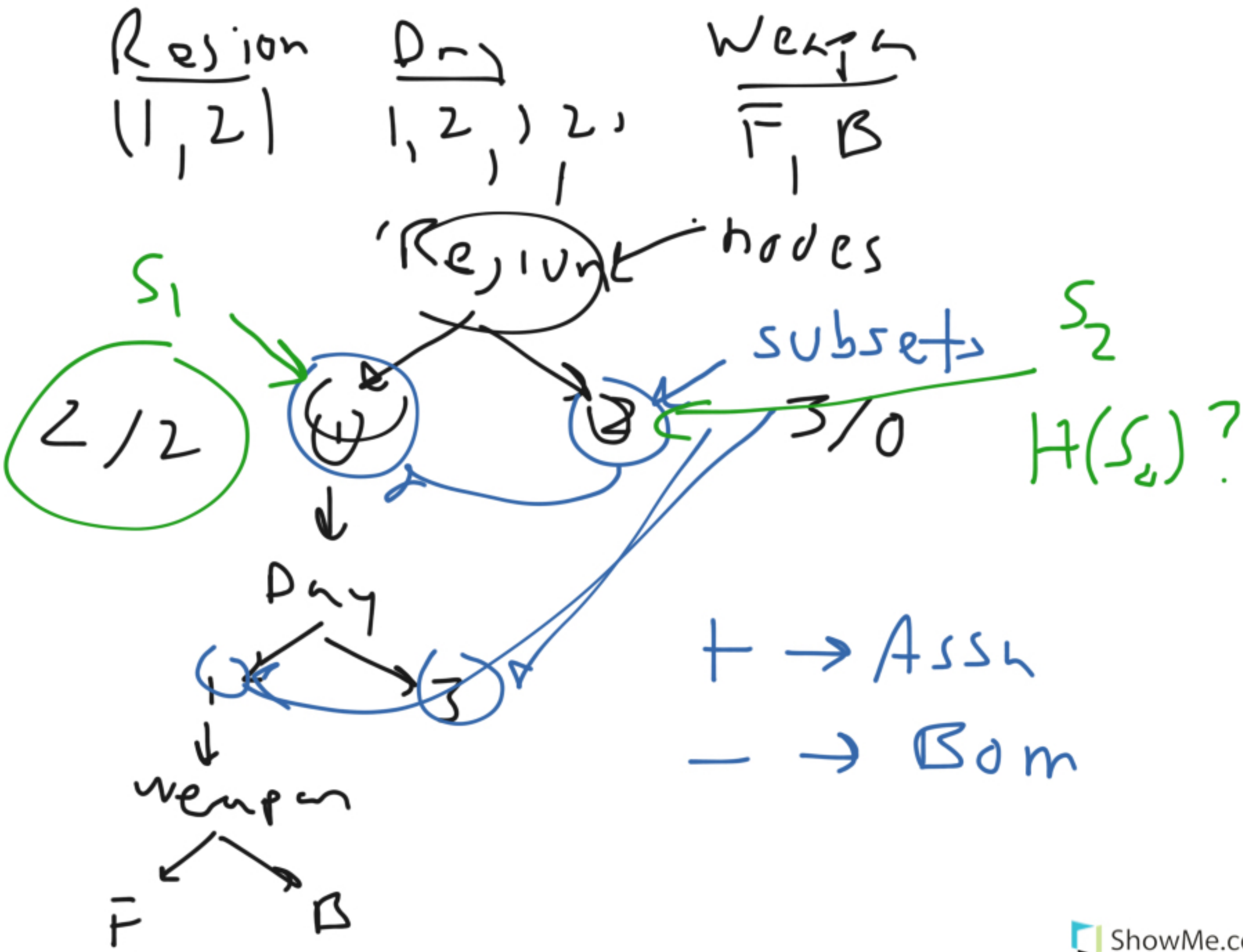

Region
1, 2
Day
Weapon
F, B
Region
nodes
S1
subsets
S2
2/2
3/0
H(S2)?
Day
3
Weapon
F
B
+ → Assn
− → Bom

Information Theory

Entropy

$S_n = \{1, \dots, N\}$

$H(S_n) = -p(+)\log_2 p(+) - p(-)\log_2 p(-)$

$H(S_1) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$

$p(+) = p(\text{Assn} \mid S_1)$ $p(-) = p(\text{Bom} \mid S_1)$

$= \frac{2}{4}$ $= \frac{2}{4}$

$$H(S_1) = 1 \text{ bit}$$

$$H(S_2) = -\frac{3}{3} lg\left(\frac{3}{3}\right) - \frac{0}{3} lg_2\left(\frac{0}{3}\right)$$

$$= 0$$

Info gain (highest) [IG]

IG

$$\text{Gain}(S, A) = H(S) - \sum_{v} \frac{|S_v|}{|S|} H(S_v)$$

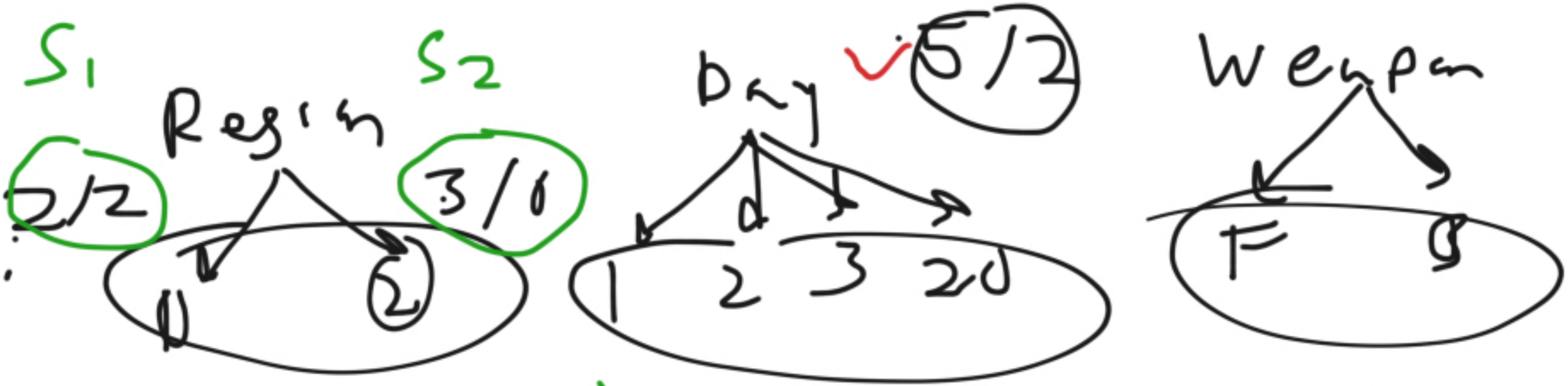
parent

child node

V - values of the feature A

S - set of example or class labels

S_v - subset of each v child node



IG(S, Region) IG(S, Day) IG(S, Weapon)

0.026

$$H(S) = -\frac{5}{7}\log_2\left(\frac{5}{7}\right) - \frac{2}{7}\log_2\left(\frac{2}{7}\right)$$

$$= 0.59$$

$V = \{1, 2\}$

$$\frac{|S_1|}{|S|} H(S_1) + \frac{|S_2|}{|S|} H(S_2)$$

$$\boxed{\frac{|S_1|}{|S|} H(S_1) + \frac{|S_2|}{|S|} H(S_2)}$$

$|S| = 7, \quad |S_1| = 4, \quad |S_2| = 3$

$H(S_1) = 1, \quad H(S_2) = 0$

$$IG(S, \text{Region}) = 0.59 - \left[\frac{4}{7}(1) + \frac{3}{7}(0)\right]$$

$$= 0.026$$

Resim

2/2 0 2

Day(Day, Weapon)

IG(S_1, Day), IG(S_1, Weapn)

What factors are associated with higher casualties in terrorist attack

$Y_i \to$ # of casualties for attack i

(dependend variable)

$X_{i1} \to$ Relig $\to$ 1 I ; $\to$ 0 B

$X_{i2} \to$ Weapon $\to$ 1 F ; $\to$ 0 B

$X_{i3} \to$ Pop density

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

$H_0 : \beta_1 = 0$

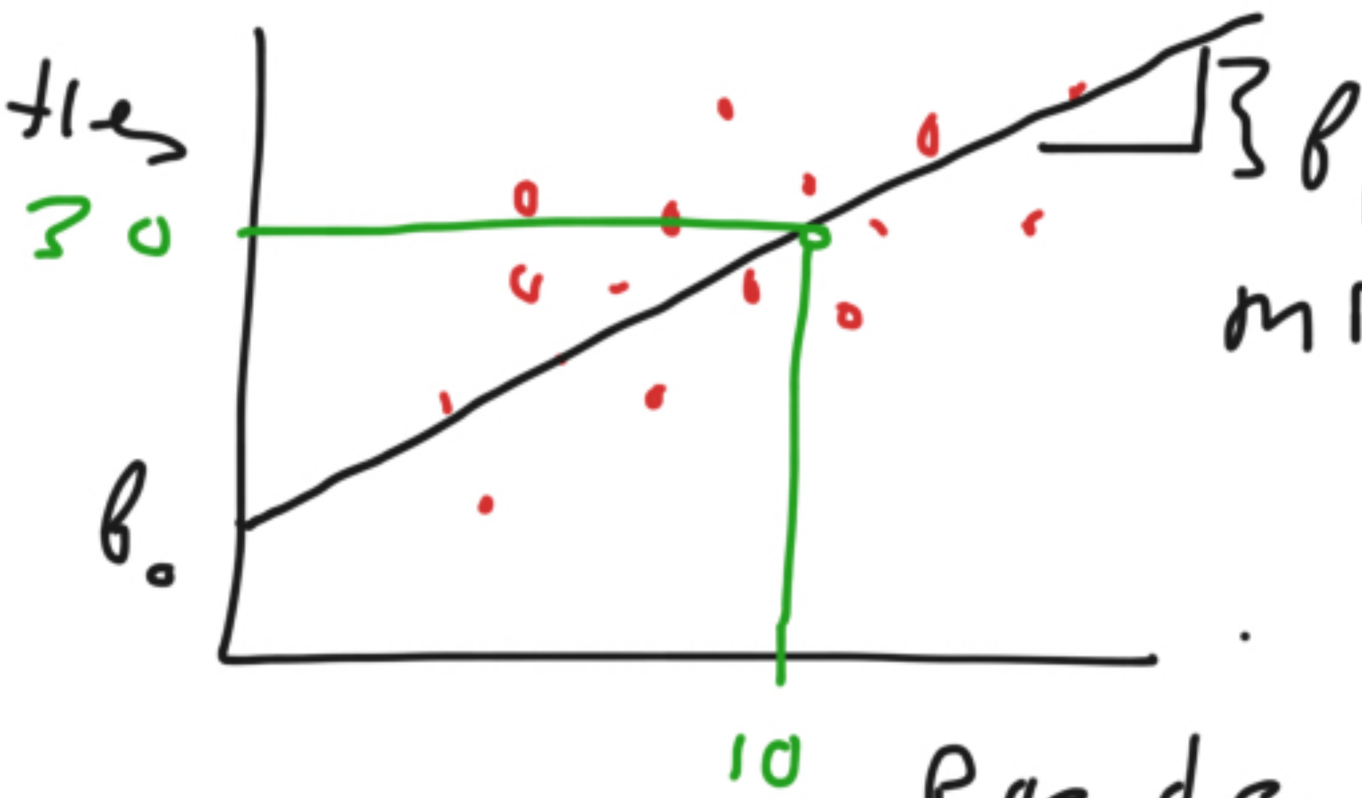
Y_i

Parameter Estimation

Pop density

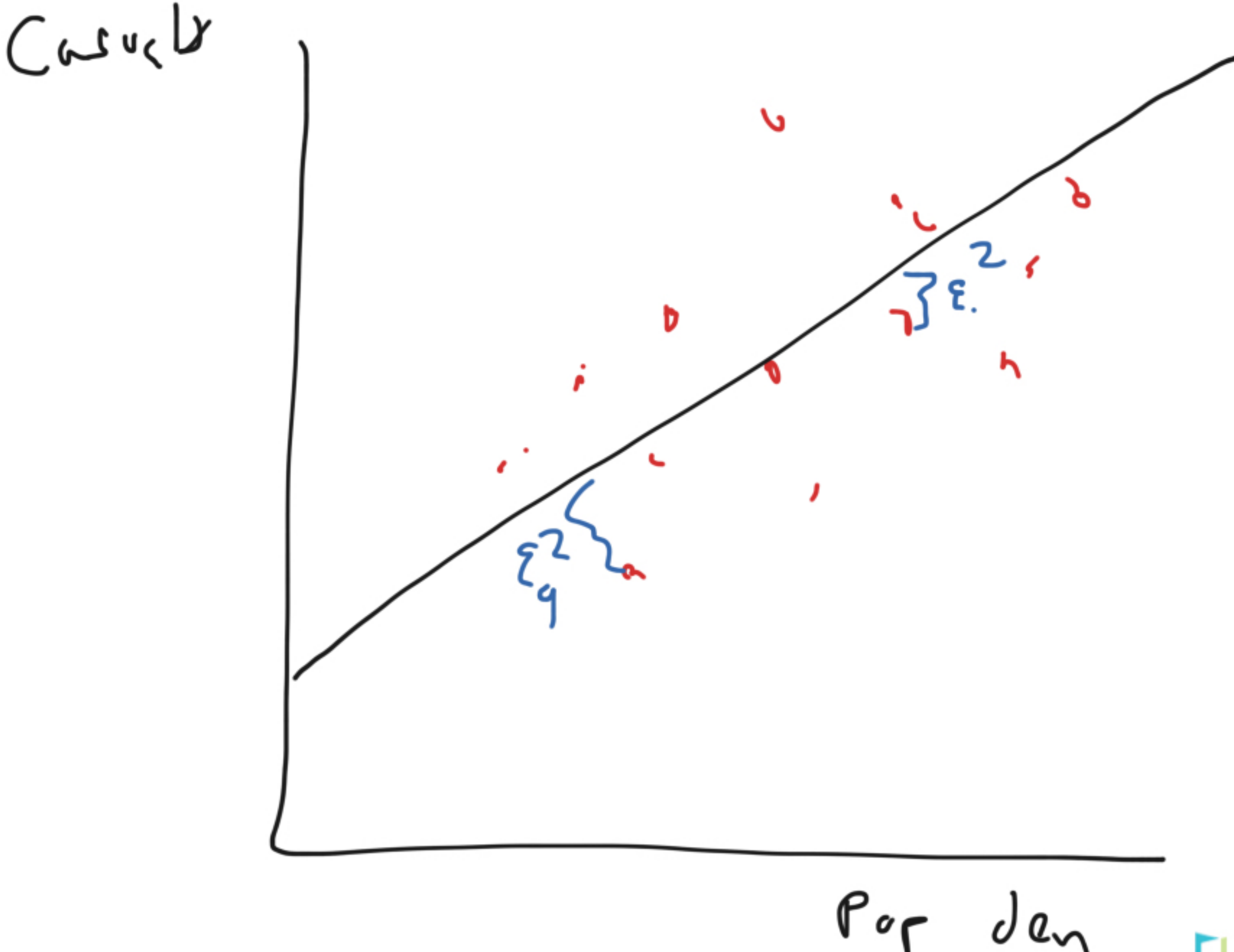
casualties $Y_i = \boxed{\beta_0} + \boxed{\beta_1} X_{i3} + \varepsilon_i$

Casualties



min SSE

Pop density


Pop den

$$Y_i = \beta_0 + \beta_1 X_{ij} + \varepsilon$$

$$\Rightarrow \underset{\beta_0,\beta_1}{\arg\min} \sum_i (Y_i - \beta_0 - \beta_1 X_{ij})^2 = L$$

$$= \sum \varepsilon_i^2$$

$$\frac{\partial \beta_0}{\partial L} = 0 \Rightarrow \cancel{2} \sum (Y_i - \beta_0 - \beta_1 X_{ij}) = 0$$

$$\Rightarrow \underline{\sum Y_i - n\beta_0 - \beta_1 \sum X_{ij}}$$

$$\Rightarrow \boxed{\beta_0 = \frac{\sum Y_i - \beta_1 \sum X_{ij}}{n}}$$

Linear Regression for prediction

- normal eqns for parameter est
- gradient decent
- asses the value or quality of the model that we build
- MSE_{Test} to asses model quality

$$\underbrace{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}_{M_1}$$

$$\underbrace{y_i = \beta_0 + \beta_3 x_{i3}}_{M_2} \quad \hat{y}_i\,?$$

$$RMSE_{M_1} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N}}$$