Module 4B
Supervised Machine Learning

Different flavors of REGRESSION and General Linear Models

Finding **a**:

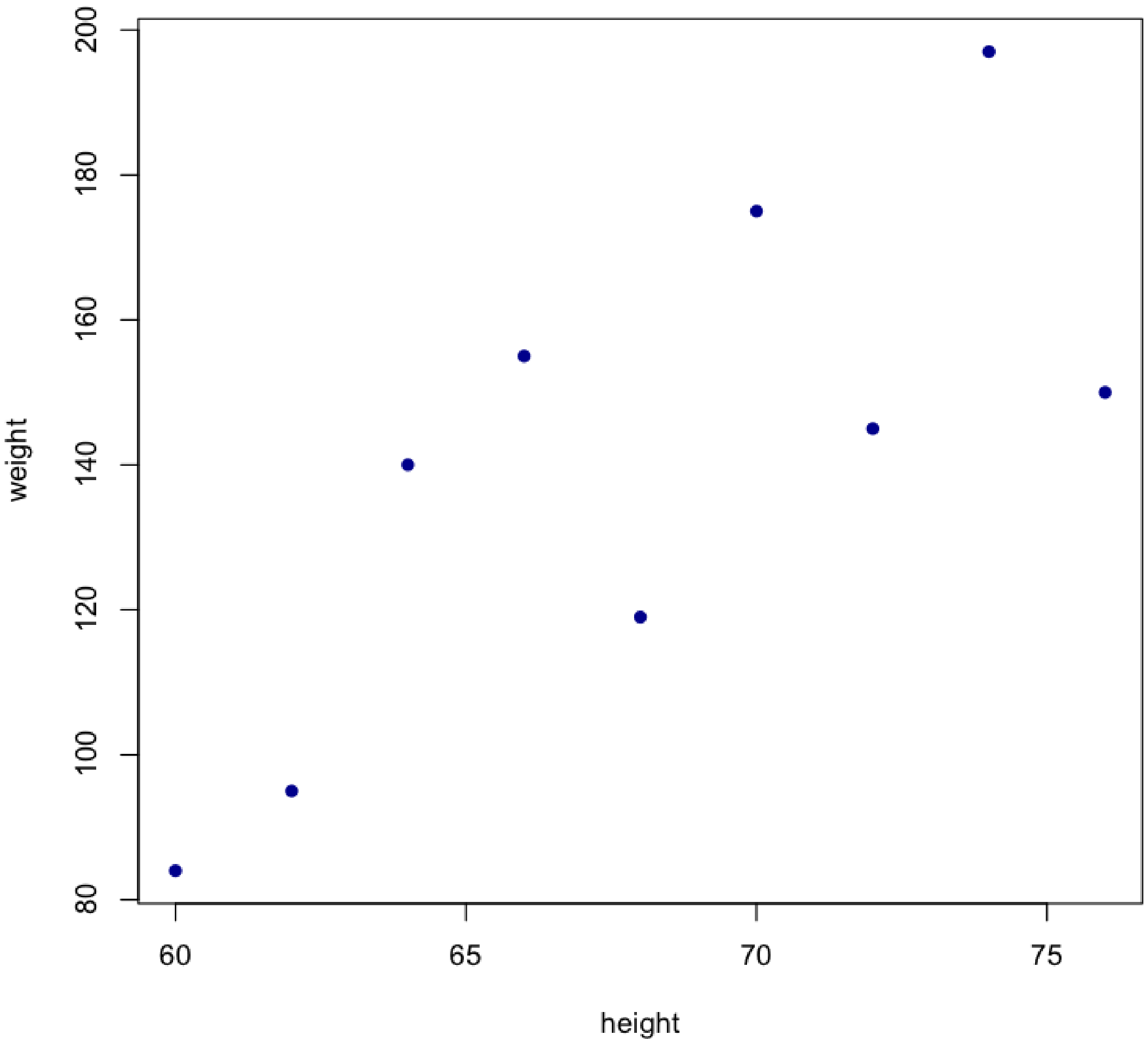
$$\boxed{\bar{Y} = a + b\bar{X}}$$

OR

$$\boxed{a = \bar{Y} - b\bar{X}}$$

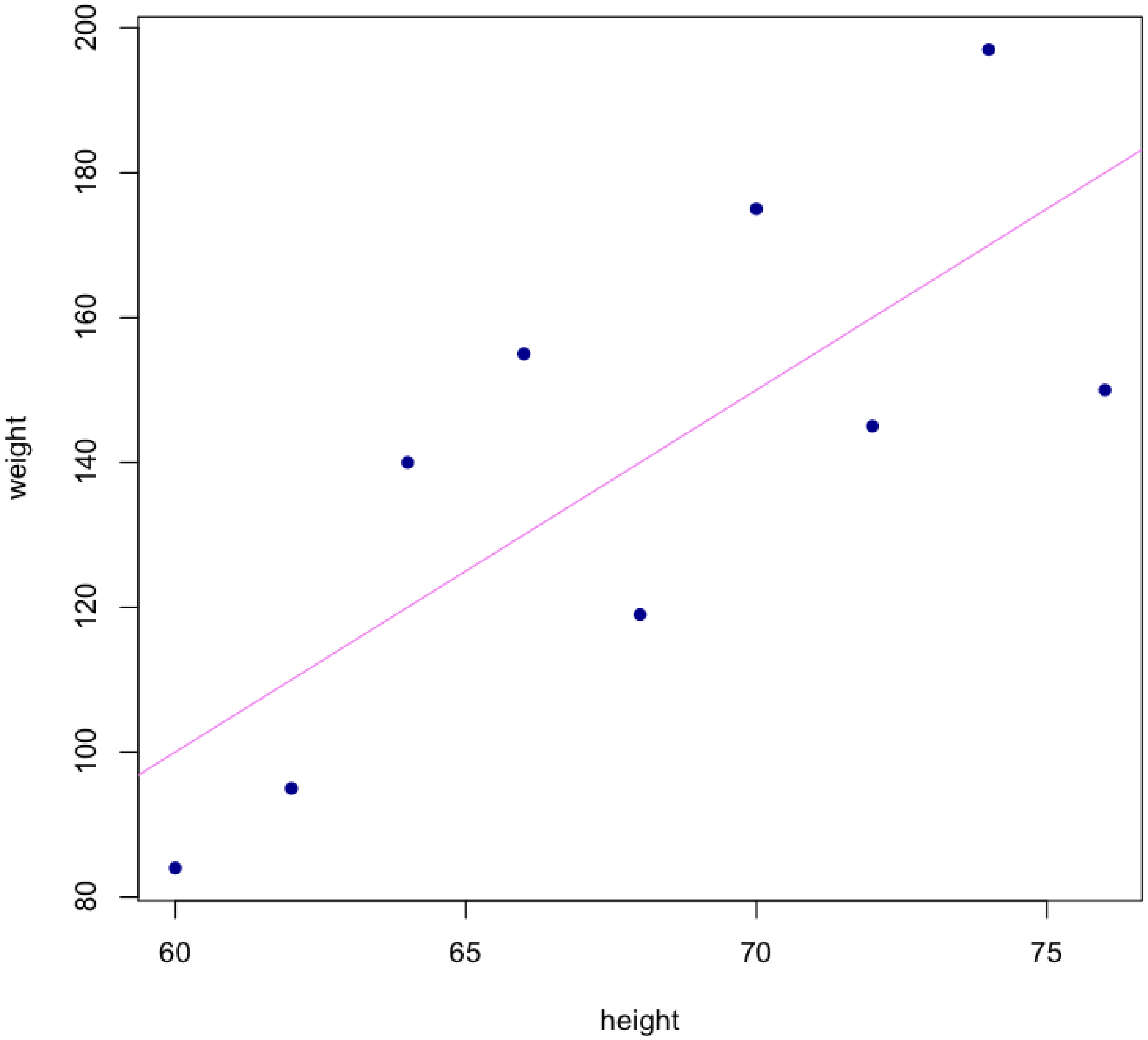
Example: Predicted weight for someone who is 65 inches tall?

Height	Weight
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150



Example: Predicted weight for someone who is 65 inches tall?

Height	Weight
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150



Height Weight data:

$\sum X = 612$

$\sum Y = 1260$

$n=9$

$\sum X^2 = 41856$

$\sum Y^2 = 186826$

$\sum (XY) = 86880$

$$\bar{Y} = 140$$

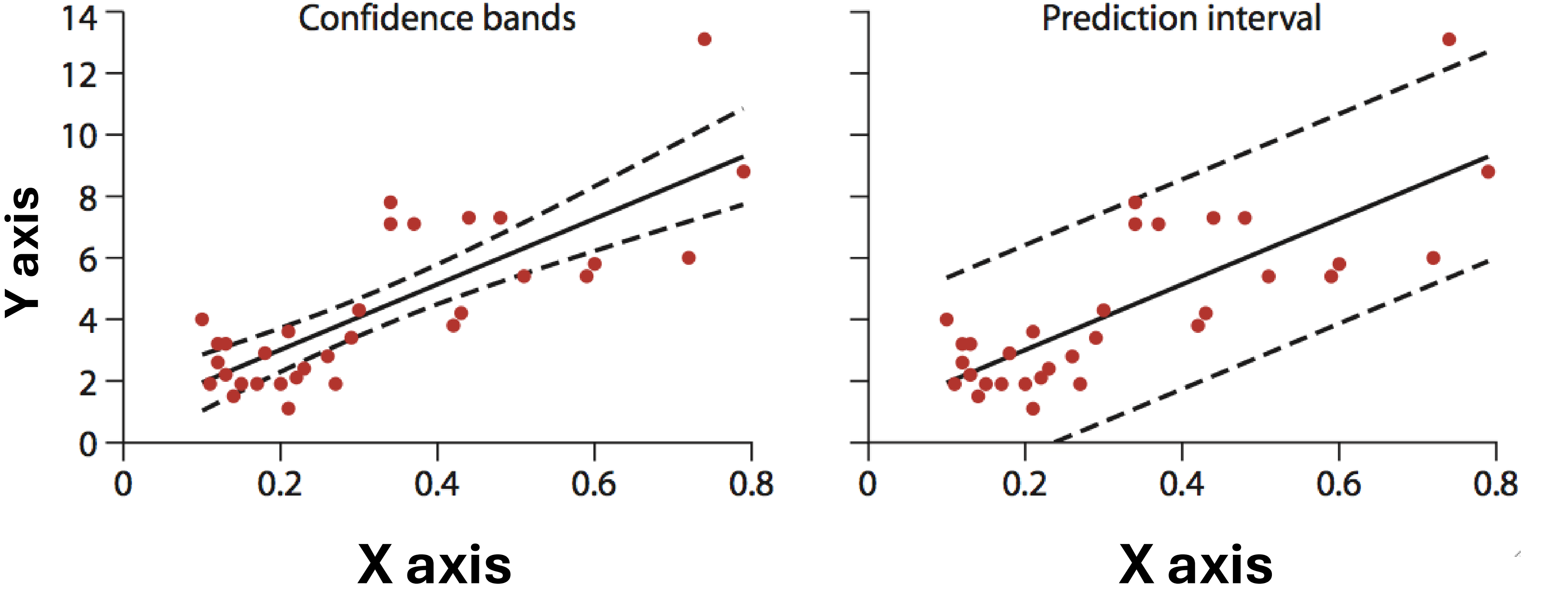
$$\bar{X} = 68$$

$b = 5$

$a = -200$

$$\boxed{\hat{Y} = -200 + 5X}$$

Prediction confidence:



Prediction confidence:

The purpose of regression is to **predict**. There are two types of prediction:

1. $\overline{Y}$ for a given X. → Confidence bands (related to Confidence Interval)

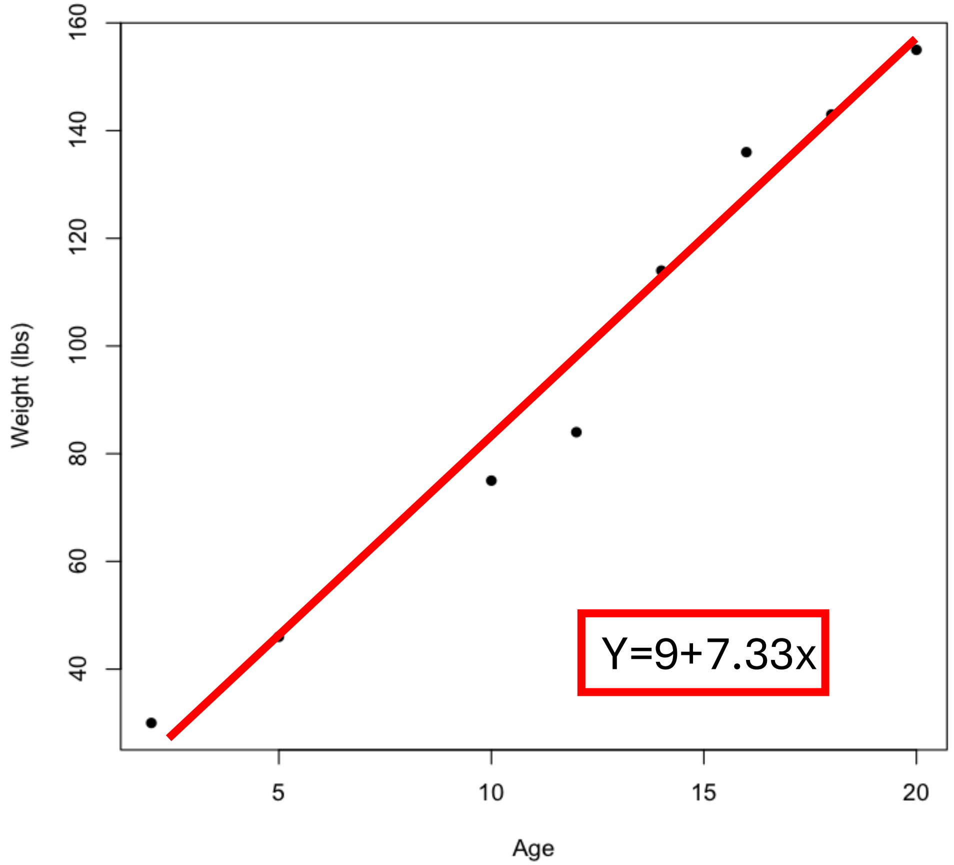
2. Single Y for a given X → Prediction bands

Both will generate $\hat{Y}$ with the same value, but the prediction of a single Y point will have a lower precision.

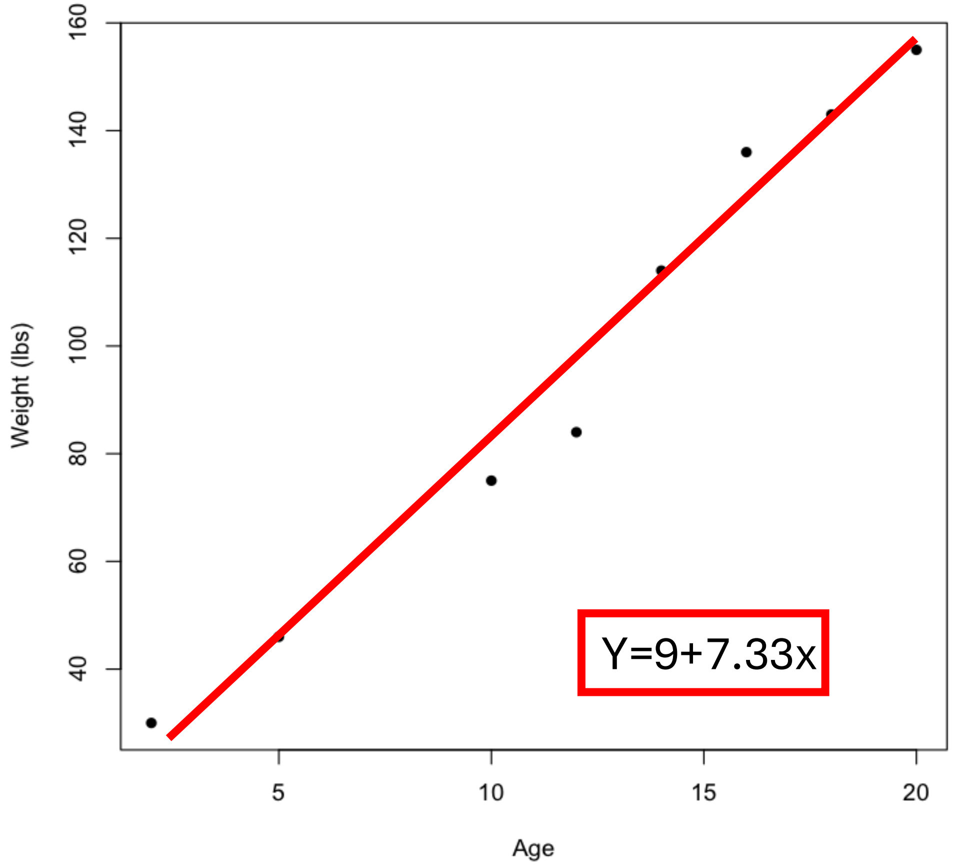
Caution! Do not extrapolate beyond the range of the data

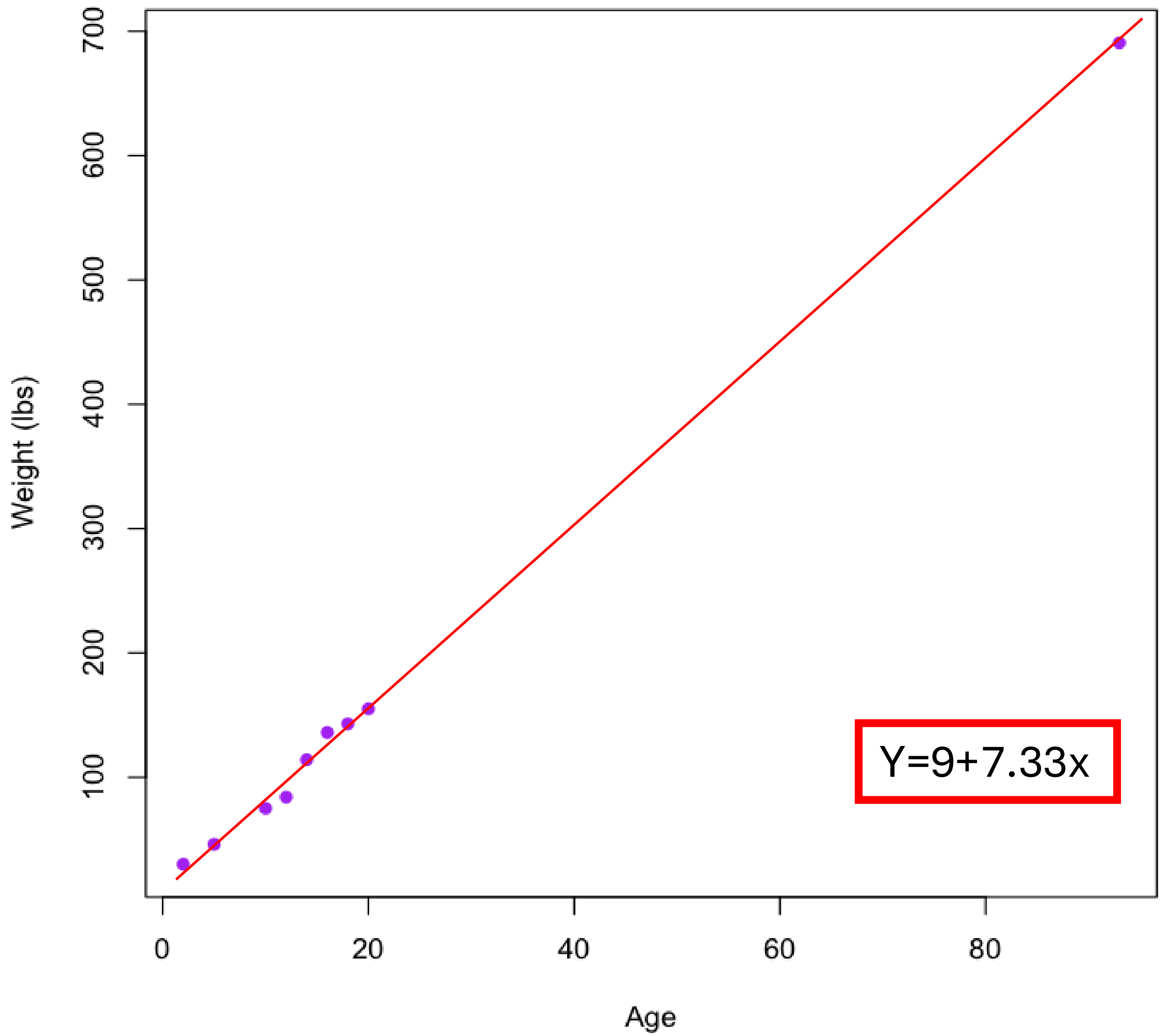
Age	Weight (lbs)	Time to run one mile	Bench Press (lbs)
2	30		
5	46		
10	75		
12	84	5:40	
14	114	5:05	
16	136	4:40	160
18	143	4:35	180
20	155	4:30	

Measurements taken over the course of an individual's life



$Y=9+7.33x$


Weight (lbs)
40
60
80
100
120
140
160
5
10
15
20
Age
Y=9+7.33x


Weight (lbs)
100
200
300
400
500
600
700
0
20
40
60
80
Age
Y=9+7.33x

Regression to the mean:

- Francis Galton invented the term to describe the observation that tall fathers had sons of average height

- He developed "regression analysis" to study this phenomenon of "regression towards mediocrity"

- results when two variables have correlation < 1
    - Individuals who are far from the mean for one of the measurements will, on average, lie closer to the mean for the other measurement

Regression fallacy:

- Tricky concept:
  - each individual has a **true** value, but the sampled value varies with time
    - the subset who scored highest on the first round included individuals who had higher values then their usual ‘true’ value
    - the second measurement captured these individuals when they happened to be closer to their own personal normal values

- failure to consider “regression towards the mean” when interpreting the results of **observational studies**

- can be a large problem when dealing with **sick** people - they are the tail of the distribution, and they might appear to improve even if the treatment applied has no real effect

Regression to the mean:

A VERY old concept:

> “You know, few sons turn out to be like their fathers;
> Most turn out worse, a few better.”
>
> (Athena speaking to Telemachus)
> - Homer, The Odyssey

Regression fallacy: Rolling a die

Student	First	Second	Second roll lower?
1	4	5	no
2	4	3	yes
3	3	-	-
4	5	5	no
5	1	-	-
6	6	5	yes
7	5	2	yes
8	6	2	yes
9	3	-	-
10	2	-	-

Remaining students have a mean value of 5 (first roll) and 3.7 (second roll)

Testing hypotheses about slope:

1. H_0: $\beta = \beta_0$ (N.B. The null hypothesis is that Y cannot be predicted from X)

   H_A: $\beta \neq \beta_0$

2. Test statistic:

$$t = \frac{b - \beta_0}{SE_b}$$

$$SE_b = \sqrt{\frac{MS_{residual}}{\Box\,(X_i - \bar{X})^2}}$$

3. significance level; df=n-2

4. Reject or FTR and:

$$\boxed{b - t_{\alpha(2),n-2} SE_b < \beta < b + t_{\alpha(2),n-2} SE_b}$$

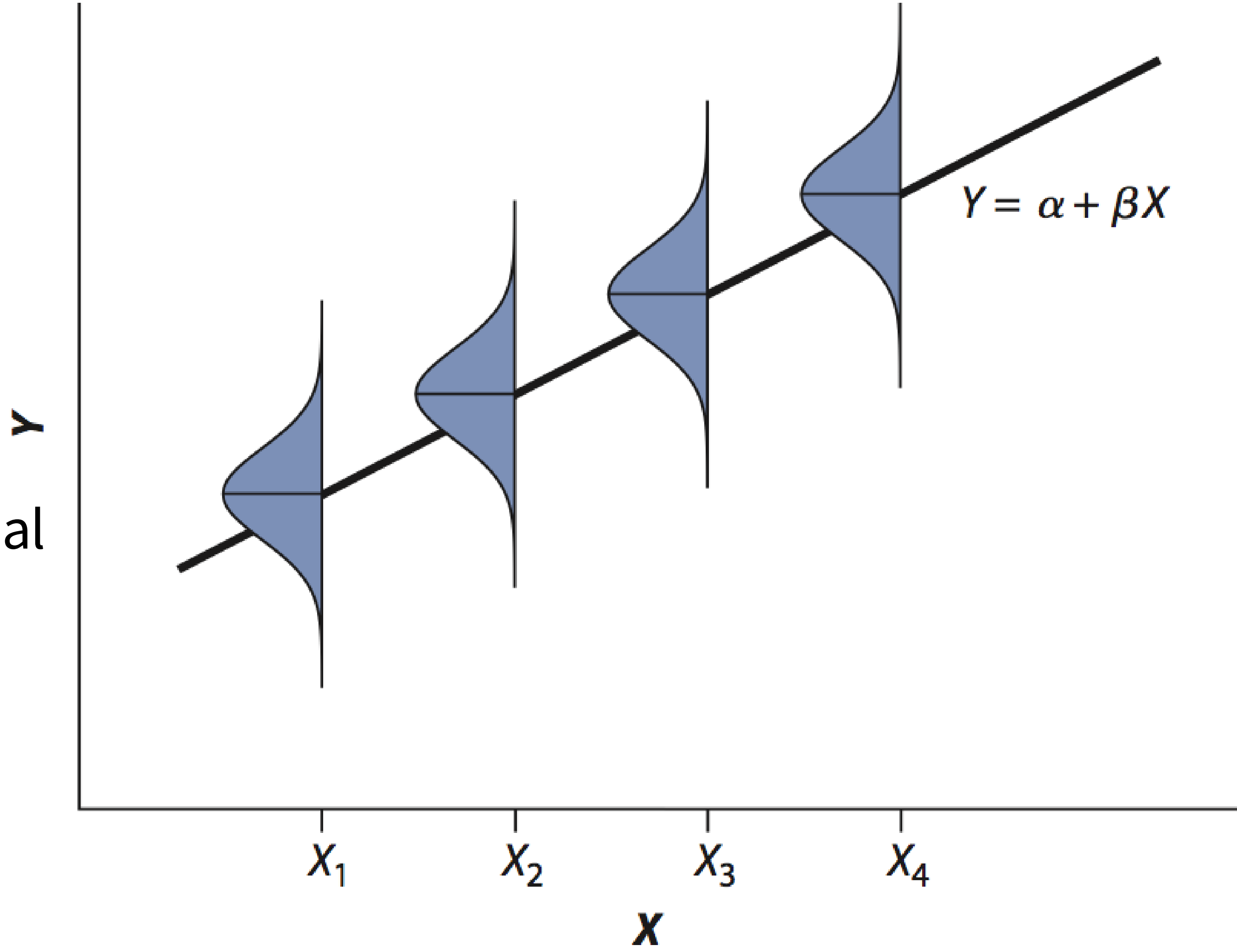
When test is two-tailed and H_0: β = 0, you can use ANOVA approach to testing regression slopes (for multiple models, too!)

- F-test versus t-test
- **If** H_0 is true, then the mean squares corresponding to the two components should be equal

Source	DF	SS	MS	F
Regression (model)	1	$\sum(\hat{Y}_i - \bar{Y}\ )^2$	$\sum(\hat{Y}_i - \bar{Y}\ )^2/1$	$MS_{regression}/MS_{residual}$
Error (residual)	N-2	$\sum(Y_i - \hat{Y}_i\ )^2$	$\sum(\hat{Y}_i - \bar{Y}\ )^2$/(n-2)	
Total	N-1	$\sum(Y_i - \bar{Y})^2$	$\sum(Y_i - \bar{Y})^2$/(n-1)	

Assumptions of Regression Analysis:

- For each X_i, there is a population of Y values whose mean lies on the ‘true’ regression line
  - For each X_i, the Y are a random sample
  - For each X_i, the Y are normally distributed
- Homoscedasticity
  - For every X_i, the variance of Y is equal
- Nothing is assumed about the distribution of X
  - It doesn’t need to be normally distributed or randomly sampled - they might be fixed by the experimenter


Y
$Y = \alpha + \beta X$
X_1
X_2
X_3
X_4
X