

Linear Regression (Review)

assumptions

1. Linearity: $E[y] = x\beta$ linear combination of the predictors

2. independence: $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$
 $i \neq j$

3. Homoscedasticity: $\text{var}(y_j) = \text{var}(\varepsilon_j) = 0$

4. Normality: $\varepsilon_j \stackrel{iid}{\sim} N(0, \sigma^2)$

What if assumptions are not met? We can't use linear regression

Case study: • response is not Normal
• constant variance (homoscedasticity) is not met

If the response is Binomial

$$Y_i \sim \text{Bin}(n, p_i) \quad i = 1, \dots, n$$

1. PMF $P(Y_i = y_i) = \binom{n}{y_i} p_i^y (1-p_i)^{n-y}$

2. mean $E[Y_i] = \mu_i = np_i$ from definition of expectation and variance

3. variance $\text{var}(Y_i) = \sigma^2 = np_i(1-p_i)$ for discrete r.v. 1

This variance can change for each y_i , so assumption 3 is not met.

On residuals vs fitted values we can see if assumptions are met. Here, not the case.

GLM (Generalized Linear Models)

if there is a correlation structure, mixed models can be used.

ACF of the AR(p)

$$\rho(h) - \phi_1 \rho(h-1) - \dots - \phi_p \rho(h-p) = 0$$

$\alpha_1, \dots, \alpha_r$: reciprocal roots

\downarrow
 m_1, \dots, m_r : multiplicity

$$\sum_{i=1}^r m_i = p$$

in this case

$$\rho(h) = \alpha_1^h p_1(h) + \dots + \alpha_r^h p_r(h)$$

$p_j(h)$ polynomial of order m

Exponential Family

Y is from exponential family if it can be written as :

$$f(y; \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

θ : the location (or natural) parameter

ϕ : the dispersion parameter

example : Binomial

$$Y \sim \text{Bin}(n, p)$$

$$\begin{aligned} f(y; n, p) &= P(Y=y; n, p) = \binom{n}{y} p^y (1-p)^{n-y} \\ &= \exp \left\{ \log \binom{n}{y} p^y (1-p)^{n-y} \right\} \\ &= \exp \left\{ \log \binom{n}{y} + y \log(p) + (n-y) \log(1-p) \right\} \\ &= \exp \left\{ \underbrace{\log \binom{n}{y}}_{\theta} + \underbrace{y \log(p)}_{b(\theta)} - \underbrace{y \log(1-p)}_{b(\theta)} + n \log(1-p) \right\} \\ &= \exp \left\{ y \underbrace{\log \left(\frac{p}{1-p} \right)}_{\theta} + n \log(1-p) + \underbrace{\log \binom{n}{y}}_{c(y, \phi)} \right\} \end{aligned}$$

Note that $\theta = \log\left(\frac{p}{1-p}\right)$, $\phi = 1$

$$\Leftrightarrow p = \frac{e^\theta}{1+e^\theta}$$

So the Binomial distribution is a member of the Exponential family.

$$\mu = E[Y] = b'(\theta)$$

$$\sigma^2 = \text{var}(Y) = b''(\theta) \phi$$

Introduction to Binomial Regression

Each y_i is a realization from a Y_i , Binomial

$$Y_i \sim \text{Bin}(1, p_i) \quad p_i \in [0, 1]$$

Binary Logistic Regression

$$Y_i \sim \text{Bin}(n_i, p_i)$$

assumption:
 Y_i are independent

Goal: predict or explain p_i using covariates

$$X_{i,1}, \dots, X_{i,p}$$

Then predict $E[Y_i] = \mu_i = n_i p_i$

Systematic component:

$$\eta_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_p x_{i,p}$$

how to link the linear predictor to the probability of success?

$$g(\mu_i) ?$$

Binomial Regression: link functions

1. Logit (logistic link):

$$\eta_i \stackrel{\text{set}}{=} g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \theta \quad \text{natural parameter}$$

2. Probit:

$$g(p_i) = \bar{\Phi}^{-1}(p_i)$$

$\bar{\Phi}^{-1}$: inverse $N(0,1)$ CDF

Binomial Regression Parameter Estimation

maximum likelihood estimation

n_i = number of trials

1. marginal p.m.f. :

$$P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$$

$$\begin{aligned} p_i &= \frac{e^{\theta_i}}{1 + e^{\theta_i}}, \quad \theta_i = \log\left(\frac{p_i}{1-p_i}\right) = \eta_i \\ &= \beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} \end{aligned}$$

2. joint p.m.f. :

because we have independence

$$\begin{aligned} P(Y_i; n_i, p_i) &= \prod_{i=1}^n \binom{n_i}{p_i} p_i^{y_i} (1-p_i)^{n_i - y_i} \\ &= \prod_{i=1}^n \binom{n_i}{p_i} \left(\frac{e^{\eta_i}}{1 + e^{\eta_i}} \right)^{y_i} \left(1 - \frac{e^{\eta_i}}{1 + e^{\eta_i}} \right)^{n_i - y_i} \end{aligned}$$

3. Likelihood function :

$$L(\beta) = \prod_{i=1}^n \binom{n_i}{y_i} \left(\frac{e^{\eta_i}}{1 + e^{\eta_i}} \right)^{y_i} \left(1 - \frac{e^{\eta_i}}{1 + e^{\eta_i}} \right)^{n_i - y_i}$$

4. log-likelihood function:

$$l(\beta) = \sum_{i=1}^n \left(y_i \eta_i - n \log(1 + e^{\eta_i}) + \log(n!) \right)$$

not linear. We use iterative technique to maximize the log-likelihood function.

5. maximize!

Interpretation of Binomial Regression

$$\eta_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} = \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)$$

2 predictors: $x_{i,1}$ and $x_{i,2}$

we have ml estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$

Definition: Let event E have probability p of occurrence.
Then the odds in favor of E is:

$$O_E = \frac{p}{1-p}$$

biased coin: $P(H) = 3/4$ and $P(T) = 1/4$

$$O_E = \frac{p}{1-p} = \frac{(3/4)^2}{1-(3/4)^2} = \frac{0.5625}{0.375} \approx \frac{1.29}{7}$$

that we get H twice in arrow.

It only holds for the logit link function.

β_0 : log odds of success when all predictors are equal to 0.

β_j : $X_{i,j}$ increase by 1 unit and all other predictors are held constant, the log-odds of success increases by β_j .

Odds of success increases by e^{β_j}

Binomial Regression in R

Occupancy : 0: not occupied
1: occupied

Temperature : in Celsius

Relative Humidity :

Light :

CO₂ measurement :

use "RCurl" package

"glm" function , family = "binomial"

format of the response : factor or
as a two-column matrix
(factor here)

Estimates computed using maximum likelihood estimation
us IRLS.

$\hat{\beta}_0 = -29.31$: . on the scale of the linear predictor
(not exponentiated)

. average log-odds of an app.
being occupied, everything else
is 0, is about 29,3%.

. $e^{\hat{\beta}_0} \approx 0$ (odd)

. $e^{\hat{\beta}_3} \approx 1.02$ (odd)

. increase is multiplicative

$$\begin{aligned}\hat{e}^{x+1} &= e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 (X_3+1) + \hat{\beta}_4 x_4} \\ &= e^{\hat{\beta}_3} \underbrace{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 x_4}}_{e^{\hat{x}}} \\ &= e^{\hat{\beta}_3} e^{\hat{x}} \quad e^{\hat{\beta}_3} \approx 1.02\end{aligned}$$

Poisson Regression : model for count data

$y_i, i=1, \dots, n$

response variable, Poisson

$x_{ij}, j=1, \dots, p$

systematic component

$y_i \sim \text{Poi}(\lambda_i)$

$y_i = 0, 1, 2, \dots$

$$P(y_i = y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

$\lambda_i > 0$ potentially different for d.f.p. measurement

mean : $\mu_i = E[y_i] = \lambda_i$

Variance : $\sigma_i^2 = \text{var}(y_i) = \lambda_i$

canonical parameter : $\theta_i = \log(\mu_i) = \log(\lambda_i)$

canonical link function : $\log(\lambda)$

systematic component :

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

has to be positive. The mean (positive) should be linked to the linear predictor by a link function

$$g(\lambda_i) = \log(\lambda_i) \quad \text{mean in a rate}$$

$$\Rightarrow \eta_i \stackrel{\text{set}}{=} g(\lambda_i) = \log(\lambda_i) = \theta \quad (\text{canonical parameter})$$

$$\lambda_i = e^{\eta_i}$$

estimation by maximum likelihood.

$$\underline{\text{Rate response}} : \mu_i = \lambda_i = \frac{\text{count}}{\text{exposure time}} = \frac{y_i}{e_i}$$

offset term : (log link)

e_i : Known period of time

$$g(\lambda_i) = \log(\lambda_i) = \log\left(\frac{y_i}{e_i}\right) = \log(y_i) - \log(e_i)$$

offset in glm formula. For different period of time of exposure.

Family = "poisson"

Poisson Regression : parameter estimation

maximum likelihood estimation

1. marginal pmf :

$$P(Y_i = y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, y_i = 0, 1, 2, \dots$$

2. joint pmf :

$$f(y_i; \lambda_i) = \prod_{i=1}^n \left(\frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \right)$$

3. likelihood function

$$L(\beta) = \prod_{i=1}^n \left(\frac{e^{\lambda_i} \lambda_i^{y_i}}{y_i!} \right)$$

$$\eta_i = \log(\lambda_i)$$

$$e^{\eta_i} = e^{\beta_0 + \beta_1 x_i + \dots}$$

$$= \prod_{i=1}^n \left(\frac{e^{e^{-\eta_i}} e^{y_i \eta_i}}{y_i!} \right)$$

$$= \prod_{i=1}^n \left(\frac{e^{y_i \eta_i - e^{\eta_i}}}{y_i!} \right)$$

4. log-likelihood function

$$l(\beta) = \sum_{i=1}^n \left(y_i \eta_i - e^{\eta_i} - \log(y_i!) \right)$$

5. maximize ! using iterative techniques.

Interpreting the Poisson Regression model

remember : we use the log link $\log(\lambda_i) = \eta_i$

- β_0 : e^{β_0} can be interpreted as the mean of the response when each predictor is set to 0.
- β_j : e^{β_j} can be interpreted as the multiplicative increase in the mean of the response for a one unit increase in $x_{i,j}$, holding all other predictors constant.

$$\begin{aligned}y_i^{+1} &= \hat{\lambda}_i^{+1} = e^{\{\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_j (x_{i,j} + 1) + \\&\quad \dots + \hat{\beta}_p x_{i,p}\}} \\&= \exp\{\hat{\beta}_j\} \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_j (x_{i,j}) + \dots + \hat{\beta}_p x_{i,p}\} \\&= \exp\{\hat{\beta}_j\} \hat{\lambda}_i\end{aligned}$$

Poisson Regression : Goodness of fit

Deviance: D $\xrightarrow{\text{mle}}$

$$D = -2 \underbrace{\ell(\hat{\beta})}_{\sum_{i=1}^n (y_i \eta_i - e^{\hat{\eta}_i} - \log(y_i!))}$$

a small deviance means a better fit.

The null deviance (deviance for the null model)

$$D_{\text{null}} = -2 \sum_{i=1}^n (y_i \hat{\eta}_i - e^{\hat{\eta}_i} - \log(y_i!))$$

if $\eta_i = \beta_0 \Leftrightarrow \hat{\lambda}_i = \bar{y}$

$$= -2 \sum_{i=1}^n (y_i \log(\hat{\lambda}_i) - \hat{\lambda}_i - \log(y_i!))$$

$$= -2 \sum_{i=1}^n (y_i \log(\bar{y}) - \bar{y} - \log(y_i!))$$

The saturated deviance (deviance for the saturated model)

A parameter for each data point.

$$D_{\text{sat}} = -2 \sum_{i=1}^n (\gamma_i \log(\gamma_i) - \gamma_i - \log(\gamma_i!))$$

The residual deviance

$$D_{\text{resid}} = D_p - D_{\text{sat}}$$

D_p is the model that we are interested

$$= -2 \sum_{i=1}^n (\gamma_i \log(\hat{\lambda}_i) - \hat{\lambda}_i - \log(\gamma_i!))$$

$$+ 2 \sum_{i=1}^n (\gamma_i \log(\gamma_i) - \gamma_i - \log(\gamma_i!))$$

$$= \dots$$
$$= 2 \sum_{i=1}^n (\gamma_i \log(\gamma_i/\hat{\lambda}_i) - (\gamma_i - \hat{\lambda}_i))$$

This quantity has a Chi-2 distribution ($n-(p+1)$) degrees of freedom

$$\chi^2_{n-(p+1)}$$

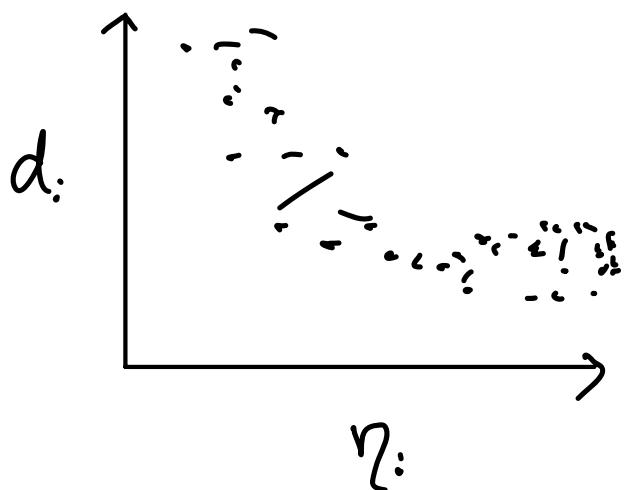
Goodness of fit test

H_0 : The model with p parameters fits the data well

Deviance residuals

$$d_i = \text{sign}(y_i - \hat{\lambda}_i) \sqrt{\varepsilon \left(y_i (\log(y_i/\hat{\lambda}_i)) - (y_i - \hat{\lambda}_i) \right)}$$

$d_i \sim N(0, 1)$ under H_0 and with a high number of counts.



if we see a nonlinear trend, this could indicate lack of fit

Other Goodness of Fit

Pearson's χ^2

$$\chi^2 = \sum_{i=1}^n (O_i - E_i)^2 / E_i$$

$$= \sum_{i=1}^n (Y_i - \hat{\lambda}_i)^2 / \hat{\lambda}_i$$

Squared standardized residual

$$\sim \chi^2_{(n-p+1)}$$

A large χ^2 is evidence against the null that the model fit is sufficient.

$$P_i = \frac{Y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}} \quad \text{pearson residual}$$

Overdispersion

$$E[Y_i] = \lambda_i \quad \text{Var}(Y_i) = \lambda_i \quad \text{if } Y_i \text{ is Poisson}$$

if $\text{var}(Y_i) > E[Y_i]$, we have overdispersion 17

Real overdispersion

vs

apparent overdispersion

- zero inflation
⇒ mixture models or hurdle models
e.g. some insurance claim models.
- spatial or temporal correlation

- outliers that are "true measurements"
⇒ detecting outliers and removing them
- missing predictors
- inappropriate link function

How to detect?

$$\mathbb{E}[Y_i] = \lambda_i, \quad \text{var}(Y_i) = \lambda_i \phi$$

ϕ is the dispersion parameter.

$$\hat{\phi} = \frac{\sum_{i=1}^n (y_i - \hat{\lambda}_i)^2 / \hat{\lambda}_i}{n - (p + 1)}$$

$$\sim \chi^2_{n-(p+1)}$$

p : number of predictors in the model

if statistically significant, then we have evidence of overdispersion.

Solutions (to overdispersion)

1. Quasi-likelihood methods

$$\hat{se}(\hat{\beta}_j) = \hat{se}(\hat{\beta}_j) \sqrt{\hat{\phi}}$$

only the estimate variances will be affected.

2. Negative Binomial regression

Introduction to Nonparametric Regression models

The Normal regression model

$$\underline{Y} \sim N(\underline{X}\underline{\beta}, \sigma^2 I_n)$$

response mean
 → ↑ ↑
 (p+1) parameters Covariance matrix
 ↓ ↓ ↓
 one parameter

Overall, we have $p+2$ parameters to estimate.

(VS)

$$\underline{Y} \sim N(f(\underline{x}), \sigma^2 I)$$



arbitrary function
say $x_i \in [-1, 1]$

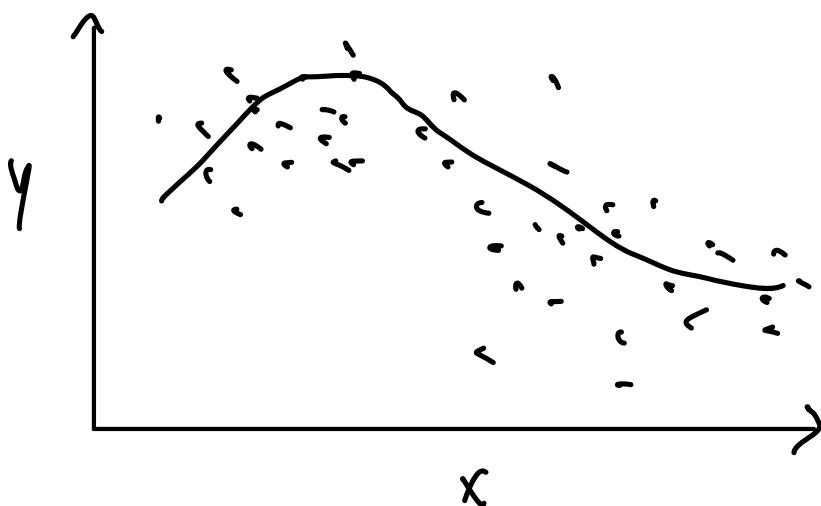
$$\mu_i = E[Y_i] = f(x_{i,1}, x_{i,2}, \dots, x_{i,p})$$

general form

Nonparametric regression allows us to be more flexible with the form of f .

We assume it comes from smooth families of functions. The set is much larger than the parametric approach.

eg. $f(x) = \sin(\pi x)$



Advantages (of Nonparametric approach)

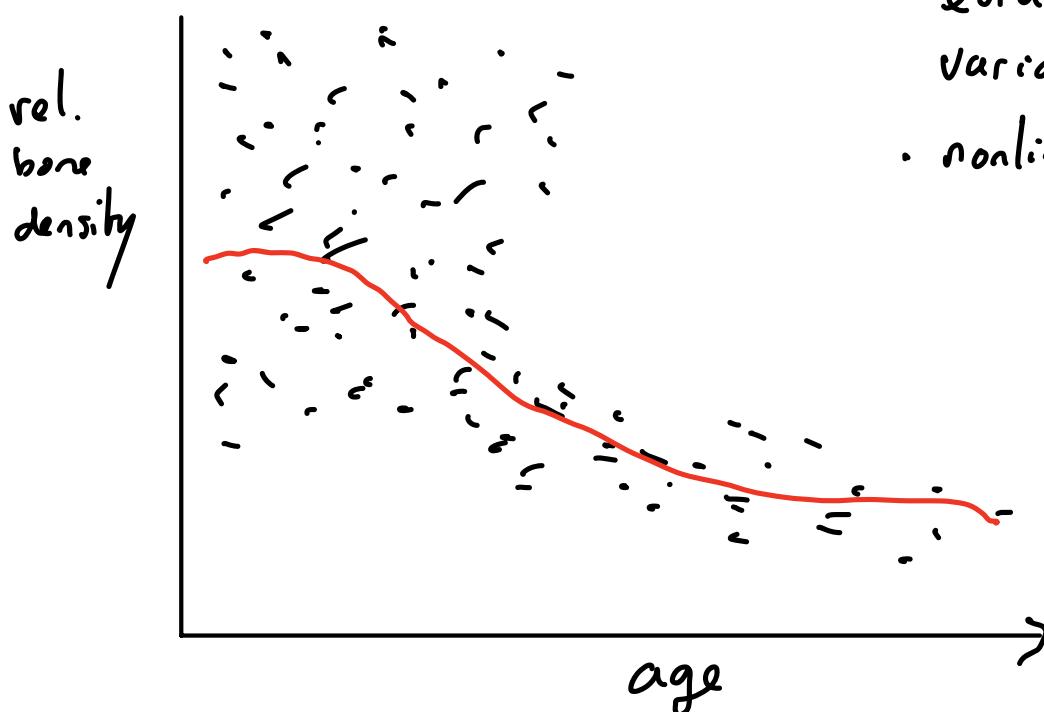
1. flexibility
2. fewer distributional assumptions

Disadvantages

1. Less efficient when the structure of the relationship is available.
2. Interpretation difficulty.

Kernel Density Estimators

e.g. Bone density data



- evidence of non constant variance
- nonlinear relationship

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_d x_i^d}_{f(x_i)} + \varepsilon_i$$

use polynomial regression?

Automated way to find the form of x_i .

Kernel smoothers (choice of the degree of the polynomial)

Kernel estimators

model : $Y_i = f(x_i) + \varepsilon_i$ (Simple weighted moving average of the response)

$$\hat{f}_\lambda(x) = \frac{\frac{1}{n\lambda} \sum_{i=1}^n K\left(\frac{x-x_i}{\lambda}\right) Y_i}{\frac{1}{n\lambda} \sum_{i=1}^n K\left(\frac{x-x_i}{\lambda}\right)}$$

$$w_i = \frac{1}{\lambda} K\left(\frac{x-x_i}{\lambda}\right) \quad \begin{matrix} \text{: weighting average} \\ w_i = \text{weights} \end{matrix}$$

Kernel : is a nonnegative, real valued function such that $K(x) = K(-x)$ for all

values of x (symmetry) and
 $\int K(x) dx = 1$ (normalization)

Commonly used Kernels

- $K = \frac{1}{2}, x \in [-1, 1]$ Uniform/rectangular Kernel
- $K = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}$ Gaussian Kernel
- $K = \frac{3}{4} (1-x)^2, x \in [-1, 1]$ Epanechnikov Kernel.

The fit is sensitive to the bandwidth (λ) or smoothing parameter. It controls the smoothness of the fit.

low λ , bumpy } estimate of \hat{f}
 high λ , smooth }

in R : Ksmooth()

How to choose the bandwidth?

The least smooth fit that does not show any implausible fluctuations. On the video, the grey curve is the best fit.
"J. Faraway"

To note:

in R, `Ksmooth()` for Kernel estimation
choose $K(x)$ and λ .

Some automatic methods to select λ , e.g. cross validation.

Smoothing Splines

$$\text{model: } Y_i = f(x_i) + \varepsilon_i$$

We might consider choosing \hat{f} by minimizing

$$TSE = \frac{1}{n} \sum_{i=1}^n (Y_i - f(x_i))^2$$

solution: $\hat{f}(x_i) = y_i$ Then we would be interpolating

so \Rightarrow smoothing splines

We minimize:

$$\underbrace{\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2}_{\text{fit}} + \lambda \underbrace{\int [f''(x)]^2 dx}_{\text{smoothness}}$$

λ : smoothing parameter

To keep it small, we need something that has low curvature on average.

A spline : is a piecewise function, where each segment is a polynomial.

A cubic spline : the segments are each of degree 3.
example from the video

Splines are continuous and have continuous derivatives.

Smoothing splines is a spline designed to balance fit with smoothness.

spar : function of λ

in R : smooth.spline(x, y, spar = ...)

LOESS : locally estimated scatterplot smoothing

$y = f(x_i) + \varepsilon_i$, we might consider choosing f by minimizing :

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

$$f(x) = ?$$

- linear regression: $\beta_0 + \beta_1 x$ (the function is linear)

- Taylor expansion around x_0 :

$$f(x) \approx f(x_0) + \frac{f'(x_0)}{1!} (x - x_0) + \frac{f''(x_0)}{2!} (x - x_0) + \dots + \frac{f^{(p)}(x_0)}{p!} (x - x_0)^p$$

$$\approx \beta_0 + \beta_1 + \dots + \beta_p$$

approximate f around a particular point x_0 using a polynomial : least squares problem.

$$MSE_{\text{Taylor}} = \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{j=0}^p \beta_j (x - x_0)^j \right)^2$$

$$MSE_{\text{Taylor}}^w = \frac{1}{n} \sum_{i=1}^n w_i (x - x_0) \left(y_i - \sum_{j=0}^p \beta_j (x - x_0)^j \right)^2$$

$$w_i(x) = \begin{cases} (1 - |x|^3)^3 & \text{if } |x| < 1 \\ 1 & \text{if } |x| \geq 1 \end{cases}$$

advantages : • flexible fit
 • simple to implement
 • uncertainty quantification

disadvantages : • computationally expensive
 • interpretation can be difficult
 • require large, densely sampled data

Generalized additive models

general model: $y_i = f(x_{i,1}, \dots, x_{i,p}) + \varepsilon_i$

sum of simpler functions

$$= f_1(x_{i,1}) + \dots + f_p(x_{i,p})$$

examples :

$$f(x_1, x_2, x_3) = x_1 + x_1^2 + x_2 + x_2^2 + \underbrace{x_1 x_2}_{f_{12}(x_1, x_2)} + \sin^2(x_3)$$

\Rightarrow not additive

$$f(x_1, x_2) = \pi + e^{5x_1} + \log(2x_2)$$

\Rightarrow additive

$$f(x_1, x_2, x_3) = 1 + \log(0.5x_1^2) - x_2 + x_3$$

\Rightarrow additive

Kernel estimation : can be generalized

$$y_i = f(x_i) + \varepsilon_i$$

$$\hat{f}(x) = \sum_{i=1}^n \frac{K_h(x - x_i) y_i}{\sum_{i=1}^n K_h(x - x_i) y_i}$$

- H is a positive definite matrix : H^{-1} exists
 $H^{1/2}$ exists

notation: $H^{-\frac{1}{2}}$

$|H|$ determinant of H

$$\bullet K_u(v) = |H|^{-\frac{1}{2}} \underbrace{K(H^{-\frac{1}{2}}v)}_{\leq}$$

Scaled Kernel

$K(\underline{z})$ is a Kernel

$$\bullet K(\underline{z}) > 0, \int K(\underline{z}) d\underline{z} = 1$$

Fit can become complex in higher dimension

\Rightarrow how to simplify $f(\cdot)$

additive models

$$E[Y_i] = f(X_{i,1}, \dots, X_{i,p}) = \beta_0 + f_1(X_{i,1}) + \dots + f_p(X_{i,p})$$

The class of functions that are additive is much more flexible than linear regression.

$$\text{eg. } \beta_0 + \beta_1 X_{i,1} + f_2(X_{i,2}) + \dots + f_p(X_{i,p})$$

we can account for categorical predictors

Additive models won't work well in the presence
of strong interactions among predictors

$$g(\lambda_i) = \log(\lambda_i) = \beta_0 + \beta_1 x_{i,1} + f_2(x_{i,2}) + \dots + f_p(x_{i,p})$$

example : Ozone data

plot marginal relationships.