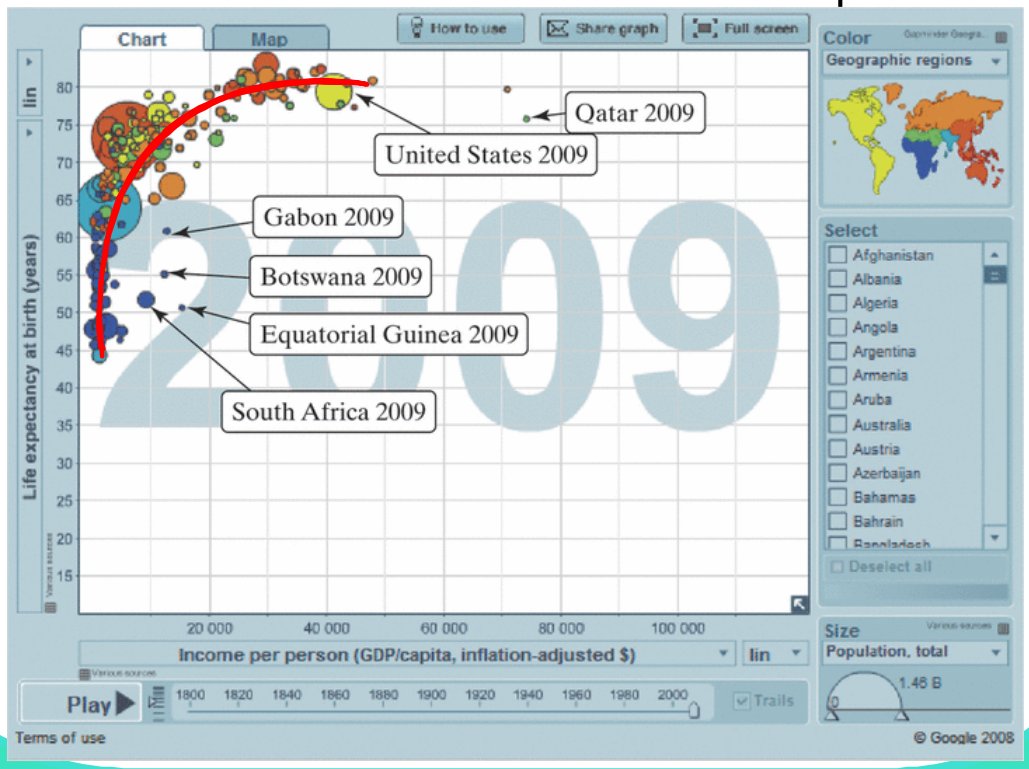
p-value: tcdf(1.23, 1E99, 10) = .1234
lower upper

12.2 Transforming to Achieve Linearity

vocab

examples

What do we do with a curved scatterplot?



What do we do with a curved scatterplot?

Transforming

applying a function, such as the logarithm or square root to a quantitative variable

we will take this approach in order to "straighten" the association

Power Model

takes the form $y=ax^p$

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

$$\text{period} = a\sqrt{\text{length}} = a(\text{length})^{1/2}$$

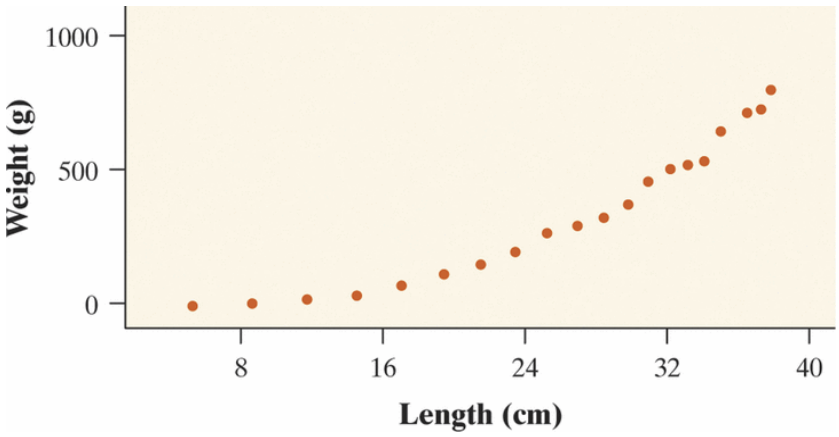
$$\text{intensity} = \frac{a}{\text{distance}^2} = a(\text{distance})^{-2}$$

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. You contact the nearby marine research laboratory, and they provide reference data on the length (in centimeters) and weight (in grams) for Atlantic Ocean rockfish of several sizes.

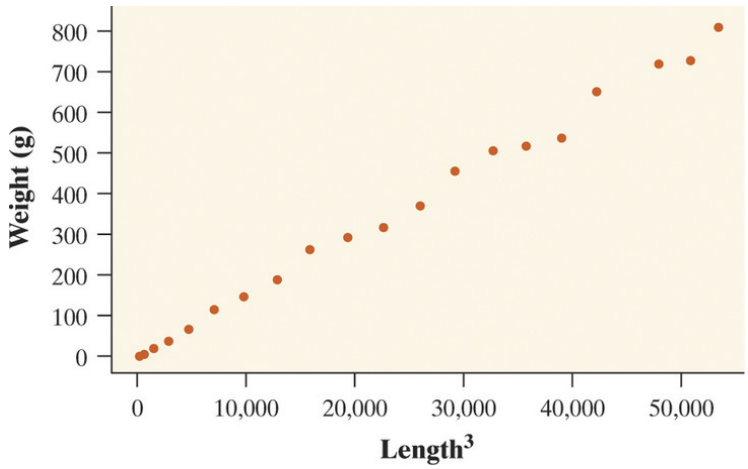
Length:	5.2	8.5	11.5	14.3	16.8	19.2	21.3	23.3	25.0	26.7
Weight:	2	8	21	38	69	117	148	190	264	293
Length:	28.2	29.6	30.8	32.0	33.0	34.0	34.9	36.4	37.1	37.7
Weight:	318	371	455	504	518	537	651	719	726	810

Go Fish!



Because length is one-dimensional and weight (like volume) is three-dimensional, a power model of the form weight = a (length)^3 should describe the relationship. What happens if we cube the lengths in the data table and then graph weight versus length^3?

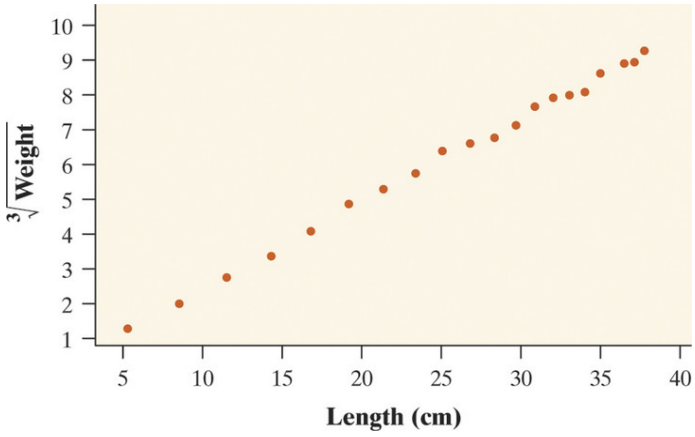
Go Fish!



There's another way to transform the data in the example to achieve linearity.

We can take the cube root of the weight values and graph weight versus length

Go Fish!



Go Fish!

x y

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

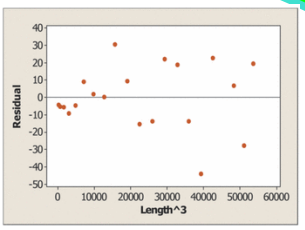
S = 18.8412 R-Sq = 99.5% R-Sq(adj) = 99.5%

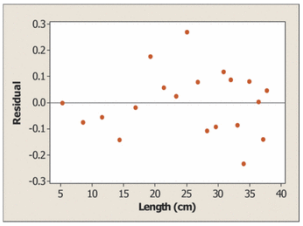
x y

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%





(1) $\hat{y} = 4.066 + .0146774x$

(a) Give the equation of the least-squares regression line. Define any variables you use. (2) $\hat{y} = -0.02204 + .246616x$

(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.

(2) The average prediction error is .124161 ∛grams.

Income vs. Child Mortality Rate

What does a country's income per person (measured in gross domestic product per person, adjusted for purchasing power) say about the under-5 child mortality rate (per 1000 live births) in that country? Here are the data for a random sample of 14 countries in 2009.

Country	Income Per Person	Under5 Mortality Rate
Switzerland	38004	4.4
Timor-Leste	2476	56.4
Uganda	1202	127.5
Ghana	1383	68.5
Peru	7859	21.3
Cambodia	1831	87.5
Suriname	8199	26.3
Armenia	4523	21.6
Sweden	32021	2.8
Niger	643	160.3
Serbia	10005	7.1
Kenya	1494	84
Fiji	4016	17.6
Grenada	8827	14.5

(a) Create a scatterplot and describe why it would not be appropriate to compute a least-squares regression line.

(b) What kind of power model might be appropriate to use for these data? $y = \frac{1}{x}$

(c) Use an appropriate power transformation to make the association linear. Sketch the resulting scatterplot, calculate the equation of the least-squares regression line, and sketch the residual plot.

$\hat{y} = 4.629 + 1102.56\left(\frac{1}{x}\right)$

L_1: x L_2: y L_3: $\frac{1}{x}$ L_4: Res.

new x

Income vs. Child Mortality Rate

What does a country's income per person (measured in gross domestic product per person, adjusted for purchasing power) say about the under-5 child mortality rate (per 1000 live births) in that country? Here are the data for a random sample of 14 countries in 2009.

Country	Income Per Person	Under5 Mortality Rate
Switzerland	38004	4.4
Timor-Leste	2476	56.4
Uganda	1202	127.5
Ghana	1383	68.5
Peru	7859	21.3
Cambodia	1831	87.5
Suriname	8199	26.3
Armenia	4523	21.6
Sweden	32021	2.8
Niger	643	160.3
Serbia	10005	7.1
Kenya	1494	84
Fiji	4016	17.6
Grenada	8827	14.5

(d) Predict the child mortality rate for the United States, who has an income per person of 41,256.

Logarithmic Model

What if you have no idea which power to choose?

1. Guess and test until you find a power that works

2. Use a logarithm (base 10 or base *e*)

	which one to log?	equation
power model	log y versus log x	log y = log a + p log x
exponential model	log y against x y against log x	log y = a + bx y = a + b log x

Income vs. Child Mortality Rate

What does a country's income per person (measured in gross domestic product per person, adjusted for purchasing power) say about the under-5 child mortality rate (per 1000 live births) in that country? Here are the data for a random sample of 14 countries in 2009.

Country	Income Per Person	Under5 Mortality Rate
Switzerland	38004	4.4
Timor-Leste	2476	56.4
Uganda	1202	127.5
Ghana	1383	68.5
Peru	7859	21.3
Cambodia	1831	87.5
Suriname	8199	26.3
Armenia	4523	21.6
Sweden	32021	2.8
Niger	643	160.3
Serbia	10005	7.1
Kenya	1494	84
Fiji	4016	17.6
Grenada	8827	14.5

(e)Use an appropriate logarithmic transformation to make the association linear. Sketch the resulting scatterplot, calculate the equation of the least-squares regression line, and sketch the residual plot. $\log y = 5.04 - .989 \log x$

(f) Predict the child mortality rate for the United States, who has an income per person of 41,256

$\log \hat{y} = 5.04 - .989 \log(41256)$

$\log \hat{y} = .475$

$\hat{y} = 10^{.475}$

$\hat{y} = 2.99$

Exponential Model

Takes the form $y = ab^x$

As x increases by 1 unit, the value of y is multiplied by b

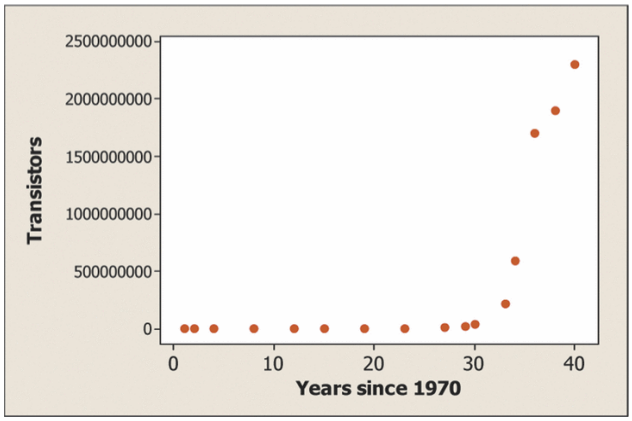
Exponential Growth: b>1

Exponential Decay: b<1

Moore's Law

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:

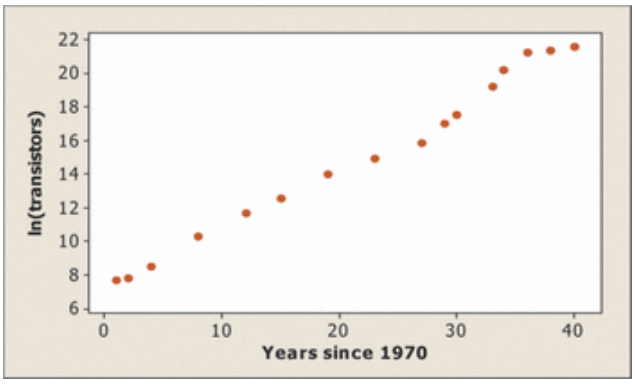
Processor	Date	Transistors
4004	1971	2,250
8008	1972	2,500
8080	1974	5,000
8086	1978	29,000
286	1982	120,000
386	1985	275,000
486 DX	1989	1,180,000
Pentium	1993	3,100,000
Pentium II	1997	7,500,000
Pentium III	1999	24,000,000
Pentium 4	2000	42,000,000
Itanium 2	2003	220,000,000
Itanium 2 w/9MB cache	2004	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



Moore's Law

(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.

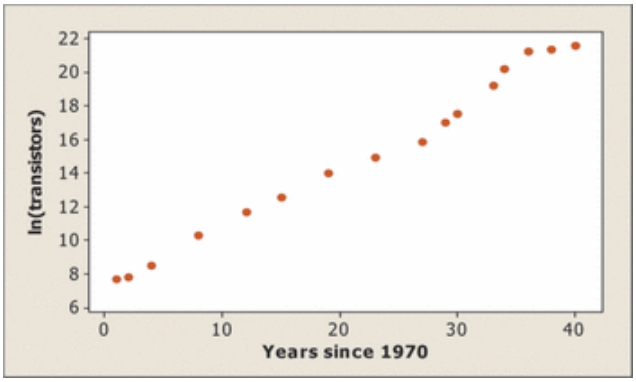
Processor	Date	Transistors
4004	1971	2,250
8008	1972	2,500
8080	1974	5,000
8086	1978	29,000
286	1982	120,000
386	1985	275,000
486 DX	1989	1,180,000
Pentium	1993	3,100,000
Pentium II	1997	7,500,000
Pentium III	1999	24,000,000
Pentium 4	2000	42,000,000
Itanium 2	2003	220,000,000
Itanium 2 w/9MB cache	2004	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



Moore's Law

(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.

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



Moore's Law

(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%

$\hat{y}$ = ln(#transistors) = 7.0647 + .36583 (yrs. since 1970)

x

Processor	Date	Transistors
	1971	2,250
	1972	2,500
	1974	5,000
	1978	29,000
	1982	120,000
	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

Moore's Law

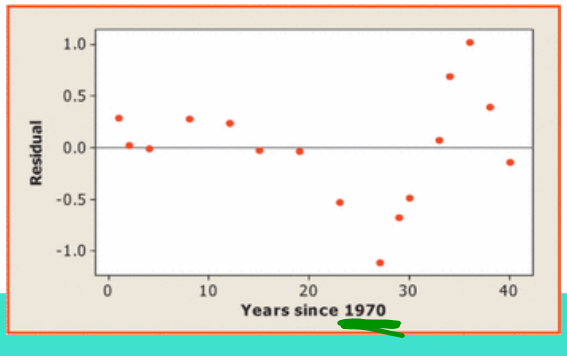
(c) Use your model from part (b) to predict the number of transistors on an Intel computer chip in 2020. Show your work.

1.029×10^{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.

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%



U of A (boo!)

In the April 1, 2011 edition of the *Arizona Daily Star*, the following data was presented about mandatory fees at the University of Arizona.

(a) Letting $x = 4$ represent 2004-05, graph a scatterplot. Does the relationship look linear?

Year	Fees
2004-05	89
2005-06	93
2006-07	160
2007-08	213
2008-09	257
2009-10	302
2010-11	623
2011-12	913

(b) Sketch a scatterplot of ln(fees) vs. year, calculate the equation of the least-squares regression line, and sketch a residual plot.

L_1: yr

~~L_2: fees~~

L_3: lnfees

L_4: resid

$\ln\hat{y} = 3.005 + .332x$

$\hat{y}$ = predicted fee

x = yr.

U of A (boo!)

In the April 1, 2011 edition of the *Arizona Daily Star*, the following data was presented about mandatory fees at the University of Arizona.

Year	Fees
2004-05	89
2005-06	93
2006-07	160
2007-08	213
2008-09	257
2009-10	302
2010-11	623
2011-12	913

(c) Could a power model be better? Sketch a scatterplot of ln(fees) vs. ln(year), calculate the equation of the least-squares regression line, and sketch a residual plot.

L_1: yr L_2: fee L_3: lnyr L_4: lnfee

(d) Based on your answers to (b) and (c), would an exponential model or a power model be more appropriate for this data? Explain.

ln of y (fee) only

(e) Use the model you chose in part (d) to predict the fees for the 2012–13 school year.

$\ln\hat{y} = 6.9902$

$\hat{y} = 1085.94$