MONDAY NOV 27

TODAY: DIAGNOSTIC PLOTS

FLEXIBLE MODELS

HW 9 DUE TOMORROW

General Fact:

Nice functions can be approximated by BASIS of

polynomials

splines

TRUE MODEL QUADRATIC, FIT LINEAR

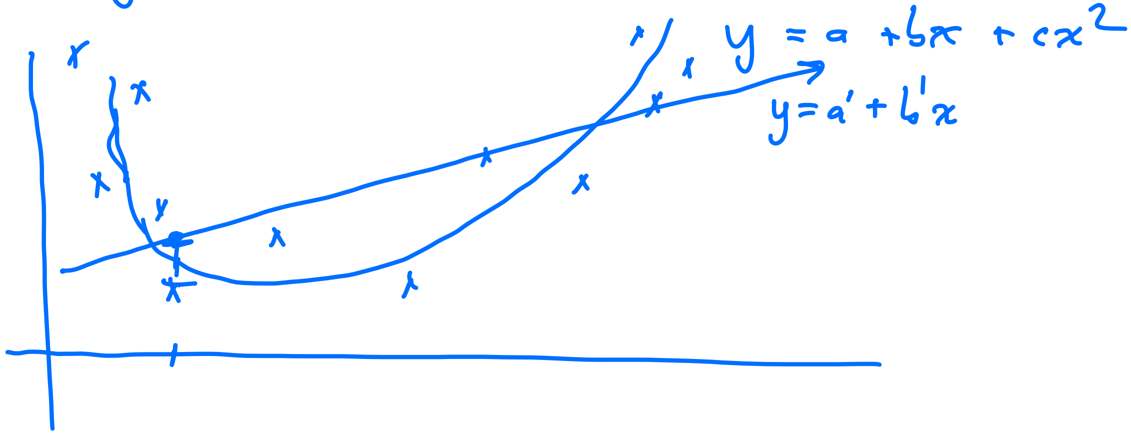
MSEP $\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$ estimate of σ^2

$\sqrt{\frac{1}{n}\sum (y_i - \hat{y}_i)^2}$ estimate of σ

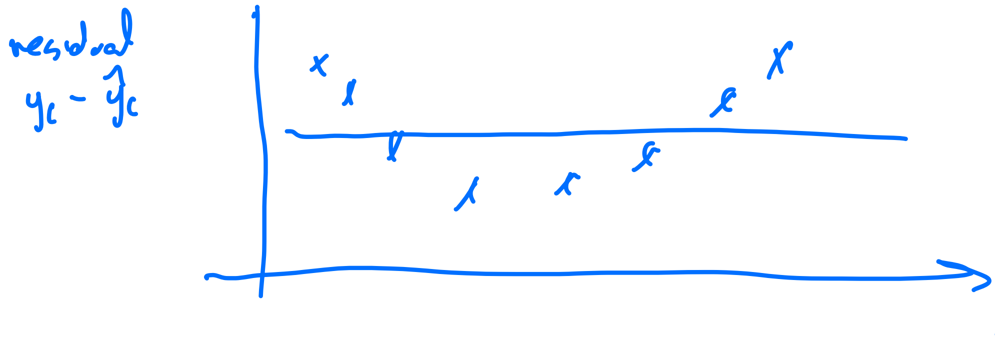
---

quadratic model

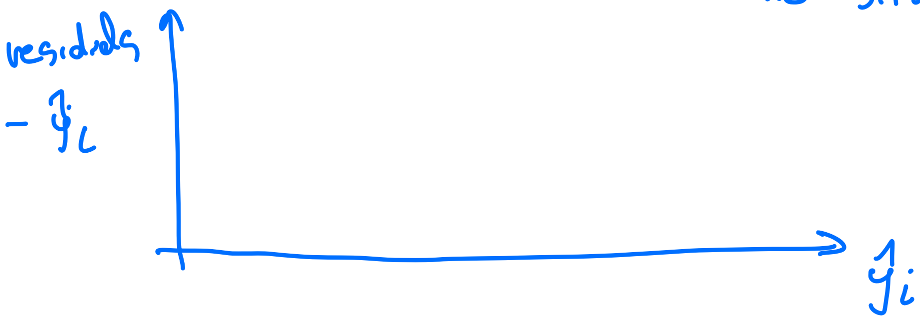
Truth $y_i = a + bx_i + cx_i^2 + \varepsilon_i$



Fit linear model: $y_i = a + bx_i + \varepsilon_i$



residuals vs fitted values : If model is adequate, no structure to residual plot



$$y_i = a + bx_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

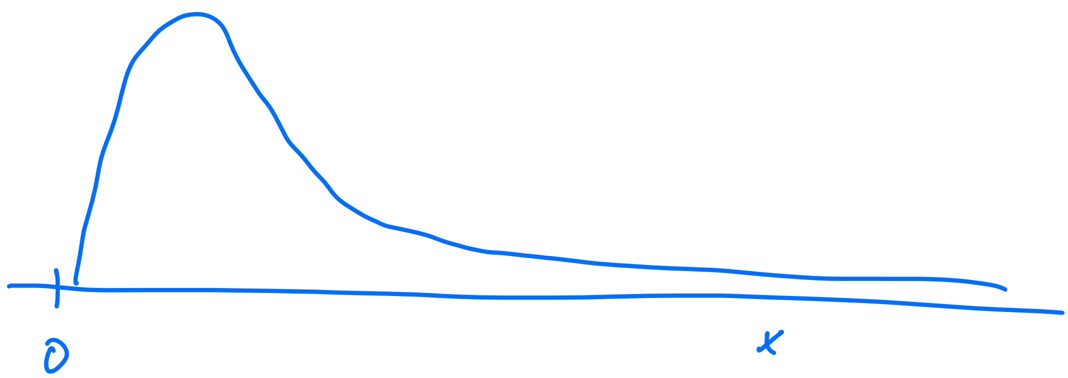
σ^2 is constant for all data points

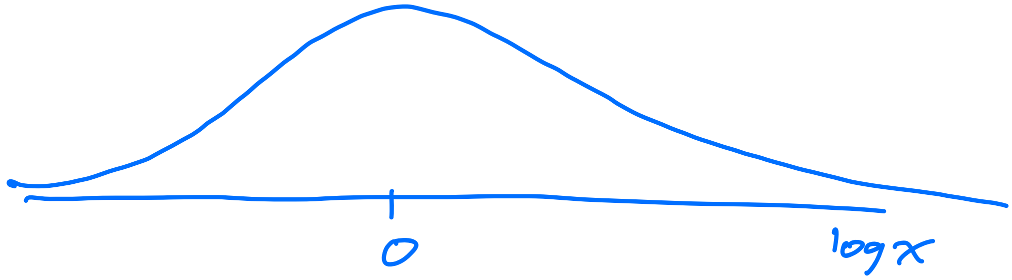
In this example

$$\hat{y}_i = \hat{a} + \hat{b} x_i$$

$$y_i - \hat{y}_i = a + bx_i + \varepsilon_i - (\hat{a} + \hat{b} x_i)$$

$$\simeq \varepsilon_i$$





Model

$$y_i = f(x_i) + \varepsilon_i \qquad \text{find } f$$

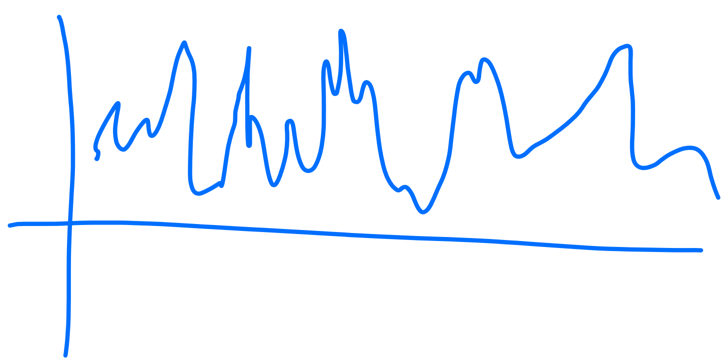
If f is smooth, can approx by

- polynomials
- sin's and cos's
- other families of simple functions
  (piecewise linear, piecewise quadratic)

Approximate $y_i = \sum_{j=1}^{K} \beta_j f_j(x_i) + \varepsilon_i$

how to choose k ($f_1, f_2, \ldots$ fixed basis of functions)

How to choose k? Cross-validation to pick optimal k



average 2nd derivative
measure of smoothness

Summarize

fit model $y_i = \sum_{j=1}^{K} \beta_j f_j(x_i) + \varepsilon_i$

$f_1(x), \ldots, f_K(x)$ fixed known functions

$x \qquad x^k$

$\sin(x), \cos(x), \sin(2x), \cos(2x), \ldots$

splines (piecewise polynomials) preferred family for mathematical reasons

Answered: how many k to use?

(1) USE CV to compare different choices

(2) USE penalty w/ fixed large K

↳ penalize too wiggly solutions

USE CV to determine weight given to penalty term!

Really no different than multiple linear regression w/ penalty term

here X matrix is determined by basis of functions

Here only one predictor x

But actually we have k predictors in the model by using the basis

$f_1(x), f_2(x), \ldots, f_k(x)$