EXAMPE1(PG 164) DOES FIDGETING KEEP YOU SLIM?

3.2 Least Squares Regression Line (LSRL)

Regression Line

A regression line is a line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x.

Regression Lines as Mathematical Models

Suppose that y is a response variable (plotted on the vertical axis) and x is an explanatory variable (plotted on the horizontal axis). A regression line relating y to x has an equation of the form

$$y = a + bx$$

In this equation, b is the **slope**, the amount by which y changes when x increases by one unit. The number a is the y **intercept**, the value of y when $x = 0$.

Example to investigate the steps to develop an LSRL equation (ANSWER THE FOLLOWING QUESTIONS)

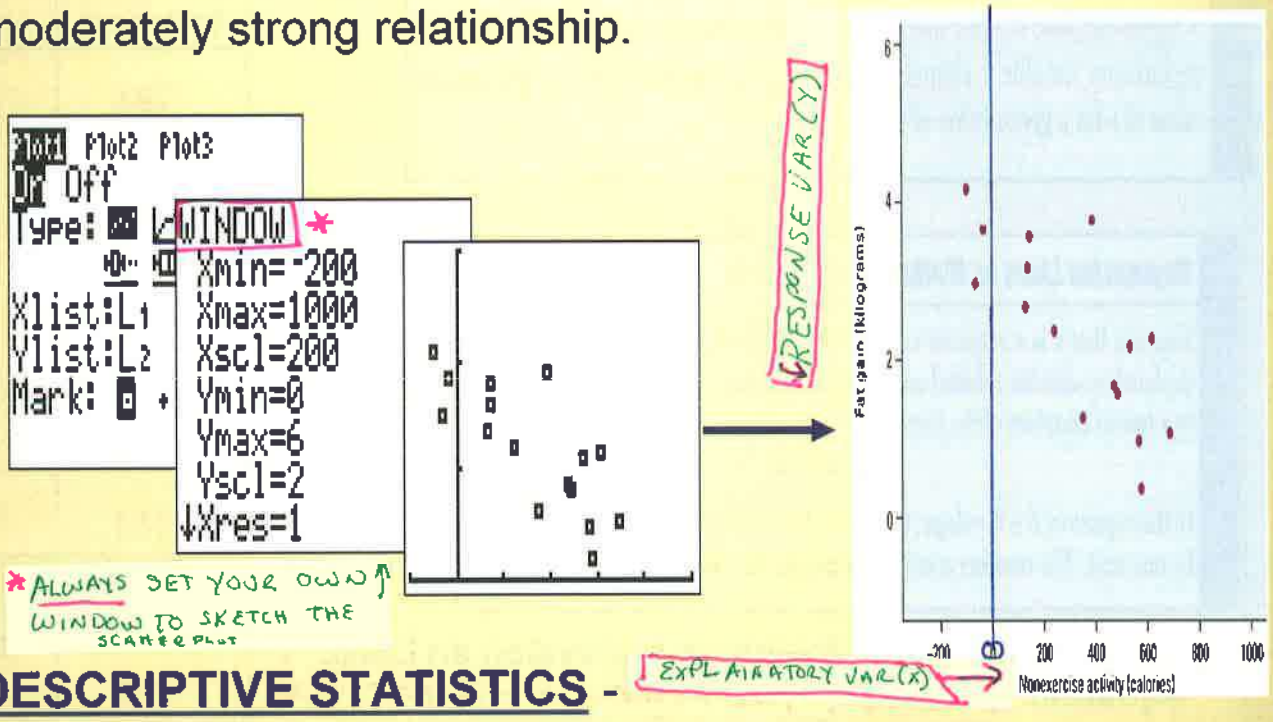
1. Enter L1 - Non-exercise activity
2. Enter L2 – Fat Gained
3. Plot the scatter plot. What is the association (direction, form, and strength)?
4. Find the mean and standard deviation for both variables in context.
5. Find the linear regression equation. What does it mean?
6. Plot the LSRL on the scatterplot. What are residuals?
7. Plot the residuals. What does this mean?
8. How do you assess the model? What does r^2 mean?
9. Use the LSRL equation to make predictions. When is it inappropriate to predict with LSRL?

NEA (calories)	Fat Gained (kilograms)
-94	4.2
-57	3.0
-29	3.7
135	2.7
143	3.2
151	3.6
245	2.4
355	1.3
392	3.8
473	1.7
486	1.6
535	2.2
571	1.0
580	0.4
620	2.3
690	1.1

Review the Data

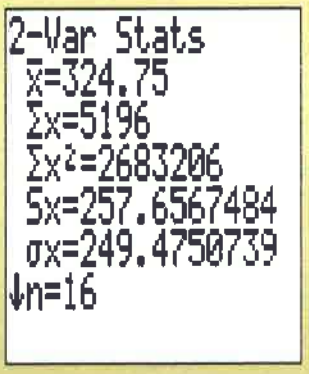
1ST STEP – ALWAYS GRAPH!!!

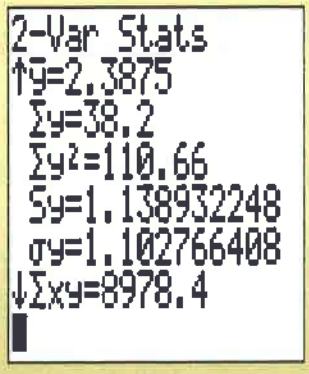
THE SCATTERPLOT - The relationship between non-exercise activity and fat shows a negative association, with a linear form, and appears to have a moderately strong relationship.

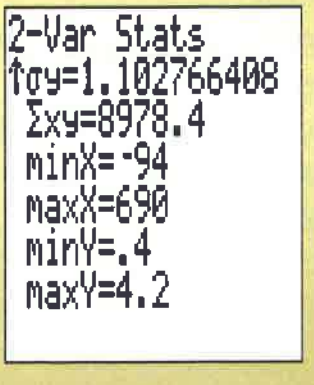


DESCRIPTIVE STATISTICS -

- The mean for non-exercise activity is about 325 calories with a standard deviation of about 258 calories with a spread (based on range) of 794 calories.
- The mean for fat gain is 2.39 kilograms with a standard deviation of 1.14 kilograms and a spread (based on range) of 3.8 calories.







Equation of LSRL

Equation of the Least-Squares Regression Line

We have data on an explanatory variable x and a response variable y for n individuals. From the data, calculate the means $\bar{x}$ and $\bar{y}$ and the standard deviations s_x and s_y of the two variables and their correlation r. The least-squares regression line is

$$\hat{y} = a + bx$$

$\hat{y} = b_0 + b_1 x$

with slope

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

$b_1 = r\,\frac{s_y}{s_x}$

that passes through the point $(\bar{x}, \bar{y})$.

DO NOT NEED TO Memorize

- These are on your AP GREEN SHEET
- Check it out!

```
LinReg
 y=a+bx
 a=3.505122916
 b=-.003441487
 r²=.6061492049
 r=-.7785558457
```

Y-INT (a)

slope (b)

COEF OF DETERMINATION

CORR COEF

The slope here B = —.00344 tells us that fat gained goes down by .00344 kg for each added calorie of NEA according to this linear model.

Our regression equation is the predicted RATE OF CHANGE in the response y as the explanatory variable x changes.

↑ FAT GAIN NONEXERCISE ↑ ACTIVITY (NEA)

The Y intercept a = 3.505kg is the fat gain estimated by this model if NEA does not change when a person overeats.

$\hat{y}$2

LSRL EQUATION: Y hat = 3.51- .0034X

*** * (better to use words) * ***

YHAT = $\hat{y}$

$\hat{y}$ - predicted y

(Fat Gain) hat = 3.51- .0034(NEA)

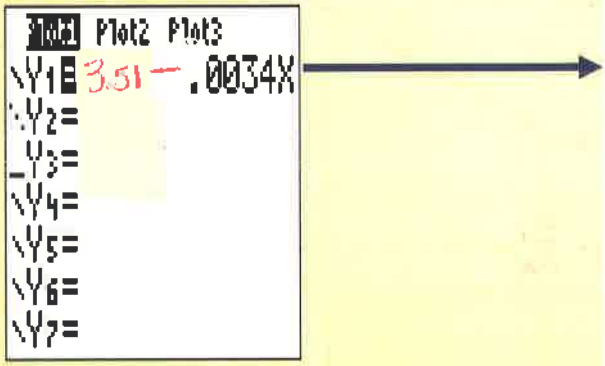
PREDICTED VARIABLES MUST WEAR A HAT!

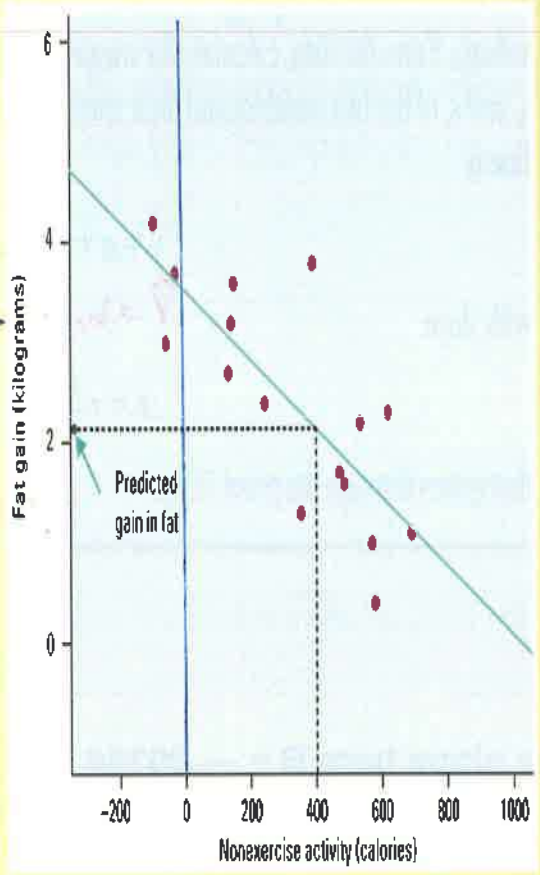
Graph the LSRL on our Scatterplot

LSRL EQUATION:

Y hat = 3.51- .0034X

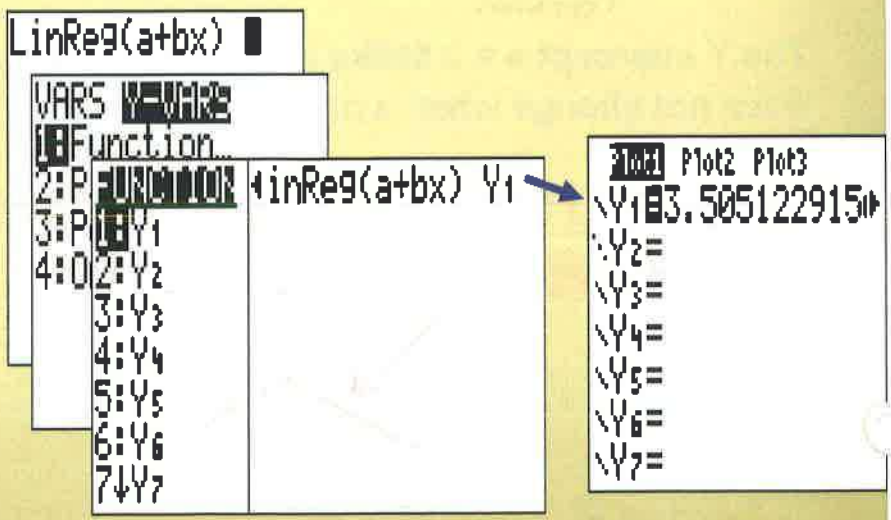
(Fat Gain) hat = 3.51- .0034(NEA)





TIP:

But in fact, we can get the LSRL equation, from our calculator by saving the equation when we calculate the LSRL. Here are the steps if you want to try on your own.



The LSRL "Line"

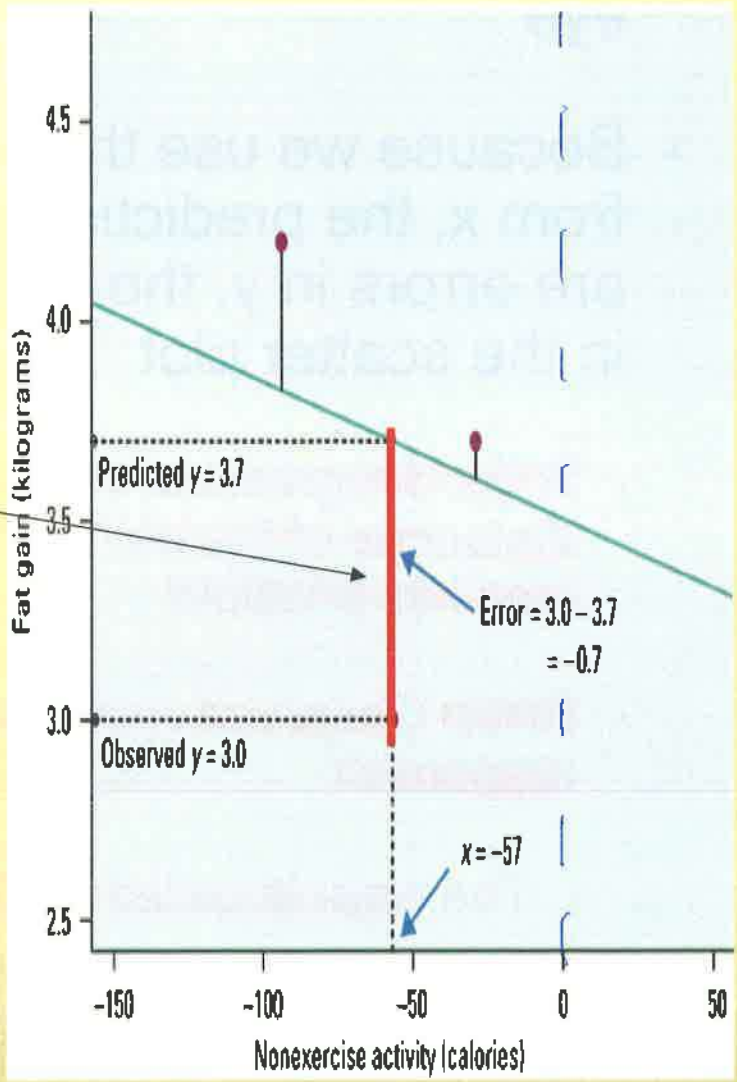
- In most cases, no line will pass exactly through all the points in a scatter plot and different people will draw different regression lines by eye.
- Because we use the line to predict y from x, the prediction errors we make are errors in y, the vertical direction in the scatter plot
  - A good regression line makes the vertical distances of the points from the line as small as possible
  - Error: Observed response - predicted response
    - The error is called RESIDUALS!

Goal of LSRL

Least-Squares Regression Line

The **least-squares regression line** of y on x is the line that makes the sum of the squared vertical distances of the data points from the line as small as possible.

- **Goal of LSRL is to minimize error.**
- **The error is called residuals.**
- **Want to minimize the sum of the residuals squared.**



Residuals

Residuals

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

$$\text{residual} = \text{observed } y - \text{predicted } y$$
$$= y - \hat{y}$$

- The error of our predictions, or vertical distance from predicted Y to observed Y, are called residuals because they are "left-over" variation in the response.

EXA MPLE: One subject's NEA rose by 135 calories. That subject gained 2.7 KG of fat. The predicted gain for 135 calories is

Predicted: ($\hat{Y}$)

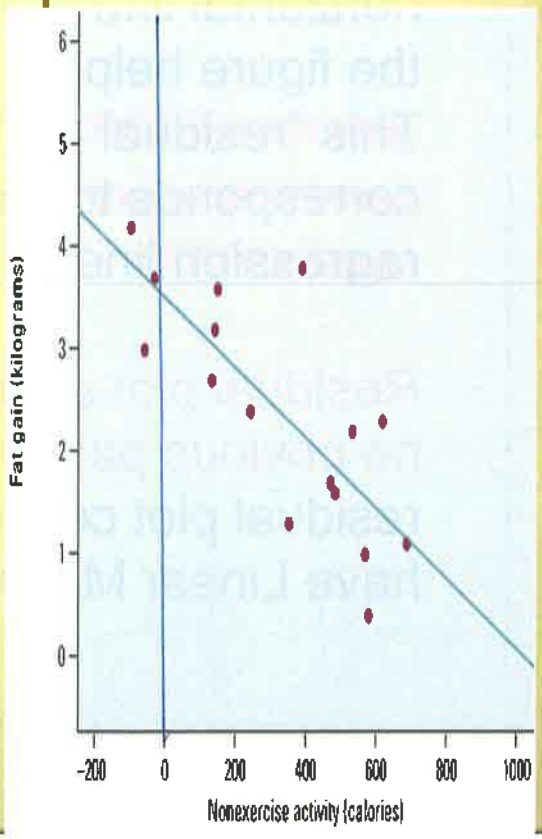
Y hat = 3.505- .00344(135) = 3.04 kg

Observed: 2.7 KG of fat

The residual for this subject is

($Y - \hat{Y}$)

y – yhat = 2.7 - 3.04 = -.34 kg

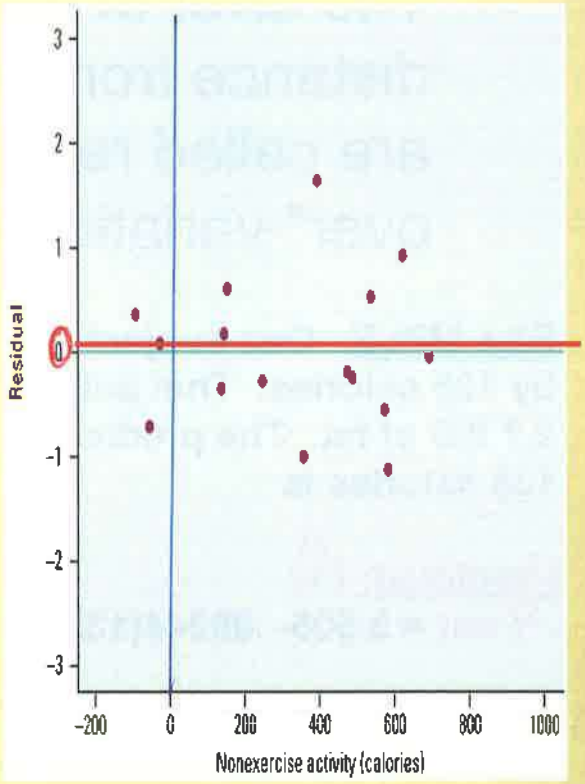


Residual Plot

Residual Plots

A **residual plot** is a scatterplot of the regression residuals against the explanatory variable (or equivalently, against the predicted y-values). Residual plots help us assess how well a regression line fits the data.

- The sum of the least-squares residuals is always zero.
- The mean of the residuals is always zero, the horizontal line at zero in the figure helps orient us. This "residual = 0" line corresponds to the regression line
- Residual plot should show no obvious pattern. Our residual plot confirms we have Linear Model.



Residuals Plots on Calc

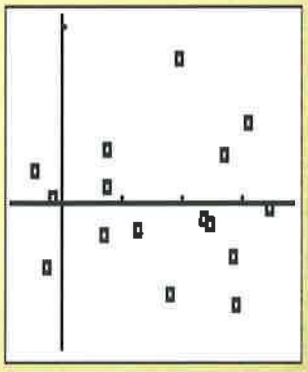
- If you want to get all your residuals listed in L3 highlight L3 (the name of the list, on the top) and go to 2nd- stat- RESID then hit enter and enter and the list that pops out is your resid for each individual in the corresponding L1 and L2. (if you were to create a normal scatter plot using this list as your y list, so x list: L1 and Y list L3 you would get the exact same thing as if you did a residual plot defining x list as L1 and Y list as RESID as we had been doing).
- This is a helpful list to have to check your work when asked to calculate an individuals residual.

2nd STAT PLOT

2nd LIST

ZOOM 9

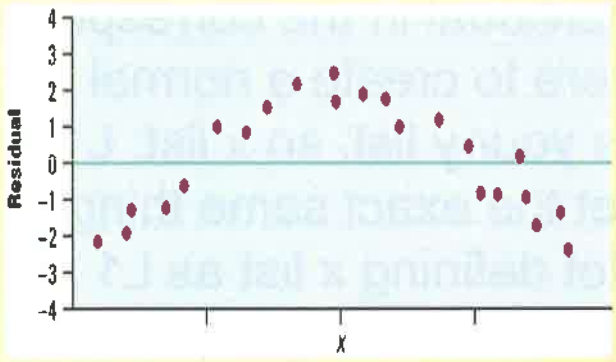
Residual Plot



Assessing Models

Examining Residuals Plots and Residual Standard Deviation

- **Residual plot should show no obvious pattern.** Residual plot should show no obvious pattern. **A <u>curved</u> pattern shows that the relationship is not linear and a straight line may not be the best model.**



- Residuals should be regression line in a model that fits the data well should come close" to most of the points.
  - A commonly used measure of this is the standard deviation of the residuals, given by:

$$s=\sqrt{\frac{\sum residds^2}{n-2}}$$

For the NEA and fat gain data, s_e = $\sqrt{\frac{7.663}{14}}=.740$

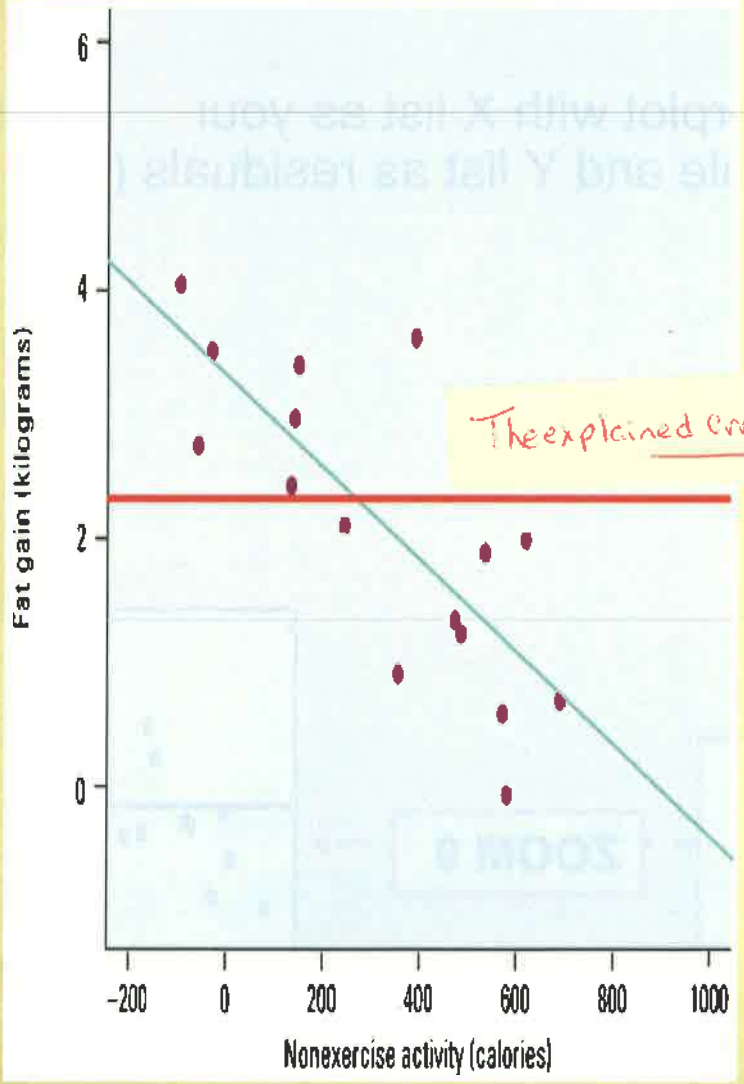
Assessing Models

R squared- Coefficient of determination

The Coefficient of Determination: r^2 in Regression

The **coefficient of determination** r^2 is the fraction of the variation in the values of y that is explained by the least-squares regression line of y on x. We can calculate r^2 using the following formula:

```
LinReg
 y=a+bx
 a=3.505122916
 b=-.003441487
 r²=.6061492049
 r=-.7785558457
```



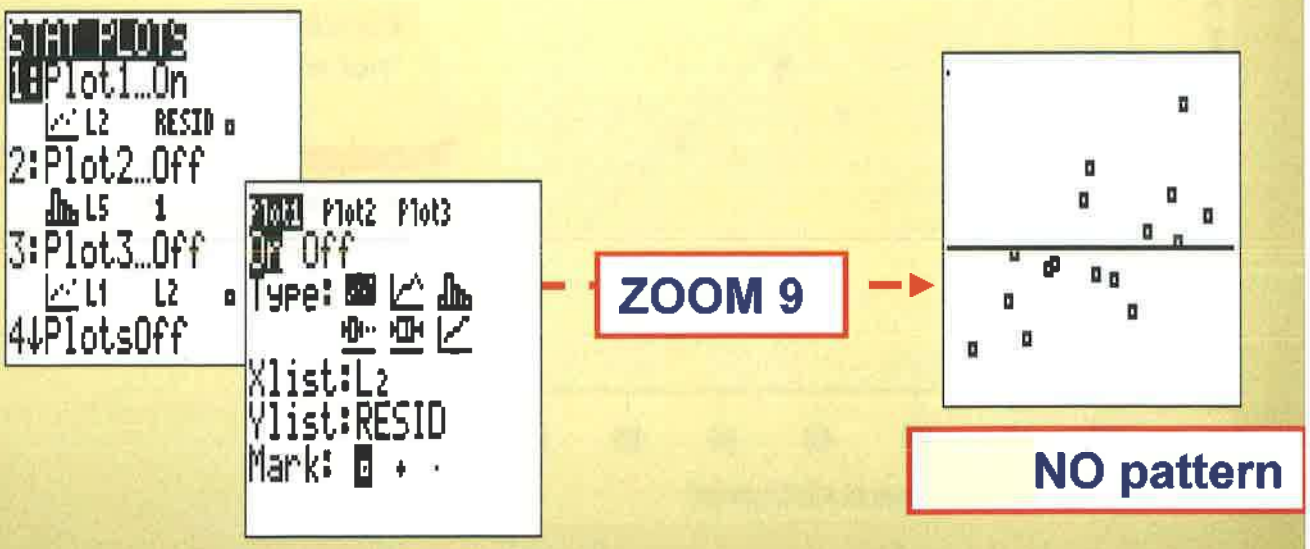
1. If all the points fall directly on the least-squares line, r squared = 1. Then all the variation in y is explained by the linear relationship with x.

2. So, if r squared = .606, that means that 61% of the variation in y among individual subjects is due to the influence of the other variable. The other 39% is "not explained".

3. r squared is a measure of how successful the regression was in explaining the response

Plot Residual vs X or Y

- Produce Scatterplot and Regression line from data (lets use BAC if still in there)
  - Turn all plots off
  - Create new scatterplot with X list as your explanatory variable and Y list as residuals (2nd stat, resid)
  - Zoom Stat



NOT PART OF THIS EXAMPLE

Prediction

- We can use a regression line to predict the response y for a specific value of the explanatory variable x (but only for the range of x values used in our LSRL).

> ***Extrapolation***
>
> **Extrapolation** is the use of a regression line for prediction outside the range of values of the explanatory variable x used to obtain the line. Such predictions are often not accurate.