4.1 Transforming Relationships

http://www.ruf.rice.edu/~lane/stat_sim/transformations/index.html

http://tools.google.com/gapminder/

Beware of rounding errors!

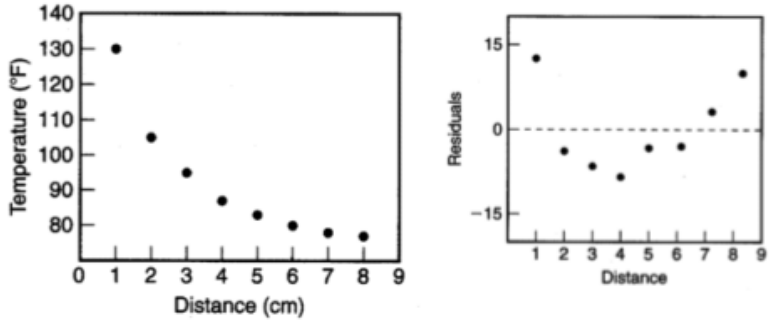
Transform data to make data more

- closely approximate a theoretical distribution
- evenly spread out/constant in variance
- symmetric
- linear

The table shows the temperature (*Temp*) of an instrument measured as its distance (*Dist*) from a heat source is varied. Although calculation would yield $r = -.894$, the scatterplot shows clearly that the data do not have a linear relationship.

Distance (cm)	Temperature (°F)
1	130
2	105
3	95
4	87
5	83
6	80
7	78
8	77

Scatterplot:



Try some very general suggestions:

clustered near origin and increasing ⇒ log x and log y

concave down and increasing ⇒ √x or log x

concave up with rapid growth ⇒ log y

concave up with rapid growth but clustered at right ⇒ log x and log y

concave up and decreasing ⇒ 1/x and 1/y

Any model may be subject to improvement.

Which model?

Linear:
- each term changes by adding a constant
- look for a (nearly) common difference between consecutive terms

Exponential:
- each term changes by multiplying by a constant
- look for a (nearly) common ratio of consecutive terms

Steps in transforming data

1. Plot the data.

2. If linear, find LSRL, r, r squared, and residual plot.

3. If not linear, use transformed data to find LSRL, r, r squared, and residual plot.

Transform the data based on whether it is...
...power: Using (log x, log y) may straighten it.
...exponential: Using (x, log y) may straighten it.
...logarithmic: Using (log x, y) may straighten it.
... related to some power: Using the "ladder of power transformations" may straighten it.

Starbucks Coffee: Number of stores in certain years

year	stores
1971	1
1987	17
1988	33
1989	55
1990	84
1991	116
1992	165
1993	272
1994	425
1995	676
1996	1015
1997	1412
1998	1886
1999	2135
2000	3501
2001	4709
2002	5886
2003	7225
2004	8337

Steps in transforming data

1. Plot the data.

2. If linear, find LSRL, r, r squared, and residual plot.

3. If not linear, use transformed data to find LSRL, r, r squared, and residual plot.

Transform the data based on whether it is...
...power: Using (log x, log y) may straighten it.
...exponential: Using (x, log y) may straighten it.
...logarithmic: Using (log x, y) may straighten it.
... related to some power: Using the "ladder of power transformations" may straighten it.

http://www.starbucks.com/aboutus/timeline.asp

Now try
exercises 7, 9,
& 11 on
page 214.

Mystery Data

t (x)	d (y)
0.2409	0.3871
0.6152	0.7323
1	1
1.881	1.524
11.86	5.203
29.46	9.555
84.01	19.22
164.8	30.11
247.7	39.81

Steps in transforming data

1. Plot the data.

2. If linear, find LSRL, r, r squared, and residual plot.

3. If not linear, use transformed data to find LSRL, r, r squared, and residual plot.

Transform the data based on whether it is...
...power: Using (log x, log y) may straighten it.
...exponential: Using (x, log y) may straighten it.
...logarithmic: Using (log x, y) may straighten it.
... related to some power: Using the "ladder of power transformations" may straighten it.

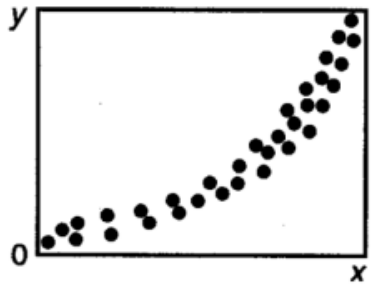
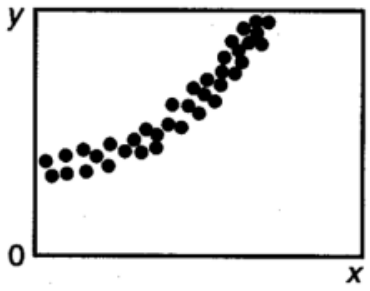
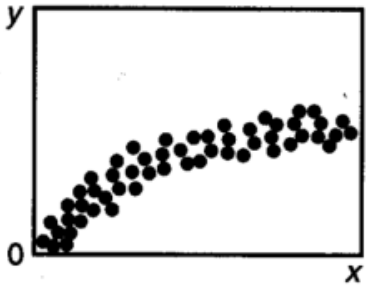
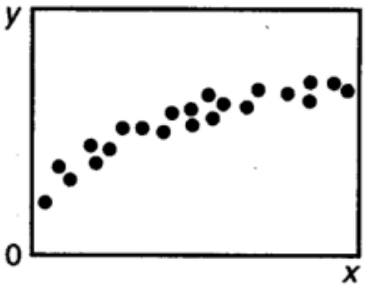
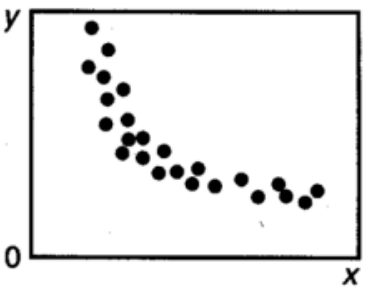
$$\log x^{p}$$
$$= p \log x$$

$$\widehat{\log y} = 1.1 + .2x$$

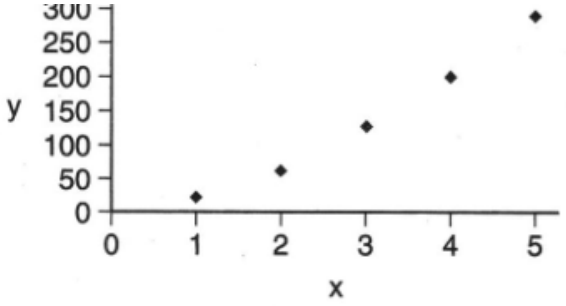
$x = .3 \quad y = ?$

p. 214 7, 9, 11

Now try
exercises 13 & 15
on page 219.

Intuitive curve of best fit	Example scatterplot	Suggested transformation
Contains (0, 0) and appears to be a power curve, or a curve asymptotic to both horizontal and vertical axes.		$(x_i, y_i) \rightarrow (\ln x_i, \ln y_i)$ $x_i > 0$ $y_i > 0$
Contains a nonzero y-intercept and appears exponential (either growth or decay).		$(x_i, y_i) \rightarrow (x_i, \ln y_i)$ $y_i > 0$
Contains (0, 0) and appears logarithmic.		$(x_i, y_i) \rightarrow (\sqrt{x_i}, y_i)$ $x_i \geq 0$
Contains a nonzero y-intercept and appears logarithmic.		$(x_i, y_i) \rightarrow (\ln x_i, y_i)$ $x_i > 0$
Has nonzero horizontal and vertical asymptotes.		$(x_i, y_i) \rightarrow \left(\frac{1}{x_i}, \frac{1}{y_i}\right)$ $x_i \neq 0$ $y_i \neq 0$

x	1	2	3	4	5
y	20	60	120	190	280



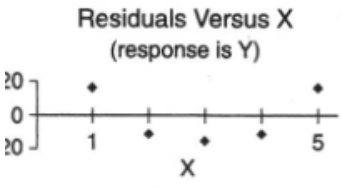
Answer: A linear fit to x and y gives $\hat{y} = 65x - 61$ with $r = .99$.
A linear fit to x and $\log y$ gives $\log y = 0.279x + 1.139$ with $r = .98$. This results in an *exponential* relationship:

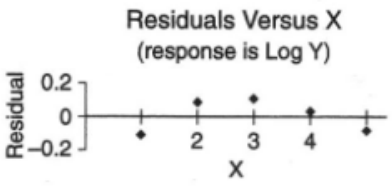
$$\hat{y} = 10^{0.279x+1.139} = 13.77(10^{0.279x}) = 13.77(1.901^{x})$$

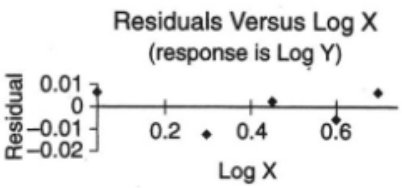
A linear fit to $\log x$ and $\log y$ gives $\log y = 1.639 \log x + 1.295$, also with $r = .99$. This results in a *power* relationship:

$$\hat{y} = 10^{1.639 \log x+1.295} = 19.72(x^{1.639})$$

All three models give high correlation and are reasonable fits. Further analysis can be done by examining the residual plots:







The first two residual plots have distinct curved patterns. Among the above three models, the power model, $\hat{y} = 19.72(x^{1.639})$, appears to be best.