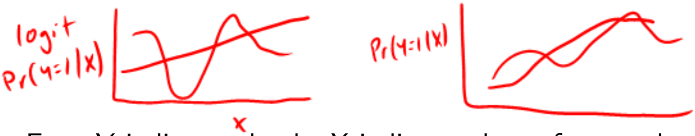
STATISTICS 538, Lecture #12

From lines to curves

December 1, 2010

GLM: linear predictor is linear in **parameters**, not necessarily linear in **predictors**



E.g., Y indicates death, X indicates dose of nasty chemical. First thought

$$Pr(Y=1|X) \quad = \quad \text{expit}(\beta_0 + \beta_1 X)$$

No good? Change link function? Or expand linear predictor from $\eta = \beta_0 + \beta_1 X$ to something more complex (but still linear in parameters).

logit $\rightarrow$ expit$(\beta_0+\beta_1 X)$

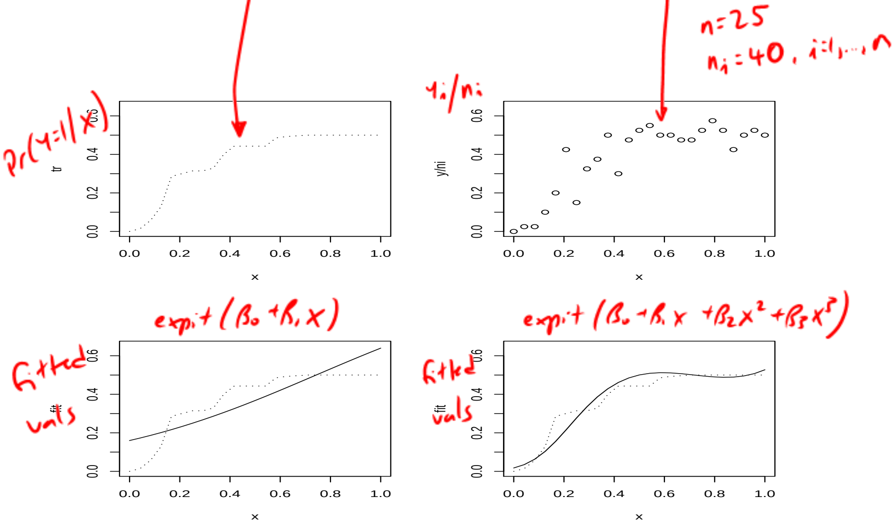
probit $\rightarrow$ $\Phi(\beta_0+\beta_1 X)$

standard normal cdf

first thought

$\eta = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots + \beta_j X^j$

Consider this true relationship spawning these data, and two fits



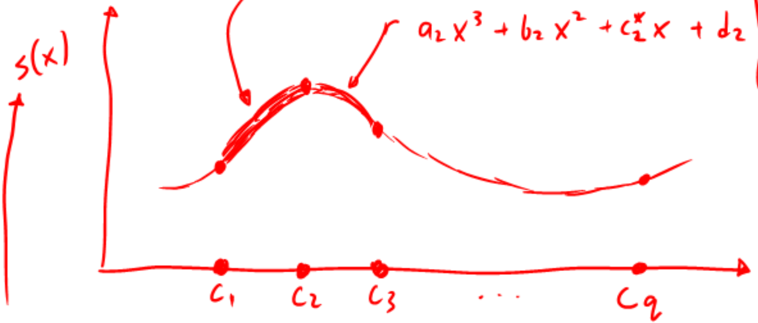
Alternative to polynomials? Splines!

Cubic splines: Choose **knots** $c_1, \ldots, c_q$. Consider piecewise cubic functions (between knots) with smooth joins (at knots)

$a_1 x^3 + b_1 x^2 + c_1^* x + d_1$

$a_2 x^3 + b_2 x^2 + c_2^* x + d_2$

$\boxed{s(\,),\, s'(\,) \text{ and } s''(\,) \text{ continuous}}$



some smooth function

$(q+1)$ cubic polynomials $\to (q+1) \times 4$ parameters

BUT $-\ 3q$ constraints $\to$ $q+4$ parameters

Splines have 'basis function' representations, i.e., linear in parameters

For instance

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{q} \beta_{j+3} \{(x - c_j)_+\}^3$$

Note that $s(x)$ is indeed a piecewise cubic function with smooth joins, described by $q+4$ parameters ✓

$$\{(x - c_j)_+\}^3 = \begin{cases} (x-c_j)^3 & \text{if } x > c_j \\ 0 & \text{if } x \leq c_j \end{cases}$$

Knots are user-chosen - possible default values?

$\left(\frac{1}{q+1}, \frac{2}{q+1}, \ldots, \frac{q}{q+1}\right)$ quantiles of the x values

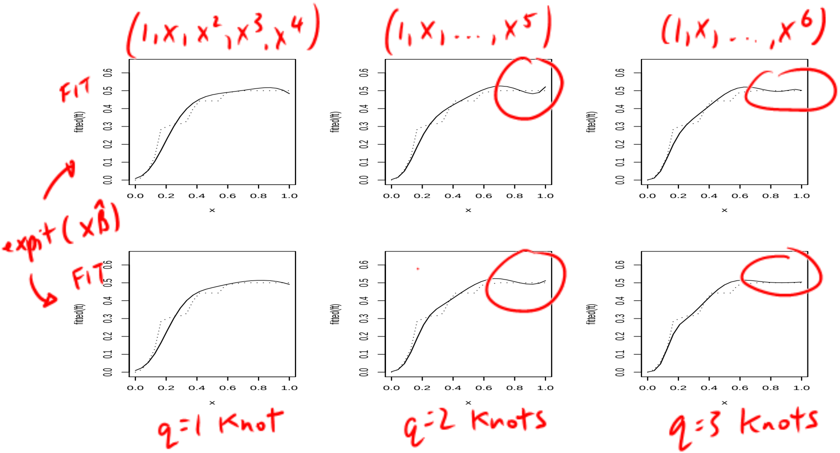
So 'cheap homemade' spline fitting...

$$\begin{pmatrix} s(x_1) \\ \vdots \\ s(x_n) \end{pmatrix} = \underbrace{\begin{pmatrix} 1, \text{xmat} \end{pmatrix}}_{n \times (q+4)} \underbrace{\beta}_{(q+4)\times 1}$$

```
xmat <- cbind(x,x^2,x^3)

for (j in 1:q) {
  xmat <- cbind(xmat, pmax(x - quantile(x,j/(q+1)),0)^3)
}

ft <- glm(cbind(y,ni-y) ~ xmat, family=binomial)

plot(x, fitted(ft)) ...
```

$$\underbrace{\texttt{quantile(x,j/(q+1))}}_{\text{choice of } c_j}$$

Add one, two, three parameters beyond the cubic model: polynomials (top) versus splines (bottom)



Further issues

Model selection: how many knots?

Usual tools available. E.g. in our example

Deviance GOF: $(1, x)$ flunks, $(1, x, x^2)$ borderline

$(1, x, x^2, x^3)$ or splines with $q = 1, \dots, 5$ poss

AIC: splines with $q = 2$ wins

BIC: $(1, x, x^2, x^3)$ wins

More refined splines: **natural cubic splines**

take cubic spline - add additional constraints

$s(x)$ linear to the left of c_1 and to the right of c_q

4 more constraints $s''(c_1) = s'''(c_1) = s''(c_q) = s'''(c_q) = 0$

now q knots $\Rightarrow$ q parameters

Further issues, continued

Lots of curve-fitting packages - don't need to do-it-yourself

```
> library(splines)
> ft <- glm(cbind(y,ni-y)~ ns(x,df=4), family=binomial)
```

natural cubic spline

$$s(x) = \beta_0 + \sum_{j=1}^{4} \beta_j s_j(x)$$

basis functions

ncs with knots at $\left(\frac{0}{4}, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, \frac{4}{4}\right)$ quantiles of x values