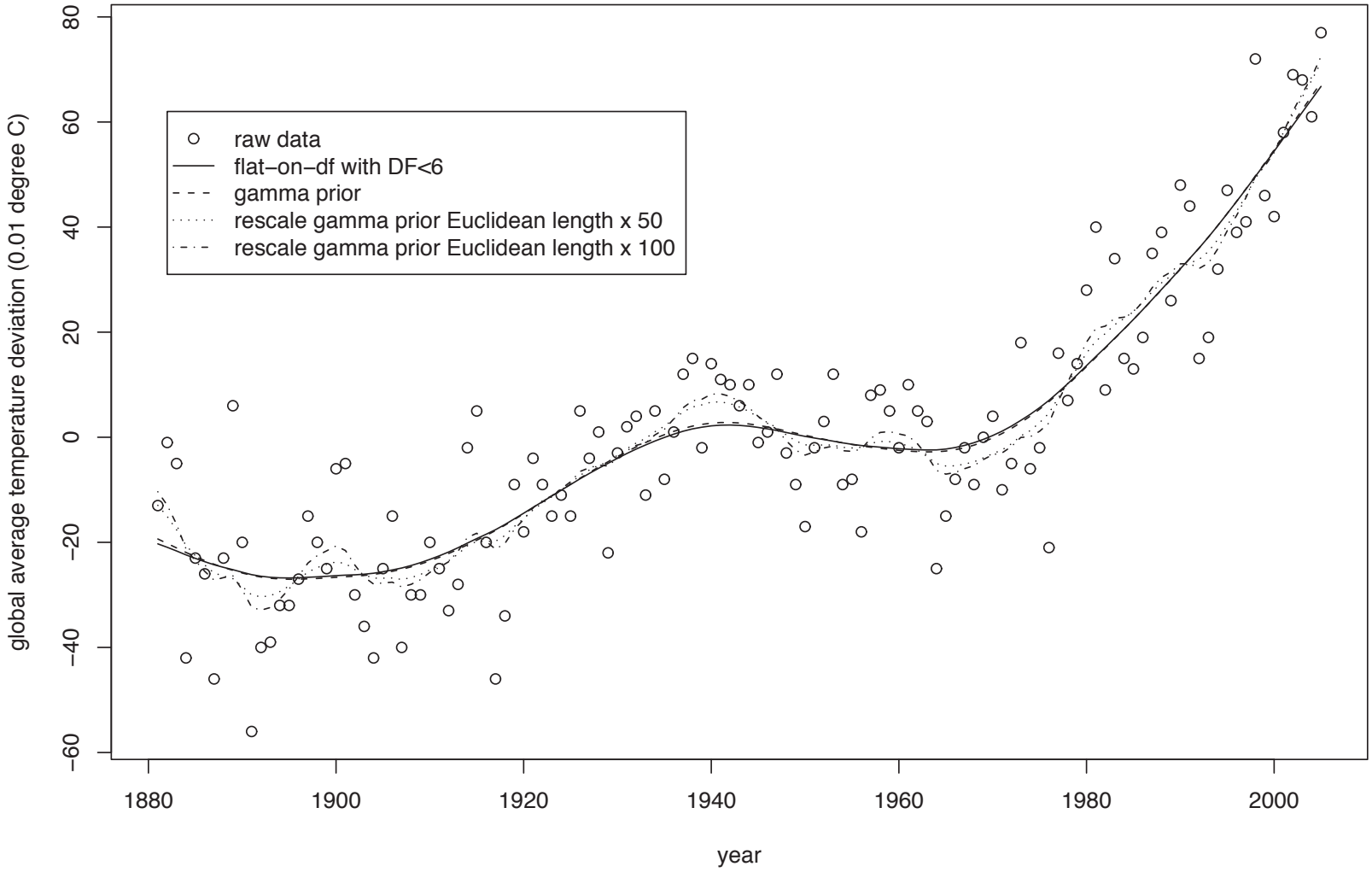
Penalized splines are one kind of smoother

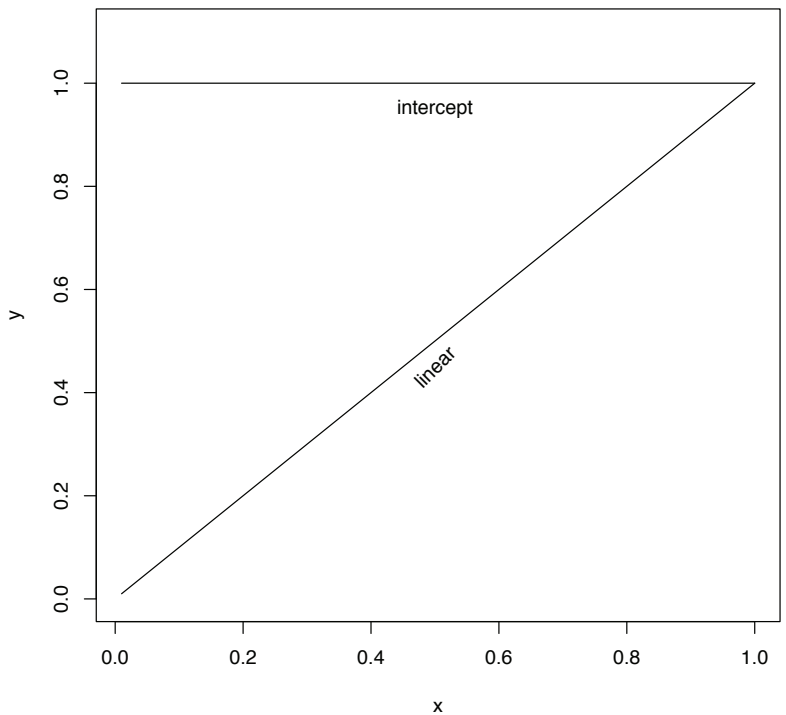
Here's an example where you might use one:



Penalized splines: Basis, knots, penalty

We'll start with a basis in the sense of linear algebra (RWC Sec. 3.2)

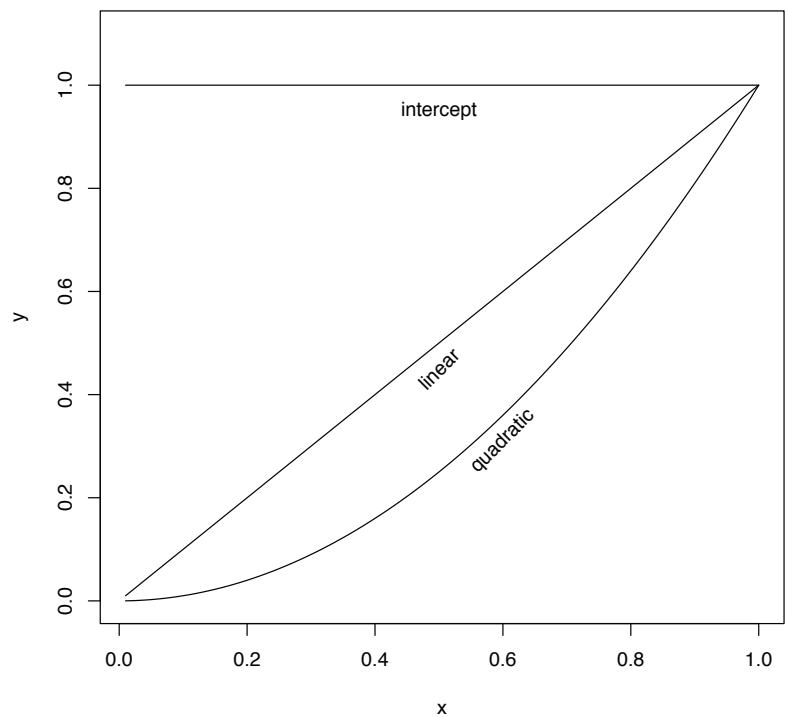
For the linear model $y_i = \beta_0 + \beta_1 x_i + [\text{error}], i = 1, \ldots, n,\ x_i \in [0, 1]$,



the fitted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
lie in a 2-D subspace of $\mathbf{R}^n$
defined by these two basis vectors:

$$\begin{matrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{matrix}$$

For the quadratic model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + [\text{error}]$,



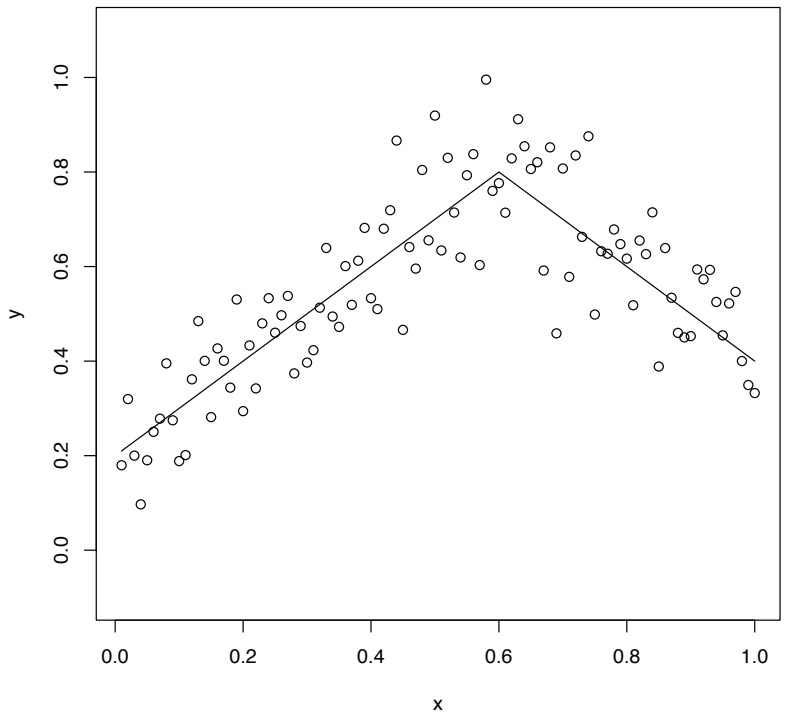
the fitted values lie
in a 3-D subspace of $\mathbf{R}^n$ defined
by these three basis vectors:

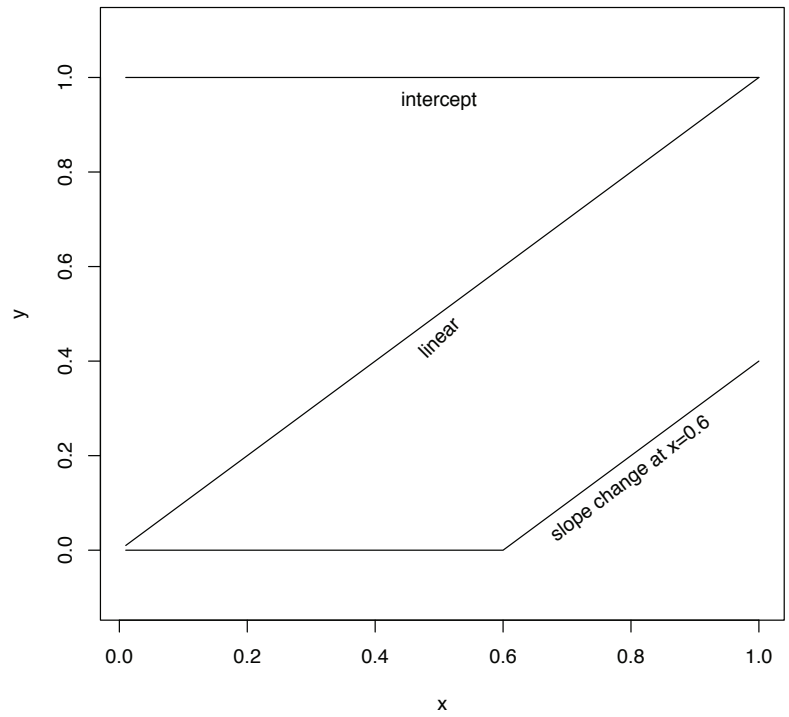
$$\begin{matrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{matrix}$$

Now for a "broken stick" regression, where the slope changes at $x = 0.6$:

$y_i = \beta_0 + \beta_1 x_i + \beta_{11}(x_i - 0.6)_+ + [\text{error}]$, where

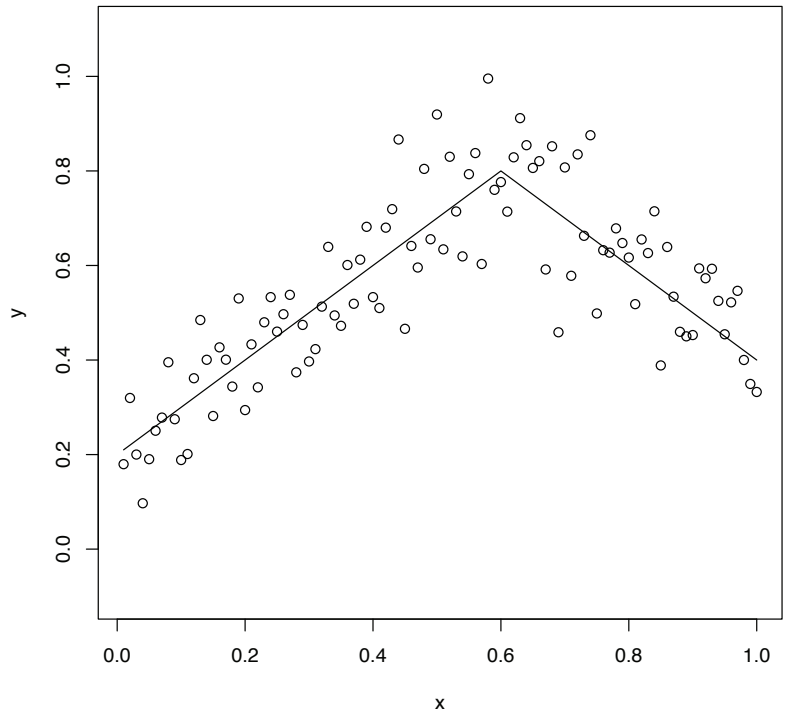
$$(z)_+ = \begin{cases} z & \text{if } z \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

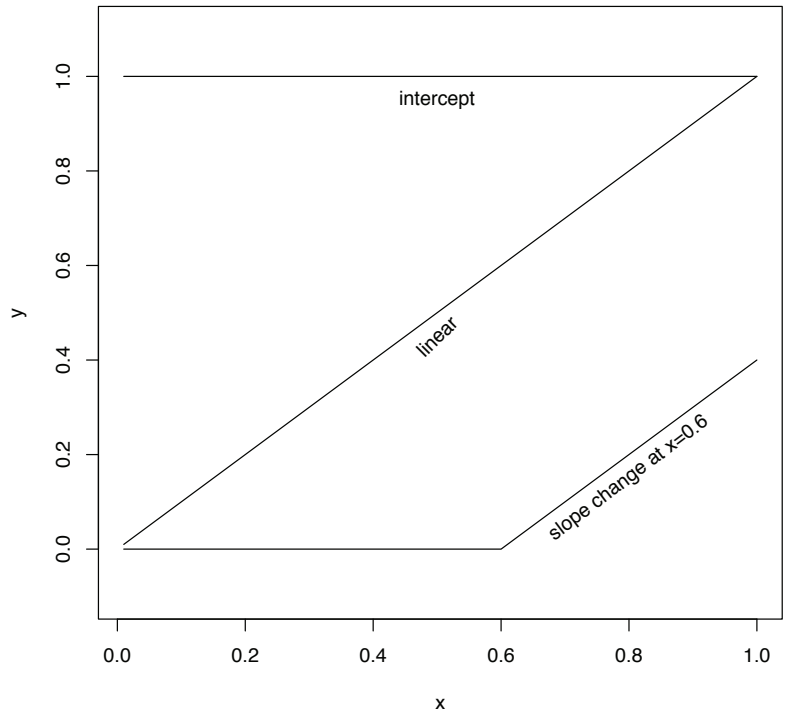




Basis for the "broken stick" regression

$$\begin{matrix} 1 & x_1 & (x_1 - 0.6)_+ \\ 1 & x_2 & (x_2 - 0.6)_+ \\ 1 & x_3 & (x_3 - 0.6)_+ \\ \vdots & \vdots & \vdots \\ 1 & x_n & (x_n - 0.6)_+ \end{matrix}$$

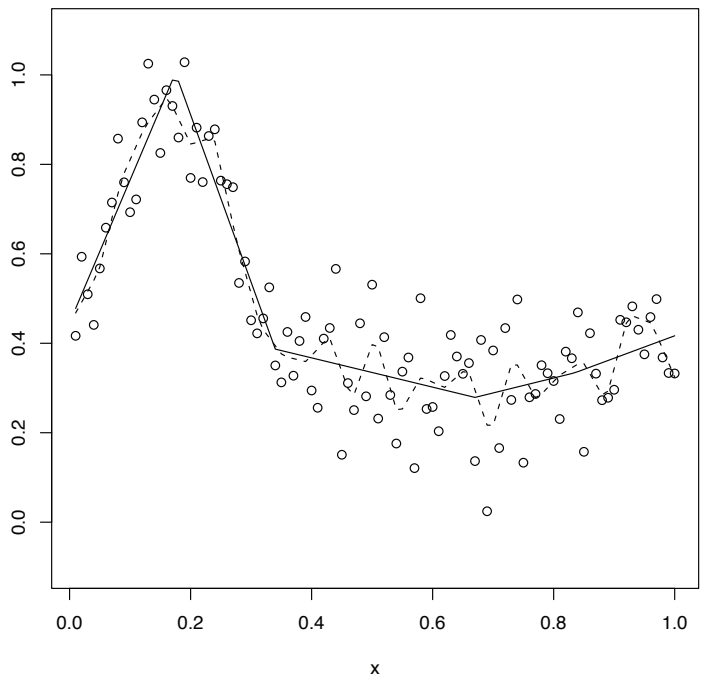




A simple spline: Let the slope change many times

A spline's *knots* $\kappa_j, j = 1, \ldots, K$ are the x's where the slope can change.

Here's a spline with 5 knots: $y_i = \beta_0 + \beta_1 x_i + \sum_{j=1}^{K} \beta_{1j}(x_i - \kappa_j)_+ + [\text{error}]$



The fitted-value space has basis

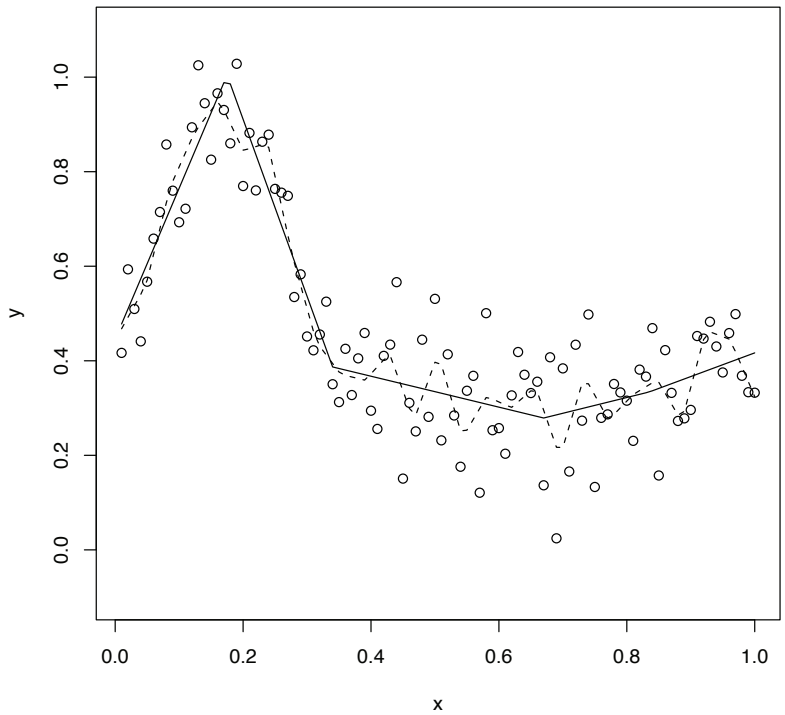
$$\begin{array}{ccccc} 1 & x_1 & (x_1-\kappa_1)_+ & \ldots & (x_1-\kappa_5)_+ \\ 1 & x_2 & (x_2-\kappa_1)_+ & \ldots & (x_2-\kappa_5)_+ \\ 1 & x_3 & (x_3-\kappa_1)_+ & \ldots & (x_3-\kappa_5)_+ \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & (x_n-\kappa_1)_+ & \ldots & (x_n-\kappa_5)_+ \end{array}$$

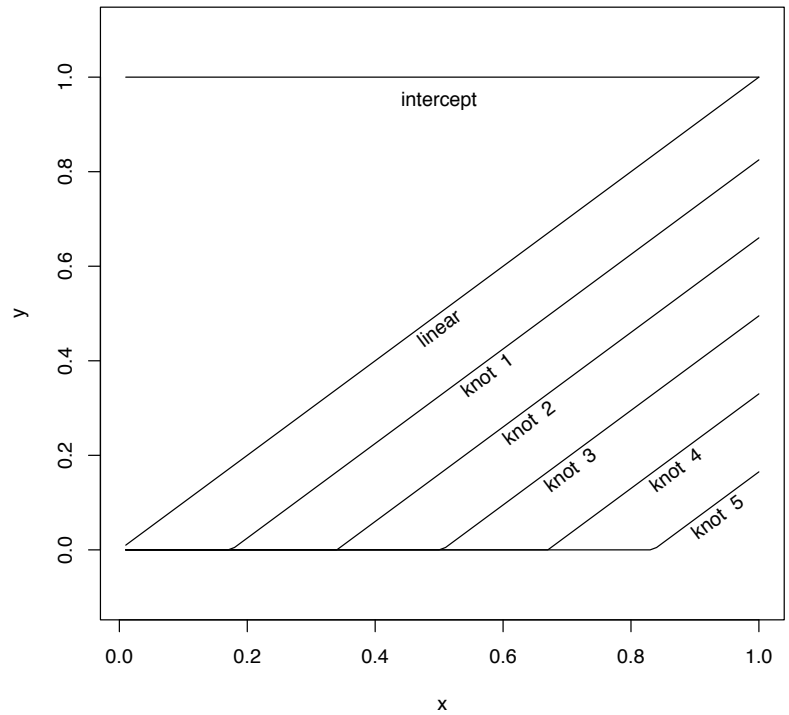
The slope at x_i is $\beta_1 + \sum_{j=1}^{K} \beta_{1j} I(x_i > \kappa_j)$

Simple spline with 5 knots

For the spline with 5 knots: $y_i = \beta_0 + \beta_1 x_i + \sum_{j=1}^{K} \beta_{1j}(x_i - \kappa_j)_+ + [\text{error}]$

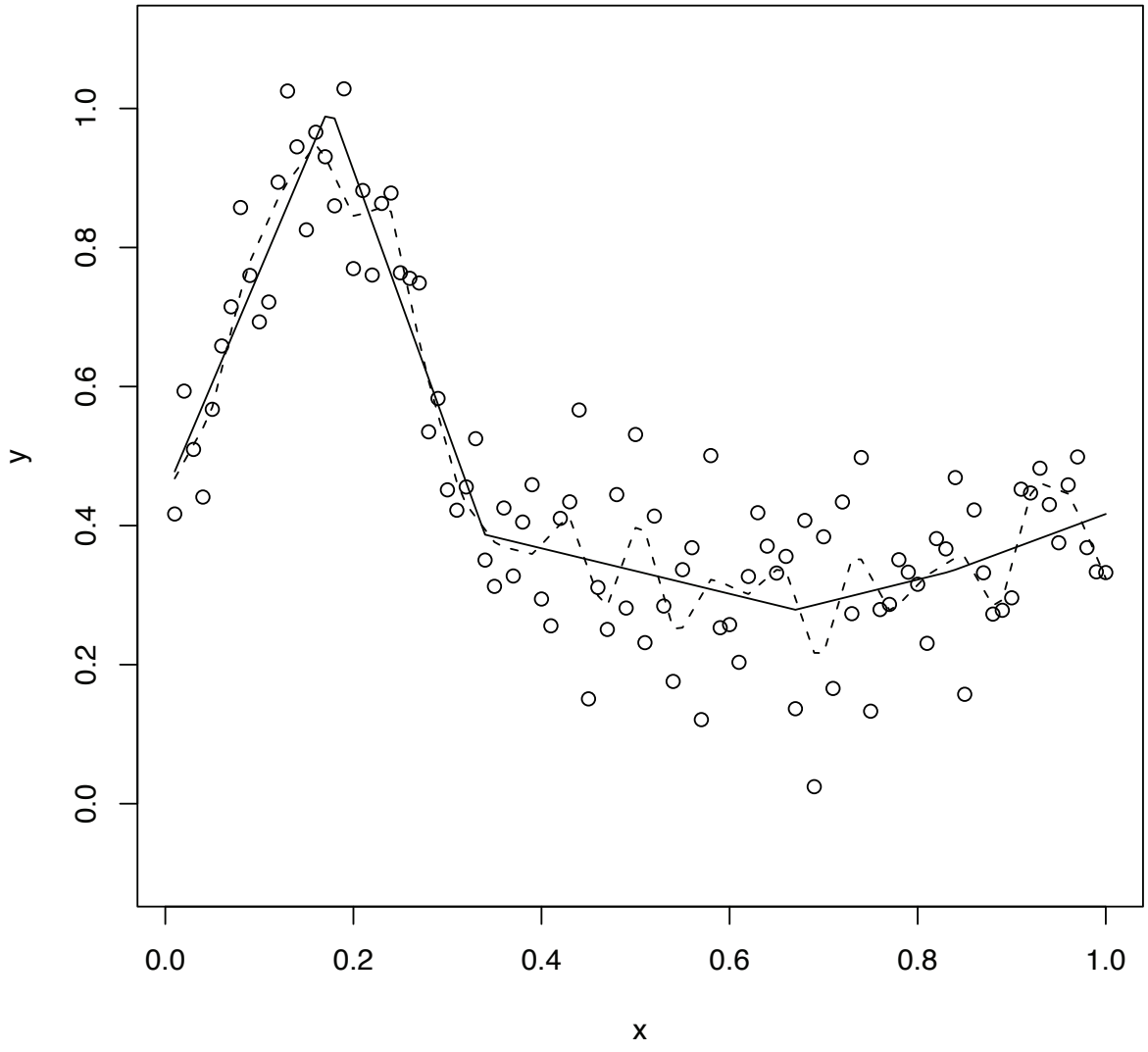
Here's the fitted spline and the basis:





We don't know where the slope *should* change $\Rightarrow$ use more knots

But the dashed fit, with 25 knots, is too wiggly.



Penalized splines compromise: flexible without overfitting

Use many knots and *penalize* changes at the knots:

Instead of this optimization problem:

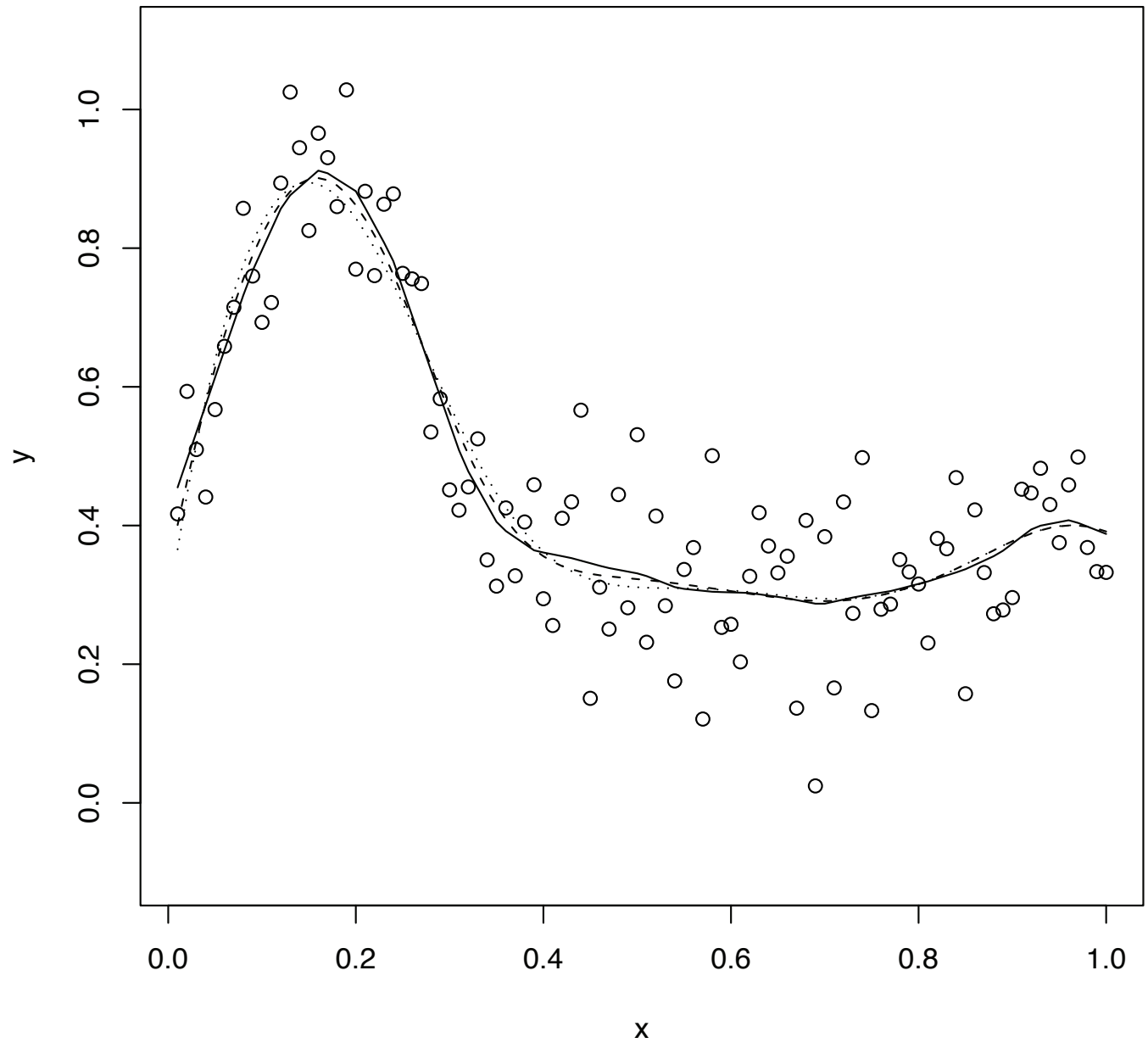
$$\text{choose } (\beta_0, \beta_1, \{\beta_{1j}\}) \text{ to minimize}$$
$$\sum_i [y_i - \beta_0 - \beta_1 x_i - \sum_{j=1}^{K} \beta_{1j}(x_i - \kappa_j)_+]^2$$

solve this optimization problem:

$$\text{choose } (\beta_0, \beta_1, \{\beta_{1j}\}) \text{ to minimize}$$
$$\sum_i [y_i - \beta_0 - \beta_1 x_i - \sum_{j=1}^{K} \beta_{1j}(x_i - \kappa_j)_+]^2 + \lambda \sum_{j=1}^{K} \beta_{1j}^2$$

We've added the penalty term $\lambda \sum_{j=1}^{K} \beta_{1j}^2$, for a positive scalar λ.

25-knot fits with penalized changes at the knots



More bases: Truncated polynomial basis

Truncated p^{th}-degree polynomial basis with knots κ_j has basis vectors corresponding to

$$1, x_i, \ldots, x_i^p, (x_i - \kappa_1)_+^p, \ldots, (x_i - \kappa_K)_+^p$$

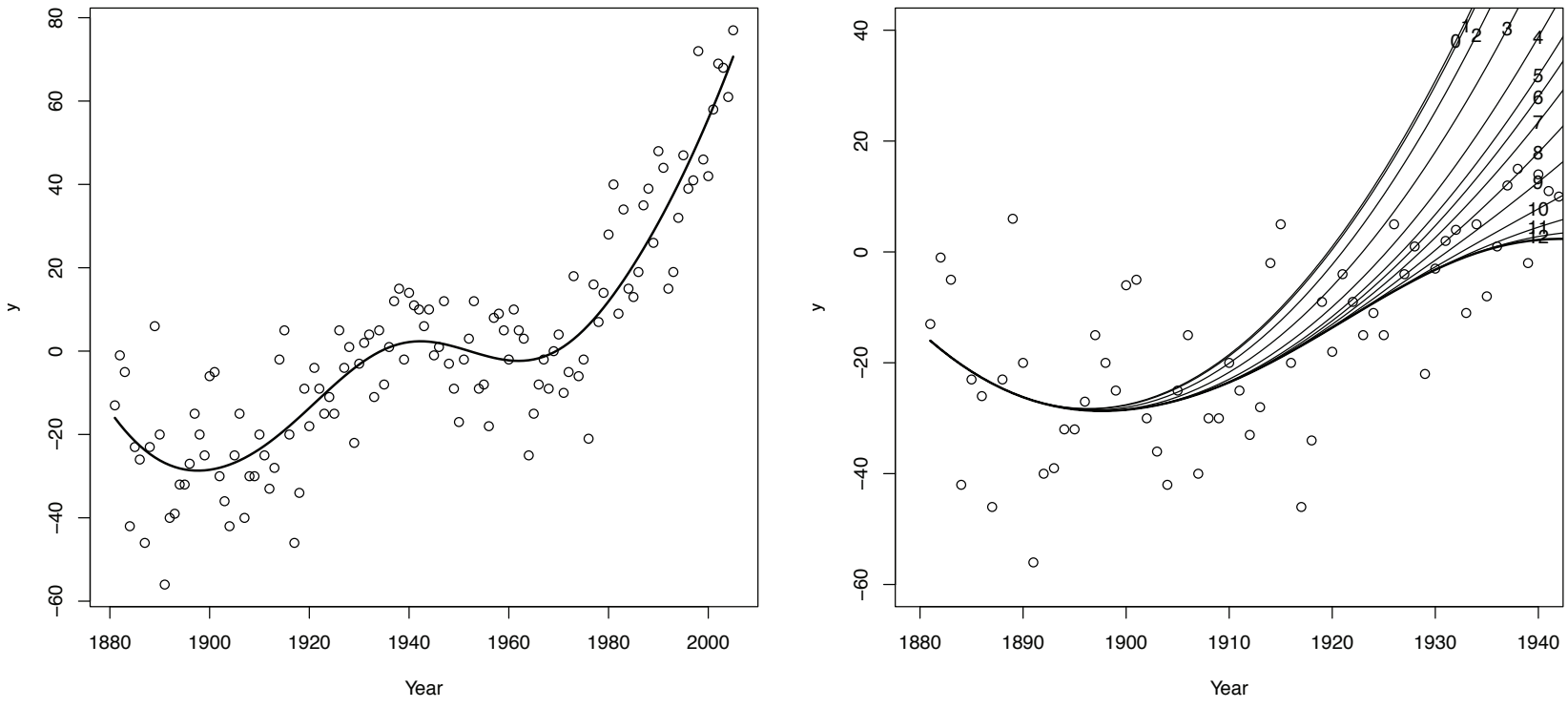
This gives a fit with $p - 1$ derivatives at each knot.

Note: Only the highest-order coefficient changes at the knots.

The previous slide shows fits for linear, quadratic (dashed), and cubic (dotted) bases.

More bases: How the fit changes at the knots

Global mean surface temp data, truncated quadratic basis with 30 knots.

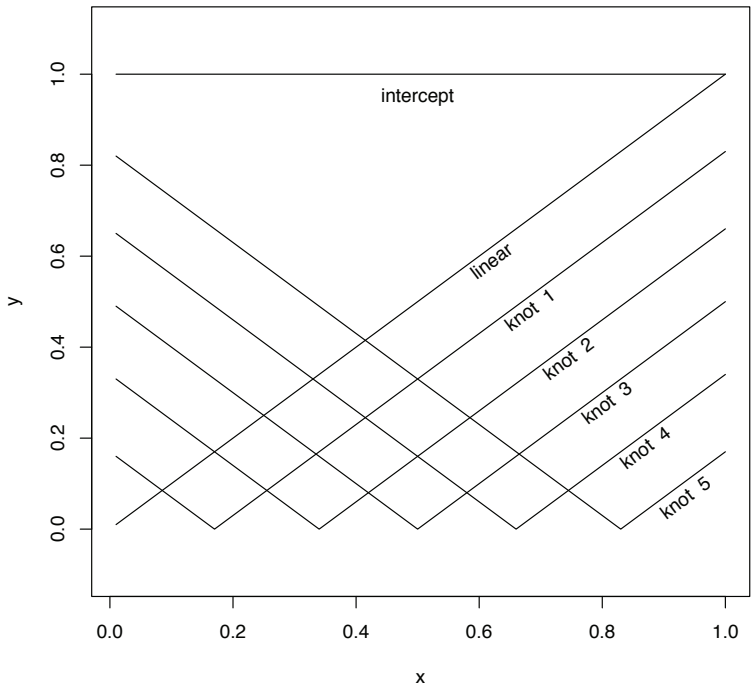


In the right panel, the horizontal axis is only 1880-1940.

More bases: Radial basis

Radial basis of order p and knots κ_j has basis vectors corresponding to

$$1, x_i, \ldots, x_i^p, |x_i - \kappa_1|_+^p, \ldots, |x_i - \kappa_K|_+^p.$$



For p $= 1$, the slope at x_i is $\beta_1 + \sum_j (\beta_{1j} I(x_i > \kappa_j) - \beta_{1j} I(x_i < \kappa_j))$

Advantage: $|x_i - \kappa_j|$ is a function of distance but not direction

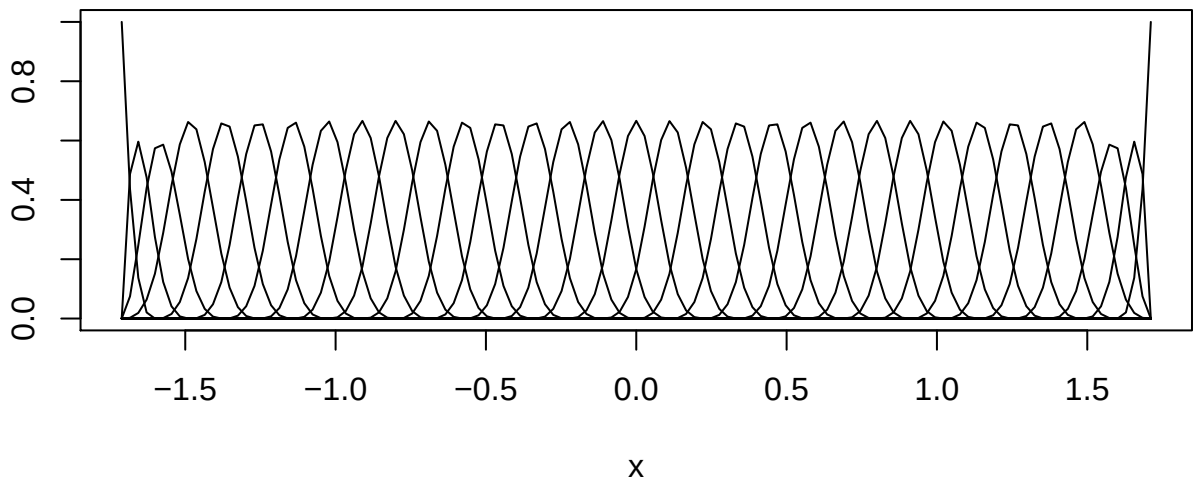
$\Rightarrow$ generalizes readily to > 1 dimension.

More bases: B-spline basis

B-spline basis vectors

- are piecewise polynomials and
- are non-zero only over the span of a few knots, so
- most pairs of basis vectors have inner product 0.

Here are the basis vectors for the GMST data and 30 knots:



The knots are at x values where one of the curves hits zero.

More on penalties

Consider a truncated quadratic basis with knots κ_j and with coefficients $\beta_0, \beta_1, \beta_2$, and $\beta_{21}, \ldots, \beta_{2K}$.

We want to constrain $\beta_{21}, \ldots, \beta_{2K}$ to avoid overfitting.

A penalized spline solves this optimization problem:

$$\text{choose } \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \{\beta_{2j}\}) \text{ to minimize}$$
$$\{\text{lack-of-fit measure}\} \text{ subject to } \{\text{penalty} \leq M\},$$

where "penalty" depends on $\boldsymbol{\beta}$ and M is a positive scalar.

Lots and lots of penalties have been proposed.

For normal-errors normal-RE models, RWC use

- "lack-of-fit measure" $=$ residual sum of squares;
- "penalty" $= \boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta}$ for diagonal $\mathbf{D}$ with 0 or 1 on the diagonal.

For any psd $\mathbf{D}$, a penalized spline solves the optimization problem

choose $\boldsymbol{\beta}$ to minimize $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ subject to $\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta} \leq M$,

where $\mathbf{X}$'s columns contain the spline's basis.

This is equivalent to:

$$\text{choose } \boldsymbol{\beta} \text{ to minimize } (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda^2\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta}.$$

For given λ, the solution to this optimization problem is

$$\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}'\mathbf{X} + \lambda^2 D)^{-1}\mathbf{X}'\mathbf{y},$$

which gives fitted values

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda^2 D)^{-1}\mathbf{X}'\mathbf{y},$$

with $\text{DF}_\lambda = \text{trace}(\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda^2 D)^{-1}\mathbf{X}')$.

Some operational considerations

How should you choose a basis?

- How many derivatives do you want?
- Polynomials are simple and interpretable but have highly collinear design matrices.
- B-splines have nice numerical properties but are more obscure.
- In higher dimensions, radial bases are invariant to rotations.

How many knots should you use and where should you put them?

- RWC default: $K = \min(35, \{\#\text{ unique } x_i\}/4)$, evenly spaced.
- More if the function being estimated has "a lot of fine detail".
- More in intervals of x that have fine detail.

Penalized splines represented as mixed linear models

For a truncated-power basis and RWC's penalty, the model is

$$y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p + \sum_{j=1}^{K} \beta_{pj} (x_i - \kappa_j)_+^p + \text{ [error]}.$$

A squared-error lack-of-fit measure $\Rightarrow$ [error] is $\epsilon_i \overset{iid}{\sim} N(0, \sigma_e^2)$.

Then the model can be formulated as a mixed linear model:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & \dots & x_1^p \\ & & \vdots & \\ 1 & x_n & \dots & x_n^p \end{bmatrix}; \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}; \qquad \mathbf{R} = \sigma_e^2 \mathbf{I}_n;$$

$$\mathbf{Z} = \begin{bmatrix} (x_1 - \kappa_1)_+^p & \dots & (x_1 - \kappa_K)_+^p \\ & \vdots & \\ (x_n - \kappa_1)_+^p & \dots & (x_n - \kappa_K)_+^p \end{bmatrix}; \qquad \mathbf{u}' = \begin{bmatrix} \beta_{p1} \\ \vdots \\ \beta_{pK} \end{bmatrix}.$$

Different basis and knots $\Rightarrow$ different $\mathbf{X}$ and $\mathbf{Z}$.

Mixed linear model representation: What is **G**?

The penalized-spline fitting problem is to choose $(\boldsymbol{\beta}, \mathbf{u})$ that minimizes

$$\frac{1}{\sigma_e^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \frac{\lambda^2}{\sigma_e^2}\mathbf{u}'\mathbf{u},$$

where

- the objective function has been divided by σ_e^2
- the penalty matrix $\mathbf{D}$ is diagonal with $p+1$ diagonal elements of 0 for $\boldsymbol{\beta}$ and K diagonal elements of 1 for $\mathbf{u}$.

This has the same form as the objective function the conventional analysis minimizes to give BLUPs.

This suggests setting $\text{cov}(\mathbf{u}) = \mathbf{G} = \sigma_s^2\mathbf{I}_K$ for $\sigma_s^2 = \sigma_e^2/\lambda^2$, giving

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \mathbf{u}'\mathbf{G}^{-1}\mathbf{u}.$$

Penalized splines now have the *form* of a MLM

The penalized spline's $\mathbf{u}$ is not an old-style random effect:

- $\mathbf{u}$'s "levels" are not a draw from a population; there is no population.
- $\mathbf{u}$ is part of the model's mean, not part of its variance.

The fit is *not*	$\hat{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$ with error $\mathbf{Z}\hat{\mathbf{u}} + \boldsymbol{\epsilon}$;
The fit is	$\hat{y} = \hat{f}(x) = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}}$ with error $\boldsymbol{\epsilon}$.

This has concrete consequences, as we'll see.

RWC (p. 138): "we [use] the mixed-linear-model formulation of penalized splines as a convenient fiction".

"fiction[:] a belief or statement that is false, but that is often held to be true because it is expedient to do so" (*New Oxford American Dictionary*).

How is this "convenient fiction" expedient?

All of MLM theory, methods, and software can be used for splines:

- Select the degree of smoothing by maximizing the RL.
- Compute confidence intervals around the fitted smooth.
- Check model fit.
- [. . . later material in this course, which you've never heard of.]

RWC (p. 138, p. 177): "[W]e used the mixed model formulation of penalized splines as a convenient fiction to estimate smoothing parameters. The mixed model is a reasonable (though not compelling) Bayesian prior for a smooth curve, and . . . [maximized RL] estimates of variance components give estimates of the smoothing parameter that generally behave well. [. . .] Although we are attracted by the automatic nature of the mixed model-REML approach to fitting additive models, we discourage blind acceptance of whatever answer it provides and recommend looking at other amounts of smoothing".

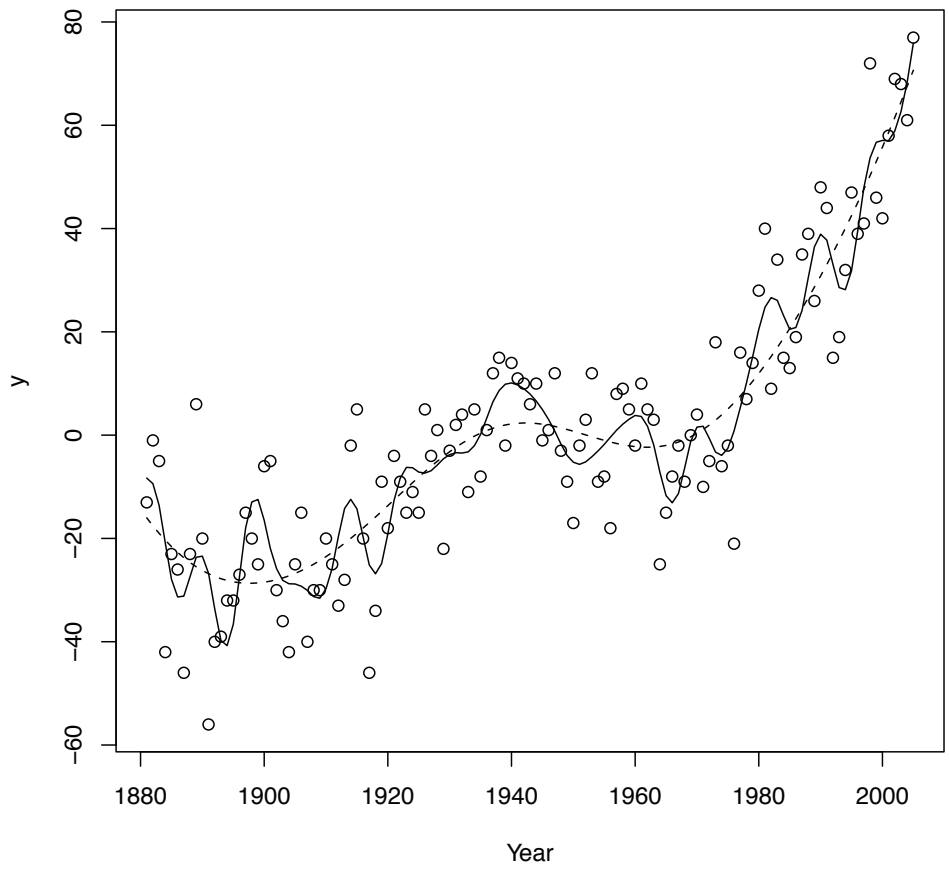
An example, and a quandary that it illustrates

Global mean surface temp (GMST), 1881 to 2005 (downloaded 2006),

$$y_t, t = 0, \ldots, 124, \text{ units } 0.01^\circ \text{ C}.$$

The dashed line is a penalized-spline fit, quadratic basis, 30 knots.

The solid line is the spline fit with $1/\sigma_s^2 = 0$, i.e., no penalty.



Fitting the penalized spline

The smooth dashed line was fit by maximizing the RL:

- $\hat{\sigma}_e^2 = 145$, so the error SD ≈ 12
- $\hat{\sigma}_s^2 = 947$, which is uninterpretable.
- The fit has 6.7 DF: fixed effects 3 DF, random effects 3.7 DF out of a possible 30.

Maximizing the RL gives a reasonable smooth here.

The wiggly line is the unpenalized fit. It's grossly overfit.

The next slide shows coefficient estimates and SEs from both fits.

The unpenalized SEs are huge because of the extreme collinearity.

	Unpenalized fit		Penalized fit	
	Estimate	SE	Estimate	SE
Intercept	–5924.8	10661.9	85.33	131.12
Linear	–6930.8	12906.3	176.37	174.76
Quadratic	–2029.7	3903.7	68.45	58.67
u1	4387.7	5258.6	–0.53	30.77
u2	–4688.3	2817.7	–3.48	30.68
u3	5679.9	2519.4	–6.61	30.35
u4	–6310.3	2465.8	–8.41	29.78
u5	4015.4	2455.8	–5.68	29.14
u6	–1305.9	2453.9	–3.93	28.62
u7	1662.1	2453.5	–5.80	28.31
u8	–3786.6	2453.5	–9.19	28.17
u9	4810.5	2453.5	–12.12	28.13
u10	–3868.4	2453.4	–15.85	28.13
u11	1859.5	2453.4	–16.95	28.12
u12	–704.7	2453.4	–15.36	28.12
u13	921.5	2453.4	–11.43	28.11
u14	–1280.9	2453.4	–4.55	28.10
u15	430.3	2453.4	5.42	28.10
u16	19.8	2453.4	14.97	28.10
u17	776.0	2453.4	20.24	28.10
u18	–500.1	2453.4	20.76	28.11
u19	–222.0	2453.4	19.35	28.11
u20	–977.6	2453.4	17.33	28.12
u21	3185.8	2453.4	12.59	28.13
u22	–3738.0	2453.4	4.26	28.13
u23	2975.1	2453.5	–3.45	28.14

A closer look at the $\hat{u}_j$

The estimated (or predicted) $\hat{u}_j$ don't look anything like an iid sample from any distribution.

The iid Normal assumption on $\mathbf{u}$ simply constrains the fit.

Other parts of this model are "wrong" too, e.g., iid errors.

Maximizing the RL is an effective way to draw a smooth curve thru these data.

But now we get into something that's not so clear ...

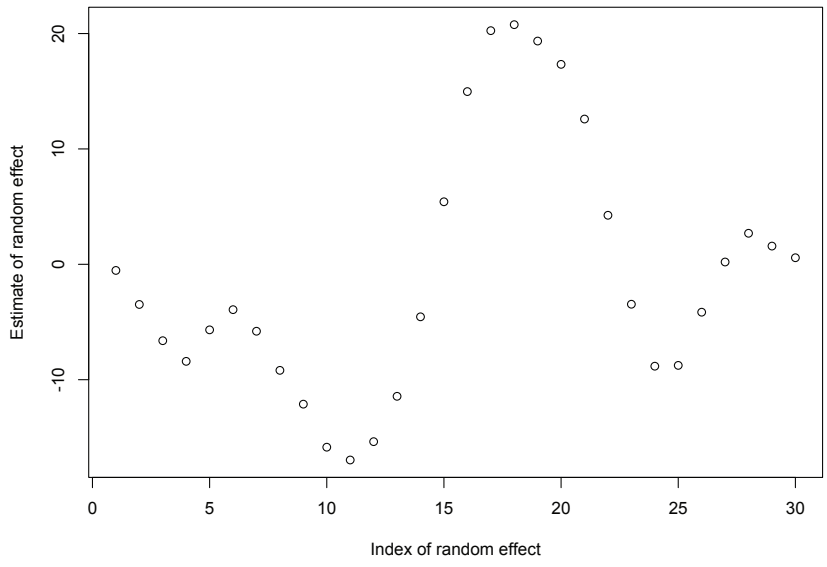


Figure: The estimated u_j

How do we draw a confidence band around the fit?

Specifically: Should you condition on the true, unknown $\mathbf{u}$?

Yes: The model is $\mathbf{y} = \mathbf{f} + \boldsymbol{\epsilon}$ for $= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$, and

- $\mathbf{f}$ is a fixed but unknown smooth function
- $\mathbf{u} \sim$ iid $N(0, \sigma_s^2)$ is just a device to shrink $\hat{\mathbf{u}}$ and give a smooth fit.
- $\mathbf{u}$ is not random, so we should treat it like any other parameter, as fixed and unknown.

No, don't condition on $\mathbf{u}$:

- (a) $\mathbf{u}$ is a random draw; in frequentist theory, we don't condition on randomly-drawn things.
- (b) Conditioning creates bias:
  - Conditional on $\mathbf{u}$, $\hat{f}(x_0)|\mathbf{u}$ is biased.
  - Unconditionally, $\hat{f}(x_0)$ is unbiased.
  - To avoid bias, we must use the unconditional approach.

Let's follow the argument as RWC present it

Recall $\mathbf{C} = [\mathbf{X}|\mathbf{Z}]$; its i^{th} row is $\mathbf{C}_i = [\mathbf{X}_i|\mathbf{Z}_i]$.

$\hat{f}(x_i) = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\mathbf{u}}$; hats indicate use of the REML estimates.

$$\text{var}\{\hat{f}(x_i)|\mathbf{u}\} = \mathbf{C}_i\text{cov}([\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}]'|\mathbf{u})\mathbf{C}_i',$$

where $\text{cov}([\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}]'|\mathbf{u})$ is a function of σ_e^2 and σ_s^2 (but not a function of $\mathbf{u}$ despite the conditioning).

$\Rightarrow$ Conditional 95% CI: $\hat{f}(x_i) \pm 1.96 \times \text{SD}\{\hat{f}(x_i)|\mathbf{u}\}$.

RWC: "If there is no appreciable bias" in the spline fit, "then $E[\hat{f}(x)|\mathbf{u}] \approx f(x)$, and this interval can be interpreted as a confidence interval for $f(x)$."

But the MLM framework allows an estimate of the bias conditional on $\mathbf{u}$:

$$\begin{aligned} E[\hat{f}(x) - f(x)|\mathbf{u}] &= E[\, \mathbf{C} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} |\mathbf{u}\,] - \mathbf{C} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{pmatrix} \\ &= -r\mathbf{C}(\mathbf{C}'\mathbf{C} + r\mathbf{I})^{-1} \begin{pmatrix} \mathbf{0}_p \\ \mathbf{u} \end{pmatrix}, \end{aligned}$$

where $r = \sigma_e^2/\sigma_s^2$, with larger values implying a smoother fit.

Bias: All smoothers fill valleys and round off peaks and corners.

Thus, the conditional CI is not a CI for $f(x)$ in general

- because its coverage is too low for x where bias is "appreciable",
- because it has the wrong center for those x.

RWC: "But, since $E(\mathbf{u}) = \mathbf{0}$, the unconditional bias [i.e., averaging the bias over $\mathbf{u}$'s distribution] is $E[\hat{f}(x) - f(x)] = 0$.

Thus, on average over the distribution of $\mathbf{u}$, $\hat{f}(x)$ is unbiased for $f(x)$.

To account for bias in the confidence intervals, the [conditional] variance $\text{var}\{\hat{f}(x)|\mathbf{u}\}$ should be replaced by the conditional mean-squared error $E[\{\hat{f}(x) - f(x)\}^2|\mathbf{u}]$ ... and then averaged over the $\mathbf{u}$ distribution."

This gives

$$E[\{\hat{f}(x_i) - f(x_i)\}^2] = \mathbf{C}_i \text{cov}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} - \mathbf{u} \end{pmatrix} \mathbf{C}_i' > \text{var}\{\hat{f}(x_i)|\mathbf{u}\}.$$

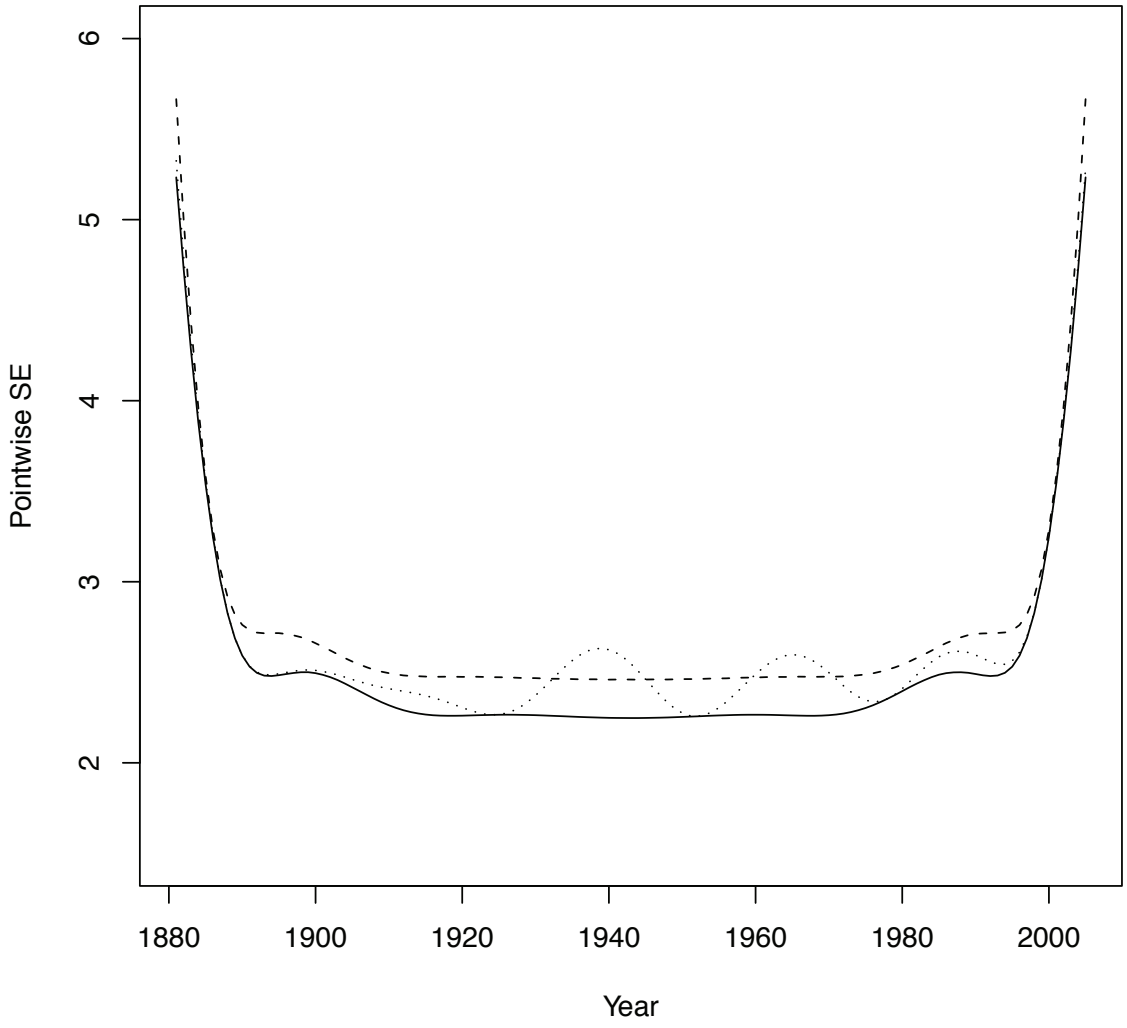
$\Rightarrow$ Unconditional 95% CI: $\hat{f}(x_i) \pm 1.96 \times E[\{\hat{f}(x_i) - f(x_i)\}^2]^{0.5}$,

which is wider than the conditional CI.

What does this actually do?

Solid: Conditional SD Dashed: Unconditional SD

Dotted: $\sqrt{\widehat{\text{conditional variance}} + \widehat{\text{bias}}^2}$



The unconditional SE doesn't address the problem

The unconditional SE is too small at x where bias is largest and too large where bias is small.

Nychka (1988): the unconditional pointwise CI gives 95% coverage *averaging over the values of the predictor*.

If $\mathbf{u}$ is understood as a fixed but unknown quantity, the unconditional SE is a *non sequitur*.

So don't widen the interval; move its center.

- Hodges (2013): Shift the center by the bias. (Works badly.)
- Dai (class project): Center the CI by re-fitting the spline with less smoothing; use the variance estimate of the corrected fit.

Other aspects of analyzing the penalized spline as an MLM

You can (RWC Sections 6.6 to 6.9):

- Compute simultaneous (as opposed to pointwise) confidence bands.
- Test whether the random-effect part of the fit should be zero. (Remember: The asymptotics are poor.)
- Test for no effect of the predictor.
- Do inference about derivatives (as the spline basis permits).
- Test for the existence of a feature.

You can also do a Bayesian analysis of this MLM:

- The analysis is a straightforward application of Bayesian methods, especially if you use MCMC.
- Re the puzzle above: Bayes offers no conceptual or practical advantage.