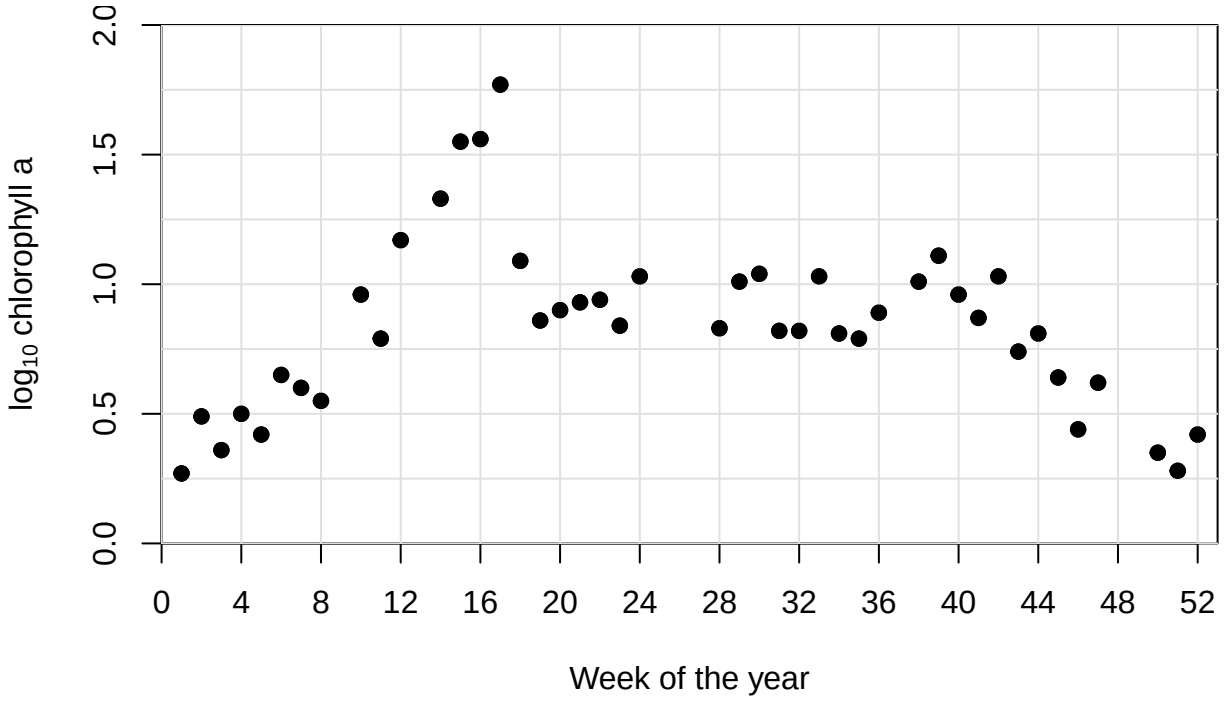
NRES 746, Week 13, Lecture 2: Spline by Hand (take-home)

Name: _____

Due Monday, Nov. 23. Work on your own. You'll need R for Parts 2-4; try Part 1 on paper first.

A water utility samples a reservoir every week for chlorophyll *a*, an index of algal biomass. Ice and wind cancelled 8 of last year's 52 visits. The data are $\log_{10}$ chlorophyll *a* (μg/L) by week of the year. The utility wants to know when the spring algal bloom peaks and how big it gets, so it can time its treatment.



Part 1: Your guess (on paper)

1a. On the figure above, sketch the smooth seasonal curve you think lies under the noise.

1b. From your sketch, the spring bloom peaks in week _____ at $\log_{10}$ chlorophyll _____.

Part 2: Splines without a penalty

Each coding question has a label:

- *Recommended: no AI*: this week's core skill. R's help pages, web search, and classmates are fine.
- *AI optional/encouraged*: the results matter, but the code isn't the point.

After each coding question, circle how you used AI.

What to turn in: your filled-in answers and the three named figures (2b, 3c, 4a) as images on WebCampus. You don't need to submit code.

Read the data with

```
chl <- read.csv(
  "https://kevintshoemaker.github.io/NRES-746/data/ReservoirChlorophyll.csv")
```

or paste them in:

```
week <- c(1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24,
          28, 29, 30, 31, 32, 33, 34, 35, 36, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 50,
          51, 52)
```

```
log_chl <- c(0.27, 0.49, 0.36, 0.5, 0.42, 0.65, 0.6, 0.55, 0.96, 0.79, 1.17, 1.33, 1.55,
             1.56, 1.77, 1.09, 0.86, 0.9, 0.93, 0.94, 0.84, 1.03, 0.83, 1.01, 1.04, 0.82,
             0.82, 1.03, 0.81, 0.79, 0.89, 1.01, 1.11, 0.96, 0.87, 1.03, 0.74, 0.81, 0.64,
             0.44, 0.62, 0.35, 0.28, 0.42)
chl <- data.frame(week, log_chl)
```

2a. *(AI optional/encouraged.)* Fit `lm(log_chl ~ bs(week, df = K), data = chl)` for each number of basis functions K below (`bs()` is in the `splines` package, which comes with R). Record each fit's AIC, and say whether its curve looks too smooth, about right, or too wiggly.

K	AIC	Too smooth, about right, or too wiggly?
4		
8		
12		
20		

AI use (circle one): no AI / AI helped (e.g., debugging) / AI wrote the code

2b. *(AI optional/encouraged.)* Plot the data with the fitted curves for $K = 4$ and $K = 20$. Name it **Figure 2b**.

AI use (circle one): no AI / AI helped (e.g., debugging) / AI wrote the code

Part 3: A penalized spline, by hand

Use 20 cubic B-spline basis functions, with the intercept built into the basis:

```
library(splines)
B <- bs(chl$week, df = 20, intercept = TRUE)     # 44 x 20 model matrix
```

The model is $y_i \sim \text{Normal}(\mu_i, \sigma)$ with $\mu_i = \sum_{k=1}^{20} \beta_k B_k(\text{week}_i)$ (the vector of all μ_i is `B %*% beta`), and the penalized NLL is

$$\text{NLL}_{\text{pen}} = -\sum_{i=1}^{44} \log p(y_i \mid \mu_i, \sigma) + \lambda \sum_{k=3}^{20} (\beta_k - 2\beta_{k-1} + \beta_{k-2})^2$$

3a. *(Recommended: no AI.)* Write a function `pnll(par, lambda)` that returns NLL_{pen}, where `par` holds the 20 weights followed by $\log \sigma$. Check it: `pnll(c(rep(1, 20), log(0.2)), lambda = 10)` should return 43.72.

Your value: _____

AI use (circle one): no AI / AI helped (e.g., debugging) / AI wrote the code

3b. *(Recommended: no AI.)* Minimize `pnll()` with `optim()` for each λ below, starting from `c(rep(1, 20), log(0.2))`. Use `method = "BFGS"` and `control = list(maxit = 2000)`, and check `$convergence`. Record $\hat{\sigma}$ and the penalty term $\lambda \sum (\beta_k - 2\beta_{k-1} + \beta_{k-2})^2$ at the optimum.

λ	$\hat{\sigma}$	**penalty term at the optimum**
0		
10		
1000		

AI use (circle one): no AI / AI helped (e.g., debugging) / AI wrote the code

3c. *(AI optional/encouraged.)* Plot the data with your three fitted curves (`B_grid <- predict(B, newx = 1:52)` gives the basis at every week). Name it **Figure 3c**.

AI use (circle one): no AI / AI helped (e.g., debugging) / AI wrote the code

3d. Which λ would you use? Why does $\hat{\sigma}$ change the way it does as λ grows?

Part 4: The same model in mgcv

4a. *(AI optional/encouraged.)* Fit `gam(log_chl ~ s(week, bs = "ps", k = 20), data = chl, method = "REML")` (package `mgcv`), which uses the same kind of basis and penalty and chooses λ itself. Plot its curve with your chosen fit from 3d. Name it **Figure 4a**.

Effective degrees of freedom (edf): _____ $\hat{\sigma}$: _____

AI use (circle one): no AI / AI helped (e.g., debugging) / AI wrote the code

4b. When does the spring bloom peak, and at what level? Compare the `gam()` curve with your sketch from 1b.

`gam()` peak: week _____ at $\log_{10}$ chlorophyll _____

4c. Where did you use AI, if anywhere, and what did you have to check or fix? If you didn't use it, what was hardest to get working?

Bonus (optional). Refit your penalized spline with $\lambda = 10^6$ and compare its curve with `lm(log_chl ~ week, data = chl)`. Why does a huge λ give this curve?