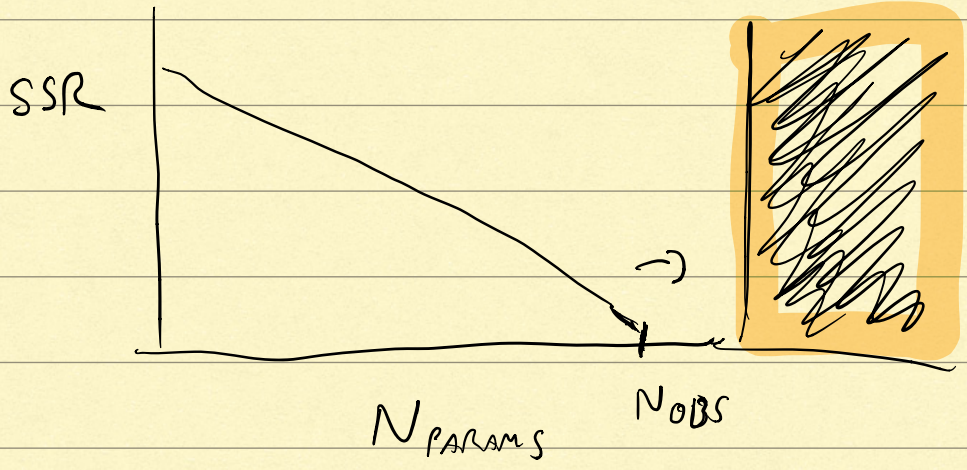
# HIGH-DIMENSIONAL DATA

SUPPOSE $N_{OBS}$ DATA POINTS

MODEL HAS $N_{PARAMS}$ PARAMETERS

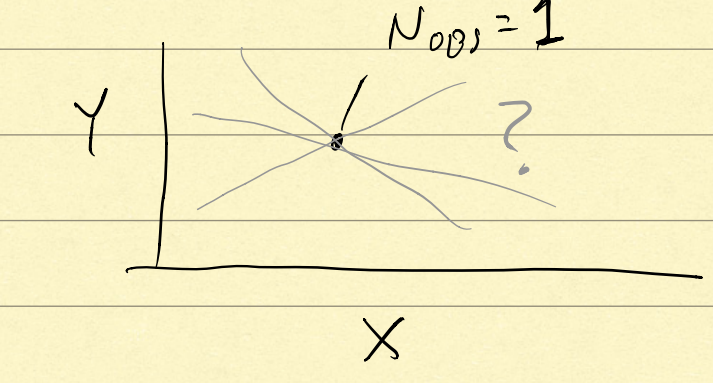


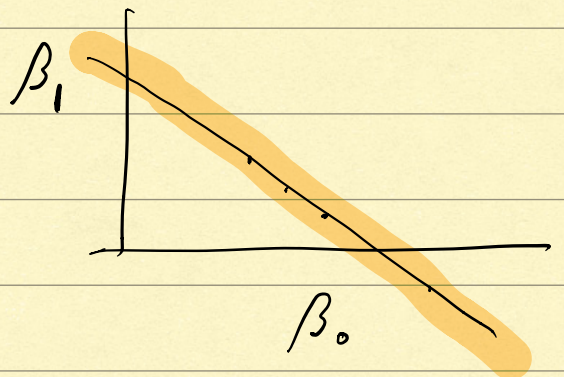

$N_{PARAMS} > N_{OBS}$

GENES ~30000 — CELLS ~100

BACTERIAL SPECIES IN GUT ~$10^4$ — PATIENTS ~10

EX




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

$N_{PARAM} = 2$

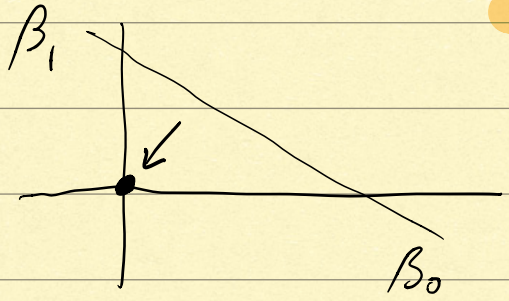
## RIDGE REGRESSION (LAME :)

PENALTY

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

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

IF $\lambda \to \infty$

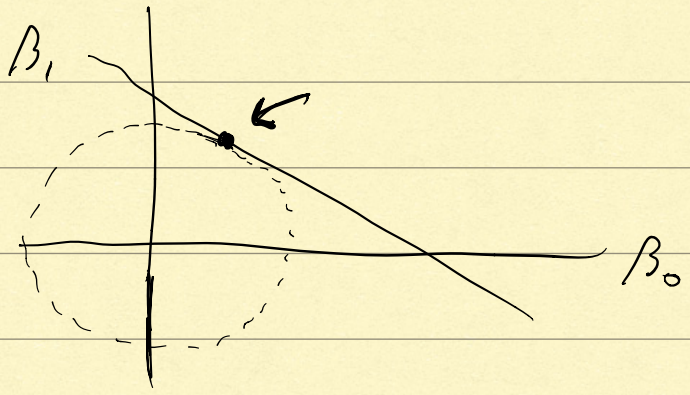


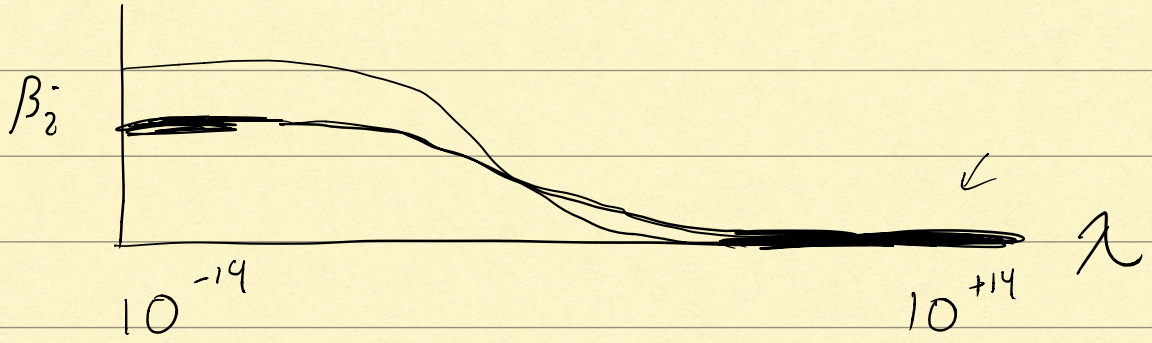

$\beta_0^2 + \beta_1^2 = C$

CIRCLE

IF $\lambda \to 0$

(BUT NOT $\lambda = 0$)



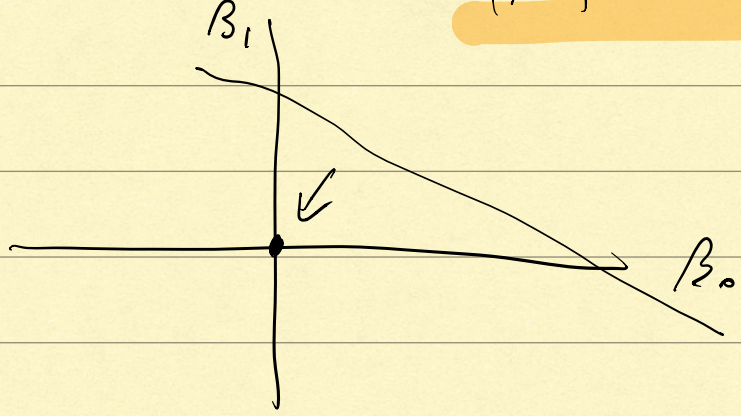

---

## LASSO REGRESSION

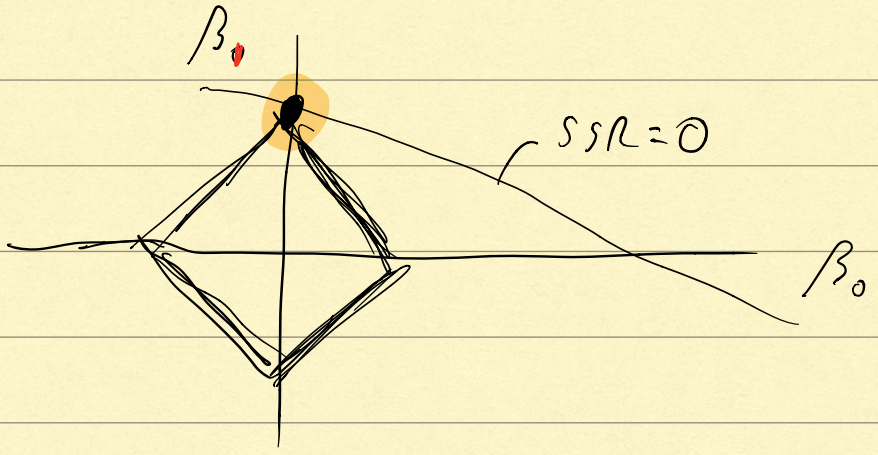
MINIMIZE $SSR + \lambda \sum_i |\beta_i|$

$|\beta_0| + |\beta_1| = C$

As $\lambda \to \infty$




$\lambda \to 0$




# GENERALIZED LINEAR MODELS

PATIENT DIABETES (YES/NO)

→ GLUCOSE LEVEL $X$

→ INSULIN LEVEL $Y$

$p_i$ - PROBABILITY PATIENT $i$ HAS DIABETES

$$p_i = \frac{e^{\beta_0 + \beta_1 X + \beta_2 Y}}{1 + e^{\beta_0 + \beta_1 X + \beta_2 Y}}$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X + \beta_2 Y$$

ODDS

LOG ODDS

↖ BINOMIAL GENERALIZED LINEAR MODEL

| | | FACTORS | |
|---|---|---|---|
| PATIENT # | DIAGNOSIS | X | Y |
| 1 | Y | $X_1$ | $Y_1$ |
| 2 | Y | $X_2$ | $Y_2$ |
| 3 | N | $X_3$ | $Y_3$ |
| 4 | Y | $X_4$ | $Y_4$ |

LIKELIHOOD

$$L = p(X_1, Y_1) \cdot p(X_2, Y_2) \cdot \left(1 - p(X_3, Y_3)\right) \cdot p(X_4, Y_4)$$

$$L(\beta_0, \beta_1, \beta_2)$$