Chapter 4 - More on Two Variable Data
4.1 Transforming Relationships - Day 1
OBJ: You will determine the relationship between two variables when they are not linear.

Two Types of Nonlinear Growth:

Exponential Func ($y = ab^x$)   Power Func ($y = ax^b$)

Steps:
1) These Functions can be transformed into Linear Functions.
2) Use Lin. Reg. methods to find LSRL on transformed data.
3) Perform an inverse transformation to obtain a curve that models the original data.

Laws of Logarithms
$\log_b X = y$ iff $b^y = X$ Solve for y: $\log y = x$
① $\log(AB) = \log A + \log B$
② $\log(A/B) = \log A - \log B$
③ $\log X^P = P \cdot \log X$

Linear Growth - Adds a fixed amount in each = time period.
Exponential Growth - Multiplies a fixed # > 1 (Growth) or a fixed # < 1 (Decay) in each = time period.

Do Example 4.4 ⟹

YR	Subscribers	Ratios	log y
1990	5283	No Predecessor	3.7229
1993	16 009	X inc by 3 not by 1 !	4.2044
1994	24 134	$\frac{24134}{16009} = 1.51$	4.3826
1995	33 135	1.40	4.5207
1996	44 043	1.30	4.6434
1997	55 312	1.26	4.7428
1998	69 209	1.25	4.8402
1999	86 047	1.24	4.9347

1) Enter Year in L_1 and Subscribers in L_2
2) Plot $L_1 + L_2$
a) Lin Reg L_1, L_2, Y_1
b) Resid Plot (Contain Lg Residuals)

Looks Exponential ∴ Log Both sides

$y = 10^x$
$\log y = \log 10^x$
$\underline{\log y = x}$

3) Log of L_2 (call Log y)
4) Plot $L_1 + \log y$ – Notice Linearity
a) Lin Reg L_1, Log y, y_2 – still has form, but high r^2 and sm. residual.
b) Resid Plot – notice r and r^2 values

This is why we are doing All this For !!

5) Find the curve that fits the orig. data –
$\log \hat{y} = -263.2033 + .1342x$
Anti log both sides
$\cancel{10^{\log \hat{y}}} = 10^{-263.2033 + .1342x}$
$\hat{y} = 10^{-263.2033} \cdot 10^{.1342x}$

Calc can't handle, so graph $\hat{y} = 10^{-263.2033 + .1342x}$ or 10^{y_1}

AP: Form: $\hat{y} = a \cdot b^x$
$a = 10^{-263.2033}$
$b = 10^{.1342}$

or

Form $\hat{y} = C \cdot 10^{Kx}$
$C = 10^{-263.2033}$
$K = .1342$

6) Deselect y_2
7) $y_3 = 10 \wedge y_2$
8) Plot $L_1 + L_2$ Again Along w/ y_3 – Not A Perfect Fit but close!

To see a perfect fit, try:

Day:	Amt:
1	.01
2	.02
3	.04
4	.08
5	.16
6	.32
7	.64
8	1.28
9	2.56
10	5.12

HW: 4.6 – Day 1
4.11 – Day 2

4.1 Transforming Relationships – Day 2

OBJ: You will transform power functions ($y = ax^b$)

Power Regression

$y = ax^b$

$\log y = \log(ax^b)$

$\log y = \log a + \log x^b$

$\log y = \log a + b \log x$ } If data is not linear or exponential... Try power – Lin Reg on log x, log y

constant

Let a = log a

* How to Find the inv. transformation (Find curve that Best Fits orig. Data)

$\log y = a + b \log x$

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

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

$y = 10^a \cdot x^b$ How?

$10^{\log x^b} = Ans$

$\log \boxed{x^b} = \log 10 \boxed{Ans}$

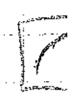
*PWR Func. Passes through (0,0)

Do Example 4.9

Length in L1, weight in L2

NOT Linear – Exponential?

Stat Plot L1, Log y 

Lin Reg L1, Log y, Y1: $\hat{y} = .4841 + .0692x$ r = .96 r^2 .93 

Resid Plot → NOT Exponential

Power?

Stat Plot Log x, Log y

Lin Reg Log x, log y, Y2: $\hat{y} = -1.8994 + .30494x$ { r = .999 r

Resid Plot  No Form Random Scatter Plot

→ ∴ Power Function!

Find inv. trans. (orig. curve)

$y = 10^a \cdot x^b$ so.....

$y = 10^{-1.8994} \cdot x^{.30494}$

HW: 4.13, 4.14

4.2 Cautions about Correlation + Regression

<u>OBJ</u>: You will explore situations that may inflate or deflate correlation.

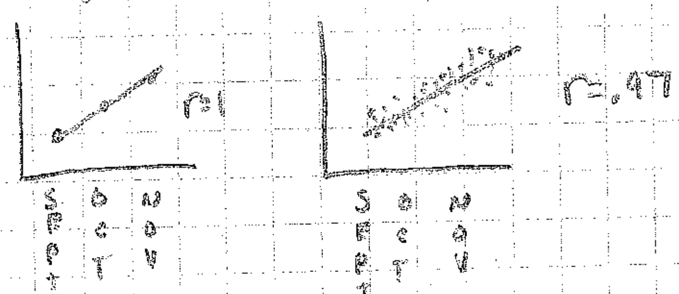
<u>Extrapolation</u> - The use of the LSRL for predictions far outside the domain of values of the explanatory variable x that you used to obtain the line or curve. Such predictions are often not accurate.

<u>Lurking Variable</u> - A variable that is not among the exp. or response vars. in a study + yet may influence the interpretation of relationships among those variables.

* A lurking variable can falsley suggest a strong relationship between x and y, or it can hide a rel that is really there.

Example 4.10

Averaged Data - Top of Pg 230



4.28-4.30

<u>Causation</u> - The goal of a study is to establish that changes in the exp. variable <u>causes</u> changes in the response var. Even when a strong assosciation is present, the conclusion is often elusive. (Smoking, Rhino stampede, etc)

<u>Confounding</u> - Two vars. are confounded when their effects on a response var. can't be distinguished from each other. The confounding variables may be exp. or lurking var.

<u>Common Response</u> - When a lurking variable is common to both Exp + Resp. vars.

<u>For Ex</u>: There is a pos. assoc. between ice cream sales and sunburns. This is caused by common response

In Class 4.33, 4.35, 4.36

4.3 Relations in Categorical Data – Day 1

<u>OBJ</u>: You will describe relationships between two or more categorical variables.

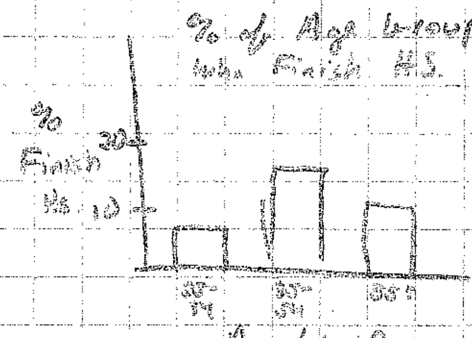
See table 4.6 on pg. 281

- A two way table describes two categorical variables.
- Arrange the variables from least to most.
- If the Row + column totals are missing, calculate them.
- <u>Marginal Distributions</u>: The distributions @ the right + bottom margins of a two-way table.
- <u>Roundoff Error</u> – See 35-54 yr olds (81, 435)
  - Add the column (81, 435)
  - Each level of education is rounded indiv.
  - Census Bureau rounded total #.

Calculate the marginal Distributions (In %) of people <u>25 or older</u> who have finished H.S.

Age Group: 25-34   35-54   55+

Percent: $\frac{11\,546}{175\,230} = 6.6\%$ 15.1%   11.4%



To Describe Relationships among cat. vars., calculate appropriate percents from the counts given.

Why? 11,066 Age 25-34 completed ≥ 4 yrs. college
10,596 Age 55+ " " " "

<u>But</u> 55+ has a larger group ∴ can't compare counts.

4.51, 4.52

<u>Ex</u>: Find the complete dist. of 55+ Age Group.

Education:	DNF	HS	1-3 Coll	≥ 4 Coll
Percent	$\frac{14\,334}{56\,088}$	$\frac{20\,030}{56\,088}$	$\frac{11\,129}{56\,088}$	$\frac{10\,596}{56\,088}$
	25.5%	35.5%	19.9%	18.8%

100% b/c all 55+ fall into 1 cat.

Conditional Dist. – Refers only to those in 55+ cond.

<u>HW</u>: 4.53, 4.55

4.3 Relations in Categorical Data – Day 2

OBJ: You will compare data that involves Simpson's Paradox.

Simpson's Paradox – The reversal of the direction of a comparison or an association when data from several groups are combined to form a single group.

On Time Arrivals

	Alaskan Airlines		America West	
City	% on Time	# Arrivals (on Time)	% on Time	# Arrivals (on-Time)
LA	88.9%	559 (497)	85.6%	811 (694)
PHX	94.8%	233 (221)	92.1%	5255 (4840)
SD	91.4%	232 (212)	85.5%	448 (383)
SF	83.1%	605 (503)	71.3%	449 (320)
SEA	85.3%	2146 (1831)	76.7%	262 (201)
Totals	86.7%	3775 (3273)	89.1%	7225 (6437)

Each indiv. city has higher on time % for Alaskan Airlines but this reverses when a single group of all cities is formed and now favors America West!

Why? – "Anchor" Cities

(Sea for Alaskan
PHX for America West)