Simple Linear Regression: Bivariate Data

Many statistical investigations center on bivariate data. Bivariate data is simply the observed values on two distinct/different population variables pertaining to unit/individual in the population of interest. Basically, it is when you sample two variables from one population.

For example when observing a selected population subset like a classroom, collect two sets of data instead of one i.e." please record your height **and** your weight."

Some examples of data that are bivariate in nature:

1. A Statistics 323 student's midterm exam mark and final exam mark.
2. The number of years of post-secondary education an individual has and their annual income.
3. A year's inflation and interest rate.
4. The average price of oil in a month and the average price of the Canadian dollar (relative to the U.S. dollar).
5. *The temperature in Celsius and in Fahrenheit*

From a notation standpoint, the two variables of interest are represented by:

$$X_i's \text{ the observed value of Variable } X \text{ from subject } i, i = 1,2,\cdots,n.$$
$$Y_i's \text{ the observed value of Variable } Y \text{ from subject } i, i = 1,2,\cdots,n.$$

Or for convenience and organization perhaps as ordered pairs from a Cartesian Plane.

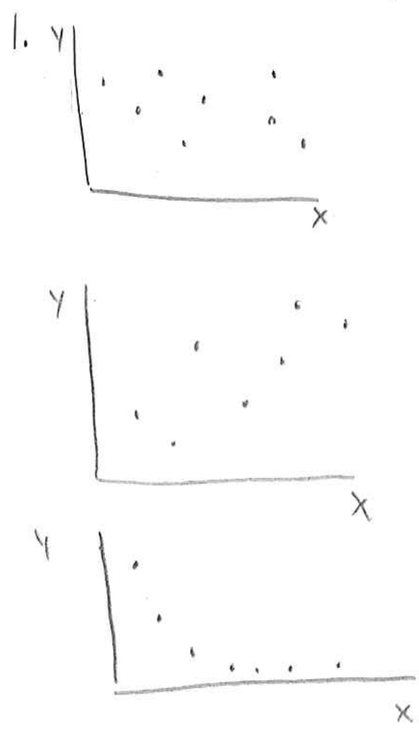
$$(X_i, Y_i) \in \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$$

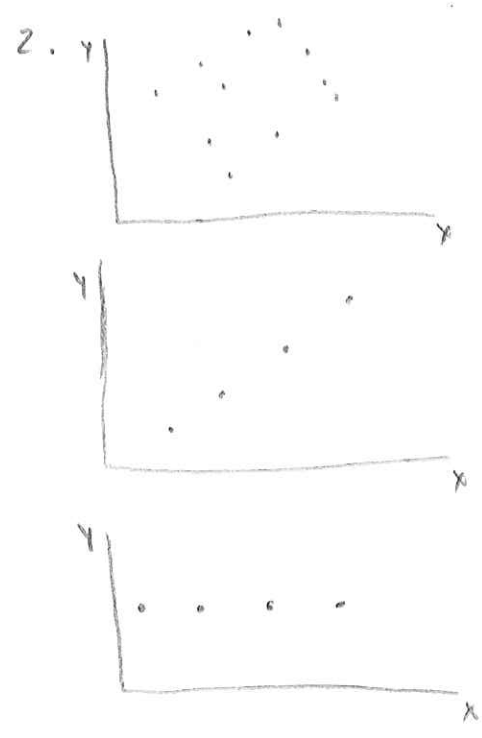
What is the motivation for studying (quanitative) bivariate data? It is all about relationships. When an experimental study or random sampling method produces data on two different variables, there are three research questions that are posed.

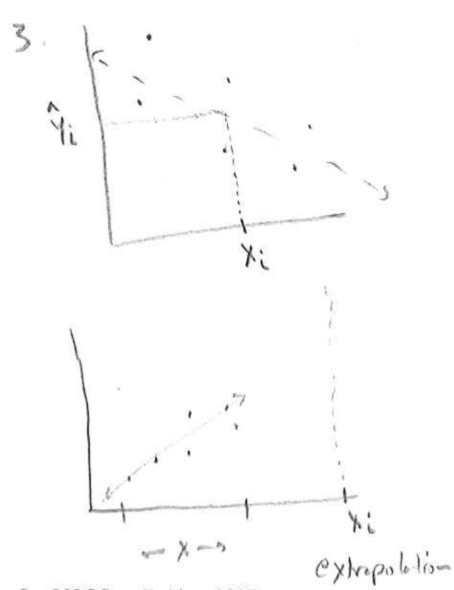
1. **Is there a relationship between the two variables**? If there is a relationship, what is the direction of the relationship? Is the relationship positive? negative (or inverse)? Does the relationship seem to be linear? non-linear?

2. **If a relationship exists between X and Y , how strong is this relationship?** Is such a relationship subtle, or strong?

3. **If the relationship between X and Y is strong, can the existing relationship be used to predict what will happen in the future?** That is, can we create a mathematical function, $y = f(x)$, which will predict one's final exam mark once the midterm exam mark has been applied to this function?







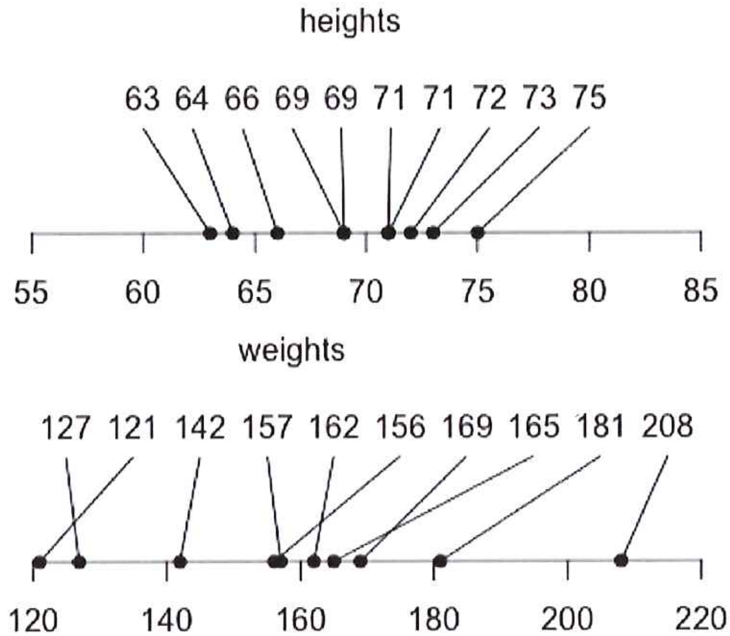
Let's consider 10 students who each provided their height and weight.

	height	weight
1	63	127
2	64	121
3	66	142
4	69	157
5	69	162
6	71	156
7	71	169
8	72	165
9	73	181
10	75	208

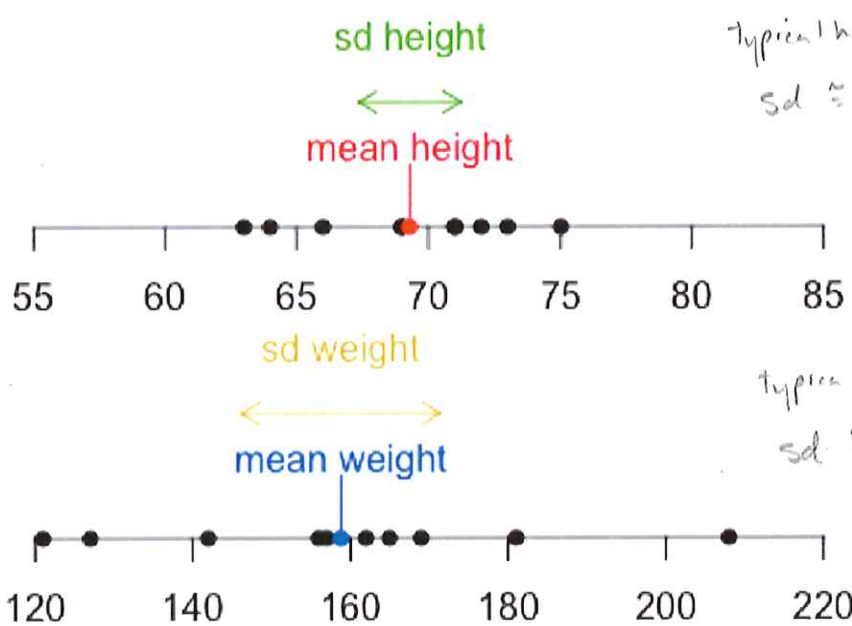
I think the natural question associated with this type of data collection is apparent. Is the height and weight of a student related?

Before studying bivariate data we studied univariate data, so let's look at the heights and weights as a separate variables and then consider how to determine possible relationships.

Observe the number lines given:



While studying univariate data we were also very interested in the expect values and spreads of the data.



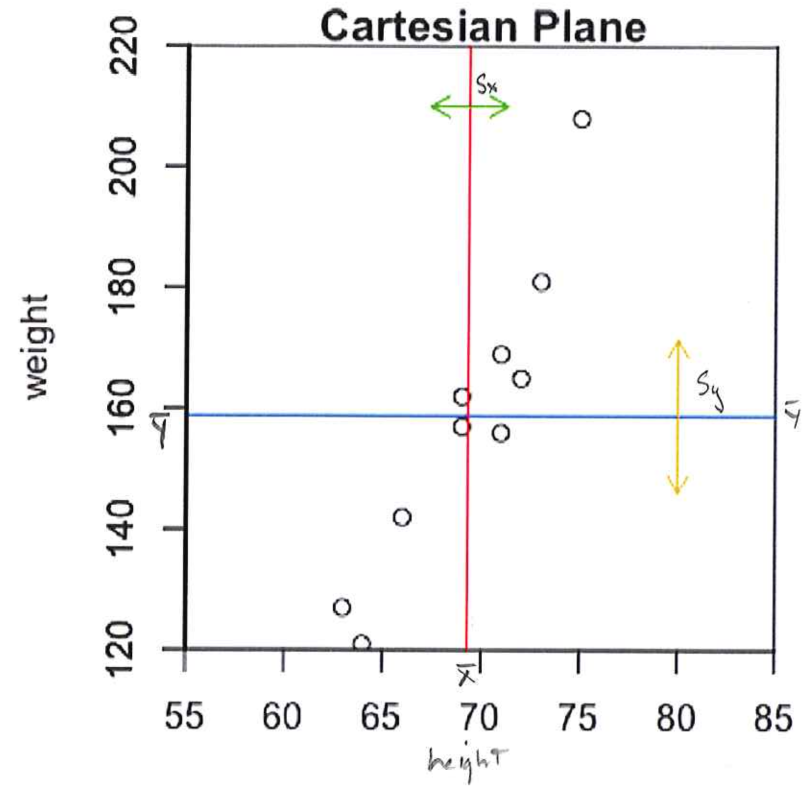
Typical height ≈ 69.3
sd ≈ 3.92

Typical weight ≈ 158.8
sd ≈ 25.38

So for a "new" person you would say...

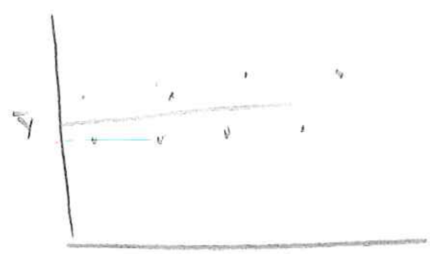
what if we first knew the "new" person's height...

But what if we consider additional information that these points are "paired," each X has a corresponding Y.



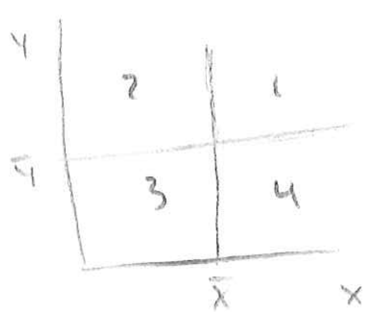
Can you see a dependence/relationship between X and Y? What would the plane look like if there were no relationship/independence present?

No relationship



Consider $\bar{x}, \bar{y}$
→ creating quadrants!

No matter the x we suggest
$\bar{y}$!



relationship if: bulk of points in quad 1 & 3
or if in quad 2 & 4
if even or if even in 1 & 2 or 3 & 4
→ No relationship

Quantifying the "linear trend"

Recall from previous course(s):

$$Cov(X,Y) = \sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - E[X]E[Y]$$

Sample covariance can be found in a similar way:

$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

What does it do? Measures dependence…

→ measures if "Particular" Points are Both far (x from $\bar{x}$, y from $\bar{y}$) if small −'s if Big +'s Product of Same signs ⇒ + with "Big" magnitudes!

→ or if X & y are unrelated you will be getting values with little sign change i.e. near 0!

Pearson's Correlation is a scaled (by the standard deviations) version of the covariance:

$$Cor(X,Y) = \rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}, \quad \boxed{-1 < \rho < 1}$$

→ Because of Scaling. Makes Apples and oranges Comparable!

Sample correlation:

$$r = \frac{S_{XY}}{S_XS_Y} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})\big/_{n-1}}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\big/_{n-1}}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2\big/_{n-1}}} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}, \quad -1 < r < 1$$

Notation: $S_{XX} = S_X^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})}{n-1} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1} \Rightarrow S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}$ which we know as the sample standard deviation of X

What does it look like?

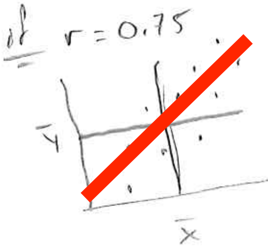
if $r \cong -1$

good x & y

very obvious "Straight" Line

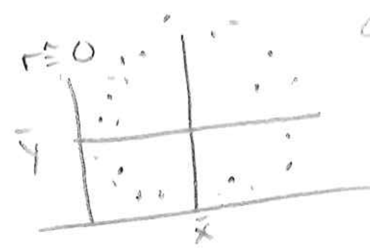
$\bar{x}$ & $\bar{y}$ should Be middle

if $r = 0.75$



less obvious "straight" Line

if $r \cong 0$



doesn't mean No pattern Just No straight line pattern!

Pearson's Correlation is a measure of the linear quality of the relationship between two variables