PROJECT 2022-04-25

T TEST

2 POPULATIONS

ARE MEANS DIFFERENT?
VAR

NULL HYPOTHESIS: MEANS SAME.

ALTERNATIVE: DIFFERENT.

YOU MIGHT GET PARTICULAR ABOUT HOW DIFFERENT.

E.G. BLUE > RED

Q: IF NO DIFF, HOW LIKELY IS WHAT YOU SAW.

ANOVA

ANALYSIS OF VARIANCE.

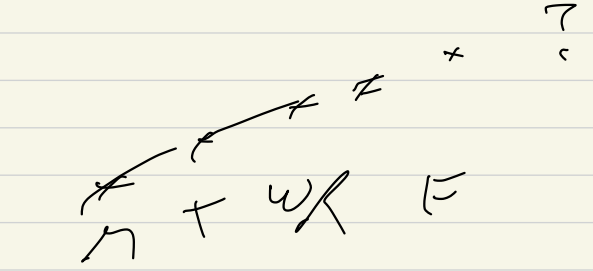
DRUG TRIALS:
DECIDE WHAT WILL BE MEASURED + WHAT SUCCESS WILL BE, BEFORE TRIALS.

BAD DATA DREDGING:
LOOK AT RESULTS TO FIND INTERESTING THINGS.

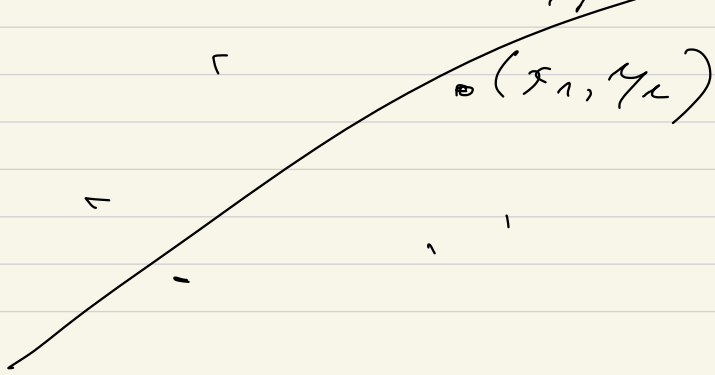
LINEAR REGRESSION

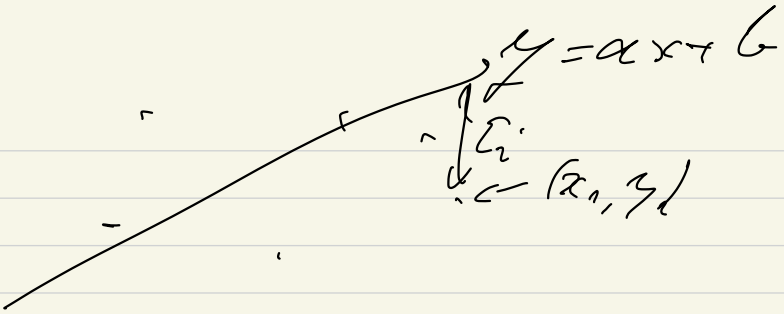
STATISTICS, FIND PARAMETERS, TEST HYPOTHESES, FIND RELATIONS. PREDICT FUTURE.

EG S&P500



FIT A LINE TO DATA - $y = Ax + B$.





FIND a, b TO MINIMIZE ERROR

$$\varepsilon_i = (\underbrace{a x_i + b}_{\text{COMPUTED FROM LINE}} - \underbrace{y_i}_{\text{ACTUAL}})^2 ?$$

MIN $\sum \varepsilon_i^2$

$$\text{MIN } E = \sum (a x_i + b - y_i)^2$$

PICK a, b TO MIN THAT.

$$E = a^2 \sum x_i^2 + N b^2 + \sum y_i^2$$
$$+ 2ab \sum x_i - 2a \sum x_i y_i$$
$$+ 2n \sum b y_i$$

$$\frac{dE}{da} = 2a\sum x_i^2 + 2b\sum x_i - 2\sum x_i y_i$$

$$= 0 \text{ at min.}$$

$$a\sum x_i^2 + b\sum x_i - \sum x_i y_i = 0$$

$$\frac{dE}{db} = 2Nb + 2a\sum x_i + 2n\sum y_i$$

$$= 0$$

$$Nb + a\sum x_i + N\sum y_i = 0$$

SOLVE FOR a, b.

COMPUTE E.

MULTILINEAR REGRESSION

$y_i = a_1 x_{i1} + a_2 x_{i2} + a_3 x_3 + \varepsilon$

STEPWISE REGRESSION

PICK MOST IMPORTANT INDEPENDENT VARIABLES

ADD TO MIX (R_i^2).

USE CORRELATION COEFFICIENT.

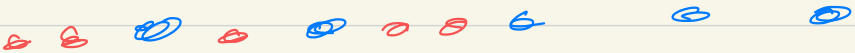
COMPUTE ρ FOR EACH INDEPENDENT VAR + DEPENDENT VAR. PICK VAR WITH BIGGEST $|\rho|$.

COMPUTE RESIDUAL ERRORS.

ADD NEW MOST IMPORTANT VAR.

NON PARAMETRIC STATS

- DON'T ASSUME ANYTHING ABOUT DISTRIBUTION.
MORE ROBUST
BUT WEAKER.



ARE RED OBSERVATIONS SMALLER. DOES RED POPULATION HAVE SMALLER MEAN?
USE ONLY THEIR ORDER.
COUNT # TIMES $R_i < B_j$
$U = 5+5+4+3+3 = 20$
"MANN-WHITNEY U STAT"
IF NULL $\bar{U} = 12.5$
$VAR(U) = \sim 23$. $\sigma \approx 5$

NULL HYPOTHESIS THAT RED
+BLUE ARE SAME MEAN
& 1.5 NORMAL VAR.
·TOOK SHORTCUTS.
THIS IS UNLIKELY.

ML
MACHINE
LEARNING

CURRENT BIG
APP OF STATS