MODULE - 1
SIMPLE LINEAR REGRESSION

Two things
- Standard linear regression
- Generalized linear models

- Effect on
- Relationship to

$Y \longleftarrow (X_1 \; X_2 \; X_3 \; \dots \; X_3)$

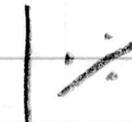
"Response"
"Dependant"
What we are trying to understand

"Predicting"
"Explanatory"
"Independant"

This is a RANDOM var

This is a FIXED var
→ Set before observing response.

SIMPLE : $Y = \beta_0 + \beta_1 X + \epsilon$



a "plane"

MULTIPLE : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

POLYNOMIAL : $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \epsilon$

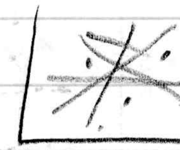
Objectives : ① Prediction
② Modeling relationships
③ Testing (associations)

(INDEP)
of each other

SIMPLE REGRESSION : $Y = \beta_0 + \beta_1 X + \mathcal{E}$

β_0 → Intercept
β_1 → slope
$\mathcal{E}$ → **DEVIANCE FROM MODEL**

GOAL : find (β_0, β_1)
to find the best fit line



which one??

DATA : We have a list of pairs
$\{(x_1, y_1), (x_2 y_2) \ldots (x_n y_n)\}$

Composed of
$X_1 = (x_1, x_2 \ldots)$

Think a table!

Y	x_1	x_2	x_3	
y_1	-	-	-	→ X_1
y_2	-	-	-	→ X_2

⇒ **MODEL** : $Y_i = \beta_0 + \beta_1 X_i + \mathcal{E}_i \; i \quad \forall i \in (1, n)$

ASSUME ① $E(\mathcal{E}_i) = 0$ → Expected Value
Linearity / Mean zero assumption.

② $Var(\mathcal{E}_i) = 0$ (Constant var. assump.)

③ $\mathcal{E}_0 \ldots \mathcal{E}_n$ are iid (Indep. assump.)

ASSUMPTIONS DETAILED

① $\boxed{\text{Linearity}}$: → Even if $Y = \beta_0 + \beta_1 X^{\bullet} + \beta_2 X^2 + \ldots$

this is still linear!

just $\boxed{\text{addition}}$!

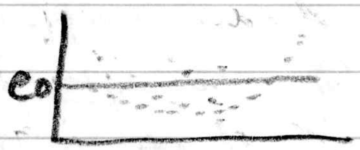
When you say "mean of errors must be zero", imagine this (you forgot negative numbers!)

$\Sigma(e) = 0$

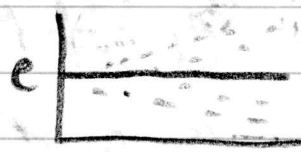
$E[e \mid x_1 x_2 \ldots] = 0$ (???)

Residual: $e = Y_i - \hat{Y}_i$ (Exp, pred.)

If you see these, indicates Systematic prob in ~~dat~~ model:



Quadratic effects



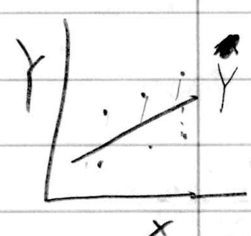
"Fanout" ⇒ unequal var ⇒ heteroscedastic!

→ Violation means problems estimating β_0

→ Means that model is too low/high for certain subgroups.

DIVERSION

You were wrong: point is to not find a ~~min~~ line where sum of errors is zero (X)



There can be ∞ such lines! This is because of negative values. This is why we use sums

So, not $\Sigma(e) = 0$ WRONG

you want $Min[\Sigma(e^2)]$ RIGHT

SSR = ESS
Expl. S of Sq

SSE = RSS
Residual Sum of Sq

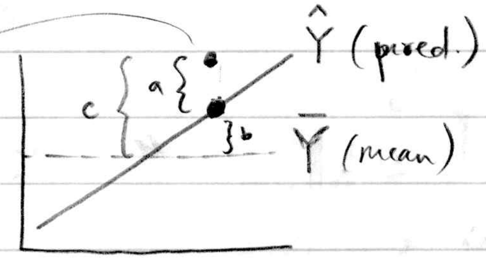
ind. var / data point

Response Var Stuff : $\bar{Y}_i$ Y_i $\hat{Y}_i$

mean — observed / Expected — predicted

Population : $Y_i = \beta_0 + \beta_1 X_1 + \mathcal{E}$

Sample : $\hat{Y} = b_0 + b_1 \cancel{X}_1 + e$ or $\hat{\beta}_0 + \hat{\beta}_1 X_1 + e$



(a) is the unexplained deviation from mean

(b) is the explained deviation from mean

(c) Total deviation from mean

So,

$$a = \sum (Y - \bar{Y}_i)^2 \quad \text{"SSE"} \rightarrow \text{Error}$$

$$b = \sum (\hat{Y} - \bar{Y})^2 \quad \text{"SSR"} \rightarrow \text{Regression}$$

$$a + b = SST \text{ — Total}$$

So! The whole point of R^2 is "how much deviation from mean has your model explained??"

$$R^2 = \frac{\text{Deviation from mean Explained by Regression/Model}}{\text{Deviation from mean Total}} = \frac{\text{Variation Expl. by Model}}{\text{Total Var.}}$$

or $\boxed{R^2 = SSR/SST}$

GOD! FINALLY!! ♡

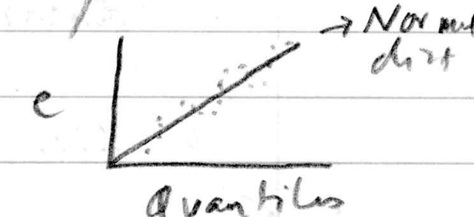
$$SST = \underbrace{\sum (Y_i - \hat{Y}_i)}_{SSE} + \underbrace{\sum (\hat{Y}_i - \bar{Y}_i)}_{SSR}$$

$$\Rightarrow SST = \sum (Y_i - \bar{Y}_i)$$

Constant Variance : $Var(\varepsilon_i) = \sigma^2$

IF NOT → Model is better for some chunks of data than others. :(

→ "Results in poorly calibrated prediction intervals" ??? maybe? →



Independence : $\varepsilon_1\ \varepsilon_2\ \varepsilon_3 \ldots \varepsilon_n$ are iid.

→ One obs. has no effect on the other

→ Prob. w/ time series data

→ "Strength of Regression" is affected ???

ε_i is Normally Dist : Needed for stat. Inference.

→ If not, hyp. ~~t~~ tests, conf. Int. affected ???

| MODEL PARAMS : $\underbrace{\beta_0\ \beta_1\ \sigma^2}$, μ $\boxed{\text{POPULATION}}$

→ These are unknown, theoretical!

→ They will always be unknown!

Theoretical → Assumed that they exist ???
(can we prove they don't?)

⟹ When you speak of $\beta_0\ \beta_1$ → "Parameters"

When you say $\hat{\beta}_0\ \hat{\beta}_1\ b_0\ b_1$ → "Estimators"

↓

Practical ...

More terms

observed pred.

→ Fitted Values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

→ Residuals : $r_i = \hat{\varepsilon}_i = e_i = y_i - \hat{y}_i$

→ Mean Sq. Error $= \dfrac{SSE}{n-2}$

—— DIVERSION ——

→ You always deal with ESTIMATES.

You SAMPLE the POPULATION and hope to get models that can describe that

STATISTIC PARAMETER

	SAMPLE		POPULATION
Mean	$\bar{x}$	Estimates of →	μ
Std dev	s	→	σ
Model.	$\hat{Y}$	→	Y
	$\hat{\beta}_0$	→	β_0
	$\hat{\beta}_1$	→	β_1

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} \qquad \sigma^2 = \frac{\Sigma(x-\mu)^2}{n}$$

"Statistic $\boxed{\text{NOT}}$ a Parameter"

↳ Think of what this means!

TODO: $MSE = SSE/_{n-2}$, $\chi^2_{n-1\ \text{dof}}$.

$\boxed{\text{You need to understand DoF}}$

This is so fucking important!

PARAM estimator $Y \propto \beta_1 X$

+ve $Y\uparrow$

-ve $\underline{Y\downarrow}$

~0 no sig. assoc.

MODEL estimator $\hat{Y} \propto \hat{\beta}_1 X + \hat{\beta}_0$

↳ Exp. Val when pred. var equal zero

Exp. change ∀ unit change in pred. var.

→ Estimators are determined by $\hat{\beta}_0\ \hat{\beta}_1$

<u>min</u> <u>SSE</u> ($SSE = \sum(Y_i - \hat{Y}_i)$)

∴ This is an optimization prob!

→ β_1 is also thought of as "gradient"

→ "Unbiasedness" means that

if E[Something] = [redacted] some other TRUE thing

⇒ "Something" is an unbiased Est. of

"some other TRUE thing"

So, if you can prove that

$$E[\hat{\beta}_0] = \beta_0 , \quad \hat{\beta}_0 \xrightarrow[\text{est. of}]{\text{unb.}} \beta_0$$

<u>Objective</u>: Minimize SSE!

$$SSE = \sum (y_i - \hat{Y}_i)$$

But <u>NO</u>! Remember the <u>GOD MODEL</u>.

Goal: $\boxed{\text{Min} \sum (y_i - Y_i)}$

You cannot use $\hat{Y}_i$ because you don't know ~~that~~ what it will be! This is in theoretical space!

GOAL: $\text{Min} \sum (y_i - (\beta_0 + \beta_1 x_i))^2$

Min [Sum of sq. err. bet. <u>observed</u> and theoretical <u>perfect</u> (God) model]

<u>SIDENOTE</u>: $S_{xx} = \sum (x_i - \bar{x})(x_i - \bar{x}) = \sum (x_i - \bar{x})^2$

$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$

Now do magic and get $\hat{\beta}_0$, $\hat{\beta}_1$ from the PDE $\left(\underset{\beta_0 \beta_1}{\text{Min}} \sum (y_i - \beta_0 - \beta_1 x_i)^2 \right)$

SOLUTIONS <u>TO PDE</u>

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$

<u>TO PROVE</u>

$\hat{\beta}_0$ is an unbiased est. of its perfect twin β_0 in the perfect world.

"Formally" and strictly speaking, <u>Later</u>, prove that the <u>solutions</u> ($\hat{\beta}$) are <u>unbiased</u> estimators of <u>THE TRUE β_0 and β_1</u>.

NOTE: $\sum(x_i - \bar{x}) = 0$ Think about it.... -ve numbers!

but! $\sum(x_i - \bar{x})^2 \neq 0$ Taking care of

ALSO: $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

$= \frac{\sum y_i(x_i - \bar{x}) + \sum \bar{y}(x_i - \bar{x})}{S_{xx}}$ ② to zero ∵ NORMALITY ASSUMPTION

① PULLOUT ∵ Constant!

$= \frac{\sum y_i(x_i - \bar{x}) + \bar{y}\sum(x_i - \bar{x})}{S_{xx}}$

$\therefore \boxed{\hat{\beta}_1 = \frac{\sum y_i(x_i - \bar{x})}{S_{xx}}}$ Y's mean vanishes!!

$N(0,1)$ $\left.\begin{array}{l}\text{Mean}\\ \text{Variance}\end{array}\right\}$ "Standard Normal Dist"

Variance is always +ve (Duh... sum of sq. of diff)

but! Covariance is +ve or -ve.

MAYBE In GOD-DIMENSION, $N = \infty$

→ So you cannot say what some random var. X will assume as a value!

∴ DoF!

→ You can with a sample for one/two datapoint

Estimation, Prediction

$\hat{y} | x^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are N.D.

$E(\hat{Y} | x^*) = \beta_0 + \beta_1 x^*$

$Var(\hat{Y} | x^*)$

$= \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)$

EST ONE SOURCE of uncertainty : ~~pred~~ estimators $\hat{\beta}_0, \hat{\beta}_1$

PRED TWO Sources : estimators & new observation.

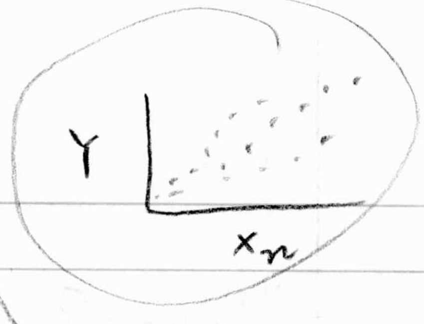
→ PRED : Regression Eq. is same as Est.
USE PRED INT not! CONF INT
wider than CONF INT in Est.

Called ANOVA but is tut on means!
Variance

Dude finish yours EFS

- Linearity / Mean Zero

$E[\varepsilon_i] = 0$

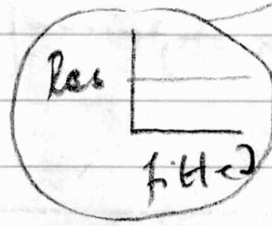


- Constant Variance $Var(\varepsilon_i) = \sigma^2$

- Independence : ε_i are iid

- Normality $\varepsilon_i \sim N(0, \sigma^2)$

Hist/QQ of residuals



Estimation	Prediction
- Formula same	Formula same
- Average ESTIMATED MEAN response for all settings of x^*	- Average Response for one setting of x^*
- Use CONF INT	- Use PRED INT ↓ **WIDER**
-	
Uncertainty └ Estim of β_0, β_1	Uncertainty └ $(n+1)^{th}$ obs └ Estimate of β_0, β_1

- "Expected value"
- "Mean profit"
- $\text{Exp}(\hat{Y}|x^*)$ =
- <u>Variance diff!</u> (uncertainty)

- Allow for error
- <u>Future</u> value
- "<u>This particular value</u>"
- "Profit in week 10"
- $\text{Exp}(\hat{Y}|x^*)$
- <u>Variance has more!</u> one more σ^2! (uncertainty $2\times\uparrow$)

~~As for~~

"Estimating a mean resp."

"Predicting a new observation"