Lecture 2 - April 9th, 2025

LINEAR REGRESSION

a tool for understanding relationship between variables:

X (independent variable, covariate, input, feature.) and Y (response, dependent variable, outcome)

Model:

$$Y \approx \beta_0 + \beta_1 X$$

intercept slope

"the regression coefficients"
"the regression parameters"

More formally,

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

the "error."

ε is a random variable where $\mathbb{E}[\varepsilon]$ (ε is mean zero.)

This is a special case of the model:

$$Y = f(X) + \varepsilon$$

Linear regression is "parametric" because f depends on parameters (β_0, β_1).

Often, we will use data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ to obtain

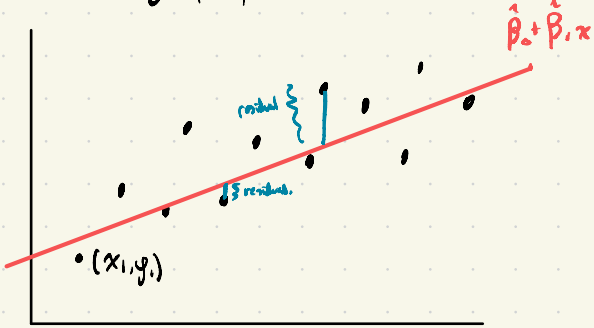
what we think are "good" estimates of β_0 and β_1

The estimates are $\hat{\beta}_0$ and $\hat{\beta}_1$, this gives rise to the fitted model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

<u>Definition:</u> a "residual" is the value $e_i := y_i - \hat{y}_i$

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$
$$= \hat{\beta}_0 + \hat{\beta}_1 x_i + y_i - \hat{y}_i$$
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$



<u>Definition:</u> the residual sum of squares:

$$RSS := e_1^2 + e_2^2 + \dots + e_n^2$$
$$= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

To fit the linear model, pick:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{\operatorname{argmin}} \ RSS$$

$$= \underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

We can use calculus to get a more explicit formula:

let's minimize $g(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$.

$$\frac{\partial g}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$0 = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$\Longrightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\beta}_1 x_i$$

$$\Longrightarrow \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i = n \hat{\beta}_0$$

$$\Longrightarrow \hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$= \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\frac{\partial g}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

$$0 = \left. \frac{\partial g}{\partial \beta_1} \right|_{\beta_1 = \hat{\beta}_1} \Longrightarrow 0 = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$\Longrightarrow \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} \hat{\beta}_0 x_i + \sum_{i=1}^{n} \hat{\beta}_1 x_i^2$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i y_i - \hat{\beta}_0 x_i)}{\sum_{i=1}^{n} x_i^2}$$

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

Q: what if I minimized:

$$\sum_{i=1}^{n} |e_i| \quad \underline{\text{OR}} \quad \sum_{i=1}^{n} |e_i|^3 \quad \underline{\text{OR}} \quad \sum_{i=1}^{n} e_i^4 \quad ?$$

We can do statistical tests like:

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

MULTIPLE LINEAR REGRESSION

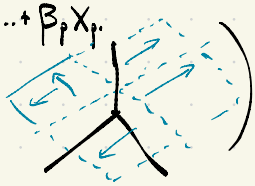
now we will have <u>multiple covariates</u>

<u>notation:</u> $X = (X_1, X_2, \ldots, X_p)$

Now, in $Y = f(X) + \varepsilon$, we have

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

(Now, $f(X)$ is a "hyperplane":


)

We'll still minimize

$$RSS = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)\right)^2$$

$$= \| \underset{n\times 1}{y} - \underset{n\times(p+1)}{X} \underset{(p+1)\times 1}{\beta} \|_2^2$$

where:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{13} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & & X_{np} \end{bmatrix}$$

helpful to remember: X_{ij} = i^{th} observation's value of the jth variable.

Recall: $\| z \|_2^2 = \sum_{i=1}^{\dim(z)} z_i^2$

Let's do the matrix multiplication:

$$X\beta = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{13} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \hat{y}$$

$$1 \cdot \beta_0 + X_{11} \cdot \beta_1 + X_{12} \cdot \beta_2 + \dots + X_{1p} \cdot \beta_p$$

$$= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \dots + \beta_p X_{1p}$$

$$= \hat{y}_1$$

$$\Longrightarrow \quad RSS = \| y - X\beta \|_2^2 = \| y - \hat{y} \|_2^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Now: we fit our model with:

$$\underset{(n\times 1)}{\hat{\beta}} = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \|y - X\beta\|_2^2$$

ASIDE: look up the "matrix cook book" online for help with these rules

to minimize: for $g(\beta) = \|y - X\beta\|_2^2$,

we have $\nabla_\beta g(\beta) = \begin{bmatrix} \frac{\partial g}{\partial \beta_0} \\ \frac{\partial g}{\partial \beta_1} \\ \vdots \\ \frac{\partial g}{\partial \beta_p} \end{bmatrix} = -2X^T(y - X\beta)$

Solve for critical value:

$$0 = -2X^T(y - X\hat{\beta}) \Longrightarrow X^T X \hat{\beta} = X^T y$$

$$\Longrightarrow (X^T X)^{-1}(X^T X)\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\Longrightarrow \boxed{\hat{\beta} = (X^T X)^{-1} X^T y}$$

CATEGORICAL VARIABLES

what if X is a "catogory."

A, B, AB, O, A+

e.g., health outcome $\approx \beta_0 + \beta_1$ (blood type)

(Y) X

IDEA: A $\longrightarrow$ 1 AB $\longrightarrow$ 3

B $\longrightarrow$ 2 O $\longrightarrow$ 4

IDEA: use "dummy variables"

$$A \longrightarrow (0, 0, 0)$$
$$B \longrightarrow (0, 0, 1)$$
$$AB \longrightarrow (0, 1, 0)$$
$$O \longrightarrow (1, 0, 0)$$

$A \to (0,0)$
$B \to (0,1)$
$AB \to (1,0)$
$O \to (1,1)$

if you have K categories, then it takes $K-1$ dummy variables to encode them.

$$Y_i = \beta_0 + \beta_1 \mathbb{1}\{i\text{th obs is } B\} + \beta_2 \mathbb{1}\{i\text{th obs is } AB\}$$
$$+ \beta_3 \mathbb{1}\{i^{th} \text{ obs is } O\} + \varepsilon_i$$

Q: what about:

$$Y_i = \tilde{\beta}_1 \mathbb{1}\{i^{th} \text{ obs is } A\} + \tilde{\beta}_2 \mathbb{1}\{i^{th} \text{ is } B\}$$
$$+ \tilde{\beta}_3 \mathbb{1}\{i\text{th obs is } AB\} + \tilde{\beta}_4 \mathbb{1}\{i^{th} \text{ obs is } O\} + \varepsilon_i$$

$\longrightarrow$ $\tilde{\beta}_2 = \beta_0 + \beta_1$, $\tilde{\beta}_3 = \beta_0 + \beta_2$, $\tilde{\beta}_4 = \beta_0 + \beta_3$

$\tilde{\beta}_1 = \beta_0$

what if we encode as:

$$A \longrightarrow (0, 0, 0, 1)$$
$$B \longrightarrow (0, 0, 1, 0)$$
$$C \longrightarrow (0, 1, 0, 0)$$
$$D \longrightarrow (1, 0, 0, 0)$$

$\Rightarrow$ coefficients not unique anymore.

if $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \mathbb{1}\{i^{th} \text{ is } A\} + \hat{\beta}_2 \mathbb{1}\{i^{th} \text{ is } B\}$
$+ \hat{\beta}_3 \mathbb{1}\{i^{th} \text{ is } AB\} + \hat{\beta}_4 \mathbb{1}\{i^{th} \text{ is } O\}$

$\hat{\beta}_0 = 3$		$\hat{\beta}_0 = 2$
$\hat{\beta}_1 = 11$	SAME AS	$\hat{\beta}_1 = 12$
$\hat{\beta}_2 = 6$	$\longleftrightarrow$	$\hat{\beta}_2 = 7$
$\hat{\beta}_3 = 8$		$\hat{\beta}_3 = 9$
$\hat{\beta}_4 = 2$		$\hat{\beta}_4 = 3$

<u>problem:</u> remember $\hat{\beta} = \underbrace{(X^T X)}_{\text{not invertible now.}}{}^{-1} X^T Y$

INCREASING FLEXIBILITY OF LINEAR MODELS

<u>Additive models:</u> $f(X) = f_1(X_1) + f_2(X_2) + \dots + f_p(X_p)$

<u>Nonadditive:</u> $\tilde{f}(X) = g_1(X_1, X_2) + g_2(X_3)$

example of nonadditive: add "interaction terms"

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

$$= \beta_0 + \underbrace{(\beta_1 + \beta_3 X_2)}_{= \tilde{\beta}_1} X_1 + \beta_2 X_2 + \varepsilon$$

$$= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \varepsilon$$

if β_3 is large $\Longrightarrow$ $\begin{cases} \text{if } X_2 \text{ large,} \\ \quad X_1 \text{ has larger regression coefficient} \\ \text{if } X_2 \text{ small,} \\ \quad X_1 \text{ has smaller coefficient} \end{cases}$

if β_3 is small $\Longrightarrow$ X_1 has same "ish" regression coefficient regardless of the size of X_2.

Can increase flexibility w/ polynomial powers:

$$f_1(X) = \beta_0 + \beta_1 X$$

$$f_2(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

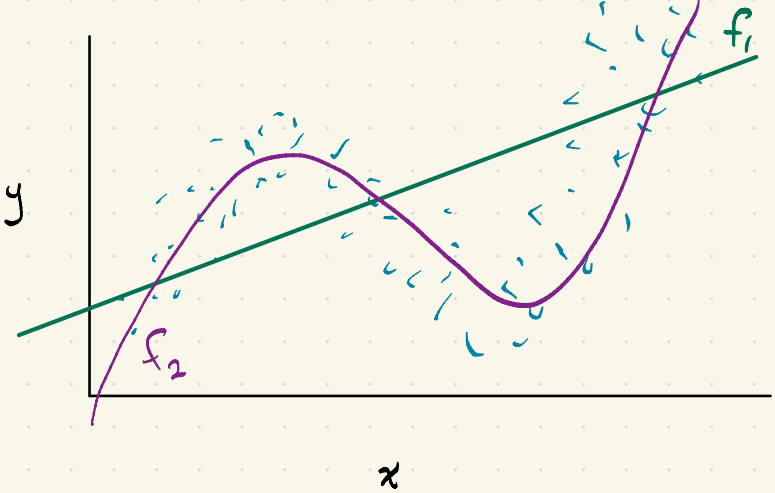
$$f_3(X) = \beta_0 + \beta_1 X + \beta_2 \sqrt{X} + \beta_3 \sin(X) + \beta_4 \exp(\tan(X^3))$$

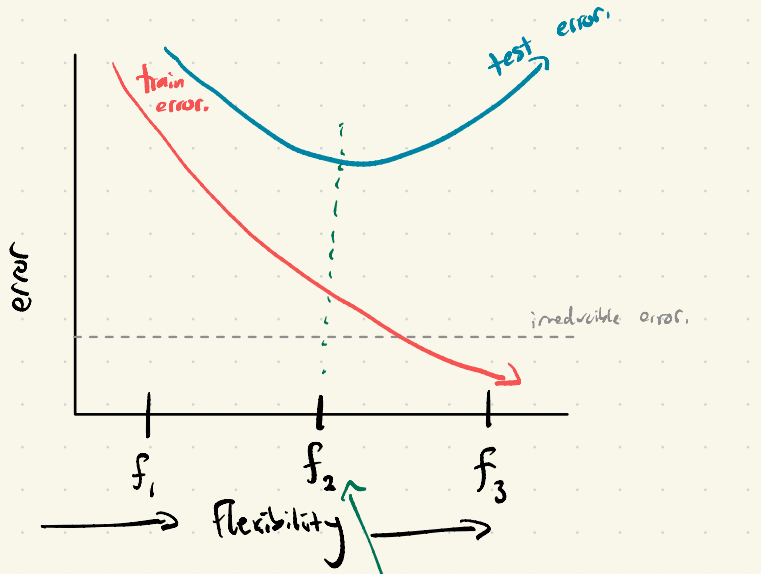
what makes a model linear?

a linear model is <u>linear in the coefficients.</u>

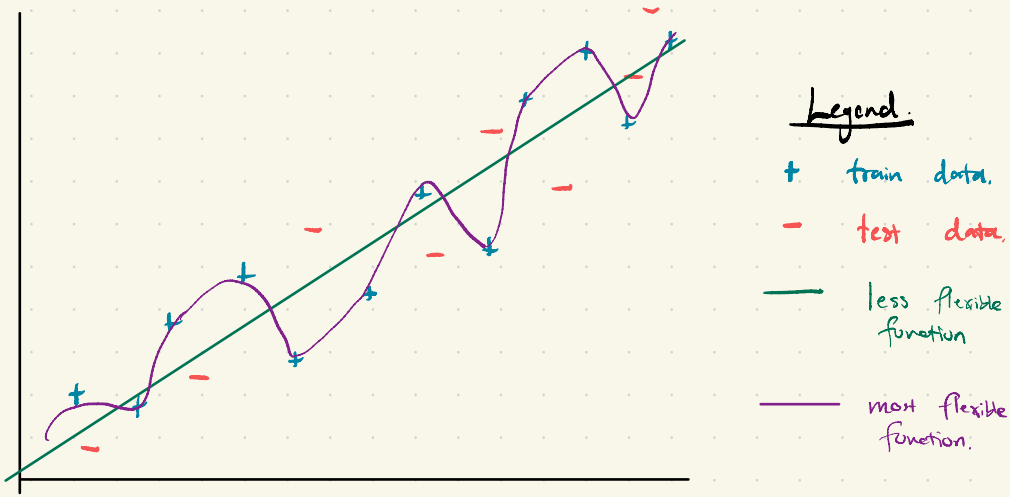
this would <u>not</u> be a linear model:

$$f_4(X) = \beta_0 + \sqrt{\beta_1 + \beta_2}\, X_1 + X_2^{\beta_3}$$





f_2 "best": it gets best error on test data.



take home message:

* adding more terms & polynomial powers in a linear model increases flexibility.

* there is a "sweet spot" of flexibility which allows your model to do best on unseen (test) data.

things that can go "wrong":

* many of these listed in Section 3.3.3 of ISL.

* one thing to think about is "multicollinearity."

e.g. X_i is correlated with X_j for $i \neq j$.

"multicollinearity" is when a linear combination of $\{X_1, \ldots, X_p\}$ is highly correlated with another linear combination of $\{X_1, \ldots, X_p\}$

e.g., $\mathrm{Cor}\left(2X_1 + 3X_2,\ X_4 - X_5 + \frac{1}{6}X_7\right)$ is high.

if $\text{weight(kg)} = \beta_0 + \beta_1(\text{height (cm)}) + \beta_2(\text{height(in)}) + \varepsilon$

here, $\mathrm{Cor}(\text{height(cm)}, \text{height(in)}) = 1$

Consider the following:

(a) $\hat{\beta}_0 = 17,\ \hat{\beta}_1 = 5,\ \hat{\beta}_2 = 0$

(b) $\hat{\beta}_0 = 17,\ \hat{\beta}_1 = 0,\ \hat{\beta}_2 = \frac{5}{2.54}$

(c) $\hat{\beta}_0 = 17,\ \hat{\beta}_1 = 10^6 + 5,\ \hat{\beta}_2 = \frac{-10^6}{2.54}$

How do I know if I have collinearity?

→ book talks about "VIF" (variance inflation factor.)
(on page 103 of textbook).

What do you do if there is collinearity?

(1) drop one of the variables correlated with others, repeat this until multicollinearity not a problem.

(2) combine the collinear variables:

e.g. replace X_1, X_2, X_3 with $\frac{X_1 + X_2 + X_3}{3}$

COMPARISON WITH K-NEAREST NEIGHBORS REGRESSION

Linear regression is parametric.

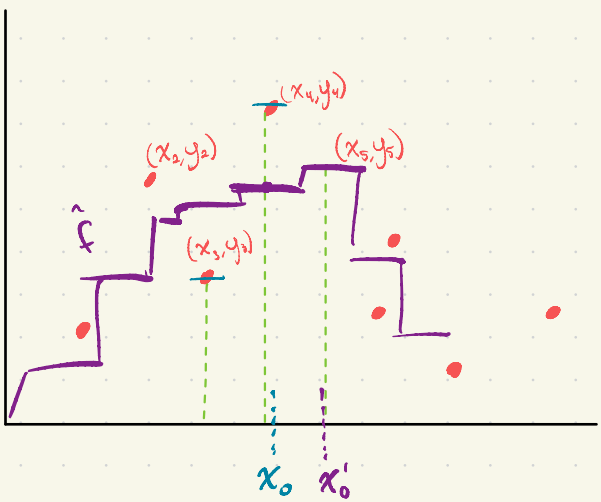
$$f(X) = \beta_0 + \beta_1 X$$

There also are nonparametric models. (we will not assume a super strict structure for $f(X)$).

<u>Example:</u> K-nearest neighbors regression.

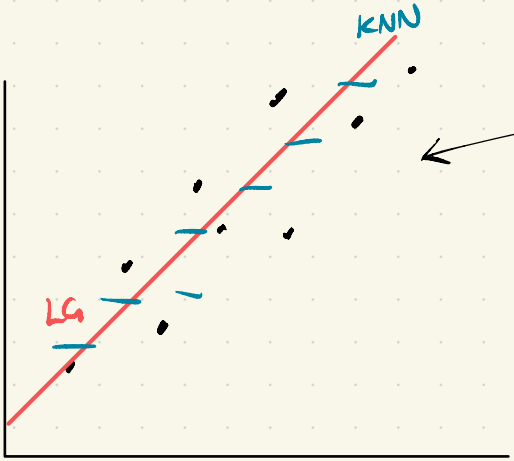
$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_K(x_0)} y_i$$

where K is specified by the user, and $N_K(x_0) := \{$ the closest K neighbors to $x_0 \}$.



$K = 3$

$$\hat{f}(x_0) = \frac{1}{3}(y_3 + y_4 + y_5)$$



better off w/ linear regression.

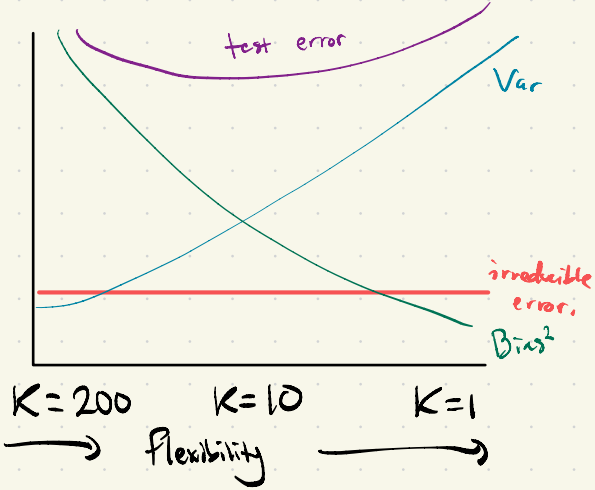
if you have knowledge of the structure of your data → parametric much better.

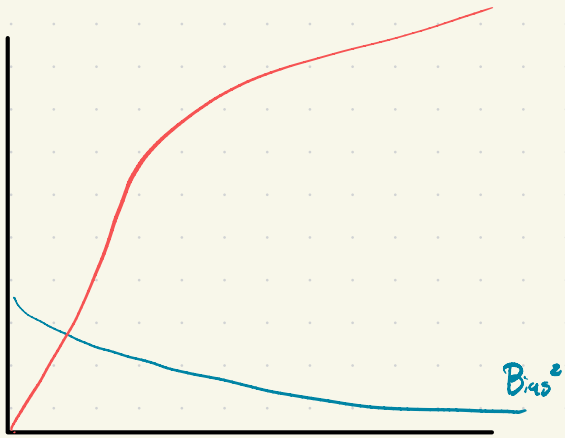
no knowledge → nonparametric much better.

---

little data → usually parametric better

<u>lots</u> of data → usually nonparametric better.


test error
Var
irreducible error.
Bias²
K=200
K=10
K=1
Flexibility


Bias²