

Lab 8, Act 2: Shrinking the Coefficients

What changes when all eight pond features compete in one model?

NRES 746, name: _____ group: _____

The scenario

In Act 1 you compared seven hand-picked models for Bd prevalence in the 60-pond survey. With eight candidate pond features there are $2^8 = 256$ possible subsets, and AIC couldn't separate temperature from elevation. Another approach is to keep all eight features in one model and **penalize** the size of their coefficients, so a feature keeps a large coefficient only if the data insist on it.

What happens to each pond feature's coefficient when all eight compete, and what can we tell the managers?

Try Part 1 on paper first, before you open a laptop or start typing code.

Part 1: Plan it on paper (~15 min)

1a. Write the ridge-penalized NLL for the logistic-binomial model with all eight standardized predictors, as a function of $\beta_0, \beta_1, \dots, \beta_8$ and the penalty λ . Should β_0 be penalized? Why or why not?

1b. Why standardize the predictors before penalizing them?

1c. In plain English (not R code), describe step by step how you would choose λ for your ridge model by 5-fold cross-validation, scoring each held-out fold by its NLL.

Checkpoint: finalize your plan (~5 min)

Look back over 1a-1c and fix anything that no longer looks right before you start coding.

Part 2: Fit it (laptop, ~40 min)

Using AI: each question has a recommendation. *Recommended:* *no AI* marks the core skill of the lab (writing a penalized likelihood and fitting it); R's help pages, web search, and classmates are fine. *AI optional* marks work where the results matter but writing the code by hand isn't the point. After each question, circle how you actually used AI.

What to turn in: your filled-in answers and your 2b figure as an image (upload in WebCampus). You don't need to submit your code.

The data are in `BdRegionalSurvey.csv` on the course website: <https://kevintshoemaker.github.io/NRES-746/data/BdRegionalSurvey.csv>. Read it and standardize the eight predictors (columns 2 to 9):

```
data_url <-  
  "https://kevintshoemaker.github.io/NRES-746/data/BdRegionalSurvey.csv"  
ponds <- read.csv(data_url)  
X <- scale(as.matrix(ponds[, 2:9]))
```

2a. (*Recommended: no AI.*) Write a function for your penalized NLL from 1a and fit it with `optim(..., method = "BFGS")` at $\lambda = 0$ and $\lambda = 10$. Check: the temperature coefficient should be about -0.64 at $\lambda = 0$ (the same as `glm()`) and -0.41 at $\lambda = 10$. Fill in those two columns of the table.

Predictor	MLE (lambda = 0)	Ridge (lambda = 10)	Lasso (2c)
water temp.	-0.64	-0.41	
elevation			
canopy			
conductivity			
area			
fish			
bullfrogs			
road distance			

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

2b. (*AI optional.*) Fit your ridge model at a grid of λ values from 0 to 100 (say 0, 1, 2, 5, 10, 20, 50, 100), and record all eight slope estimates at each one. Then make a **ridge path**: one plot with λ on the x-axis, the coefficient estimate on the y-axis, and one line per predictor, so you can follow how each estimate changes as the penalty grows. Mark $\lambda = 10$. Turn in this figure.

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

2c. (*AI optional.*) Fit the lasso with the `glmnet` package: `cv.glmnet()` with `family = "binomial"` and `nfolds = 10`, run right after `set.seed(1)`. For binomial counts, `glmnet` expects the response as `cbind(failures, successes)`. Fill in the lasso column with the coefficients at `lambda.1se` (write 0 for any it dropped).

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

2d. 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?

Part 3: Interpret (~20 min)

3a. Temperature and elevation are strongly correlated ($r = -0.94$): on average, a pond 1 SD warmer than average is also about 0.94 SD lower. Here are two ways to divide their shared effect:

- Option A: $\beta_{temp} = -0.6$, $\beta_{elev} = 0$
- Option B: $\beta_{temp} = -0.3$, $\beta_{elev} = 0.3$

For a pond 1 SD warmer and 0.94 SD lower than average, how much does each option change the log-odds of infection? What does each option add to the ridge penalty ($\beta_{temp}^2 + \beta_{elev}^2$)? Use your answers to explain what the elevation line does in your 2b figure, and why the other six predictors behave differently.

3b. Compute the lasso penalty, $|\beta_{temp}| + |\beta_{elev}|$, for Options A and B. Does the lasso reward splitting the way ridge does? Your lasso kept both predictors. The table shows `glm()` fits with and without elevation. Why does adding elevation triple temperature’s standard error, and how much would you trust either method’s split between the two?

Model	Temperature (SE)	Elevation (SE)
temp + bullfrogs + canopy	-0.74 (0.08)	not in model
temp + elevation + bullfrogs + canopy	-0.54 (0.24)	0.21 (0.24)

Reflection (~10 min)

The managers asked which pond features predict where Bd is common. Write two or three sentences they could act on, including what these data can and can’t tell them about temperature versus elevation.

Also record any changes you made, new insights, and any “muddiest points” or topics you want to review.

If you finish early. (*Recommended: no AI.*) Code your cross-validation plan from 1c and use it to choose λ for your ridge model from a grid of values. Which λ did you pick?

$\lambda =$ _____

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