4.1 Transforming Relationships

We will use Example 4.4 from P. 205 to discuss what happens when bivariate data is NOT linear and why residual plots are important...

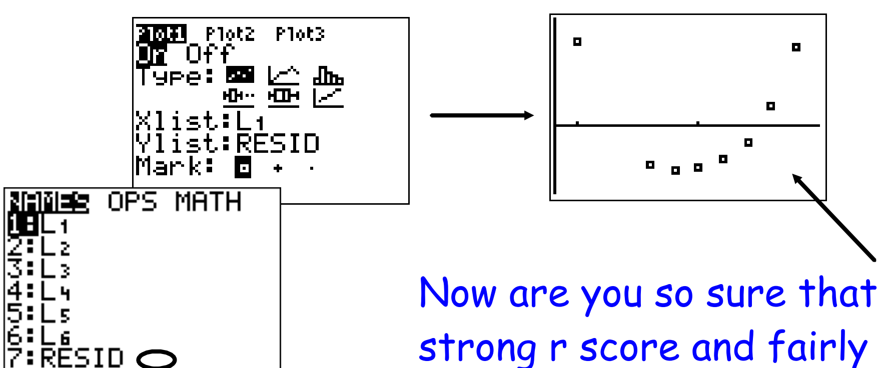
TABLE 4.1 The number of cell phone subscribers in the United States, 1990–1999

Year	1990	1993	1994	1995	1996	1997	1998	1999
Subscribers (thousands)	5283	16,009	24,134	33,786	44,043	55,312	69,209	86,047

Source: Statistical Abstract of the United States, 2000 and the Cellular Telecommunications Industry Association, Washington, D.C.

Enter the data in L1 and L2 and do two things: perform a linear regression (with the equation in Y1) and create a scatterplot with the LSRL. What do you think?

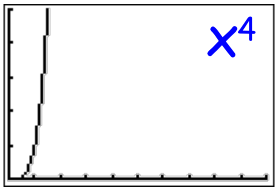
That could be pretty convincing as a linear relationship UNTIL you inspect the residual plot. This is why residual plots are an important part of assessing the appropriateness of a linear model for your data!

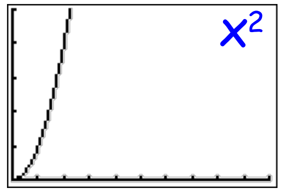


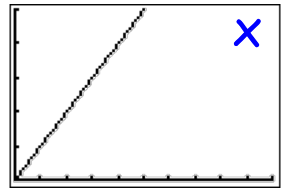
Whenever you perform a linear regression (STAT|CALC|8) on your calculator, the residuals are automatically calculated and stored in your lists. NEVER delete this list!

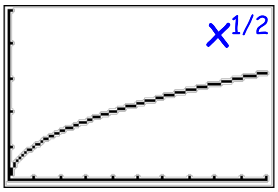
Now are you so sure that the strong r score and fairly linear looking model meant that you should use a straight line regression equation for this data?

In chapter 4, we will learn how to create mathematical models for data that have a CURVED form. In order to investigate this further, we need to refresh our memories about the different types of curved functions from Algebra...

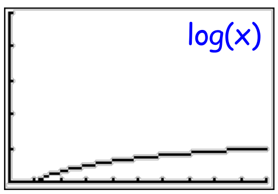








And the very special...

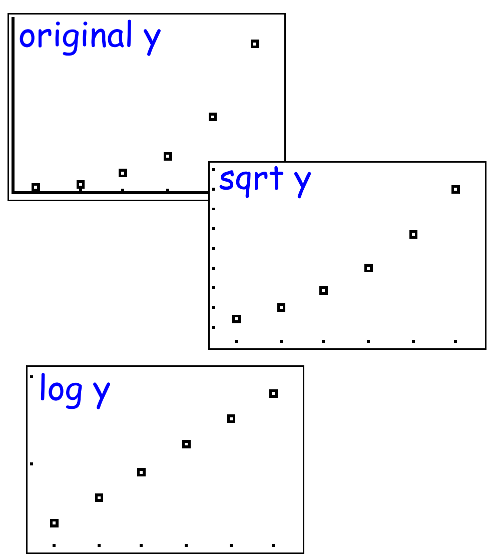


if we "do" one of these things to a set of values, the values will take on the shape of that curve. If data is curved upward and we chose a log function to transform it, we might just straighten out our x/y relationship

These "ladder of power" functions are all monotonic functions for values of x >0, meaning that their y values always increase with x or always decrease with x. We can use these to transform data in order to straighten it so that an LSRL can be created and used effectively.

We'll start with learning to identify when a set of data displays exponential growth

x	y	sqrt y	log y
1	2	1.4142	.30103
2	4	2	.60206
3	8	2.8284	.90309
4	16	4	1.2041
5	32	5.6569	1.5051
6	64	8	1.8062



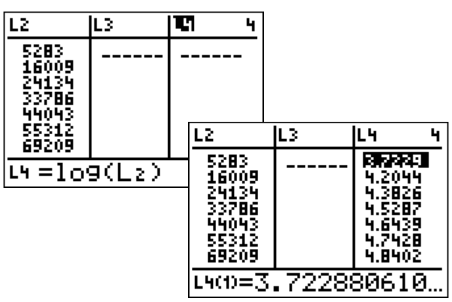
Exponential equations take the form: $y = ab^x$

where a and b are constants

x	y	ratios of y change	log (y)
1990	5283	-	3.72288
1993	16,009	-	4.20436
1994	24,134	1.51	4.38263
1995	33,786	1.40	4.52874
1996	44,043	1.30	4.64388
1997	55,312	1.26	4.74282
1998	69,209	1.25	4.84016
1999	86,047	1.24	4.93474

these ratios being approximately the same is how we know the data has exponential form and we need a $y = ab^x$ model.

PROBLEM: we don't know what a and b are so that we can write the equation. We now know that a logarithmic transformation will straighten the data, so we will take the log of the y values and then do a regression of log(y) on x.



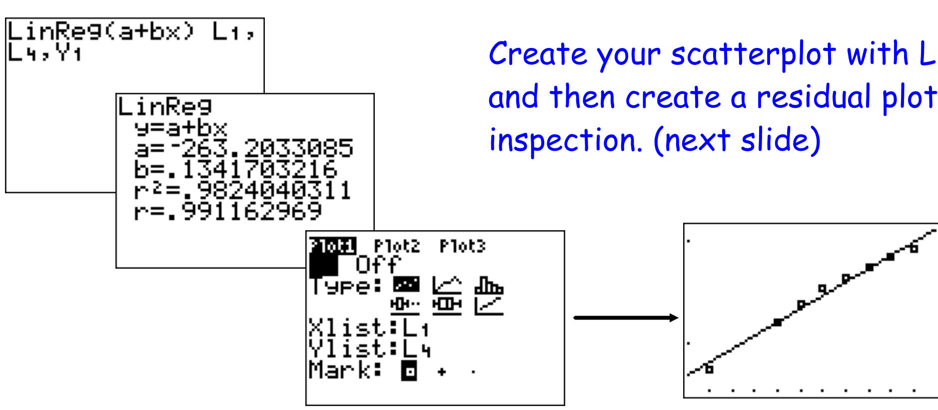
take the log of each y value by typing log(L2) in the header of L4.

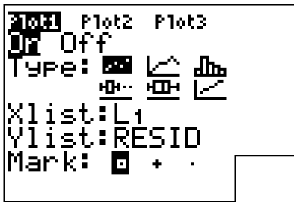
We will leave L3 blank for now.

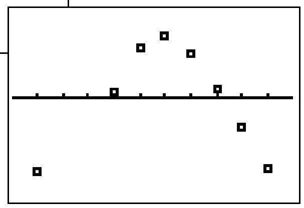
Once you have L4 filled, perform STAT|CALC|8 on L1,L4,Y1

You are regressing (x,logy)

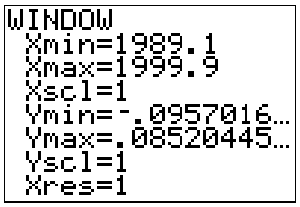
Create your scatterplot with LSRL and then create a residual plot for inspection. (next slide)







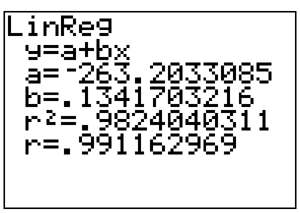
This residual plot looks problematic, but before we toss our procedure, we want to consider the data and whether it would be appropriate to edit the data. The book suggests omitting the first four observations and recalculating. For purposes of time, we will continue on as is.



look at the spread of Ymin to Ymax: the errors are very small, so the remaining pattern may not be terribly significant.

You really want your r to be very very strong at this point and your r^2 to be almost 1

Now for the predictions:



our regression equation will look like this:

y = -263.2033085 + .1341703216(X)

When we make a prediction for the year 2000, will our y value be the number of cell phones?

Y1(2000) = 5.137334711

to fix this we will "undo" the log by raising 10 to the power of our answer.

$10^{5.137334711}$ = 137,193.8712

Why do we do this? ....

Logarithmic transformation and antilog by hand:

Logarithm rules:
$\log(AB) = \log A + \log B$
$\log(A/B) = \log A - \log B$
$\log X^p = (p)\log X$

$\hat{y} = ck^x$ exponential model

$\log\hat{y} = \log c + \log k \cdot x$ logarithmic transformation gives LSRL of (x,logy)

① $\log\hat{y} = a + b \cdot x$ calculator gives a and b with regression of x,logy substitute here as logc = a and logk = b

$10^{\log\hat{y}} = 10^{a+bx}$ with constants in place in LSRL, back-transform to get exponential model

② $\hat{y} = (10^a)(10^b)^x$ this is the exponential model - you can go straight to this step with a and b from your STAT|CALC|8|L1,L4,Y1

$\hat{y} = ck^x$ if logc = a, then $10^a = c$, and if logk = b, then $10^b = k$

Power Models:

x	y
1	1
2	4
3	9
4	16
5	25

1. check the ratios between y values
2. if ratios are steadily increasing or decreasing, then the data is a power curve $\hat{y} = cx^k$
3. with x data in L1 and y data in L2 use L3 for LogL1 (Logx) and L4 for LogL2 (Logy)
4. STAT|CALC|8|L3,L4,Y1 to generate LSRL for transformed data and discover constants a and b to use in power model
5. $\hat{y} = (10^a)x^b$ OR $\hat{y} = cx^k$ [where $10^a = c$ and b = k]

see next screen for transformation

Logarithmic transformation for power models

$$\hat{y} = cx^k$$ power model

$$\log\hat{y} = \log c + \log x \cdot k$$ logarithmic transformation gives LSRL of (logx,logy) where logc is the intercept and k is the slope

① $$\log\hat{y} = a + \log x \cdot b$$ calculator gives a and b with regression of (logx,logy) substitute here as logc = a and k = b

$$10^{\log\hat{y}} = 10^{a+\log x \cdot b}$$

$$\hat{y} = (10^a)(10^{\log x})^b$$

with constants in place in LSRL, back-transform to get exponential model

② $$\hat{y} = (10^a)x^b$$ this is the power model - you can go straight to this step with a and b from your STAT|CALC|8|L1,L4,Y1

$$\hat{y} = cx^k$$ if logc = a, then $10^a = c$, $10^{\log x} = x$, and b = k