STATISTICS FOR DATA ANALYSIS

EMPHASIS: METHODS THAT ARE INSENSITIVE TO HOW DATA OR ERRORS ARE DISTRIBUTED.

I. FITTING
- LINEAR LEAST SQUARES (ADVANTAGES & LIMITATIONS)
- MINIMUM ABSOLUTE DEVIATION; NONLINEAR FITTING
(LATER WE'LL TOUCH ON A NONPARAMETRIC ALTERNATIVE)

II. CONFIDENCE INTERVALS
- χ^2 (BRIEFLY)
- STATISTICAL BOOTSTRAP
→ INSTRUMENT FOCUS TEST EXAMPLE (FITTING W/UNKNOWN ERRORS)

III. NONPARAMETRIC (DISTRIBUTION FREE) METHODS
→ EXAMPLE: MEDIAN VS. MEAN
- MEASURES OF CORRELATION
- NONPARAMETRIC FITTING
- KOLMOGOROV-SMIRNOV TEST
→ EXAMPLE: DISTRIBUTION OF NEARBY STARS

WE TAKE K-S TEST AS AN EXAMPLE OF FREQUENTIST HYPOTHESIS TESTING.

MODELING OF DATA (NR CH. 15)

LINEAR LEAST SQUARES — VERY FAST! $\mathcal{O}(N)$

FIT $y_i' = m x_i + b$

GIVEN MEASUREMENTS $y_i = y_i' + \varepsilon_i$, $x_i = x_i'$.

IF $P(\varepsilon)$ IS NO WORSE THAN GAUSSIAN (I.E. NO FAT TAILS)
THEN LINEAR LEAST SQUARES IS A GOOD CHOICE.

σ_i = VARIANCE OF ε_i

THE GENERAL IDEA IS THAT WE WANT TO MINIMIZE χ^2 WITH RESPECT TO m AND b:

$$\chi^2 = \sum_i \frac{(y_i^{MODEL} - y_i)^2}{\sigma_i^2} = \sum_i \frac{(m x_i + b - y_i)^2}{\sigma_i^2}$$

MINIMIZE WRT m:

$$0 = \frac{\partial \chi^2}{\partial m} = 2 \sum_i \frac{x_i (m x_i + b - y_i)}{\sigma_i^2}$$

MINIMIZE WRT b:

$$0 = \frac{\partial \chi^2}{\partial b} = 2 \sum_i \frac{m x_i + b - y_i}{\sigma_i^2}$$

(I.E. ADJUST b SO THAT $\bar{y} = \bar{y}^{MODEL}$.)

THESE ARE EASY TO SOLVE ANALYTICALLY: (NR EQ. 15.2.6)

$$b = \frac{S_{xx} S_y - S_x S_{xy}}{\Delta}$$

$$m = \frac{S S_{xy} - S_x S_y}{\Delta}$$

$$\Delta = S S_{xx} - (S_x)^2 \quad \text{WHERE...}$$

$$S \equiv \sum_i \frac{1}{\sigma_i^2} \qquad S_x \equiv \sum_i \frac{x_i}{\sigma_i^2} \qquad S_y \equiv \sum_i \frac{y_i}{\sigma_i^2}$$

$$S_{xx} \equiv \sum_i \frac{x_i^2}{\sigma_i^2} \qquad S_{xy} \equiv \sum_i \frac{x_i y_i}{\sigma_i^2}$$

THE UNCERTAINTIES IN THE FIT ARE

$$\sigma_b = \sqrt{S_{xx}/\Delta}$$

$$\sigma_m = \sqrt{S/\Delta}$$

(NR 15.2.9)

(SEE DERIVATION IN NR)

QUESTION: HOW GOOD IS THE FIT? I.E. IS THE LINEAR MODEL SUFFICIENT TO DESCRIBE THE RELATIONSHIP BETWEEN x AND y ?

TRADITIONAL ANSWER: MUST KNOW σ_i WELL TO ANSWER. IF $\chi^2_R = \chi^2/N \sim 1$, THEN THE LINEAR MODEL FULLY EXPLAINS THE DATA.

FAILING THAT, ASK WHETHER $y_i - m x_i$ IS STATISTICALLY INDEPENDENT OF x_i ...

FITTING MORE GENERAL FUNCTIONS — GENERAL LINEAR LEAST SQUARES

DATA (x_i, y_i), UNCERTAINTY σ_i

MODEL $y^*(x) = \sum_j a_j f_j(x)$

THE FUNCTIONS $f_j(x)$ NEED NOT BE ORTHOGONAL, BUT SHOULD BE LINEARLY INDEPENDENT.

$$\chi^2 = \sum_i \frac{(y_i - y^*(x_i))^2}{\sigma_i^2}$$

$$= \sum_i \frac{\left(y_i - \sum_j a_j f_j(x_i)\right)^2}{\sigma_i^2}$$

NOW MINIMIZE WITH RESPECT TO a_k:

$$0 = \frac{\partial \chi^2}{\partial a_k} = \sum_i \frac{2 f_k(x_i)\left(y_i - \sum_j a_j f_j(x_i)\right)}{\sigma_i^2}$$

$$\sum_j \sum_i \frac{a_j f_j(x_i) f_k(x_i)}{\sigma_i^2} = \sum_i y_i f_k(x_i)$$ "NORMAL EQUATIONS"

LET $M_{jk} \equiv \sum_i \frac{f_j(x_i) f_k(x_i)}{\sigma_i^2}$, $b_k \equiv \sum_i y_i f_k(x_i)$

SO $\sum_j M_{jk} a_j = b_k$ --- $M\vec{a} = \vec{b}$

SOLVE FOR $\vec{a}$ ... NR `lfit()` USES GAUSS-JORDAN ELIMINATION,
`svdfit()` USES SINGULAR VALUE DECOMPOSITION.

AN ALTERNATIVE: MINIMUM ABSOLUTE DEVIATION (MORE ROBUST)

$$y^*(x) = f(\vec{a}, x) \quad \text{--- FULLY GENERAL,}$$

$$M(\vec{a}) \equiv \frac{\sum_i |y^*(x_i) - Y_i|}{\sigma_i}$$

MINIMIZE M WRT $\vec{a}$. (WE KNOW HOW TO DO THIS!

THE MAIN QUESTION IS: WHAT IS THE UNCERTAINTY IN THE PARAMETERS $\vec{a}$?

BOOTSTRAP

(1) FROM THE N DATA POINTS (X_i, Y_i), RESAMPLE WITH REPLACEMENT N POINTS.

(2) DO THE FIT AGAIN, RESULT $\vec{a}^1$.

(3) REPEAT $m(\sim 10^4)$ TIMES $\rightarrow$ $a_i^1, a_i^2, \ldots a_i^j, \ldots a_i^m$
AND SORT THE a_i^j ASCENDING.

(4) ESTABLISH CONFIDENCE LIMITS ON a_i.
E.G., HERE IS THE 98% (2-SIDED) CONFIDENCE RANGE ON a_i:

$$a_i^{m/100} < a_i \leq a_i^{m-\frac{m}{100}}$$

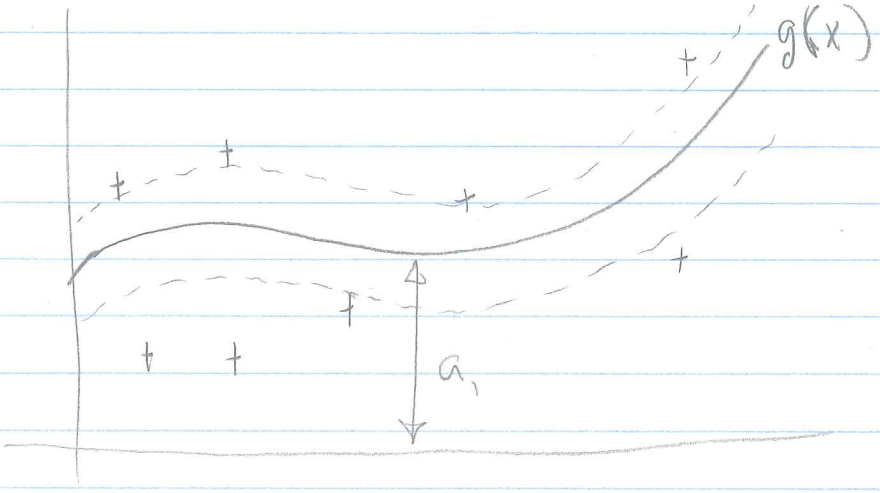
THIS CAN BE APPLIED TO ANY FITTING EXERCISE, REGARDLESS OF THE TECHNIQUE.

Broader context: HYPOTHESIS TESTING

PROBLEM WITH MINIMUM ABSOLUTE DEVIATION:
IT DOES NOT GIVE ONE UNIQUE SOLUTION.

SIMPLE EXAMPLE: MODEL $y = a_1 + g(x)$

SUPPOSE WE HAVE AN EVEN NUMBER OF DATA POINTS:



$M(a_1)$ IS MINIMIZED AS LONG AS

THERE'S AN EQUAL # OF DATA POINTS ABOVE AND BELOW THE CURVE. (EZ TO SHOW).
AS A RESULT, THE ANSWER IS <u>ANYWHERE</u> WITHIN THE DASHED LINES.

SIMILAR UNCERTAINTIES TYPICALLY APPLY TO ALL PARAMETERS a_i IN A MORE COMPLICATED MODEL.

<u>ASIDE:</u>
- NONLINEAR LEAST SQUARES
- NONLINEAR FITTING USING OTHER GOF CRITERIA