Bivariate Data Summary

Bivariate data – data that examines the relationship between two variables

- What individuals to the data describe?
- What are the variables and how are they measured
- Are the variables quantitative or categorical

Types of bivariate data

- Response variable – measures the outcome of a study
- Explanatory variables – attempts to explain (not cause) the response variable

to determine which is explanatory and which is response, think about which one seems to be a possible explanation of the other.

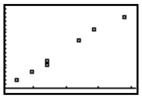
if it is not obvious which one is explanatory and which is response, it very well be that it doesn't matter.

Scatterplots – shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appears on the horizontal (x) axis and the other axis appears on the vertical (y) axis. Each individual appears as a point in the plot. The explanatory variable is placed on the x-axis and the response variable is placed on the y-axis. If there is no explanatory-response relationship, either variable can go on the horizontal axis.

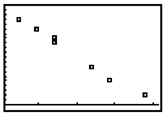
Scatterplots *must* be labeled (both axes). The intervals on each axis must be uniform.

Interpreting scatterplots

- Look for the overall pattern and deviations from that pattern.
  - Form – does the data appear linear or curved or have distinct clusters?
  - Direction – the term used is association
  - Positive association – low values of the explanatory variable accompanies low values of the response variable and high values of the explanatory variable accompanies high values of the response variable



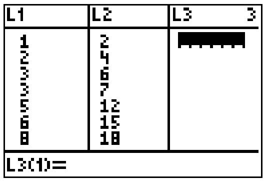
  - Negative association – low values of the explanatory variable accompanies high values of the response variable and high values of the explanatory variable accompanies low values of the response variable



  - Strength of an association – how closely the points follow a clear form. Both of the associations above are strong.
  - Outliers – a point that falls outside the overall pattern of the relationship. As usual, outliers should be eliminated if it can be justified.

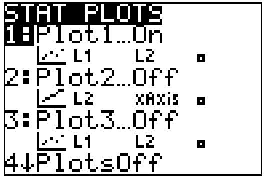
Using the calculator:

Place the data in your lists: STAT EDIT

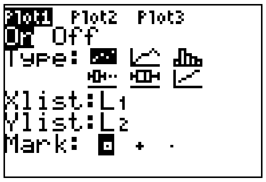


L1	L2	L3 3
1	2	
2	4	
3	6	
3	7	
5	12	
6	15	
8	18	

L3(1)=



STAT PLOTS
1:Plot1…On
L1 L2 □
2:Plot2…Off
L2 xAxis □
3:Plot3…Off
L1 L2 □
4↓PlotsOff



On Off
Type:
Xlist:L1
Ylist:L2
Mark: □ + ·

Set up your plots: 2nd Y =

Then: ZOOM 9 : ZoomStat Be sure that you have no graphs in your Y = list.

Correlation –measures the direction and strength of the linear relationship between two quantitative variables. The variable to denote correlation is *r*.

Facts about correlation:

1. *r* is a number between -1 and 1 inclusive. Positive values of *r* indicates a positive association between variables. Negative values of *r* indicates a negative association between variables.

2. Values of *r* near 1 or -1 mean a very strong association (the points are very close to forming a straight line). Values of *r* near 0 mean a very weak association. If *r* is exactly 1 or -1, the association is perfect (rarely happens).

3. Correlation makes no difference between explanatory and response variables. It makes no difference which variable is *x* and which is *y* when calculating *r*.

4. When calculating *r*, both variables must be quantitative.

5. *r* does not change when we change the units of measurements of *x*, *y*, or both.

6. Correlation measures the strength of a linear relationship between variable. It does not relationships that are curved no matter how strong they appear to be.

7. *r* is non-resistant … it is affected strongly by outliers.

8. In general, we will make this claim:

$$\begin{cases} |r| \geq .9\text{...Very strong association} \\ .7 \leq |r| < .9\text{...Fairly strong association} \\ .5 \leq |r| < .7\text{...Moderately strong association} \\ .2 \leq |r| < .5\text{...Fairly weak association} \\ |r| < .2\text{...Very weak association} \end{cases}$$

The formula for correlation: $r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$. You are not responsible for this formula.

We will find r using the calculator (below).

Least Squares Regression – a method for finding a line that summarizes the relationship between two variables.

Regression line – a straight line that describes how a response variable y that changes as an explanatory variable x changes. Regression lines re use to predict the value of y for a given value of x. Regression requires an explanatory variable and a response variable.

Least squares regression line (LSRL) – the line that makes the sum of the squares of the vertical distances of the of the vertical distances of the data points from the line as small as possible. (Fathom Demo)

Important: Every LSRL goes through the point $(\bar{x}, \bar{y})$

Equation of the LSRL – The regression line is in the form $\hat{y} = a + bx$ (similar to what you used in algebra. The $\hat{y}$ (read y-hat) is used to denote the *predicted* value of y based on the value of x.

The values of a and b are based on the means $\bar{x}$ and $\bar{y}$ and the standard deviations s_x and s_y as well as the correlation r.

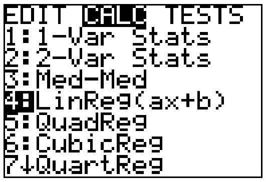
slope $b = r\frac{s_y}{s_x}$

y-intercept $a = \bar{y} - b\bar{x}$ You will not have to find the LSRL using these formulas.

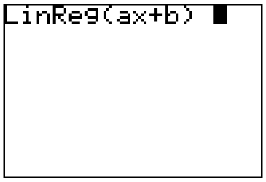
Using the Calculator:

Have your data in your lists as shown above and create a scatterplot.

STAT | CALC | 4 : LinReg(ax + b)



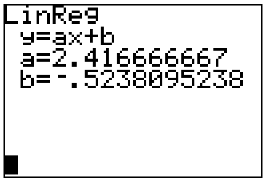
ENTER



```
LinReg(ax+b)
```

If you do not specify the lists, the calculator will automatically do regression on L1 and L2.
If you want regression on lists other than L1 and L2 you must specify (ex: LinReg(ax+b) L2,L4)

This is your result:

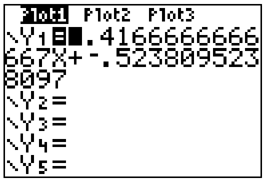


```
LinReg
 y=ax+b
 a=2.416666667
 b=-.5238095238
```

If you want to actually graph the line, do these steps:

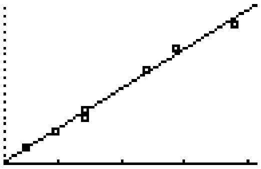
STAT | CALC | 4 : LinReg(ax + b) | VARS | Y - VARS | Function | Y1 | Enter

Press Y = and your equation will be in Y1: .

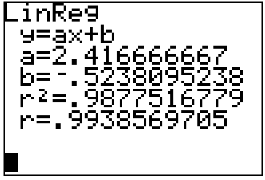


```
Plot1 Plot2 Plot3
\Y1=█.4166666666
667X+-.523809523
8097
\Y2=
\Y3=
\Y4=
\Y5=
```

Now GRAPH



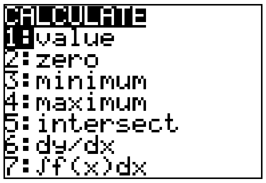
If you want to find the value of *r* (correlation), do this:
2nd | CATALOG | D | DiagnosticOn | ENTER . When you do LSRL on the calculator, the value of *r* will be present: . Note that r^2 (below is also given. Once you turn diagnostic on, it will stay on (so just leave it on – it doesn't hurt anything).



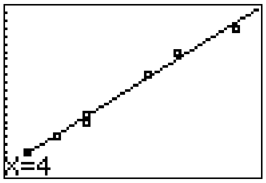
```
LinReg
 y=ax+b
 a=2.416666667
 b=-.5238095238
 r²=.9877516779
 r=.9938569705
█
```

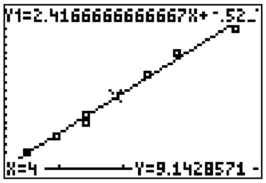
Using the LSRL: Suppose we want to predict the value of *y* when *x* = 4. You could plug *x* = 4 into the equation above, but it is easier to let the calculator do the work.

Graphic way: 2nd | CALC | Value | ENTER | 4 | ENTER

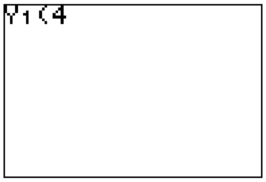


```
CALCULATE
1:value
2:zero
3:minimum
4:maximum
5:intersect
6:dy/dx
7:∫f(x)dx
```

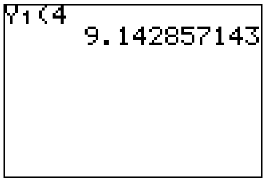




Function method: Vars | Y - Vars | Function | Y1 | ENTER | (4 | Enter



```
Y1(4
```



```
Y1(4
        9.142857143
```

This says that the predicted value of *y* when *x* = 4 is 9.14.

Coefficient of determination - r^2. This represents the percentage of the variation in y that can be explained by the LSRL. Memorize these words.

If for instance, you are measuring the relationship to the amount of studying one does compared to the grades in a test, and you get $r = .9$ (strong positive association), then $r^2 = .81$. You say that 81% of the variation in grades in the test can be explained by the LSRL of study time on grades. That means that 19% of the variation in grades can be explained by other factors other that study time.

Residuals – The different between an observed value of the response variable and the value predicted By the LSRL. The formula for a residual is:

$$\boxed{\text{Residual} = \text{observed } y - \text{ predicted } y = y - \hat{y}}$$

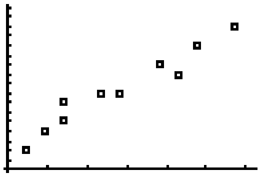
For any LSRL, the sum of the residuals is always zero.

Residual Plots – a scatterplot of the regression residuals against the explanatory variable x. Residual plots help us to assess how good a line is in describing the data.

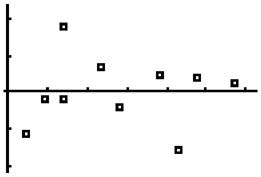
You look at the scatterplot of the residuals and determine information about the original data.

Important:
If a residual scatterplot shows no pattern, then the original data has a good linear fit.

Original data

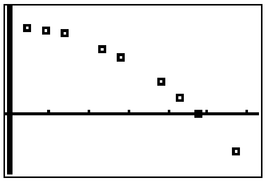


Residual plot

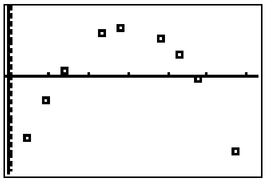


If a residual scatterplot shows a curved pattern, then the original data does *not* have a good linear fit.

Original data

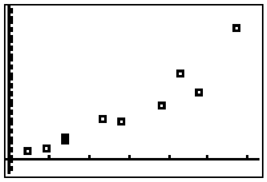


Residual plot

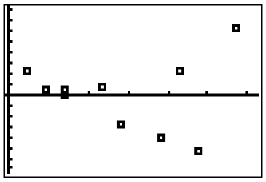


If the residual scatterplot shows no pattern but has an increasing or decreasing spread about the line, a linear fit can be used but the LSRL will be less accurate for larger values of y.

Original data



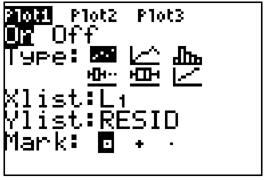
Residual plot



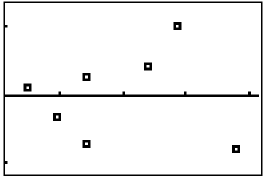
Using the Calculator:

For the data above, the way you create a residual plot is to follow these steps:

1) Put the data in your lists
2) Generate your scatterplot
3) Perform Least-squares regression on your data and graph the line
4) Go to StatPlot and go to your plot. Change the Ylist: to LRESID. This list is found in your List menu 2nd STAT :
5) Then press Zoom 9 : ZoomStat (You might want to turn your Y1 off)

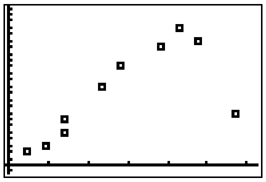


```
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Ylist:RESID
Mark: ▪ + ·
```



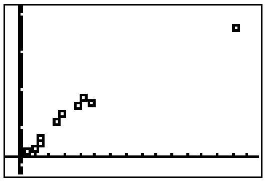
Influential observations:

Outliers – observations that lies outside the overall pattern of the other observations.



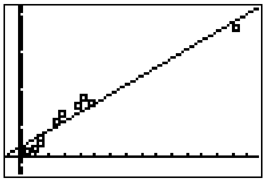
- the last data point is clearly an outlier.

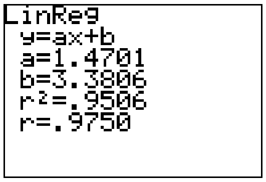
Influential observations – an observation that, if removing it would markedly change the results of the calculation of the LSRL and *r*.



- the last data point is an influential point.

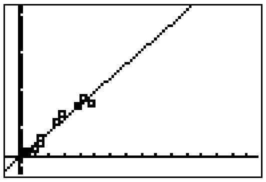
With the influential point:

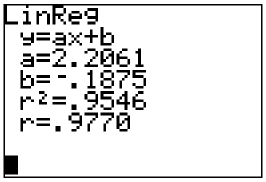




```
LinReg
 y=ax+b
 a=1.4701
 b=3.3806
 r²=.9506
 r=.9750
```

Without the influential point:





```
LinReg
 y=ax+b
 a=2.2061
 b=-.1875
 r²=.9546
 r=.9770
```

An outlier is influential. An influential point is not necessarily an outlier.