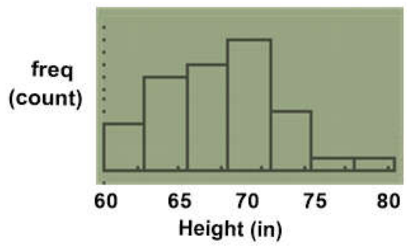
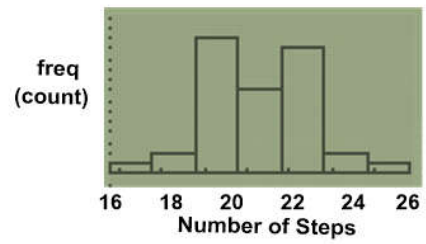
AP Statistics – Lesson Notes - Chapter 7,8,9: Correlation, Linear Regression

Relationships between two quantitative variables

Last unit (chapters 3-6) focused on analyzing a **single variable** and considering some variation of **count vs. category**:

Consider our height and step data. We could make histograms to display either variable individually to see how each variable is distributed...

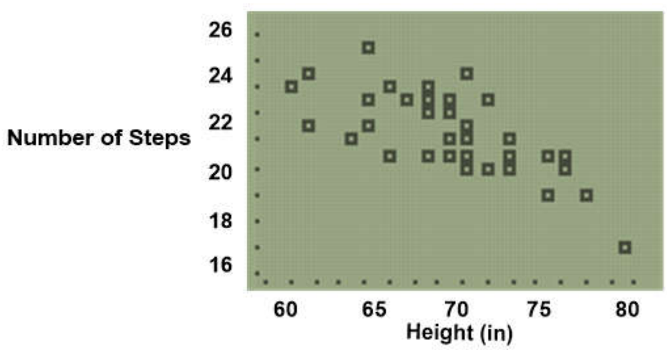




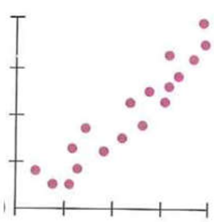
But this ignores the fact that each data point in height is tied to a specific data point in number of steps (matched by person).

Unit 2 focuses on analyzing **two variables** which are both **quantitative**. We will not be dividing into 'bins' and looking at counts. Instead, we will look at how the value of one variable is **related to, or associated with,** the value of the other variable.

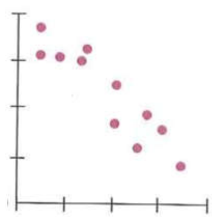
Our main tool to visualize relationships between two variables is a **Scatterplot**:



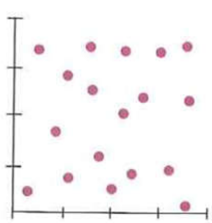
Scatterplots give us a visual indication of whether there is an **association** between the variables, the **direction** of the association, and a rough idea of the strength. The 'shape' of the cluster of points is called the **form**.



Positive Association
(y increases as x increases)

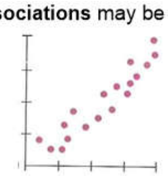


Negative Association
(y decreases as x increases)

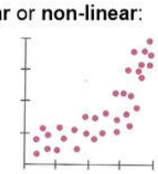


No Association

Associations may be **linear** or **non-linear**:

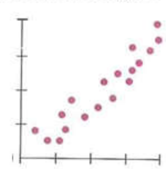


Linear Association

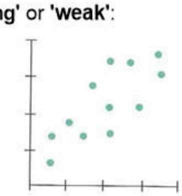


Non-Linear Association

Associations may be **'strong'** or **'weak'**:



'stronger'



'weaker'

<u>To describe an association, include</u>: form, direction, strength

"There is a linear, moderately strong, negative association between steps and height."

Roles of the variables

Which variable should be x, and which y? It depends upon how we think about the variables and if we believe variation in one can *predict* variation in the other.

x-variable (a.k.a. explanatory, predictor, independent variable)
The variable we can vary and expect that the other variable will depend upon the value of this variable. The variable associated with 'cause' if there is a causal relationship. Often this variable is 'time' or 'year'.

y-variable (a.k.a. response, dependent variable)
The variable we expect will be affected by the value of the x-variable. Often, this is the variable we are studying in an analysis (e.g. if studying auto accidents per year, auto accidents would be the y-variable). The variable associated with 'effect' if there is a causal relationship.

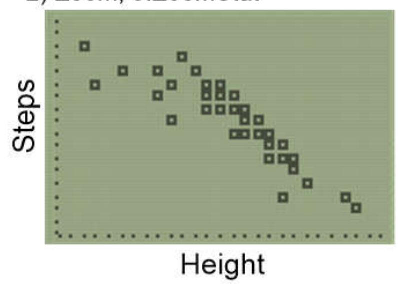
Scatterplots using a calculator

Enter data for x variable in L1, y variable in L2. Let's use our height vs steps data:

Scatterplot

1) 2nd, Y (Stat Plot), enter

2) Zoom, 9:ZoomStat



Correlation and Association

Association is a general term meaning there appears to be some relationship between the variables.

Correlation is a precise term describing the strength and direction of a linear relationship.

The strength of association (correlation) can be found by computing the **correlation coefficient** of a dataset:

correlation coefficient: (Pearson's)

$$r = \frac{\sum z_x z_y}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{\left(x_i - \bar{x}\right)}{s_x}\frac{\left(y_i - \bar{y}\right)}{s_y}$$

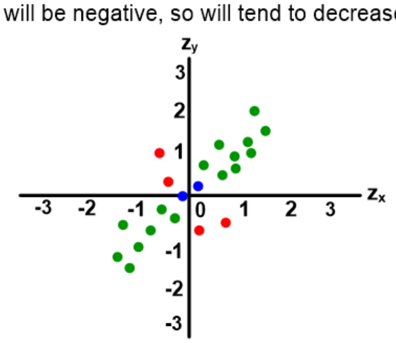
Computationally, the correlation coefficient is the average of the product of the z-scores.

Correlation coefficient is a number between -1 (perfect negative correlation) and +1 (perfect positive correlation). Correlation coefficient of 0 means there is no association between the variables.

How does correlation coefficient work?

Points in quadrants 1 and 3 tend to indicate positive direction and $z_x z_y$ will be positive, so will tend to increase the sum.

Points in quadrants 2 and 4 tend to indicate negative direction and $z_x z_y$ will be negative, so will tend to decrease the sum.



Correlation Coefficient on Ti84

Calculator can compute correlation coefficient, but you need to run linear regression (explained more fully later):

1) Mode, Stat Diagnostics: set to ON (only done once)

2) Data entered into L1 and L2 (or any list)

3) Stat, -> Calc, 8: LinReg(a+bx)

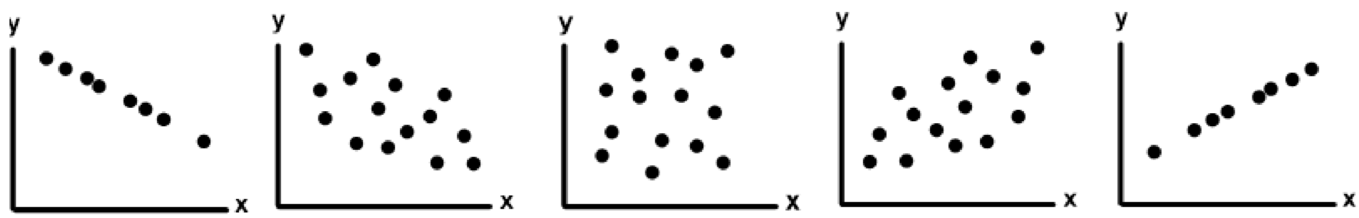
4)

5) Screen should display r value.

Correlation Coefficient properties

range of possible r values

-1 -0.5 0 0.5 1



Drawing a 'trend line' (in this course, called a Least-Squares Regression Line, LSRL)

- **always a line (we never use curves in this course)**
- **about half the points above and half below**
- **average distance of the points from the line (in the y-direction) is always zero.**

-1 to -0.8	-0.8 to -0.5	-0.5 to -0.2	-0.2 to 0.2	0.2 to 0.5	0.5 to 0.8	0.8 to 1
strong association	moderate association	weak association	no association	weak association	moderate association	strong association

- There is no formal agreement about what constitutes 'strong' or 'weak' association. It depends upon context (but rules of thumb shown above)

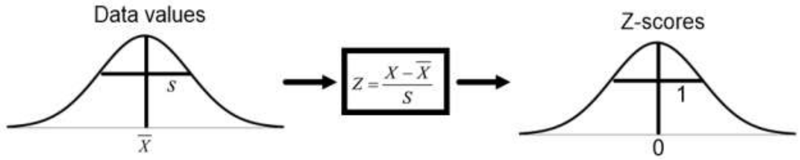
r is a measure of how close the points are to the LSRL

Correlation Conditions and Properties

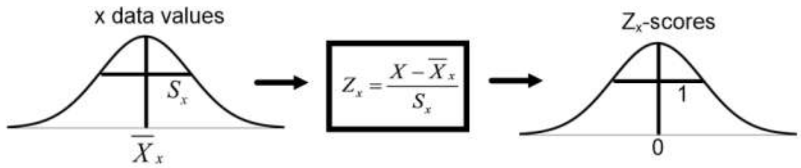
- Both variables must be quantitative.
- Correlation coefficient is very sensitive to outliers. Consider computing with and without outliers included.
(Check for linearity and outliers with a scatterplot).
- r has no units (although some people, incorrectly, report it as a percentage).
- r is unaffected by changes in center or scale of either variable.

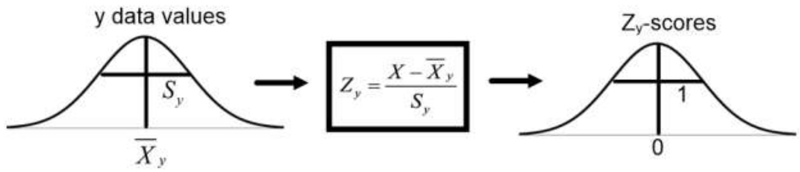
Although the 'x' and 'y' data are connected, they are also each separate data sets

Remember, with a data set we can define things like the mean and standard deviation, and we can also convert values to z-scores...



With (x,y) data, each variable has a mean, and standard deviation, and could be converted to Z-scores...



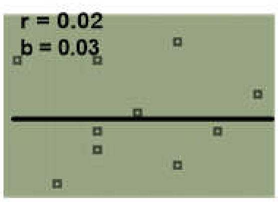


Correlation coefficient only measures the strength of <u>linear</u> association and it becomes invalid when slope approaches 0 or infinity

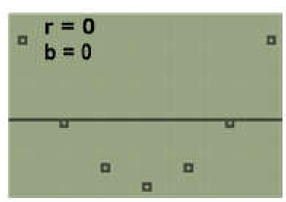
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{\left(x_i - \bar{x}\right)}{s_x}\frac{\left(y_i - \bar{y}\right)}{s_y} = \frac{1}{n-1}\sum_{i=1}^{n} z_x z_y$$

Since the correlation coefficient is a sum of the products of z scores, any point that has a z score near 0 in either x or y doesn't contribute to r.

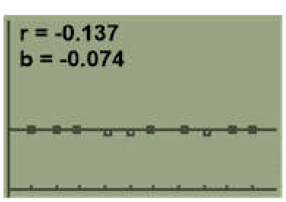
Some ways correlation coefficient, r, can be close to zero...



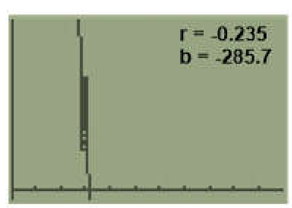
r is nearly 0 because there is no discernable trend of any kind.
(LSRL is a horizontal line at the average y)



r is nearly zero because even though there is a strong association, it isn't linear.



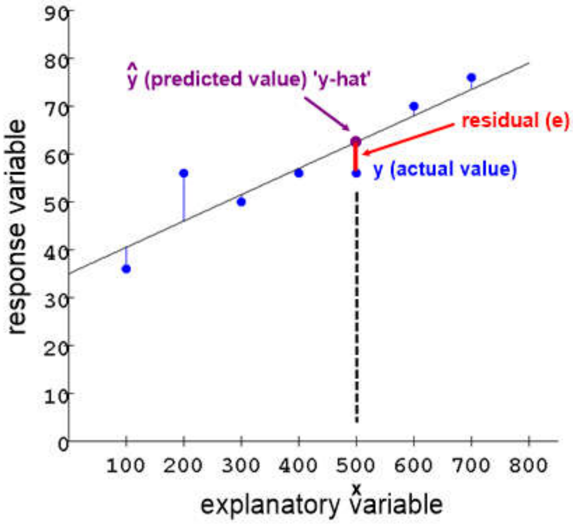
z_y is nearly 0 for all these points,so r is nearly 0 even though there is a strong linear association.



z_x is nearly 0 for all these points,so r is nearly 0 even though there is a strong linear association.

Linear Regression

Associations which are approximately linear on a scatterplot can be modelled with a line called the **Least Squares Regression Line (LSRL).** (Also known less accurately as 'linear model', 'line of best fit', or 'trend line').



A **residual** is the error between what the LSRL line predicts the y value will be (for a given x) and the actual y value.

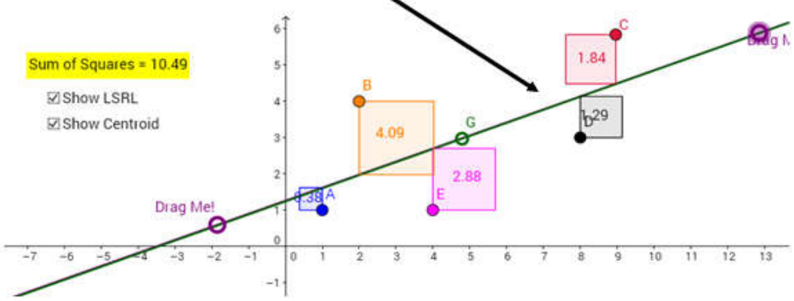
$$e = y - \hat{y}$$

Residuals can be positive or negative. To minimize the overall error, we square each residual and sum all the squared residuals. We then adjust the line so that the total sum of squares is minimized.

Let's look at a specific, small data set:

X	Y
1	1
2	4
4	1
8	3
9	6

The LSRL is the equation of the line which has the smallest sum of squares of the residuals (errors)



https://www.geogebra.org/m/PQfrXvW9#material/crBa6TAWhtt

Linear Regression on Ti84

If you have the full data set, you can find the LSRL with a calculator:

1) Mode, Stat Diagnostics: set to ON (only done once)

2) Data entered into L1 and L2.

3) Stat, -> Calc, 8: LinReg(a+bx) (careful, there is an ax+b also)

4)  or LinReg(a+bx) L1,L2,Y1

5) 

$$\hat{y} = (53.8471) - (0.5728)x$$

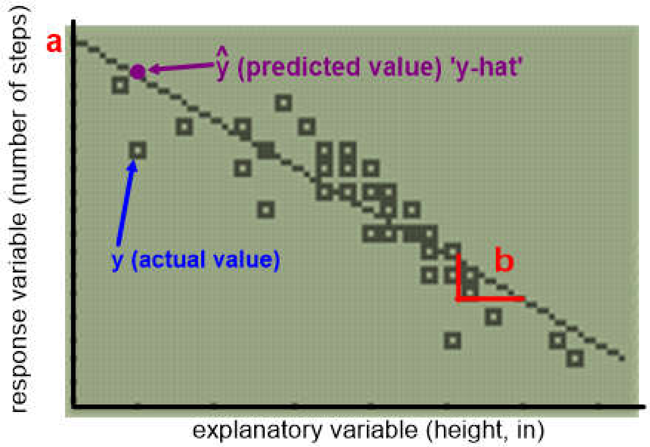
$$\widehat{steps} = (53.8471) - (0.5728)\,height$$

Linear Regression

Equation of the LSRL:

$$\hat{y} = a + bx \quad (calculator)$$
$$\hat{y} = b_0 + b_1 x \quad (textbook)$$

$$\hat{y} = (53.8471) - (0.5728)x$$



b: slope

For each increase of 1 *explanatory variable* unit, the LSRL predicts an increase (or decrease) of b units of the *response variable*.

"For every 1 added inch in height, the number of steps decreases by 0.5728 steps, on average."

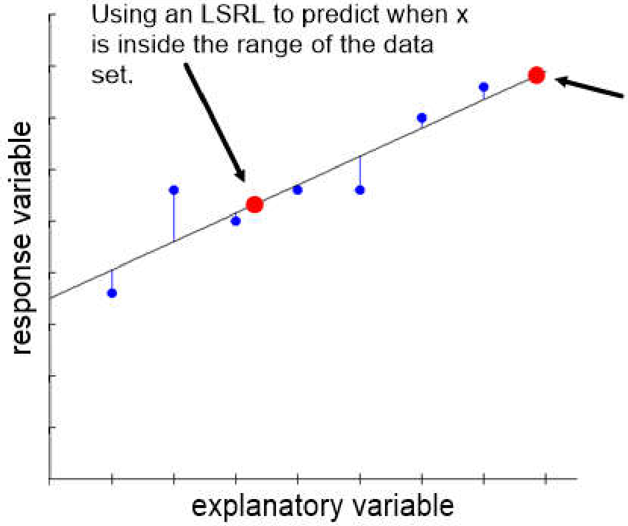
a: y-intercept

Most of the time, a doesn't have any particular meaning.

"A person who is zero inches tall is predicted to take 53.8471 steps, on average."

Linear Regression : Using the LSRL

The LSRL is a model that can be used to predict the y value for a given x value.



Interpolation

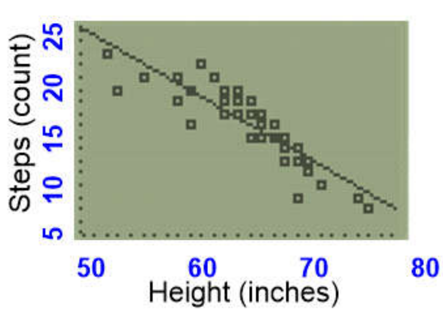
Using an LSRL to predict when x is inside the range of the data set.

Extrapolation

Using an LSRL to predict when x is outside the range of the data set.

Caution: extrapolation makes the assumption that data in a range we have no evidence about will continue following the same pattern.

What does the LSRL predict will be the number of steps a person 59" tall will take?



$$\hat{y} = (53.847) + (-.5728)x$$

What does the LSRL predict will be the number of steps a person 82" tall will take?

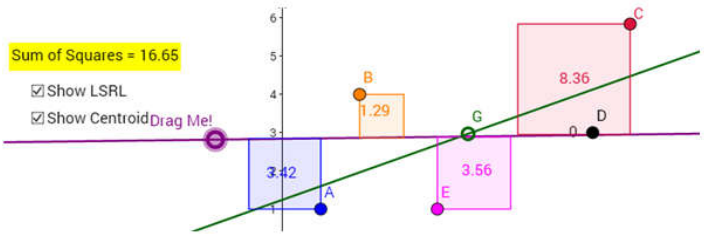
r^2: Coefficient of determination

The value r^2 is called the **coefficient of determination** and it is a measure of how 'good' the LSRL is at explaining the variation in y as x varies.

You can think of r^2 (x100) as the percentage of variation in the *response variable* that is explained by the LSRL which relates *explanatory variable* to *response variable.*

The reason why this is true is explained very well in the math box on p.172-173 of your textbook, but we can also explain intuitively here with our LSRL applet.

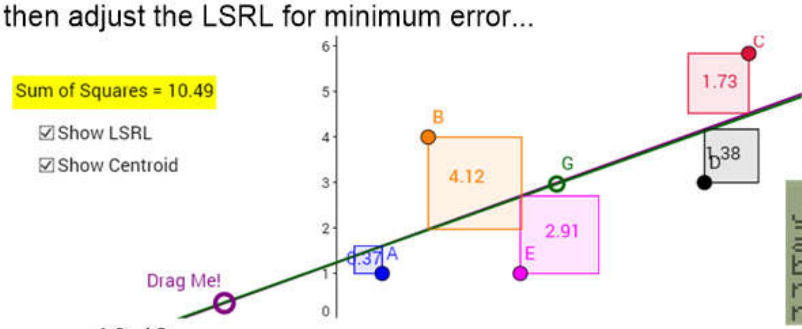
One way to explain this is to consider if there were no association between x and y:



This would mean that the y values are varying randomly, but not in any way connected with the x values. It would also mean **<u>if there is no association, the LSRL slope=0</u>**.

The sum of squares of the areas then represents the total 'error' or variation in y (16.65).

If we then adjust the LSRL for minimum error...

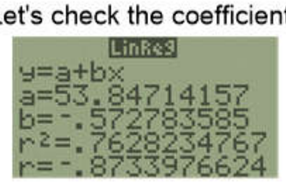


That means $\frac{10.49}{16.65} = 0.63$ or 63% of the variation in y is still present after we account for as much as we can with the LSRL. **<u>Which means about 37% of the variation in y is explained by the LSRL relating y to x</u>**. (The calculator shows this is 39.5%).

(If you want to see a derivation showing why r^2 works this way, look at "Analysis: What does r-squared tell us" on the www.mrfelling.com site)

r^2 = 1.0 would be a perfect model (every data point is exactly on the LSRL).
r^2 = 0.0 would be a terrible model.

Real models are between these extremes. Let's check the coefficient of determination for our steps vs. height model.



This means that approximately 76% of the variation in number of steps, from person to person, is associated with a person's height. The remaining 24% of variation in number of steps must be due to other factors.

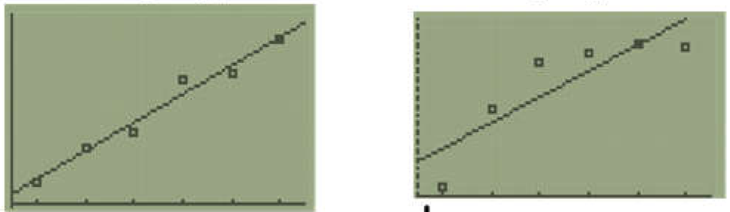
Residual Plots

A **residual plot** for a given linear regression shows the residual (r) vs. x.

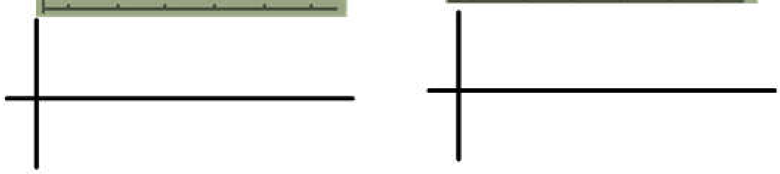
Examples: Sketch residual plots by hand for each data set

x	y
1	8
2	22
3	28
4	48
5	51
6	64

x	y
1	2
2	18
3	28
4	30
5	32
6	31



Scatter plots



Residual plots

If the residuals are randomly scattered around '0' then you know that a linear model is appropriate. (Residuals make it easier to see non-linearity compared to scatterplots.)

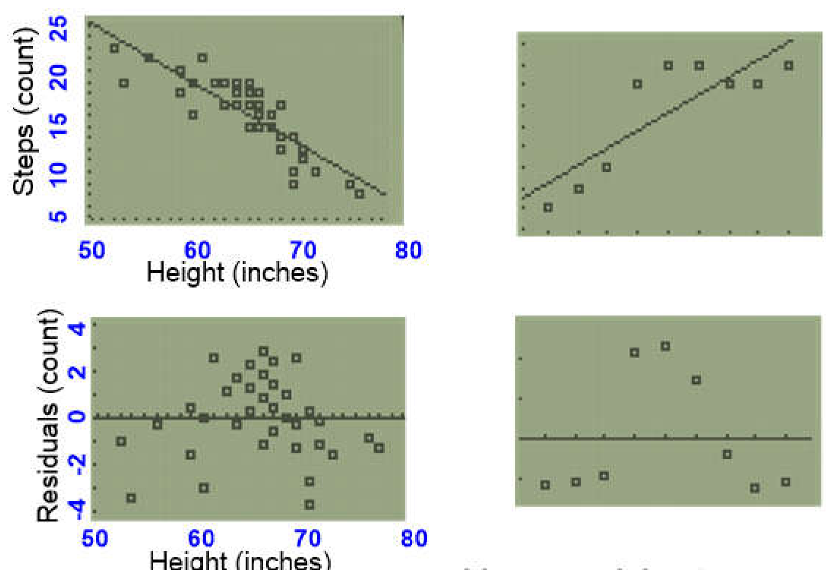
<u>To do residual plot on Ti84, after running a LinReg as normal</u>:

- 2nd, Y= (Stat Plots)
- Re-enter the stat plot you are using, or enable Stat Plot 1
- Change the Ylist for RESID (2nd LIST)
- Zoom, 9:ZoomStat

A **residual plot** for a given linear regression shows the residual (r) vs. x.

Scatterplot

Residual plot



Linear model appropriate **Linear model not appropriate**

- **Slope of an LSRL through residuals is always zero.**
- **Mean of the residuals is always zero.**
- **Standard deviation of the residuals is a measure of how points are spread around the LSRL.**

Getting Equation of LSRL from software output

One way to get an LSRL equation (if you have the data set) is to perform a LinReg and the calculator will provide the equation, r, and r^2 value. But sometimes, a problem will not give you the data set, but instead provide the output of analysis performed by someone else using a software program:

Dependent variable is: **Concentration**
No Selector
R squared = 90.8% *R squared (adjusted) = 90.6%*
s = 3.472 with 43 - 2 = 41 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	*4900.55*	*1*	*4900.55*	*407*
Residual	*494.199*	*41*	*12.0536*	

	Variable	Coefficient	s.e. of Coeff	t-ratio	prob
this row about intercept term (a) →	*Constant*	*40.3266*	*1.295*	*31.1*	*≤ 0.0001*
this row about the x variable term (b) →	*Time*	*-5.95956*	*0.2956*	*-20.2*	*≤ 0.0001*

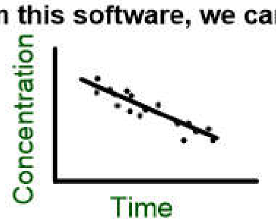
$$\hat{y} = a + bx$$

Dependent variable is: $\boxed{\text{Concentration}}$ y variable
No Selector
$\boxed{\textit{R squared = 90.8\%}}$ *R squared (adjusted) = 90.6%*
s = 3.472 with 43 - 2 = 41 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	*4900.55*	*1*	*4900.55*	*407*
Residual	*494.199*	*41*	*12.0536*	

	Variable	Coefficient	s.e. of Coeff	t-ratio	prob
this row about intercept term (a) →	*Constant*	*40.3266* = a	*1.295*	*31.1*	*≤ 0.0001*
this row about the x variable term (b) →	$\boxed{\textit{Time}}$ x variable	*-5.95956*= b	*0.2956*	*-20.2*	*≤ 0.0001*

From this software, we can get the following info...



$\widehat{\textit{concentration}}$ **= 40.3266 - 5.95956 (*time*)**

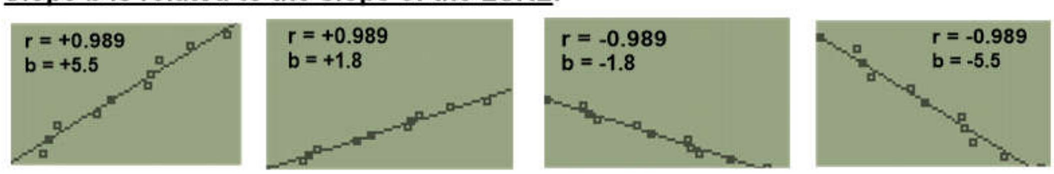
$r^2 = 0.908$ $\quad r = \pm\sqrt{0.908} = \pm 0.953$

and slope is negative, so r is also negative:

$$r = -0.953$$

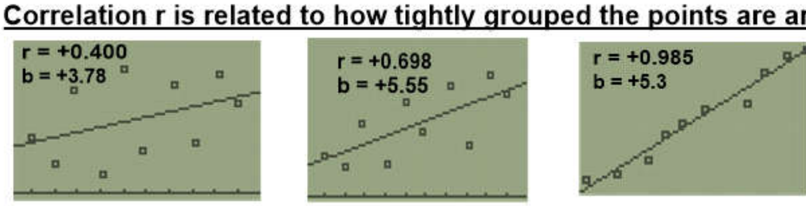
Relationship between slope b, r, and Scatter plot shape:

<u>Slope b is related to the slope of the LSRL</u>:



(r value varies independently of slope, but it *does match the sign of the slope)*

<u>Correlation r is related to how tightly grouped the points are around the LSRL</u>:



Using the slope equation

(If you don't have the full data set, or a software analysis output, you can still calculate various useful information.)

A very useful equation: $\boxed{b = r\frac{s_y}{s_x}}$

(If you want to know more about this equation, look at "Derivation of the b=r(sy/sx) equation" on the www.mrfelling.com site)

Here, we'll look at a few things that you can do with this equation.

1) Given summary statistics (but no data) find the slope and LSRL

$\overline{height} = 65.8333\ in$ $\quad s_{height} = 4.9598\ in$ $\qquad \overline{steps} = 16.1389$ $\quad s_{steps} = 3.2527$ $\qquad r = -.8734$

a) Use formula to calculate b.

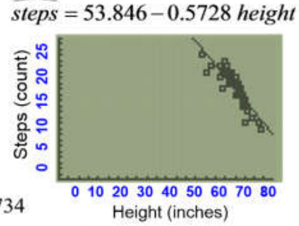
b) Find y-intercept a - the centroid $(\bar{x}, \bar{y})$ will *always* lie on the LSRL.

c) Write out the LSRL.

2) Find LSRL if x and y are swapped

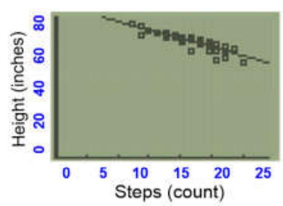
$\widehat{steps} = 53.846 - 0.5728\ height$

$\overline{steps} = 16.1389$



$r = -.8734$

$\overline{height} = 65.8333\ in$



Does r change? **no (r measures strength of association, swap does not affect r)**

Does b change? **yes** Before swap: $b = r\frac{s_y}{s_x}$ After swap: $b = r\frac{s_x}{s_y}$

3) Given data value std dev in x, find corresponding std dev in y

If a student's height is 1.2 standard deviations above the mean height of the class, how many standard deviations above the mean number of steps would you predict this student's number of steps to be?

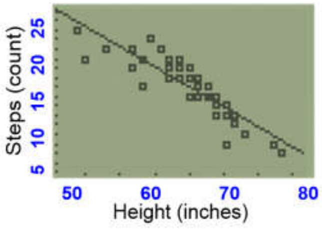
(When considering standard deviations, these are z-scores, so $s_y = s_x = 1$)

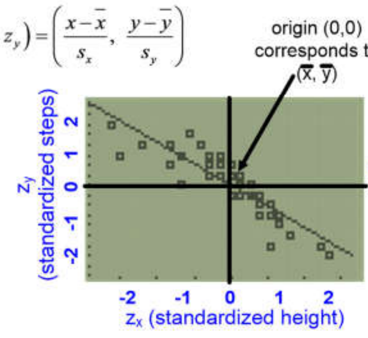
Why does this work? Standardizing a scatterplot...

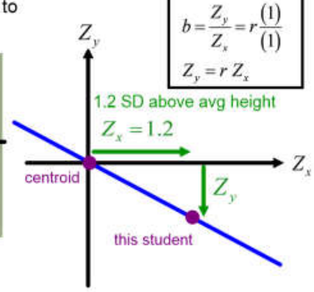
The general shape of a scatterplot remains the same even if the units are changed, but may look stretched depending upon the scale size. But a consistent scatterplot shape can be produced by using the z-score standardized values for each variable, so that each axis plots number of standard deviations a value is away from that variable's mean:

$$\left(z = \frac{x-\bar{x}}{s}\right) \qquad (z_x, z_y) = \left(\frac{x-\bar{x}}{s_x}, \frac{y-\bar{y}}{s_y}\right)$$

$$b = \frac{Z_y}{Z_x} = r\frac{s_y}{s_x} \qquad b = \frac{Z_y}{Z_x} = r\frac{(1)}{(1)} \qquad Z_y = r\,Z_x$$







The 3 ways to find the equation of an LSRL...

1) **<u>If you have the full data set</u>:**
   Use calculator. Enter data in L1, L2 and run LinReg.

2) **<u>If you have the output of a software analysis</u>:**
   Use the values in the table:
   - **Look for the word 'coefficient' or 'estimate':**
   - **One row is for the y-intercept, labelled 'constant' or 'intercept'.
     (The coefficient of this row is the y-intercept, 'a')**
   - **One row is for the x term, labelled the name of the x-variable.
     (The coefficient of this row is the slope, 'b')**

3) **<u>If you have no data or software output, but summary data on x and y</u>:**
   - **Use the formula** $b = r\dfrac{s_y}{s_x}$ **to find the slope, b.**
   - **Solve for y-intercept, a, by plugging in the centroid** $\left(\overline{x}, \overline{y}\right)$
     (the only point that is always on the LSRL) and solving for a.

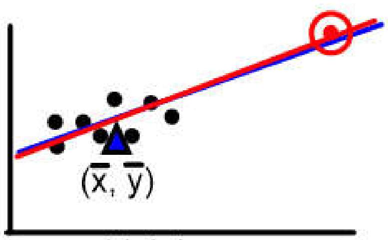
The effect of outliers on linear regression https://www.geogebra.org/m/NZpWpCW8

Any outlier deserves attention due to the unduly large effect it may have on results.
Outliers may (or may not) affect correlation and/or slope.

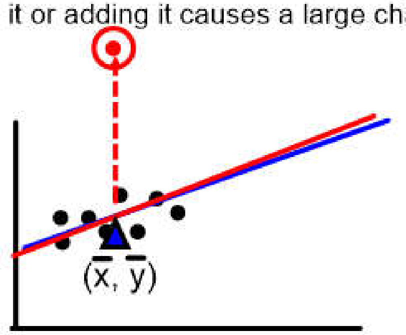
Outlier effect on slope - the concept of 'leverage'

The point $(\bar{x}, \bar{y})$ is always on the LSRL and you can think of it like a 'fulcrum' of a lever.
A data point whose x value is the same as the fulcrum will have no effect on the LSRL, but a line whose x value is far away from the fulcrum has a large effect and we say this is a **high leverage point.**

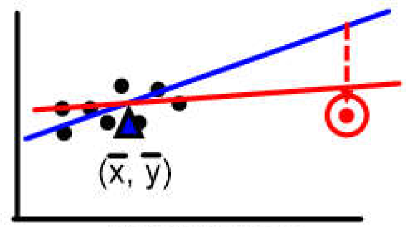
But for the point to cause a change in slope of the LSRL, it needs to have a high residual. A point already near the LSRL won't 'push' the LSRL and cause much change. If a point has high leverage and high residual (so that removing it or adding it causes a large change in the LSRL slope) we say that point is **influential.**



high leverage
small residual, not influential



high residual but
low leverage, not influential

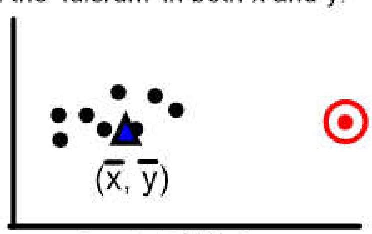


high leverage
large residual, influential

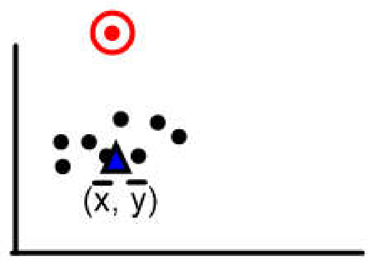
Outlier effect on correlation

- Correlation is the measure of the strength of the association and is high (close to + 1 or -1) if the points are grouped tightly around the LSRL.
- Correlation is calculated as sum of products of standardized distances x and y from the mean, so a point has a large effect on correlation if it is far from the 'fulcrum' in both x and y.

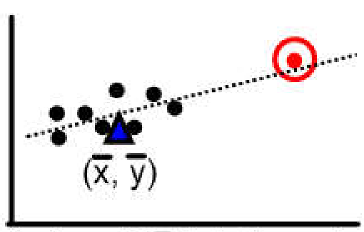
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{\left(x_i-\bar{x}\right)}{s_x}\frac{\left(y_i-\bar{y}\right)}{s_y}$$



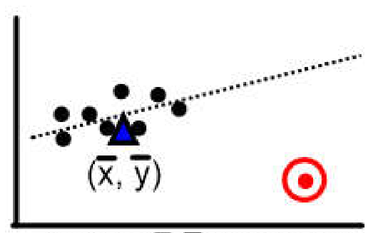
close to $(\bar{x}, \bar{y})$ in y
little effect on correlation



close to $(\bar{x}, \bar{y})$ in x
little effect on correlation

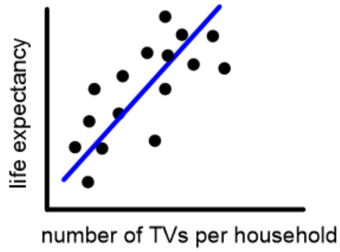


far from $(\bar{x}, \bar{y})$ in both x and y
strongly affects correlation
including this point strengthens correlation



far from $(\bar{x}, \bar{y})$ in both x and y
strongly affects correlation
including this point weakens correlation

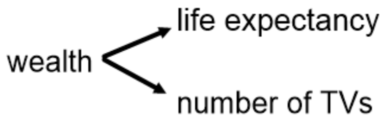
Association (correlation) does not imply Causation



So, if we want people to live longer, we should just give them more TVs, right?

A strong association between two variables **does not mean that one variable is *causing*** a change in the other variable.

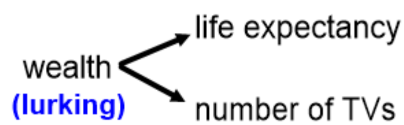
What might be going on here to explain this?



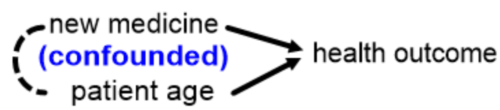
Both life expectancy and number of TVs might be changing according to the value of another variable such as wealth.

Terminology

Lurking variable: When one variable causes two other variables to change together, making them appear associated. (Used by our textbook, not an official term)



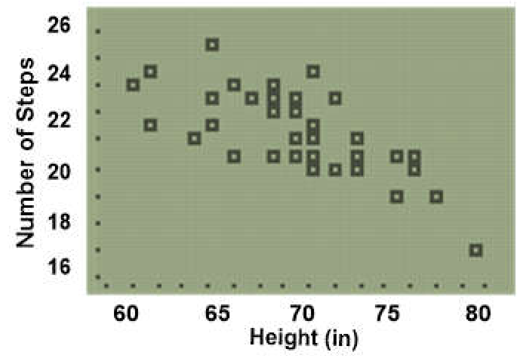
Confounded variables: When the effect of multiple explanatory variables on a response variable can't be separated. (Official term, used on AP Statistics Exam)



Association: General term meaning there appears to be some relationship between variables.
(Official term, used on AP Statistics Exam)

Correlation: Precise term describing the strength and direction of a linear relationship (usually taken to mean the correlation coefficient, r).
(Official term, used on AP Statistics Exam)

Standard explanation wordings:



$$\hat{y} = (53.8471) - (0.5728)x$$

$$r = -0.873 \qquad r^2 = 0.763$$

$$s = 1.58$$

slope, b:

"For every 1 added inch in height, the number of steps decreases by 0.5728 steps, on average."

intercept, a:

"A person who is zero inches tall is predicted to take 53.8471 steps, on average."

association / explaining r (correlation coefficient):

"There is a linear, negative, fairly strong association between number of steps and height".

r^2 (coefficient of determination):

"About 76% of the variation in number of steps is explained by the LSRL which relates number of steps to height."

s (standard deviation of the residuals):

"The actual number of steps (for a given height) are 1.58 steps away from the predicted number of steps, on average." or "The average error between actual and predicted number of steps for a given height is 1.58 steps."