RISK DIFFERENCE :

$X =$ Exposed $\begin{cases} \text{Yes } E \\ \text{No } \sim E \end{cases}$

$Y -$ Disease $\begin{cases} \text{Yes}, D \\ \text{No}, \sim D \end{cases}$

	D	~D	
E	30	70	100
~E	20	80	100
	50	150	

$$P(D \mid E) = \frac{30}{100} = 0.30$$

$$P(D \mid \sim E) = \frac{20}{100} = 0.20$$

On An Additive scale : Risk Difference (RD) or Attributable Risk (AR) $= \underbrace{P(D \mid E) - P(D \mid \sim E)} = 0.30 - 0.20 = 0.10 = 10\%$

Prob. that an exposure leads to 10% inc. in D. (Extra Risk)

On Relative Scale :

$$\text{Relative Risk} = \frac{P(D \mid E)}{P(D \mid \sim E)} = \frac{0.30}{0.20} = 1.5 = 15\%$$

15% inc. on a relative scale

10% inc. on additive scale

odds(A) = $\dfrac{\text{Probability of A happening}}{\text{Prob. of A not happening}}$

$$\text{Odds Ratio (OR)} = \frac{\text{odds}(D \mid E)}{\text{Odds}(D \mid \sim E)} = \frac{P(D \mid E)/P(D \mid \sim E)}{P(D \mid \sim E)/P(\sim D \mid \sim E)}$$

$$= \frac{0.30/0.70}{0.20/0.80} = 1.71\%$$

odds of D for E is 1.71 times of some one ~E

and $1 - 1.71 = 0.71 \Rightarrow$ odds of D inc by 71% .

Odds Ratio for Case-control Study Design: $OR = \dfrac{\text{Odds}(D \mid E)}{\text{Odds}(D \mid \sim E)}$

	D	~D	
E	a	b	a+b
~E	c	d	c+d
	a+c	b+d	a+b+c+d

$$OR = \frac{P(D \mid E)/P(D \mid \sim E)}{P(D \mid \sim E)/P(\sim D \mid \sim E)} = \frac{\left(\frac{a}{a+b}\right)/\left(\frac{b}{a+b}\right)}{\left(\frac{c}{c+d}\right)/\left(\frac{d}{c+d}\right)} = \frac{ad}{bc}$$

In a disease study, we select people disease and no-disease and ask them the exposure and can't estimate prevalance i.e. these probabilities . We can still conveniently use OR :-

We can estimate prevalance of exposure

$$OR = \frac{\text{odds}(E \mid D)}{\text{odds}(E \mid \sim D)} = \frac{P(E \mid D)/P(\sim E \mid D)}{P(E \mid \sim D)/P(\sim E \mid \sim D)}$$

$$= \frac{(a/a+c)/(c/a+c)}{(b/b+d)/(d/b+d)} = \frac{ad}{bc}$$

OR ≈ Rate Ratio (RR) — For RARE DISEASE!
(prevalence 5% or less)

Non-Linearity In REGRESSION:

Assumptions in Linear Reg:
1. Linearity
2. Constant Variance (Homoscedaticity)

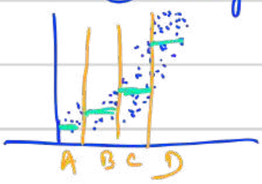
1. Transform Y eg. ln(Y)
   - Can make linear relation
   - can address non-constant variance

Can cause Problem in interpretability

2. Transform X ⇒ $\sqrt{X}$, ln(X). Also
   *Ladder of transformations

3. Polynomial/Quadratic fitting $\hat{Y} = b_0 + b_1X + b_2X^2$
   (Adding powers, more inflection points)
   Again loose Interpretability
   X^2 ∩
   X^3 ~

4. Categorizing X
   Loss of Information

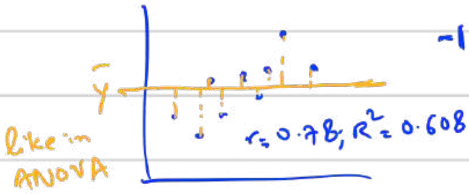


5. Non-linear Regression Model eg. SPLINE
   Disadv. No Interpretability like coeff for X feature Imp

R^2 = coff of Determination:

In a simple linear reg. R^2 is (Pearson Corrl. coeff)^2
$= r$

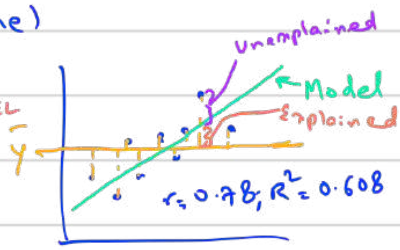
$-1 \leq r \leq +1$ $\quad 0 < R^2 \leq 1$



R^2 = %age of variability in Y explained by model.
(61% here)

Total Var: $SS_{Total} = \sum_{all} (Y_i - \bar{Y})^2$

SS_{Total}
- Explained by Model SS_{EXPL} / SS_{MODEL}
- Unexplained by Model → SS_{UNEXPL} or SS_{ERROR}



$$R^2 = \frac{SS_{EXPL}}{SS_{TOTAL}} = \frac{SS_{MODEL}}{SS_{MODEL} + SS_{ERROR}} = 1 - \frac{SS_{ERROR}}{SS_{TOTAL}}$$

Adjusted R^2 = R^2 − penalty for number of x's in Model.

Used in multiple Linear Reg — to counteract some correl^n bet^n x's & y.

$$y = b_0 + b_1 x_1 + b_2 x_2$$

and Adj R^2 doesn't anymore explains % of variation as R^2.

NOTE: R^2 is generally calculated on the same data.
Use validation

MEASURES OF VARIABILITY:

— RANGE (Max − Min.)

— IQR $Q_3 - Q_1$
Range middle 50% of data ordered
Insensitive to outliers ✓

— Sample variance (s^2)

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

If x in kg
Unit = kg^2

* sensitive to outliers ✓

— Sample S.D (s) = $\sqrt{s^2}$

Monty hall PROBLEM:

Door

	1	2	3
Possibility 1	$	x	x
2	x	$	x
3	x	x	$

You choose	Host	Outcome if you switch
1	2 or 3	Loose
1	has only 1 choice	Win
1	-"-	Win

$\}$ 2/3

If you don't switch = Prob. of winning = 1/3