Local likelihood

Patrick Breheny

November 13

Moving beyond least squares

- Thus far, we have fit local least squares models
- More generally, we may allow the outcome Y_i to follow a distribution $f(y|\theta_i)$, *e.g.*,
  - Exponential: $f(y|\theta) = \theta^{-1}\exp(y/\theta), \quad y \geq 0$
  - Binomial: $f(1|\theta) = \theta, \quad f(0|\theta) = 1 - \theta$
- For regression problems, θ_i depends on some covariate x_i
- A parametric model would involve the specification $\theta_i = \alpha + \beta x_i$; today we will let $\theta_i = \theta(x_i)$ represent an unknown smooth function we wish to estimate

Local likelihood

- One way to achieve that flexibility is by fitting separate, local models at each target point x_0 and smoothing those models together using kernel weighting
- Specifically, at x_0, we estimate $\hat{\alpha}$ and $\hat{\beta}$ by maximizing

$$\sum_i K_h(x_0, x_i) l(\alpha + \beta x_i | y_i)$$

  where $l(\theta|y) = \log\{f(y|\theta)\}$
- In principle, any distribution and likelihood could be extended to this approach, but in practice it is usually applied to generalized linear models

Fitting local GLMs

- Letting the ith row of the design matrix be $(1, x_i - x_0)$ as in local linear regression, the local likelihood estimate $\widehat{\boldsymbol{\beta}}$ at x_0 can be found by solving

$$\mathbf{X}'\mathbf{W}\mathbf{u} = \mathbf{0},$$

  where $\mathbf{W}$ is the diagonal matrix of kernel weights and $\mathbf{u} = \frac{\partial}{\partial\theta} l(y_i, \hat{\theta}_i)$ is the score vector
- Unlike local linear regression, this equation typically does not have a closed form solution and must be solved by iterative methods

Linearization of the score

- As with regular GLMs, we may proceed by constructing a linear approximation to the score via Taylor series expansion around the current estimate, $\tilde{\boldsymbol{\theta}}$:

$$\mathbf{u} \approx \mathbf{V}(\mathbf{z} - \boldsymbol{\theta}),$$

  where $\mathbf{V}$ is a diagonal matrix with entries $-\frac{\partial^2}{\partial \theta^2} l(\hat{\theta}_i | y_i)$ (*i.e.*, the observed information) and $\mathbf{z} = \tilde{\boldsymbol{\theta}} + \mathbf{V}^{-1}(\mathbf{y} - \tilde{\mu})$ is the "pseudoresponse"
- The solution to our local maximum likelihood solution is therefore

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}\mathbf{W}\mathbf{V}\mathbf{z}$$

- It is important to keep in mind, however, that both $\mathbf{z}$ and $\mathbf{V}$ depend on $\tilde{\boldsymbol{\theta}}$, and thus we need to update them via $\tilde{\boldsymbol{\theta}} \leftarrow \mathbf{X}\tilde{\boldsymbol{\beta}}$ and iterate until convergence

Deviance and degrees of freedom

- The analogous concept to the residual sum of squares for generalized linear models is the *deviance*:

$$D(\mathbf{y}|\hat{\boldsymbol{\theta}}) = 2\left\{l(\boldsymbol{\theta}_{\max}|\mathbf{y}) - l(\hat{\boldsymbol{\theta}}|\mathbf{y})\right\},$$

  where $\boldsymbol{\theta}_{\max}$ is the vector of parameters that maximize $l(\boldsymbol{\theta}_{\max}|\mathbf{y})$ over all $\boldsymbol{\theta}$ (the "saturated" model)
- Continuing with the analogy to local linear regression, we may define our two effective degree of freedom terms:

$$\begin{aligned}\nu &= \mathrm{tr}(\mathbf{R}) \\ \tilde{\nu} &= 2\mathrm{tr}(\mathbf{R}) - \mathrm{tr}(\mathbf{R}'\mathbf{V}\mathbf{R}\mathbf{V}^{-1}),\end{aligned}$$

  where $\mathbf{R} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{V}$

Caveats

- Unlike the residual sum of squares, the deviance is not χ^2 distributed, not even asymptotically
- Nevertheless, inference based on deviance and approximate degrees of freedom is useful in practice, aids with interpretation, and usually provides adequate empirical accuracy in terms of preserving coverage and type I error rates

Selection of h

- As always, there is the issue of how to choose the bandwidth h
- One approach is to carry out leave-one-out cross-validation with deviance replacing squared error loss:

$$\mathrm{CV} = \sum_i D\left(y_i|\hat{\theta}_{-i}(x_i)\right)$$

- However, unlike local linear regression, non-gaussian GLMs are not linear smoothers and there is no convenient way to calculate $\hat{\theta}_{-i}(x_i)$ without refitting the model
- For this reason, it is customary to use a criterion such as AIC instead:

$$\mathrm{AIC} = \sum_i D(y_i|\hat{\theta}_i) + 2\nu$$

Confidence intervals

One can obtain confidence intervals for $\theta(x_0)$ via quadratic approximations, as is often done with GLMs themselves:

$$\hat{\theta}(x_0) = \mathbf{R}\mathbf{z}$$

Thus,

$$\begin{aligned}\mathbb{V}\{\hat{\theta}(x_0)\} &= \mathbf{R}\mathbb{V}(\mathbf{z})\mathbf{R}' \\ &= \mathbf{R}\mathbf{V}^{-1}\mathbf{R}'\end{aligned}$$

Generalized likelihood ratio tests

- Finally, we can carry out hypothesis testing between two nested models via approximate generalized likelihood ratio tests:

$$\Lambda = 2\left\{l(\hat{\boldsymbol{\theta}}_1|\mathbf{y}) - l(\hat{\boldsymbol{\theta}}_0|\mathbf{y})\right\}$$

  or equivalently,

$$\Lambda = 2\left\{D(\mathbf{y}|\hat{\boldsymbol{\theta}}_0) - D(\mathbf{y}|\hat{\boldsymbol{\theta}}_1)\right\}$$

- Under the null hypothesis that model 0 is correct, Λ follows a distribution very similar to a χ^2 distribution with $\tilde{\nu}_1 - \tilde{\nu}_0$ degrees of freedom

Syntax

- The syntax for fitting generalized linear models in `R` is straightforward; both `locfit` and `gam` provide a `family` argument that works exactly the same as it does in `glm`
- Thus, for `locfit`:

```
locfit(chd~lp(sbp), data=heart, family="binomial")
```

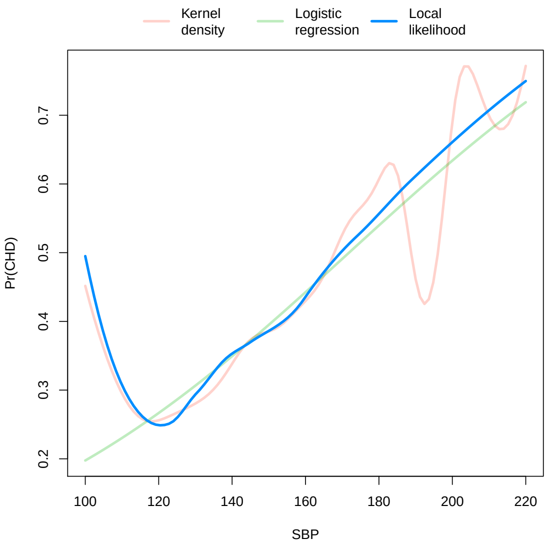
and for `gam`:

```
gam(chd~lo(sbp), data=heart, family="binomial")
```

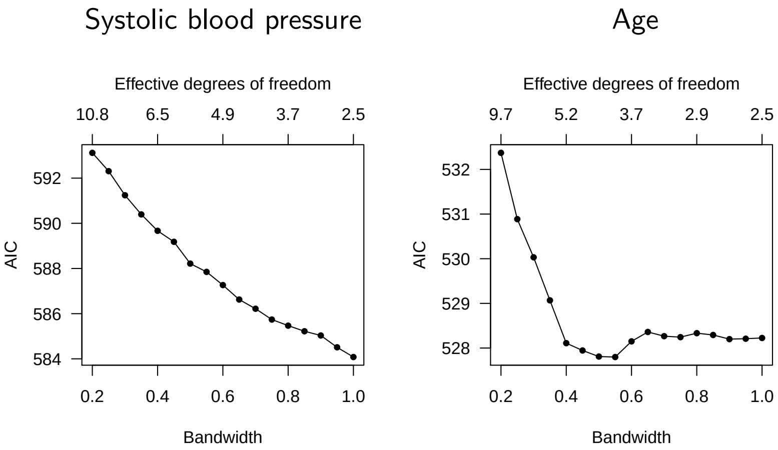
Local logistic regression

- By default, both `gam` and `locfit` incorporate a *link function*; rather than model $\mathbb{E}(Y)$ directly, they model
$$g\left\{\mathbb{E}(Y|x)\right\} = \theta(x),$$
where g is a known function
- For logistic regression, g is usually chosen to be the logit function:
$$g(\pi) = \log\left\{\frac{\pi}{1-\pi}\right\},$$
where $\pi = \mathbb{P}(Y=1)$, thus implying
$$\pi = \frac{e^{\theta}}{1+e^{\theta}}$$
- This is the *canonical link* for a binomial likelihood; in general, canonical links have many attractive statistical properties, such as ensuring that $\mathbb{E}(Y)$ stays within the support of Y

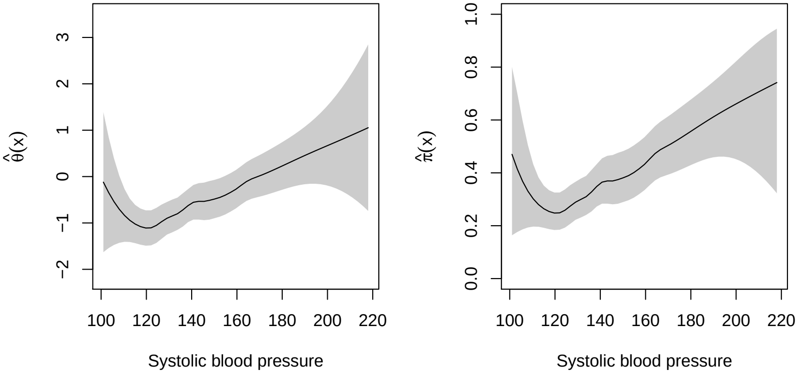
Comparison of local likelihood with other methods



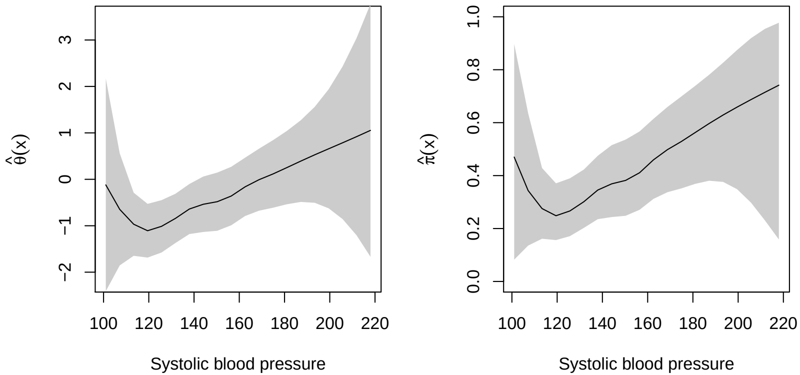
Using AIC to choose bandwidth



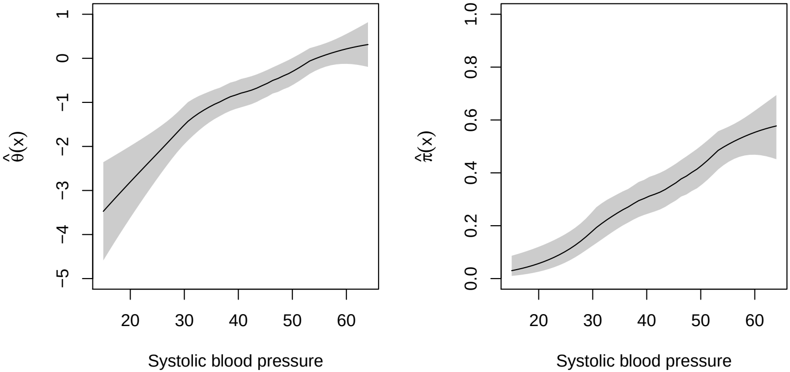
SBP: Pointwise bands



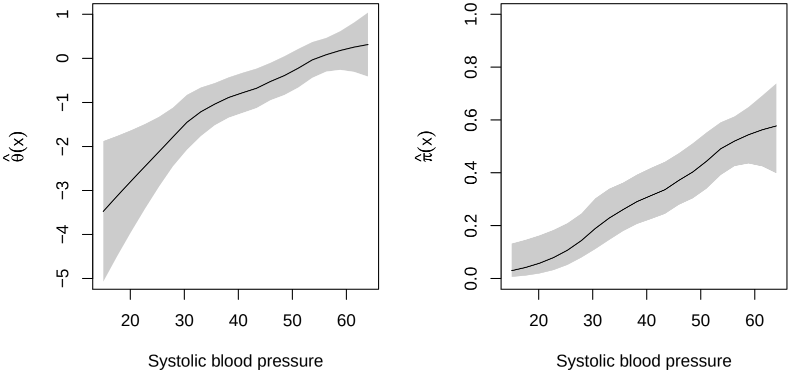
SBP: Simultaneous bands



Age: Pointwise bands



Age: Simultaneous bands



ANOVA table: SBP

	Resid. df	Deviance	$\tilde{\nu}$	ΔDev	p
Null	461	596.1			
Linear	460	579.3	1	16.79	< 0.0001
Local	457.4	577.7	2.6	1.60	0.58

ANOVA table: Age

	Resid. df	Deviance	$\tilde{\nu}$	ΔDev	p
Null	461	596.1			
Linear	460	525.6	1	70.55	< 0.0001
Local	457.5	519.9	2.5	5.65	0.09