$$\alpha_i = \mu_{i\cdot} - \mu_{\cdot\cdot}$$

$$\beta_{j(i)} = \mu_{ij} - \mu_{i\cdot}$$

$$\sum_{i=1}^{a} \alpha_i = 0 \qquad \sum_{j=1}^{b} \beta_{j(i)} = 0 \quad \text{for } i = 1, \ldots a$$

ANOVA

partitioning the SS (balanced design)

$$y_{ijk} = \bar{y}_{\cdot\cdot\cdot} + (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot}) + (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot}) + (y_{ijk} - \bar{y}_{ij\cdot})$$

$$SSTO = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{\cdot\cdot\cdot})^2$$

$$= \sum_i \sum_j \sum_{k=1}^{n} (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2 + \sum_i \sum_j \sum_k (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot})^2 + \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij\cdot})^2$$

$$= bn \sum_i (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2 + n \sum_i \sum_j (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot})^2 + \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij\cdot})^2$$

$$= SS_A + SS_{B(A)} + SSE$$

Note in Two-way ANOVA

$$SSTO = bn \sum_i (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2 + an \sum_j (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot})^2 + n \sum_i \sum_j (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot})^2$$

$$+ \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij\cdot})^2$$

$$SS_{B(A)} = SS_B + SS_{AB}$$

Under Normality, all SS_s/σ^2 are indep.

Computationally.

$$SS_{TO} = \sum_i \sum_j \sum_k y_{ijk}^2 - \frac{y_{...}^2}{abn} \qquad y_{...} = \sum\sum\sum y_{ijk}$$

$$SS_A = \frac{1}{bn} \sum y_{i..}^2 - \frac{y_{...}^2}{abn}$$

$$SS_{B(A)} = \frac{1}{n} \sum\sum y_{ij.}^2 - \frac{1}{bn} \sum y_{i..}^2$$

$$SS_E = \sum\sum\sum y_{ijk}^2 - \frac{1}{n} \sum\sum y_{ij.}^2$$

ANOVA table.

Source	SS	df	MS
A	SS_A	$a-1$	
B(A)	$SS_{B(A)}$	$(b-1)+(a-1)(b-1)$ $= a(b-1)$	
Error	SS_E	$ab(n-1)$	
total	SS_T		

- For nested design.

  For any bracketed subscripts in the model, place a 1 under those subscripts that are inside the brackets.

- If A and B fixed.

	F	F	R
	a	b	n
	i	j	k
α_i	0	b	n
$\beta_{j(i)}$	1	0	n

$$EMS_S \qquad EMSE = \sigma^2$$

$$EMSA = \sigma^2 + bn \sum_{i=1}^{a} \frac{\alpha_i^2}{a-1}$$

$$EMSB(A) = \sigma^2 + n \frac{\sum_i \sum_j \beta_{j(i)}^2}{a(b-1)}$$

F test A: $\frac{MSA}{MSE}$, B(A) $\frac{MSB(A)}{MSE}$

- If A Fixed, B random.

	F	R	R
	a	b	n
	i	j	k
α_i	0	b	n
$\beta_{j(i)}$	1	1	n

$$E MSA = \sigma^2 + bn \frac{\sum \alpha_i^2}{a-1} + n\sigma_\beta^2$$

$$E MSB(A) = \sigma^2 + n\sigma_\beta^2$$

$$E(MSE) = \sigma^2$$

F test A: $\frac{MSA}{MSB(A)}$, B(A): $\frac{MSB(A)}{MSE}$

$$\hat{\sigma}_\beta^2 = \frac{MSB(A) - MSE}{n} \quad V\ 0$$

- If A, B random

	R	R	R
	a	b	n
	i	j	k
α_i	1	b	n
$\beta_{j(i)}$	1	1	n

$$E MSA = \sigma^2 + bn\sigma_\alpha^2 + n\sigma_\beta^2$$

$$E MSB(A) = \sigma^2 + n\sigma_\beta^2$$

$$E MSE = \sigma^2$$

F-test A: $\frac{MSA}{MSB(A)}$ B(A): $\frac{MSB(A)}{MSE}$

$\hat{\sigma}_\beta^2 =$ $\hat{\sigma}_\alpha^2 =$

Fractional factorial design (Chap 29)

2^k factorial design

- k factors, each factor has two levels (often labeled + & −)
- very useful for preliminary analysis
- can "remove" unimportant factors
- all interactions.

$k = 2$

general 2-factor factorial model

$$\mu_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

$$1 + (a-1) + (b-1) + (a-1)(b-1)$$

4 parameters

$$\mu, \quad \alpha_1 = -\alpha_2, \quad \beta_1 = -\beta_2$$

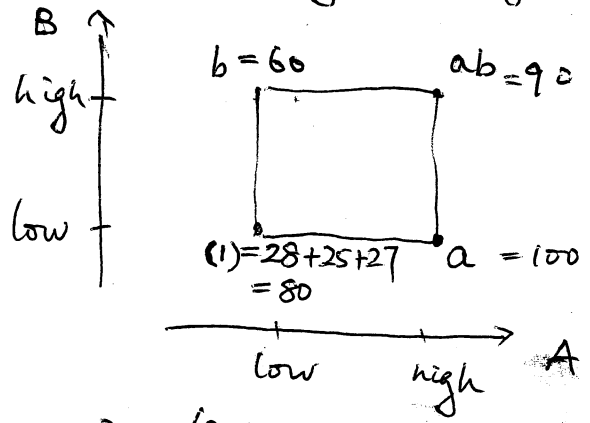
$$\begin{cases} (\alpha\beta)_{11} + (\alpha\beta)_{12} = 0 \\ (\alpha\beta)_{21} + (\alpha\beta)_{22} = 0 \\ (\alpha\beta)_{11} + (\alpha\beta)_{21} = 0 \end{cases} \qquad (\alpha\beta)_{11} = (\alpha\beta)_{22}$$

Design for 2^2 factorial

- Label the levels of factors A and B using + and −
- There are 4 experimental combinations labeled.

Symbol	Factor A	Factor B
(1)	−	−
a	+	−
b	−	+
ab	+	+

	I	II	III
A low B low	28	25	27
A high B low	36	32	32
A low B high	18	19	23
A high B high	31	30	29



- Can express combination in terms of model parameters.

$ab = (+,+)$: $Ey = \mu + \alpha_2 + \beta_2 + (\alpha\beta)_{22}$

$a = (+,-)$: $Ey = \mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21}$

$b = (-,+)$: $Ey = \mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12}$

$(1) = (-,-)$: $Ey = \mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11}$

ab $\quad E y = \mu + \alpha_2 + \beta_2 + (\alpha\beta)_{22}$

a $\quad E y = \mu + \alpha_2 - \beta_2 - (\alpha\beta)_{22}$

b $\quad E y = \mu - \alpha_2 + \beta_2 - (\alpha\beta)_{22}$

(1) $\quad E y = \mu - \alpha_2 - \beta_2 + (\alpha\beta)_{22}$

Estimate parameters (4)

$$\hat{\mu} = \frac{ab + a + b + (1)}{4n}$$

n: replications of each combination

$$\hat{\alpha}_2 = \frac{1}{2}\left(\frac{ab+a}{2n} - \frac{b+(1)}{2n}\right) = \frac{ab + a - b - (1)}{4n}$$

$$\hat{\beta}_2 = \frac{1}{2}\left(\frac{ab+b}{2n} - \frac{a+(1)}{2n}\right) = \frac{ab - a + b - (1)}{4n}$$

$$(\hat{\alpha\beta})_{22} = \frac{1}{2}\left(\frac{ab+(1)}{2n} - \frac{a+b}{2n}\right) = \frac{ab - a - b + (1)}{4n}$$

Nonparametric Regression (kernel smoothing, spline smoothing)

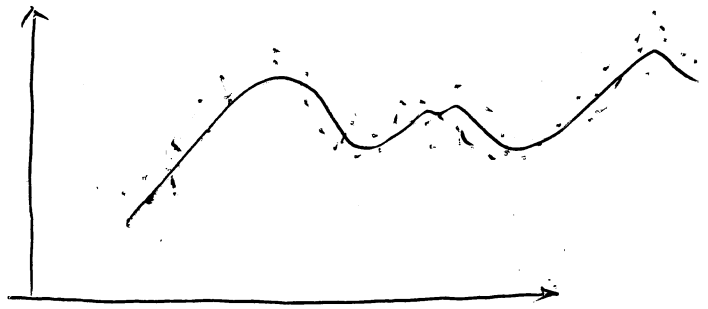
- Regression problem: Given fixed $(x_1, \cdots x_n)$, we observe $(y_1, \cdots y_n)$

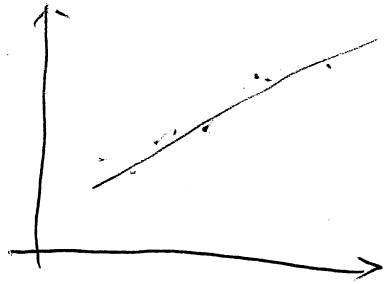
where $y_i = m(x_i) + \varepsilon_i$

ε_i : iid

$E(\varepsilon_i) = 0, \quad Var(\varepsilon_i) = \sigma^2$

m: unknow, the problem is to estimate m.





Assumption:

the nonparametric approach is to choose m from some smooth family of functions, "some degree of smoothness".

- advantage and shortcoming

  pros: flexibility

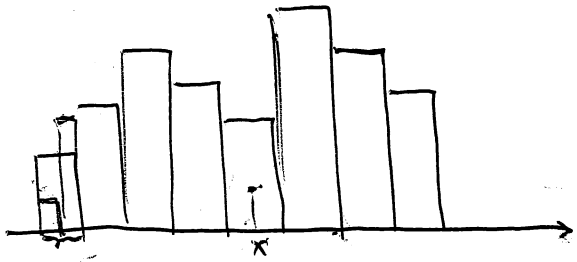
  useful when little past experience is avaible

  con: . parametric approach is more effective if the model is correct.

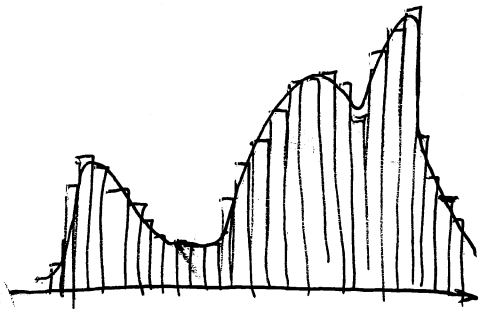
  . do not have a formulaic way to describe the relationship (graphically)

- kernel smoothing

  histogram



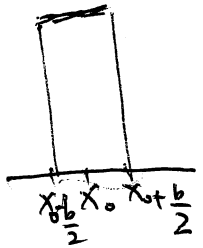
bin width : $b \to 0$



$f(x)$. prob. density func. at x.

$$\frac{\#\text{ of } X_i \text{ in the } Bin(x)}{n} \approx P(X \in B(x))$$

$$\int_a^b f(x)\,dx = P(a \le X \le b)$$



$$\int_{x_0-\frac{b}{2}}^{x_0+\frac{b}{2}} f(x)\,dx \approx f(x_0)\cdot b$$

$$\hat{f}(x_0, b) = \frac{P(X \in (x_0 - \frac{b}{2},\ x_0 + \frac{b}{2}))}{b}$$

$$= \frac{\#\text{ of } X_i \text{ in } (x_0-\frac{b}{2},\ x_0+\frac{b}{2})}{nb}$$

$$= \frac{\sum_{i=1}^{n} I[X_i \in (x_0-\frac{b}{2},\ x_0+\frac{b}{2})]}{nb}$$

$n \to \infty$ ①

$b \to 0$ ②

$$E \sum_{i=1}^{n} I[X_i \in (x_0-\frac{b}{2},\ x_0+\frac{b}{2})] \approx nb\,f(x_0)$$

If $n \to \infty$, $nbf(x_0) \to \infty$

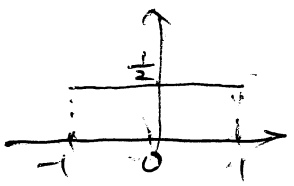
So $nb \to \infty$ ③

for example $b \sim \frac{1}{\sqrt{n}}$

h : bandwidth.

$$\hat{f}_h(x) = \frac{\sum_{i=1}^{n} I(X_i \in (x-h,\ x+h))}{n\ 2h}$$

$$= \frac{\sum_{i=1}^{n} \frac{1}{2h} I\left(\left|\frac{X_i - x}{h}\right| < 1\right)}{n}$$



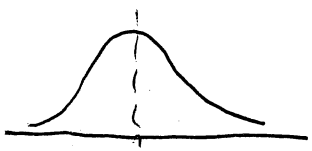
$$= \frac{\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x}{h}\right)}{n} \leftarrow \text{here } K(u) = \frac{I(|u|<1)}{2}$$

unif kernel

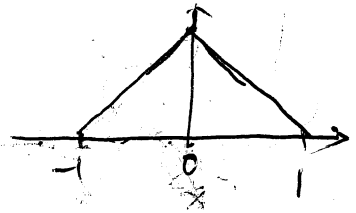
$$= \frac{\sum_{i=1}^{n} K_h(X_i - x)}{n} \leftarrow \text{here } K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$$

· Choice of kernel function

$K(u) = \varphi(u)$ standard normal



$K(u) = (1-|u|)\, I(|u|<1)$ triangular kernel



• kernel function

i). $h \to 0$, $nh \to \infty$

ii) K is symmetric $\int_{-\infty}^{\infty} uK(u)\,du = 0$

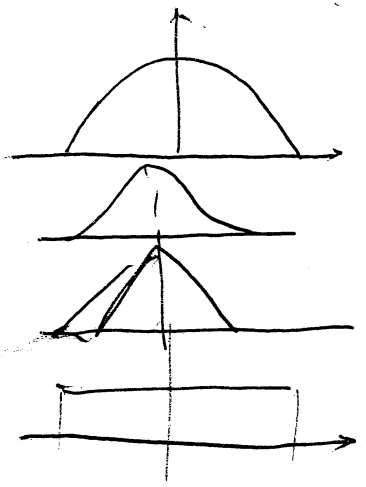
$$\int_{-\infty}^{\infty} K(u)\,du = 1$$

$$K(u) \geq 0 \quad , \quad \int u^2 K(u)\,du < \infty$$

kernel function seledion.

Epanechinikov is the "best" kernel

Normal	.951
triangular	.986
uniform	.93



· Density estimation

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-X_i}{h}\right) \longrightarrow f(x)$$

i) $\hat{f}_h(x) \geq 0$

ii) $\int_R \hat{f}_h(x)\,dx = \frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{h} k\left(\frac{x-X_i}{h}\right) dx = 1$

· $MSE\{\hat{f}_h(x)\} = Bias^2 + Variance$

here $Bias = E\,\hat{f}_h(x) - f(x)$

$Variance = Var(\hat{f}_h(x))$

$Bias = E\,\hat{f}_h(x) - f(x)$ → $\{X_1, \cdots X_n\}$ ~ $f(x)$

$= E\,\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-X_i}{h}\right) - f(x)$

$= \int \frac{1}{h} k\left(\frac{x-u}{h}\right) f(u)\,du - f(x)$

Let $v = \frac{u-x}{h}$ ⟹ $\int \frac{1}{h} k(-v)\, f(x+vh)\, h\, dv - f(x)$

$$= \int K(v)\, f(x+vh)\, dv - f(x)$$

$$\overset{\text{Taylor}}{=} \int K(v)\left[f(x) + vh f'(x) + \frac{(vh)^2}{2} f''(x) + o(h^2)\right] dv - f(x)$$

$$= f(x) + f'(x)h \int v K(v)\, dv + \frac{f''(x)h^2}{2}\int v^2 K(v)\, dv + o(h^2) - f(x)$$

$$= \frac{h^2 f''(x)}{2}\mu_2(K) + o(h^2)$$

$$\mu_m(K) = \int_{-\infty}^{\infty} K(v)\, v^m\, dv$$