EX POLYNOMIAL FITTING

K-NN

SSR

← CROSS VALIDATION
← NEW DATA

← TRAINING DATA

n_{POLY}

→ "COMPLEX"
"FLEXIBLE"

TRAINING DATA

100%

ACC 50%

NEW DATA

1 K ∞ n_{OBS}

¿?   CROSS VALIDATION

¿?? NEW DATA

¿V CROSS VALIDATION
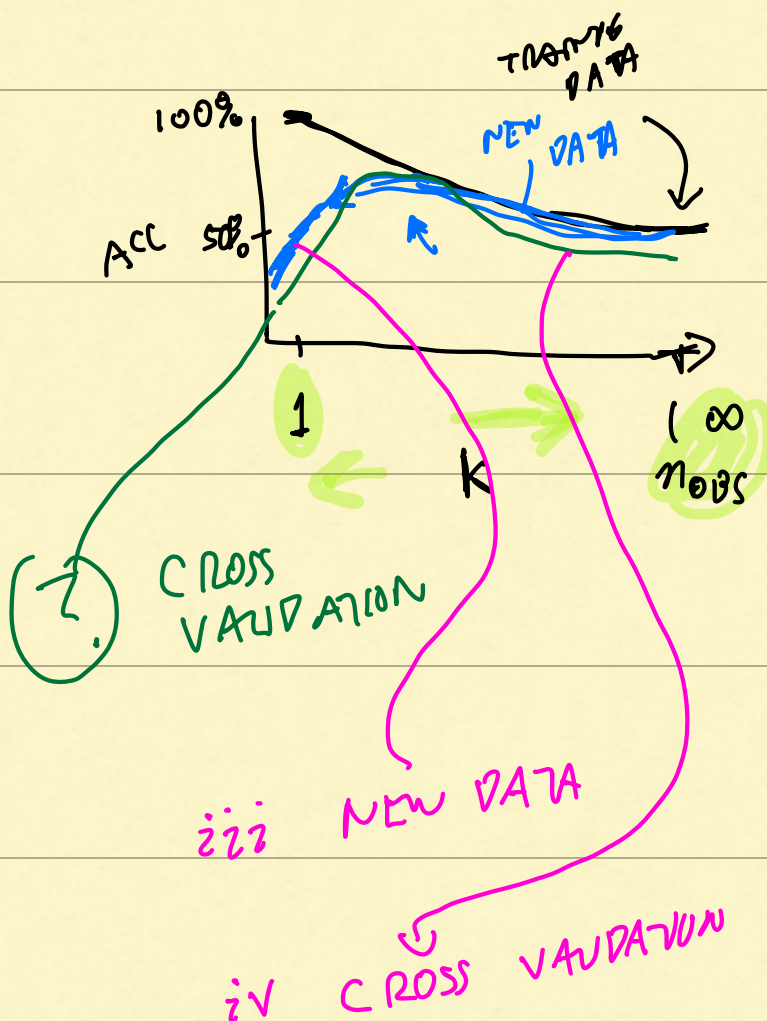
HIGH DIMENSIONAL DATA

DATA: $N_{CELLS} \approx 1000$ ← OBSERVATIONS

$N_{GENE} \approx 20000$ ← PARAMS

$N_{PATIENTS} = 10$

$N_{SPECIES} \approx 10000$ ← SPECIES IN GUT MICROBIOME

SSR

$n_{POLY} = n_{OBS}$

n_{POLY}

EX $n_{OBS} = 2,$ $n_{POLY} = 1$

Y

X

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

β_1

β_0

SSR = 0
$L = \infty$

$n_{OBS} = 3$ $n_{POLY} = 1$

β_1

β_0

SSR SMALL

$n_{OBS} = 1$ $n_{POLY} = 1$

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

β_1

β_0

SSR = 0

NON-IDENTIFIABILITY
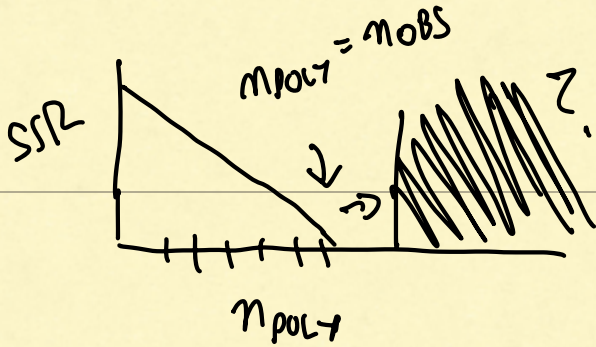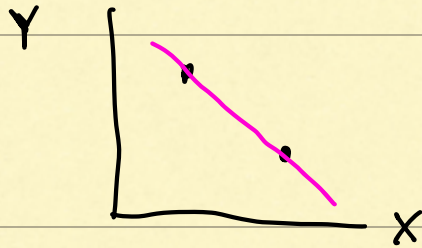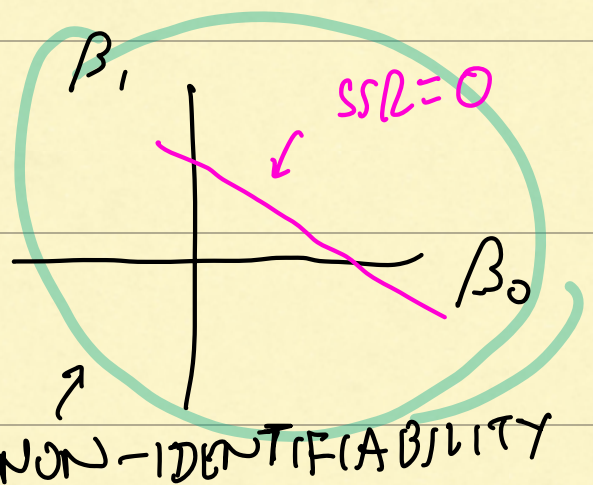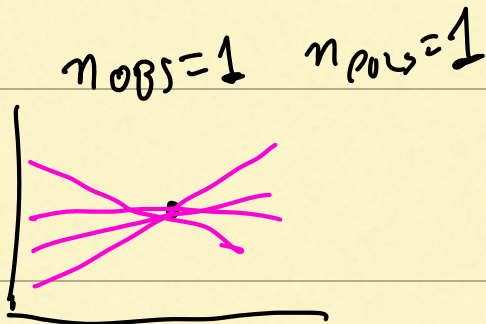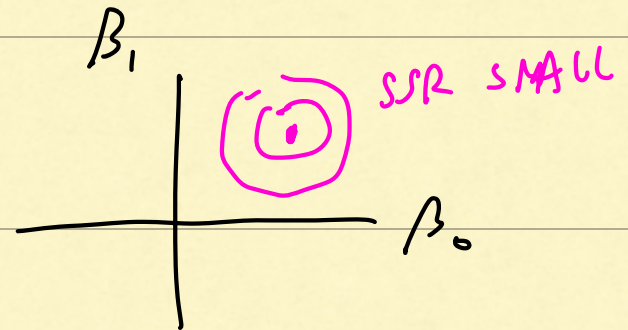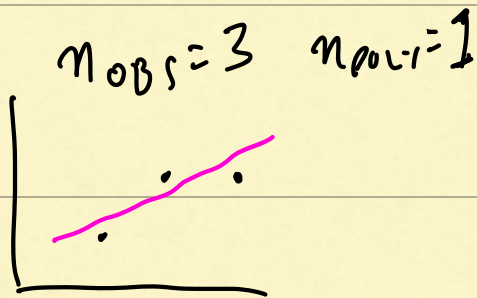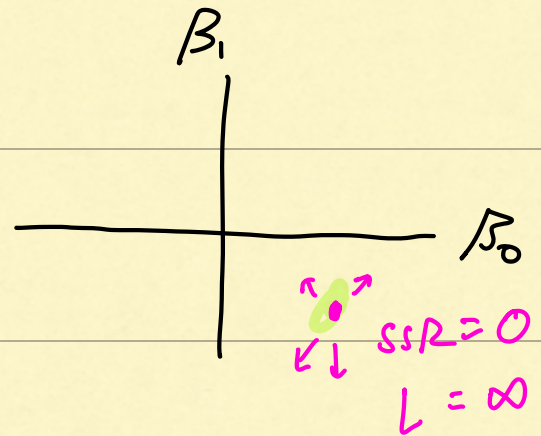
RIDGE   REGRESSION

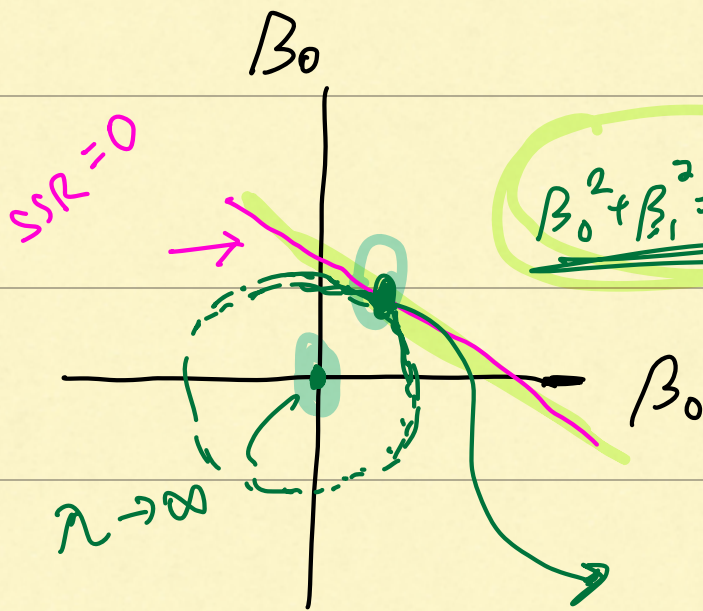MINIMIZE $SSR + \lambda \sum \beta_i^2$

MAXIMIZE $L(\beta_i) - \lambda \sum \beta_i^2$

↑ LIKELIHOOD   ↑ PENALTY
                         REGULARIZATION

$\lambda (\beta_0^2 + \beta_1^2)$

B_0

SSR=0

$\beta_0^2 + \beta_1^2 = CONST$

B_0

$\lambda \to \infty$

$B_0,$
β_1

λ
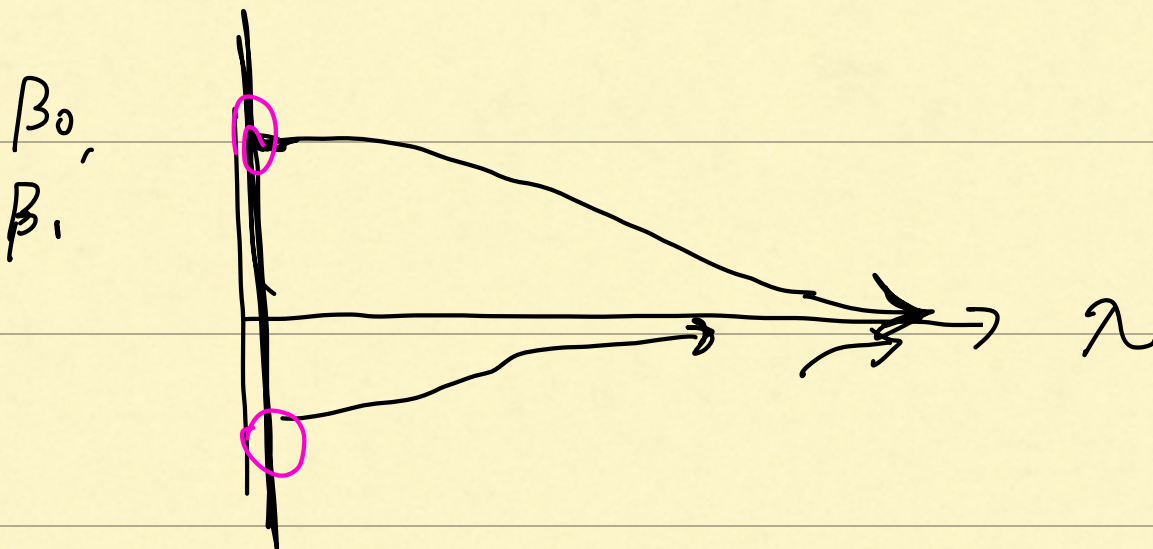
LASSO REGRESSION

MINIMIZE $\quad SSR + \lambda \sum_i |\beta_i|$

MAXIMIZE $\quad L(\beta) - \lambda \sum_i |\beta_i|$

$\underbrace{}$

REGULARIZATION

$|\beta_0| + |\beta_1| = CONST$

SSR=0

β_1

β_0

$\lambda \to \infty$

β_i
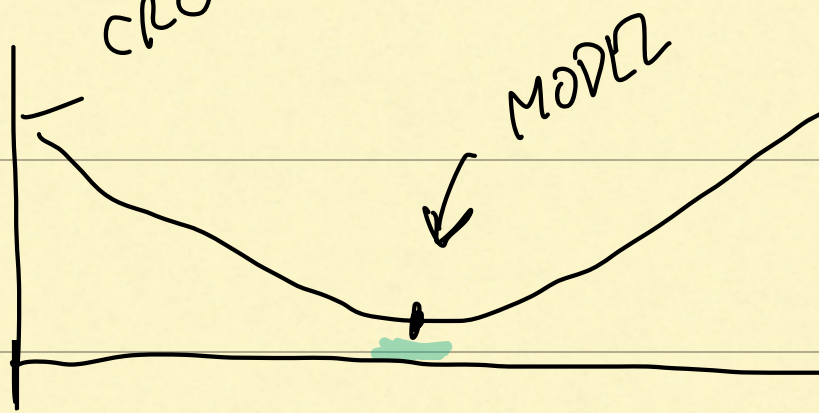
λ
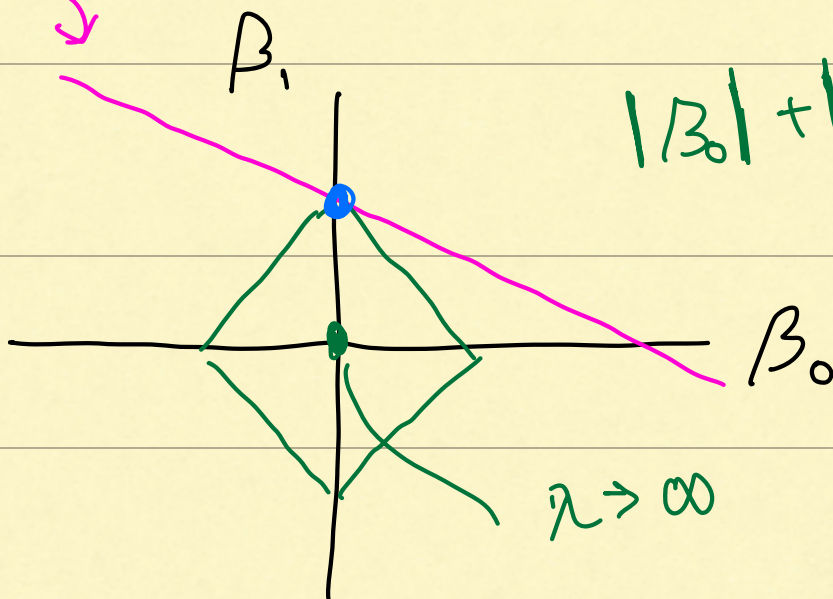
FEWER (NONZERO)

MORE
(NON ZERO)
PARAMETERS

PARAMETERS

CROSS VALIDATION

MODEL

MODEL
COMPLEXITY