Lecture Ten

High-Dimensional Linear Regression

① **Linear Regression:**

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$$

$$y_t = \beta_0 + \beta_1 \cos\frac{2\pi t}{12} + \beta_2 \sin\frac{2\pi t}{12} + \varepsilon_t$$

② **Non Linear Regression:**

$$y_t = \beta_0 + \boxed{\beta_1} \cos 2\pi \boxed{f} t + \boxed{\beta_2} \sin 2\pi \boxed{f} t + \varepsilon_t$$

$$RSS(f) = \min_{\beta_0, \beta_1, \beta_2} \sum \left(y_t - \cdots\right)^2$$

③ **High-Dimension ~~Linear~~ Regression:**

$$y_t =$$

Change of Slope Model

California yearly popution at time t.

$t \downarrow$ year

$$y_t = \log(\text{CA population at time } t)$$ 1900 to 2024

① $$y_t = \beta_0 + \underbrace{\beta_1}_{} t + \varepsilon_t$$

$100 \times \beta_1$: percent growth rate

② $$y_t = \beta_0 + \beta_1 t + \beta_2 (t - c)_+ + \varepsilon_t$$

grid of length

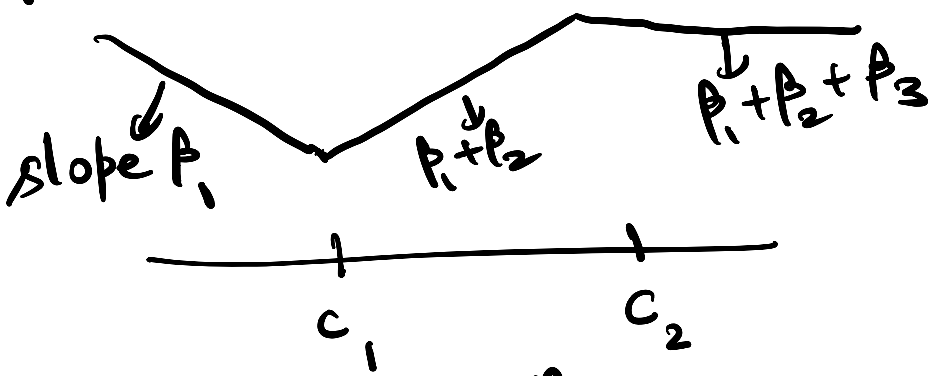
$t \le c$: slope β_1

$t > c$: slope $\beta_1 + \underbrace{\beta_2}$

length 1000 for c

$\hat{c}$, β_1, $\beta_1 + \beta_2$

③ $y_t = \beta_0 + \beta_1 t + \beta_2 (t - c_1)_+ + \beta_3 (t - c_2)_+ + \varepsilon_t$



$$RSS(c_1, c_2) = \min_{\beta_0, \beta_1, \beta_2, \beta_3} \sum_{t=1}^{n} \left[y_t - \beta_0 - \beta_1 t - \beta_2 (t - c_1)_+ - \beta_3 (t - c_2)_+ \right]^2$$

$$(\hat{c}_1, \hat{c}_2) = \min_{c_1, c_2} RSS(c_1, c_2)$$

c_1, c_2 } both on grid of length 1000

Alternative way of fitting two (or more) changes of slope model:

$$RSS(\underbrace{\beta_0, \beta_1, \beta_2, \beta_3, c_1}, c_2) = \underbrace{\sum_{t=1}^{n} \left[y_t - \beta_0 - \beta_1 t - \beta_2 (t - c_1)_+ - \beta_3 (t - c_2)_+ \right]^2}$$

↳ Put this in some optimization software
→ PyTorch.

$$\left. \begin{aligned} y_t = \beta_0 + \beta_1 t + \beta_2 (t - c_1)_+ + \beta_3 (t - c_2)_+ + \\ + \dots + \beta_{k+1} (t - c_k)_+ + \varepsilon_t \end{aligned} \right\}$$

① How to choose k?

② Estimation of parameters (??)

③ $\boxed{k \text{ large then there will be overfitting.}}$

We create a high-dimensional model by using all possible choices of $c_1, c_2, \dots$

$$y_t = \beta_0 + \beta_1 t + \beta_2 (t - c_1)_+ + \beta_3 (t - c_2)_+ + \beta_4 (t - c_3)_+ + \dots$$

$t = 1, \dots, n$

$$(t - 5.7)_+ = (0.3)(t-5)_+ + (0.7)(t-6)_+$$

for all $t \leq 5$
& all $t \geq 6$

If $t \geq 6$,
$lhs = t - 5.7$, $rhs = (0.3)(t-5) + (0.7)(t-6)$
$= t - 5.7$

We shall consider:

$$M_1: \quad y_t = \beta_0 + \beta_1 t + \beta_2 (t-2)_+ + \beta_3 (t-3)_+ + \beta_4 (t-4)_+ + \dots + \beta_{n-1} (t-(n-1))_+ + \varepsilon_t$$

$$\Updownarrow$$

$$M_2: \quad y_t = \beta_0 + \beta_1 t + \beta_2 (t-c_1)_+ + \beta_3 (t-c_2)_+ + \dots + \beta_k (t-c_k)_+ + \varepsilon_t$$

1. M_1: knots are fixed at $2, 3, \dots n-1$

   M_2: knots are also unknown parameters

2. M_2 uses k (number of knots) that needs to be determined

3. M_1: Linear Regression $\longrightarrow$ HIGH DIMENSIONAL (# parameters ~ n)

   M_2: Nonlinear Regression $\longrightarrow$ LOW DIMENSIONAL

How to do Inference in M_1?

$$\sum_{t=1}^{n} \left[y_t - \beta_0 - \beta_1 t - \beta_2 (t-2)_+ - \dots - \beta_{n-1} (t-(n-1))_+ \right]^2$$

Slight change in notation for M_1:

$$\beta_0 + \beta_1(t-1) + \underset{\sim}{\beta_2}(t-2)_+ + \underset{\sim}{\beta_3}(t-3)_+ + \cdots + \underset{\sim}{\beta_{n-1}}(t-\widetilde{n-1})_+$$
$$= y_t \,,\quad t = 1,\ldots,n$$

$t=1 \rightarrow \beta_0 = y_1$

$t=2 \rightarrow \beta_0 + \beta_1 = y_2 \Rightarrow \beta_1 = y_2 - y_1$

$t=3 \rightarrow \beta_0 + 2\beta_1 + \beta_2 = y_3 \overset{\text{CHECK}}{\Longrightarrow} \beta_2 = (y_3 - y_2) - (y_2 - y_1)$

More generally,

$$\beta_j = (y_{j+1} - y_j) - (y_j - y_{j-1})$$

$j \geq 2$

$y_t = \log P_t \rightarrow$ CA population for year t.

β_0 = Population in year 1 on log scale $= y_1$

$$\beta_1 = y_2 - y_1 = \log \frac{P_2}{P_1} \approx \frac{P_2 - P_1}{P_1}$$

$100 \times \beta_1$ = percent change in Population from year 1 to year 2

$$\beta_2 = (y_3 - y_2) - (y_2 - y_1)$$

$100 \times \beta_2 =$ change in percent change between years 2 to 3 & 1 to 2.

Use this model but change estimation strategy.

① Frequentist → Regularized MLE

② Bayesian → Change the prior.

$$\boxed{\beta_0, \beta_1, \ldots, \beta_m \overset{iid}{\sim} \text{Unif}(-C, C)}$$

change this