Math 2311

Bekki George – bekki@math.uh.edu

Office Hours: MW 11am to 12:45pm in 639 PGH

Online Thursdays 4-5:30pm

And by appointment

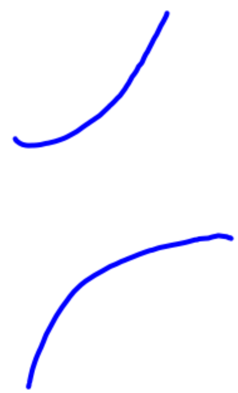
Class webpage: http://www.math.uh.edu/~bekki/Math2311.html

Popper 12

1. If data follows a trend that is not linear, we cannot make a prediction about it.

a. True

b. False



5.5 – Non-Linear Methods

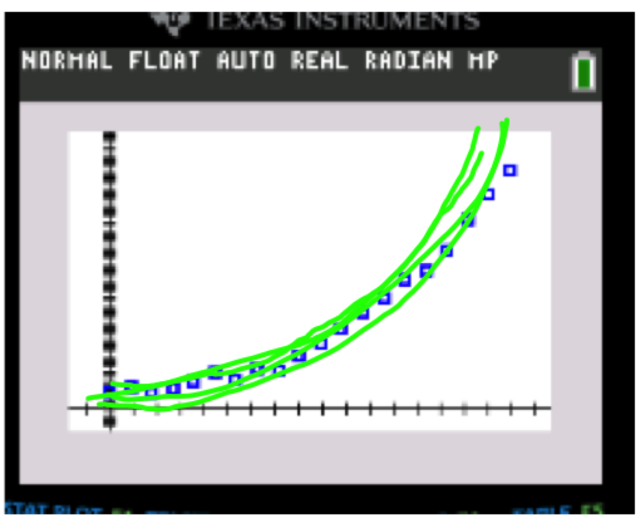
Many times a scatter-plot reveals a curved pattern instead of a linear pattern.

We can **transform** the data by changing the scale of the measurement that was used when the data was collected. In order to find a good model we may need to transform our *x* value or our *y* value or both.

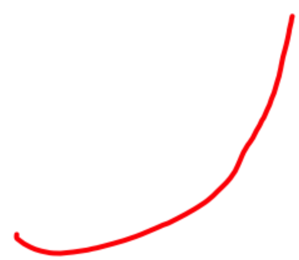
In this example from section 5.4, we saw that the linear model was not a good fit for this data:

		0	10	20	30						90
L1	Year x	1790	1800	1810	1820	1830	1840	1850	1860	1870	1880
L2	People per square mile	4.5	6.1	4.3	5.5	7.4	9.8	7.9	10.6	10.09	14.2
L1	Year x 100	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980
L2	People per square mile	17.8	21.5	26	29.9	34.7	37.2	42.6	50.6	57.5	64

Let's investigate other models.



$y = x^2$



$y = e^x$

$y = e^x$ is the inverse of $y = \ln x$

$\ln e^x = x$

$y = x^2$

resid ↓

transformed y for $y = x^2$

L_1	L_2	L_3	L_4
x	y		$\sqrt{y}$

$L_4 = \sqrt{(L_2)}$

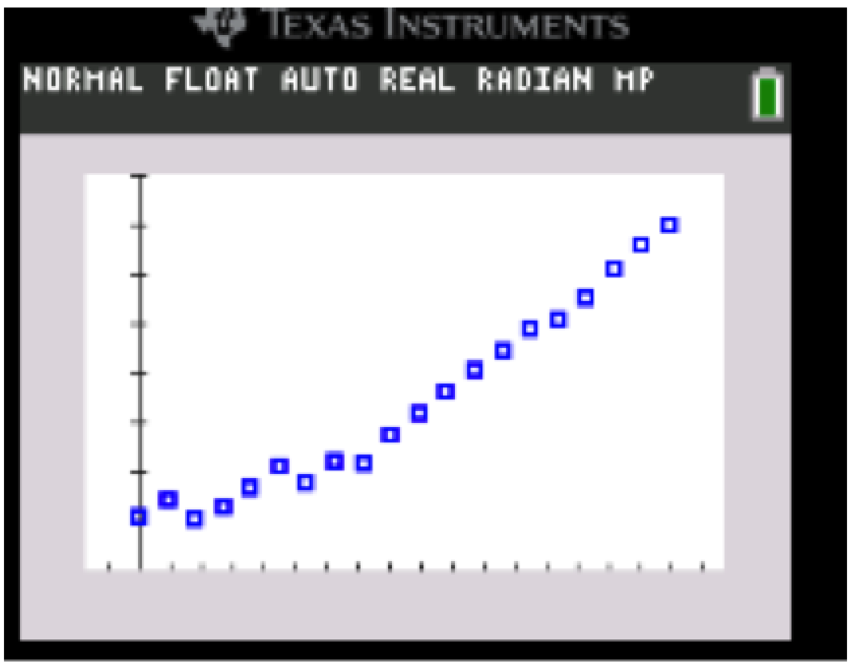
$\boxed{\text{LinReg}(a+bx)\ L_1, L_4, Y_1}$

L_1, L_4 $r = .977$

('old' r was $.9431$)

L_1, L_2

$\sqrt{y} = x \Rightarrow y = x^2$



Stat plot

X list : L_1

Y list : L_4

Graph

Zoom 9

L_1	L_2	L_3	L_4
x	y	$\downarrow$	$\sqrt{y}$

resid
for $\sqrt{y} \sim x$ $(y = x^2)$

$$\boxed{L_3 = L_4 - Y_1(L_1)}$$

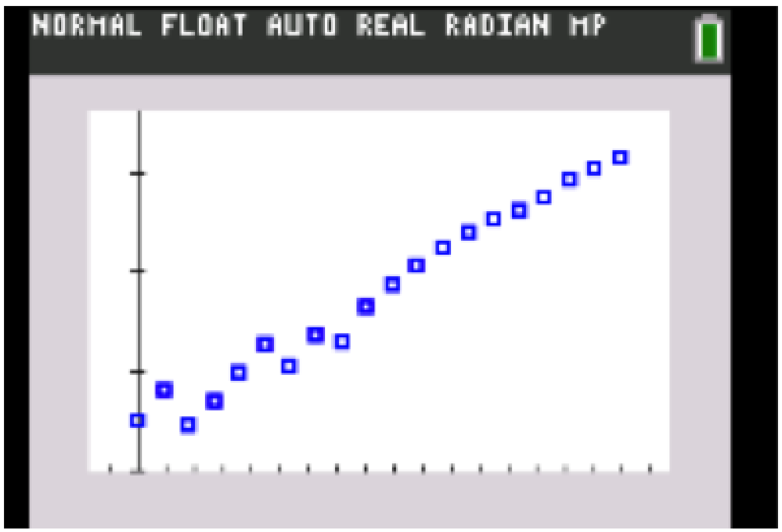
Y_1 - vars - YVARS - ENTER

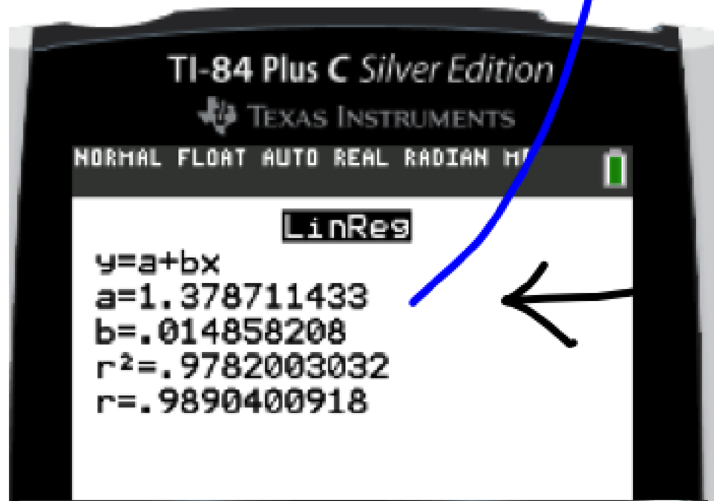
$y = e^x$ ★

L_1	L_2	L_3	L_4	L_5
x	y	Res	$\sqrt{y}$	$\ln(y)$

$L_5 = LN(L_2)$

$\ln \hat{y} = 1.3787 + .01486x$

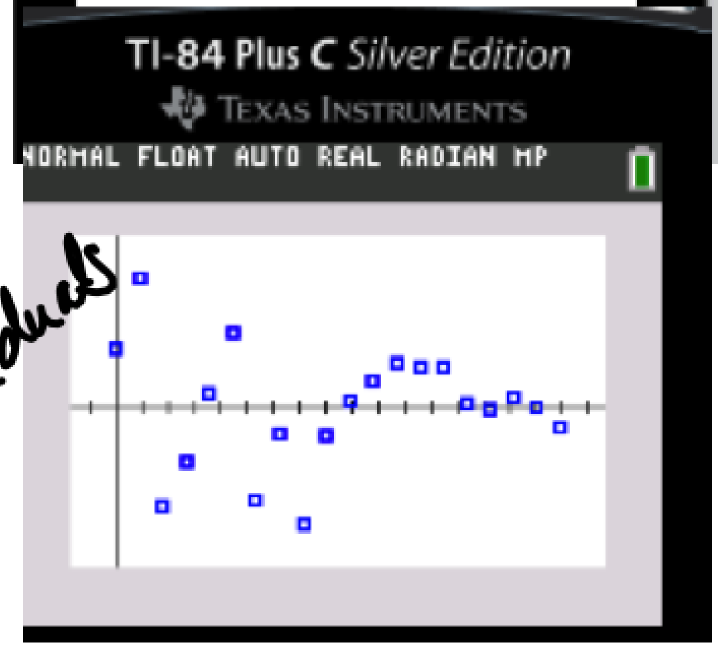




Stat calc
8
L_1, L_5, Y_1

$L_3 = L_5 - Y_1(L_1)$

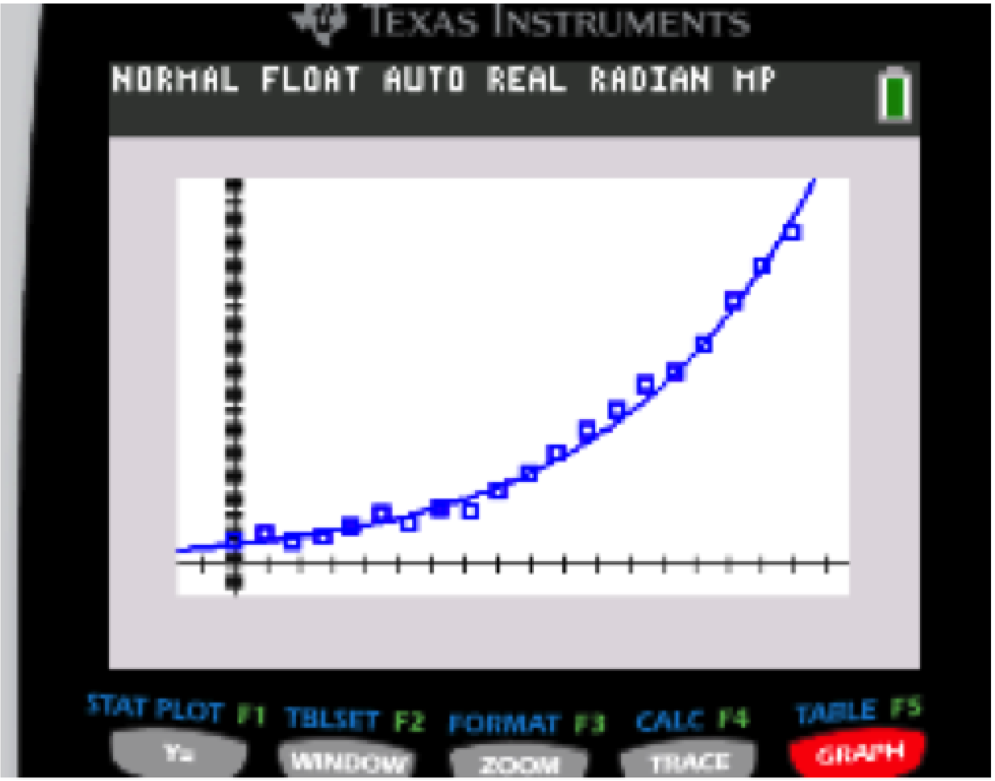
Residuals



$\ln y = 1.3787 + .01486x$

$$\boxed{\hat{y} = e^{(1.3787 + .01486x)}}$$

model



5.6 – Relations in Categorical Data

A **two-way table** organizes the data for two categorical variables.

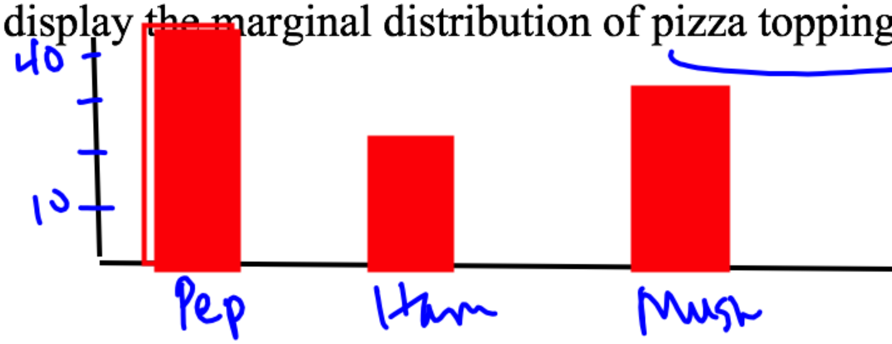
The totals of each row and column are considered **marginal distributions** because they appear in the margins of the table.

Example:
The following two-way table describes the preferences in movies and pizza toppings for a random sample of 100 people.

Movie	Pepperoni	Hamburger	Mushrooms	
Jurassic Park	20	5	10	35
Star Wars	15	15	12	42
Gone with the Wind	8	2	13	23
	43	22	35	100

Enter the marginal distributions in the table.

Draw a bar chart to display the marginal distribution of pizza topping preference.



What percent of our sample likes Gone with the Wind? $23/100 = 23\%$

What percent of pepperoni lovers like Star Wars? $15/43 = 34.88$

A **conditional distribution** is made up of the percentages that satisfy a given condition.

Compare the conditional distributions of movie preference for hamburger lovers and mushroom lovers. Back up your description with percentages.

	Hamb	Mush
JP	$\frac{5}{22} = 22.7\%$	$10/35 = 28.6\%$
SW	$\frac{15}{22} = 68.2\%$	$12/35 = 34.3\%$
GW	$\frac{2}{22} = 9\%$	$13/35 = 37.1\%$
	(99.9%)	

Popper 12

The following two-way table describes the preferences in music genre and pizza toppings for a random sample of 100 people.

	Cheese	Peperoni	Mushrooms
Techno	20	5	15
Country	8	12	11
Rock	14	2	13

2. What percent of the sample likes rock music?
   a. 35%
   b. 43%
   c. 30%
   d. 29%
   e. none of these

3. What percent of cheese lovers like techno music?
   a. 18.6%
   b. 68.2%
   c. 22.7%
   d. 47.6%
   e. none of these

Always be careful if combining data to make a comparison. **Simpson's Paradox** is the reversal of the direction of a comparison or an association when data from several groups are combined to form a single group.

This is adapted from Subsection 2.3.2 of A. Agresti (2002), *Categorical Data Analysis*, 2nd ed., Wiley, pp. 48-51.
In a 1991 study by Radelet and Pierce of the effect of race on death-penalty sentences, the following table was obtained tabulating the death-penalty sentences (Death) and non-death-penalty sentences (No death) in murder convictions in the state of Florida.

Defendant's race	Death	No death	Percent death
Caucasian	53	430	11.0
African-American	15	176	7.9

From this table, we see Caucasian defendants received the death penalty more often than African-American defendants.

Now, we consider *the very same data*, except that we stratify according to the **race of the victim** of the murder. Below is the table.

Victim's race	Defendant's race	Death	No death	Percent death
Caucasian	Caucasian	53	414	11.3
Caucasian	African-American	11	37	22.9
African-American	Caucasian	0	16	0.0
African-American	African-American	4	139	2.8

Here we see that when considering the cases involving Caucasian victims separately from the cases involving African-American victims, that the African-American *defendants* are more likely than Caucasian ones to receive the death penalty in both instances (22.9% vs 11.3% in the first case and 2.8% vs. 0.0% in the second case).

Example 3 (from text): A drug company tests two new treatments for an illness. In trial 1, drug A cures 45 out of 200 of the patients with the illness and drug B cures 32 out of 200 patients with the illness. In trial 2, 100 patients with the illness are given drug A and 85 of them are cured. Drug B is given to 500 patients in trial 2 and 400 are cured.

A. Create a table for each trial and compare results. Which treatment would you conclude is better based on the data in the tables?

Trial 1	Cured	Total	Percent
Drug A	45	200	22.5%
Drug B	32	200	16%

Trial 2	Cured	Total	Percent
Drug A	85	100	85%
Drug B	400	500	80%

A: $\frac{45}{200}$ $\frac{85}{100}$

$\neq$

B. Put the data together into one table and calculate the percentage cured for the aggregated data. Which treatment would you conclude is better based on the data in this table?

Trials 1 and 2 combined	Cured	Total	Percent
Drug A	130	300	43.3%
Drug B	432	700	61.7%

$A = \frac{45+85}{200+100}$

Popper 12

4. Simpson's Paradox occurs when
    a. There is a reversal in direction of a comparison when data is transformed with powers.
    b. There is a reversal in direction of a comparison when data from several groups is combined.
    c. A conditional distribution gives a contradictory answer.
    d. The LSRL is used to predict data that is far from the other explanatory values.

5. LSRL stands for:
    a. Linear squares right line
    b. Least squares regression line
    c. Line standard regression lower