Applied Regression Analysis

Week 1

1. Review of basic statistical concepts
   - Central tendency and variability
   - Sampling distributions
   - Bias
   - Confidence intervals
   - *p*-values
2. Regression and correlation
3. Introduction to STATA

Stanley Lemeshow, Professor of Biostatistics
College of Public Health, The Ohio State University

THE OHIO STATE UNIVERSITY

Topics to be discussed:

- Measures of Central Tendency
- Measures of Dispersion
- Degrees of Freedom
- Population Parameters vs Sample Statistics
- Sampling Distributions
    - Expected Values, Standard Errors
  - Unbiased vs Biased Estimators
  - Confidence Intervals
  - Hypothesis Testing
    - *p*-values

Measures of Central Tendency

The Population Mean

Given a set of N values, $X_1, X_2, \ldots, X_N$,
the population mean is computed as:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

The Sample Mean

Given a set of n values, $x_1, x_2, \ldots, x_n$, their mean, denoted by $\bar{x}$, is defined by their sum divided by the number of observations, n:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

e.g.

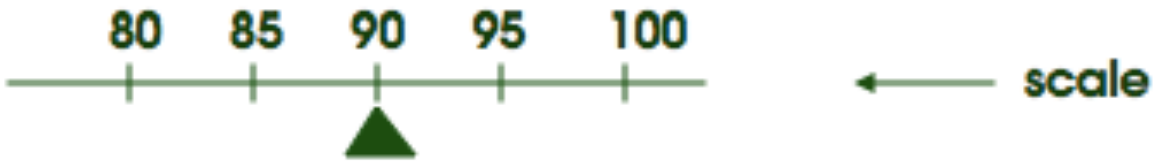
Suppose the scores on five exams are as follows:

$$X_1: \quad 90$$
$$X_2: \quad 80$$
$$X_3: \quad 95$$
$$X_4: \quad 85$$
$$X_5: \quad 100$$

What is the average?

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{5} = \frac{90+80+95+85+100}{5} = \frac{450}{5} = 90$$

The mean is the "center of gravity" of the observations.



note also: $\sum (x_i - \bar{x}) = 0$

The Variance and Standard Deviation

The value $\sum_{i=1}^{N}\left(x_i - \mu\right)$ = sum of deviations about the sample mean. This ≡ 0.

The value $\sum_{i=1}^{N}\left(x_i - \mu\right)/N$ = mean deviation. This also ≡ 0.

We can avoid this difficulty by taking absolute values of the deviations.

- i.e., we can ignore the algebraic signs. Then, we could use the "mean absolute deviation".

$$\frac{1}{N}\sum_{i=1}^{N}\left|x_i - \mu\right|$$

where $\left|x_i - \mu\right|$ is read "the absolute value of $x_i - \mu$"

We don't use the mean absolute deviation because working with absolute values discourages further mathematical or theoretical treatment.

We also avoided the difficulty by squaring each deviation since, in the process, all negative signs disappear.

$$\frac{1}{N}\sum_{i=1}^{N}\left(x_i-\mu\right)^2 = VARIANCE$$

This provides a MEASURE OF DISPERSION

The value $\sum_{i=1}^{N}\left(x_i-\mu\right)^2$ = sum of square deviations about the sample mean. (This is ≥ 0.)

The value $\sum_{i=1}^{N}\left(x_i-\mu\right)^2 / N$ = mean square deviation.

(i) if observed values are identical,

$$\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 = 0$$

(ii) if observed values are close together,

$$\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 \text{ will be small}$$

(iii) if observed values scatter over a wide range,

$$\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 \text{ will be correspondingly large.}$$

Computations from sample

x_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
12	12	0	0
6	12	–6	36
15	12	3	9
3	12	–9	81
12	12	0	0
6	12	–6	31
21	12	9	81
15	12	3	9
18	12	6	36
12	12	0	0
$\sum x_i = 120$		☺	228

$$\bar{x} = \frac{\sum x_i}{10} = 12$$

check on work
if not = 0, there is an error

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 288$$

and

$$\text{the variance } = s^2 = \frac{\sum_{i=1}^{10} (x_i - \bar{x})^2}{10 - 1} = 32$$

Note that the "sample variance" is defined as:

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}{n-1}$$

where "*n*" is the number of observations in the "sample".

In the previous discussion *N* was the number of observations in a "population".

Q. Why do we use "*n* - 1" in the denominator of the sample variance instead of "*n*"?

A. (i) if we selected many samples from a population, and computed a variance for each, then the average of the sample variance will not equal the population variance, σ^2, if *n* is used in the denominator.

In fact, the average of these sample variances will be too small.

i.e., $\sum_{i-1}^{k} s_i^2 \Big/ k < \sigma^2$, where k = # samples drawn.

Alternatively, if we make each s^2 larger (i.e., by using $n-1$ in the denominator), we would find that $\sum_{i=1}^{k} s_i^2 \Big/ k = \sigma^2$.

(ii) $n-1$ represents "degrees of freedom". To calculate the variance, we first calculated $\bar{x}$. But, given $\bar{x}$, we have lost a degree of freedom since if you tell me $\bar{x}$, only $n-1$ of the n observations are free to vary.

e.g., suppose I tell you that $\bar{x} = 90$ and $n = 5$. Then,

$x_1 = 80$

$x_2 = 85$

$x_3 = 90$

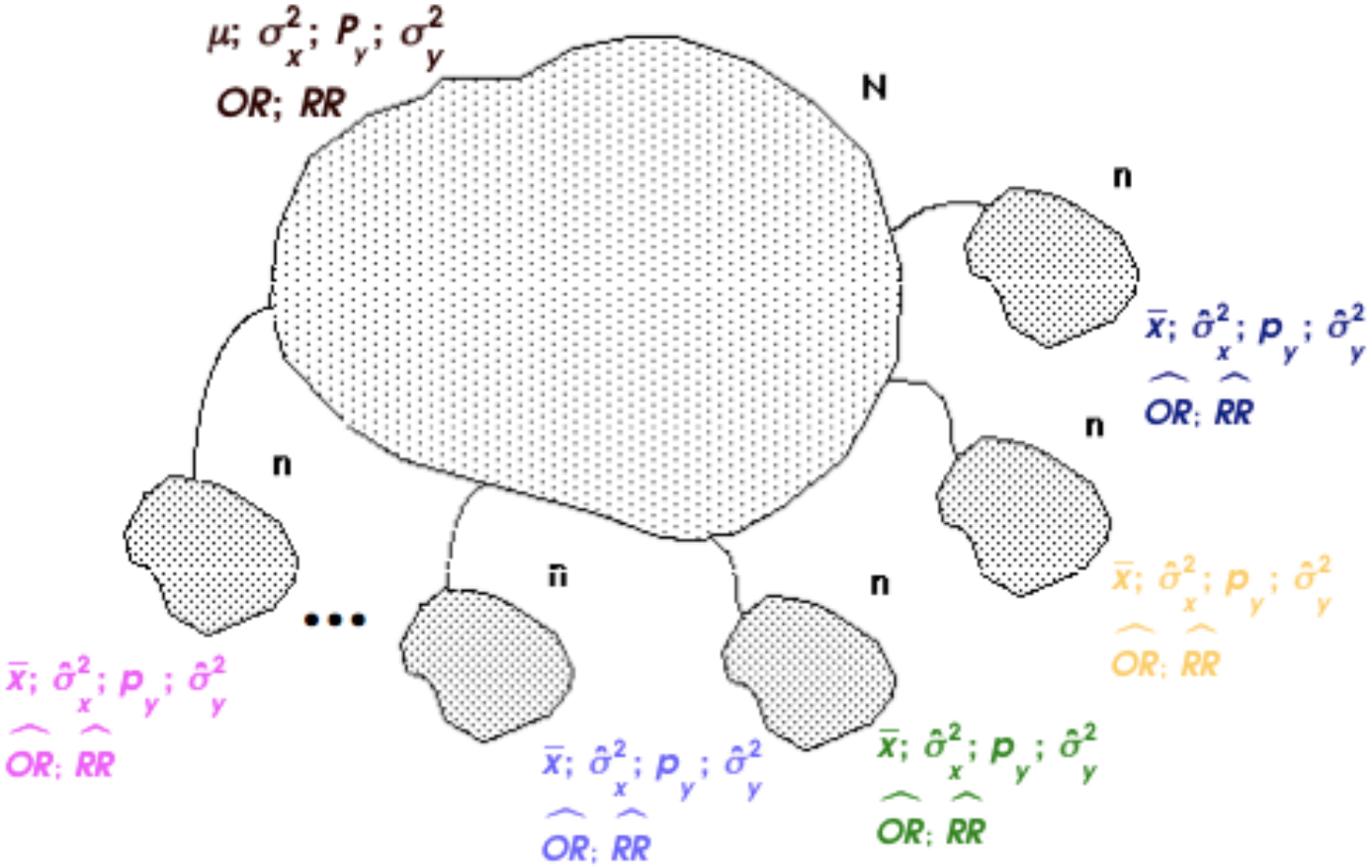
$x_4 = 95$

$x_5 =$ ← must = 100

since $\bar{x} = \frac{\sum x_i}{5} = 90 \Rightarrow \sum x_i = 450$

Sampling Distributions

We consider all possible samples that can be generated using a particular sampling plan.



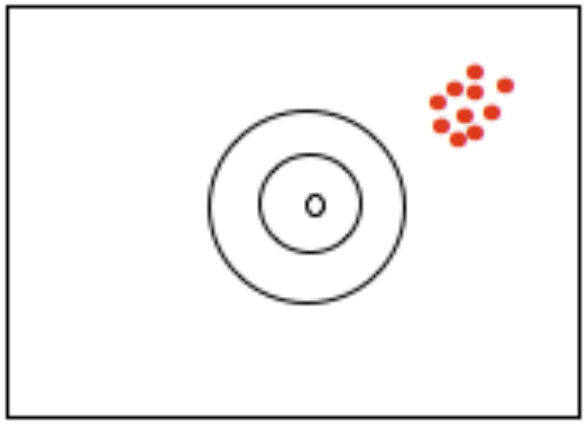
We select 10,000,000 samples from this population, each time estimating the population parameters.

Sample	$\bar{x}_t$
1	$\bar{x}_1$
2	$\bar{x}_2$
3	x_3
⋮	⋮
t	$\bar{x}_t$
⋮	⋮
10,000,000	$\bar{x}_{10,000,000}$

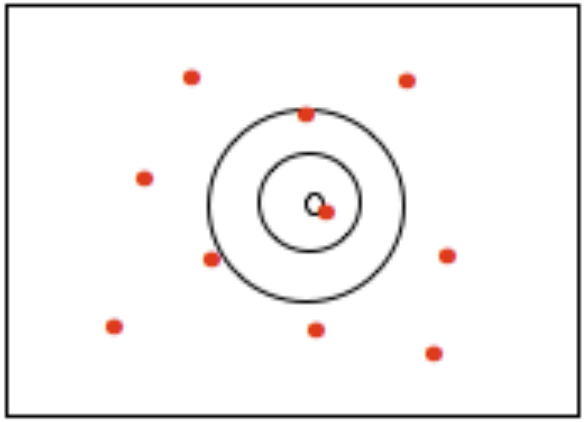
This could be any statistic

If, on average, the sample estimate is equal to the population parameter, then the estimate is said to be "unbiased".

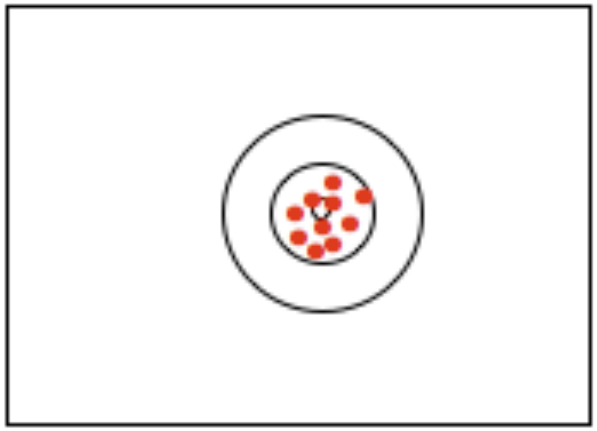
We not only look for estimators that are unbiased, but we also want estimators that have minimum variance.



high bias
low variance



low bias
high variance



low bias
low variance

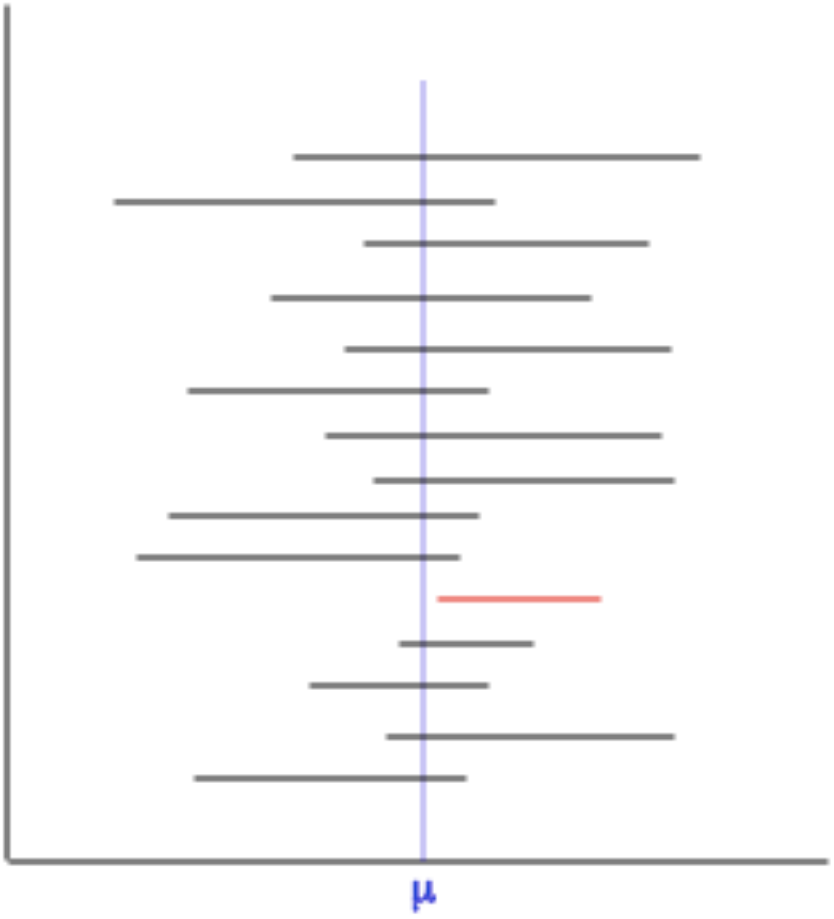
Confidence intervals:

Take the form:

$$\bar{x} - 1.96\widehat{SE}\left(\bar{x}\right) \le \mu \le \bar{x} + 1.96\widehat{SE}\left(\bar{x}\right)$$

some multiplier

could be any parameter



def: 95% Confidence Interval:

Upon repeated sampling, 95% of intervals constructed in the same way will "cover" the true population parameter.

note: Once an interval is specified, it is either right or wrong

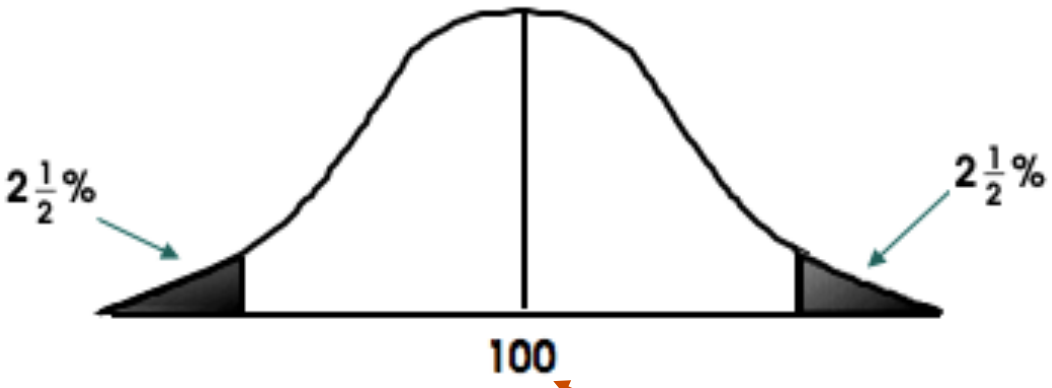
Hypothesis Testing:

$$H_0 : \mu = 100$$

$$H_a : \mu = 100$$

Statistician's role: Believe the null hypothesis until the evidence is so strong that the only reasonable response is to reject it.

Sampling distribution of $\bar{x}$:



The "Central Limit Theorem" tells us that the sampling distribution of the sample mean is normal...irrespective of the distribution of the original data.

If the null hypothesis is true, then the mean of the sampling distribution of the sample mean will = 100

If the sample mean is far from the hypothesized mean, then chances are that the hypothesis is false.

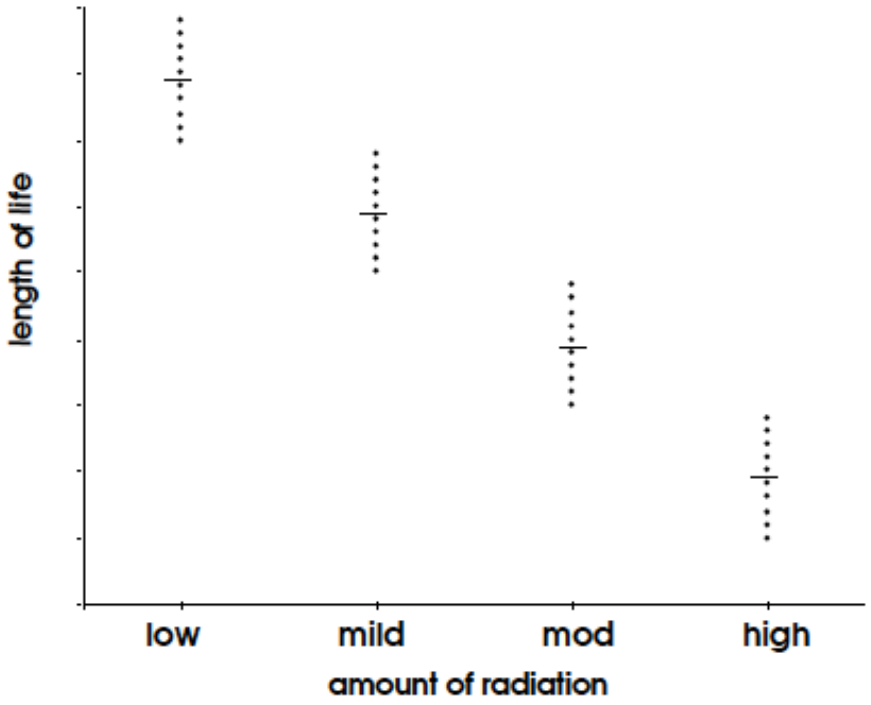
Type I error: Probability of rejecting the H_0 when H_0 is true.
Type II error: Probability of failing to reject H_0 when H_0 is false.

***p*-value: Probability of observing a result as extreme or more extreme than the one observed given H_0 is true.**

- **a small *p*-value (e.g., $p \leq .05$) suggests that the results of a study are "statistically significant"**
  - **i.e., null hypothesis is probably false**
- **a large *p*-value (e.g., $p > .05$) suggests that the results of a study are not significant**
  - **i.e., there is no evidence to reject the null hypothesis**

REGRESSION: considers the frequency distribution of one variable when another is held fixed at each of several levels

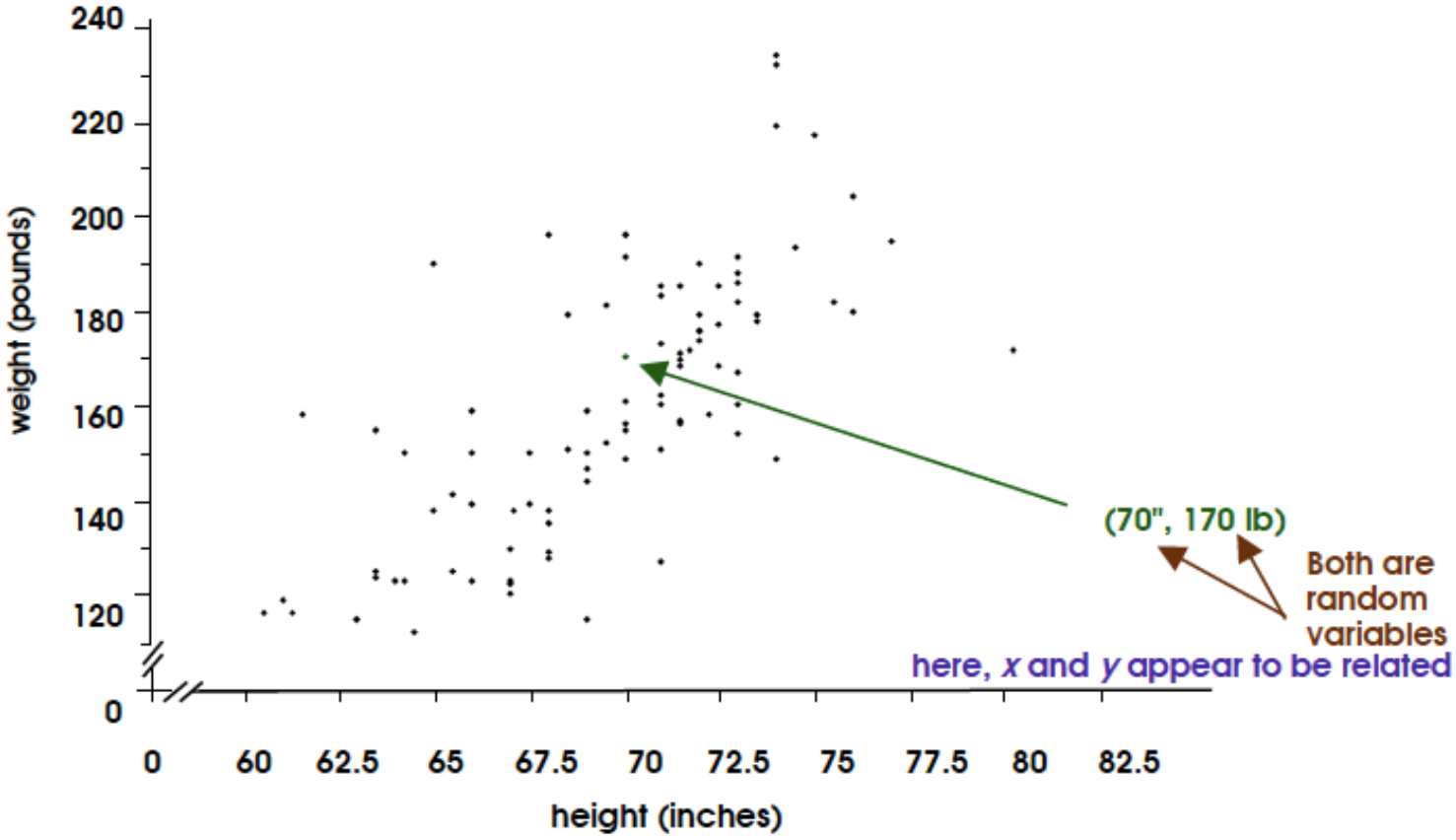
e.g. Suppose we have 40 mice and we expose them to varying amounts of radiation (4 levels)



In this problem, the variability in length of life (y = dependent variable) is studied with respect to particular levels of the amount of radiation (x = independent variable).

CORRELATION: considers the association of two random variables

e.g., Suppose we're studying height and weight

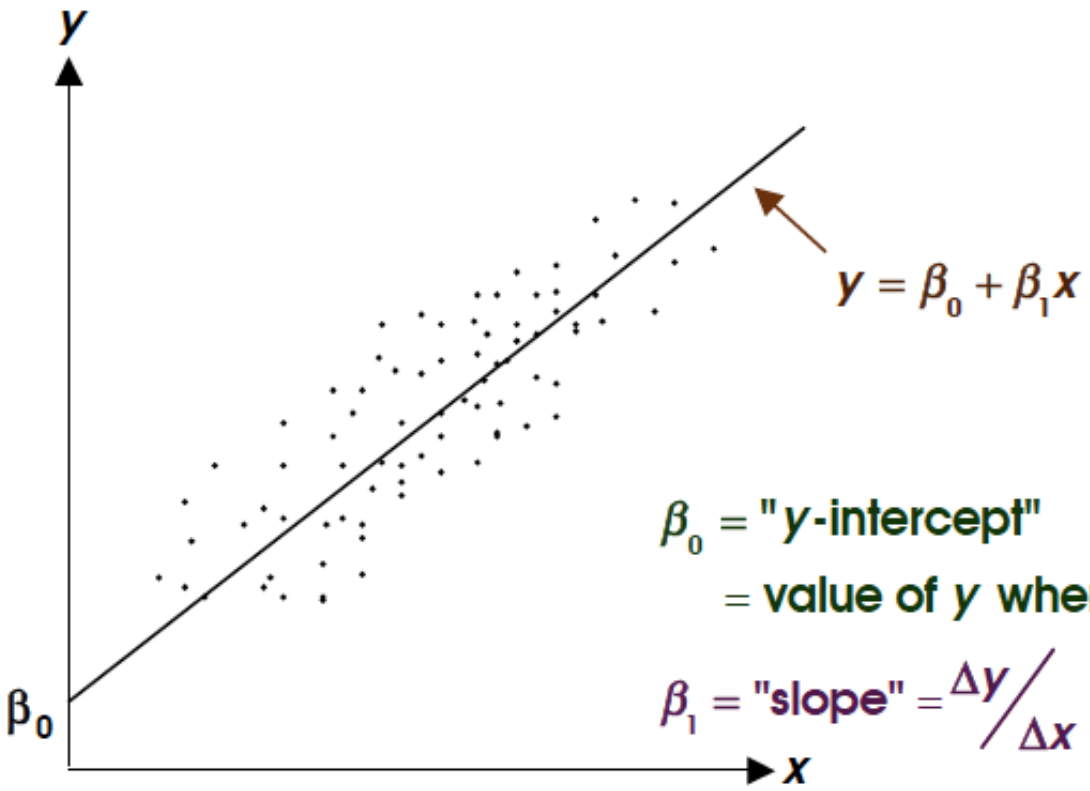


We will use similar techniques for the two types of problems but we should keep the distinction in mind.

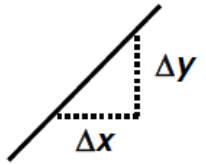
For both the correlation and regression problems it makes sense to describe the relationship between the two variables by fitting a line to the points.

- This line exhibits the "trend" in the data.

Let us review some elementary algebra:

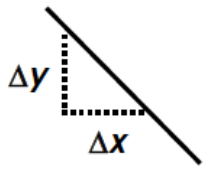


e.g.,



Slope > 0

Slope = 0



Slope < 0