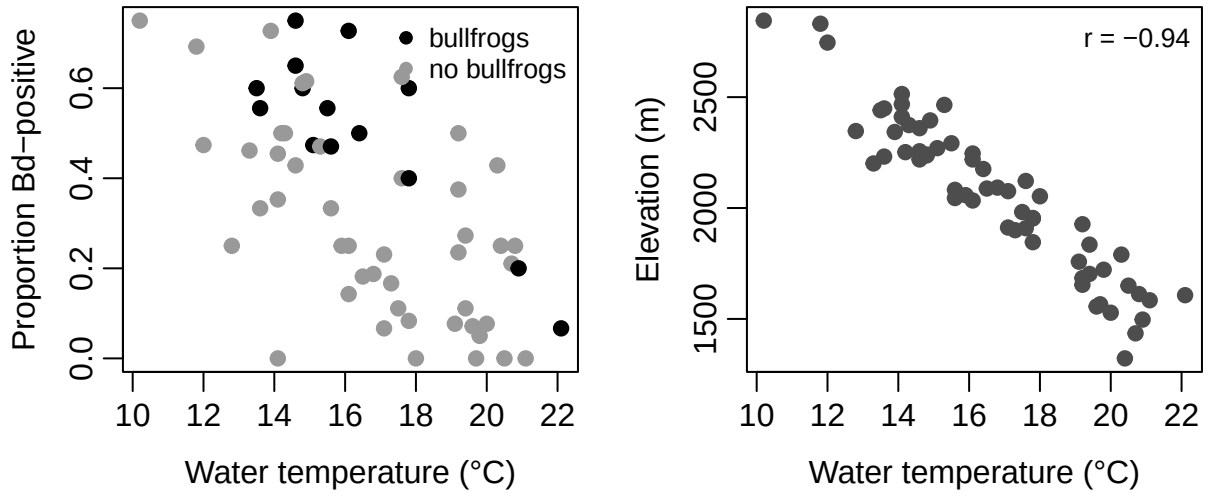
Lab 8, Act 1: Which Pond Features Matter?

Which pond features predict chytrid, and can the data tell temperature from elevation?

NRES 746, name: _____ group: _____

The scenario

The summer after the 12-pond survey from Lab 7, the field crew swabbed frogs at **60 ponds** across the region and recorded eight features of each pond: water temperature, elevation, canopy cover, conductivity, pond area, whether fish are present, whether bullfrogs are present (an invasive species that carries Bd without dying from it), and distance to the nearest road. Managers can afford to monitor only some ponds and want to know which features predict where Bd is common.



The model for every candidate is the logistic-binomial model from Lab 7, with more predictors: the number of Bd-positive frogs at pond i is Binomial(n_i, p_i), with

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$$

Every predictor was **standardized** (mean 0, SD 1) before fitting, so each β is the change in log-odds of infection per standard deviation of that predictor.

Which pond features predict Bd prevalence?

Please keep laptops closed for this Act.

Part 1: Comparing candidate models (individual, ~20 min)

The table lists seven candidate models, the number of parameters k (including β_0), and each model's negative log-likelihood (NLL) at its MLE.

1a. Fill in the missing AIC values ($\text{AIC} = 2\,\text{NLL} + 2k$) and every ΔAIC.

Model	Predictors	k	NLL	AIC	dAIC
1	intercept only	1	179.40	360.80	
2	temp	2	137.42		
3	elevation	2	137.73		
4	temp + elevation	3	136.39	278.78	
5	temp + bullfrogs + canopy	4	106.58		
6	temp + elev. + bullfrogs + canopy	5	106.20	222.40	
7	all eight	9	104.59		

1b. Which model is best by AIC? Model 6 adds elevation to Model 5 and sits within 2 AIC units of it. Does that mean the data support an effect of elevation? Does the difference in the two models' raw NLLs help you answer this question, and if so, how?

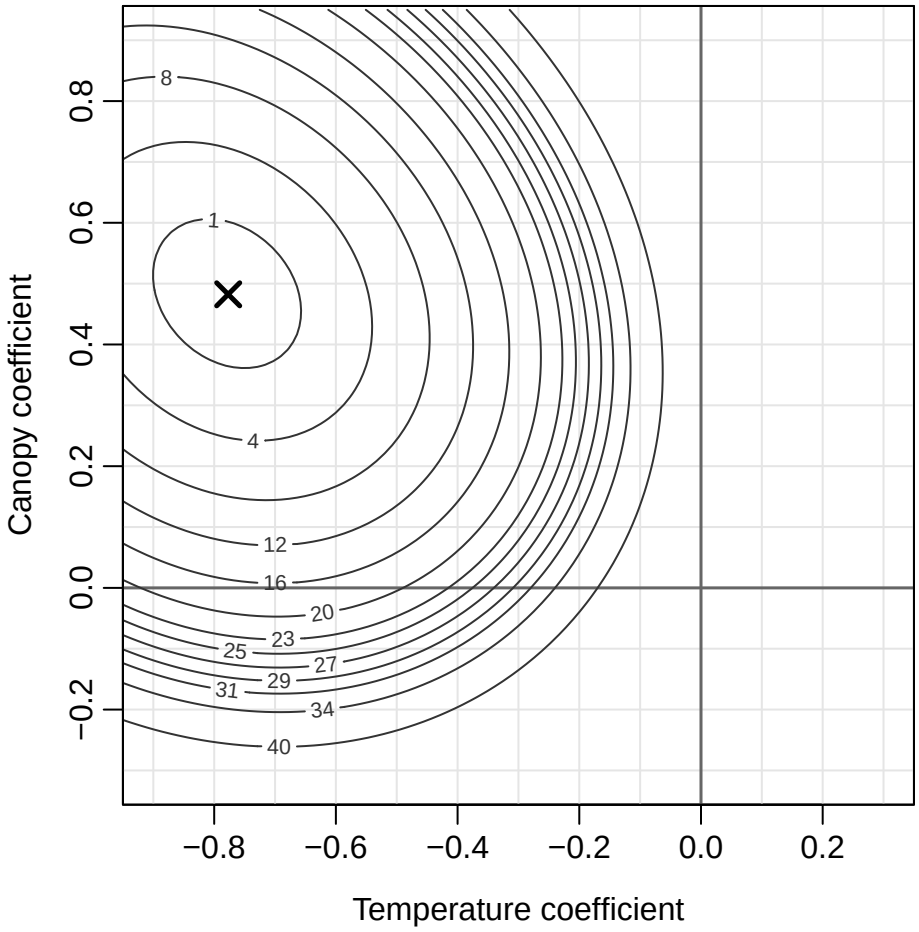
1c. Use likelihood ratio tests to compare Model 5 with Model 6, and Model 5 with Model 7. For each, give the test statistic, its degrees of freedom, and your conclusion. Critical values of χ^2 at the 0.05 level: 3.84 (1 df), 5.99 (2), 7.81 (3), 9.49 (4), 11.07 (5).

1d. Models 2 and 3 are nearly tied. Can these data tell whether water temperature or elevation is what drives Bd? Why or why not? What kind of ponds, or what kind of study, would help?

Part 2: Penalties as budgets (individual, ~15 min)

A ridge penalty, $\text{NLL}(\beta) + \lambda \sum_j \beta_j^2$, gives the same estimate as finding the lowest-NLL point inside a **budget** $\sum_j \beta_j^2 \le t^2$, for some t that shrinks as λ grows. The lasso budget is $\sum_j |\beta_j| \le t$. Neither budget includes β_0.

The figure shows NLL contours for a model with only two predictors, water temperature and canopy cover, labeled by how far each lies above the minimum. The × marks the MLE.



2a. On the figure, draw the ridge budget $\beta_{temp}^2 + \beta_{canopy}^2 \le 0.3^2$ and the lasso budget $|\beta_{temp}| + |\beta_{canopy}| \le 0.3$. Mark the ridge estimate and the lasso estimate, and read off their approximate coordinates.

Ridge: (_____ , _____) Lasso: (_____ , _____)

2b. Did either penalty set a coefficient to exactly 0? Could a ridge budget ever set a coefficient to exactly 0? Use the shapes in your sketch to explain.

2c. In general, when does the lasso set a coefficient to exactly 0? Describe what the contours and the budget have to look like.

Group discussion and revision (~10 min)

Compare your answers with your group, especially 1b, 1d, and your sketch in 2a, and revise as needed. Record any changes you made, new insights, and any "muddiest points" or topics you want to review.