

Dependence: ...

$$① (Y, \underline{X}) = (Y, X_1, X_2, \dots, X_k)$$

residual

$$Y = E(Y) + (Y - E(Y))$$

$$Y = E(Y) + \varepsilon$$

"erro" ou "resíduo" a média

$$(Y|X) \sim f(Y|X)$$

Queremos ~~pre~~ prever Y a partir do conhecimento das variáveis (features) $\underline{X}$.

→ Precisamos conhecer $(Y|X) \sim f(Y|X)$.

→ Como especificar isto para toda configuração $\underline{X}$ e todo valor Y ??

$Y =$ Preço de ap - ^{to}

$\underline{X} = (\text{área}, \text{n.º quartos}, \text{n.º suites}, \dots, \text{piscina?})$
 $\quad \quad \quad X_1 \quad \quad \quad X_2 \quad \quad \quad X_3 \quad \quad \quad X_{30}$

Precisamos da distribuição

$(Y | X_1 = 250 \text{ m}^2, X_2 = 2 \text{ q.ºs}, X_3 = 3 \text{ suites}, \dots)$

mas tb. para

$(Y | X_1 = 120 \text{ m}^2, X_2 = 2 \text{ q.ºs}, X_3 = 1 \text{ suite}, \dots)$

e todas as outras ^{infinitas} possibilidades para $\underline{X}$.

Ao invés de obter a distribuição. ②

$f(y|x)$ para todo $\tilde{x}$, vamos

tentar algo mais simples;
prever Y a partir de $\tilde{x}$ condicionado
por saber $\tilde{x} = \tilde{x}$
uma função matemática $g(\tilde{x})$.
Com

$$g: \mathbb{R}^{30} \rightarrow \mathbb{R}$$

$$\tilde{x} \rightarrow g(\tilde{x}) \approx$$

tal que o erro

$$\varepsilon = Y - g(\tilde{x})$$

seja pequeno.

ε é var. ?? O que é aleatório em ε ?

~~Resposta: $\tilde{x}$ e Y~~

~~Mas queremos nos incomodar a variações de $\tilde{x}$. Queremos olhar para o erro fixando $\tilde{x}$~~

ε pode ser grande ou pequeno, positivo ou negativo. Queremos olhar tanto para o posit. quanto para o negat.

Vamos olhar o erro absoluto (3)
erro quadrático:

$$e^2 = (Y - g(X))^2$$

Qual é a função g que minimiza
este erro aleatório??

Impossível resolver assim: vamos

Resolver o erro aleatório num

único número: seu valor esperado

Problema achar g que minimize

$$E(Y - g(X))^2$$

Esperança com resp a uma dist. O que
é aleatório aqui?

$$E(Y - g(X))^2 = \int (y - g(x))^2 \underbrace{f(x, y)}_{\text{joint density}} dx dy$$

$$= \int \underbrace{\left[\int (y - g(x))^2 \underbrace{f(y|x)}_{\text{conditional density}} f(x) dx \right]}_{\text{marginal density of } x} dx$$

Fixado $\underline{x}$ Variação mínima

(4)

$$\int (y - g(\underline{x}))^2 f(y | \underline{x}) dy$$

$$= E_{(Y | \underline{X} = \underline{x})} \left[(Y - g(\underline{x}))^2 \right]$$

Qual é g ??

Chame $\mu(\underline{x}) = E(Y | \underline{X} = \underline{x})$

$$\text{Então } E_{(Y | \underline{X} = \underline{x})} (Y - g(\underline{x}))^2 = E_{(Y | \underline{X} = \underline{x})} \left[(Y - \mu(\underline{x}) + \mu(\underline{x}) - g(\underline{x}))^2 \right] =$$

$$= E \left[(Y - \mu(\underline{x}))^2 + 2(Y - \mu(\underline{x}))(\mu(\underline{x}) - g(\underline{x})) + (\mu(\underline{x}) - g(\underline{x}))^2 \right]$$

$$= E \left[(Y - \mu)^2 \right] + 2 E \left[(Y - \mu)(\mu - g(\underline{x})) \right] + E \left[(\mu - g(\underline{x}))^2 \right]$$

$$= \text{Var}(Y | \underline{X} = \underline{x}) + 2(\mu(\underline{x}) - g(\underline{x})) \underbrace{E[(Y - \mu)]}_{=0} + (\mu(\underline{x}) - g(\underline{x}))^2$$

$$= \text{Var}(Y | \underline{X} = \underline{x}) + \underbrace{(\mu(\underline{x}) - g(\underline{x}))^2}_{\geq 0 \text{ para todo } g(\underline{x})} \quad \text{e} = 0 \text{ se } g(\underline{x}) = \mu(\underline{x})$$

(5)

Exercício

Problema Que a func $g(\tilde{x})$ que
minimiza o erro (quadrático) de
predição $E(Y - g(\tilde{x}))^2$?

$$E(g(\tilde{x}) - \mu(\tilde{x}))^2 = E(Y - g(\tilde{x}))^2$$

Fantásticos: Nosso objetivo ao buscar
as melhores minimizar o erro de predição
deve ser sempre encontrar uma
boa aproximação para $E(Y|\tilde{x})$.

Seremos
Amostra gigante: queremos predições
de Y para $\tilde{x}$ fixado em
 $x_1 = 250 m^2$, $x_2 = 39$ ^{ros}, $x_3 = 3$ milte. —

pegue todos os itens de sua base que
estaticamente tenham estas features. ~~Elegeremos essas variáveis.~~
média aritmética desses valores.
Isto será uma boa aproximação

para $E(Y|\tilde{x} = \tilde{x}_i)$.

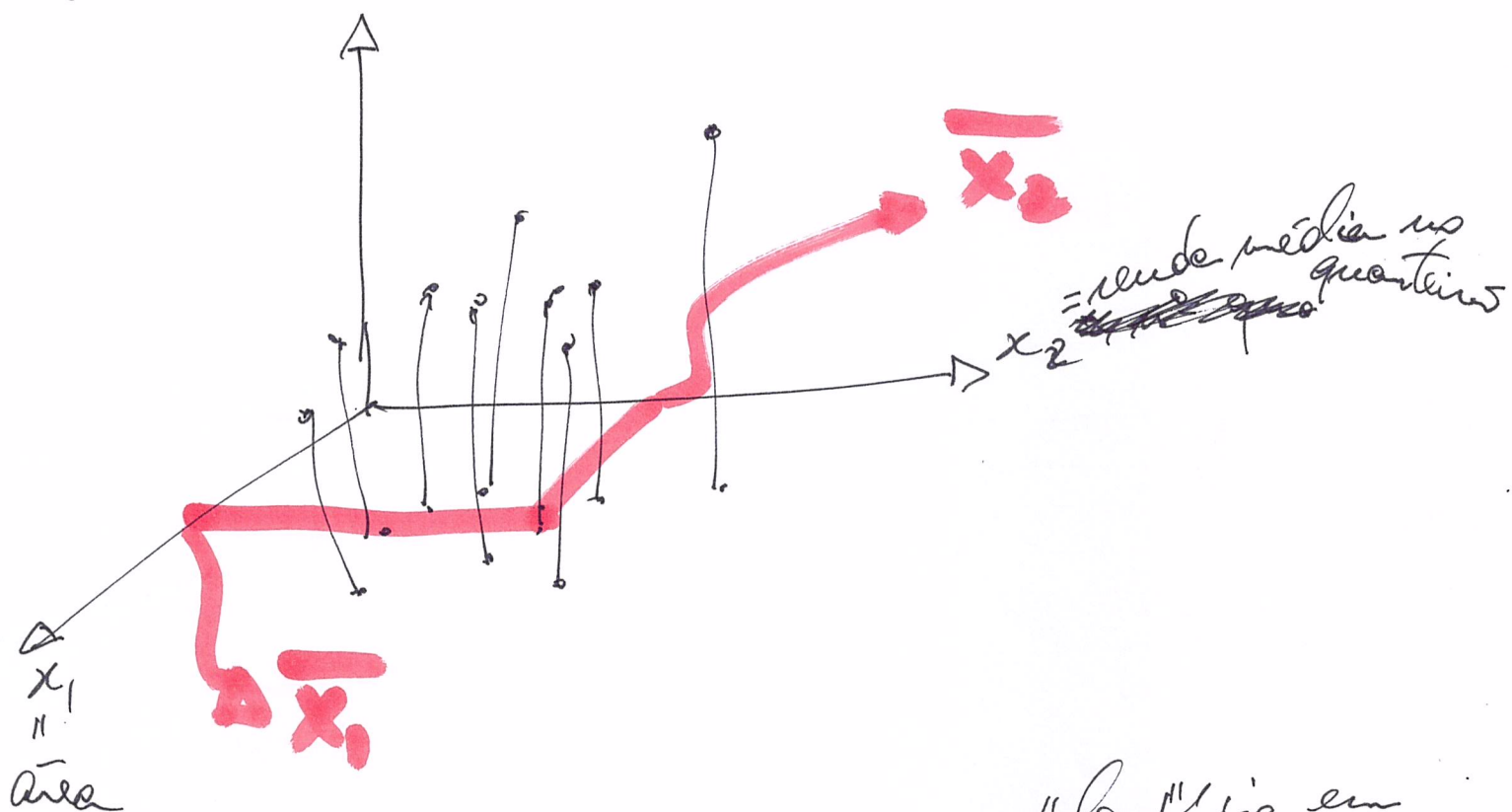
Trabalho, pode não ter nenhuma ou poucas x_i
com estaticamente as características x_i .

Vamos imaginar que

(6)

$$\mu(\underline{x}) = E(Y | \underline{X} = \underline{x})$$

seja uma função contínua das variáveis $\underline{x}$. Imagine que temos apenas duas variáveis em $\underline{x} = (x_1, x_2)$.



Imaginamos

$$(Y | \underline{X} = \underline{x}) = \underbrace{E(Y | \underline{X} = \underline{x})}_{\mu(\underline{x})} + \epsilon$$

"erro" em relac a $\mu(\underline{x})$

$$(Y | \underline{X} = \underline{x}) = \underbrace{\mu(\underline{x})}_{\text{fixado } \underline{x}, \text{ é uma constante}} + \underbrace{\epsilon}_{\text{aleatório}}$$

no é aleatório apenas uma função matemática de $\underline{x}$

Fórmula de Taylor :

(7)

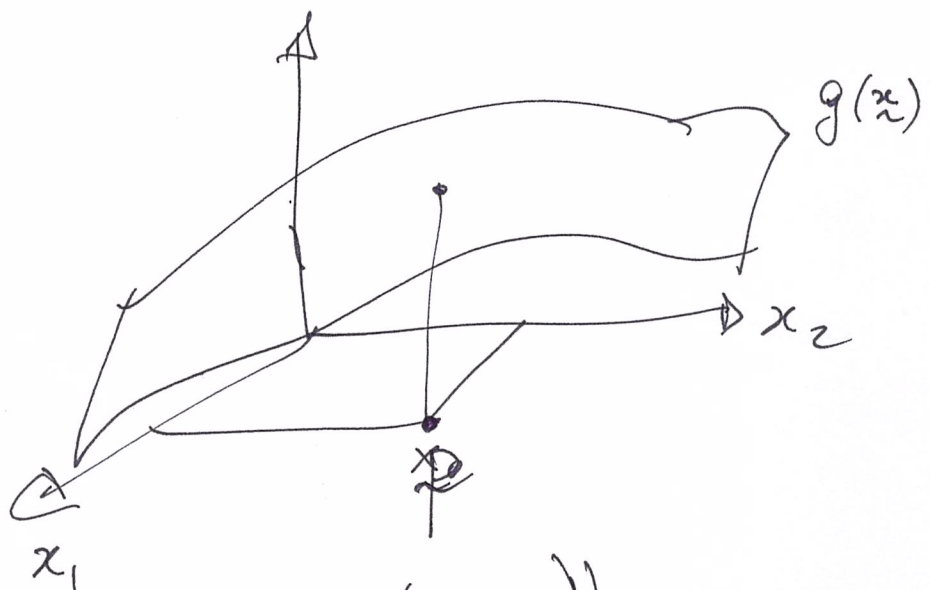
$$g: \mathbb{R}^{30} \rightarrow \mathbb{R}$$

~~g(x)~~

$$x \rightarrow g(x)$$

best

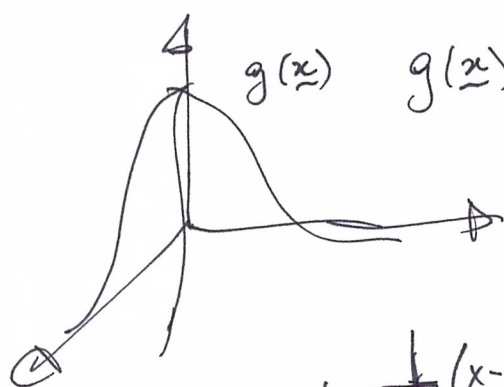
Aproximar $g(x)$ por um polinômio multivariado em torno de um ponto



$$g(x) \approx g(p) + \left(\nabla g \right)_p \cdot (x - p) + \frac{1}{2} (x - p)^t \left(D^2 g \right)_p (x - p) + \dots$$

(Hessian)

Por exemplo, $f(x) = e$ em torno de $p = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$



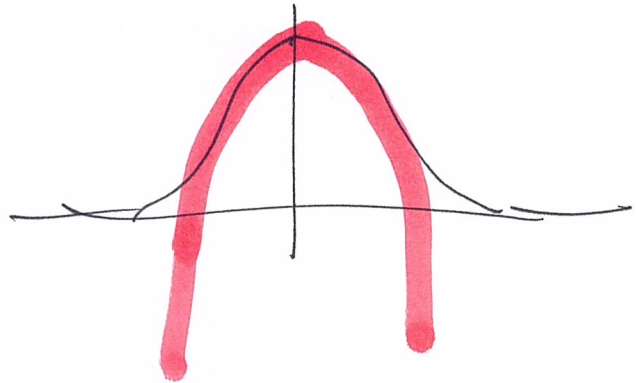
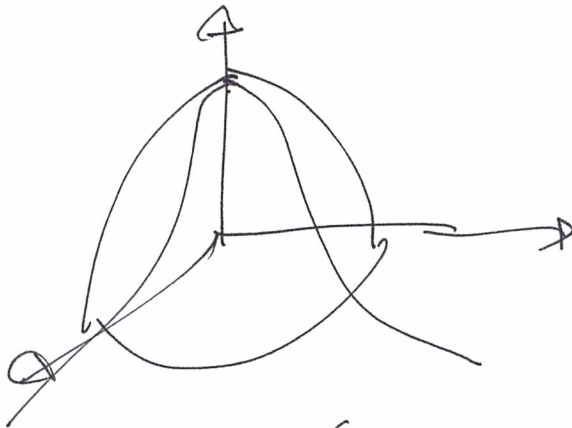
$$g(x) \approx e - (0+0) + \begin{pmatrix} -2xe^{-(x^2+y^2)} \\ -2ye^{-(x^2+y^2)} \end{pmatrix} \bigg|_{p=\begin{pmatrix} 0 \\ 0 \end{pmatrix}} \cdot \begin{pmatrix} x-0 \\ y-0 \end{pmatrix} + \frac{1}{2} (x-0, y-0) \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix} \bigg|_{p=\begin{pmatrix} 0 \\ 0 \end{pmatrix}} \cdot \begin{pmatrix} x-0 \\ y-0 \end{pmatrix}$$

$$g(\underline{x}) = 1 - (x^2 + y^2)$$

(8)

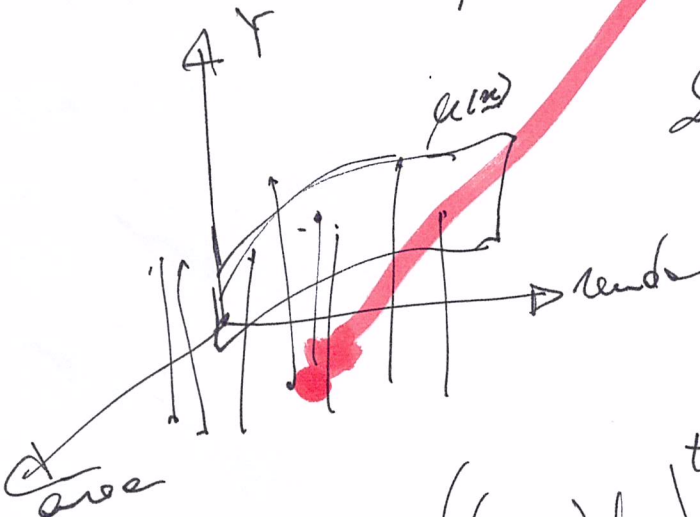
$$\approx 1 + (0, 0) \begin{pmatrix} x-0 \\ y-0 \end{pmatrix} + \frac{1}{2} (x \ y) \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

$$= 1 - x^2 - y^2$$



Back Aproximam
 $\mu(\underline{x}) = E(Y | \underline{x} = \underline{x})$ por Taylor

em torno do ponto $\underline{\bar{x}} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{30})$



Suponha que
 $\underline{\bar{x}} = \left(\underbrace{122 \text{ m}^2}_{\text{área}}, \underbrace{5 \text{ K}}_{\text{renda}} \right)$

$$\mu(\underline{x}) \approx \underbrace{\mu(\underline{\bar{x}})}_{\omega_0} + \left(\underbrace{\nabla \mu}_{\omega_1} \right) \Big|_{\underline{\bar{x}}} \cdot \begin{pmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \end{pmatrix} + \frac{1}{2} (\underline{x} - \underline{\bar{x}})^t \left(\underbrace{\nabla^2 \mu}_{\omega_2} \right) \Big|_{\underline{\bar{x}}} \cdot (\underline{x} - \underline{\bar{x}})$$

$$\omega_0 + \left(\omega_1 \right)^t \cdot \begin{pmatrix} \text{área} - 122 \\ \text{renda} - 5K \end{pmatrix} + \dots$$

$$\mu(\underline{x}) \approx w_0 + w_1(\text{área} - 122) + w_2(\text{renda} - 5K) \quad (9)$$

$$\approx \underbrace{(w_0 + 122w_1 - 5Kw_2)}_{\beta_0} + \underbrace{w_1}_{\frac{\partial \mu(\underline{x})}{\partial x_1}} \text{área} + \underbrace{w_2}_{\frac{\partial \mu}{\partial x_2}} \text{renda}$$

$$\mu(\underline{x}) = E(Y | \underline{x} = \underline{x}) \approx \beta_0 + w_1 \left(\frac{\text{área}}{x_1} \right) + w_2 \left(\frac{\text{renda}}{x_2} \right)$$

E se tirássemos mais a expansão de Taylor até a 2ª ordem ??

$$\begin{aligned} \mu(\underline{x}) &= E(Y | \underline{x} = \underline{x}) \\ &\approx \beta_0 + w_1 \left(\frac{\text{área}}{x_1} \right) + w_2 \left(\frac{\text{renda}}{x_2} \right) + w_3 \left(\frac{\text{área}^2}{x_1^2} \right) + w_4 \left(\frac{\text{renda}^2}{x_2^2} \right) \\ &\quad + w_5 \left(\frac{\text{área} * \text{renda}}{x_1 * x_2} \right) \end{aligned}$$

(termos de interação entre x_1 e x_2)

⊕ Assim, o modelo de regressão linear pode ser visto como a aproximação de Taylor ^{de 1ª ordem} para o best predictor $\mu(x) = E(Y|X=x)$.

⊕ Imagine que a lista de variáveis importante é enorme.

$$\underline{x} = (x_1, x_2, \dots, x_{523})$$

⊕ Teríamos $Y = \mu(\underline{x}) + \underbrace{(Y - \mu(\underline{x}))}_{\text{v.a.} + \text{desvio}}, \varepsilon$

$$\mu(\underline{x}) = E(Y|X=\underline{x})$$

$$\approx w_0 + \beta_1 x_1 + \dots + \beta_{30} x_{30} + \beta_{31} x_{31} + \dots + \beta_{523} x_{523}$$

$$Y = \mu(\underline{x}) + \varepsilon$$

$$\approx w_0 + \beta_1 x_1 + \dots + \beta_{30} x_{30} + \beta_{31} x_{31} + \dots + \beta_{523} x_{523}$$

$$+ \cancel{\phi_1 x_1^2} + \dots + \cancel{\phi_{523} x_{523} * x_1} + \dots + \underline{\varepsilon}$$

apenas 30 variáveis

$\beta_{31} = \frac{\partial \mu}{\partial x_{31}}$ $\approx$ como o valor esperado de Y muda qdo x_{31} passa de x_{31} para $x_{31} + 1$

Se passarmos para $x_{31} + 5$, o preço esperado muda em $5 * \beta_{31}$

⊕ Uma variável X pouco importante ⑪
 terá $\beta \approx 0$: o peso praticamente
 é nulo qdo alteramos X .

⊕ Pegue todos as variáveis pouco
 importantes e ~~some com~~
~~some com~~ ~~para~~ o
 desvio:

$$Y = \mu(x) + \varepsilon$$

$$\approx w_0 + \beta_1 X_1 + \dots + \beta_{30} X_{30} + \underbrace{(\beta_{31} X_{31} + \dots)}_{\text{novo desvio } \varepsilon^*} + \varepsilon$$

Assumamos que os desvios ε^* oscilam
 em torno de zero de forma gaussiana)
 $\varepsilon^* \sim N(0, \sigma^2)$

Chegamos então ao modelo
 de regressão linear (na a prox.)

$$E(Y|X=x) = \underbrace{(1 \ x_1 \ x_2 \ \dots \ x_{30})}_{E(Y|X=x)} \cdot \underbrace{\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{30} \end{pmatrix}}_{\substack{\beta_0 \\ \vdots \\ \beta_{30}}} + \underbrace{\varepsilon^*}_{N(0, \sigma^2)}$$

Em notação matricial:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_{5000} \end{pmatrix} = \underset{\sim}{Y} = \begin{pmatrix} 1 & X_{11} & \dots & X_{30,30} \\ \vdots & \vdots & & \vdots \\ 1 & X_{5000,1} & \dots & X_{5000,30} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{30} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{5000} \end{pmatrix}$$

$\underbrace{\begin{pmatrix} 1 & X_{11} & \dots & X_{30,30} \\ \vdots & \vdots & & \vdots \\ 1 & X_{5000,1} & \dots & X_{5000,30} \end{pmatrix}}_{\substack{\downarrow \\ \text{areas}}}$
risco?

$$\begin{pmatrix} Y | X=x \\ \sim \end{pmatrix} = \underbrace{X}_{5000 \times 31} \cdot \underbrace{\beta}_{31 \times 1} + \underbrace{\epsilon}_{5000 \times 1}$$

Relembre:
 $\underset{\sim}{Y} \sim N_p(\mu, \Sigma)$
 $A \underset{\sim}{Y} \sim N_K(A\mu, A\Sigma A^T)$
 $K \times p$ $p \times 1$

$$\underset{\sim}{\epsilon} \sim N_{5000}(\mathbf{0}, \sigma^2 I)$$

Assim, $(Y|X=x) \sim N_{5000}(X\beta, \sigma^2 I)$

⊕ Como encontrar a aproximação para β ?

⊕ Mínimos quadrados:

Ache β que minimiza $\|\underset{\sim}{Y} - X\beta\|^2$.

Função de custo: $J(\beta) = \|\underset{\sim}{Y} - X\beta\|^2 = \sum_{i=1}^n (y_i - x_i' \beta)^2$

fixo $\rightarrow$ com uma Alg. linear

Já encontramos esta solução (sem cálculos)

$\hat{\beta} = (X'X)^{-1} X' \underset{\sim}{Y}$ e o vetor predito

$$\hat{\underset{\sim}{Y}} = X \hat{\beta} = X (X'X)^{-1} X' \underset{\sim}{Y} = H \underset{\sim}{Y}$$

$$\hat{\beta} = \underbrace{(X'X)^{-1}X'}_A Y$$

↓
 $N_{5000}(X\beta, \sigma^2 I)$

(13)

$$\hat{\beta} = N_{5000} \left(\underbrace{(X'X)^{-1}X'(X\beta)}_{\beta}, \underbrace{A(\sigma^2 I)A'}_{\sigma^2 (X'X)^{-1}} \right)$$

$\sigma^2 (X'X)^{-1}$
 $\sigma^2 \alpha I$

not needed
para estimar β