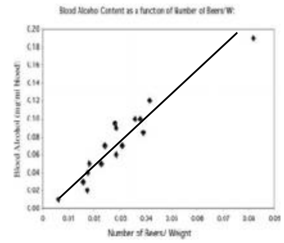
Least-Squares Regression Line (LSRL)

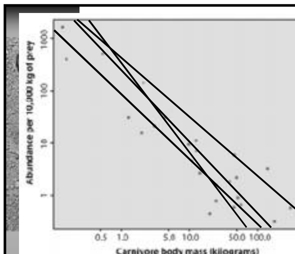
The Regression Line

- 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***.
- . A model for the data, much like the density curves.



The Least-Squares Regression Line

- We all got slightly different lines, although they agree well enough that our predictions only disagree by a few cm.
- How can we define a process so that we all get the same answer from the same data?
- So let's find a way to get an equation that automatically minimizes the sum of the areas of the squares.
  - That equation is called the ***Least Squares Regression Line*** and is abbreviated **LSRL**



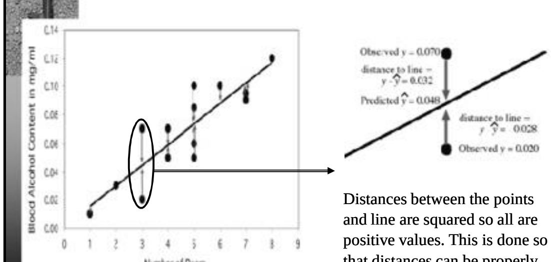
Correlation tells us about *strength* (scatter) and *direction* of the linear relationship between two quantitative variables.

In addition, we would like to have a numerical description of how both variables vary together. For instance, is one variable increasing faster than the other one? And we would like to make predictions based on that numerical description.

But which line best describes our data?

The Regression Line

The least-squares regression line is the unique line such that the sum of the squared vertical (*y*) distances between the data points and the line is the smallest possible.



Distances between the points and line are squared so all are positive values. This is done so that distances can be properly added .

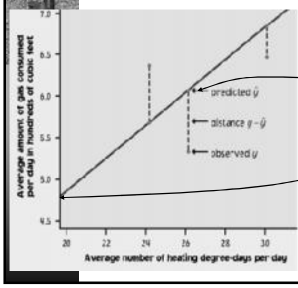
Facts about least-squares regression

1. The distinction between explanatory and response variables is essential in regression.
2. There is a close connection between correlation and the slope of the least-squares line.
3. The least-squares regression line always passes through the point $(\bar{x}, \bar{y})$.
4. The correlation ***r*** describes the strength of a straight-line relationship. The square of the correlation, r^2, **is the fraction of the variation in the values of *y that is explained by the least-squares regression of y on x.***

Properties

The least-squares regression line can be shown to have this equation:

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



$\hat{y}$ is the predicted y value (y hat)

b_1 is the slope

b_0 is the y-intercept

"b_0" is in units of y

"b_1" is in units of y/units of x

Regression Line as Mathematical Models

- Suppose ***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

$$y \quad b_0 \quad b_1 x$$

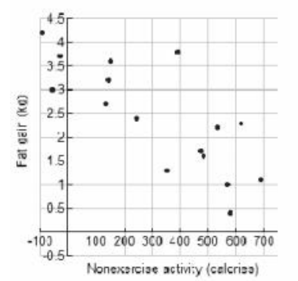
- In this equation, b_1 is the ***slope***, the amount by which ***y*** changes when ***x*** increases by one unit. The number b_0 is the ***y-intercept***, the value of ***y*** when ***x* = 0.**

Example : Does fidgeting keep you slim?

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

Example : Does fidgeting keep you slim?

L1	L2
94	4.2
-57	3
-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
580	0.4
020	2.3
690	1.1



Example : Does fidgeting keep you slim?

L1	L2
94	4.2
-57	3
-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
580	0.4
020	2.3
690	1.1

Two-Variable Statistics

$\bar{x} = 324.75$; $\bar{y} = 2.3875$

$\Sigma x = 5196.$; $\Sigma x^2 = 2.68321e6$

$Sx = 257.657$

$\sigma x = 249.475$

$n = 16.$

$\Sigma y = 38.2$; $\Sigma y^2 = 110.66$

$Sy = 1.13893$

$\sigma y = 1.10277$

$\Sigma xy = 8978.4$

$minX = -94.$; $maxX = 690.$

$minY = .4$; $maxY = 4.2$

Example : Does fidgeting keep you slim?

L1	L2
94	4.2
-57	3
-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
580	0.4
020	2.3
690	1.1

Linear Regression (a+bx)

regEQ(x) = 3.50512 + -.003441x

a = 3.50512

b = -.003441

r = -.778556

fat gain b_0 b_1***(NEA change)***

Equation of the Least-Squares Regression Line

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

Slope

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

Intercept

$$b_0 = \bar{y} - b_1\bar{x}$$

How to:

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

First we calculate the ***slope of the line***, ***b*,** from statistics we already know:

- ***r*** is the correlation
- **s_y** is the standard deviation of the response variable *y*
- **s_x** is the standard deviation of the explanatory variable *x*

Once we know ***b_1***, the slope, we can calculate ***b_0*,** the ***y*-intercept**:

$$b_0 = \bar{y} - b_1\bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the *x* and *y* variables

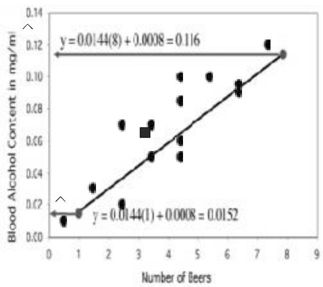
This means that we don't have to calculate a lot of squared distances to find the least-squares regression line for a data set. We can instead rely on the equation.

*But typically, we use a **2-var stats calculator** or a stats software.*

The equation completely describes the regression line.

To plot the regression line, you only need to plug two *x* values into the equation, get *y*, and draw the line that goes through those two points.

Hint: The regression line always passes through the mean of x and y.



The points you use for drawing the regression line are derived from the equation.

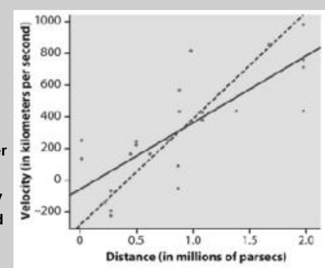
They are NOT points from your sample data (except by pure coincidence).

The distinction between explanatory and response variables is crucial in regression. **If you exchange *y* for *x* in calculating the regression line, you will get the wrong line**.

Regression examines the distance of all points from the line <u>in the y direction only</u>.

Data from the Hubble telescope about galaxies moving away from Earth:

These two lines are the two regression lines calculated either correctly (x = distance, y = velocity, solid line) or incorrectly (x = velocity, y = distance, dotted line).



Formulas for LSRL

- The LSRL is linear, so it follows the form ***y* = *mx* + *b***
  - In Statistics, we say $\hat{y} = b_0 + b_1 x$
  - In this context, $\hat{y}$ is called the ***predicted*** value
- There are specific formulas for the slope and intercept of the LSRL.
- To use these formulas, we need ***b_1***, ***S_x***, ***b_0***, and ***S_y***, plus the correlation constant ***r***.
  - We also need a clear indication of which is the explanatory variable; that will be the ***x*** data.
- Once you have calculated the 5 items above, here are the formulas:

$$\text{Slope:} \quad b_0 = r\frac{S_y}{S_x} \qquad \text{Intercept:} \quad b_0 = \bar{Y} - b_1\bar{X}$$

Example: *Classifying Fossils*

Femur:	38	56	59	64	74
Humerus:	41	63	70	72	84

***a)** Make a scatterplot. Do you think that all five specimens come from the same species?*

***b)** Find the correlation r step-by-step. That is, find the mean and standard deviation of the femur lengths and of the humerus lengths. Then find the five standardized values for each variable and use the formula for r.*

Finding a "*Best-fit*" Line

- Consider the archaeopteryx data from problem 3.13

Femur length:	**38**	**56**	**59**	**64**	**74**
Humerus length:	**41**	**63**	**70**	**72**	**84**

- Draw axes on graph paper with scales appropriate to these data and plot the points
  - Assume (unrealistically) that femur length is the explanatory variable and that humerus length is the response variable
- Use a straightedge to draw a straight line that appears to fit the data as well as possible
- Using two points on your line (not necessarily data points), find the equation of the line
  - Write the equation in terms of femur and humerus, not *x* and *y*

Making Predictions from the Line

- My best-fit line had this equation:

  humerus length **= 1.197 (*femur*) – 3.660**
  - Note that I have rounded to the nearest tenth, which is one more decimal place than the source data.
  - Note also that my equation is IN CONTEXT. I don't use "*x*" or "*y*".
- Based on my equation, how long would you expect the humerus to be of a specimen with a femur length of 47 cm?

  humerus length **= 1.197 ò 47 – 3.660 = 52.599 cm**
- Based on YOUR equation, how long should the humerus be?
- ***Caution:*** Predictions are only valid for ***x*** values WITHIN the range of actual ***x*** data

Finding the LSRL for Archaeopteryx

- Yesterday we ran 2-Var Stats on the archaeopteryx data and calculated that r = 0.994.
- Run 2-Var Stats again on the lists FEMUR and HUMER to be sure you have the correct ☑ etc.
- Use the formula to find the slope of the LSRL (round to .001)

  $$b_1 = .994\left(\frac{15.890}{13.198}\right) = 1.197$$
- Use the formula to find the intercept of the LSRL

  $$b_0 = 66 - 1.197 \times 58.2 = -3.665$$
- Write the LSRL equation *in context.*

  humerus = -3.665 + 1.197 × femur
- Predict humerus length for a femur length of 47 cm

  $$\text{humerus} = -3.665 + 1.197 \times 47 = 52.6 \text{ cm}$$

The Meaning of Slope

- In a simple algebraic function like ***y*** = 2***x*** + 17, what is the real meaning of the slope?
  - ***For every increase in x of 1 unit, y increases by 2***
- In the context of the archaeopteryx problem, what is the meaning of the slope?
  - ***For every increase of femur length by 1 cm the predicted humerus length increases by 1.197 cm.***
- In the function *y* = 2*x* + 17, what is the meaning of the *y* intercept?
  - ***It is the value y takes on when x = 0***
- In the context of the archaeopteryx problem, what is the meaning of the intercept?
  - ***When the femur length = 0, the humerus length = -3.665 (huh!)***
- What's wrong with that answer?
  - ***It makes no sense because the femur value is outside of the range of the data used to get the LSRL equation.***

BEWARE !!!

Not all calculators and software use the same convention:

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

Some use instead:

$$\hat{y} = ax + b$$

Make sure you know what YOUR calculator gives you for a and b before you answer homework or exam questions.